Statistical Science
1993, Vol. 8, No. 2, 82–119

Backcalculation of HIV Infection Rates

Peter Bacchetti, Mark R. Segal and Nicholas P. Jewell

Abstract. Backcalculation is an important method for reconstructing past rates of human immunodeficiency virus (HIV) infection and for estimating current prevalence of HIV infection and future incidence of acquired immunodeficiency syndrome (AIDS). This paper reviews the backcalculation technique, focusing on the key assumptions of the method, including the necessary information regarding incubation, reporting delay, and models for the infection curve. A summary is given of the extent to which the appropriate external information is available and whether checks of the relevant assumptions are possible through use of data on AIDS incidence from surveillance systems. A likelihood approach to backcalculation is described and implemented on AIDS incidence data in the United States. New features of the approach include incorporation of seasonal variation in diagnosis rates, smooth nonparametric estimation of both the HIV infection curve and nonstationary aspects of the incubation period and reporting delay distributions, and an analysis of residuals from backcalculation fits. Unexplained lack of fit is examined and discussed. A fundamental concern is the appropriate acknowledgment of uncertainty associated with backcalculation estimates caused by misspecified assumptions and inaccurate external estimates of key components of the technique. Such uncertainty limits the usefulness of backcalculation and highlights the need for complementary approaches.

Key words and phrases: Acquired immunodeficiency syndrome (AIDS), HIV, backcalculation, deconvolution, seasonality, overdispersion, projections.

1. INTRODUCTION AND OVERVIEW

Throughout the AIDS epidemic, there has been substantial interest in understanding its current state and predicting its future path. Specifically, attention has focused on estimating the current prevalence of HIV infection in populations and projecting the number of AIDS cases that will occur in future time periods, together with assessment of the uncertainty of such estimates. These quantities are of substantial concern to policymakers, health care systems and epidemiologists. Statisticians have evaluated the data sources available for prevalence and projection estimates and have developed estimation methods based on such data.

Backcalculation is the main method used to reconstruct the past pattern of HIV infections in the United States, and it is also widely used to predict future numbers of AIDS cases (Brookmeyer and Gail, 1986, 1988; Brookmeyer, 1991; Rosenberg et al., 1991a). The method uses an assumed known distribution for the incubation period in a given population to deduce the numbers of HIV infections that must have occurred in the past to give rise to the observed pattern of AIDS cases over time. Because the incubation period is defined as the time between infection and diagnosis, the distribution of diagnoses is the convolution of the infection and incubation distributions, as described formally in Section 2.1. Backcalculation exploits this relationship and thus depends on three key components: the assumed incubation distribution; the observed counts of AIDS diagnoses over time; and a model for the distribution of infections.

Backcalculation is known to be sensitive to the incubation period used (Brookmeyer and Gail, 1988), but

Peter Bacchetti and Mark R. Segal are Assistant Professors in the Department of Epidemiology and Biostatistics, University of California, San Francisco, California 94143-1347. Nicholas P. Jewell is a Professor in the Program in Biostatistics and Department of Statistics, University of California, Berkeley, California 94720.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *Statistical Science.*
www.jstor.org

it is usually assumed that adequate data are available for accurate estimation and that incubation does not vary across populations (Brookmeyer, 1991; Rosenberg et al., 1991a). In fact, incubation data are limited by the fact that exact infection times are usually unknown or are known only for highly selected groups of individuals. Many new methods have been developed specifically to estimate incubation distributions while accounting for the complex censoring and truncation patterns often found in AIDS data. In Section 2.2, we review the data and methods that have been used, along with methods that allow for lengthening of incubations due to recent introduction of effective preventive therapies. We also discuss the difficulty of formulating sensitivity analyses that adequately span the range of uncertainty.

The observed diagnosis counts are subject to some incompleteness, which must be accounted for if backcalculation is to produce sensible estimates. Section 2.3 discusses AIDS surveillance in the United States and its two sources of incompleteness: cases who are never reported to the system; and reporting delays, which affect recent counts. Section 2.4 discusses the final component, models for the infection distribution and reviews the parametric and nonparametric approaches that have been used. We conclude the review of backcalculation by briefly discussing other approaches that have been tried or proposed for estimating past HIV infection rates and projecting future AIDS counts.

To examine further some key aspects of backcalculation, we develop a specific implementation of the backcalculation strategy, as described in Section 3. Our method has the same basic structure as those used extensively before, and we use maximum penalized likelihood to obtain smooth estimates of the HIV infection curve. New features of the method include modeling seasonal patterns in diagnoses and smooth secular trends in incubation or reporting delay. We also discuss methods for choosing the amount of smoothness of the nonparametric components and for estimating confidence intervals. Section 4 summarizes the results of applying the methods to U.S. AIDS counts. These analyses confirm the strong influence of incubation on estimated infections and show that uncertainty about underreporting is also potentially important. These influences swamp the small amount of inherent stochastic uncertainty that would be present even if the inputs were known perfectly, as reflected by relatively narrow estimated confidence intervals. Greatly improved fits to the observed counts are obtained by modeling seasonal patterns and secular trends. Analysis of residuals shows that modeling seasonal effects also produces more homoscedasticity and reduced correlations between different subgroups' residuals. Section 5 discusses the implications of our findings, emphasizing backcalculation's large uncertainty and the need for complementary methods.

2. REVIEW OF THE BACKCALCULATION TECHNIQUE

2.1 The Basic Idea

We begin by briefly describing the convolution equation that forms the basis of the backcalculation method. Suppose that Y_j cases of AIDS occur in a given population between times T_{j-1} and T_j. Now, suppose infections are occurring in chronological time according to some form of arrival process that yields an infection curve, denoted by $I(\cdot)$; that is, the expected total number of infections that have occurred by chronological time s is given by $\int_{-\infty}^{s} I(u)du$. Further, let the incubation distribution be $F(\cdot)$, so that $F(t)$ is the probability that an AIDS diagnosis occurs within t time units after infection. To allow for the possibility that the incubation distribution may depend on the chronological time of infection, we let $F(\cdot\,|s)$ denote the incubation distribution for individuals infected at chronological time s. Then, the expected number of AIDS cases occurring in $[T_{j-1}, T_j]$ is given by

$$(1)\quad E(Y_j) = \int_{-\infty}^{T_j} I(s)\{F(T_j - s\,|s) - F(T_{j-1} - s\,|s)\}ds,$$

where we set $F(t|s) = 0$ for $t \le 0$ and all s. In practice, the time $-\infty$ can be replaced by the time of the beginning of the epidemic; we will use the time origin, $s = 0$, to denote this time, If AIDS diagnosis counts are available for a series of n (nonoverlapping) intervals, then we have a vector $(Y_1, \ldots, Y_n)$ of observations to analyze, where the expectation of each component satisfies (1). The basic idea of backcalculation is to use external information on the incubation distribution, F, together with the observed $(Y_1, \ldots, Y_n)$ to estimate I through some form of deconvolution. Once an estimate of the infection curve, $\hat{I}$, is obtained, future projections of AIDS case counts can be estimated using (1). For example, we can estimate the number of AIDS cases, Y_{AB}, in some future interval $[T_A, T_B]$ by

$$\hat{Y}_{AB} = \int_{0}^{T_B} \hat{I}(s)\{F(T_B - s\,|s) - F(T_A - s\,|s)\}ds.$$

Note that backcalculation of I on the basis of AIDS cases up to the current time T can only yield an estimate of $I(s)$ for $s \le T$. Therefore, the "estimated" values of $I(s)$ for $T < s < T_B$ depend on extrapolation. Early application of backcalculation (Brookmeyer and Gail, 1986, 1988) set $I(s) = 0$ for $T < s < T_B$, so that $\hat{Y}_{AB}$ predicts the number of future cases based only on individuals infected by time T. Other options include extrapolation of the estimated I beyond time T under various scenarios. We note that the quantity estimated is the *expected* number of AIDS cases, so interval estimates should account for random variation about this expected number in addition to the likely accuracy of the estimated number itself (Jewell, 1990).

HIV prevalence estimates can also be based on backcalculated estimates, $\int_{-\infty}^{T} \hat{I}(s)ds$, of the cumulative number of infections by time T. The current number of HIV-infected persons who are still alive can then be estimated by reducing this number by the estimated number of individuals, infected with HIV by time T, who have died by time T. The basis of the latter estimate is the AIDS death surveillance system of the Centers for Disease Control (CDC), which provides the number of AIDS-related deaths by calendar time. Because reporting of deaths is known to be incomplete and to suffer from reporting delay, it may be necessary to use other information on survival of HIV-infected individuals to obtain more accurate estimates of the current number still alive.

2.2 Incubation Estimates

2.2.1 *Data sources and methods*

Because HIV infection is generally not immediately detected, good data on incubation times are available for only a few cohorts whose infections can be traced to specific blood transfusions or whose times of infection can be determined (approximately) by tests of multiple serum specimens that bracket the time of infection. An important limitation of these data sources is that the cohorts may not be representative of the more general populations for which backcalculation is to be performed. In addition, even these "good" data have some important limitations. The infection times for transfusion-associated AIDS cases are retrospectively ascertained only after AIDS is diagnosed. This results in severe right truncation that greatly reduces the usefulness of the data, allowing estimation of the distribution only up to a constant of proportionality (Kalbfleisch and Lawless, 1988, 1989; Lagakos, Barraj and De Gruttola, 1988).

Repeated serum testing only determines the time of seroconversion (which can be a few months later than the time of infection) to within an interval between the last negative test and the first positive test, resulting in what is known as "doubly censored" data, and early stored serum specimens were available for only a few cohorts, including the San Francisco City Clinic (SFCC) Cohort (Hessol et al., 1989), a group of hemophiliacs from three U.S. cities (Brookmeyer and Goedert, 1989) and hemophiliacs from two French hospitals (De Gruttola and Lagakos, 1989). In addition, seroconverters from these and other studies have been gathered in an international registry (Biggar, 1990). The censoring has been handled by imputing the seroconversion time as the interval midpoint (Lui, Darrow and Rutherford, 1988; Hessol et al., 1989; Biggar, 1990) and by parametric (Brookmeyer and Goedert, 1989) and nonparametric (De Gruttola and Lagakos, 1989; Bacchetti, Segal and Jewell, 1992a) modeling of both the seroconversion and incubation distributions. Although it may seem unnecessary to estimate distributions of seroconversion dates in order to estimate incubation distributions, it has been demonstrated (De Gruttola and Lagakos, 1989; Jewell, 1990) that failing to do so can result in bias. At best, the midpoint approach will only be appropriate if the seroconversion density is uniform over the censoring interval. Often, however, the intervals are sufficiently long to make this assumption unreasonable.

Many cohorts of homosexual and bisexual men were recruited in 1984, by which time most of the subjects were already infected (seroprevalent). Such prevalent cohorts include the Multicenter AIDS Cohort Study (MACS) (Muñoz et al., 1989; Kuo, Taylor and Detels, 1991), the San Francisco General Hospital Cohort and the San Francisco Men's Health Study (Bacchetti and Jewell, 1991). Although subjects who are already seropositive at recruitment can be regarded as having interval-censored seroconversion times, the extreme width of the intervals (from the beginning of the epidemic to recruitment) precludes the fully nonparametric or midpoint-imputation approaches that are possible with less extreme, doubly censored data. An additional difficulty is the fact that the intervals are nearly the same for all prevalent subjects in a cohort, so that the intervals are not only wide but also almost completely overlapping. A number of strategies for overcoming these difficulties have been attempted. Fully parametric (Kuo, Taylor and Detels, 1991) and semiparametric (Taylor, Kuo and Detels, 1991) approaches have been pursued. Bacchetti and Jewell (1991) use a prior seroconversion-density estimate from external data. Another approach is to impute a seroconversion time for each subject based on laboratory markers at recruitment and knowledge of how the markers change with time since seroconversion (Muñoz et al., 1989).

Bacchetti (1990) used population data on AIDS incidence in gay men in San Francisco to obtain a deconvolution estimate of the incubation distribution. This relied on cohort-based estimates of seroconversion rates in the population and so did not escape the possibility that the cohorts might not be representative.

2.2.2 *Results*

There have been numerous estimates of AIDS incubation distributions from many sources. We summarize here only a few estimates from some influential data sources in the United States.

The first analysis (Lui, Darrow and Rutherford, 1988) of data from the SFCC cohort utilized the midpoint between the last negative and first positive specimen as an estimated seroconversion date for 84 subjects for whom this interval was less than 24 months. Subjects were considered to be right censored as of January 1987, and a parametric Weibull model produced an estimated median incubation time of 7.6 years. A subsequent analysis (Hessol et al., 1989) used the same estimated seroconversion dates but used followup through May 1989 and expanded the sample by also using a model-based seroconversion estimate (Byers et al., 1988) for individuals with an interval greater than 24 months. After 9.2 years the cumulative proportion with AIDS was 39% from a Kaplan-Meier estimate. An analysis of 1,171 seroconverters from various sources found significant evidence for heterogeneous incubation distributions in different groups (Biggar, 1990), as did an analysis of three cohorts with publicly available data (Bacchetti et al., 1993).

Bacchetti (1990) and Bacchetti and Jewell (1991) provide smooth nonparametric estimates of AIDS-incubation distributions using data from the San Francisco cohorts. Estimates of the median time to AIDS are between 10 and 11 years. An important feature of the nonparametric hazard estimates is that they flatten after approximately 7 years, contradicting the steadily increasing hazard assumed by Weibull models. Muñoz et al. (1989) also observed flattening of the hazard, but after only 4 years. By projecting the constant hazard rate, the median incubation time was estimated as 10.7 years.

As part of a joint analysis of covariate effects on both seroconversion and development of AIDS in a cohort of hemophiliacs, Brookmeyer and Goedert (1989) developed a Weibull regression estimate for their incubation distribution. This estimate has a median incubation time of 10 years and has been widely used as an input to backcalculation procedures and as a starting point for sensitivity analyses.

2.2.3 *Lengthening incubation*

Estimates of mean or median incubation times have lengthened steadily since an early estimate of 4.5 years (Lui et al., 1986). This may be a result of the wide use of Weibull models, which fit steadily increasing hazard functions. If the hazard in fact flattens after 4 to 7 years (Muñoz et al., 1989; Bacchetti, 1990), then median incubation times estimated under a Weibull assumption will lengthen as more followup becomes available. More recently, however, effective preventive treatments have truly lengthened incubation times for people receiving them. Thus, some estimate of the populationwide influence of this lengthening should be incorporated into realistic backcalculations (Brookmeyer, 1991; Rosenberg, Gail and Carroll, 1992).

Ideally, one would like to use available cohort data to directly estimate the influence of calendar time, for example, by including it as a time-varying cofactor in proportional hazards incubation models. Unfortunately, highly concentrated seroconversion densities for most cohorts mean that changes over time can be modeled fairly well using only stationary distributions, so that there is limited information about nonstationarity. To date, we are aware of only one attempt to directly estimate the influence of calendar time (Taylor, Kuo and Detels, 1991); this found weak evidence of lengthening. In addition, using nonstationary estimates from cohorts for backcalculation is problematic for two reasons: (1) cohorts can become less representative of general populations over time; for example, none of the San Francisco cohorts now have any subjects under age 30; (2) cohorts are usually carefully monitored for progression of HIV disease and counseled about treatment options, so they may be more likely to use preventive therapy and therefore show more lengthening of incubation times than a wider population.

The primary strategy to date has been to develop models of the chronological factors that may cause lengthening of incubation periods. Typically, this will involve assumptions or external estimates regarding the availability of treatment, treatment practices and the effects of treatment on delaying the onset of AIDS. Examples of this approach can be found in Solomon and Wilson (1990), Brookmeyer and Liao (1990b), Gail, Rosenberg and Goedert (1990), Brookmeyer (1991), Brookmeyer and Liao (1992) and Rosenberg, Gail and Carroll (1992). In some ways, this development runs counter to the original spirit of backcalculation, which largely avoided the need for assumption-dependent, explicit modeling. See Section 3.3 for the alternative approach used here.

Other factors may also influence the length of time between infection and an AIDS diagnosis. Few attempts have been made to incorporate covariate effects into backcalculation, although Becker and Marschner (1993) have used age, which is known to influence incubation times (Brookmeyer and Goedert, 1989; Biggar, 1990) and also clearly influences risk of infection from sexual activity or needle use. Here we allow for the risk of onset of AIDS to depend explicitly on calendar month and find considerable evidence of seasonal effects of this type.

2.2.4 *Sensitivity analyses*

Although backcalculation treats the incubation distribution as known, the limitations of the available data and the uncertainties about nonstationarity imply that it is far from known. Indeed, some investigators have found that uncertainty about incubation is the main source of uncertainty in backcalculation (Taylor, 1989; Gail and Rosenberg, 1992). Furthermore, the evidence for substantially differing incubations in different populations (Bacchetti et al., 1993; Biggar,

1990) is an additional source of uncertainty. This especially adds uncertainty for populations that lack specific incubation data, such as intravenous drug users, heterosexuals, and women. Thus, realistic sensitivity analyses are crucial for accurate assessment of the uncertainty of backcalculated estimates.

The complexity of the dependence of backcalculated estimates on the incubation used as an input, along with the lack of any low-dimensional structure for the set of possible incubation distributions, make choosing a comprehensive set of alternative incubations for sensitivity analyses a difficult task. Nevertheless, one basic insight can be gleaned from examination of Equation (1). Because the epidemic is less than 15 years old, the existing data do not involve all of the incubation distribution, and there is therefore scope for adjusting the scale of F. From Equation (1), we see that multiplying F by a constant will divide the estimate of I by the same constant, leaving fitted and projected diagnosis counts unchanged. Thus, varying the scale is important for assessing the uncertainty in cumulative infections but useless (or even misleading) for assessing uncertainty in projections. For example, sensitivity analyses based mainly on Weibull models with similar shapes may be too narrow to realistically assess uncertainty in projected diagnosis counts (Gail, Rosenberg and Goedert, 1990; Segal and Bacchetti, 1990). In addition, it must be kept in mind that estimates of I are influenced by the shape and nonstationarity of F, in addition to its scale. Parameters such as prevalence and effectiveness of treatment in nonstationary incubation models can be systematically varied (Brookmeyer, 1991; Rosenberg, Gail and Carroll, 1992), but it is not clear that such variation fully covers the range of uncertainty about incubation. In Section 4, we use four incubation inputs that differ in a number of respects, but we do not claim that this set of four is comprehensive. This is an area that would benefit from further development.

2.3 Surveillance Data and Reporting Delays

According to the HIV/AIDS surveillance reports issued by the Centers for Disease Control, "all 50 states, the District of Columbia, U.S. dependencies and possessions, and independent nations in free association with the U.S. [Puerto Rico] report AIDS cases to the CDC using a uniform case definition and case report form." The original list of opportunistic infections and malignancies that qualify as an AIDS diagnosis (CDC, 1982) was broadened slightly in 1985 (CDC, 1985) and again in 1987 (CDC, 1987a)..Information supplied for each reported AIDS diagnosis includes age at diagnosis, geographic location, risk category, and cause or condition of diagnosis. Completeness of reporting of diagnosed cases apparently varies across geographic regions and patient characteristics. Based on mortality studies, CDC surveillance reports state that 70 to 90% of HIV-related deaths in men 25 to 44 years of age are identified through the AIDS diagnosis surveillance system (Buehler et al., 1990). Many authors use 15% as an estimate of the extent of underreporting as a basis for projecting actual AIDS cases rather than merely reported cases (Rosenberg, Gail and Carroll, 1992). Rosenberg and Gail (1990) also examined a few possible scenarios in which underreporting changed over time.

A key aspect of the surveillance system is the delay between the date of diagnosis and the date that a report is received at the CDC. Fewer than 10% of cases are reported in the month of diagnosis; about 50% are reported within 2 months, 85% within 1 year, and 95% within 2 years (Karon, Devine and Morgan, 1989). There is evidence that the length of this *reporting delay* varies across geographic region (Brookmeyer and Damiano, 1989).

Information on the distribution of reporting delays is available from the data on reported cases, because both date of diagnosis and date of report are provided for each case. Note, however, that for a given period of diagnosis, reporting delay data is right truncated; for example, diagnoses occurring in the first month of 1991 with a reporting delay of over a year will not be observed in the surveillance system as of the first month of 1992. Nevertheless, there has been a substantial statistical methodology developed to estimate a distribution function based on truncated data (Wang, Jewell and Tsai, 1986; Lagakos, Barraj and De Gruttola, 1988; Kalbfleisch and Lawless, 1989) that has been successfully applied to estimation of the reporting delay distribution (Brookmeyer and Damiano, 1989; Harris, 1990).

We note that reporting delays have been accommodated in backcalculation in different ways. First, knowledge of the reporting delay distribution can be used to impute "true" AIDS counts from reported counts, and then the adjusted AIDS counts are used as if there were no reporting delay. Second, the reporting delay distribution can be simultaneously estimated along with the infection curve in the backcalculation technique through use of the report dates in addition to diagnosis dates (Harris, 1990; Lawless and Sun, 1992). This permits an assessment of how uncertainty about reporting delay adds to uncertainty in backcalculated estimates. In Sections 3 and 4, we take an intermediate approach, namely, allowing for variation in reporting delays (rather than using single imputations) in the likelihood but assuming that the appropriate reporting delay distributions are known.

Finally, we note that some authors have allowed for the possibility that reporting delays may themselves be changing in magnitude as the epidemic evolves (Harris, 1990; Brookmeyer and Liao, 1990a; Pagano et al., 1992b). There is growing suspicion that reporting lags are lengthening in comparison to earlier in the epidemic. The methods described in Section 3 allow for this possibility by incorporating a nonstationary effect as part of the modeling of nonstationary incubation properties.

2.4 Infection Model

Equation (1) only specifies the first moment of the Y_j. Further distributional assumptions will be necessary to develop a full likelihood approach to estimation and inference for I. Alternatively, at least second-moment structure [that is, the covariance structure of $(Y_1, \ldots, Y_n)$] is necessary for a generalized linear model approach to estimation of I. A standard assumption is that the (unobserved) times of infection for different individuals are independent. This follows from a stronger assumption (but only slightly stronger) that infections arrive according to a nonhomogeneous Poisson process with intensity function I. This will be the approach taken in the likelihood development we describe in Section 3. In fact, both the independence and Poisson assumptions may be violated. Because Equation (1) holds even if infection times are dependent, estimates based on independence are likely to remain consistent even when this assumption is violated, but model-based estimates of variability may be underestimated, as in other similar situations (Liang and Zeger, 1986). It is, however, unlikely that this phenomenon is a major concern in light of much greater uncertainties associated with other aspects of the backcalculation approach. Most researchers have found substantial extra-Poisson variation or overdispersion. Some have handled this by estimating overdispersion in a quasi-likelihood approach (Brookmeyer and Liao, 1990b; Lawless and Sun, 1992), but overdispersion is often ignored in the estimation process, because it also does not invalidate Equation (1). Overdispersion must be taken into account, however, when assessing stochastic variation of backcalculated estimates.

To avoid an ill-posed inverse problem (O'Sullivan, 1986), it is necessary to impose some kind of structure on the infection curve, I. Some backcalculations used smooth parametric models for I (Day et al., 1989). Brookmeyer and Gail (1986, 1988) and Rosenberg et al. (1991a) used somewhat more flexible parametric models for I, namely, step functions where I is assumed to be piecewise constant. Typically, these models for I had four or five steps with jumps occurring at fixed known points in time. Later, rapid fitting techniques (Rosenberg and Gail, 1991) were used to select optimum cutpoints between the four steps. Rosenberg, Gail and Pee (1991) provide simulation studies suggesting that step-function models are satisfactory in terms of mean-square error of estimated HIV reconstructions and AIDS projections. Other investigations have allowed an arbitrary number of steps but have constrained the form of I using smoothness assumptions. Becker, Watson and Carlin (1991) introduced smoothness by using the EMS algorithm (Silverman et al., 1990) whereas Brookmeyer (1991) and Bacchetti, Segal and Jewell (1992a) used a penalized likelihood approach, and Pagano et al. (1992a) employed ridge regression. Section 3 describes the penalized likelihood used here.

2.5 Complementary Methods

Several alternative approaches to estimation of the past HIV infection curve, current HIV prevalence and future AIDS case counts have been suggested and may be very useful as comparisons to or corroborations of the respective estimates produced through backcalculation. Mathematical epidemic models for HIV transmission have been and continue to be widely studied. Excellent introductory reviews can be found in Isham (1988) and Anderson et al. (1986). Such models have not been broadly effective in providing quantitative estimates of HIV prevalence or AIDS incidence projections, largely because there is little accurate data available for estimates of key parameters. Nevertheless, they remain valuable tools for assessing the effects of various potential interventions and for contributing plausible explanations of observed patterns of AIDS incidence. For further comments, see Gail and Brookmeyer (1988) and Jewell (1990).

At the other end of the spectrum, simple extrapolation techniques have often given useful indications of the short-term prognosis of the epidemic. Statistical issues concerning the useful implementation of extrapolation techniques have been discussed by a variety of authors including Morgan and Curran (1986), Healy and Tillet (1988), Zeger, See and Diggle (1989) and Karon, Devine and Morgan (1989). The advantage of extrapolation over backcalculation is that it does not require external information on incubation distributions, a considerable source of uncertainty. There are two major disadvantages, however: (1) extrapolation does not provide any information on the HIV incidence curve; and (2) the method uses less information and so may be less efficient than backcalculation. Regarding the latter point, it remains to be seen how much bias the backcalculation method can tolerate (through misspecification of incubation properties) before any gains in efficiency are negated.

Cohorts of seronegatives can be followed to obtain seroconversion rates, but those who remain in followup may be more conscientious and less likely to risk infection than those who drop out or the general population. Two other strategies are (1) estimation of seroprevalence from a populationwide random survey sample; and (2) estimation of seroprevalence in specified risk subgroups by a variety of means, and then combination of these estimates to provide a population estimate through weighting by estimates of the size of the risk groups. Repetition of such surveys over time allows for estimation of HIV incidence. De Gruttola and Fineberg (1989) discuss issues surrounding implementation of large-scale HIV prevalence surveys. A particularly difficult problem is nonresponse bias (CDC, 1991a). We discuss in Section 5 a recent proposal to use laboratory measurements on samples of seropositives to deduce past infection rates.

3. BACKCALCULATION METHODS

The particular approach to be used here for estimation of the infection rates is based on maximizing a roughness-penalized likelihood. We make some simplifying assumptions, but we use only observed data with no imputation. We also generalize the parameterization to allow for estimation of secular trends (such as the influence of preventive treatment) and seasonal effects as part of the backcalculation procedure. We assume that diagnosis counts are available by month, so that a discrete monthly time scale is appropriate, in contrast to the continuous notation used in Section 2.1.

3.1 Likelihood Formulation

We make the usual assumption that the number of new infections in month i follows a Poisson distribution with parameter θ_i and that given θ, the numbers of infections in different months are independent. This corresponds to assuming that infections arise according to a nonhomogeneous Poisson process as described in Section 2.4. We consider all cases diagnosed by some month n and reported by a possibly later month n^*. Thus, we observe

$$y_{jk} = \#\text{ diagnosed month } j \text{ and reported month } k,$$
$$0 \le j \le n, j \le k \le n^*,$$

and we wish to estimate θ using knowledge of

$$a_{ijk} = Pr\{\text{diagnosed at } j \text{ and reported at } k \mid \text{infected at } i\}.$$

The numbers a_{ijk} reflect external knowledge about the incubation, reporting delay and underreporting distributions. The assumptions concerning the infection rates θ imply that the y_{jk} are also independent Poisson. Letting $y_j = \sum_{k=j}^{n^*} y_{jk}$, we write the likelihood of θ given **a** as

$$L(\theta|\mathbf{a}) = L_c \cdot L_m,$$

where L_c is the likelihood of the y_{jk} conditional on the y_j and L_m is the marginal likelihood of the y_j. We have

$$L_c \propto \prod_{j=0}^{n} \prod_{k=j}^{n^*} A_{jk}^{y_{jk}}, \text{ where } A_{jk} = \frac{\sum_{i=0}^{j} a_{ijk}\theta_i}{\sum_{m=j}^{n^*} \sum_{i=0}^{j} a_{ijm}\theta_i},$$

and

$$L_m \propto \prod_{j=0}^{n} \left[\sum_{k=j}^{n^*} \sum_{i=0}^{j} a_{ijk}\theta_i\right]^{y_j} \exp\left[-\sum_{k=j}^{n^*} \sum_{i=0}^{j} a_{ijk}\theta_i\right].$$

Suppose that $a_{ijk} = D_{ij}R_{jk}$, where D_{ij} is the probability of diagnosis at j given infection at i and R_{jk} is the probability of reporting at k given diagnosis at j. (D_{ij} and R_{jk} are assumed to be known from external information.) This assumes that the length of incubation does not influence the subsequent reporting delay or chance of no report, leading to a useful simplification:

$$A_{jk} = \frac{R_{jk}}{\sum_{m=j}^{n^*} R_{jm}},$$

so that L_c does not depend on θ. We therefore estimate θ using the log-likelihood

$$\text{(2)} \quad \log(L_m) = \sum_{j=0}^{n} \left[y_j \log\left(\sum_{i=0}^{j} R_j D_{ij}\theta_i\right) - \sum_{i=0}^{j} R_j D_{ij}\theta_i \right],$$

where $R_j = \sum_{k=j}^{n^*} R_{jk}$, the probability that a case diagnosed at time j is reported by time n^*. Note that R_j depends on both reporting delay and underreporting.

Because confidentiality concerns may prevent release of exact diagnosis and reporting times of early cases, say those diagnosed up to time g, we may observe only $y^* = \sum_{j=0}^{g} y_j$. (For example, the CDC does not release month of diagnosis for cases diagnosed before 1982.) In this case, we have

$$\text{(2a)} \quad \begin{aligned} \log(L_m) = {} & y^* \log\left(\sum_{j=0}^{g} \sum_{i=0}^{j} R_j D_{ij}\theta_i\right) - \sum_{j=0}^{g} \sum_{i=0}^{j} R_j D_{ij}\theta_i \\ & + \sum_{j=g+1}^{n} \left[y_j \log\left(\sum_{i=0}^{j} R_j D_{ij}\theta_i\right) - \sum_{i=0}^{j} R_j D_{ij}\theta_i \right]. \end{aligned}$$

3.2 Estimation Method

We estimate θ by optimizing a penalized-likelihood criterion:

$$\text{(3)} \quad \begin{aligned} \log(L_m) - \frac{\lambda_\theta}{2} \sum_{i=0}^{n-2} \Big(\log(\theta_i) & - 2\log(\theta_{i+1}) \\ & + \log(\theta_{i+2})\Big)^2. \end{aligned}$$

Penalizing roughness on the log scale forces $\theta_i > 0$ for any $\lambda_\theta > 0$, so that constrained optimization is not needed. The ultrasmooth case ($\lambda_\theta = \infty$) is exponential growth or decline. Although the criterion (3) can be optimized directly by the Newton-Raphson method, an

EM algorithm is useful for reducing the computational burden and for providing good starting values for fitting the more complicated models described below in Section 3.3. To define the EM algorithm, consider the complete data

$$x_{ijk} = \#\,\text{infected month } i,\ \text{diagnosed month } j \text{ and reported month } k,$$

which are distributed Poisson with means $\theta_i D_{ij} R_{jk}$. Simple manipulations show that the complete-data log-likelihood, up to a constant, is

$$(4)\qquad \sum_{i=0}^{n}\left[x_i \log(\theta_i) - \theta_i \sum_{j=i}^{n} D_{ij} R_j\right],$$

where $x_i = \Sigma_j \Sigma_k x_{ijk}$. The EM algorithm begins with an initial guess for θ and proceeds by calculating the expectation of the complete-data log-likelihood (E step), using the formula (see Green, 1990, p. 450):

$$E[x_i \,|\mathrm{y}, \theta] = y^* \frac{\theta_i \sum_{j=i}^{g} D_{ij} R_j}{\sum_{r=0}^{g} \theta_r \sum_{s=r}^{g} D_{rs} R_s} + \sum_{\substack{j=i \\ j>g}}^{n} y_j \frac{\theta_i D_{ij}}{\sum_{r=0}^{j} \theta_r D_{ij}},$$

with the first term appearing only if $i \leq g$. The algorithm then continues by finding a new value of θ that optimizes the penalized expected log-likelihood (M step). This optimization can be done quickly by the Newton-Raphson method, because the log-likelihood (4) has a diagonal matrix of second derivatives, and the matrix for the roughness penalty is banded (Bacchetti, 1990).

With an initial guess of constant (and positive) infection rates, the algorithm defined by alternating the E and M steps described above, until convergence, is substantially faster than direct use of Newton-Raphson optimization on the incomplete-data criterion (3).

3.3 Modeling the Influence of Calendar Time

The formulation in Section 3.1 allows for the D_{ij}'s and R_{jk}'s to depend on more than just the elapsed time, $j - i$ or $k - j$. Thus, external estimates of the influence of calendar time itself, such as secular trends and seasonal effects, can be incorporated, if available. This is particularly important because it is widely believed that the incubation distribution is nonstationary, that is, that it has been changing over time due to preventive treatment and other factors (Section 2.2.3). Good external estimates of such nonstationarity, however, are generally not available as of this writing, and it is useful in any case to be able to estimate calendar time effects as part of the backcalculation to correct for possibly inaccurate or inappropriate external estimates. We allow for multiplicative time effects by replacing the known terms $R_j D_{ij}$ in (2) or (2a) by $R_j D_{ij} e^{\beta_j + S(j)}$ and by adding another roughness penalty to the overall criterion (3) to ensure that the β_j are smooth. Here the nonstationarity factors β_j are unknown and model smooth secular trends (see Section 5 concerning the handling of jumps), and the $S(j)$ are unknown seasonal effects such that, for monthly data, $S(j + 12) = S(j)$ for all j and $S(1) = 0$. The seasonal effects thus consist of 11 parameters unless the $S(j)$'s are further restricted. Note that β_j and $S(j)$ can be thought of as modifying R_j, as modifying the D_{ij} for all i, or both, because these terms only appear multiplied together in (2) and (2a). Thus, with this model, lengthening incubation times, longer reporting delays, and increased underreporting cannot be separately estimated from the y_j; only their combined effect on diagnosis counts contributes to the likelihood.

We use a roughness penalty for β equal to the sum of squared second differences, so the criterion optimized is

$$\log(L_m) - \frac{\lambda_\theta}{2} \sum_{i=0}^{n-2} \left(\log(\theta_i) - 2\log(\theta_{i+1}) + \log(\theta_{i+2})\right)^2$$
$$- \frac{\lambda_\beta}{2} \sum_{j=0}^{n-2} (\beta_j - 2\beta_{j+1} + \beta_{j+2})^2.$$

Because incubation times are thought to have been influenced mainly by recent use of preventive treatment and recent changes in the AIDS case definition, we set β_j equal to zero for j before January 1986, noting that the penalty includes terms to ensure that the β_j's diverge from zero gradually after that date. In particular, the ultrasmooth case ($\lambda_\beta = \infty$) is $\beta = 0$. The joint optimization in θ, β and the $S(j)$'s does not appear to be amenable to an EM approach, so we instead use the Newton-Raphson method. For a starting value, we find the optimal θ with β and the $S(j)$'s set to zero, using the EM algorithm described in the preceding section.

3.4 Diagnosis Counts, Reporting Delay, Fitted Values and Incubation

We apply our backcalculation methods to diagnosis counts from the AIDS Public Information Data Set provided by the CDC. We use data for all adult or adolescent cases diagnosed by December 1990 (which is n in the notation of Section 3) and reported by June 1991 (n^* in Section 3). (These cases all meet the 1987 case definition of AIDS.) For some analyses, we also apply our methods to the following specific risk groupings: homosexual and bisexual men who did not use intravenous drugs; heterosexual or female intravenous drug users; homosexual or bisexual male intravenous drug users; and all others. The data set includes estimates of reporting delay corrections, which we combine with assumptions about underreporting in order to calculate the reporting completeness factors R_j. (We note that the nonstationarity estimation described in Section 3.3 can correct for some inaccuracy in the reporting delay distribution used.) The fitted values for the observed y^* and y_j's are

$$\hat{y}^* = \sum_{j=0}^{g} e^{\hat{S}(j)} \sum_{i=0}^{j} R_j D_{ij} \hat{\theta}_i$$

$$\hat{y}_j = e^{\hat{\beta}_j + \hat{S}(j)} \sum_{i=0}^{j} R_j D_{ij} \hat{\theta}_i,$$

and the standardized residuals are $(y_j - \hat{y}_j)/\hat{y}_j^{1/2}$. For some purposes, it is more natural to consider the actual number of diagnoses, whether reported or not, so we impute these numbers as y_j/R_j. By using the *a priori* R_j's, we make the optimistic assumption that the β_j's modify only the D_{ij}'s and not the R_j's. Fitted values for these imputed counts (and predicted future counts) are calculated by the same formula given above for $\hat{y}_j$ but without the factor of R_j, that is, by $\hat{y}_j/R_j$. For future counts, we extrapolate the β_j's linearly. Goodness of fit is measured by the Poisson deviance

$$-2\left[y^* \log \frac{\hat{y}^*}{y^*} - (\hat{y}^* - y^*) + \sum_{j=g+1}^{n} \left(y_j \log \frac{\hat{y}_j}{y_j} - (\hat{y}_j - y_j) \right) \right].$$

Note that the $\hat{y}$'s, and hence the deviance, depend on the externally estimated D_{ij}'s and R_j's, in addition to the backcalculated estimates.

We use D_{ij}'s corresponding to four different incubation distributions that have been previously described. Briefly, these are: a nonstationary model that attempts to reflect the influence of preventive treatment beginning in mid-1987 (Brookmeyer, 1991), a stationary Weibull model fitted to data from a random sample of the San Francisco City Clinic Cohort, and stationary three-parameter generalizations of the Weibull fitted to data from the San Francisco Hepatitis B Vaccine Trial Cohort and the Multicenter Hemophilia Cohort (Bacchetti et al., 1993). A deterministic seasonal effect reflecting variation in the lengths of calendar months is incorporated a priori into the D_{ij}'s for each of the distributions. (Thus, the estimated $S(j)$'s reflect seasonal effects beyond, e.g., the effect caused by February's having fewer days than January.) We feel that all of these distributions are reasonable, since they are based on either actual cohort data or plausible assumptions (the treatment model). They differ in scale, shape and stationarity and so provide some insight into backcalculation's sensitivity to the assumed D_{ij}'s, although they certainly do not span the entire range of plausible incubation distributions.

3.5 Choosing the Amount of Smoothness and Estimating Confidence Intervals

The estimates obtained depend on the values used for λ_θ and λ_β. The impact of the choice of these smoothness parameters can be explored by plotting the resulting deviance for a range of possible values and by examining the estimates that result from different values. In addition, it is useful to have a data-driven method for choosing the amounts of smoothness. Because the value of one parameter, say λ_β, can influence the data-driven choice of the other, say λ_θ, we alternately refine the choice of each smoothness parameter while holding the other fixed, until the parameters appear to be near their simultaneously optimal values.

The choice of one smoothness parameter, say λ_θ, with the other fixed, can be based on a method for testing $\lambda_\theta = \lambda_0$ versus $\lambda_\theta = \lambda_1$, with $\lambda_1 < \lambda_0$. The method creates simulated values of y^* and of the y_j's based on the estimates that result from setting $\lambda_\theta = \lambda_0$. These simulated epidemics are then fitted using both $\lambda_\theta = \lambda_0$ and $\lambda_\theta = \lambda_1$, and the improvement in the deviance from using λ_1 instead of λ_0 is calculated. This improvement is due to the greater flexibility allowed with a smaller λ_θ, even though the "true" underlying model (the one used to create the simulated data) is based on the larger λ_0. Thus, the hope is that a large number of such simulations will define how much improvement could be expected merely from the greater flexibility of λ_1, even though λ_0 is adequate. This knowledge is then used to decide if the actual improvement (using the real y^* and y_j's) from λ_1 is likely to reflect a genuinely better fit to the data, for example, if it is more than twice the average of the simulated improvements (analogous to the Akaike Information Criterion) or if fewer than 10% of the simulated improvements are as large.

To select a value for λ_θ, we begin with a value that is too large and test it against an alternative whose log is 0.5 smaller. The logs of the test values λ_0 are decreased by 0.5 until one is found where 100 simulations produce no strong evidence that the improvement from the next smaller value is genuine. The same approach is then used to choose a value for λ_β and is then repeated for λ_θ and so on until the choices no longer change.

Realistic simulations are needed for this approach to produce good choices of λ_θ and λ_β. In particular, simulations that do not reflect overdispersion (see Section 2.4) are likely to allow improvements that result from overfitting the actual data to appear to be better than random, because the overfitting produces bigger improvements than it would for simulated data that were not overdispersed. We therefore generate simulations that reflect overdispersion by generating normally distributed random deviates with the means from our model and variances equal to the means times the estimated overdispersion, and then rounding to the nearest integer. We estimate the overdispersion as the deviance divided by the approximate error degrees of freedom [number of data points minus number of parameters estimated (McCullagh and Nelder, 1989)]. The nonparametric estimation methods for θ and β make the effective number of parameters for these estimates difficult to ascertain, but because there are over 100 data points, some inaccuracy in this number does not change the estimated overdispersion much.

Once λ's have been chosen, we use simulations as described above, with fixed λ's, to produce pointwise confidence intervals using the percentile method (Efron, 1982). We do not use these simulations to correct for bias (Efron, 1985) because we want the estimates to be biased toward smoothness (hence the roughness penalties). This approach accounts for overdispersion because the simulations reflect it. We also produce confidence intervals by using the inverse of the observed information matrix for $(\theta, \beta, S(\cdot))$ – that is, the second derivatives of the penalized log likelihood – and assuming normality. [See Segal, Bacchetti and Jewell (1993), for a discussion of the asymptotic covariance matrix in penalized settings. Although this matrix is very high-dimensional, it can be efficiently calculated analytically and inverted by using standard routines.] This approach ignores overdispersion. An additional source of variation is the choice of the smoothness parameters. A preliminary examination of this issue (Bacchetti, Segal and Jewell, 1992b) finds that this does not appear to influence the widths of confidence intervals as much as other factors.

4. RESULTS

We present here some analyses that explore a number of methodological issues in backcalculation. For the sake of brevity, most analyses are performed using only the treatment model (Brookmeyer, 1991) of incubation and a constant 10% underreporting rate. The exception is Section 4.1, where we specifically explore the importance of these choices.

4.1 Sensitivity Analyses

4.1.1 *Choice of incubation distribution*

Figures 1 and 2 show the estimates of θ and β that result from using the four different incubation distributions described in Section 3.4, with underreporting assumed constant at 10%. (See Section 4.3 for details of how λ_θ and λ_β are chosen for each incubation.) The estimated infection rates are similar in that they all peak in 1984 or early 1985, but they vary quite a bit in magnitude and shape. In particular, two of the estimates, those based on incubation distributions whose hazards eventually decline, have very gradual declines in infection rates and continuing high rates of infection. The other two show steeper declines and lower recent rates. The estimated nonstationarity factors β also show some overall similarity, with all reflecting slowing of reported diagnoses after mid-1987. For the treatment model, the β_j's are *additional* nonstationarity factors that reflect slowing beyond that caused by the modeled treatment effects. In this case, an adequate fit to the data can be obtained with

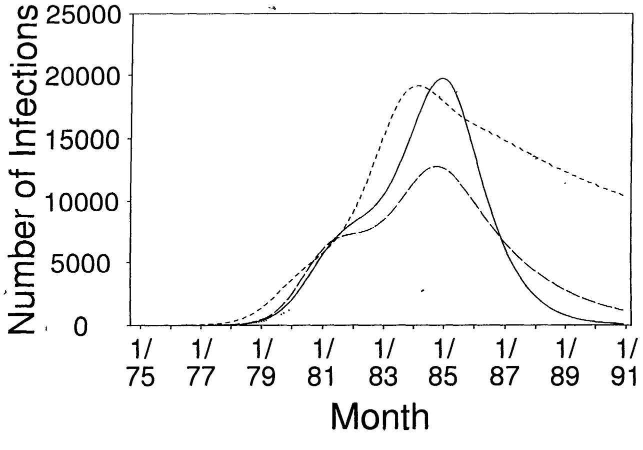

FIG. 1. *Estimated numbers of new infections each month from backcalculations using different assumed incubation periods, with adaptively chosen smoothness parameters and underreporting assumed constant at* 10%. (*Solid line*) *Treatment model.* (*Long dashed line*) *Weibull model fitted to random sample data.* (*Dots*) *Three-parameter model fitted to hepatitis B vaccine trial data.* (*Short dashed line*) *Three-parameter model fitted to hemophiliac data.*

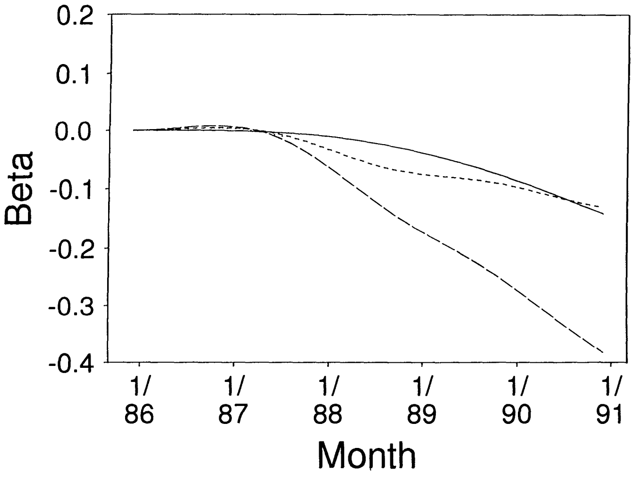

FIG. 2. *Estimated nonstationarity factors for each month, beginning in* 1986, *from backcalculations using different assumed incubation periods, with adaptively chosen smoothness parameters and underreporting assumed constant at* 10%. (*Solid line*) *Treatment model.* (*Long dashed line*) *Weibull model fitted to random sample data.* (*Dots*) *Three-parameter model fitted to hepatitis B vaccine trial data.* (*Short dashed line*) *Three-parameter model fitted to hemophiliac data.*

TABLE 1
Summary of backcalculations using the different incubation periods from Section 3.4

Incubation model	Adaptively chosen		Month of peak in infections	Cumulative infections (in thousands by month/year)			Deviance	Predicted diagnoses in 1/95*
	$\log(\lambda_\theta)$	$\log(\lambda_\beta)$		12/84	6/87	12/90		
Treatment model	10	17	12/84	620	986	1,031	198.8	3,600
Random sample	9	14	10/84	505	781	897	188.8	3,000
Hepatitis B vaccine trial	10.5	14.5	4/85	522	1,033	1,505	191.3	3,700
Hemophiliac cohort	10	14.5	3/84	752	1,229	1,742	190.5	3,800

* All cases, whether reported or not, that meet the 1987 case definition.

smoother β_j's that are closer to zero. Use of the Weibull incubation model from the Random Sample cohort, which has hazard rates that increase rapidly, requires very large amounts of estimated slowing in order to fit the data. Estimated seasonal effects from the four models are very similar (not shown).

Table 1 summarizes some aspects of the backcalculation models based on the four incubation distributions. There is an almost twofold difference in estimated cumulative infections through 1990, but better agreement earlier, particularly through 1984. The models show fairly good agreement for the projections of new AIDS diagnoses in January 1995. This agreement, however, depends on assuming for all four incubations that the β_j's modify only diagnosis probabilities and not reporting completeness. If the β_j's are instead assumed to modify only reporting completeness, then there would be wider disagreements, with projected numbers as high as 6,800. The values of λ_β for each model are all highly statistically significant when tested directly against $\lambda_\beta = \infty$ ($\beta = 0$) by the method described in Section 3.5. Requiring $\beta = 0$ would produce lower estimated infection rates, particularly for the Random Sample-based model, leading to cumulative infection estimates that range from 0.5 to 1.5 million by December 1990. All four models produce excellent fits to the observed diagnosis counts, as shown in Figure 3. (The counts and fits have been adjusted for seasonal effects and incomplete reporting as described in the figure legend to produce a visually simpler plot.) The differences in the deviances are fairly small, consistent with the close agreement shown in Figure 3, and they partly reflect differences in the effective numbers of parameters.

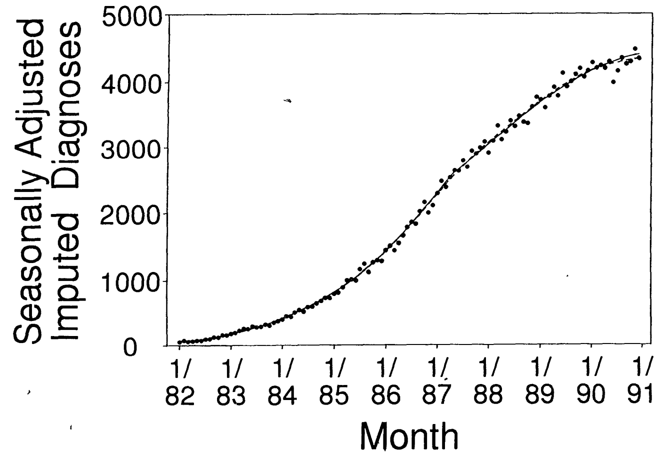

FIG. 3. *Fits to seasonally adjusted, reporting-adjusted diagnosis counts for all adult/adolescent U.S. AIDS cases, from backcalculations using different assumed incubation periods, with adaptively chosen smoothness parameters and underreporting assumed constant at* 10%. *(Dots) Imputed diagnosis counts calculated by dividing y_j by $R_j e^{S(j)}$ and then correcting for the lengths of calendar months by multiplying by* 365.25 *and dividing by twelve times the number of days in the month. (Solid line) Adjusted fits using the treatment model of incubation, calculated by applying the same corrections described above to the raw fitted $\hat{y}_j$. (Dashed line) Adjusted fits using the other three incubation models (all three visually indistinguishable).*

4.1.2 *Choice of underreporting assumptions*

For this analysis, we restrict attention to backcalculations using the treatment model of incubation with $\lambda_\theta = \exp(10)$ and $\lambda_\beta = \exp(17)$. We examine a fairly narrow set of possibilities that each have underre-

TABLE 2
Influence of underreporting assumptions on estimates of cumulative HIV infections and projections. Models are based on the treatment model (Brookmeyer, 1991) *of incubation, with λ_θ =* exp(10) *and λ_β =* exp(17)

Underreporting (month/year)		HIV infections (in thousands by month/year)			Predicted diagnoses in 1/95*
6/87	12/90	12/84	1/85–6/87	7/87–12/90	
10%	10%	620	366	45	3,600
15%	15%	657	387	47	3,800
10%	15%	616	406	61	4,100
15%	10%	661	345	36	3,400

* All cases, whether reported or not, that meet the 1987 case definition.

TABLE 3
Deviances and estimated overdispersions from backcalculation models with λ_θ = exp(8.5) *and* λ_β = exp(15), *using the treatment model (Brookmeyer, 1991) of incubation*

		Deviance			Estimated overdispersion	
Group	N	No seasonal effects	Quarter effects (3 df)	Month effects (11 df)	No seasonal effects	Month effects
Homosexual/bisexual men	105,741	306.6	204.9	160.7	3.1	1.8
Intravenous drug users	39,904	203.4	170.8	140.0	2.0	1.6
Homosexual/bisexual male intravenous drug users	11,823	169.5	154.8	146.8	1.7	1.6
All others	22,226	135.4	98.4	88.0	1.4	1.0
Total adult or adolescent cases	179,694	411.3	255.7	185.6	4.1	2.1

porting constant up to June 1987 and then changing linearly (or remaining constant) from June 1987 to December 1990. Table 2 shows that allowing for changing underreporting rates can have larger impacts than simply altering an assumed constant level. In particular, line three of the table shows fewer early infections than line one and more recent infections and higher projections than line two. Also, line four shows fewer recent infections and lower projections than line one but more early infections than line two. The estimated nonstationarity factors (not shown) and projections are influenced fairly directly by changing assumed rates of underreporting.

4.2 Overdispersion and Seasonal Effects

In this subsection we again restrict attention to backcalculations that use the treatment model for incubation. We use smaller values for the smoothness parameters than would be chosen using the method of Section 3.5 (in particular, we here use λ_θ = exp(8.5) and λ_β = exp(1.5)). This is to reduce the influence of any systematic lack of fit on the residuals that we wish to analyze. Estimating the overdispersion of the residuals requires knowing the effective number of parameters estimated (see Section 3.5). We treat the use of λ_θ = exp(8.5) and λ_β = exp(1.5) as being equivalent to estimating 9 parameters, because scatterplot smooths of the diagnosis counts with approximately 9 degrees of freedom (Buja, Hastie and Tibshirani, 1989) produce comparable or slightly better fits to the data. We correspondingly treat models that estimate month effects as having 20 effective parameters.

Table 3 summarizes the model fits for four subgroups and overall, both with and without seasonal effects. With no seasonal effects, the estimated overdispersion for all cases is large, and Figure 4 shows that there is also heteroscedasticity in the standardized residuals. These findings suggest that the Poisson model is somewhat unrealistic, at least when no seasonal effects are modeled. A further anomaly is the fact that the overdispersion is larger for all cases than for any of the subgroups. Because it estimates the ratio of the variance to the mean, the overdispersion for all cases would be a weighted average of the overdispersions in the subgroups, if the subgroups were independent. Table 4 confirms that the standardized residuals from separate backcalculations on the subgroups are not independent.

Allowing for seasonal effects substantially improves the fit to the data, resulting in much lower estimates of the overdispersion (Table 3). (The estimates of θ and β, however, are not changed much.) These improvements appear to be highly statistically significant, because the improvements in the deviances are much larger than would be expected for the appropriate null chi-square distributions. (See Buja, Hastie and Tibshir-

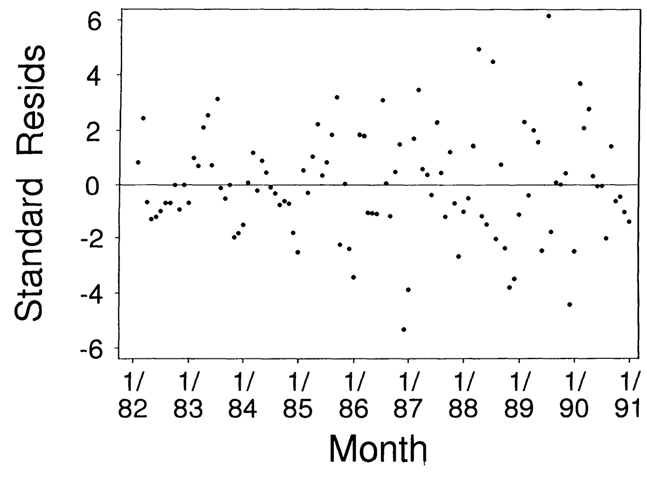

FIG. 4. *Standardized residuals* $(y_j - \hat{y}_j)/\hat{y}_j^{1/2}$ *for the fit to all adult/adolescent U.S. AIDS cases, with the* $\hat{y}_j$ *from a backcalculation model using the treatment model of incubation,* λ_θ = exp(8.5), λ_β = exp(15), *underreporting assumed constant at* 10% *and no monthly seasonal effects* ($S(j)$ = 0 *for all* j).

TABLE 4
Correlations among monthly standardized residuals for different risk groups, from backcalculation models with the treatment model (Brookmeyer, 1991) of incubation, $\lambda_\theta = \exp(8.5)$, $\lambda_\beta = \exp(15)$, and underreporting assumed constant at 10%

	No seasonal effects		
Correlation (P-value)	Intravenous drug users	Homosexual or bisexual male intravenous drug users	All others
Homosexual or bisexual men	0.369 (0.0001)	0.422 (0.0001)	0.347 (0.0002)
Intravenous drug users		0.247 (0.0097)	0.230 (0.016)
Homosexual or bisexual male intravenous drug users			0.141 (0.14)
	Month effects		
Correlation (P-value)	Intravenous drug users	Homosexual or bisexual male intravenous drug users	All others
Homosexual or bisexual men	0.080 (0.41)	0.342 (0.0003)	0.086 (0.37)
Intravenous drug users		0.156 (0.11)	0.182 (0.058)
Homosexual or bisexual male intravenous drug users			0.039 (0.69)

ani, 1989, for a discussion of using chi-squared distributions to compare deviances in nonparametric or semiparametric settings.) Quarter effects [from requiring that $S(1) = S(2) = S(3)$, $S(4) = S(5) = S(6)$, etc.], rather than month effects appear to be adequate for the two smaller groups. Figure 5 shows that the standardized residuals are less heteroscedastic, although they are still smaller before 1985 than after. The heteroscedasticity is reduced because the multiplicative form of the month effects causes larger improvements for larger fitted values, even when standardized by the square root of the fitted values. We also see from Table 4 that the intergroup correlations in standardized residuals are smaller, although still all positive and still large enough to make the overall overdisperion larger than in the subgroups (Table 3). This reduction in correlations results from similarity of the estimated month effects in the different groups (not shown). In general, modeling monthly effects brings us substantially closer to meeting the Poisson assumption concerning infection rates, although it requires estimation of an uncomfortably large number of parameters. (Estimated autocorrelations in the residuals in Figure 5 are small, and modeling them does not produce any significant further reductions in the overdispersion.)

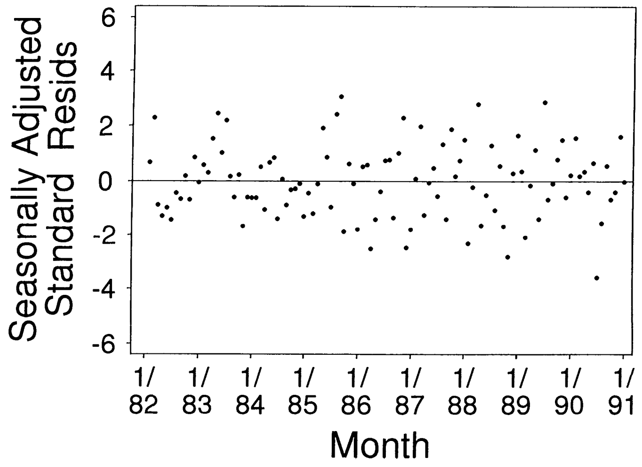

FIG. 5. *Standardized residuals $(y_j - \hat{y}_j)/\hat{y}_j^{1/2}$ for the fit to all adult/adolescent U.S. AIDS cases, with the $\hat{y}_j$ from a backcalculation model using the treatment model of incubation, $\lambda_\theta = \exp(8.5)$, $\lambda_\beta = \exp(15)$, underreporting assumed constant at 10% and monthly seasonal effects.*

The seasonal effects are also of some interest in themselves. Table 5 gives the standardized residuals from

TABLE 5
Actual reported counts, fitted values and standardized residuals from Figure 4, for three calendar months

	January			June			November		
Year	y_j	$\hat{y}_j$	$(y_j-\hat{y}_j)/\hat{y}_j^{1/2}$	y_j	$\hat{y}_j$	$(y_j-\hat{y}_j)/\hat{y}_j^{1/2}$	y_j	$\hat{y}_j$	$(y_j-\hat{y}_j)/\hat{y}_j^{1/2}$
1982	61	55	0.87	78	87	−0.96	135	135	0.02
1983	177	164	1.05	277	229	3.17	287	319	−1.79
1984	375	373	0.11	481	482	−0.06	586	631	−1.79
1985	738	722	0.58	919	893	0.88	1.043	1,123	−2.39
1986	1,333	1,267	1.87	1,597	1,477	3.13	1,565	1,789	−5.30
1987	2,064	1,986	1.74	2,340	2,231	2.31	2,309	2,440	−2.66
1988	2,583	2,607	−0.47	2,952	2,718	4.48	2,707	2,893	−3.47
1989	3,181	3,052	2.34	3,401	3,058	6.19	2,839	3.084	−4.41
1990	3,380	3,170	3.72	2,921	2,922	−0.01	2,559	2,610	−0.99

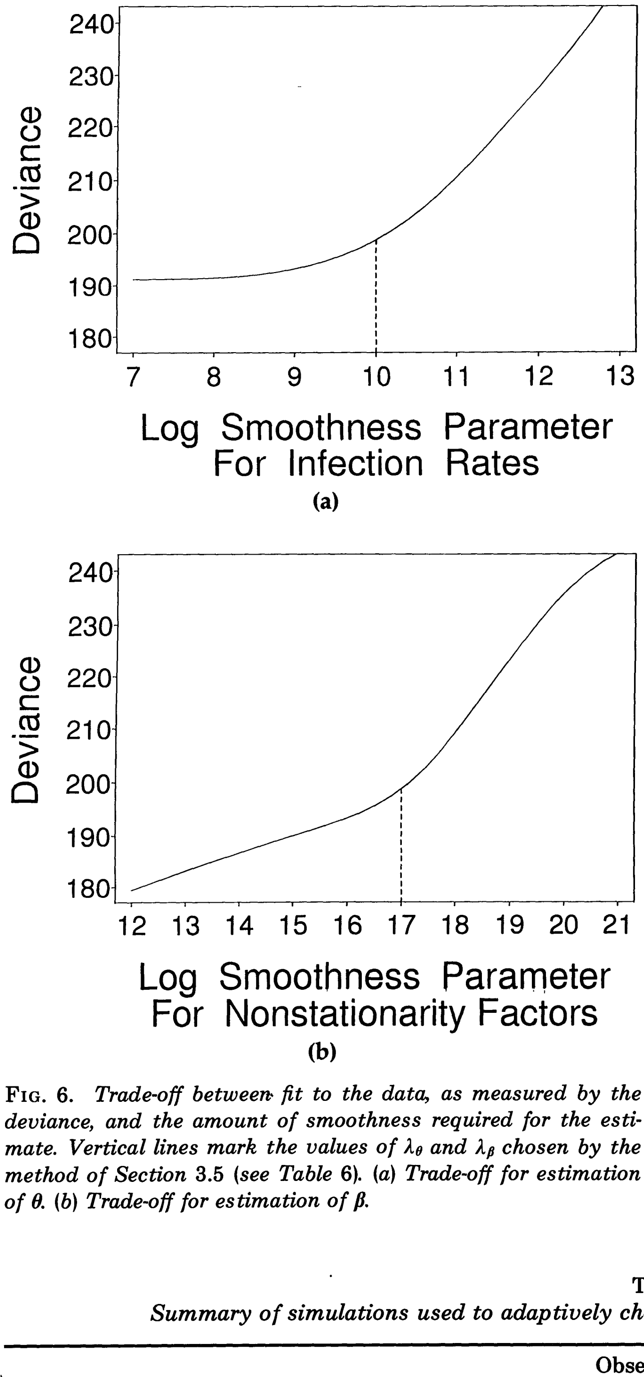

FIG. 6. *Trade-off between fit to the data, as measured by the deviance, and the amount of smoothness required for the estimate. Vertical lines mark the values of λ_θ and λ_β chosen by the method of Section* 3.5 (*see Table* 6). (*a*) *Trade-off for estimation of θ.* (*b*) *Trade-off for estimation of β.*

Figure 4 for three of the months. The generally increasing magnitude of the residuals in later years suggests that the multiplicative form of the month-effect model is appropriate for reducing heteroscedasticity. A downward trend over the calendar year is common to all four risk groups, but it is interrupted by a spike in June. This spike is fairly consistent across the groups, and Table 5 shows that the standardized residuals for June are quite high for five different years, so the spike appears to be real. We have not yet systematically examined estimated seasonal effects in subgroups defined by initial diagnosis, geographic area, and primary risk factor. We are currently developing methods for gauging the statistical significance of differences in seasonal effects in different groups in order to facilitate such a systematic investigation. This may provide further insights into possible causes of the month effects.

4.3 Smoothness Selection

The smoothness parameters λ_θ and λ_β are key ingredients in the estimation methods of Section 3. The choice of these parameters determines the fit to the data and the plausibility of the estimates. Here we examine the choice of these values for the treatment model of incubation. Subjective choices can be based on plots of the deviance as a function of the parameters (Figures 6a and 6b), along with assessments of whether the resulting estimates of θ and β are sufficiently smooth. Because these parameters are so important, however, it is also desirable to have an automatic, data-driven method for choosing their values. We therefore use the methods of Section 3.5 to obtain the results given in Table 6. If we go to a smaller value only when fewer than 10 of the 100 simulated improvements are larger than the observed improvement, then we end up choosing $\lambda_\theta = \exp(10)$ and $\lambda_\beta = \exp(17)$. This is the method that produces the choices in Table 1. If we go to a smaller value when the observed improvement is more

TABLE 6
Summary of simulations used to adaptively choose values for the smoothness parameters λ_θ and λ_β

Comparison	Observed improvement in deviance	Average of simulated improvements	Proportion larger than observed
Choice of λ_θ with λ_β fixed at exp(17) (*100 simulations each*)			
$\lambda_\theta = \exp(11)$ versus $\lambda_\theta = \exp(10.5)$	6.70	2.53	0.01
$\lambda_\theta = \exp(10.5)$ versus $\lambda_\theta = \exp(10)$	5.06	2.47	0.04
$\lambda_\theta = \exp(10)$ versus $\lambda_\theta = \exp(9.5)$	3.45	2.18	0.12
$\lambda_\theta = \exp(9.5)$ versus $\lambda_\theta = \exp(9)$	2.12	1.76	0.25
Choice of λ_β with λ_θ fixed at exp(10) (*100 simulations each*)			
$\lambda_\beta = \exp(18)$ versus $\lambda_\beta = \exp(17.5)$	5.92	2.38	0.03
$\lambda_\beta = \exp(17.5)$ versus $\lambda_\beta = \exp(17)$	4.50	2.37	0.03
$\lambda_\beta = \exp(17)$ versus $\lambda_\beta = \exp(16.5)$	3.18	2.21	0.10
$\lambda_\beta = \exp(16.5)$ versus $\lambda_\beta = \exp(16)$	2.24	1.92	0.29

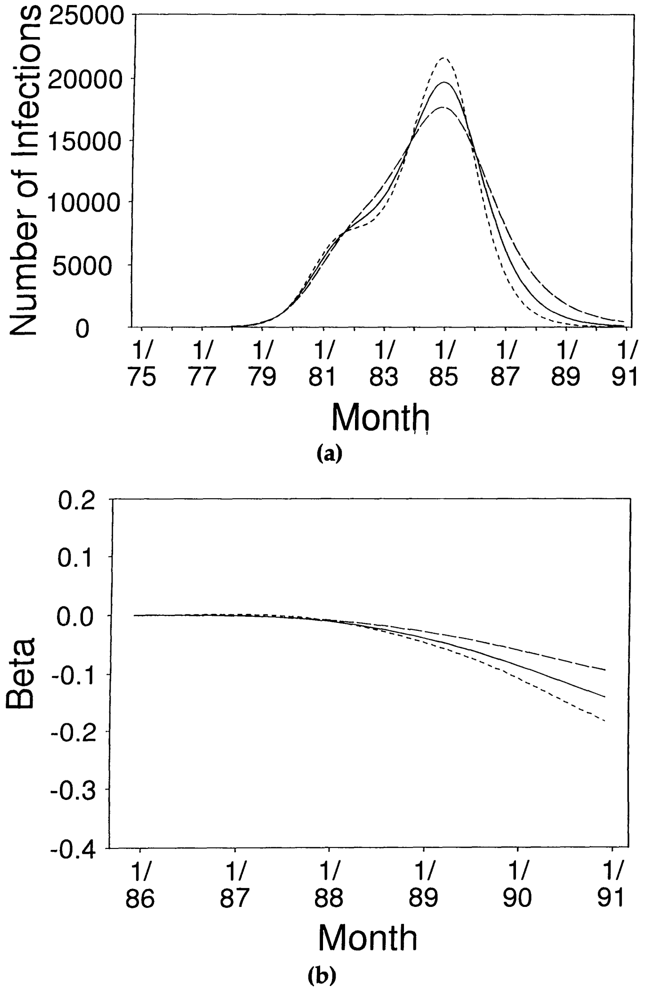

FIG. 7. *Influence of smoothness parameters on estimates. (a) Influence of λ_θ on estimation of θ, with λ_β fixed at* exp(17). *(Short dashed line)* λ_θ = exp(9). *(Solid line)* λ_θ = exp(10). *(Long dashed line)* λ_θ = exp(11). *(b) Influence of λ_β on estimation of β, with λ_θ fixed at* exp(10). *(Short dashed line)* λ_β = exp(16). *(Solid line)* λ_β = exp(17). *(Long dashed line)* λ_β = exp(18).

than twice the average of the simulated improvements, then we would choose λ_θ = exp(10) and λ_β = exp(17.5). These choices seem sensible in relation to the shapes of the curves shown in Figures 6a and 6b. Figures 7a and 7b show how the estimates vary depending on the amount of smoothness. The estimates seem plausible (and similar) throughout the ranges of λ_θ and λ_β examined in Table 6.

4.4 Confidence Intervals

Backcalculated estimates have some inherent stochastic uncertainty, in addition to the uncertainty resulting from imprecise knowledge of the inputs discussed in Section 4.1. We assess this uncertainty using the methods of Section 3.5 for the treatment model of incubation and 10% constant underreporting, with λ_θ = exp(10) and λ_β = exp(17). These confidence intervals reflect only the uncertainty that would be present if we actually knew the incubation and reporting inputs and the proper values for λ_θ and λ_β. Widths of 95% confidence intervals are not nearly as large as the uncertainties from the inputs described in Section 4.1. Using simulations, the intervals for estimated cumulative infections are about 10% wider without month effects than with and range from 48,000 wide for infections through 12/84 to 94,000 wide through 12/90. Those based on the information matrix are about 10% wider with month effects than without and range from 44,000 wide by 12/84 to 259,000 wide by 12/90.

5. DISCUSSION

The methods used here extend previous methods in some important ways, leading to improved fits to the observed diagnosis counts. In particular, modeling nonstationarity and seasonal effects results in significantly improved fits for all of the different inputs considered in Section 4.1. In addition, the methods used to determine the model complexity (λ_θ and λ_β) appear to give sensible results for the situations considered. There remains, however, substantial overdispersion, due at least in part to unexplained correlations between different subgroups. Also, models that include monthly effects estimate a fairly large number of parameters relative to the number of data points. The usefulness of the results is mainly limited by uncertainty about the AIDS incubation distribution and about how to allocate the estimated nonstationarity between diagnosis and reporting.

Uncertainty about what incubation distribution to use with the backcalculation method is known to be a major source of uncertainty in the resulting estimates (Gail and Rosenberg, 1992; Bacchetti et al., 1993). The large variations from different incubation inputs shown in Table 1 and Figures 1 and 2 reflect this fact. Nevertheless, the variations are not as large as in our previous report using the same inputs but not modeling seasonal effects or additional nonstationarity (Bacchetti et al., 1993). In this case, adaptively choosing the amount of nonstationarity, while accounting for seasonal effects, brings the estimates using the different incubation assumptions into better agreement. This reduction in disparity suggests that the estimates may also be more accurate and that failure to use flexible methods to obtain optimal fits to the observed data may introduce errors beyond those that are inevitable due to stochastic variability and uncertainty about backcalculation's inputs.

The advantages of modeling nonstationarity as described are balanced by a major drawback—its lack of interpretability. As noted in Section 3.3, the β_j's can be thought of as modifying either or both of the incubation and reporting distributions. With a highly concentrated infection curve, the β_j's could even be interpreted as correcting a misspecified stationary incubation distribution. In addition, reductions in the reporting factors R_j can be attributed to either longer delays or increasing underreporting. We note that using the full likelihood of Section 3.1 does not solve this problem, because the factors β_j do not appear in L_c (they cancel out), even if they are interpreted as modifying the R_{jk}'s. Thus, external information is necessary for interpretation of nonstationarity. L_c does provide information about changes in the shape of the reporting delay distribution, which can be used with additional assumptions (e.g., that reporting is complete after four years) to infer lengthening reporting delays (Harris, 1990; Brookmeyer and Liao, 1990a; Pagano et al., 1992b).

The interpretation of the nonstationarity has a large influence on projected numbers of future diagnoses, particularly because the β's must be extrapolated if they are interpreted as modifying incubation. If we were to assume that the β_j's modified only the R_j's, then projections would be much higher (up to 6,800) than shown in Tables 1 and 2, where we assume that only the diagnosis probabilities are modified. The variations shown in Table 2 further reinforce the potential importance of changes in reporting. Although we only examine very minor variations in underreporting, the scenarios with changing rates are more extreme for every estimate than those with constant rates. Larger variations than those examined, including changing rates, are probably consistent with what is known about underreporting (Buehler, Berkelman and Stehr-Green, 1992).

Even if we ignore the difficulties of interpretation, the backcalculated estimates of nonstationarity vary substantially depending on the assumed incubation (Figure 2) and are therefore very uncertain. This reflects the fact that diagnosis counts provide relatively weak information about nonstationarity and highlights the desirability of obtaining external estimates. Despite the difficulties noted in Section 2.2.3, estimates based on actual incubation data might be more reliable than either treatment models or estimated β's from nonstationary backcalculation models. With accurate external estimates, the need to estimate β's could be eliminated, which would improve the plausibility of backcalculated estimates while also reducing their statistical variability.

Estimates of overdispersion previously reported [e.g., a value of 5.6 in Rosenberg, Gail and Carroll (1992)] are large enough to cast doubt on the backcalculation model. Here, we reduce the estimated overdispersion, mainly by modeling seasonal effects (Table 3). Modeling month effects produces small changes in the estimates of θ and β, but failure to recognize seasonal patterns can influence projections—the pattern of decreasing effects over the course of the calendar year with a substantial spike in June implies that projections based on case series ending in March or June will be higher than those based on series ending in December. For the models considered here, accounting for seasonal effects has the disadvantage of requiring 11 additional parameters, but more parsimonious models may be possible. In addition, the causes of the effects are not yet clear. Examination of such effects for more subgroups may reveal additional information. For example, if the most pronounced increases between November–December and January–February are found among cases with Kaposi's sarcoma, then year-end procrastination and new-year initiative might be implicated, since this disease is more often mild enough to allow postponing diagnosis than are other AIDS manifestations. Examination of variation in monthly effects between different regions of the country may also provide insights.

The four risk groups examined here have similar estimated month effects, which accounts for much of the intergroup correlation of residuals from the models without seasonal effects (Table 4). Even after modeling month effects, however, we still see positive correlations in residuals, which make the overall estimate of overdispersion larger than for any of the subgroups. The cause of these correlations is also unknown. One possibility currently being investigated is that fluctuations in the number of workdays per month create similar fluctuations in different groups. These fluctuations would not be captured by seasonal effects because the number of workdays in a given month varies from year to year. Another example of a possible cause might be sporadic temporary increases in media coverage of AIDS issues, causing individuals in different groups to all seek health care in greater numbers, leading to temporary increases in diagnoses for all groups. Such temporary increases, however, would have to be very short-lived to avoid producing detectable serial correlations (which we do not see here).

The choices of the smoothness parameters λ_θ and λ_β directly influence the resulting model. Although a subjective choice can be based on examination of a range of possibilities (and such a range should be examined in any case), it is also desirable to have a data-driven method for choosing the smoothness. The method used here produces apparently appropriate choices, based on subjective assessments of the deviance plots in Figures 6a and 6b, the resulting estimates of θ and β, and the resulting fits to the observed diagnosis counts. In addition, we can see from Figures 7a and 7b that possible variation in the chosen values for λ_θ and λ_β would not lead to substantially different estimates of θ and β. We have not, however, systematically assessed the method's performance. This would require generating simulated data sets for which the optimal (in some sense) values of λ_θ and λ_β are known, and then employing the method to automatically choose values of λ_θ and λ_β and comparing the chosen values to the optimal ones. Because the method itself is so computationally intensive, this undertaking would require considerable computing time.

The confidence intervals briefly discussed in Section 4.4 show that backcalculation's stochastic uncertainty, when all of its assumptions are met, is much smaller than the uncertainty from the assumptions examined in Section 4.1, no matter how the intervals are calculated. Modeling seasonal effects has some influence on the widths of the intervals, but the largest influence on the widths depends on whether the intervals are based on simulations or on the information matrix.

A frequently used alternative to our likelihood formulation in Section 3.1 is to calculate corrected diagnosis counts $\tilde{y}_j = y_j/R_j$ and assume that each $\tilde{y}_j$ is Poisson with mean $\sum_{i=0}^{j} D_{ij}\theta_i$. Our log likelihood (2) can be rewritten, up to a constant, as

$$\sum_{j=0}^{n} R_j \left[\frac{y_j}{R_j}\log\left(\sum_{i=0}^{j} D_{ij}\theta_i\right) - \sum_{i=0}^{j} D_{ij}\theta_i\right],$$

which differs from the Poisson log likelihood for the imputed $\tilde{y}_j$ by downweighting the term for each $\tilde{y}_j$ by its reporting completeness, R_j. This downweighting accounts for the uncertainty in $\tilde{y}_j$'s that are largely imputed because of small R_j's, for example, very recent counts. Neglecting the weighting overstates what is really known and gives too much weight to recent, still preliminary, data. This, however, would not result in any large errors for the data considered here. For the treatment model with $\lambda_\vartheta = \exp(10)$ and $\lambda_\beta = \exp(17)$, estimates of cumulative infections are within 0.5% of the numbers in Table 1 and the asymptotic 95% confidence intervals are only 0–3% narrower using imputation. There is not much more influence on estimates of the β_j and $S(j)$, with the magnitude of the $\hat{\beta}_j$ about 5% larger with imputation and small differences in the $S(j)$'s, along with confidence intervals up to 5% narrower.

The β_j's could be used to capture a nonsmooth impact of the revision of the AIDS case definition in 1987 (CDC, 1987a), in addition to smooth secular trends. This would be accomplished by modifying the form of the roughness penalty for β so that it takes on smaller values for the terms with j near the time of the change. [This would be an alternative to the method of Rosenberg, Gail and Carroll (1992) that eliminates all new-definition cases who were diagnosed before the official time of the change and then estimates a parameter to account for the jump at the time of the change.] We have not done this here because there does not appear to be much systematic lack of fit around the time of the change, suggesting that retrospective reporting has substantially smoothed the impact of the revision on diagnosis counts. Thus, the globally smooth β's seem to adequately capture the influence of the change. Presumably, the estimated β's would have been more negative if the revision had not taken place.

A proposed revision of the AIDS case definition to include HIV seropositive persons with CD4 cell counts <200 μL would make backcalculation from AIDS diagnosis counts much more difficult. This revision would most likely have a larger impact on diagnosis rates than previous revisions, making backcalculation methods based primarily on previous estimates of incubation distributions questionable. It seems unlikely that modifications to allow for simple modeling of the change's influence as part of the backcalculation model, such as the nonstationarity estimated here or the single-definition change parameter estimated by Rosenberg, Gail and Carroll (1992), would be accurate enough to avoid substantial errors. Very little information would be gained from the postchange diagnosis counts if backcalculation were made flexible enough to accurately fit the counts without using external data. In our framework, AIDS-incidence projections from such backcalculations would rely on extrapolating large and rapidly changing β's and so would essentially be extrapolations themselves. Quantification of the impact of the change from external data, however, would be difficult, because diagnosis for many people will depend on how closely CD4 counts are monitored, making heavily monitored cohorts (that have provided much previous information on incubation) very unrepresentative of wider populations. In addition, variations in monitoring of CD4 counts between different HIV-positive populations (including variations in the proportions who know that they are positive) could make incubation distributions more heterogeneous. As we have argued previously (Bacchetti et al., 1993), heterogeneity could substantially increase the already considerable uncertainty about incubation.

Backcalculation using counts of HIV-related deaths avoids some of these difficulties. Definitional changes are not an issue (although ascertainment of HIV relatedness could change), and death is an important and biologically meaningful endpoint. Cohorts are more likely to be representative in terms of their times from infection to death than to a detected drop in CD4 count, so more information is available about time to death. In addition, heterogeneity in time to death may be less severe, particularly if early treatment increases AIDS-free survival more than overall survival. Although death surveillance may not currently be as good as diagnosis surveillance, death certificates are a valuable resource that could be used to study carefully the completeness of death surveillance and to improve it.

The proposed definition change gives added importance to alternatives to backcalculation. Brookmeyer and Liao (1990b) have proposed methods for synthesizing backcalculation and prevalence estimates from surveys. Unfortunately, direct surveys of HIV prevalence are hampered by the difficulty of obtaining samples with representative mixes of HIV positives and negatives, particularly when subjects self-select into the sample. A recently proposed method avoids this difficulty by focusing only on positives, deducing past rates of infection from the mix of laboratory-marker values (Satten and Longini, 1992). This requires samples of seropositives who are representative of seropositives generally at the same time. How such samples could be obtained, however, is not clear. In addition, the method is, like backcalculation, indirect, and it depends heavily on detailed knowledge of how marker values change over time in the population of interest. Such detailed knowledge is probably even more difficult to obtain than knowledge of incubation distributions (Jewell and Kalbfleisch, 1992).

A more direct alternative would be to observe seroconversion rates in samples of seronegatives. This requires that the sampled seronegatives be representative of seronegatives generally at the same time, but they can be part of a larger sample that either over- or underrepresents seronegatives. Obtaining a representative sample could be made easier by employing anonymous, nonvoluntary methods, in which initial samples are stored untested, with the subject's identifying information, until an additional later specimen is obtained, at which time the identifiers are removed and the pair of specimens is tested. Such an approach could, for example, be used for many of the anonymous surveys currently used to assess prevalence (Pappaioanou et al., 1990). This is a more powerful method for measuring seroconversion rates than repeated cross-sectional surveys, and it does not rely on obtaining samples with representative prevalence. A weakness of this approach is that the subjects providing repeat samples may not be representative of those who provide only one. Nevertheless, this method directly attacks the area where backcalculation is weakest, estimation of recent infection rates, so it may be a valuable complement to backcalculation.

The value of the backcalculation technique in the assessment of the future course of the AIDS epidemic is evident from its past applications. The method possesses a clear advantage over simple extrapolation techniques or complex transmission models in that with apparently modest assumptions, and some available incubation input, the pattern of HIV infections can be reconstructed in a systematic manner that allows appropriate consideration of stochastic uncertainties. The issues regarding the methodology, therefore, center on whether the necessary assumptions hold up to scrutiny, whether the required external data are sufficiently valid and the extent to which simple interpretations of the results of backcalculation are appropriate. The answer to these questions is necessarily complex, as we have shown. To a great extent, we do not currently possess the information on the infection process, incubation period distributions and aspects of reporting delay and underreporting that we would like. It is therefore incumbent on users of backcalculation to allow for these modeling uncertainties in reporting and interpreting backcalculation estimates. This can be done by performing sufficiently broad sensitivity analyses, possibly including simultaneous variation of several uncertain inputs. This, however, will be computationally demanding, and it may be difficult to delineate broad alternatives for the inputs, as noted in Section 2.2.4 for the case of incubation. Alternatively, detailed Bayesian approaches might prove fruitful. Acknowledgment of uncertainty can complicate even the simplest interpretations one would like to draw from backcalculation. For example, the evidence for nonstationarity of incubation distributions or quantification of the "treatment effect" is significantly affected by the assumptions regarding the underlying incubation and reporting delay distributions and chronological changes in underreporting rates. Nevertheless, the method of backcalculation will continue to be useful for AIDS projections and prevalence estimation. Further, as the technique is critically applied, it will continue to raise important questions about AIDS surveillance and HIV incubation that might otherwise be missed.

Challenges to the effective use of backcalculation will continue as our knowledge of the natural history of HIV infection grows and as factors within the surveillance system change. Further, the nature of the uncertainties surrounding AIDS projections and HIV prevalence estimates dictate both continued vigilance with regard to the quality of the surveillance systems and the need to collect precise natural history data, as well as a careful use of other methods and data sources to provide validation and context for backcalculation analyses.

ACKNOWLEDGMENTS

This research was supported by grants R29-AI31331, R01-AI29162, R01-DA04722 and R29-GM45543 from the National Institutes of Health.

REFERENCES

Anderson, R. M., Medley, G.F., May, R. M. and Johnson, A. M. (1986). A preliminary study of the transmission dynamics of the Human Immunodeficiency Virus (HIV), the causative agent of AIDS. *IMA J. Math. Appl. Med. Biol.* 3 229–263.

Bacchetti, P. (1990). Estimating the incubation period of AIDS by comparing population infection and diagnosis patterns. *J. Amer. Statist. Assoc.* **85** 1002–1008.

Bacchetti, P. and Jewell, N. P. (1991). Nonparametric estimation of the incubation period of AIDS based on a prevalent cohort with unknown infection times. *Biometrics* **47** 947–960.

Bacchetti, P., Segal, M. R., Hessol, N. A. and Jewell, N. P. (1993). Impact of uncertain AIDS incubation period on reconstruction of HIV epidemics. *Proc. Nat. Acad. Sci. U.S.A.* To appear.

Bacchetti, P., Segal, M. R. and Jewell, N. P. (1992a). Uncertainty about the incubation period of AIDS and its impact on backcalculation. In: *AIDS Epidemiology: Methodological Issues* (N. P. Jewell, K. Dietz and V. T. Farewell, eds.) 61–80. Birkhäuser, Boston.

Bacchetti, P., Segal, M. R. and Jewell, N. P. (1992b). Backcalculation of HIV infection rates. Technical Report, Dept. Epidemiology and Biostatistics, Univ. California, San Francisco.

Becker, N. G. and Marschner, I. C. (1993). A method for estimating the age-specific relative risk of HIV infection from AIDS incidence data. *Biometrika* **80** 165–178.

Becker, N. G., Watson, L. F. and Carlin, J. B. (1991). A method of nonparametric back-projection and its application to AIDS data. *Statistics in Medicine* **10** 1527–1542.

Biggar, J. (1990). AIDS incubation in 1891 HIV seroconverters from different exposure groups. *AIDS* **4** 1059–1066.

Brookmeyer, R. (1991). Reconstruction and future trends of the AIDS epidemic in the United States. *Science* **253** 37–42.

Brookmeyer, R. and Damiano, A. (1989). Statistical methods for short-term projections of AIDS incidence. *Statistics in Medicine* **8** 23–34.

Brookmeyer, R. and Gail, M. H. (1986). Minimum size of the acquired immunodeficiency syndrome (AIDS) epidemic in the United States. *Lancet* **2** 1320–1322.

Brookmeyer, R. and Gail, M. H. (1988). A method for obtaining short-term predictions and lower bounds on the size of the AIDS epidemic. *J. Amer. Statist. Assoc.* **83** 301–308.

Brookmeyer, R. and Goedert, J. J. (1989). Censoring in an epidemic with an application to hemophilia-associated AIDS. *Biometrics* **45** 325–335.

Brookmeyer, R. and Liao, J. (1990a). The analysis of delays in disease reporting: methods and results for the acquired immunodeficiency syndrome. *American Journal of Epidemiology* **132** 355–365.

Brookmeyer R. and Liao, J. (1990b). Statistical modelling of the AIDS epidemic for forecasting health care needs. *Biometrics* **46** 1151–1163.

Brookmeyer, R. and Liao, J. (1992). Statistical methods for reconstructing infection curves. In *AIDS Epidemiology: Methodological Issues* (N. P. Jewell, K. Dietz and V. T. Farewell, eds.) 39–60. Birkhäuser, Boston.

Buehler, J. W., Berkelman, R. L. and Stehr-Green, J. K. (1992). The completeness of AIDS surveillance. *Journal of Acquired Immune Deficiency Syndromes* **5** 257–264.

Buehler, J. W., Devine, O. J., Berkelman, R. L. and Chevarley, F. (1990). Impact of the human immunodeficiency virus epidemic. *American Journal of Public Health* **80** 1080–1086.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17** 453–510.

Byers, R. H., Morgan, W.M., Darrow, W. W., Doll, L., Jaffe, H. W., Rutherford, G., Hessol, N. and O'Malley, P. M. (1988). Estimating AIDS infection rates in the San Francisco cohort. *AIDS* **2** 207–210.

Centers for Disease Control (1982). Update on acquired immune deficiency syndrome (AIDS)—United States. *Morbidity and Mortality Weekly Reports* **31** 507–508.

Centers for Disease Control (1985). Revision of the case definition of acquired immune deficiency syndrome for national reporting—United States. *Morbidity and Mortality Weekly Reports* **34** 373–375.

Centers for Disease Control (1987a). Revision of the CDC surveillance case definition for acquired immunodeficiency syndrome. *Morbidity and Mortality Weekly Reports* **36** 3S–15S.

Centers for Disease Control (1991a). Pilot study of a household survey to determine HIV seroprevalence. *Morbidity and Mortality Weekly Reports* **40** 1–5.

Day, N.E., Gore, S. M., McGee, M. A. and South, M. (1989). Predictions of the AIDS epidemic in the U.K.: The use of the back projection method. *Philos. Trans. Roy. Soc. London Ser. B* **325** 123–134.

De Gruttola, V. and Fineberg, H. V. (1989). Estimating prevalence of HIV infection: Considerations in the design and analysis of a national seroprevalence survey. *Journal of Acquired Immune Deficiency Syndromes* **2** 472–480.

De Gruttola, V. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics* **45** 1–11.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans.* SIAM, Philadelphia.

Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika* **72** 45–58.

Gail, M. H. and Brookmeyer, R. (1988). Methods for projecting the course of the AIDS epidemic. *Journal of the National Cancer Institute* **80** 900–911.

Gail, M. H. and Rosenberg, P. S. (1992). Perspectives on using backcalculation to estimate HIV prevalence and project AIDS incidence. In *AIDS Epidemiology: Methodolgical Issues* (N. P. Jewell, K. Dietz and V. T. Farewell, eds.) 1–38. Birkhäuser, Boston.

Gail, M. H., Rosenberg, P. S. and Goedert, J. J. (1990). Therapy may explain recent deficits in AIDS incidence. *Journal of Acquired Immune Deficiency Syndromes* **3** 296–306.

Green, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation. *J. Roy. Statist. Soc. Ser. B* **52** 443–452.

Harris, J. E. (1990). Reporting delays and the incidence of AIDS. *J. Amer. Statist. Assoc.* **85** 915–924.

Healy, M. J. R. and Tillet, H. E. (1988). Short-term extrapolation of the AIDS epidemic. *J. Roy. Statist. Soc. Ser. A* **151** 50–65.

Hessol, N. A., Lifson, A. R., O'Malley, P. M., Doll, L. S., Jaffe, H. W. and Rutherford, G. W. (1989). Prevalence, incidence, and progression of human immunodeficiency virus infection in homosexual and bisexual men in hepatitis B vaccine trials, 1978–1988. *American Journal of Epidemiology* **130** 1167–1175.

Isham, V. (1988). Mathematical modeling of the transmission dynamics of HIV infection and AIDS: A review (with discussion). *J. Roy. Statist. Soc. Ser. A* **151** 5–30; 120–123.

Jewell, N. P. (1990). Some statistical issues in studies of the epidemiology of AIDS. *Statistics in Medicine* **9** 1387–1416.

Jewell, N. P. and Kalbfleisch, J. D. (1992). Marker models in survival analysis and applications to issues associated with AIDS. In *AIDS Epidemiology: Methodological Issues* (N. P. Jewell, K. Dietz and V. T. Farewell, eds.) 211–230. Birkhäuser, Boston.

Kalbfleisch, J. D. and Lawless, J. F. (1988). Estimating the incubation period for AIDS patients. *Nature* **333** 504–505.

Kalbfleisch, J. D. and Lawless, J. F. (1989). Inference based on retrospective ascertainment: An analysis of the data on transfusion-related AIDS. *J. Amer. Statist. Assoc.* **84** 360–372.

Karon, J. M., Devine, O. J. and Morgan, W. M. (1989). Predicting AIDS incidence by extrapolating from recent trends. In: *Mathematical and Statistical Approaches to AIDS Epidemiology* (C. Castillo-Chavez, ed.) 58-88. Springer, New York.

Kuo, J-M., Taylor, J. M. G. and Detels, R. (1991). Estimating the AIDS incubation period from a prevalent cohort. *American Journal of Epidemiology* **133** 1050-1057.

Lagakos, S. W., Barraj, L. M. and De Gruttola, V. (1988). Nonparametric analysis of truncated survival data with applications to AIDS. *Biometrika* **75** 515-523.

Lawless, J. F. and Sun, J. (1992). A comprehensive backcalculation framework for the estimation and prediction of AIDS cases. In *AIDS Epidemiology: Methodological Issues* (N. P. Jewell, K. Dietz and V. T. Farewell, eds.) 81-104. Birkhäuser, Boston.

Liang, K-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13-22.

Lui, K-J., Darrow, W. W. and Rutherford, G. W. (1988). A model-based estimate of the mean incubation period for AIDS in homosexual men. *Science* **240** 1333-1335.

Lui, K-J., Lawrence, D. N., Morgan, W. M., Peterman, T. A., Haverkos, H. W. and Bregman, D. J. (1986). A model-based approach for estimating the mean incubation period of transfusion-associated acquired immunodeficiency syndrome. *Proc. Nat. Acad. Sci. U.S.A.* **83** 3051-3055.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* Chapman and Hall, London.

Morgan, W. M. and Curran, J. W. (1986). Acquired immunodeficiency syndrome: Current and future trends. *Public Health Reports* **101** 459-465.

Muñoz, A., Wang, M-C., Bass, S., Taylor, J. M. G., Kingsley, L. A., Chmiel, J. S. and Polk, B. F. (1989). Acquired immunodeficiency syndrome (AIDS)-free time after Human Immunodeficiency Virus Type 1 (HIV-1) seroconversion in homosexual men. *American Journal of Epidemiology* **130** 530-539.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statist. Sci.* **1** 502-527.

Pagano, M., De Gruttola, V., MaWhinney, S. and Tu, X. M. (1992a). The HIV epidemic in New York City: Statistical methods for projecting AIDS incidence and prevalence. In *AIDS Epidemiology: Methodological Issues* (N. P. Jewell, K. Dietz and V. T. Farewell, eds.) 123-142. Birkhäuser, Boston.

Pagano, M., Tu, X. M., De Gruttola, V. and MaWhinney, S. (1992b). Analysis of censored and truncated data: Estimating reporting delay distributions and AIDS incidence from surveillance data. Technical report, Harvard School of Public Health, Harvard Univ.

Pappaioanou, M., Dondero, T. J., Peterson, L. R., Onorato, I. M., Sanchez, C. D. and Curran, J. W. (1990). The family of HIV seroprevalence surveys: Objectives, methods, and uses of sentinel surveillance for HIV in the United States. *Public Health Reports* **105** 113-119.

Rosenberg, P. S., Biggar, R. J., Goedert, J. J. and Gail, M. H. (1991a). Backcalculation of the number with human immunodeficiency virus infection in the United States. *American Journal of Epidemiology* **133** 276-285.

Rosenberg, P. S. and Gail, M. H. (1990). Uncertainty in estimates of HIV prevalence derived by backcalculation. *Annals of Epidemiology* **1** 105-115.

Rosenberg, P. S. and Gail, M. H. (1991). Backcalculation of flexible linear models of the human immunodeficiency virus infection curve. *J. Roy. Statist. Soc. Ser. C* **40** 269-282.

Rosenberg, P. S., Gail, M. H. and Carroll, R. J. (1992). Estimating HIV prevalence and projecting AIDS incidence in the United States: A model that accounts for therapy and changes in the surveillance definition of AIDS. *Statistics in Medicine* **11** 1633-1655.

Rosenberg, P. S., Gail, M. H. and Pee, D. (1991). Mean square error of estimates of HIV prevalence and short-term AIDS projections derived by backcalculation. *Statistics in Medicine* **10** 1167-1180.

Satten, G. A. and Longini, I. (1992). Estimation of incidence of HIV infection using cross-sectional marker surveys. *Biometrics.* To appear.

Segal, M. R. and Bacchetti, P. (1990). Deficits in AIDS incidence. *Journal of Acquired Immune Deficiency Syndromes* **3** 832-833.

Segal, M. R., Bacchetti, P. and Jewell, N. P. (1993). Variances for maximum penalized likelihood estimates obtained via the EM algorithm. *J. Roy. Statist. Soc. Ser. B.* To appear.

Silverman, B.W., Jones, M. C., Wilson, J. D. and Nychka, D. W. (1990). A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography. *J. Roy. Statist. Soc. Ser. B* **52** 271-324.

Solomon, P. J. and Wilson, S. R. (1990). Accommodating change due to treatment in the method of back projection for estimating HIV infection incidence. *Biometrics* **46** 1165-1170.

Taylor, J. M. G. (1989). Models for the HIV infection and AIDS epidemic in the United States. *Statistics in Medicine* **8** 45-58.

Taylor, J. M. G., Kuo, J-M. and Detels, R. (1991). Is the incubation period of AIDS lengthening? *Journal of Acquired Immune Deficiency Syndromes* **4** 69-75.

Wang, M-C., Jewell, N. P. and Tsai, W. Y. (1986). Asymptotic properties of the product limit estimate under random truncation. *Ann. Statist.* **14** 1597-1605.

Zeger, S. L., See, L.-C. and Diggle, P. J. (1989). Statistical methods for monitoring the AIDS epidemic. *Statistics in Medicine* **8** 3-21.