Modeling Reporting Delays and Reporting Corrections in Cancer Registry Data

Douglas N. MIDTHUNE, Michael P. FAY, Limin X. CLEGG, and Eric J. FEUER

The Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute is an authoritative source of cancer incidence statistics in the United States. The SEER program is a consortium of population-based cancer registries from different areas of the country. Each registry is charged with collecting data on all cancers that occur within its geographic area. As with any disease registry, there is a delay between the time that the disease (cancer) is first diagnosed and the time that it is reported to the registry. The SEER program has allowed for reporting delays of up to 19-months before releasing data for public use. Nevertheless, additional cases are discovered after the 19-month delay, and these cases are added in subsequent releases of the data. Further, any errors discovered are corrected in subsequent releases. Such reporting delays and corrections typically lead to underestimation of the cancer incidence rates in recent diagnosis years, making it difficult to monitor trends. In this article we study models that account for reporting delays and corrections in predicting eventual cancer counts for a diagnosis year from the preliminary counts. Previous models of this type have been studied, especially as applied to AIDS registries. We offer several additions to existing models. First, we explicitly model the reporting corrections. Second, we model the delay distribution with very general models, combining aspects of previous nonparametric-like models (i.e., models that have a separate parameter for each delay time) with more parametric models. Third, we allow random reporting-year effects in the model. Practical issues of model selection and how the data are classified are also discussed, particularly how the definition of a reporting correction may change depending on how subpopulations are defined. An example with SEER melanoma data is studied in detail.

KEY WORDS: Cancer surveillance; Delay-adjusted rates; Incurred but not reported; Random effects; Surveillance, Epidemiology, and End Results program; Truncated data.

1. INTRODUCTION

Cancer registries have the ambitious goal of recording every cancer diagnosed in a particular population over a given time period. From such data, one can calculate cancer incidence rates in the population and monitor trends over time. In the United States the authoritative source for trends in cancer incidence is the National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) program, a consortium of U.S. cancer registries. SEER began collecting data in 1973, expanded by 1992 to cover approximately 14% of the U.S. population, and expanded further in 2001 to cover about 26% of the population. The remainder of the U.S. is now covered by a system of state registries sponsored by the Centers for Disease Control and Prevention (CDC) National Program of Cancer Registries (NPCR). In this article we use data from the nine registries that have been part of SEER since 1975 and that cover approximately 10% of the U.S. population.

Although the SEER registries are not a random sample of the U.S. population, they were selected to be fairly representative of the U.S., while at the same time providing an oversample of minority populations and meeting the high standards set by SEER for timely and accurate reporting. The population of the nine SEER registries in 2000 was 76.8% white, 12.2% black, 9.6% Asian/Pacific Islander, and 1.4% American Indian/Alaska Native, whereas in the U.S. population, these percentages were 81.7%, 13.0%, 4.2%, and 1.1%. Similarly, the SEER population in 2000 was 88% urban, 85% high school graduates, and 11% living below the poverty line, whereas in the U.S. population these percentages were 79%, 80%, and 13%. The representativeness of SEER has been assessed by Frey, McMillen, Cowan, Horm, and Kessler (1992), who compared cancer mortality rates and trends in the combined nine SEER registries to those in the U.S. as a whole and found them to be comparable for most cancers. Even for cancers for which SEER cannot be considered representative of the U.S. population, SEER is still an invaluable resource for identifying emerging trends in a fixed set of catchment areas, and as such is extensively referenced by cancer control planners and policy makers. Trends in SEER are part of an annual report by the NCI (Ries et al. 2004) used by the director to brief Congress concerning progress against cancer, and are also part of the Annual Report to the Nation on Cancer (Jemal et al. 2004), a collaborative publication by the American Cancer Society, the NCI, the CDC, and the North American Association of Central Cancer Registries.

SEER is the gold standard for timeliness and accuracy of reporting in cancer registries in the U.S. (Fritz 2001). SEER registries have extensive infrastructure to enable them to provide high-quality data in a timely fashion. Registry staff make periodic visits to some or all of the facilities that can potentially contribute cases. Unvisited facilities are required to report cases to the registries, which monitor and address noncompliance. The registries periodically audit facilities to assess completeness of reporting and data quality. Even in SEER, however, delay time—the time between diagnosing a cancer and reporting it to the NCI—can lead to underreporting. During the period analyzed in this article, the SEER program allowed for a 19-month delay before diagnosed cancers were due to be submitted to the NCI (currently, 22 months are allowed). For example, cancers diagnosed in 1997 were required to be reported by August 1, 1999. Nevertheless, additional cases are inevitably discovered after the allowed delay period and are added in subsequent years. Some of the determinants of delay time tend to be systematic and predictable, such as the periodic nature of visits to a facility by registry staff, the lag time before new facilities can be identified and integrated into the reporting process, and

Douglas N. Midthune is Mathematical Statistician, Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD 20892 (E-mail: *dm76q@nih.gov*). Michael P. Fay is Mathematical Statistician, Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda, MD 20892. Limin X. Clegg is Mathematical Statistician and Eric J. Feuer is Chief, Statistical Research and Applications Branch, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD 20892.

Journal of the American Statistical Association
March 2005, Vol. 100, No. 469, Applications and Case Studies
DOI 10.1198/016214504000001899

the changing nature of the health care system (e.g., the trend from hospital to outpatient care). Other determinants tend to be sporadic and unpredictable, such as the timeliness of reporting from a particular facility that has agreed to report cases, changes in staff, and special research studies based on SEER cases that identify inconsistencies after intensive medical record review.

Reporting delays can make monitoring trends difficult, because the most recent diagnosis years are most prone to underreporting. For example, Horm and Kessler (1986) reported that the age-adjusted lung cancer incidence rate for white men had declined by 4% from 1982 to 1983. Such a decline might have indicated the beginning of a long-awaited downturn in lung cancer rates consistent with an earlier decline in smoking rates. The current data, however, show that the rate was generally flat from 1980 to 1990, followed by a steady decline, and that the report of a decline in 1983 was due primarily to reporting delay (Ries et al. 2004). Today, SEER researchers are aware of such potential problems and are careful not to overinterpret apparent changes in trends in the most recent diagnosis years.

Statistically speaking, the problem is to estimate the distribution of the delay time when the observed delay times are right-truncated, because only cases reported before some given time are observed. This problem has been studied in many other applications: to estimate the incubation period for AIDS, that is, the time from HIV infection to AIDS (e.g., Kalbfleisch and Lawless 1989); to predict AIDS cases from reported cases (Sellero et al. 1996); to predict claims made on products under warranty (Kalbfleisch, Lawless, and Robinson 1991); and to predict "incurred but not reported" (IBNR) insurance claims (e.g., Verrall 2000; Doray 1996). Another potential application, suggested by a referee, is to allow an interim data-monitoring committee to predict the number of events in an ongoing clinical trial when there is a lag in reporting.

Nonparametric methods for estimating a delay distribution can be derived from either a conditional likelihood or an unconditional likelihood (Kalbfleisch and Lawless 1989; Harris 1990), or from a random truncation model (Keiding and Gill 1990; Herbst 1999). Cox and Medley (1989) and Doray (1996) studied parametric continuous-time models, Pagano, Tu, De Gruttola, and MaWhinney (1994) studied parametric discrete-time models, and Kalbfleisch and Lawless (1991) discuss both types. Lawless (1994) proposed a random-effects model to address the problem of overdispersion.

Despite the ample literature on this problem, two main aspects of modeling delay times have not been adequately addressed and are important for our application. First, the registry may contain nonnegligible reporting errors, requiring that the model be able to handle such errors and their periodic correction. A common error occurs when a cancer is reported from several different data sources and is reported as multiple cancers until the registry is able to match records to correct the duplication. Errors can also occur due to uncertainty as to whether a cancer is primary or metastatic. For example, a cancer originally coded as a liver cancer may later be deleted because it was found to be a breast cancer that had metastasized to the liver. In SEER, only primary cancers are included. The net effect of modeling the reporting delays and corrections may be a positive or negative change in the cancer count for a particular subpopulation; such changes cannot be modeled with simple multinomial or Poisson distributions. To our knowledge, Verrall (2000) described the only model that addresses these concerns. Verrall proposed a method for IBNR estimation when there are negative incremental claims by modeling the conditional cumulative claims as normally distributed. Because our incremental changes are often small counts, this normal approximation is not adequate.

The second aspect important to our application is the model selection process. The *SEER Cancer Statistics Review* (Ries et al. 2004), published annually, reports for different subgroups of the population the incidence rates for many different cancer sites, each of which may have a different delay distribution. For example, we estimate that more than 99% of diagnosed lung cancers are reported within 4 years of diagnosis but only 86% of diagnosed melanoma cases are reported within 4 years, and that it takes more than 10 years for 99% of the melanomas to be reported. The longer reporting delays for melanoma may be explained in part by the fact that the percentage of cases diagnosed outside of a hospital is much higher for melanoma (26% in 1995) than for lung cancer (5% in 1995), and that cases diagnosed in hospitals tend to be reported more quickly than cases diagnosed in nonhospital settings such as physicians' offices or pathology labs. Pagano et al. (1994) offered graphical aids for the model selection process, but we require a more automatic method because we must produce reproducible reports for many different subgroups and cancer sites annually.

We propose a new reporting model that explicitly models both reporting delays and corrections. An earlier version of this model has been previously described by Clegg, Feuer, Midthune, Fay, and Hankey (2002) and used to calculate delay- and correction-adjusted incidence rates (Ries et al. 2004). In Section 2 we describe the model, considering both fixed-effects and random-effects models. In Section 3 we present simulations to assess the model and evaluate the use of the Akaike information criterion (AIC) for model selection. In Section 4 we apply the model to SEER melanoma incidence data.

2. METHODS

2.1 Data Structure

Table 1 summarizes the SEER invasive melanoma data for whites for diagnosis years 1981–1997 and reporting years 1983–1999. For each diagnosis year, the table shows the number of cases reported in the initial reporting year and the number of cases added to and subtracted from the previous count for each subsequent reporting year.

We divide the SEER population into subpopulations based on registry and the usual subgroups used for reporting, that is, one for each combination of levels of the following variables: year of diagnosis (17 years), gender, race (for melanoma we use only whites, because melanoma is rare among other racial groups), and 5-year age groups (0–4, 5–9, ..., 80–84, ≥85). We include only the diagnosis years 1981–1997 and the nine registries having data back to 1981 (San Francisco-Oakland, Connecticut, Metropolitan Detroit, Hawaii, Iowa, New Mexico, Seattle, Utah, and Metropolitan Atlanta), because 1981 is the earliest year for which we have complete data on reporting delays and corrections. Thus our example dataset has 9 registries $\times 17$ diagnosis years $\times 2$ genders $\times 18$ age groups $= 5{,}508$ subpopulations.

Table 1. SEER Invasive Melanoma Incidence for Whites

Reporting	*Diagnosis year*																
year	*1981*	*1982*	*1983*	*1984*	*1985*	*1986*	*1987*	*1988*	*1989*	*1990*	*1991*	*1992*	*1993*	*1994*	*1995*	*1996*	*1997*
1983	1,817																
1984	+84 −51	1,797															
1985	+31 −12	+51 −18	1,762														
1986	+40 −23	+78 −21	+118 −31	1,861													
1987	+25 −28	+38 −46	+36 −66	+89 −57	2,117												
1988	+38 −10	+53 −12	+79 −19	+78 −10	+92 −24	2,132											
1989	+40 −4	+34 −5	+55 −5	+62 −22	+92 −16	+212 −44	2,334										
1990	+18 −9	+23 −13	+30 −18	+36 −13	+44 −21	+57 −12	+104 −33	2,189									
1991	+43 −7	+84 −8	+102 −11	+90 −13	+89 −16	+120 −20	+168 −33	+234 −20	2,530								
1992	+10 −3	+14 −2	+10 −4	+16 −7	+42 −8	+76 −14	+66 −10	+93 −19	+154 −17	2,615							
1993	+12 −33	+5 −58	+15 −80	+17 −57	+27 −71	+18 −77	+20 −82	+37 −84	+37 −149	+73 −146	2,721						
1994	+20 −4	+33 −3	+8 −4	+9 −3	+13 −2	+10 −3	+33 −11	+65 −5	+72 −11	+123 −14	+152 −26	2,689					
1995	+34 −3	+61 −3	+89 −4	+92 −5	+76 −2	+97 −7	+100 −5	+117 −5	+165 −7	+190 −13	+251 −19	+456 −50	2,960				
1996	+4 −6	+5 −1	+6 −6	+11 −7	+8 −8	+11 −6	+21 −9	+19 −9	+25 −15	+34 −14	+36 −19	+69 −24	+146 −24	3,261			
1997	+5 −1	+11 −2	+5 −3	+12 −2	+17 −7	+28 −5	+21 −8	+16 −6	+32 −11	+29 −12	+31 −14	+47 −17	+86 −28	+172 −46	3,514		
1998	+7 −2	+6 −4	+8 −4	+8 −5	+9 −3	+22 −2	+12 −4	+17 −3	+14 −1	+21 −10	+15 −7	+28 −10	+30 −8	+51 −15	+124 −35	3,716	
1999	+22 −11	+27 −26	+31 −22	+23 −16	+40 −32	+44 −29	+43 −28	+18 −8	+31 −11	+21 −9	+24 −12	+44 −27	+58 −23	+71 −23	+106 −28	+227 −53	3,860
Initial	1,817	1,797	1,762	1,861	2,117	2,132	2,334	2,189	2,530	2,615	2,721	2,689	2,960	3,261	3,514	3,716	3,860
Total +	433	523	592	543	549	695	588	616	530	491	509	644	320	294	230	277	
Total −	207	222	277	217	210	219	223	159	222	218	97	128	83	84	63	53	
Net	2,043	2,098	2,077	2,187	2,456	2,608	2,699	2,646	2,838	2,888	3,133	3,205	3,197	3,471	3,681	3,940	3,860

NOTE: Number of cases initially reported for diagnosis years 1981–1997, and number added to (+) and subtracted from (−) previous count in each subsequent reporting year.

Let S be the number of subpopulations, and let J_i be the number of observed reporting years for subpopulation i. Let the delay time be the reporting year minus year of diagnosis, and let the delay times in subpopulation i be $t_1, t_2, \ldots, t_{J_i}$. For the ith subpopulation, we observe $x_{ij}, y_{ij}, j = 1, 2, \ldots, J_i$, where x_{ij} (y_{ij}) is the number of reported cases that were added to (removed from) subpopulation i at delay time t_j. The delay times (rounded to the nearest year) for SEER are $t_1 = 2, t_2 = 3, \ldots$. The data are right-truncated in that only cancers reported and corrections made by the current reporting year are observed.

Consider one example showing how reporting corrections can affect the subpopulation rates. Suppose that in reporting year m, the age at diagnosis for a previously reported case diagnosed in year k is changed from 43 to 48, that is, moved from the 40–44 age group to the 45–49 age group. Because this change causes the case to be moved, say, from subpopulation i to subpopulation i', it causes both y_{ij} and $x_{i'j}$ to increase by 1, where $t_j = m - k$, and we say that a reporting correction has occurred. If, however, the age stratification were defined in 10-year groups (i.e., 0–9, 10–19, ...), then the subpopulation would not change, and the x_{ij} and y_{ij} would not be affected. We would not consider this to be a reporting correction.

The net count for subpopulation i at delay time t_j is $n_{ij} = \sum_{k=1}^{j}(x_{ik} - y_{ik})$. Our primary objective is to predict the eventual net count, ideally after an infinite delay, based on the data collected up until the current reporting year. A secondary interest is in the reporting delays and corrections themselves, for data quality control purposes. We would like to know, for example, the average reporting delay, the percentage of diagnosed cancers reported within 2 years, the percentage of reporting errors in the data, and whether these measures have improved or worsened in recent years.

2.2 The Reporting Model

Although x_{ij}, y_{ij}, and n_{ij} are observed only for $j \le J_i$, we let the corresponding random variables X_{ij}, Y_{ij}, and N_{ij} be defined for all positive integers j. We assume that the X_{ij}'s are independent and that $X_{ij} \sim \text{Poisson}(\lambda_i p_{ij})$, where $0 \le p_{ij} \le 1$ and $\sum_{j=1}^{\infty} p_{ij} = 1$. The parameter λ_i is the expected number of cancers that will eventually be reported for subpopulation i, whereas p_{ij} is the probability that a reported cancer is reported at delay time t_j. The distribution defined by $p_{ij}, j = 1, 2, \ldots$, represents the delay distribution for reporting cancers in subpopulation i.

The random variable Y_{ij} depends on both X_{ik} and Y_{ik}, $k = 1, \ldots, j-1$, because a case must be added before it can be removed and cannot be removed more than once. For simplicity, we assume that

$$(Y_{ij}|X_{ik}, Y_{ik}, k = 1, \ldots, j-1) \sim (Y_{ij}|N_{i,j-1}) \\ \sim \text{binomial}(N_{i,j-1}, g_{ij}),$$

where $0 \le g_{ij} \le 1$ and g_{ij} is the conditional probability of a reported case being removed at delay time t_j, given that the case was reported and not removed before delay time t_j. The likelihood function for subpopulation i is

$$L_i(x_{i1}, y_{i1}, \ldots, x_{iJ_i}, y_{iJ_i}) \\ = \prod_{j=1}^{J_i} \binom{n_{i,j-1}}{y_{ij}} g_{ij}^{y_{ij}} (1 - g_{ij})^{n_{i,j-1} - y_{ij}} \left(\frac{e^{-\lambda_i p_{ij}} (\lambda_i p_{ij})^{x_{ij}}}{x_{ij}!} \right).$$

When the $g_{ij} \to 0$, the model reduces to Harris' (1990) reporting delay model.

Because $N \sim \text{Poisson}(\theta)$, $(Y|N) \sim \text{binomial}(N, h) \Rightarrow Y \sim \text{Poisson}(\theta h)$, we can show by induction that N_{ij} and Y_{ij} have marginal Poisson distributions, with means $\text{E}(N_{ij}) = \lambda_i \sum_{k=1}^{j} [p_{ik} \prod_{r=k+1}^{j} (1 - g_{ir})]$ and $\text{E}(Y_{ij}) = g_{ij} \text{E}(N_{i,j-1})$, where we define $\prod_{r=k+1}^{k} a_r = 1$.

This model is not identifiable unless we further restricted the p_{ij}. Cox and Medley (1989) restricted the p_{ij} by assuming parametric delay distributions in continuous time. For discrete delay times, a simple distribution is the geometric, where $p_{ij} = \rho(1-\rho)^{j-1}$. Parametric delay distributions allow prediction of the cumulative counts after an infinite delay. In practice, however, these parametric delay distributions are often unstable, and one often assumes that the probability of reporting is negligible after a certain number of years (see, e.g., Cox and Medley 1989). In the remainder of this article we assume that $\sum_{j=1}^{J} p_{ij} = 1$ (i.e., $p_{ij} = 0$ for $j > J$), where $J \equiv \max(J_i)$ is the maximum observed delay time. As noted by Brookmeyer and Damiano (1989), under such an assumption one should interpret $\{p_{i1}, p_{i2}, \ldots, p_{iJ}\}$ as a truncated delay distribution, conditional on the delays being less than or equal to t_J. Similarly, we assume that $g_{ij} = 0$ for $j > J$. We parameterize p_{ij} in terms of its hazard $h_{ij} \equiv p_{ij} / \sum_{k=j}^{J} p_{ij}$, using a truncated version of the complementary log-log model,

$$h_{ij} = \left(1 - \exp(-e^{\mathbf{z}_{Xij}\boldsymbol{\beta}_X})\right) \Big/ \left(1 - \exp\left(-\sum_{k=j}^{J} e^{\mathbf{z}_{Xik}\boldsymbol{\beta}_X}\right)\right), \\ j \le J,$$

where $\mathbf{z}_{X_{ij}}$ is a $(1 \times d_1)$ vector of covariates and $\boldsymbol{\beta}_X$ is a $(d_1 \times 1)$ vector of regression coefficients. We also model the conditional probability g_{ij} using a complementary log-log model, $g_{ij} = 1 - \exp(-e^{\mathbf{z}_{Yij}\boldsymbol{\beta}_Y})$, $j \le \text{J}$, where $\mathbf{z}_{Yij}$ is a $(1 \times d_2)$ vector of covariates and $\boldsymbol{\beta}_Y$ is a $(d_2 \times 1)$ vector of regression coefficients. The parameterizations of h_{ij} and g_{ij} are very general, allowing one to specify nonparametric delay distributions, (truncated) geometric distributions, discrete-time proportional hazards models, and generalizations of these, depending on one's choice of covariates.

We calculate maximum likelihood estimates (MLEs) of the parameters $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Y$, $\widehat{\boldsymbol{\beta}}_X$ and $\widehat{\boldsymbol{\beta}}_Y$, using a Newton–Raphson method. The MLEs of p_{ij} and g_{ij}, $\widehat{p}_{ij}$ and $\widehat{g}_{ij}$, $j \le J$, are obtained by replacing $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Y$ with $\widehat{\boldsymbol{\beta}}_X$ and $\widehat{\boldsymbol{\beta}}_Y$ in the equations for p_{ij} and g_{ij}. The MLE of the ith component of $\boldsymbol{\lambda} \equiv (\lambda_1, \ldots, \lambda_S)$ is $\widehat{\lambda}_i = \sum_{j=1}^{J_i} x_{ij} / \sum_{j=1}^{J_i} \widehat{p}_{ij}$. An SAS/IML macro (SAS Institute 2000) to perform these calculations is available on request.

2.3 Estimating Variances

If the postulated model is correct, then one can use $\mathcal{I}^{-1}$, the inverse of the observed information, to estimate the covariance of $(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\beta}}_X, \widehat{\boldsymbol{\beta}}_Y)$. But our models of SEER data indicate overdispersion; that is, the variances of X_{ij} and $(Y_{ij}|N_{i,j-1})$ exceed their nominal Poisson and binomial variances. In the next section we consider a random-effects model to account for the extra variability. For the present model, we follow the general approach of McCullagh and Nelder (1989), who adjust for overdispersion by multiplying $\mathcal{I}^{-1}$ by a scalar estimate of overdispersion.

The covariates $\mathbf{z}_X$ and $\mathbf{z}_Y$ chosen for the model implicitly partition the S subpopulations into G groups ($G \le S$), such that subpopulations in the same group have the same delay distribution and same diagnosis year, whereas subpopulations in different groups have different delay distributions or diagnosis years. It is apparent from the likelihood function that $\widehat{\boldsymbol{\beta}}_Y$ is independent of $\widehat{\boldsymbol{\lambda}}$ and $\widehat{\boldsymbol{\beta}}_X$ and that $\mathcal{I}^{-1}$ can be partitioned as a block-diagonal matrix with elements $\mathcal{I}_1^{-1}$ and $\mathcal{I}_2^{-1}$, where $\mathcal{I}_1$ and $\mathcal{I}_2$ are the information matrices for $(\boldsymbol{\lambda}, \boldsymbol{\beta}_X)$ and $\boldsymbol{\beta}_Y$. If we assume that $\mathrm{var}(X_{ij}) = \phi_X \mathrm{E}(X_{ij})$ and $\mathrm{var}(Y_{ij}|N_{i,j-1}) = \phi_{Y|N} N_{i,j-1} g_{ij}(1 - g_{ij})$, then overdispersion parameters ϕ_X and $\phi_{Y|N}$ can be estimated from generalized Pearson statistics,

$$\widehat{\phi}_X = \frac{1}{M - d_1} \sum_{g=1}^{G} \sum_{j=1}^{J_i} \frac{(\sum_{i \in S_g} (x_{ij} - \widehat{\lambda}_i \widehat{p}_{ij}))^2}{\sum_{i \in S_g} \widehat{\lambda}_i \widehat{p}_{ij}} \qquad \text{and}$$

$$\widehat{\phi}_{Y|N} = \frac{1}{M - d_2} \sum_{g=1}^{G} \sum_{j=2}^{J_i} \frac{(\sum_{i \in S_g} (y_{ij} - n_{i,j-1} \widehat{g}_{ij}))^2}{\sum_{i \in S_g} n_{i,j-1} \widehat{g}_{ij}(1 - \widehat{g}_{ij})},$$

where $M = \sum_{i=1}^{G} (J_i - 1)$, and S_g is the set of all subpopulation indices for subpopulations in group $g, g = 1, \ldots, G$. The covariances of $(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\beta}}_X)$ and $(\widehat{\boldsymbol{\beta}}_Y)$ can then be estimated as $\widehat{\phi}_X \mathcal{I}_1^{-1}$ and $\widehat{\phi}_{Y|N} \mathcal{I}_2^{-1}$.

For model selection with overdispersed data, Burnham and Anderson (2002, pp. 67–70) suggested a modified AIC statistic that uses a quasi-likelihood function that incorporates the overdispersion. In Section 3 we report on simulations designed to evaluate the performance of this model selection criterion.

2.4 Random Reporting-Year Effects

As mentioned in the previous section, Pearson statistics for models of SEER melanoma data indicate overdispersion, that is, extra-Poisson and extra-binomial variation. Lawless (1994) discussed overdispersion in AIDS registry data and proposed a random-effects model to address the problem. Lawless's model assumes that for each diagnosis year, the delay probabilities follow a Dirichlet distribution, so that counts are correlated within diagnosis year but not within reporting year. Table 1 suggests that the problem with the SEER melanoma data is that counts are correlated within reporting year. In particular, reporting years 1991, 1993, 1995, and 1999 appear to have unusually large positive or negative counts. Such effects, associated with a particular reporting year across diagnosis years, appear to be sporadic rather than systematic and would seem to be good candidates for modeling as random effects.

We define a model that adds random reporting-year effects to the linear predictors for h_{ij} and g_{ij} when $j \ge 2$: $h_{ij} = (1 - \exp(-e^{\mathbf{z}_{Xij}\boldsymbol{\beta}_X + \gamma_{1,i+j-2}}))/(1 - \exp(-\sum_{k=j}^{J} e^{\mathbf{z}_{Xik}\boldsymbol{\beta}_X + \gamma_{1,i+k-2}}))$, $g_{ij} = 1 - \exp(-e^{\mathbf{z}_{Yij}\boldsymbol{\beta}_Y + \gamma_{2,i+j-2}})$, where $\gamma_{1,k} \sim \mathrm{N}(0, \sigma_1^2)$, $\gamma_{2,k} \sim \mathrm{N}(0, \sigma_2^2)$, $k = 1, \ldots, J - 1$, and all $\gamma_{m,k}$'s are independent. We use the relatively easy-to-implement Laplace approximation to numerically integrate the resulting likelihood over the unobserved random effects, although other methods are available (see Pinheiro and Bates 1995). The Laplace approximation of the log-likelihood of the random reporting-year effects model is

$$\begin{aligned}
&\widehat{\ell}_R(\lambda, \boldsymbol{\beta}_X, \boldsymbol{\beta}_Y, \sigma_1^2, \sigma_2^2; X, Y) \\
&\quad = \ell_C(\lambda, \boldsymbol{\beta}_X, \boldsymbol{\beta}_Y; \widehat{\boldsymbol{\gamma}}_1, \widehat{\boldsymbol{\gamma}}_2, X, Y) \\
&\qquad - \sum_{m=1}^{2} \left(\frac{1}{2\sigma_m^2} \sum_{k=1}^{J-1} \widehat{\gamma}_{m,k}^2 + \frac{J-1}{2} \log(\sigma_m^2) \right. \\
&\qquad + \frac{1}{2} \log \left| \frac{1}{\sigma_m^2} \mathbf{I} \right. \\
&\qquad \left. \left. - \left(\left. \frac{\partial^2 \ell_C(\lambda, \boldsymbol{\beta}_X, \boldsymbol{\beta}_Y; \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, X, Y)}{\partial \boldsymbol{\gamma}_m \, \partial \boldsymbol{\gamma}_m^T} \right|_{\boldsymbol{\gamma}_1 = \widehat{\boldsymbol{\gamma}}_1, \boldsymbol{\gamma}_2 = \widehat{\boldsymbol{\gamma}}_2} \right) \right| \right),
\end{aligned}$$

where $\boldsymbol{\gamma}_m = (\gamma_{m,1}, \gamma_{m,2}, \ldots, \gamma_{m,J-1})^T$; ℓ_C is the log-conditional likelihood given $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$; $\widehat{\boldsymbol{\gamma}}_1$ and $\widehat{\boldsymbol{\gamma}}_2$ maximize the penalized log-likelihood $\ell_C(\lambda, \boldsymbol{\beta}_X, \boldsymbol{\beta}_Y; \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, X, Y) - \sum_{m=1}^{2} (\frac{1}{2\sigma_m^2} \times \sum_{k=1}^{J-1} \gamma_{m,k}^2)$ for given λ, $\boldsymbol{\beta}_X$, $\boldsymbol{\beta}_Y$, σ_1, and σ_2; $|\mathbf{A}|$ is the determinant of $\mathbf{A}$, and $\mathbf{I}$ is an identity matrix. We maximize $\widehat{\ell}_R$ by a quasi-Newton method.

Let $p_{ij}(\boldsymbol{\gamma}_1)$ and $g_{ij}(\boldsymbol{\gamma}_2)$ denote random variables p_{ij} and g_{ij} as functions of $\boldsymbol{\gamma}_1$ or $\boldsymbol{\gamma}_2$. Using a Taylor series approximation, $p_{ij}(\boldsymbol{\gamma}_1) \approx p_{ij}(0) + \boldsymbol{\gamma}_1^T \mathbf{p}'_{ij}(0) + \frac{1}{2} \boldsymbol{\gamma}_1^T \mathbf{p}''_{ij}(0) \boldsymbol{\gamma}_1$, where $\mathbf{p}'_{ij}(0) = (\frac{\partial p_{ij}(\boldsymbol{\gamma}_1)}{\partial \boldsymbol{\gamma}_1}|_{\boldsymbol{\gamma}_1 = \mathbf{0}})$ and $\mathbf{p}''_{ij}(0) = (\frac{\partial^2 p_{ij}(\boldsymbol{\gamma}_1)}{\partial \boldsymbol{\gamma}_1 \, \partial \boldsymbol{\gamma}_1^T}|_{\boldsymbol{\gamma}_1 = \mathbf{0}})$, we approximate the mean and variance of $p_{ij}(\boldsymbol{\gamma}_1)$ as $\mathrm{E}(p_{ij}(\boldsymbol{\gamma}_1)) \approx p_{ij}(0) + \frac{\sigma_1^2}{2} \sum_{k=1}^{J-1} p''_{ij,kk}(0)$ and $\mathrm{var}(p_{ij}(\boldsymbol{\gamma}_1)) \approx \sigma_1^2 \sum_{k=1}^{J-1} (p'_{ij,k}(0))^2 + \frac{\sigma_1^4}{2} \sum_{k=1}^{J-1} \sum_{m=1}^{J-1} (p''_{ij,km}(0))^2$, and obtain estimates by replacing model parameters with MLEs. We estimate the mean and variance of $g_{ij}(\boldsymbol{\gamma}_2)$ similarly. In practice, we found that adding random reporting-year effects reduced, but did not eliminate, overdispersion. For this reason, and to make the random reporting-year model comparable to the nonrandom model, we adjust variance estimates for overdispersion and calculate a modified AIC statistic, as described in Section 2.3.

2.5 Predicting Net Counts

Under the reporting model, the predictor of N_{ij} that minimizes the mean squared prediction error, given the observed data up to delay time t_k, $k < j$, is the conditional expectation, $\mathrm{E}(N_{ij}|N_{ik}) = N_{ik} \prod_{s=k+1}^{j} (1 - g_{is}) + \lambda_i \sum_{r=k+1}^{j} p_{ir} \prod_{s=r+1}^{j} (1 - g_{is}) = N_{ik} a_{jk}(g_i) + \mathrm{E}(M_{ik}) b_{jk}(p_i, g_i)$, where $M_{ik} = \sum_{r=1}^{k} X_{ir}$, $a_{jk}(g_i) = \prod_{s=k+1}^{j} (1 - g_{is})$, and $b_{jk}(p_i, g_i) = \sum_{r=k+1}^{j} p_{ir} a_{jr}(g_i) / \sum_{r=1}^{k} p_{ir}$. We obtain predictor $\widehat{N}_{ij}(k)$ of N_{ij} by replacing $\mathrm{E}(M_{ik})$ with M_{ik} in the expression for $\mathrm{E}(N_{ij}|N_{ik})$, and replacing p_i and g_i with their MLEs, $\widehat{N}_{ij}(k) = N_{ik} a_{jk}(\widehat{g}_i) + M_{ik} b_{jk}(\widehat{p}_i, \widehat{g}_i)$. In particular, $\widehat{N}_{iJ}(J_i)$ is the predicted eventual count given the data up to the current reporting year.

The variances and covariance of $a_{jk}(\widehat{g}_i)$ and $b_{jk}(\widehat{p}_i, \widehat{g}_i)$ can be estimated using the delta method. Under the model assumptions, $\mathrm{var}(M_{ik}) = \mathrm{E}(M_{ik})$, $\mathrm{var}(N_{ik}) = \mathrm{E}(N_{ik})$, and $\mathrm{cov}(M_{ik}, N_{ik}) = \mathrm{E}(N_{ik})$. From these variances and covariances, one can estimate the variance of $\widehat{N}_{ij}(k)$ under the assumption that

(M_{ik}, N_{ik}) is independent of $(\widehat{p}_i, \widehat{g}_i)$, which is true asymptotically. Prediction for the model with random reporting-year effects is similar, except that we replace p_{ij} and g_{ij} with $\mathrm{E}(p_{ij})$ and $\mathrm{E}(g_{ij})$ in the expression for $\mathrm{E}(N_{ij}|N_{ik})$.

3. SIMULATIONS

We performed five sets of simulations with 1,000 datasets each. Each set of simulations is based on a different true model; the simulated datasets each have 17 subpopulations corresponding to 17 diagnosis years as in our melanoma example. The five models, defined in terms of $\mathbf{z}_{Xij}$ and $\mathbf{z}_{Yij}$, are as follows:

(a) Nonparametric tails: $\mathbf{z}_{Xij} = \{I(j = 1), I(j = 2), \ldots, I(j = 16)\}$, $\mathbf{z}_{Yij} = \{I(j = 2), I(j = 3), \ldots, I(j = 17)\}$

(b) Constant tails: $\mathbf{z}_{Xij} = \{I(j = 1), I(j = 2), I(j = 3), I(j = 4), I(j \geq 5)\}$, $\mathbf{z}_{Yij} = \{I(j = 2), I(j = 3), I(j = 4), I(j \geq 5)\}$

(c) Trend in diagnosis year: $\mathbf{z}_{Xij} = \{I(j = 1), I(j = 2), I(j = 3), I(j = 4), I(j \geq 5), (i - 1)\}$, $\mathbf{z}_{Yij} = \{I(j = 2), I(j = 3), I(j = 4), I(j \geq 5)\}$

(d) Random effects: $\mathbf{z}_{Xij}$ and $\mathbf{z}_{Yij}$ as in (b), but fitting the random reporting-year effects model

(e) Random effects, trend in diagnosis year: $\mathbf{z}_{Xij}$ and $\mathbf{z}_{Yij}$ as in (c), but fitting the random reporting-year effects model.

Here $I(a) = 1$ if a is true and 0 if otherwise. The corresponding parameters for the true models, chosen to be similar to those estimated from the melanoma data, are as follows:

(a) $\boldsymbol{\beta}_X = \{.4, -1, -1.5, -2, -2.1, -2.2, \ldots, -3.3\}^T$, $\boldsymbol{\beta}_Y = \{-4.5, -5, -5.5, -5.6, -5.7, \ldots, -6.8\}^T$

(b) $\boldsymbol{\beta}_X = \{.4, -1, -1.5, -2, -3\}^T$, $\boldsymbol{\beta}_Y = \{-4.5, -5, -5.5, -6\}^T$

(c) $\boldsymbol{\beta}_X = \{.32, -1, -1.5, -2, -3, .01\}^T$, $\boldsymbol{\beta}_Y = \{-4.5, -5, -5.5, -6\}^T$

(d) $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Y$ as in (b), $\sigma_1 = \sigma_2 = .5$

(e) $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Y$ as in (c), $\sigma_1 = \sigma_2 = .5$.

In all simulations, we set $\lambda_i \equiv 2{,}000$.

For each simulation, we fit all five models and selected the model with the smallest AIC value, as described in Section 2.3. This selection criteria performed better than minimizing a Pearson statistic or using the first 14 reporting years as a training sample to predict the counts in the last 3 reporting years.

Table 2 shows that when the true model is fit, $\widehat{N}_{17,17}(1)$ appears unbiased, where $\widehat{N}_{17,17}(1)$ is the predicted eventual count for diagnosis year 17, given the count after a t_1-year delay. The estimator of the variance of $\widehat{N}_{17,17}(1)$ is slightly conservative compared with the simulated variance. Simulations 1, 3, and 5 show that when an incorrect model is fit, the predicted count can be biased. When there are no reporting-year effects, the random-effects models perform nearly as well as their nonrandom counterparts. On the other hand, when there are reporting-year effects, the nonrandom models underestimate the variance of the predicted count (even after adjusting for overdispersion) and may have much larger mean squared errors (MSEs) than the random-effects models; see model (c).

The AIC selection procedure performed well in the simulations; the MSE for the selected model was moderately larger than that for the true model but was much smaller than that for some of the incorrect models in simulations 3–5. The estimated variance for the selected model was only slightly smaller than the simulated variance, due to the conservative nature of the estimated variance when there is no model selection.

4. EXAMPLE ANALYSIS OF MELANOMA INCIDENCE

We analyze the SEER invasive melanoma incidence data for whites for diagnosis years 1981–1997 and reporting years 1983–1999, summarized in Table 1. We begin with three simple models. Model 1 is the nonparametric tails model, and model 2 is the constant tails model, defined in Section 3. We considered various values for the time after which the tails are constant but present results only for constant tails for $j \geq 5$. In model 3 we modify model 2 by adding the linear term $(j - 5) \times I(j \geq 5)$ to $\mathbf{z}_{Yij}$. Model 3 allows the probability of a correction to decrease over time.

In Table 3 we see no great difference between models 1–3 with respect to AIC or the predicted count; although model 1 has a slightly smaller AIC value than models 2 and 3, the Pearson statistics $\widehat{\phi}_X$ and $\widehat{\phi}_{Y|N}$ indicate a huge lack of fit for all three models. The estimated regression coefficient for the linear tail in model 3 was $-.112$ [standard error (SE) $= .073$], not statistically significantly different from 0.

Models 4–6 address the question of whether the delay distributions are stationary by adding a trend in diagnosis year to models 1–3. We add the covariate year$(i) - 1981$ to $\mathbf{z}_{Xij}$ in models 1–3, where year(i) is the year of diagnosis for reported cancers in subpopulation i. We attempted to add the trend covariate to $\mathbf{z}_{Yij}$, but this produced unstable results. Table 3 indicates a marginal improvement in fit over models 1–3, and a moderate (not statistically significant) decrease in the predicted count (about 9% greater than the observed count of 3,860, compared with about 18% greater for models 1–3). The estimated coefficient for the trend was .021 (SE $= .012$) in model 4 and was similar in models 5–6, again not statistically significantly different from 0.

In models 7–12 we add random reporting-year effects to models 1–6. Table 3 shows that the random-effects models fit much better than their nonrandom counterparts, with dramatically smaller AIC and Pearson statistics, although they still exhibit some lack of fit, because a correct model would have Pearson statistics close to 1. The estimated standard deviations of the random effects in model 9, the best model according to AIC, were $\widehat{\sigma}_1 = .52$ (SE $= .17$) and $\widehat{\sigma}_2 = .86$ (SE $= .23$), both significantly different from zero. Except for model 7, the estimated SEs of the predicted counts are substantially larger for the random-effects models than for the corresponding nonrandom models; from the simulations, we know that when the random-effects model is true, the nonrandom models underestimate SEs. A troubling point is that the estimated regression coefficients for trend in models 4–6 have the opposite sign of those in models 10–12, although none is statistically significantly different from 0; the coefficient for trend is .021 (SE $= .012$) in model 4 and $-.013$ (SE $= .009$) in model 10. We note the possibility of confounding between delay time, reporting year, and diagnosis year effects in nonstationary models, and that such confounding could have a significant effect on the prediction of eventual counts.

Although the number of melanoma cases with unknown race is relatively small, these cases account for a large proportion of

Table 2. Simulation Results

Simulation number	Fitted model (true model in bold)	Simulated mean of AIC	% times model selected	True $E(N_{17})$	Simulated mean of $\widehat{N}_{17,17}(1)$	Simulated standard deviation of $\widehat{N}_{17,17}(1)$	Root mean estimated variance of $\widehat{N}_{17,17}(1)$	Simulated root mean squared error
1	**(a) Nonparametric tail**	**348.3**	**84.6**	**1,910**	**1,909**	**48.0**	**48.0**	**48.0**
	(b) Constant tail	367.2	4.6		1,897	47.6	47.7	49.3
	(c) Trend in diagnosis year	368.2	.8		1,896	49.9	50.6	51.8
	(d) Random effects	365.0	8.8		1,897	47.6	47.8	49.3
	(e) Random effects, trend diagnosis year	366.3	1.2		1,895	49.9	50.9	52.1
	Selected model				1,907	48.4	48.0	48.5
2	(a) Nonparametric tail	347.4	.3		1,903	46.6	47.5	46.6
	(b) Constant tail	**324.5**	**54.3**	**1,902**	**1,903**	**46.7**	**47.4**	**46.7**
	(c) Trend in diagnosis year	325.4	12.1		1,902	49.0	49.9	49.0
	(d) Random effects	325.0	30.7		1,903	46.7	47.6	46.7
	(e) Random effects, trend diagnosis year	326.2	2.6		1902	49.0	50.5	49.0
	Selected model				1,902	48.0	47.9	48.0
3	(a) Nonparametric tail	359.3	0		1,955	47.8	48.6	72.1
	(b) Constant tail	336.5	1.3		1,955	47.6	48.5	72.0
	(c) Trend in diagnosis year	**324.9**	**65.9**	**1,901**	**1,901**	**47.7**	**49.3**	**47.7**
	(d) Random effects	330.5	6.7		1,959	48.1	49.0	75.3
	(e) Random effects, trend diagnosis year	325.6	26.1		1,901	47.7	49.8	47.7
	Selected model				1,905	50.0	49.4	50.2
4	(a) Nonparametric tail	1,499.4	0		1,891	65.3	51.3	65.3
	(b) Constant tail	1,487.7	0		1,891	64.5	50.6	64.5
	(c) Trend in diagnosis year	1,462.8	0		1,895	112.1	65.6	112.2
	(d) Random effects	**334.2**	**86.8**	**1,890**	**1,890**	**58.0**	**60.6**	**58.0**
	(e) Random effects, trend diagnosis year	335.3	13.2		1,888	64.9	66.1	64.9
	Selected model				1,889	61.0	61.3	61.0
5	(a) Nonparametric tail	1,576.4	0		1,949	65.0	53.1	88.5
	(b) Constant tail	1,566.8	0		1,949	64.1	52.2	87.8
	(c) Trend in diagnosis year	1,528.4	0		1,898	110.5	65.2	110.9
	(d) Random effects	341.3	20.5		1,958	57.0	68.0	89.5
	(e) Random effects, trend diagnosis year	**336.4**	**79.5**	**1,889**	**1,889**	**60.5**	**62.3**	**60.5**
	Selected model				1,897	64.1	62.0	64.6

Note: $\widehat{N}_{17,17}(1)$ is the predicted eventual count for the most recent diagnosis year, given the observed count after a t_1-year delay; the selected model is the model with the smallest AIC.

the reporting corrections in our data; many cases moved from unknown race to white and from white to unknown. Discussions with SEER personnel indicated that this problem was due to changing policies in some of the registries on how to handle melanoma cases with unknown race. Because 99% of diagnosed melanoma cases with a known race are coded as white, some registries were undecided as to whether or not to assume that unknown cases were white. To avoid this problem, we decided to combine white and unknown into a single race category, so that we did not have to consider such changes to be reporting corrections. This resulted in a 35% reduction in the number of reporting corrections. We then refit the 12 models to the white/unknown data. All types of models gave similar results; we present results for the linear models, labeled 3(b)–12(b) in Table 3. The table indicates better fit for the nonrandom models compared with the whites only data, with Pearson statistics <8, and smaller reporting-year effects in the random-effects models, with estimated standard deviations $\widehat{\sigma}_1 = .33$ (SE $= .10$) and $\widehat{\sigma}_2 = .37$ (SE $= .10$) in model 12(b). The predicted counts are also estimated with more precision, and the discrepancy that we saw between models 6 and 12 disappears in models 6(b) and 12(b). Estimated regression coefficients in model 12(b), the best model according to the AIC, are .020 (SE $=$ .006) for trend in diagnosis year and $-.089$ (SE $=$ .021) for linear tail in Y, both statistically significant.

Table 4 gives the estimated percentage of reported cases that are reported within 2, 4, and 10 years of diagnosis, as well as the estimated percentage of reported cases with reporting errors, according to some of the models. We see that the reporting statistics for the white/unknown race data [model 9(b)] are substantially better than those for the white data (model 9). In model 9 only 79% of reported cases are reported within 2 years, compared with 86% in model 9(b); similarly, 6.6% of reported cases in model 9 have reporting errors, compared with only 4.5% in model 9(b). Model 12(b) indicates that reporting delay is getting shorter over time; in 1981, only 84% of cases were reported within 2 years, compared with 92% in 1997.

Figure 1 presents observed and predicted age-adjusted melanoma incidence rates for whites in diagnosis years 1990–1997 (before 1990, the observed and predicted rates are virtually identical), based on the white/unknown data and estimating predicted rates from model 12(b). Age-adjusted rates, commonly used in surveillance to compare rates over time, are weighted means of 5-year age-specific rates, with weights proportional to the size of the age groups in a reference population (in this case the 1970 U.S. population). The figure shows that

Table 3. SEER Melanoma Incidence in Diagnosis Years 1981–1997; AIC and Generalized Pearson Statistics for Different Models; Predicted Eventual Count for Diagnosis Year 1997, Given Observed Count in Reporting Year 1999

Model	*Random reporting-year effects*	*Trend in diagnosis year*	*Shape of delay distribution*	*AIC*	*Pearson statistic for X*	*Pearson statistic for Y*	*Predicted count for diagnosis year 1997*
Models for race = white, observed melanoma count for diagnosis year 1997 = 3,860							
1	No	No	Nonparametric	2,910	30.5	33.3	4,572$^{*}_{(128)}$
2	No	No	Constant tail	2,995	28.8	34.1	4,521$_{(116)}$
3	No	No	Linear tail	2,929	28.8	32.7	4,566$_{(116)}$
4	No	Yes	Nonparametric	2,886	29.9	33.3	4,202$_{(194)}$
5	No	Yes	Constant tail	2,973	28.3	34.1	4,178$_{(191)}$
6	No	Yes	Linear tail	2,908	28.3	32.7	4,219$_{(191)}$
7	Yes	No	Nonparametric	345.9	2.77	1.74	4,593$_{(112)}$
8	Yes	No	Constant tail	348.3	3.19	2.33	4,527$_{(167)}$
9	Yes	No	Linear tail	337.4	3.19	2.27	4,565$_{(163)}$
10	Yes	Yes	Nonparametric	345.2	2.74	1.74	4,965$_{(289)}$
11	Yes	Yes	Constant tail	349.2	3.21	2.33	4,710$_{(277)}$
12	Yes	Yes	Linear tail	338.3	3.21	2.27	4,749$_{(275)}$
Models for race = white/unknown, observed melanoma count for diagnosis year 1997 = 4,106							
3(b)	No	No	Linear tail	566.2	7.38	2.82	4,511$_{(76)}$
6(b)	No	Yes	Linear tail	520.3	5.68	2.82	4,259$_{(78)}$
9(b)	Yes	No	Linear tail	330.5	2.79	1.74	4,580$_{(84)}$
12(b)	Yes	Yes	Linear tail	328.1	2.77	1.74	4,273$_{(102)}$
Models for race = white/unknown, stratified by reporting source (hospital/nonhospital)							
3(c)	No	No	Linear tail	1,273	10.0	4.26	4,476$_{(82)}$
6(c)	No	Yes	Linear tail	1,088	6.62	4.39	4,221$_{(92)}$
9(c)	Yes	No	Linear tail	684.9	3.03	2.59	4,658$_{(141)}$
12(c)	Yes	Yes	Linear tail	653.9	2.54	2.59	4,355$_{(170)}$

*Estimated standard errors are in parentheses.

the predicted rates rose constantly from 1993 to 1997, whereas the observed rates declined in 1997, although the observed rate for 1997 was within the 95% confidence interval for the predicted rate. Figure 1 also includes the initial observed rate based on cases reported within the first 2 years after diagnosis year. The initial rate, although biased, often gives a better indication of trend than the observed rate.

Finally, because a large number of melanomas are diagnosed outside of hospitals, and because we believe that these nonhospital reporting sources tend to have longer delay times than hospitals, we also fit some models that allow hospital and nonhospital reporting sources to have different delay distributions. This step was complicated by two factors. First, to fit such a model, it is necessary to divide the subpopulations by reporting source, thereby considering reporting source changes to be reporting corrections. This increased the number of reporting corrections by about 80%. Second, even though by 1995, 26% of melanomas were being diagnosed outside of hospitals, in 1981 the number was only 12% (243 cases), which may not be sufficient to obtain accurate estimates of the delay distributions. Nevertheless, we fit the white/unknown data to stratified versions of models 3(b)–12(b) that allow $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Y$ to differ by reporting source. The results, presented as models 3(c)–12(c) in Table 3, indicate that these stratified models do not fit as well or provide as precise prediction as the unstratified models. Table 4 suggests, however, that delay distributions do differ by reporting source; the estimated percentage of cases reported within 2 years is 88% for hospitals, compared with only 59% for nonhospital sources. Similarly, the estimated percentage of reporting corrections is much higher for nonhospital sources (25%). Such stratified models can help one understand the nature of reporting delay distributions for quality control purposes, even if they are not used in prediction.

5. DISCUSSION

We have jointly modeled reporting delays and reporting corrections. Correction of reporting errors would be common in any registry database; the record of those corrections, however, is not routinely made available to those modeling the reporting delay. Without such historical records, it appears as if the errors never occurred, which can lead to biased prediction if the number of cases that were erroneously added to a subpopulation at

Table 4. SEER Melanoma Incidence in Diagnosis Years 1981–1997: Percentage of Reported Cases Reported Within 2, 4, and 10 Years, and Percentage of Reported Cases Having Reporting Errors, According to Different Models

Model	*Diagnosis year*	*Reporting source*	*% reported within 2 years*	*% reported within 4 years*	*% reported within 10 years*	*% reporting errors*
9			79$^{*}_{(1.8)}$	86$_{(.7)}$	94$_{(.5)}$	6.6$_{(1.7)}$
9(b)			86$_{(.6)}$	92$_{(.5)}$	97$_{(.5)}$	4.5$_{(.6)}$
12(b)	1981		84$_{(.8)}$	91$_{(.7)}$	98$_{(.5)}$	4.5$_{(.5)}$
12(b)	1997		92$_{(1.5)}$	96$_{(1.2)}$	99$_{(.4)}$	4.7$_{(.6)}$
9(c)		Hospital	88$_{(.8)}$	93$_{(.5)}$	97$_{(.4)}$	3.6$_{(.6)}$
9(c)		Nonhospital	59$_{(3.7)}$	71$_{(1.6)}$	89$_{(1.4)}$	25.2$_{(4.4)}$

*Estimated standard errors are in parentheses.

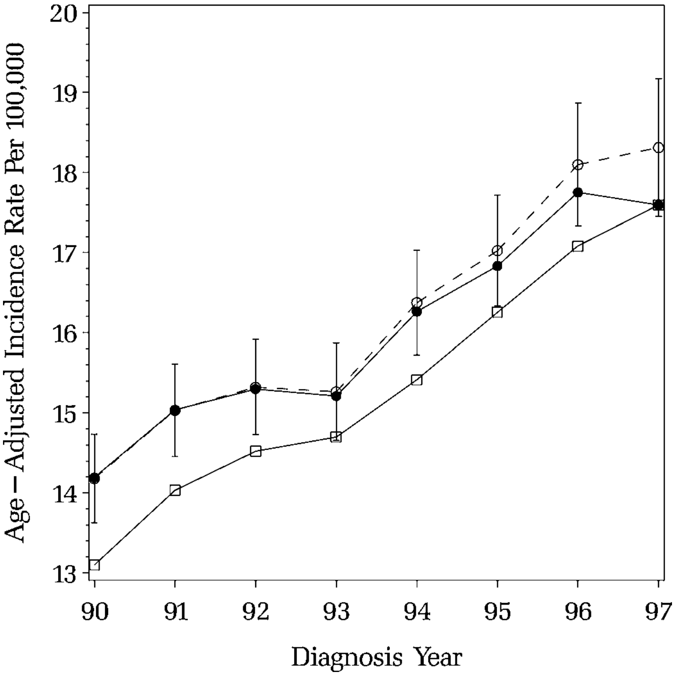

Figure 1. Observed and Predicted SEER Melanoma Age-Adjusted Incidence Rates in Diagnosis Years 1990–1997 for Race White/Unknown. Initial observed rates, based on cases reported within first 2 years after diagnosis year (□); observed rates in reporting year 1999 (●); predicted eventual rates and 95% confidence intervals (+/ − 1.96 × SE), based on observed cases in diagnosis years 1981–1997 and reporting years 1983–1999, and estimated under model 12(b) (○).

each delay time is not balanced by the number of cases that were erroneously excluded from that subpopulation. For ease of exposition, we began our discussion as if data in the form of Table 1 were readily available, but in fact some work was done with archived datasets to get the data in the form that we have described.

All of our reporting models assume that the reporting process is relatively stable. Abrupt, significant changes in the reporting process, such as changes in reporting personnel or reporting procedures, may cause difficulty for these models; a minor example of this type of change is the change from 19 to 22 months for the allowable delay in SEER. This is because the reporting models try to predict the future reports based on the past, but the recent past may be different from the distant past if these abrupt changes have occurred. In fact, as we begin to use and disseminate these reporting models, we will be faced with the statistical prediction problem that in response to this information, the registries may modify their practices to decrease their reporting delays and reporting errors. These modifications are a desired result from the standpoint of timely and accurate reporting of vital rates, but they will create modeling challenges for future applications of our reporting models.

The modified AIC model selection criterion performed well in simulations, but we noted that some care must be used when fitting nonstationary models, because there is a risk of confounding diagnosis year, reporting year, and delay time effects, which could have a significant effect on prediction of eventual counts and might not be detected by the AIC.

Our analysis of the melanoma data shows that reporting-year effects sometimes may be necessary. In some situations, large reporting-year effects may be avoided by combining subpopulations so that some reporting corrections are eliminated (e.g., combining white and unknown race subpopulations so that changes from white to unknown no longer count as corrections). We introduced a random reporting-year effects version of the reporting model that fit our data much better than did the nonrandom model, but still exhibited some lack of fit. For simplicity, we assumed that the random effects were normally distributed and independent; alternatively, one could allow the random effects to be correlated or use nonparametric methods to estimate the distribution of effects, as was done by Laird (1978). Lawless' (1994) model allows counts to be correlated within diagnosis year, whereas our model allows correlation within reporting year. In reality, both types of correlation are likely to occur, and a model that could incorporate both is desirable.

In this article we have extended reporting delay models to include reporting corrections. This important addition allows the use of this class of models to properly adjust trends from population-based cancer registries, which is important for the evaluation of national cancer control efforts.

[Received June 2002. Revised December 2003.]

REFERENCES

Brookmeyer, R., and Damiano, A. (1989), "Statistical Methods for Short-Term Projections of AIDS Incidence," *Statistics in Medicine*, 8, 23–34.

Burnham, K. P., and Anderson, D. R. (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.), New York: Springer-Verlag.

Clegg, L. X., Feuer, E. J., Midthune, D. N., Fay, M. P., and Hankey, B. F. (2002), "Impact of Reporting Delay and Reporting Error on Cancer Incidence Rates and Trends," *Journal of the National Cancer Institute*, 94, 1537–1545.

Cox, D. R., and Medley, G. F. (1989), "A Process of Events With Notification Delay and the Forecasting of AIDS," *Philosophical Transactions of the Royal Society of London*, 325, 135–145.

Doray, L. G. (1996), "UMVUE of the IBNR Reserve in a Lognormal Linear Regression Model," *Insurance: Mathematics and Economics*, 18, 43–57.

Frey, C. M., McMillen, M. M., Cowan, C. D., Horm, J. W., and Kessler, L. (1992), "Representativeness of the Surveillance, Epidemiology, and End Results Program Data: Recent Trends in Cancer Mortality Rates," *Journal of the National Cancer Institute*, 84, 872–877.

Fritz, A. (2001), "The SEER Program's Commitment to Data Quality," *Journal of Registry Management*, 28, 35–40.

Harris, J. E. (1990), "Reporting Delays and the Incidence of AIDS," *Journal of the American Statistical Association*, 85, 915–924.

Herbst, T. (1999), "An Application of Randomly Truncated Data Models in Reserving IBNR Claims," *Insurance: Mathematics and Economics*, 25, 123–131.

Horm, J. W., and Kessler, L. G. (1986), "Falling Rates of Lung Cancer in Men in the United States," *Lancet*, 8478, 425–426.

Jemal, A., Clegg, L. X., Ward, E., Ries, L. A. G., Wu, X., Jamison, P. M., Wingo, P. A., Howe, H. L., Anderson, R. N., and Edwards, B. K. (2004), "Annual Report to the Nation on the Status of Cancer, 1975–2001, With a Special Feature on Survival," *Cancer*, 101, 3–27.

Kalbfleisch, J. D., and Lawless, J. F. (1989), "Inference Based on Retrospective Ascertainment: An Analysis of the Data on Transfusion-Related AIDS," *Journal of the American Statistical Association*, 84, 360–372.

——— (1991), "Regression Models for Right Truncated Data With Applications to AIDS Incubation Times and Reporting Lags," *Statistica Sinica*, 1, 19–32.

Kalbfleisch, J. D., Lawless, J. F., and Robinson, J. A. (1991), "Methods for the Analysis and Prediction of Warranty Claims," *Technometrics*, 33, 273–285.

Keiding, N., and Gill, R. D. (1990), "Random Truncation Models and Markov Processes," *The Annals of Statistics*, 18, 582–602.

Laird, N. (1978), "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution," *Journal of the American Statistical Association*, 73, 805–811.

Lawless, J. F. (1994), "Adjustments for Reporting Delays and the Prediction of Occurred but not Reported Events," *Canadian Journal of Statistics*, 22, 15–31.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall.

Pagano, M., Tu, X. M., De Gruttola, V., and MaWhinney, S. (1994), "Regression Analysis of Censored and Truncated Data: Estimating Reporting-Delay Distributions and AIDS Incidence From Surveillance Data," *Biometrics*, 50, 1203–1214.

Pinheiro, J. C., and Bates, D. M. (1995), "Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model," *Journal of Computational and Graphical Statistics*, 4, 12–35.

Ries, L. A. G., Eisner, M. P., Kosary, C. L., Hankey, B. F., Miller, B. A, Clegg, L., Mariotto, A., Feuer, E. J., and Edwards, B. K. (eds.) (2004), *SEER Cancer Statistics Review, 1975–2001*, Bethesda, MD: National Cancer Institute, available at *http://seer.cancer.gov/csr/1975_2001*.

SAS Institute Inc. (2000), *SAS/IML User's Guide, Version 8*, Cary, NC: Author.

Sellero, C. S., Fernandez, E. V., Manteiga, W. G., Otero, X. L., Hervada, X., Fernandez, E., and Taboada, X. A. (1996), "Reporting Delay: A Review With a Simulation Study and Application to Spanish AIDS Data," *Statistics in Medicine*, 15, 305–321.

Verrall, R. J. (2000), "An Investigation Into Stochastic Claims Reserving Models and the Chain-Ladder Technique," *Insurance: Mathematics and Economics*, 26, 91–99.