UW Biostatistics Working Paper Series

5-26-2006

A Marginalized Diffusion Model for Estimating Age at First Endoscopy Examination from Current Status Data

Diana Miglioretti
Group Health Cooperative, miglioretti.d@ghc.org

Elizabeth Brown
University of Washington, elizab@u.washington.edu

1 Introduction

Adoption of cancer screening by at risk populations affects cancer incidence and mortality and the demand for health services including screening and diagnostic tests as well as treatment for detected disease. Understanding trends in screening use can inform our interpretation of observed changes in cancer incidence and mortality rates and may aid in the projection of future health care utilization needs. Longitudinal estimates of the proportion of people who have been screened for cancer are available from large, publicly available databases including surveys and public health records. We can use these estimates to evaluate whether program goals have been met and to predict whether future goals such as those set by the Healthy People 2010 (http://www.healthypeople.gov/) are obtainable. In addition, models describing screening dissemination may be used by microsimulation modelers such as the Cancer Intervention and Surveillance Modeling Network (CISNET: http://cisnet.cancer.gov/) to estimate the contribution of screening to observed changes in cancer incidence and mortality (e.g., Berry, 2005).

In this paper, we develop a model for the dissemination of endoscopy examinations in the United States from 1975 to 2003. Our long term goal is estimation of the contribution of endoscopy to observed changes in national colorectal cancer incidence and mortality rates. To meet this goal, our endoscopy dissemination model will be applied to microsimulation models developed by three CISNET groups: Group Health Cooperative (Rutter, Miglioretti, Savarino, in prep), Sloan-Kettering Institute for Cancer Research/MISCAN (Loeve, *et al.*, 1999), and Harvard School of Public Health (Knudsen, 2005; Frazier, *et al.*, 2000) in future research. Descriptions and comparisons of these microsimulation models can be found at http://cisnet.cancer.gov/profiles/.

1.1 Available Data Sources

To assess screening behaviors in the United States, Epidemiologists rely heavily on two surveys: the National Health Interview Survey (NHIS) and the Behavioral Risk Factor Surveillance System (BRFSS). The NHIS is a household in-person interview conducted by the Center for Disease Control's National Center for Health Statistics. The NHIS uses a multistage area probability design which permits representative sampling of households to provide national estimates of health behaviors (http://www.cdc.gov/nchs/nhis.htm). The NHIS covers the civilian noninstitutionalized population of the United States and includes approximately 43,000 households and 106,000 persons per year. The annual response rate is greater than 90% of eligible households.

The BRFSS is a large (approximately 160,000 participants per year) telephone survey developed to monitor state-level health risks and behaviors (http://www.cdc.gov/brfss/). It is administered and supported by the Center for Disease Control's National Center for Chronic Disease Prevention and Health Promotion. The BRFSS is limited to families with land-line telephones and the response rate is lower than the NHIS, ranging across states from 46% to 93% (median 77%) in 1997. Most states use disproportionate stratified sampling to allow estimation at the region as well as the state level. All states and Washington, D.C. adopted this sampling approach by 2001.

The NHIS and BRFSS have complementary advantages that make it desirable to include both data sources for estimating the age at first endoscopy. The high response rate of the NHIS and the fact that it is an in-person interview likely make it less biased than the BRFSS. However, the BRFSS is a larger survey and asked about endoscopy examinations in more years.

1.2 Modeling Challenges

In developing our model for age at first endoscopy, we faced three main challenges: First, we can not easily obtain empirical estimates of the age at first endoscopy over the entire time-period of interest (1975-2003), because available survey data consists of questions of the form "Have you ever had [*the screening test under study*]." Thus, at each survey year, we observe the proportion of participants that have had an endoscopy prior to the time of the survey. This type of data is referred to as current status data (Jewell and van der Laan, 2002). Second, because national survey data was not available until 1987, there is a substantial time period during which we have no information about endoscopy use. A third challenge was including data from both the NHIS and BRFSS to increase the amount of information about the dissemination of endoscopy. While the NHIS is designed for estimation at the national or regional level, the BRFSS was designed to make inference at the state level and most states did not ask about endoscopy in all years; therefore, state-level differences must be taken into account to adjust for missing data as well as for clustering within states.

To overcome these challenges, we use a theoretically motivated parametric model from the diffusion of innovation literature (Mahajan and Peterson, 1985) to extrapolate from information observed in 1987 to 2003 to estimate the rate of adoption in the earlier years. Cronin *et al.* (2005) used a mixed-influence diffusion model to model time to first mammography examination separately by birth cohort based on NHIS data only. We extend their model by incorporating covariate effects to pool information across birth cohorts to better estimate the rate of endoscopy adoption over time. Further, to combine information from the NHIS and BRSS to make inference about the national-level diffusion model parameters, we link a state-specific model to the national-level diffusion model of interest using a marginalized modeling approach (Heagerty and Zeger, 2000; Heagerty, 2002; Miglioretti and Heagerty, 2004).

In the next section, we introduce our national-level and state-specific mixed-influence diffusion models. In section 3, we describe our methodology for estimating the national-level diffusion model parameters of interest. In section 4, we fit the model to the NHIS and BRFSS data to estimate the diffusion of endoscopy. We close with a discussion of our approach.

2 Marginalized diffusion model

We sought to combine information from the NHIS and the BRFSS at multiple survey years to obtain national-level estimates of the diffusion of endoscopy use over time by birth year and gender. *Diffusion of innovations* models are widely used in economics and social research to model the adoption of new products or behaviors by a population. We propose using a specific type of diffusion model, the *mixed influence diffusion model* (Mahajan and Peterson, 1985), to model the age at first endoscopy examination in the United States. This model accounts for influences from both external sources, such as mass media and physicians, and internal sources, such as discussions with friends and family. In addition, the model can be parameterized to allow a proportion of the population to never be screened. The regression model we use is based on the following differential equation that describes the rate of change in endoscopy use at time T as:

$$\frac{dp^N(t<T)}{dt} = \left(a + bp^N(t<T)\right)\left(c - p^N(t<T)\right) \tag{1}$$

where $p^N(t<T)$ is the proportion of individuals in the United States that have had an endoscopy examination before time T, $a > 0$ measures the *external influence*, $b > 0$ measures the *internal influence*, and $0 < c < 1$ is the proportion of people who will ever have an endoscopy. Thus, $c - p^N(t<T)$ is the proportion of potential adopters remaining at time T. Integration of equation (1) gives a model for the national cumulative incidence of endoscopy use over time:

$$p^N(t<T) = \frac{ac(1 - \exp\left[-(a+bc)(T-t_0)\right])}{a + bc\exp\left[-(a+bc)(T-t_0)\right]} \tag{2}$$

where t_0 is the year endoscopy was first introduced for general use. Figure 1 displays mixed-influence diffusion curves for a range of values of a, b, and c. Increasing the external influence parameter a has an immediate effect on increasing the diffusion curve, because external influences do not depend on how many people have had an endoscopy. In contrast, increasing the internal influence parameter b has a more delayed effect on the diffusion curve. As more people have an endoscopy, the diffusion rate increases at a faster rate for larger values of b. A delay in the rise of the diffusion curve could also occur if there is a delay in use of endoscopy until a certain age. For example, for younger birth cohorts, there may be a delay in the use of endoscopy until the cohort reaches the age at which guidelines recommend use of endoscopy for screening (typically age 50). This would result in an increase in b with increasing birth year. As the asymptote c increases, a larger proportion of people eventually have an endoscopy.

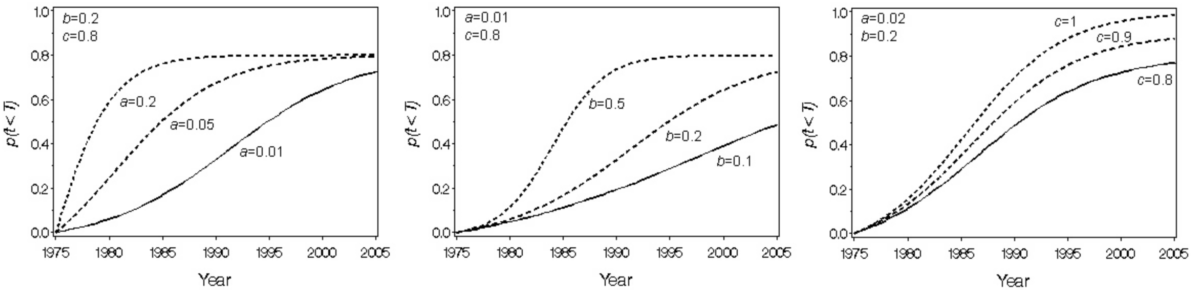

Figure 1: Mixed-influence diffusion curves, $p^N(t < T)$, for different values of a, b, and c.

We allow the cumulative incidence of endoscopy use to depend on covariates such as birth year by modeling each diffusion parameter as a function of covariates $\mathbf{x}$. Substituting $f_{\boldsymbol{\alpha}}(\mathbf{x})$, $f_{\beta}(\mathbf{x})$, and $f_{\boldsymbol{\eta}}(\mathbf{x})$ for a, b, and c, respectively, we use a log-link for the influence parameters $f_{\boldsymbol{\alpha}}(\mathbf{x})$ and

$f_{\boldsymbol{\beta}}(\mathbf{x})$ and a logit-link for the asymptote $f_{\boldsymbol{\eta}}(\mathbf{x})$ to maintain the parameter constraints:

$$\begin{aligned}
\log\left(f_{\boldsymbol{\alpha}}(\mathbf{x})\right) &= \alpha_0 + \alpha_1\mathbf{x} \\
\log\left(f_{\boldsymbol{\beta}}(\mathbf{x})\right) &= \beta_0 + \beta_1\mathbf{x} \\
\log\left(\frac{f_{\boldsymbol{\eta}}(\mathbf{x})}{1 - f_{\boldsymbol{\eta}}(\mathbf{x})}\right) &= \eta_0 + \eta_1\mathbf{x}.
\end{aligned}$$

Incorporating these functions into equation (2) gives a cumulative incidence model conditional on covariates:

$$p^N\left(t < T|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{x}_k, \mathbf{x}\right) = \frac{f_{\boldsymbol{\alpha}}(\mathbf{x})\, f_{\boldsymbol{\eta}}(\mathbf{x})\left(1 - \exp\left[-\left(f_{\boldsymbol{\alpha}}(\mathbf{x}) + f_{\boldsymbol{\beta}}(\mathbf{x})\, f_{\eta}(\mathbf{x})\right)(T - t_0)\right]\right)}{f_{\boldsymbol{\alpha}}(\mathbf{x}) + f_{\boldsymbol{\beta}}(\mathbf{x})\, f_{\eta}(\mathbf{x}) \exp\left[-\left(f_{\boldsymbol{\alpha}}(\mathbf{x}) + f_{\boldsymbol{\beta}}(\mathbf{x})\, f_{\eta}(\mathbf{x})\right)(T - t_0)\right]} \qquad (3)$$

For parameter estimation, equation (3) can be directly applied to data from the NHIS when accounting for survey weights, because the NHIS was designed to provide estimates that are representative at the national level. The NHIS sample size is too small to make inferences about states, and the public data set does not contain information about an individual's state of residence. In contrast, the BRFSS was designed to provide estimates at the state level; however, in some years, only some states asked about endoscopy use. Because states that did not ask about endoscopy use are likely to be different from states that did, missing data are probably not missing completely at random and thus estimates based only on available data will not be nationally representative. Therefore, state differences must be taken into account when estimating national-level parameters. To do this, we introduce a state-level model for the probability of having a previous endoscopy examination by survey year j for the ith state and the kth covariate combination $\mathbf{x}_k$ given state-specific effects θ_i :

$$\begin{aligned}
p^S\left(t < j|\Delta_{jk}\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \theta_i, \mathbf{x}_k\right), \theta_i\right) &= \operatorname{expit}\left(\Delta_{jk}\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \theta_i, \mathbf{x}_k\right) + \theta_i\right) \qquad (4) \\
\theta_i &\sim \operatorname{Normal}\left(0, \sigma^2\right)
\end{aligned}$$

where expit$(\cdot)$ is the inverse-logit function and $\Delta_{jk}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta},\theta_i,\mathbf{x}_k)$ is a tractable function of the diffusion model parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$,and $\boldsymbol{\eta}$, the state-specific effects θ_i, the survey year j and the covariates $\mathbf{x}_k$. In the following, we drop the notational dependence of Δ_{jk} on $(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta},\theta_i,\mathbf{x}_k)$ for simplicity, unless it is useful for clarification. Note that all 50 states plus Washington, D.C. are included in the BRFSS for a total of 51 state-specific effects.

The state-level model intercepts $\boldsymbol{\Delta}$ link the state-level model to the national-level model parameters and are fully determined by the relationship between the national and state-specific curves. The expected value of the state-specific probabilities $p^S(t<j|\Delta_{jk}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta},\theta_i,\mathbf{x}_k),\theta_i)$ taken with respect to the states (i) must equal the national-level probability for that survey year and covariate combination $p^N(t<j|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta},\mathbf{x}_k)$:

$$\begin{aligned} p^N(t<j|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta},\mathbf{x}_k) &= E_{(i)}\left[p^S(t<j|\Delta_{jk}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta},\theta_i,\mathbf{x}_k),\theta_i)\right] \qquad (5)\\ &= \frac{1}{\sum_{i=1}^{51} N_{ijk}}\sum_{i=1}^{51} p^S(t<j|\Delta_{jk}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta},\theta_i,\mathbf{x}_k),\theta_i)\,N_{ijk}. \end{aligned}$$

where N_{ijk} is the ith state's population size for the kth covariate combination during the jth survey year. We solve equation (5) for the state-level model intercepts $\boldsymbol{\Delta}$ using Newton-Raphson with N_{ijk} taken from the census (http://www.census.gov).

3 Model Estimation

For unbiased national and state-level parameter estimates, we need to account for the survey sampling designs. Because the design variables are not available in the NHIS or BRFSS public data sets, we must rely on the survey weights provided in the public data files. Let $h=1$ for the NHIS and $h=2$ for the BRFSS. Let n_{hijk} be the number of individuals surveyed with survey h from state i during year j with the kth covariate combination, $i=1,\ldots,51; j=1,\ldots,J_{hi}; k=1,\ldots,K$, and let

y_{hijk} be the number of surveyed individuals that said they have ever had an endoscopy examination. For notational purposes, we replace the state-level subscript i with $\cdot$ when $h = 1$, because state of residence is not available in the NHIS public use data sets. We use the survey weights w_{hijk} to create a *pseudo-sample* with $y^*_{hijk} = w_{hijk} y_{hijk}$ positive responses and $n_{hijk} - y^*_{hijk} = w_{hijk}\left(n_{hijk} - y_{hijk}\right)$ negative responses for each survey h, state i, survey year j, and covariate combination k (Cronin *et al.*, 2005). We take the weight w_{hijk} to be the sum of the standardized survey weights, summed over the number of individuals surveyed and standardized to sum to the total survey sample size $\sum_{k=1}^{K} n_{hijk}$. This *pseudo-sample* approach results in a weighted version of the likelihood that is equivalent to the pseudo-likelihood commonly used in the analysis of survey data (Korn and Graubard, 1999; Chambers and Skinner, 2003).

As discussed in the previous section, the NHIS provides nationally representative data; therefore, assuming the weighted observations are conditionally independent given the model, the likelihood component for the NHIS *pseudo-data* y_1^* may be written as a function of the national model (3):

$$p\left(\mathbf{y}_1^*|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{x}\right) \propto \prod_{j=1}^{J_{h\cdot}} \prod_{k=1}^{K} p^N\left(t < j|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{x}_k\right)^{y^*_{1\cdot jk}} \left(1 - p^N\left(t < j|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{x}_k\right)\right)^{n_{1\cdot jk} - y^*_{1\cdot jk}},$$

where $p^N\left(t < j|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{x}\right)$ is defined in equation (3). The BRFSS provides data that are representative at the state level, and as discussed in the previous section, differences between states must be taken into account due to missing data for some states in some survey years; therefore, we write the likelihood component for the BRFSS *pseudo-data* y_2^* in terms of the state-specific model (4). Assuming outcomes within a state are independent conditional on the state-specific effects $\boldsymbol{\theta}$, the likelihood for the BRFSS may be written as follows:

$$p\left(\mathbf{y}_2^*|\alpha, \beta, \eta, \mathbf{x}_k, \boldsymbol{\theta}\right) \propto \prod_{i=1}^{51} \prod_{j=1}^{J_i} \prod_{k=1}^{K} p^S\left(t < j|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{x}_k, \theta_i\right)^{y^*_{2ijk}} \left(1 - p^S\left(t < j|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{x}_k, \theta_i\right)\right)^{n_{2ijk} - y^*_{2ijk}},$$

where $p^S\left(t < j|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{x}_k, \theta_i\right)$ is defined in equation (4). We take the full likelihood to be the product of the likelihoods for the NHIS and BRFSS. While we recognize that observations within the same state may not be independent across the two surveys, we do not have information about an individual's state for the NHIS, so we can not take this dependence into account.

We use Markov Chain Monte Carlo (MCMC) to sample from the posterior distribution, which is proportional to the product of the prior distributions and the likelihood:

$$\begin{aligned} p\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\theta}, \sigma|\mathbf{y}^*, \mathbf{x}\right) \quad \propto & \quad p\left(\boldsymbol{\alpha}\right) p\left(\boldsymbol{\beta}\right) p\left(\boldsymbol{\eta}\right) p\left(\sigma\right) \prod_{i=1}^{51} p\left(\theta_i|\sigma\right) \\ & \quad p\left(\mathbf{y}_1^*|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{x}\right) p\left(\mathbf{y}_2^*|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{x}, \boldsymbol{\theta}\right) \end{aligned} \tag{6}$$

We use standard prior distributions, taking $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and η to be Normal$(0, 100)$ and σ to be Uniform$(0, 10)$. In the style of Gibbs sampling (Gelfand and Smith, 1990), we update each set of parameters conditional on the remaining parameters using Metropolis (random walk) steps (Metropolis, *et al.*, 1953; Gilks, Richardson, and Spiegelhalter, 1996).

4 Application

We used the marginalized diffusion model to estimate the age at first endoscopy examination from current status data collected by the NHIS and BRFSS. The NHIS asked about any prior endoscopy examination in 1987, 1992, 1998, 2000, and 2003. For the BRFSS, all states asked about prior endoscopy in 1997, 1999, 2001, and 2002, while only some states asked about endoscopy in 1989-1996, 1998, 2000, and 2003. The wording of the question varied by year and survey, but in general, asked "Have you ever had a [*exam-type*]?" where [*exam-type*] was one of the following: "proctoscopic exam," "sigmoidoscopy or colonoscopy," "sigmoidoscopy or proctoscopy," or "sigmoidoscopy, colonoscopy, or proctoscopy." In most years, women and men ages 40 and over were asked about prior endoscopy. In 2001, the BRFSS started only asking the question to people 50 and over. For analysis, if the person answered that they did not know or were unsure if they ever had an examination (<1% of observations), we assumed they never had the examination. We excluded people that refused to answer the question (<1%).

*** TABLE 1 ABOUT HERE ***

Tables 1 displays the sample sizes and weighted percentages of men and women that reported ever having an endoscopy. In 1987, the first year the NHIS asked about endoscopy, weighted estimates of the percentage of people that had a prior endoscopy from the NHIS ranged from 13% for women and 15% for men born 1940-1949 to 27% for women and 33% for men born 1910-1929. By 2003, this number increased to 40% for women and 43% for men born 1940-1949 and to 43% for women and 55% for men born 1910-1929.

We fit separate models for males and females. Because endoscopy was first used in the mid 1970's, we take t_0 to be 1975. To increase computational speed, we grouped people into 10 year birth cohorts for analysis, including the mean birth year for each group as the covariate value. Because the sample size is very small for the oldest birth cohorts, we grouped people born from 1910 to 1919 with those born in 1920. The diffusion curves are expected to be similar for these birth cohorts, because they would have all been of screening age (50 or older) at the time endoscopy was introduced.

For each model, we ran three samplers starting at dispersed values for 600,000 iterations each, discarded the first 100,000 iterations for burn-in, and kept every 20th iteration for analysis. Results are based on the 75,000 remaining iterations from the three combined samplers. To check convergence, we examined trace plots and compared the three samplers to verify convergence to the same posterior modes.

Table 2. Parameter estimates and 95% highest posterior density (HPD) intervals.

	Females		Males	
Parameter	Estimate	95% HPD	Estimate	95% HPD
α_0	-3.14	(-3.22, -3.06)	-2.58	(-2.68, -2.46)
β_0	-2.20	(-2.28, -2.12)	-2.20	(-2.38, -1.98)
η_0	0.24	(0.19, 0.29)	0.30	(0.25, 0.33)
α_1	-0.092	(-0.096, -0.088)	-0.087	(-0.092, -0.082)
β_1	0.0030	(0.0004, 0.0053)	-0.0084	(-0.013, -0.0019)
η_1	0.096	(0.089, 0.10)	0.064	(0.054, 0.073)
σ	0.15	(0.13, 0.20)	0.21	(0.17, 0.26)

Diffusion model parameter estimates (posterior modes) along with 95% highest posterior density intervals are shown in Table 2. Note that the HPD intervals are approximate, because we did not adjust for clustering within strata and probability sampling units (PSUs), and we used a *pseudo-sample* for estimation. None-the-less, these intervals provide a general idea of the magnitude of variability in these estimates, assuming there is not a large amount of correlation within unacknowledged clusters.

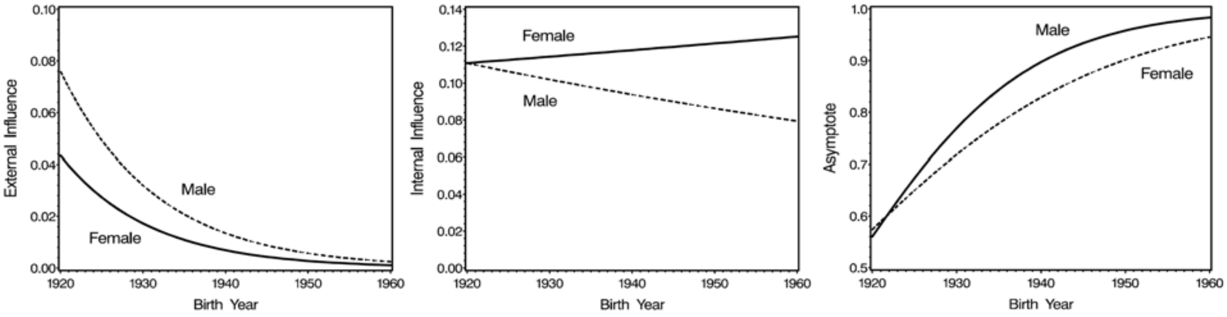

Figure 2: Diffusion model parameter estimates by birth year.

Figure 2 displays how the diffusion model parameter estimates change with birth year. For both males and females, the external influence parameter decreases rapidly with increasing year of birth. Males in the oldest cohort have a larger external influence parameter compared to women, but

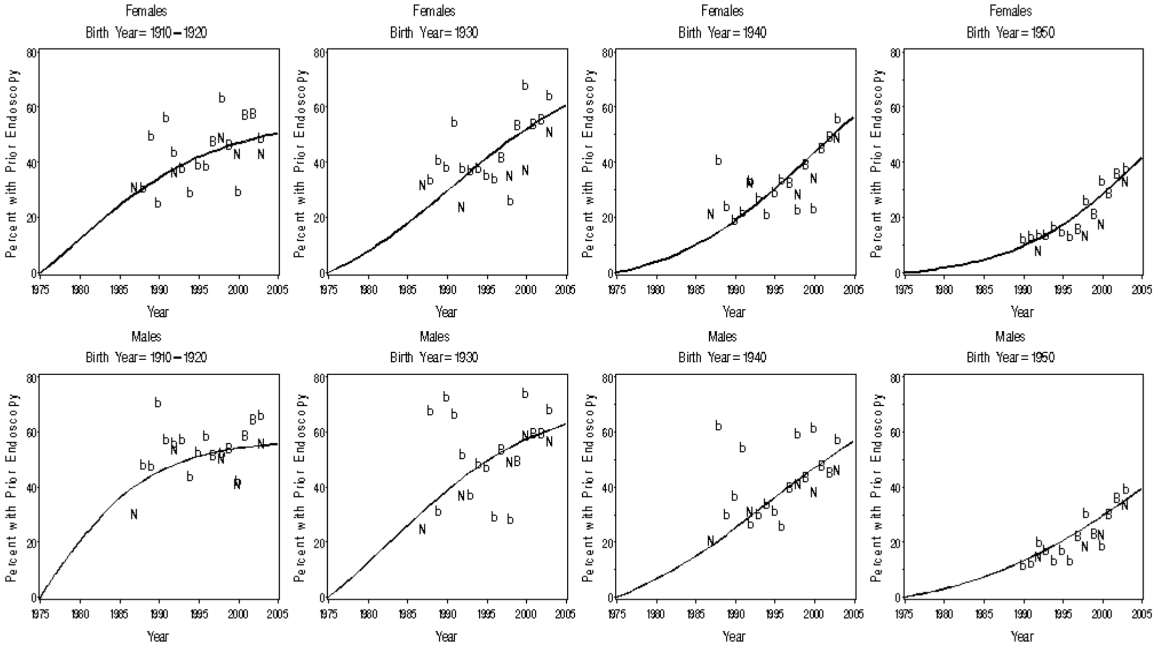

Figure 3: Estimated endoscopy diffusion curves for males and females born from 1910-1920 and in 1930, 1940, and 1950. N = weighted percentages from NHIS, B = weighted percentages from the BRFSS in years that all states were surveyed, and b = weighted percentages from the BRFSS in years when only some states were surveyed.

this difference diminishes with increasing year of birth. In contrast, the oldest males and females have a similar internal influence parameter, but with increasing birth year, the internal influence parameter increases for females and decrease for males, resulting in a larger internal influence parameter for younger females relative to males. Men and women in the oldest birth cohort have similar asymptotes, however, these asymptotes diverge as birth year increases, suggesting more males will eventually receive an endoscopy than females among the younger birth cohorts. However, care must be taken when interpreting the asymptote parameter, especially for the younger birth cohorts, because it requires extrapolation outside the observed range of data. Variability between states (Table 2) is somewhat larger for males than females.

Figure 3 shows the endoscopy diffusion curves for females and males born in 1920, 1930, 1940 and 1950. The fitted curves fit the observed data very well in years that all states were surveyed.

5 Discussion

In this paper, we present an approach for modeling the diffusion of a cancer test within a population based on current status data collected via national surveys. We incorporate covariate effects into a theoretically motivated mixed-influence diffusion model (Mahajan and Peterson, 1985) to pool information across multiple birth cohorts to estimate the rate of test adoption based on current status data observed during a limited time-period. We use marginalized modeling methodology to make inference about national-level parameters, combining data from two surveys: the NHIS, which is representative at the national or regional level and the BRFSS, which is representative at the state level. Our methodology differs from other marginalized random effects models (Heagerty and Zeger 2000; Diggle, Heagerty, Liang, and Zeger, 2002; Miglioretti and Heagerty, 2004) in that the conditional (state-level) and marginal (national-level) models are linked through a weighted average across a fixed number of state-specific models as opposed to an integral over the entire random effect distribution.

One advantage of the marginalized modeling approach is that the structure of the mean model is not constrained by the study design or the desired fitting method. By using this approach, we were able to directly model the marginal (national) mean of interest while separately specifying a state-specific model used in the likelihood for the BRFSS. Generalized estimating equations (GEE; Liang and Zeger, 1986) are commonly used when substantive interest is in the marginal or population-average regression structure. However, without modification, GEE may give biased results when data are not missing completely at random (Laird, 1988; Robins *et al.*, 1995). The marginalized modeling approach uses likelihood-based methods, which are robust when data are missing at random.

We estimated model parameters using data from two surveys: the NHIS and BRFSS. The NHIS is generally believed to be less biased than the BRFSS, because the BRFSS excludes households without land-line telephones and has a lower response rate than the NHIS. The rates of endoscopy use reported here are higher for the BRFSS compared to the NHIS, suggesting some bias in the BRFSS. However, we include the BRFSS for estimation, because the NHIS only asked about colorectal cancer testing in a limited number of years late in the diffusion process. There has been some research on adjusting for bias in the BRFSS when combining data from these two surveys for small area estimation. Elliott and Davis (2005) use a propensity score approach while Raghunathan *et al.* (2006) directly adjust for an effect of having a telephone. It may be possible to incorporate a *telephone effect* into our model, however, we are not interested in separate diffusion curves for people with and without telephones. Therefore, we would need to marginalize over the telephone effect for estimation of national-level parameters of interest, which would add an additional level of complexity to our model.

We acknowledge that the use of a pseudo-sample is *ad hoc.* Design variables are not available in the NHIS and BRFSS public use data sets, and accounting for complex survey designs using likelihood methods is challenging without design variables (Korn and Graubard, 1999; Chambers and Skinner, 2003). The pseudo-likelihood is commonly used in the analysis of complex surveys and has good design-based properties (*e.g.*, consistency). In addition, the pseudo-likelihood can be justified from an analytic perspective (Binder and Roberts, 2003). Because our interest is in estimation of the finite-population (*i.e.*, national) values of the diffusion model parameters, the pseudo-likelihood should perform well; however, more research into the use of weighted likelihoods in Bayesian methods is needed.

For our microsimulation models, we only require estimates of model parameters and not variance estimates. Unbiased variance estimation is difficult in this case for several reasons. First, we are combining data from two possibly overlapping national surveys and information about common clusters (*e.g.*, states, counties, strata, and PSUs) between the two surveys is not available. Second, we are combining information across multiple years and information about common clusters across years is not available. Last, design variables are not available in the public use data sets. To improve variance estimation, we could incorporate strata and PSU-specific effects into another level of conditional models for the NHIS and BRFSS and marginalize over these effects; however, this would further complicate our model and is not necessary for our purposes.

Three CISNET groups will use these diffusion model parameter estimates as microsimulation model inputs. We will combine the results from these models with a model for the time between endoscopy examinations. From this two-part model, we will generate endoscopy histories for individuals in our simulated populations. Examples of such input parameter generators used by CISNET groups, including one based on the model described in this paper, can be found at http://cisnet.cancer.gov/interfaces/. By comparing outcomes of interest (*e.g.*, cancer incidence and mortality) from the same simulated population with and without endoscopy use, we will estimate the effects of endoscopy on these outcomes. In this way, microsimulation models can greatly contribute to our understanding of the benefits of endoscopy without the cost and time associated with large randomized trials.

6 Acknowledgements

The authors thank the CISNET colorectal cancer group and Dr. Eric (Rocky) Feuer for their helpful comments and suggestions throughout the development of this model. This work was supported by the NCI-funded U01 CA-97427 and NCI contract #263-MQ-320084.

7 References

Behavioral Risk Factor Surveillance System, available at http://www.cdc.gov/brfss/. Accessed on May 19, 2006.

Berry, D. A., Cronin, K. A., Plevritis, S. K., Fryback, D. G., Clarke, L., Zelen, M., Mandelblatt, J. S., Yakovlev, A. Y., Habbema, J. D., Feuer, E. J., Cancer Intervention and Surveillance Modeling Network (CISNET) Collaborators (2005) Effect of screening and adjuvant therapy on mortality from breast cancer. *New England Journal of Medicine,* **353(17)**, 1784-92.

Cancer Intervention and Surveillance Modeling Network, available at http://cisnet.cancer.gov/. Accessed on May 7, 2006.

Binder, D. A. and Roberts, R. R. (2003) Design-based and Model-based Methods for Estimating Model Parameters, in Chambers R. L. and Skinner C.J. (eds.) *Analysis of Survey Data.* Chichester: Wiley.

Chambers, R. L. and Skinner, C. J. (eds.) (2003) *Analysis of Survey Data.* Chichester: Wiley.

Collins, J. F., Lieberman, D. A., Durbin, T. E., Weiss, D. G., Veterans Affairs Cooperative Study #380 Group (2005) Accuracy of screening for fecal occult blood on a single stool sample obtained by digital rectal examination: a comparison with recommended sampling practice. *Annals of Internal Medicine*, **142(2)**, 81-85.

Cronin, K. A., Binbing, Y., Krapcho, M., Miglioretti, D. L., Fay, M. P., Izmirlian, G., Ballard-Barbash, R., Geller, B. M., Feuer, E. J. (2005) Modeling the dissemination of mammography in the United States. *Cancer Causes and Control*, **16(6)**, 701-712 .

Diggle, P. J., Heagerty, P. J., Liang, K.-Y., and Zeger, S. L. (2002). *The Analysis of Longitudinal Data, 2nd Edition.* New York: Oxford University Press.

Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine*, **7**, 305-315.

Elliott, M. R. and Davis, W. W. (2005) Obtaining cancer risk factor prevalence estimates in small areas: combining data from two surveys. *Applied Statistics,* **54(3)**, 595-609.

Frazier, A. L., Colditz, G. A., Fuchs, C. S., Kuntz, K. M. (2000) Cost-effectiveness of screening for colorectal cancer in the general population. *Journal of American Medical Association,* **284**, 1954-1961.

Gelfand, A. E., Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association,* **85,** 398–409.

Gilks, W. R., Richardson, S., Spiegelhalter, D. J., (1996). *Markov Chain Monte Carlo in Practice.* New York: Chapman and Hall.

Healthy People 2010, available at http://www.healthypeople.gov/. Accessed on October 10, 2005.

Heagerty, P. J. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics,* **58**, 342-351.

Heagerty, P. J. and Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference.*Statistical Science*, **15(1)**, 1-26.

Jewell, N. P. and van der Laan, M. J. U.C. (2002) Current status data: Review, recent developments, and open problems. *Berkeley Division of Biostatistics Working Paper Series.* University of California, Berkeley.

Korn, E. L. and Graubard, B. I. (1999). *Analysis of Health Surveys.* New York: Wiley.

Knudsen, A. B. (2005). *Explaining secular trends in colorectal cancer incidenc and Mortality with an empirically-calibrated microsimulation model.* PhD doctoral thesis, Harvard University.

Liang, K. Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73(1)**, 13-22.

Loeve, F., Boer, R., van Oortmarssen, G. J., van Ballegooijen, M., Habbema, J. D. F. (1999) The MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. *Computers* *and Biomedical Research.* **32**, 13-33.

Mahajan, V. and Peterson, R. A. (1985) *Models for Innovation Diffusion.* Newbury Park, CA: Sage Publications.

Metropolis, N., Rosenbluth, A. W. , Rosenbluth, M. N. , Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machine. *Journal of Chemical Physics,* **21**, 1087-1091.

Miglioretti, D. L., and Heagerty, P. J. (2004) Marginal modeling of multilevel binary data with time varying covariates. *Biostatistics*, **5(3)**, 381-398.

National Health Interview Survey. Available at: www.cdc.gov/nchs/nhis.htm. Accessed on May 19, 2006.

Raghunathan, T. E., Xie, D., Schenker, N., Parsons, V., Davis, W., Dodd, K. W., Feuer, E. J. (Submitted) Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening.

Robins, J., Rotnitzky A., and Zhao L.-P. (1995). Analysis of semi-parametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association,* **90**, 106–121.

Rutter, C. M., Miglioretti, D. L., Savarino J. (In preparation) Bayesian microsimulation model calibration.

Table 1. Survey sample size and weighted percentage reporting they had ever had an endoscopy exam from the NHIS and BRFSS by birth cohort and gender.

		Females										Males									
	N	1910-1929		1930-1939		1940-1949		1950-1959		1960-1963		1910-1929		1930-1939		1940-1949		1950-1959		1960-1963	
Survey	States*	*N*	(%)	*N*	(%)	*N*	(%)	*N*	(%)	*N*	(%)	*N*	(%)	*N*	(%)	*N*	(%)	*N*	(%)	*N*	(%)
NHIS																					
1987	51	2,935	(27%)	1,414	(22%)	1,451	(13%)	0	N/A	0	N/A	1,827	(33%)	1,120	(22%)	1,178	(15%)	0	N/A	0	N/A
1992	51	1,367	(35%)	769	(29%)	995	(21%)	387	(11%)	0	N/A	815	(44%)	590	(31%)	810	(25%)	295	(16%)	0	N/A
1998	51	3,003	(38%)	1,941	(33%)	2,585	(25%)	3,024	(13%)	0	N/A	1,770	(50%)	1,502	(47%)	2,048	(30%)	2,542	(17%)	0	N/A
2000	51	2,533	(43%)	1,822	(41%)	2,367	(30%)	3,272	(14%)	350	(09%)	1,442	(49%)	1,366	(47%)	1,947	(33%)	2,660	(15%)	310	(08%)
2003	51	2,023	(43%)	1,704	(47%)	2,292	(40%)	3,118	(20%)	1,365	(11%)	1,012	(55%)	1,194	(55%)	1,860	(43%)	2,526	(21%)	1,153	(11%)
BRFSS																					
1988	7	1,169	(45%)	580	(35%)	725	(18%)	0	N/A	0	N/A	645	(55%)	416	(37%)	499	(34%)	0	N/A	0	N/A
1989	7	1,008	(48%)	553	(30%)	757	(19%)	0	N/A	0	N/A	589	(56%)	375	(41%)	546	(29%)	0	N/A	0	N/A
1990	6	909	(46%)	479	(31%)	734	(22%)	102	(13%)	0	N/A	512	(58%)	355	(44%)	535	(37%)	73	(12%)	0	N/A
1991	7	1,171	(49%)	680	(34%)	905	(26%)	234	(16%)	0	N/A	609	(57%)	459	(45%)	642	(30%)	165	(18%)	0	N/A
1992	4	715	(45%)	478	(37%)	649	(24%)	257	(14%)	0	N/A	432	(61%)	338	(42%)	450	(30%)	183	(20%)	0	N/A
1993	50	12,190	(39%)	6,545	(34%)	9,052	(20%)	4,804	(12%)	0	N/A	6,875	(47%)	4,777	(39%)	7,042	(24%)	3,868	(14%)	0	N/A
1994	4	1,101	(38%)	607	(34%)	844	(21%)	542	(13%)	0	N/A	555	(45%)	459	(40%)	689	(25%)	432	(13%)	0	N/A
1995	49	12,089	(38%)	7,208	(33%)	9,415	(21%)	7,729	(13%)	0	N/A	6,689	(52%)	5,074	(43%)	7,111	(26%)	5,977	(14%)	0	N/A
1996	4	1,026	(34%)	698	(33%)	870	(20%)	842	(12%)	0	N/A	559	(45%)	464	(41%)	645	(25%)	658	(14%)	0	N/A
1997	51	13,192	(44%)	8,769	(40%)	11,409	(27%)	12,517	(15%)	0	N/A	6,971	(52%)	6,061	(47%)	8,585	(32%)	9,798	(16%)	0	N/A
1998	2	550	(47%)	350	(34%)	421	(27%)	550	(14%)	0	N/A	239	(48%)	248	(43%)	310	(33%)	424	(16%)	0	N/A
1999	51	13,422	(48%)	10,474	(45%)	13,544	(32%)	18,414	(16%)	0	N/A	6,706	(56%)	6,965	(52%)	9,933	(36%)	13,833	(16%)	0	N/A
2000	4	739	(46%)	711	(44%)	969	(36%)	1,298	(17%)	158	(09%)	385	(63%)	430	(51%)	674	(36%)	1,022	(17%)	127	(14%)
2001	51	14,764	(54%)	13,214	(52%)	17,707	(42%)	4,594	(29%)	0	N/A	7,219	(59%)	8,804	(53%)	12,497	(41%)	3,264	(28%)	0	N/A
2002	51	17,812	(56%)	17,368	(56%)	22,079	(44%)	8,296	(33%)	0	N/A	8,516	(61%)	10,863	(57%)	15,512	(45%)	6,061	(31%)	0	N/A
2003	10	3,171	(58%)	3,444	(58%)	4,633	(48%)	2,296	(36%)	0	N/A	1,584	(62%)	2,251	(61%)	3,060	(49%)	1,624	(35%)	0	N/A

*Includes Washington D.C.