Methods

Data Sources

Numbers of cases of cancer reported as first diagnosed during 1999 and reported to the NCI by the SEER cancer registries on November 1, 2001, were stratified by sex, race (white, black, other), age group (0–4, 5–14, 15–24, …, 75–84, 85+) and county. Cancers of the lung and bronchus, colon and rectum, female breast, and prostate were analyzed separately; all other types of cancer were grouped together for analysis (for International Classification of Disease codes for these sites, see **http://seer.cancer.gov/siterecode/icdo2_d04152002**). Statistics for all cancers combined are the aggregation of these five cancer groups. Only malignant tumors were included; in situ and other benign tumors were excluded. These incidence data were available for the 480 counties included in the SEER Program in 1999, including 10 rural counties in Georgia and the first SEER data submission by the registries in Greater California, Kentucky, New Jersey, and Louisiana.

The numbers of deaths that occurred in 1999 were provided by the National Center for Health Statistics. Mortality was available for all 3074 U.S. counties, stratified by county, sex, race, age, and underlying cause of death. Stratified rates for death due to lung and bronchus, colorectal, breast, prostate, and other cancer were used as predictors of incidence for those cancers.

Population intercensal estimates for 1999, modified after the 2000 Census, were provided by the Census Bureau (see **http://seer.cancer.gov/popdata/methods.pdf** and **http://www.cancer.gov/newscenter/pressreleases/Census2000**). These counts were stratified in the same way as the incidence and mortality counts above.

Sociodemographic variables constructed for each county from the Area Resource File (Bureau of Health Professions 1999) and Census data (GeoLytics Inc. 1998) included urban/rural status (Butler and Beale 1994), household characteristics, income, education, occupation, medical facilities, and the percentage distribution of the population by race and ethnicity. The percentages of state and county residents who ever smoked cigarettes (males and females separately), who were at risk of obesity, who had no health care coverage, and female residents aged 50–64 who had had a mammogram in the last two years were lifestyle covariates calculated by aggregating public-use data for 1992–1998 from the CDC Behavioral Risk Factor Surveillance System (BRFSS) surveys (see **http://www.cdc.gov/brfss**; Pickle and Su 2002) at the state and county level. Age and race were available for each individual case but were grouped into the strata defined above for computational convenience. Geographic units for the analysis were county, state, and Census Region (Northeast, South, Midwest, West).

Statistical Methods

A hierarchical Poisson regression model was used to estimate the number of cases for all U.S. counties by their demographic and lifestyle profiles, based on the association of these profiles with cancer occurrence in the SEER counties. Specifically, the number of new cancer cases in county i (i = 1,...,3074), age group j (j = 1,...,10), denoted d_{ij}, was assumed to be distributed as a Poisson random variable, with mean $n_{ij}\lambda_{ij}$ where n_{ij} is the corresponding population at risk and λ_{ij} is the incidence rate in county i, age group j. We assumed a log-linear rate structure, i.e.,

$$\ln(\lambda_{ij}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\delta}, \boldsymbol{\zeta}) = \alpha_r + f(a_j)\boldsymbol{\beta} + \ln(m_{ij})\,\gamma + \mathbf{X}'_i\,\boldsymbol{\delta} + \mathbf{Y}'_i\,\boldsymbol{\zeta}$$

where α_r is the intercept for region r (r = 1,2,3,4) where county i is located, a_j is the centered midpoint of age group j, and for county i m_{ij} is the age j-specific mortality rate, $\mathbf{X}_i$ is a vector of demographic covariates, and $\mathbf{Y}_i$ is a vector of lifestyle covariates. A cubic function of age (a_j) was used to accommodate possible downturns in some cancer rates among the oldest groups.

Because the self-reported lifestyle covariates (smoking, obesity, health insurance and mammography use) from the BRFSS telephone surveys were thought to be fairly stable estimates of state values but likely to be measured with more error at the county level, an additional variance term was included for the "county residuals," i.e., the differences in county and state percentages for each of these covariates. That is, the vector $\mathbf{Y}_i$ was decomposed into state effect $\mathbf{Y}_{s(i)}$ and county residual $\mathbf{Y}^*_i$. Then the observed (BRFSS) county residuals $\boldsymbol{\gamma}^*_i$ were assumed to be normally distributed with mean 0 and variance $\sigma^2_{\gamma_i}$ where $\sigma^2_{\gamma_i}$ is inversely proportional to the population. This is equivalent to assuming that the observed county values vary randomly about their respective state values, with greater variation in small counties than in larger ones. This type of model is referred to as an errors-in-covariates model (Carroll et al. 1995).

The incident cases of cancer were analyzed separately by gender and location of the primary malignancy: breast, colon and rectum, prostate, lung and bronchus, and all other. Because of the computational difficulty in estimating the parameters when many of the age-county strata had no cases, we constrained the ages for analysis to be a minimum of 25 for breast cancer, 35 for lung and colorectal cancer, and 45 for prostate cancer. No age constraints were needed for other cancers. These age restrictions deleted 1.75% of the total cases from the analysis.

Covariates listed in the previous section were entered into the model as either scaled continuous variables or a series of binary variables. Collinearity diagnostics were used to select representative variables from each of the broad variable groups to include in the model. For example, only three of the four lifestyle covariates could be included in any one model; we kept smoking but excluded obesity in the lung cancer model, but did the reverse for the other sites. All main effects and two-way interactions were first included in the model but only very significant interactions ($p < 0.0001$)

were selected for the final models using backward stepwise fixed effects regression (SAS 1999). A Markov Chain Monte Carlo iterative process was then used to estimate the parameters of the full errors-in-covariates model structure described above (Spiegelhalter et al. 1999). With the inclusion of so many predictor variables, it was not necessary to include spatial correlation in the covariance structure.

This model was validated in several ways. First, the set of SEER counties with data available for 1995–1996 was split randomly into a training half and a validation half. Observed counts from the validation set were compared to predictions for these counties derived from the model on the training data. Results demonstrated the validity of the model and suggested ways to improve it. Then, for 1999 data from all SEER counties, predictions were compared to the observed SEER data; the model explained most of the variation in counts by age, sex, race, and county, and fewer outliers than expected were seen. Finally, predictions for other states (not in the SEER Program) were compared to the data reported to CDC (USCS 2002). All comparisons showed that this method provides accurate estimates of state incidence counts and rates. More detail on the parameter estimation methods and validation studies is available (Pickle et al. 2001; **http://srab.cancer.gov/incidence**).

The posterior mean predicted numbers of cases of each type of cancer were calculated for each combination of age, race, sex, and county. These estimates were summed to provide corresponding estimates for each state and region and for all cancers combined. Age-adjusted predicted incidence rates were calculated using the direct method of adjustment and the 2000 standard million population (Fleiss 1981). All rates are shown as cases per 100,000 population. The model predictions were also adjusted for reporting delay, as recently suggested by Clegg et al., in order to provide the numbers of cases that would be expected after data collection is complete at some time in the future (Clegg et al. 2002).

Graphical Methods

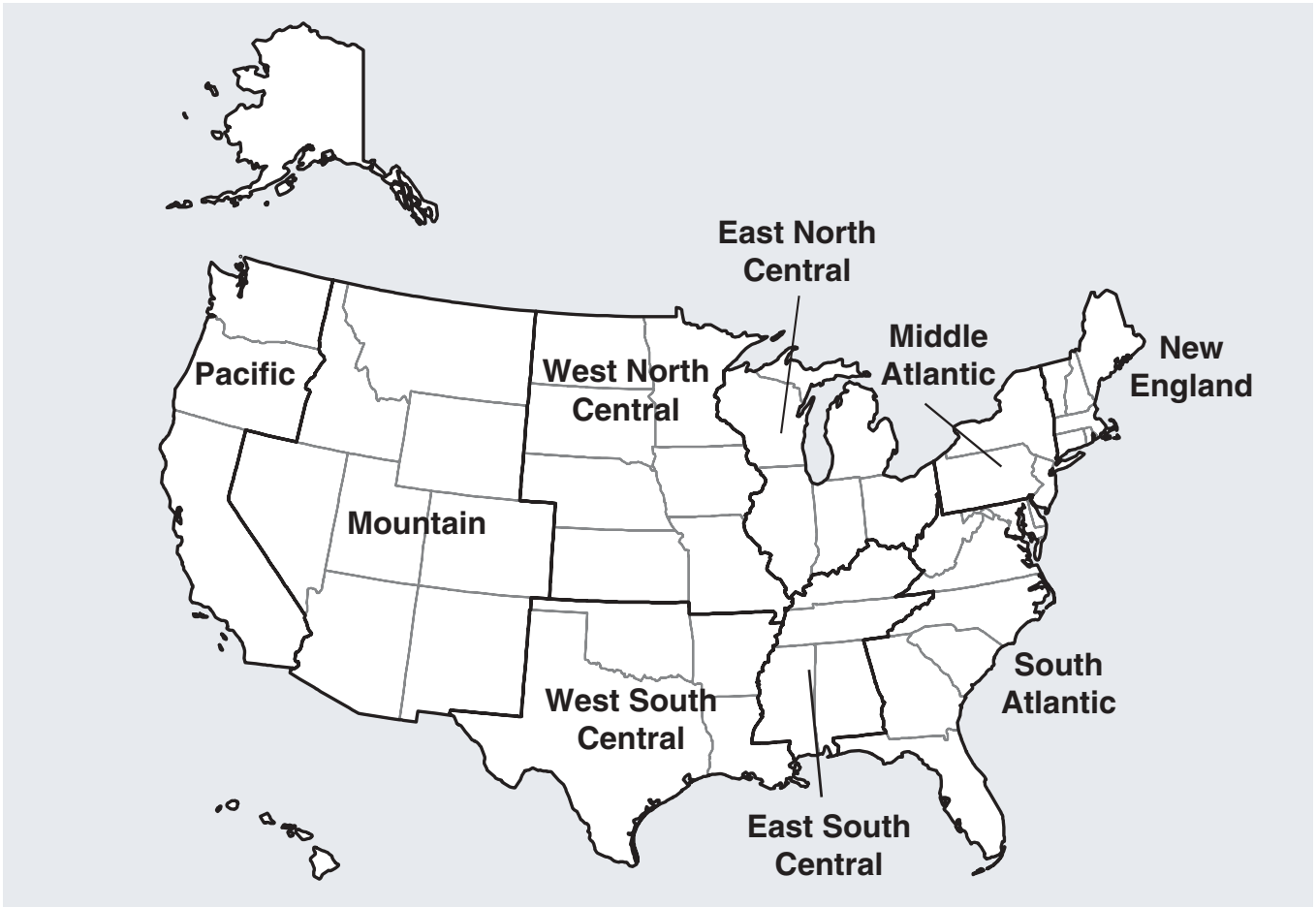
Results are presented in tables, maps, and graphs. All maps are shaded by county or state using colors chosen to permit use by color-blind readers (Brewer et al. 2003). Colors for state maps are assigned according to quintiles, i.e., about 10 states fall into each color category. A second series of maps shows these same age-adjusted state rates relative to the overall U.S. predicted rate. In this presentation, colors are assigned to equal intervals representing the proportional difference of each state's rate from the U.S. rate.

Although the basic geographic unit of the model was county, many counties have small populations that lead to a high degree of uncertainty about their expected number of cases. This uncertainty is greatly reduced by summing the predictions to the state level. However, interesting within-state patterns of incidence were apparent in maps of the county predictions. As a compromise, we present smoothed maps of age-adjusted county rates.

A nonparametric algorithm that included population weights was used to smooth away some of the underlying random variation of the county rates while highlighting broad patterns in the data (Mungiole et al. 1999). This algorithm is a two-dimensional version of a median-based moving average that readers may be familiar with from time series graphs. Because these maps present the same statistic as the state quintile maps, the same color scheme was used.

A graphic combining predicted counts and rates with maps of the rates is included for comparison of the relative (rate) and absolute (count) measures of the cancer burden by state. States are ordered by rates. In this graphic, the statistical estimates are shown as dots on the graphs, linked to the maps in the leftmost panel by color. Ninety-five percent confidence limits are shown as bars for each predicted rate and count, although the large dot size masks the bars for all but the most uncertain predictions. Note that the standard errors for model-based rates are generally smaller than those for empirical rates as shown in USCS. Guidance on the use of this and the other graphics is provided in the next section.

Figure 1. Definitions of Census Divisions in the United States



Reader's Guide

Tables

The predicted rates and counts for males and females are presented in separate tables, each ordered by state within Census Division (see Figure 1, page 8). Within a table, three columns of data for each of the cancer sites list the original prediction (rate or count), the delay-adjusted prediction, and the state's reported data from *United States Cancer Statistics: 1999 Incidence* (USCS 2002). The original predictions may be compared to the USCS report to judge the reasonableness of the model. The model predictions may also be used to supplement the USCS report where state and regional reports were unavailable. Data from high-quality cancer registries in 37 states and the District of Columbia were included in the USCS report (see USCS 2002, p. 4–5, for eligibility criteria).

Comparison of the predictions with and without delay adjustment can provide an estimate of the change in the numbers of cases or in rates that will occur in the future as more cancer cases that were diagnosed in 1999 are identified. As discussed by Clegg et al. (2002), the delay-adjusted figures provide a more accurate measure of the cancer burden in an area by removing variations due to reporting delay and updates in the records over time. Since these models are fit to SEER data, the modeled predictions implicitly project the counts assuming a reporting delay equivalent to that in the SEER registries. Variations from this assumed timing of data collection will affect the closeness of the predictions and the USCS reported figures. However, the delay-adjusted predictions do reflect what each state registry ultimately should report as data collection continues, assuming that the ultimate level of completeness is equivalent to that in SEER registries and that the ecologic associations inherent in the model hold for that area.

The reader will note that not all regions show counts or rates in the USCS report. Count totals were only published for two regions where all states reported data; rates were not computed if an insufficient number of states reported data (see Appendix L, USCS 2002, for details). The model predictions help to fill in these gaps and thus provide estimates for all regions. Since no delay adjustment is available for our "other cancer" group, only rates without adjustment are shown for this aggregated site; delay-adjusted counts were calculated by subtracting the sum of lung, colorectal, and prostate or breast cancer from this total count.

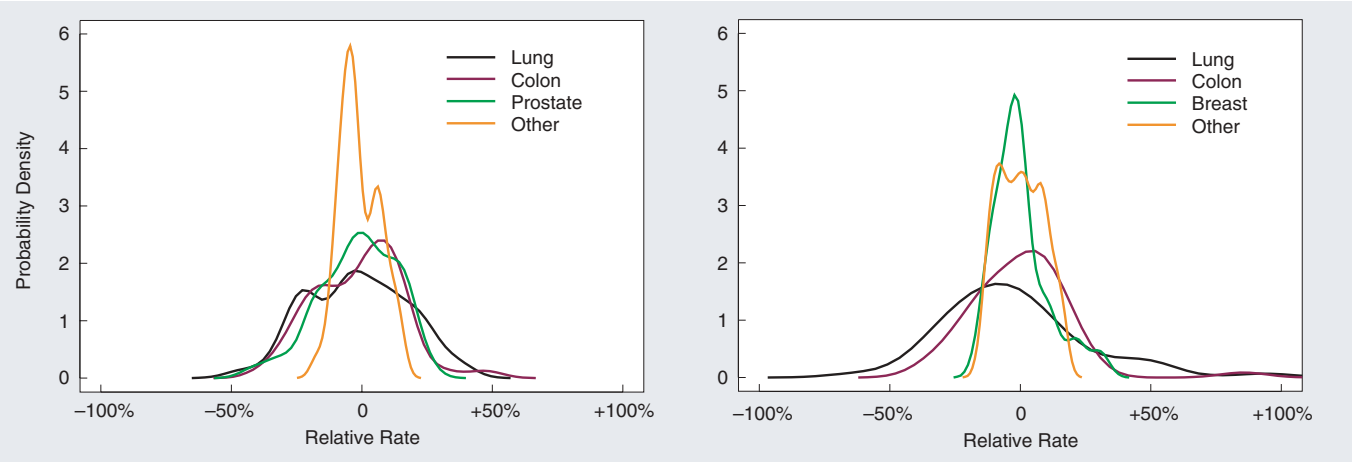
What might account for any differences between the predicted and reported cancer incidence? The prediction model assumes that the associations between the covariates and incidence rates is the same in all states as in the SEER areas; if this is not the case, the predictions will be inaccurate. Sudden spikes in screening

rates, as were observed for prostate cancer in the early and mid 1990s, perturb incidence and are difficult to capture accurately in models of this type. At the top of the table, predicted rates are given for the aggregated SEER and NPCR states for comparison. Also, regions with more population coverage by the SEER Program, e.g., Pacific, are expected to be estimated more accurately than those with lower SEER coverage. On the other hand, there is natural year-to-year variation in cancer incidence, especially in small population areas, and the model smooths over these to provide a more stable estimate of incidence than the observed data itself. Also, even though all the states included in the USCS report are certified as high quality, differences may arise from variations in registry operations such as completeness, timeliness, and specificity in coding the cancer site (Wingo et al. 2003). Finally, it should be noted that the USCS age-adjusted rates were calculated using 1999 population estimates extrapolated from the 1990 census, whereas we used updated estimates interpolated between the 1990 and 2000 censuses. These denominator differences will affect the calculated rates and their comparisons.

State Maps

State rates are presented as a series of small maps to facilitate the comparison of patterns across cancer type and gender and between predicted incidence and observed mortality rates (Tufte 1983). The map design is uncluttered; e.g., the legend is not shown on each map so the reader can focus on the patterns. A reader who wishes to know the actual rate predicted for a state should refer to the tables. Predicted incidence rates are presented both as age-adjusted rates and relative rates, i.e., the age-adjusted rate for the state divided by the corresponding U.S. rate. The quintile color categorization of the age-adjusted rates illustrates the patterns of rankings of the states whereas the equal interval color categorization of the relative rates illustrates patterns of the actual levels of the rates. For example, the age-adjusted rate map for other cancer among males shows a strong cluster of highest-ranking rates in the Northeast and low rates in the South but the relative rate map shows that these are all within 15% of the U.S. rate. This comparison highlights the small differences in age-adjusted rates that can appear to be striking on a rank-based map. It is important for the reader to remember that the colors are assigned for each map independently, so that the same color represents different ranges of actual rates for each type of cancer, although these ranges correspond to the same quintile category (lowest 20% of states, etc.). The rank-based quintile maps can best be used to answer the question, "Where are there rate differences?", while the relative maps best answer the question, "How large are these differences?". The relative maps illustrate the range of rates in comparison to the overall U.S. rate. Figure 2, page 11, shows the distribution of these predicted state rates for the four cancer sites overlaid on one density graph (a smoothed histogram). From these graphs, it is obvious that the breast and "other" cancer rates have narrower ranges than those of the lung, colon/rectum, and prostate.

Figure 2. Distribution of Predicted State Relative Incidence Rates by Cancer Site for Males (Left) and Females (Right), 1999



County Maps

The purpose of the smoothed county maps is to show within-state patterns of the predicted rates. These maps are shown in half-page size to facilitate identification of patterns at this scale. As noted above in the discussion of state maps, it can be misleading to compare similar colors across different types of maps or cancers. For example, Montana is classified as an average-rate state for prostate cancer incidence although many of its counties are in the highest quintile categories. This is the result of different distributions of state and county rates; the range of rates for the middle color category is 154.2–163.2 for states and 128.8–142.7 for counties.

The predicted county rates have been smoothed to remove some of the inherent variability in rates calculated for small populations. An example of a proper use of these maps would be to characterize the lung cancer rates among Texas males as being higher in the eastern than western parts of the state. It would be incorrect to try to identify the rate predicted for a particular county because its original prediction from the model may have been changed by the smoothing algorithm to be more like rates in neighboring counties.

Micromap Plots

The micromap plots summarize the results of the state maps and tables, but provide more detail than is possible in the color-categorized maps. For example, it is clear from this graphic that Utah's lung cancer rate is predicted to be much lower than New Mexico's, the second lowest state, but the state map categorizes all of the southwestern states into the low color category. Comparison of the rate and count panels demonstrates the dependence of the cancer count on population size—the highest number of male lung cancer cases is predicted in Florida, whose rate ranks only 9th. A glance at the series of small maps can identify clusters of similar-rate states, such as the band of high rates of male lung cancer along the Mississippi and Ohio rivers.