NAACCR Town Hall Meeting
May 9, 2006
3:00 PM Eastern

Present – Representatives from the following registries and organizations:

3 Canada: Alberta Cancer Registry, Manitoba Cancer Registry, Ontario Cancer Registry

31 U.S.: Alabama Statewide Cancer Registry, Alaska Cancer Registry, American Cancer Society, Arizona Cancer Registry, Arkansas Central Cancer Registry, Centers for Disease Control and Prevention, Colorado Central Cancer Registry, Delaware Cancer Registry, Florida Cancer Data System, Cancer Data Registry of Idaho, Illinois State Cancer Registry, Information Management Services, Louisiana Tumor Registry, Minnesota Cancer Surveillance System, Missouri Cancer Registry, National Cancer Institute, New Hampshire State Cancer Registry, New Jersey State Cancer Registry, New York State Cancer Registry, North Carolina Central Cancer Registry, Ohio Cancer Incidence Surveillance System, Oklahoma State Department of Health, Oregon State Cancer Registry, Pennsylvania Cancer Registry, Puerto Rico Central Cancer Registry, Rhode Island Cancer Registry, South Carolina Central Cancer Registry, Texas Cancer Registry, Utah Cancer Registry, Virginia Cancer Registry, Wisconsin Cancer Reporting System

3 NAACCR Staff: Dr. Holly Howe, Moderator – Executive Director, Joellyn Ellison – Program Manager of Data Evaluation and Publication, Royale Anne Hinds – Assistant to the Executive Director

Welcome **Holly Howe**

Holly welcomed all registries and organizations to the Town Meeting. The focus of the Town Meeting is Linda Pickle's project called Spatio-Temporal Estimation of Cancer Incidence in the United States for the American Cancer Society.
Holly mentioned a couple of other items before the investigators started their presentation:

- Holly asked for volunteers to complete testing on the Asian/Pacific Islander algorithm. The testing should take place in June and be completed in July. Contact Holly by email at hhowe@naaccr.org if you are interested in participating.
- Holly discussed the inclusion of county at diagnosis (NAACCR data item 90) for analysis. The county at diagnosis data item would not be used to produce county specific rates or data, but used as linkage to census data or for other purposes. In the past, an active consent has been used to obtain permission to use this data item for CINA Deluxe projects. There are registries that need to go through an IRB review or policy review before they have permission to use the county identifier. Non-response and non-consent have affected two projects negatively. We want to overcome whatever hurdles exist. If you need IRB review or other special step before granting permission, Holly would like to know these requirements so the information can be shared with researchers. The researchers can begin the necessary steps with these registries early in the process of getting access to the datafile.. Holly would also like us to think about implementing a passive consent process for projects that request the county data item. Please contact her at hhowe@naaccr.org with questions or comments.

- Linda Pickle's project plan has been updated and American Cancer Society wants to use her method to improve their process of estimating cancer cases they use in Cancer Facts and Figures. Cancer Facts and Figures is the most frequently cited reference for cancer incidence data. NAACCR would like this project to become a primary use of the data submitted to NAACCR, beginning with the file submitted for the Call for Data 2007.

Presentation to the Membership **Linda Pickle**

Linda introduced the NCI staff that will participate on the project.

- Dr. Barnali Das
- Dr. Ram Tiwari
- Dave Stinchcomb
- Dr. Rocky Feuer

The Spatio-Temporal Estimation of Cancer Incidence in the United States for the American Cancer Society's project information can be reviewed on the PowerPoint Presentation on pages 4-11.

Questions & Comments

1) Are there differences between a cancer site that does not have a poor prognosis and one that has very poor prognosis? Age is the strongest predictor of prognosis for all sites. Mortality was not a strong predictor of prognosis for any site, except lung. So far, Linda has not seen a difference by prognosis, but mortality can help one predict estimates if the prognosis is poor. If prognosis is good, Linda suggests reliance on other covariates.
2) Is the median based smoothing used? The median based smoothing is used county estimates. The predictions are better on the state level without smoothing.
3) Were data used from the 2000 census? How do you think it will affect the model the further we get away from those data? The 1995-2002 dataset was used, along with census data for 1990 and 2000. Linda hopes the census will have additional data to fill in gaps in the future.
4) Can state level data be included in the model for those states that cannot submit county data identifier? Theoretically, state level data could be used, but Linda chose to use county data at the beginning of the project. If county data form small state with a small population, such as Hawaii or Rhode Island, were not available, she would consider it. For the larger states, she does not think this procedure would work.
5) Does the model take into account that some registry data may not be as complete as others? The presentation does not include delay adjustment, but that is an ongoing discussion. The problem, with using delay adjustment, is that they do not have adjustment factors beyond SEER data. Linda would like to hear opinions from the membership on using a delay adjustment on the data.
6) Does the model take disease latency into account? Linda has struggled with this because locations where of people lived 20 or 30 years in the past is not known. Dr. Pickle has learned from professionals working in health disparities that socioeconomic factors claim more of an immediate role.
7) Is Hispanic information used separately or is it combined with White race information? This dataset combines Hispanic ethnicity information with White race information. A participant comments that the use of Hispanic ethnicity information combined with White race information could be a flaw in the model because

registries and NAACCR have been carefully separating race and ethnicity information.

8) What data were used for small counties when data were suppressed? When aggregated over time, there were not many counties with suppressed data. For these few counties, she used the state level data.

Closing Remarks — Holly Howe

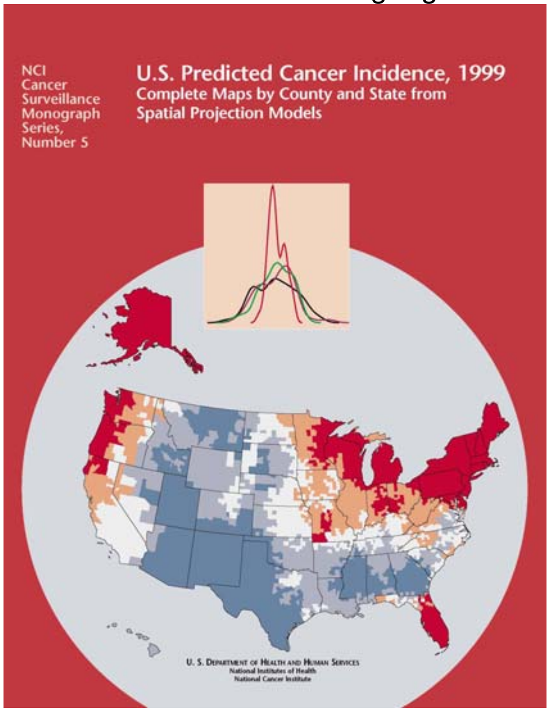
Dr. Pickle's project has been approved by the NAACCR IRB. On the previous consent, she received permission from 30 out of 40 states on the file, which is less than she had hoped. In preparation for the ACS data, she wants to use the new 1995-2003 data file, which will be available at the beginning of July. Every registry that met the CINA Deluxe criteria this year has already submitted a county variable (N.B., MN submitted, but then data were destroyed as per the agreement with them). Holly asked that each state respond to the active consent form that Joellyn will be sending for Dr. Pickle's project with an affirmative answer or information on the necessary steps to use the county of residence data item. Dr. Pickle can be contacted directly at picklel@mail.nih.gov or 301-402-9344. We hope to get a greater number of consents and believe that everyone will benefit from their participation.

Outline of presentation

- Background: spatial model results published in monograph
- Initial validation studies
- Extension of spatial to spatio-temporal model
- Expansion of input data to CINA Deluxe file from NAACCR
- Model covariates
- Results of application of new method to lung & esophageal cancers - Comparison to observed data, USCS reports, *Cancer Facts & Figures*
- Conclusions, next steps

Spatial projection models

- In 2003, NCI published a monograph showing 1999 cancer incidence rate maps for major cancers, based on input from SEER registries
- Could this same method be applied over time
  - to examine changing geographic patterns?
  - to extrapolate rates & counts to the current calendar year?
- Would this be an improvement over methods now used
  - In *Cancer Facts & Figures*?
  - For calculating registries' % completeness of case ascertainment (NAACCR)?



NCI
Cancer
Surveillance
Monograph
Series,
Number 5

U.S. Predicted Cancer Incidence, 1999
Complete Maps by County and State from
Spatial Projection Models

U. S. Department of Health and Human Services
National Institutes of Health
National Cancer Institute

Initial validation studies of case count prediction

- Applied NCI spatial model to SEER data (including 4 new SEER areas) for each year, predicted # cases in all states
- Compared predictions to known results (counts observed for SEER or published in *U.S. Cancer Statistics* (USCS))
- Compared model results to predictions from other methods, including the current ACS method
- Questions answered by these studies:

- Can we use a spatial model to accurately predict counts at the state & national level?

YES, SPATIAL MODEL BETTER THAN CURRENT ACS METHOD

- Which method is best for temporal projection ahead 4 years beyond the data? JOINPOINT (LINEAR EXTRAPOLATION)
- Can we improve predictions by using a mix of observed data and modeled predictions?

NO – ALL MODEL PREDICTIONS OR ALL DATA BETTER

Extension of spatial to spatio-temporal model

$dijt \sim$ Poisson $(nijt\ \lambda ijt)$ where $dijt$ = # cases in county i, age group j, time t

(separate models for each cancer site & gender combination)

OLD MODEL: $\ln \lambda\ ij=\beta o+f(aj)\beta +mij\gamma + Xi\delta + Zi\ ui$
NEW MODEL: $\ln \lambda ijt=\beta o+f(aj)\beta +mijt\gamma + Xit\delta +Zi\ ui + g(t)vt$

$f(aj)$ = centered cubic polynomial of age

$mij = \ln$(mortality rate), county i, age group j

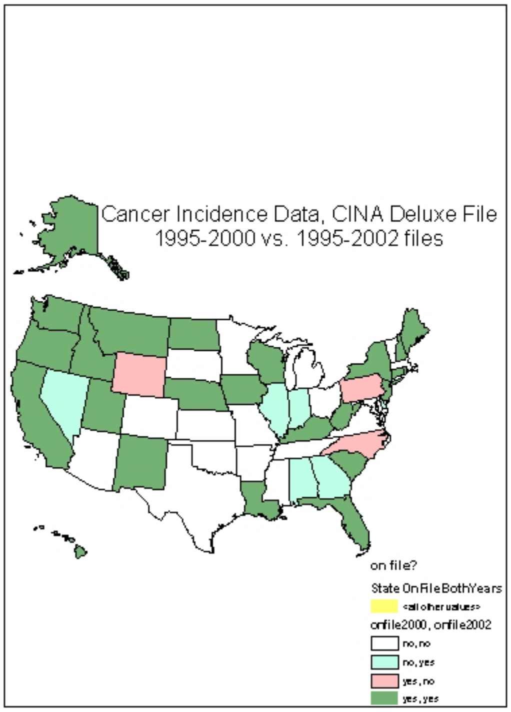
Xi = SES, demographic, lifestyle covariates
Zi = county location (centroid) or indicator of its Census Region
$g(t)$ = centered quadratic of year t

$ui \sim N(0, \Sigma sp)$ spatial autocorrelation
$vt \sim N(0, \Sigma temp)$ temporal autocorrelation

Proposal for *Cancer Facts & Figures* prediction of case counts 4 years ahead

- Use CINA Deluxe dataset for 1995+ to maximize spatial coverage (models are better with data than without)
- Select a parsimonious set of covariates that can reasonably predict the observed case counts
- Apply the NCI spatio-temporal model to the available time span to predict annual case counts for every state (summed from county predictions), using the new SAS PROC GLIMMIX (about 4 minutes each to run)
- Apply Joinpoint method to extrapolate from this time span 4 years ahead
- Tested this process using CINA 1995-2002 data for lung & esophageal cancers

What data are available in CINA 1995-2002 file?



Covariate selection

- Criteria for consideration as a covariate:
  - Available for every US county
  - Available over at least most of time span 1995-2002
  - Measured in a consistent manner over the time span
- Drop covariates:
  - to remove collinearity
  - that are highly correlated with others in topic group
  - that are not good predictors of any of the cancers
- 2-way interactions selected using a stepwise forward logistic (fixed effects) model from those that were significant at p<0.0001

Covariates selected for models by topic group

- **Personal characteristics**: age, race (W,B,O), sex
- **Year**
- **Geographic definition** (Census Division, county, state, centroid lat/long)
- **Mortality rate**
- **Density of medical facilities**: # MDs & # hospitals/1000 pop.
- **Ethnicity/origin**: % Hispanic, Black, Asian/Pac Islander, Amer Indian/Alaska native
- **Urban/rural indicators**: urban/rural continuum code aggregated to 5 groups
- **Socioeconomic status**: % living in poverty, % with 4+ years of college
- **Registry** is in SEER or NPCR program
- **Health insurance**: % adults with no health plan or insurance
- **Lifestyle**: % adults who ever smoked (M & F separately), % adults who met guidelines for vigorous exercise

- **Cancer screening behavior**
  - breast cancer: % women ages 50-64 who had a mammogram in last 2 yrs
  - Colon cancer: % ages 40+ who had colon-, sigmoid-, or proctoscopy in last 5 yrs
  - Other cancer models:
    - F: % adult women who had Pap smear in past 5 yrs
    - M: % men ages 40+ who ever had a PSA test

Model checking

- Goodness of fit (GOF) statistics
  - Hosmer-Lemeshow X2 from fixed effects logistic model
  - Usual GOF statistic:

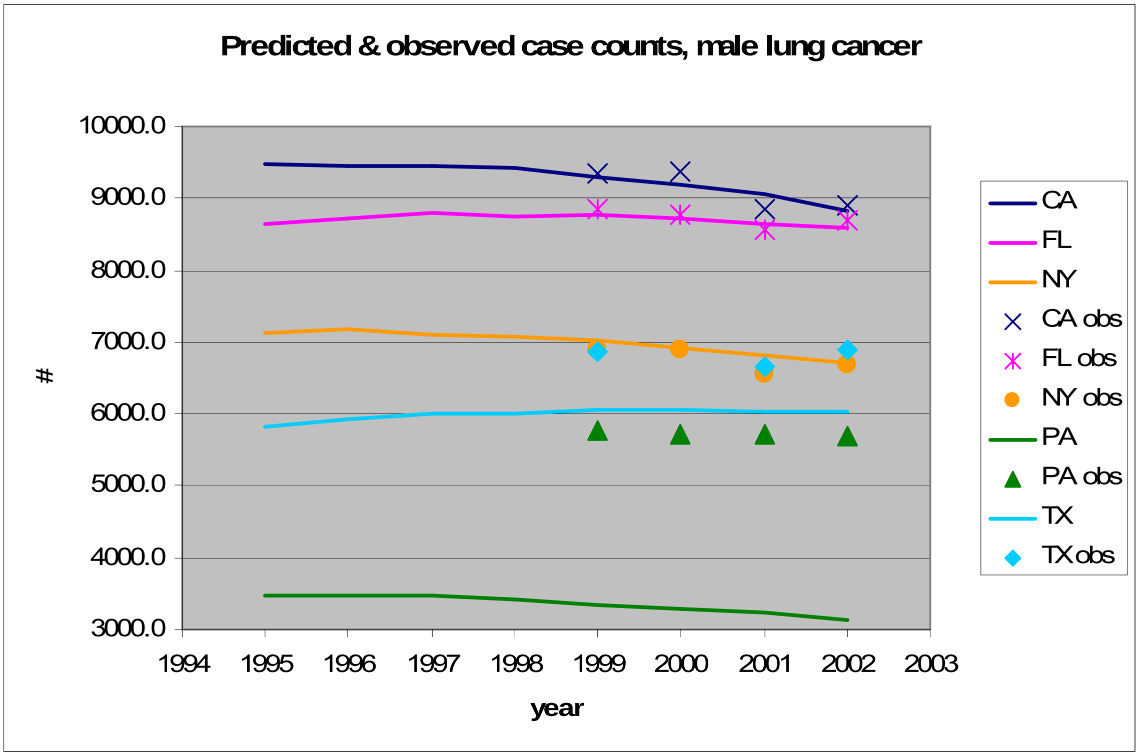
  - % of poorly fit data points (% |normalized residuals| > 2 & > 3)
  - Are the poorly fit data points clustered geographically?
  - Check that there are no unbelievably high predicted rates
  - Comparison of predicted #s by year to USCS reports

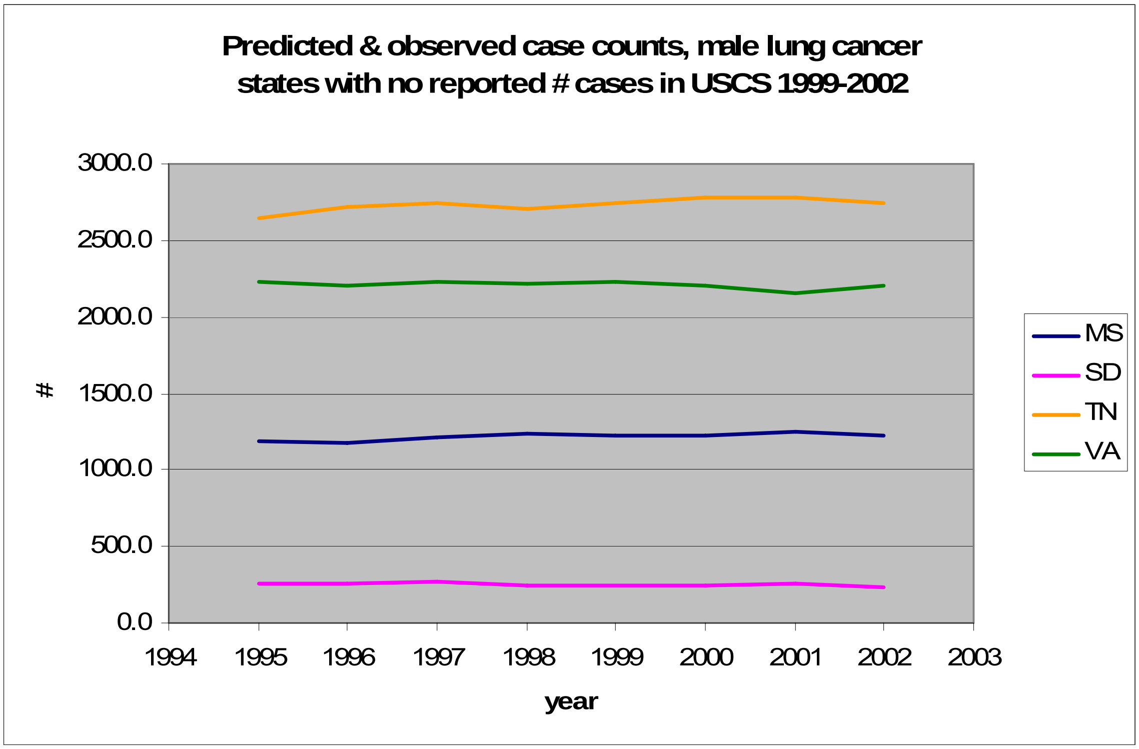
(some states not contributing data to the model are in USCS)

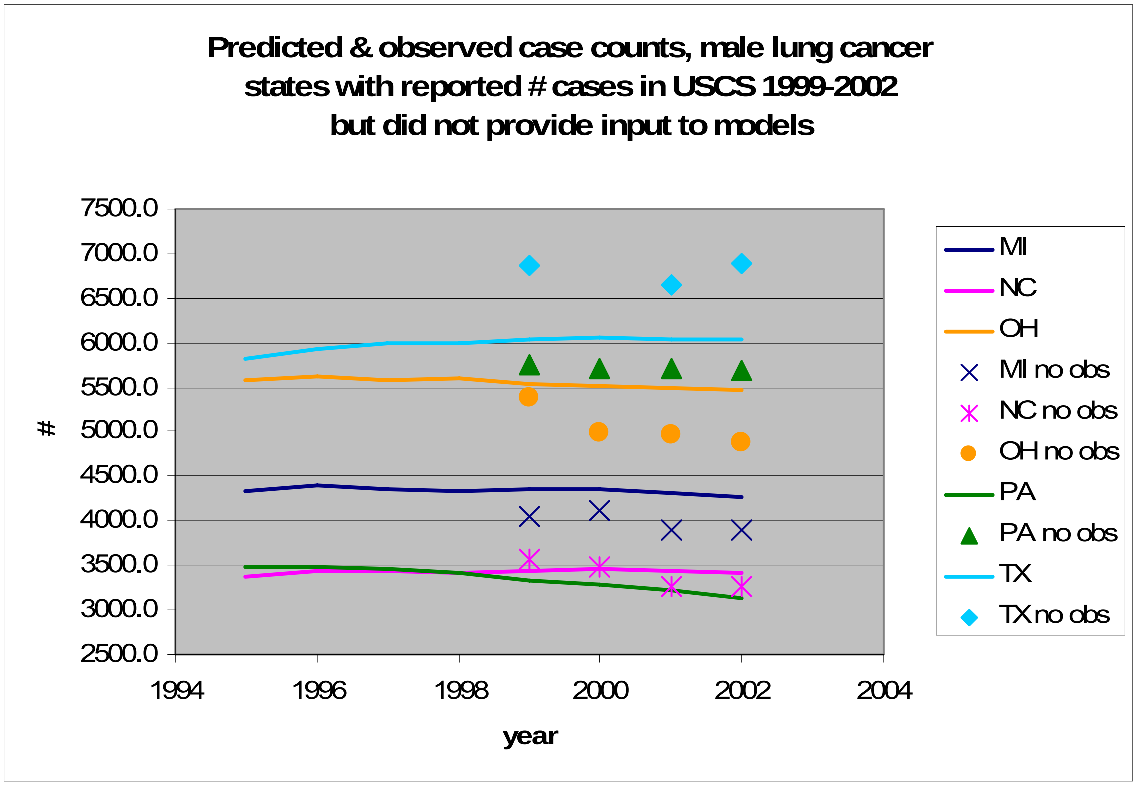
- Variance components
  - Reduction of overdispersion parameter toward 1.0
  - Significance of spatial and temporal autocorrelation

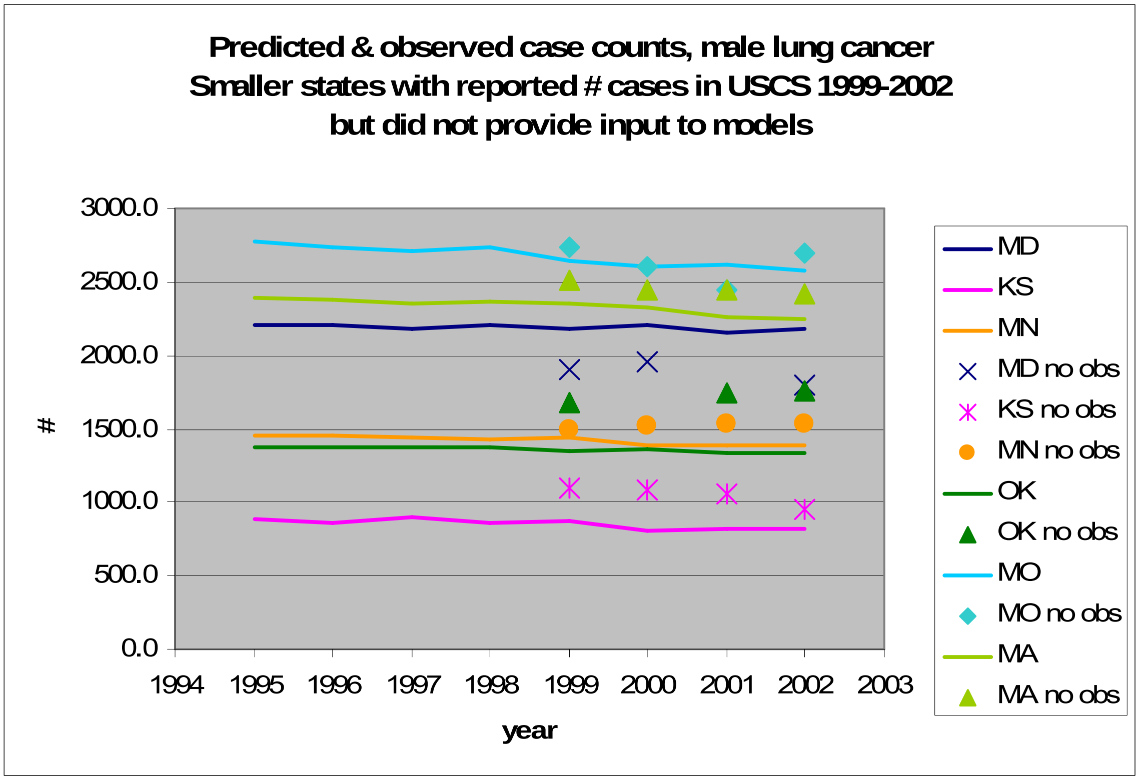
Results – Male lung cancer incidence

Predicted vs. USCS reported # for largest states




Predicted & observed case counts, male lung cancer
states with no reported # cases in USCS 1999-2002
3000.0
2500.0
2000.0
1500.0
1000.0
500.0
0.0
#
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
year
MS
SD
TN
VA

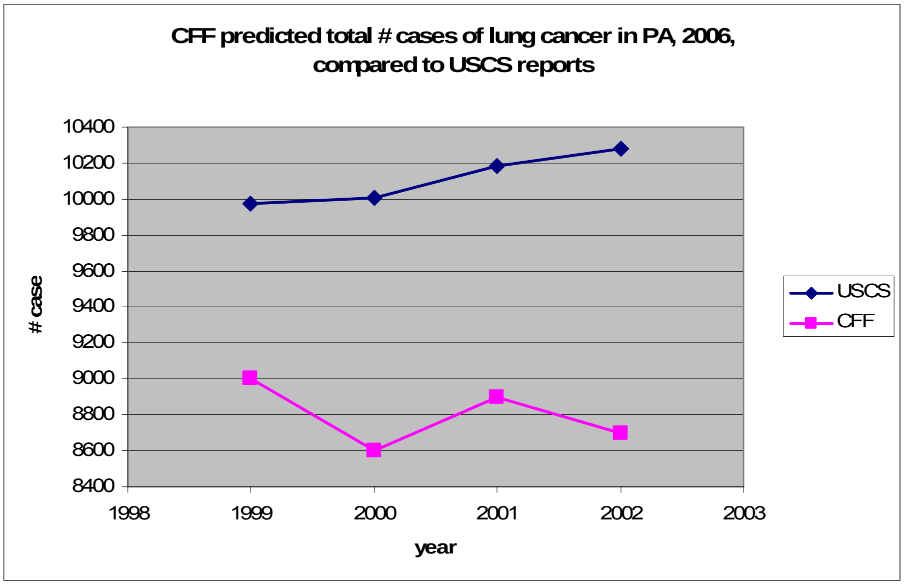

Predicted & observed case counts, male lung cancer
states with reported # cases in USCS 1999-2002
but did not provide input to models
7500.0
7000.0
6500.0
6000.0
5500.0
5000.0
4500.0
4000.0
3500.0
3000.0
2500.0
#
1994
1996
1998
2000
2002
2004
year
MI
NC
OH
MI no obs
NC no obs
OH no obs
PA
PA no obs
TX
TX no obs



Predictions of lung cancer in PA

- Male lung cancer in 2000
  - USCS report = 5709
  - Model prediction
    - Using CINA 2000 file = 5572 (incl. PA)
    - Using CINA 2002 file = 3295 (excl. PA)
    - Single year model, using SEER17 data = 7294
- M+F lung cancer:

USCS vs. CFF



Comparison of % error* in 2002 for states that did & did not provide data for model

	States providing model data	States not providing data
Average	0.13%	-6.30%
Minimum	-9.79%	-45.00%
Maximum	+7.80%	+21.34%

* % error = 100*(predicted # - USCS #)/USCS #

Comparisons with 2006 *Cancer Facts & Figures* projections

- Projected model predictions for 1995-2002 to 2006 using Joinpoint method
- Comparison to CFF projection for 2006
  - LUNG:
    - Total # M+F projected to = 187,910 by model, 174,480 by CFF
    - model is 7.7% higher for total U.S. count
    - Differences by state ranged from -20% to +26%
  - ESOPHAGUS (only U.S. total shown in CFF):
    - Total # M+F = 14,327, very close to CFF 14,550
    - State plots of #s over time are very variable

Conclusions & Next steps

- New spatio-temporal model is an improvement over previous methods for predicting case counts for next calendar year
- Counts for missing registries well predicted if state's rate is much like other registries in its region &/or those with similar covariate values (e.g., NC); otherwise registry estimate can be poor (e.g., PA)

- Advantages of model-based prediction
  - Can fill in gaps of missing registries or years within registry
  - Provides smoothed estimates of annual registry counts
  - Unlike current CFF method, permits rates to vary by geography and local characteristics, providing more accurate predictions

- Next steps
  - Run models using same CINA 1995-2002 data for 20 sites (top 15 for M & F in Annual Report to the Nation), judge feasibility of modeling all of these sites for CFF
  - Add 2003 data when available, repeat process for CFF 2007 estimates

Initial set of covariates

- Medical facilities
  - # physicians per 1000 population
  - # hospitals per 1000 pop
  - # screening mammography facilities per 1000 pop
- Ethnicity/origin
  - % of Hispanic origin
  - % of total pop who are black
  - % of total pop who are American Indian or Alaskan Natives
  - % of total pop who are Asian/Pacific Islanders
  - % of total pop who are foreign born
  - % of households in which no person ages 14+ speaks only English and who does not speak English very well
- Socioeconomic status
  - per capita income
  - % living below federal poverty line
  - % adults over 25 with < 9 years of education
  - % adults over 25 with high school education
  - % adults over 25 with 4+ years of college education
  - % unemployed
  - % adults employed in white collar jobs
  - median house value
  - coefficient of income inequality (gini)
- Cancer screening
  - % women ages 50-64 who had a mammogram in past 2 years
  - % women ages 20+ who had a Pap smear in past 5 years
  - % of persons ages 40+ who had a colonoscopy, sigmoidoscopy or proctoscopy in past 5 years
  - % of men ages 40+ who ever had a PSA test
- Health insurance
  - % of persons ages 18+ who do not have a health plan or health insurance
- Lifestyle
  - % of males ages 18+ who ever smoked cigarettes
  - % of females ages 18+ who ever smoked cigarettes
  - % of persons ages 18+ who have body mass index >=30
  - % of persons ages 18+ who met guidelines for vigorous exercise in 2001-3
- Urban/rural indicators
  - % urban pop
  - USDA urban/rural code (0=most urban, 9=most rural)
  - # persons/square mile (annual populations/land area)
- Other
  - Geography: census division (n=9), state, county, lat, long
  - SEER or NPCR registry