National Cancer Institute

Holey Data: Prediction and Mapping of **US** Cervical Cancer Incidence Rates

Marie-Jo Horner (NCI)

Dave Stinchcomb (NCI)

Joe Zao (IMS)

James Cuccinelli (IMS)

hornerm@mail.nih.gov

September 28th, 2008

ESRI GIS Conference

Washington, DC

Analytic issues in conducting study of US cancer incidence

1. Intro to cervical cancer study
2. Problem of holey data
3. Spatial & temporal statistical modeling
4. Smoothing with Headbang
5. Results

1. Introduction

- Goal: establish a baseline of cervical cancer incidence for entire USA prior to introduction of HPV vaccine
- New vaccine- be able to assess vaccine uptake

- Methods: GIS and spatial analytic methods to examine spatial-temporal trends
  - → incidence, mortality, screening
  - → regional, state, county level

Goals of Study

- Geographic baseline for assessing vaccine impact:
  - Incidence and mortality
  - Differences in Racial/Ethnic groups
  - Differences in age profiles
  - Differences in temporal trends
  - Differences in screening rates
  - Differences by socio-economic factors
  - Availability of health service providers
- Foundation for future research- e.g. Vaccine Uptake and Policy Guidance

Background

- Cervical cancer
  - US
    - 13th most common cancer among women
    - Incidence and mortality rates higher for Blacks and Hispanics
    - Median age of diagnosis is 48
      - 15.5% between 20 and 34
      - 26.2% between 35 and 44
    - 5-year relative survival rate is about 72%
      - Importance of mapping incidence, not just mortality

Risk Factors, Screening

- Risk factors:
  - Primary risk factor: human papillomavirus (HPV)
    - Very common STD; many types
    - Types with high risk for cancer: 16, 18, 31, 33, 35, 39, 45, 51-52, 56, 58-59, 68-69
  - Other risk factors: smoking, HIV, chlamydia, diet, etc. ADD STUFF FROM MODEL
- Screening
  - Cervical cytology (Pap smear)
  - Recommended for adult women every 3 years

HPV Vaccine

- New HPV vaccine
  - HPV types 16 and 18 – cause of most cervical cancers (70 percent)
  - Recommended for girls 11-12 years of age
  - Estimated cost: 360 $US per course of treatment

Data Sources

- Primary data sources
  - Cervical cancer incidence
    - NCI / SEER (17 areas)
    - CDC / NPCR through NAACCR CINA Deluxe
  - Cervical cancer mortality from NCHS

Time period: 1995-2004

Data sources
Methods (cont.)

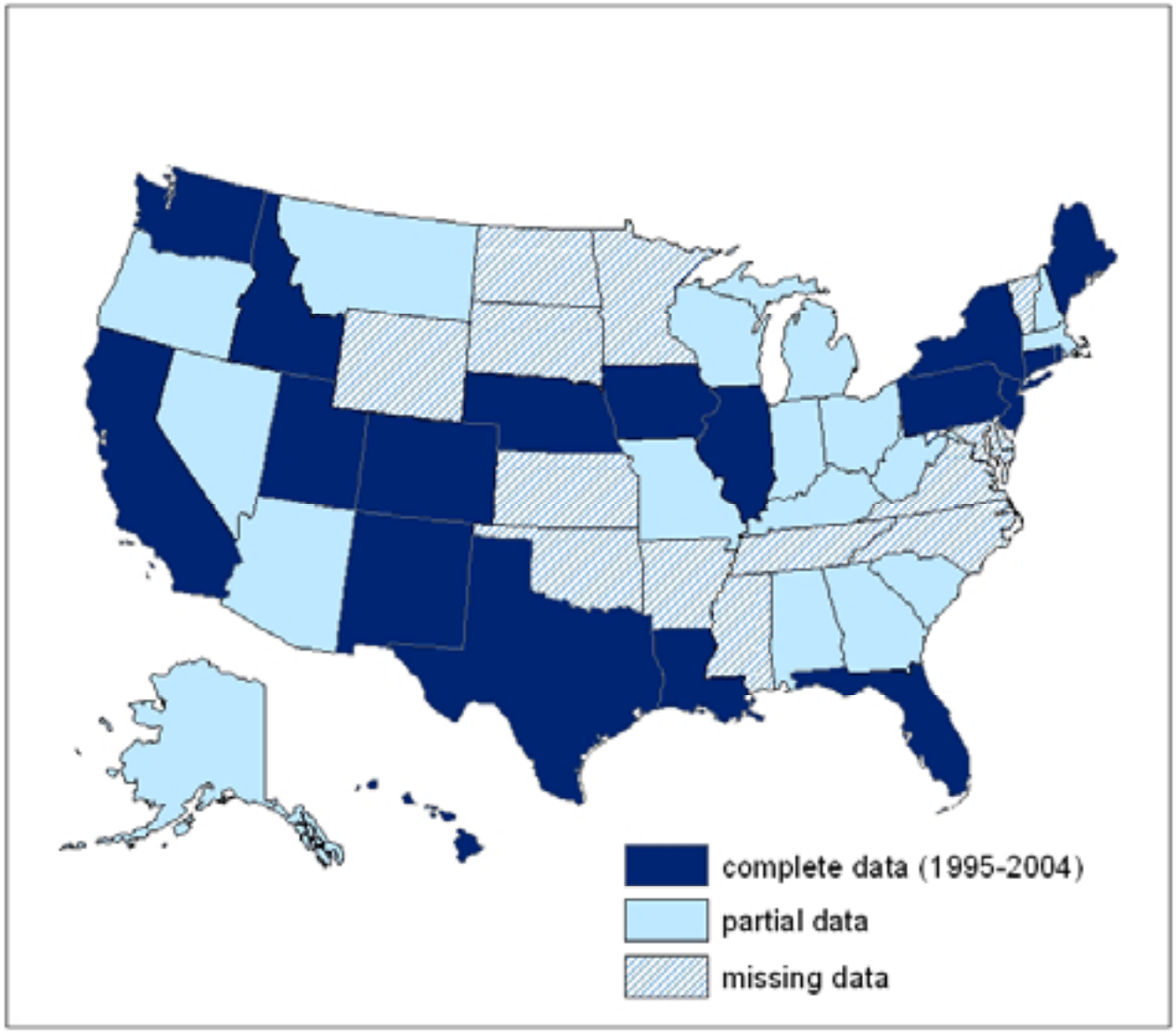
- Other data sources
  - PAP smear screening rates (BRFSS)
  - Demographic & SES data (US Census)
  - General and Ob/Gyn physicians per capita from HRSA (Area Resource File)

- Analysis & visualization methods
  - Spatial/temporal incidence model
  - Spatial smoothing (Headbang)

2. Analytic issues: Holey Data

Problem

- Missing / non-participation in study: 15 states
- Partial data: 16 states & DC (missing 1-4 years)
  - Based on QC certification / completeness of data



North American Association for Central Cancer Registries
CINA Deluxe File 1995-2004

Partial data (missing years)

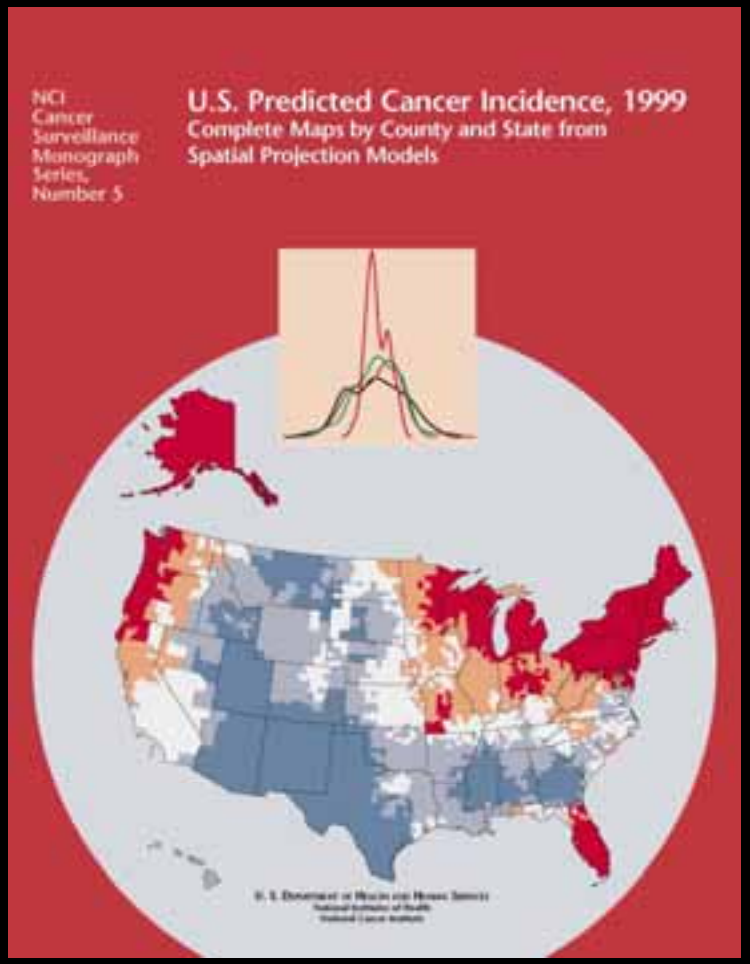
1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
			Alabama	Alabama	Alabama	Alabama	Alabama	Alabama	Alabama
	Alaska	Alaska	Alaska	Alaska	Alaska	Alaska	Alaska	Alaska	Alaska
Arizona	Arizona	Arizona	Arizona	Arizona	Arizona	Arizona	Arizona	Arizona	
			Wash Dc	Wash Dc	Wash Dc	Wash Dc	Wash Dc	Wash Dc	Wash Dc
				Georgia	Georgia	Georgia	Georgia	Georgia	Georgia
			Indiana	Indiana	Indiana	Indiana	Indiana	Indiana	Indiana
Kentucky	Kentucky	Kentucky		Kentucky	Kentucky	Kentucky	Kentucky	Kentucky	Kentucky
Michigan	Michigan	Michigan	Michigan		Michigan	Michigan	Michigan	Michigan	Michigan
			Missouri	Missouri	Missouri	Missouri	Missouri	Missouri	Missouri
	Montana	Montana	Montana	Montana	Montana	Montana	Montana	Montana	Montana
		Nevada	Nevada	Nevada	Nevada	Nevada	Nevada	Nevada	Nevada
Etc…									

What to do with the holes?

?

3. Spatial-temporal projection model

- Work of Linda Pickle & others
- 2003 NCI monograph of 1999 cancer incidence rate maps for major cancers, based on input from SEER registries
- This same method be applied over time
  - to examine changing geographic patterns
  - Also used to extrapolate rates & counts to the current calendar year (ACS Facts and Figures)



Spatial-temporal

- Previously, assumed that a state's INCIDENCE : MORTALITY ratio was the same as that of combined SEER regions.

- *We know this is not always true. Why?*

Spatial-temporal

- *Ok way*: INCIDENCE : MORTALITY ratio
  → Survival/screening differences.

- *Better way:* State- and County-level maps of estimated incidence Based on: Statistical modeling of county-level demographic and lifestyle characteristics, screening rates & cancer data.

- Unlike pure spatial smoothing, also takes into account similarities of places that may not be neighbors, but <u>similar in terms of other factors</u>

Filling in the gaps

- Predictions fill in gaps where state cancer registries are incomplete or choose not to participate in study.

- Map incidence and mortality
  → allows exploring of effects of survival and screening

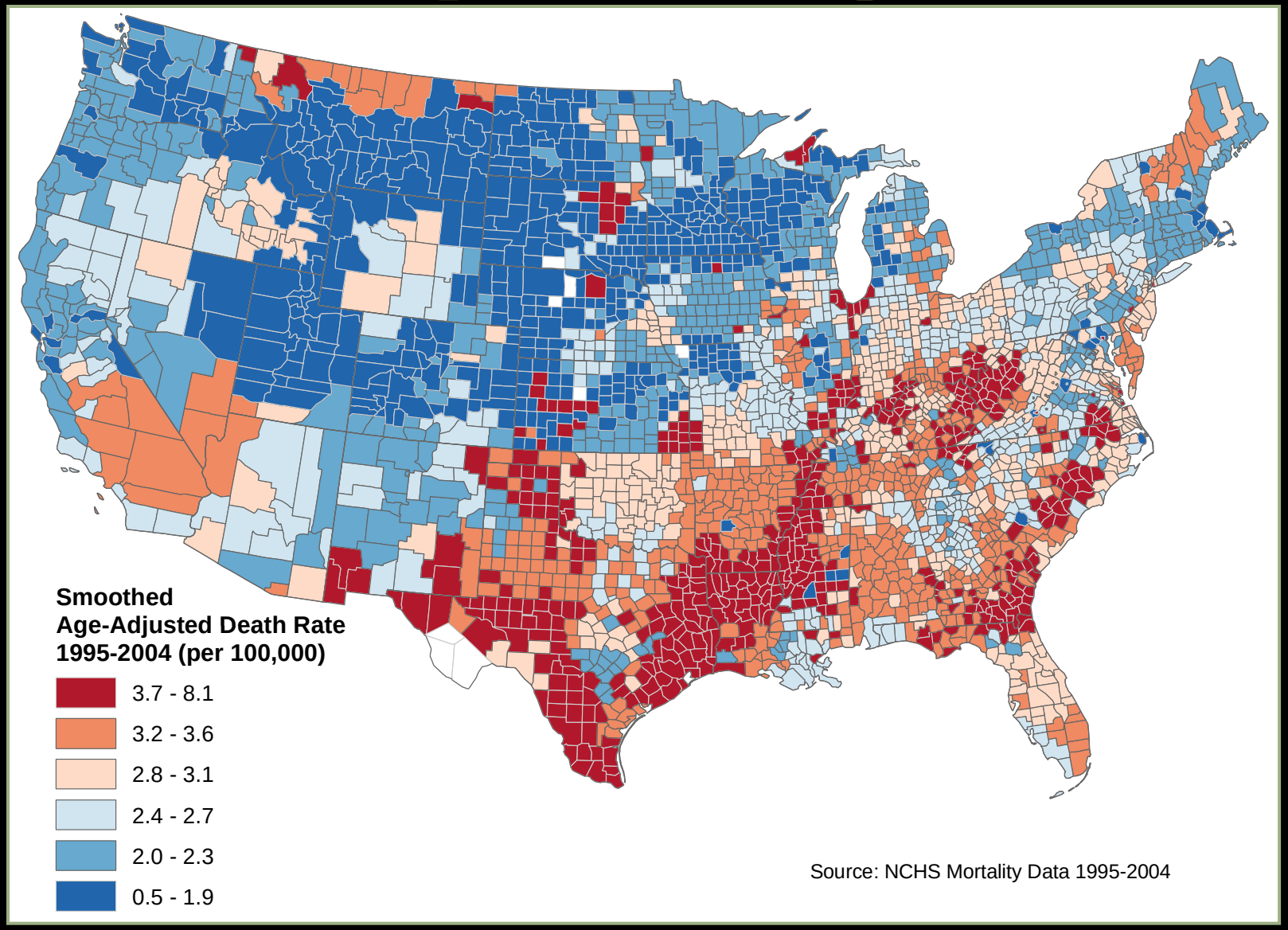
Benefits of predictions

- Smoothed maps of county-level incidence allow us to see regional differences other than by state alone.
- Administrative borders may not be most accurate or meaningful method for tabulating differences in cancer rates.

Inputs to the model

- CINA Cases stratified by race, sex, age group, county
- Deaths rates from NCHS stratified by county, sex, race, age, cause of death.

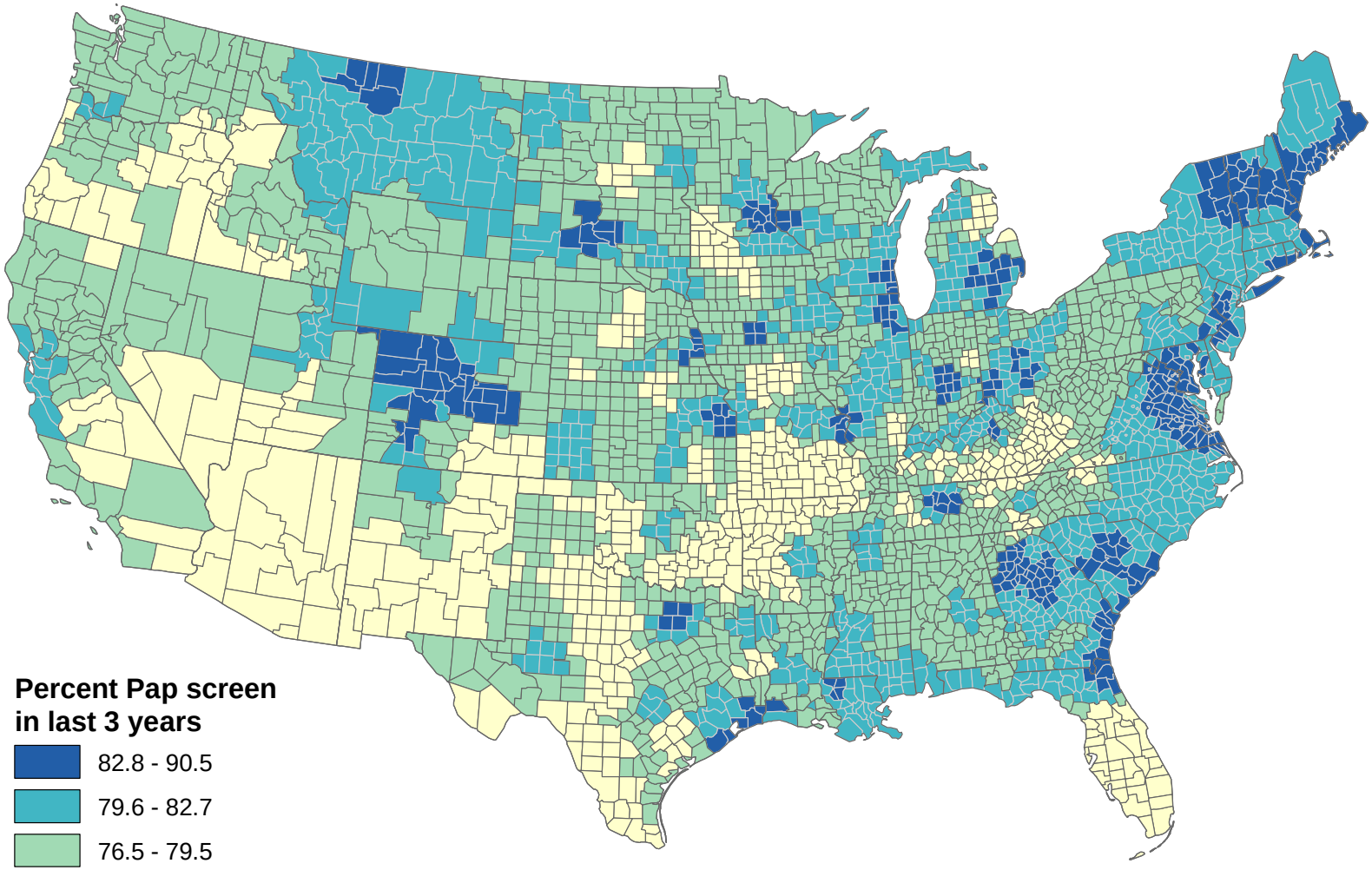
Mortality – County Level



Inputs to the model (cont.)

- From BRFSS surveys % state and county level:
  - ever smoked cigarettes,
  - risk of obesity,
  - no health coverage ,
  - female residents who had pap (starting in 2000+), used as lifestyle covariates

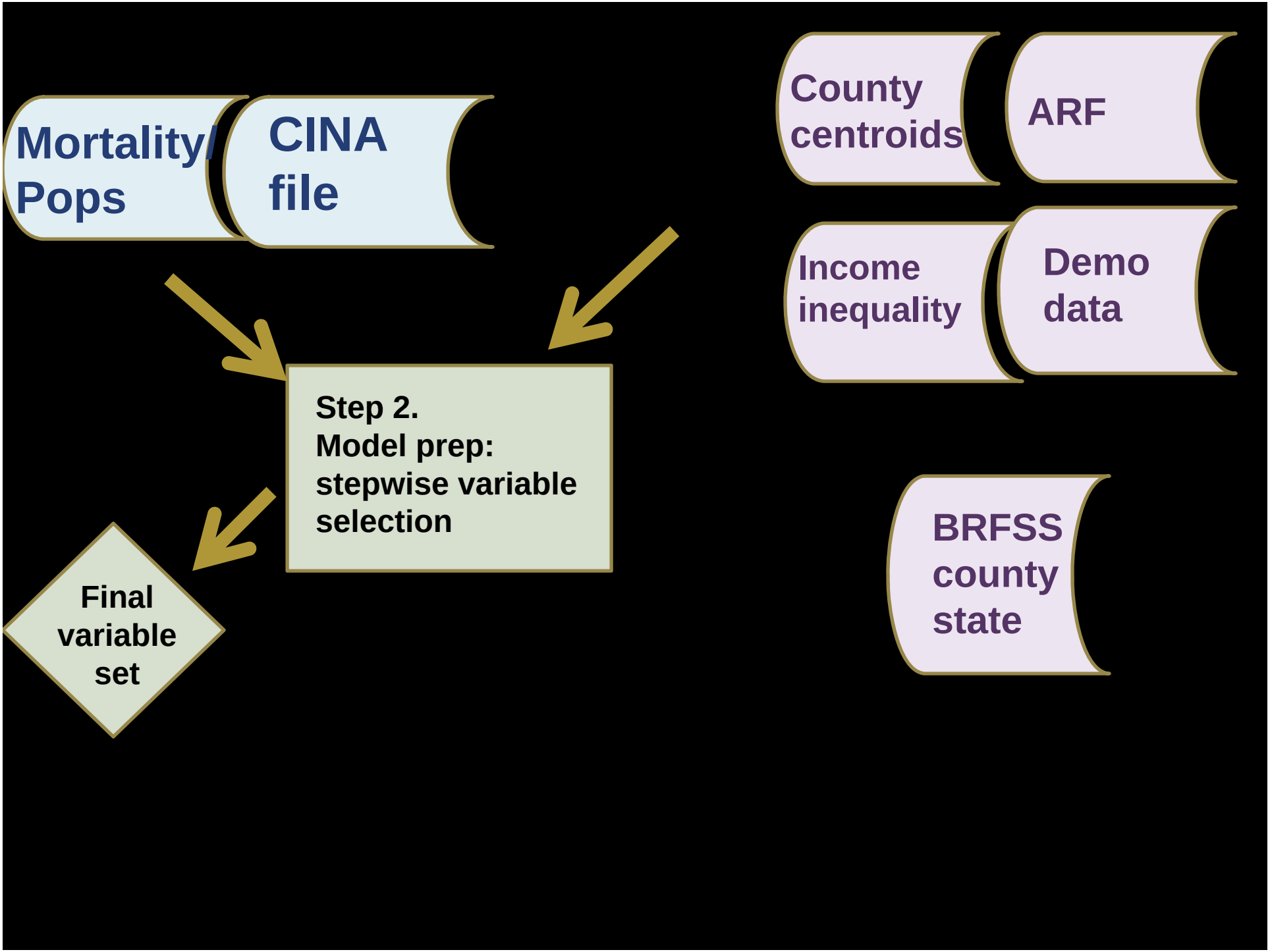
Screening Rates – County Level Estimates

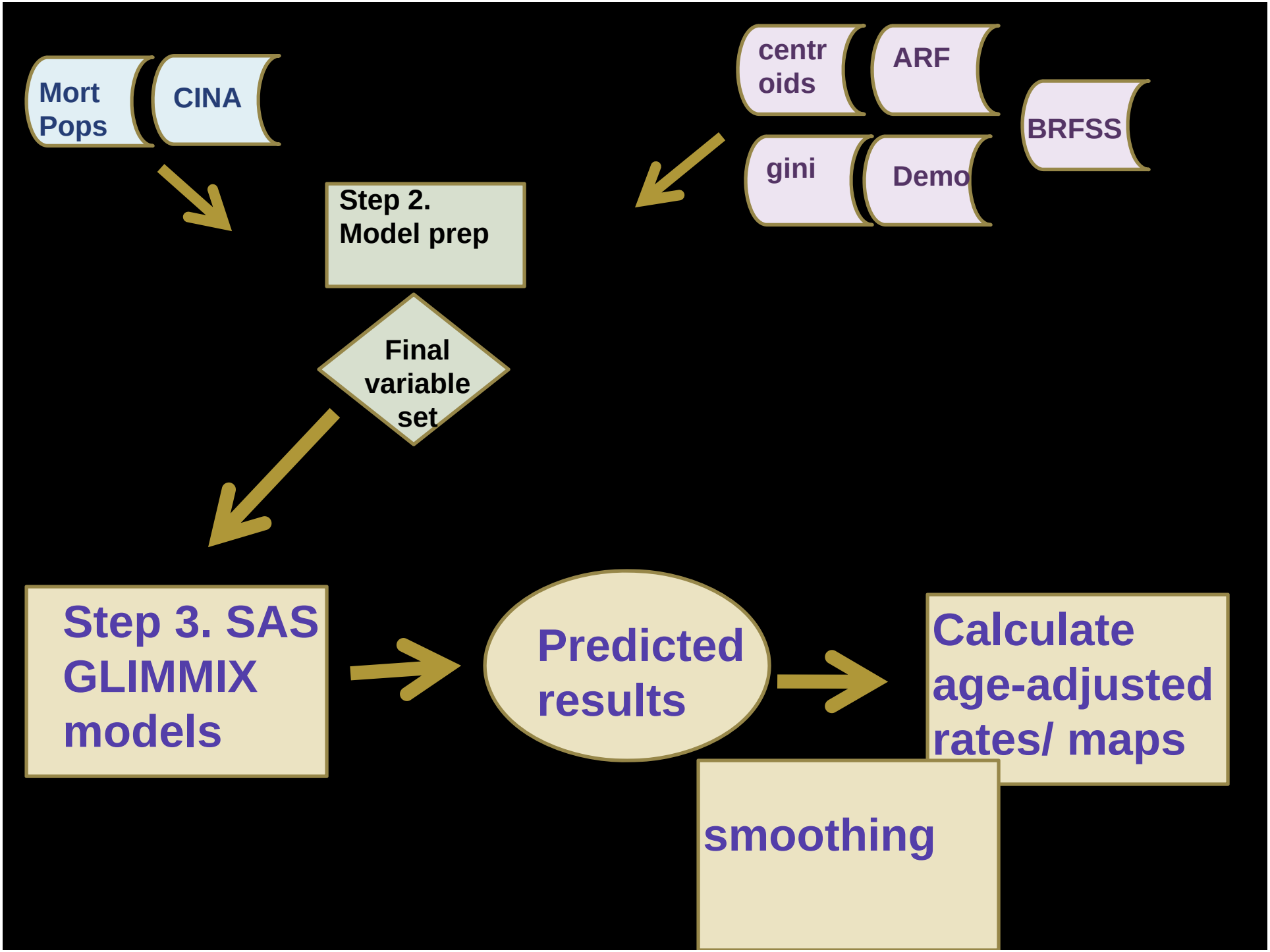


Source: NCI small area estimates based on BRFSS and NHIS data 2000-2003

Inputs to the model (cont.)

- County level SES variables from ARF file and Census Data.
  - urban/rural, household characteristics,
  - income,
  - education ,
  - occupation,
  - medical facilities,
  - % pop by race/ethnicity.


Mortality/ Pops
CINA file
County centroids
ARF
Income inequality
Demo data
Step 2.
Model prep:
stepwise variable selection
BRFSS county state
Final variable set


Mort Pops
CINA
centr oids
ARF
BRFSS
gini
Demo
Step 2. Model prep
Final variable set
Step 3. SAS GLIMMIX models
Predicted results
Calculate age-adjusted rates/ maps
smoothing

Validation

- Models have been validated to provide reasonable fit across entire nation
  - Pickle LW, Hao Y, Jemal A, Zou Z, Tiwari RC, Ward E, Hachey M, Howe HL, Feuer EJ. A new method of estimating United States and state-level cancer incidence counts for the current calendar year. *CA Cancer J Clin* 2007 57:30-42
  - More at http://srab.cancer.gov/
  - Same method used for ACS Facts and Figures.

Problems and Solutions

- Small Area Problems
  - High uncertainty about predicted number of cases in small population counties

- Solutions
- *Ok way*: resolve by summing to state level.
- *Better way*: within-state patterns of expected counts were interesting so smooth maps of age-adjusted county rates using HEADBANG

3. Headbang

3. Headbang

Not this.

Source: Google search

http://srab.cancer.gov/headbang/

Statistical Research & Applications Branch - Head-Bang PC Software - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://srab.cancer.gov/headbang/ headbang

Most Visited Getting Started Latest Headlines

National Cancer Institute
U.S. National Institutes of Health | www.cancer.gov

Statistical Research and Applications
Cancer Control and Population Sciences

Statistical Methods | Software | Funded Research | GIS | Publications | About SRAB Research

Search: GO

Cancer Control and Population Sciences Home

Surveillance Research

Statistical Research & Applications Branch

Head-Bang:

Head-Bang Home
Download Head-Bang Software
Frequently Asked Questions
Getting Help

Related Materials:

Methods & Software for Population-based Cancer Statistics
SEER*Stat
HeadBang for ArcMap

This page links to some files in Portable Document Format (PDF).

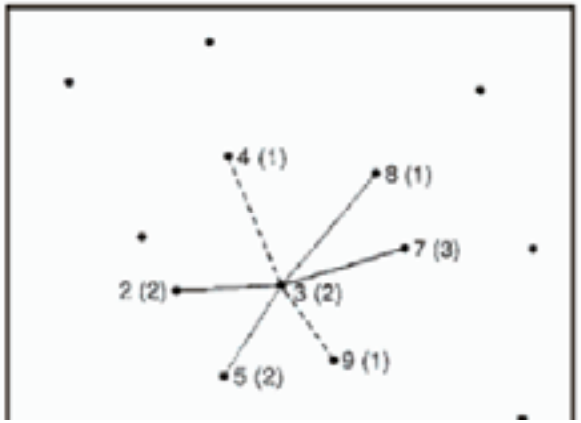
Head-Bang PC Software

"Head-banging" is a weighted two-dimensional median-based smoothing algorithm, developed to reveal underlying geographic patterns in data where the values to be smoothed do not have equal variances. The original idea was proposed by Tukey and Tukey (1981), then studied and implemented by Hansen (1991). The staff at the National Center for Health Statistics worked with Hansen to add weights to the original algorithm (Mungiole, Pickle, Simonson, 1999).

The process is implemented as follows. We use population-weighted county values for illustration, but it can be applied to any geographic units with any user-specified weights.

For each county, up to NN (default=30) neighboring counties comprise its "smoothing window", which is of sufficient size to begin to show regional patterns while retaining patterns apparent in the raw data maps.

Three nearly collinear triples are used for smoothing at the center point of this figure. Observed values are printed at each data location, with corresponding weights given in parentheses. The low screen is the weighted median of the low values (2, 4, 5)=4. The high screen is the weighted median of the high values (7, 8, 9)=8. The observed value at the center point lies below the low screen, so the weights are used to determine whether adjustment is needed. Because the sum of endpoint weights (10) exceeds the number of triples times the center point weight (3x2=6), the smoothed value at the center point is set to the low screen.



3. Headbang

- Nonparametric algorithm using population weights to smooth away some of underlying random variation,

3. Headbang

- Nonparametric algorithm using population weights to smooth away some of underlying random variation,
- **While retaining <u>broad patterns in data</u>**

3. Headbang

- Nonparametric algorithm using population weights to smooth away some of underlying random variation,
- While retaining broad patterns in data

- 2-D version of median-based moving average (~ time-series)

- Refs available on http://srab.cancer.gov/headbang/

- Geographic smoothing algorithms "borrow information" from neighboring areas to stabilize results from sparsely populated areas.
- 1. reduces variability in the data, allowing patterns to emerge
- 2. But… but increases the bias in the estimates for each small area.

Recall…

- Purpose of smoothed maps is to show within state variation of predicted rates

- Not to identify predicted rate for a given county

Why? Original prediction may have been changed by smoothing algorithm– county rate looks more like its neighbor's rate.

5. Results

Disparities is not just an issue of race: Age-profile maps

Highlights

- Preliminary Results
  - Current geographic patterns of mortality are similar to previous patterns
  - Age profiles vary geographically;
  - Screening rates are low in some areas of high mortality but not all
  - Racial / ethnic rate differences

Summary

- Discussed approaches to analyzing partial US data to establish a baseline of cervical cancer incidence
- Spatial temporal model
- HeadBang
- Geographic patterns by Race
- Geographic patterns in age profiles

Questions

Contact:
hornerm@mail.nih.gov
National Cancer Institute