



**COMMUNITY SURVEILLANCE
MANUSCRIPT DEVELOPMENT
WORKSHOP SERIES**

Session Two:
Introduction to Surveillance Data Analytic Issues
March 6, 2012

Surveillance Workshop Series

- Session One
 - Background and manuscript development
 - Webinar December 14
- Session Two
 - Analytic issues with surveillance data
 - Webinar March 6
- Session Three
 - Refining proposals and starting analysis
 - Conference call or in-person sessions at CC

Objectives for Today

- Describe ARIC community surveillance sampling design and discuss implications for analysis
- Describe ARIC surveillance data structures and available data
- Describe the process for initiating analysis

ARIC Community Surveillance

- Retrospective continuous monitoring and validation of hospitalized acute MI, CHD death and heart failure
- Goal is to evaluate trends in mortality, incidence, case fatality and medical care
- Sample CHD hospitalizations and deaths in four communities

Sampling Frame for CHD

- Fatal CHD and hospitalized MI events identified from death certificates and hospital discharge indexes
- Eligibility for sampling based on age, race, place of residence, date of discharge or death, discharge diagnosis code
- Sampled for abstraction and validation of the event

Sampling Frame for HF

- Hospitalized HF events identified from hospital discharge indexes
- Eligibility for sampling based on age, race, place of residence, date of discharge, discharge diagnosis code
- Sampled for abstraction and validation of the event

Sampling Design

- Stratified random sample
- Stratification by field center, gender, race, age, discharge diagnosis
- Sampling probabilities vary by strata and year
 - Can vary within strata
- Sampling strategies evolved over time
- Described in Manual 3 (CHD) and 3a (HF)

Accounting for the Design in Analysis

- Sampling Weights
 - Calculated for each event
 - Relevant for all events, regardless of type
 - Inverse of the probability of being sampled
 - Variable names
 - SAMWT (CHD)
 - SAMWTHF (HF)
- Stratification
 - Collapsed to avoid strata of size one
 - NESTVAR2, NESTVAR_OLD, NESTVAR_COMBN (CHD)
 - SAMSTRAT3 (HF)

Reducing Bias in Population Estimates

- Estimates of population parameters (means, counts, etc.) must account for differential sampling probabilities
- Incorporate sampling weights into all analysis
- Standard statistical software (SAS, Stata, SUDAAN) will do this
 - Include variable with sampling weights
 - Adequate for estimates but not necessarily for standard errors

Sampling Weights: A Simple Example

- Population 3000
 - 90% Whites
 - 10% African-Americans
- Sample 100 Whites and 100 African-Americans
- Mean SBP:
 - Whites = 130 mmHg
 - Afr-Am = 150 mmHg
- Simple mean = 140 mmHg is clearly incorrect
- Apply sampling weights (inverse of the sampling probability) to each sampled individual to calculate weighted mean
 - $(100/2700)^{-1} = 27$ for Whites
 - $(100/300)^{-1} = 3$ for African-Americans
- Equivalent to weighted average = 143 mmHg

Standard Error Estimation

- Standard error (se) is the standard deviation of the sampling distribution
 - Reflects variation expected based on repeated sampling of the population
 - More independent data => lower standard error
- Used to construct confidence intervals and tests of significance
- Calculation of se must account for design
 - Ignoring stratification will produce se estimates that do not apply to the population sampled

Standard Error: A Simple Example

- Return to previous example (sampling 100 Whites and 100 African-Americans)
- Standard error applies to all the possible samples from the sampling design
 - Simple random sample of 200 could result in many combinations of W/A:
 - 200/0, 150/50, 100/100, etc.
 - Stratified sample will always be 100/100
- Impacts standard error estimation

Standard Error (cont.)

- Weighted analysis is not enough!
- Software must be “survey-aware”
- SUDAAN
- SAS
 - SURVEYFREQ, SURVEYLOGISTIC, SURVEYMEANS, SURVEYPHREG, and SURVEYREG procedures
 - STRATA statement and WEIGHT statement
- Stata
 - SVY prefix

Standard Error Estimation Methods

- Taylor linearization
 - Usually the default
 - Works for “smooth” statistics (means, regression parameters, ratios)
 - Breaks down with stratum size of one
- Resampling methods
 - Bootstrap, jackknife, BRR
 - May be required for non-smooth statistics, such as percentile estimates (including median)
 - Computationally intense

Domain (Subpopulation) Analysis

- Special attention should be given to subpopulation analysis
 - Accounting for sampling weights and design is inadequate
 - Software must specifically account for subpopulation analysis
- Failure to do so will result in standard errors that do not apply to the population sampled

Domain Analysis: A Simple Example

- Continue prior example
- Interest focuses on mean SBP in smokers only
- Smoking status is known only in the sample
 - Smoking status not accounted for in design
 - Sampling fractions are unknown for White smokers and African-American smokers
- Repeated samples of 100 Whites / 100 African-Americans would have varying numbers of White smokers and African-American smokers
- Estimation of standard error must account for this extra variability

Domain Analysis (cont.)

- Statistical packages mentioned so far will calculate standard errors appropriately, if specified
 - SAS (recent versions): DOMAIN statement
 - SUDAAN: SUBPOP statement
 - Stata: SUBPOP and OVER options
- DO NOT subset out subpopulation of interest prior to analysis

Case Fatality Analysis

- National Death Index data for deaths following an event
- NDI data lags one year behind ARIC
- Same analytical issues apply
 - Sample weights
 - Domain analysis on deaths

ARIC Population Estimates

- Required for calculation of event rates
- Based on US Census for ARIC catchment areas
 - 1990, 2000, 2010
 - Years between are interpolated / extrapolated
- For analysis purposes, assumed to be non-random
- Updating population estimates for 2010
- Contact CC for population estimates

- Finding variables we mentioned
 - <http://www.csc.unc.edu/aric/>
- Data structure (events vs occurrences, event ids, etc)
- Manuscript proposals, analysis requests, queue, support for non-CC analysis

Next Steps

- Submit a manuscript proposal
 - ARIC website for procedures and form
Home -> Publications -> Policies & Forms
 - Contact Wayne, Chris and Lisa for help
- Who will do the analysis?
 - Coordinating Center
 - Investigator
- High priority papers call