# THE PUBLIC HEALTH DISPARITIES GEOCODING PROJECT

geocoding, census data & social disparities in health

Harvard TH Chan School of Public Health

N Krieger, JT Chen, PD Waterman, DH Rehkopf, SV Subramanian, L Haughton

# CASE EXAMPLE & ANALYTIC METHODS

Analysis of all cause mortality rates in Suffolk County, Massachusetts, 1989-1991, by CT poverty

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

Our primary analytic approach for describing socioeconomic gradients by area-based socioeconomic measures has been to use geocodes to append area-based socioeconomic data to case records, to stratify these records into discrete categories based on ABSM, and to aggregate numerators and denominators over areas, within levels defined by ABSM. This method avoids the problem of unstable rates arising from small areas by assuming that cases and population denominators from areas with similar socioeconomic characteristics can be legitimately combined into the same strata. An alternative approach, which preserves the spatial information of the geocodes, is discussed in the section on multilevel analyses (see website).

The following steps are used to generate age-standardized disease rates stratified by area-based socioeconomic measures, once the case data have been geocoded and appropriate ABSMs have been generated from census data.

- Aggregate the case data into numerators (age cells within areas/geocodes).
- Aggregate population denominator data into age cells within areas/geocodes.
- Merge the numerators and denominators with ABSMs, by area/geocode.
- Aggregate over areas into strata defined by categorical ABSM and age category.
- Generate age-standardized rates and other summary measures.

We've created this case example as an opportunity for you to try out our methods. The example draws on all cause mortality data from Suffolk County, Massachusetts, between 1989 and 1991. You'll have a chance to analyze these data by census tract poverty to see the socioeconomic gradient in mortality in this county. We've divided the case example into clearly defined tasks to highlight the process of moving from raw data to summary measures of the socioeconomic disparity.

The raw data files and SAS program can be downloaded from our website: www.hsph.harvard.edu/thegeocodingproject/.

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

## STEP 1. Aggregate the numerator

### ANALYTIC METHODS

Data from public health databases are typically formatted such that each record represents one person (or case report). Once these data have been geocoded, they need to be aggregated before linking to denominator and ABSM data. Before aggregating, however, one should exclude all records that are not geocoded, do not meet the case definition, or are missing data on the important covariates (e.g. age, in the case of simple age-standardized analyses; age, sex, and race/ethnicity in the case of more complex stratified analyses).

One can think of the basic unit of aggregation as a cell, defined by age and other covariates, within an area/geocode. Once aggregated, this cell within an area can be linked to a relevant population denominator. The cell contains a count of all cases within that area that meet the specified age and other covariate criteria. Since our goal is eventually to create rates, we call this count of cases the “numerator.”

### CASE EXAMPLE

The rawcase.csv file contains all deaths occurring in Suffolk County, Massachusetts, between 1989 and 1992. Each person who died is represented by one line in the data file. The variable “AGE” gives the age at death. The variable “AREAKEY” is the geocode to the census tract level.

### SAS PROGRAMMING

```
PROC IMPORT OUT= rawcase
DATAFILE= "G:\monograph\example\rawcase.csv"
DBMS=CSV REPLACE;
GETNAMES=YES;
DATAROW=2;
RUN;

DATA Step1a ;
SET rawcase ;

IF 0<=AGE<15 THEN AGECAT=1 ;
IF 15<=AGE<25 THEN AGECAT=2 ;
IF 25<=AGE<45 THEN AGECAT=3 ;
IF 45<=AGE<65 THEN AGECAT=4 ;
IF AGE>=65 THEN AGECAT=5 ;
RUN ;

PROC FREQ DATA=Step1a NOPRINT ;
TABLES AREAKEY*AGECAT /OUT=Step1b ;
RUN ;
```

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

## STEP 2. Aggregate the denominator data.

### ANALYTIC METHODS

Denominator data at the census tract level typically come from the decennial census, which gives population counts in 31 age categories (<1, 1-2, 3-4, 5, 6, 7-9, 10-11, 12-13, 14, 15, 16, 17, 18, 19, 20, 21, 22-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-61, 62-64, 65-69, 70-74, 75-79, 80-84, 85+). For the purposes of age standardization, these age categories need to be re-aggregated to match the age categories used for categorizing case data (numerators, above) and the age categories from the standard million reference population. Additionally, when using case data from multiple years, in order to calculate an average annual incidence rate, one needs to use a person-time denominator (population count multiplied by number of years of case data).

### CASE EXAMPLE

The rawdenom.csv file contains the estimated population count in 31 age categories for the 189 census tracts in Suffolk County, from the 1990 U.S. Census. Each census tract is represented by one line in the data file, with the 31 age categories arrayed horizontally.

a. Aggregate the population counts into the five broad age categories listed above.

b. Transpose the structure of the data, so that is one record for each age stratum within a census tract, with a corresponding categorical age variable and population count. You should end up with 5 records for each census tract, with each record represented by one line of your output dataset.

c. Multiply the population count by 3, to yield a person-time denominator for three years worth of death data.

### SAS PROGRAMMING

```
PROC IMPORT OUT= rawdenom
DATAFILE= "G:\monograph\example\rawdenom.csv"
DBMS=CSV REPLACE;
GETNAMES=YES;
DATAROW=2;
RUN;

DATA Step2a ;
SET rawdenom ;

AGECAT1= SUM(OF P0130001-P0130009) ;
AGECAT2= SUM(OF P0130010-P0130017) ;
AGECAT3= SUM(OF P0130018-P0130021) ;
AGECAT4= SUM(OF P0130022-P0130026) ;
AGECAT5= SUM(OF P0130027-P0130031) ;

LENGTH AGECAT 3 ;
```

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

```
ARRAY AGES [5] AGECAT1-AGECAT5 ;

DO I=1 TO 5 ;
AGECAT=I ;
DENOM=3*AGES[I] ;
OUTPUT ;
END ;

DROP I AGECAT1-AGECAT5 P0130001-P0130031 ;
RUN ;
```

## STEP 3. Merge the numerators and denominators by AGECAT and AREAKEY.

### ANALYTIC METHODS

Once the numerators and denominators have the same structure (AREAKEY x AGECAT), they can be merged together, along with the ABSM data (by AREAKEY). For age cells within areas where no cases were reported, we set the numerator to zero. This type of merging is an example of a SAS matched merge.

### CASE EXAMPLE

For age cells in census tracts where no cases were reported, set the numerator to zero.

### SAS PROGRAMMING

```
PROC SORT DATA=Step1b ;
BY AREAKEY AGECAT ;
RUN ;

PROC SORT DATA=Step2a ;
BY AREAKEY AGECAT ;
RUN ;

DATA Step3 ;
MERGE Step1b (KEEP=AREAKEY AGECAT COUNT) Step2a ;
BY AREAKEY AGECAT ;

NUMER=COUNT ;
IF DENOM>0 THEN LOGDEN=LOG(DENOM) ;
ELSE LOGDEN=. ;
IF NUMER=. AND DENOM=>0 THEN NUMER=0 ;
IF DENOM>=0 THEN OUTPUT ;

DROP COUNT ;
```

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

```
RUN ;
```

STEP 4. Now merge the combined numerator and denominator data from Step 3 with the ABSM data, by AREAKEY.

### ANALYTIC METHODS

Next, in order to generate rates for categories of a specific ABSM, it is necessary to aggregate OVER areas into strata defined by AGECAT and ABSM. Numerators and denominators from census tracts with missing ABSM data for a particular ABSM are typically excluded from that analysis.

### CASE EXAMPLE

The file rawabsm.csv containS the 189 census tracts in Suffolk County, and the % of persons living below poverty for each tract, categorized into 4 categories (1=0-4.9%, 2=5-9.9%, 3=10-19.9%, 4=20-100%).

### SAS PROGRAMMING

```
PROC IMPORT OUT= rawabsm
DATAFILE= "G:\monograph\example\rawabsm.csv"
DBMS=CSV REPLACE;
GETNAMES=YES;
DATAROW=2;
RUN;

PROC SORT DATA=rawabsm ;
BY AREAKEY ;
RUN ;

DATA Step4 ;
MERGE Step3 rawabsm ;
BY AREAKEY ;
RUN ;
```

STEP 5. For each category of CT poverty, calculate the age-standardized incidence rate, using the year 2000 standard million.

### ANALYTIC METHODS

*1. Age-standardized incidence rates*

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

The standard practice of public health departments in reporting population rates of mortality and disease incidence is to calculate age-standardized rates, which facilitates comparisons between regions or subgroups of interest. The age-standardized rate is interpretable as the rate that would be observed in a population if that population had the same age distribution as a given reference population. Direct standardized rates are obtained by applying the age-specific incidence rates observed in the area or subgroup of interest to a standard age distribution, such as the year 2000 standard million.[1]

For our project, we used five broad age categories to age standardize, in order to obtain more stable rates in each age stratum, particularly for outcomes with sparse data.

If $\text{cases}_j$ represents the number of cases in age group *j* of the group or region of interest and $pop_j$ represents the population associated with that age group, then the standardized rate $IR_{st}$ for the group or region is

$$IR_{st} = \frac{\sum_j w_j \left(\frac{cases_j}{pop_j}\right)}{\sum_j w_j} = \frac{\sum_j w_j IR_j}{\sum_j w_j}$$

where $w_j$ is the weight associated with category *j* in the reference (standardizing) population (e.g. the population size or the proportion of the total population). The estimated variance of the standardized rate is given by:

$$\text{Var}(IR_{st}) = \frac{\sum_j w_j^2 \left(\frac{cases_j}{pop_j^2}\right)}{\left(\sum_j w_j\right)}$$

(When the $w_j$s are proportions, then

$$IR_{st} = \sum_j w_j\, IR_j$$

and

$$\text{Var}(IR_{st}) = \sum_j w_j^2 \left(\frac{cases_j}{pop_j^2}\right)$$

.)

*2. Confidence intervals for directly standardized rates*

Traditional confidence limits for the direct standardized rates are based on the normal distribution and require large cell counts. In our analyses, we found that they can also occasionally result in “impossible” lower limits that are less than zero. Because of this, we adopted an alternate method for calculating the confidence limits based on the inverse gamma function.[2] This method assumes that the direct standardized rate is a linear combination of independent Poisson random variables. Assuming that this linear combination is also follows a Poisson distribution, the age-standardized rate $E(X) = x$ follows a gamma distribution $\Gamma(a,b)$ as follows:

$$X \sim \Gamma\left(\frac{x^2}{v}, \frac{v}{x}\right)$$

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

where *x* is the age-standardized rate ($IR_{st}$ as estimated above) and v is its variance, as described above. Converting this to the gamma distribution in its standard form, i.e. where b=1, this yields

$$\frac{X}{b} \sim \Gamma\left(\frac{x^2}{v}, 1\right)$$

which greatly simplifies calculations. Then the lower 100(1-α) confidence limit for $\frac{x^2}{v}$ is given by

$$L\left(\frac{x^2}{v}\right) = \Gamma^{-1}\left(\frac{x^2}{v}, 1\right)\left(\frac{\alpha}{2}\right)$$

and the upper 100(1-α) confidence limit for $\frac{x^2}{v}$ is given by

$$U\left(\frac{x^2}{v}\right) = \Gamma^{-1}\left(\frac{(x+k_M)^2}{(v+k_M)}, 1\right)\left(1-\frac{\alpha}{2}\right)$$

where

$$k = k_M = max_{je(1,\dots,j)}(k_j)$$

is a continuity correction necessitated by using a continuous distribution to estimate confidence limits for a discrete random variable. Increasing the number of events by 1 in an age stratum i results in a

$$k_j = \frac{w_j}{pop_j}$$

increase in the age-standardized rate. If $k_j$ is constant for all age intervals, then $k_j = k$. However, since the values for $w_j$ and $pop_j$ typically vary across age strata, it is unclear what value of *k* to use. A very conservative upper limit can be obtained by using the maximum value of $k_j = k_m$.

However, following the recommendation of the NCHS, we used a close approximation that alleviates the need to calculate $k_m$:

To transform these intervals to obtain the desired confidence limits for *X*, we use

$$L(X) = \frac{L\left(\frac{x^2}{v}\right)}{\frac{x}{v}} \quad \text{and} \quad U(X) = \frac{U\left(\frac{x^2}{v}\right)}{\frac{x}{v}} \quad .$$

## CASE EXAMPLE

In order to do this:

a. Aggregate the numerator and denominator within each age X CT poverty stratum, across all census tracts.

b. Exclude cases and denominator where CT poverty is missing.

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

c. Merge with the year 2000 standard million in five age categories.

d. Calculate the age-standardized incidence rate standardized to the year 2000 standard million, and the corresponding “gamma” confidence intervals for the direct standardized rates.

## SAS PROGRAMMING

```
CREATE DATASET WITH STANDARD MILLION FOR AGE STANDARDIZATION (IN FIVE CATEGORIES)
0-14
15-24
25-44
45-64
65+

*************************************************************;

data stdrd ;
input agecat y1940 y1970 y1980 y1990 y2000 ;

cards ;
1 250416 284926 226401 215383 214700
2 181677 174405 187542 147860 138646
3 301303 236183 276838 324695 298186
4 198105 205746 196440 186446 222081
5 68499 98740 112779 125616 126387
;

RUN ;

PROC SORT DATA=Step4 ;
BY AGECAT CINDPOV ;
run ;

DATA Step5a ;
SET Step4 ;
WHERE CINDPOV^=. ;
BY AGECAT CINDPOV ;

retain NMR DNM ;

if first.CINDPOV then do ;

NMR=0 ;
DNM=0 ;
end ;

NMR=NMR+NUMER ;
DNM=DNM+DENOM ;
```

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

```
if last.CINDPOV then DO ;
output ;
END ;

DROP AREAKEY NUMER DENOM ;
RUN ;


DATA Step5b ;
MERGE Step5a (in=ina) stdrd (in=inb) ;
BY AGECAT ;
if ina and inb ;

w_i=y2000/1000000 ;

IR_i=NMR/DNM ;

varpy_i=NMR/DNM**2 ;

RUN ;

proc sort data=Step5b ;
BY CINDPOV ;
run ;

data Step5c ;
SET Step5b ;
by CINDPOV ;

*************************************

IRW=weighted incidence rate
VARPY=part of person-time variance
VARPYW=weighted person-time variance
SUMWI=sum of weights
CRDEN=crude denominator
WMAX=maximum weight (for gamma CI)

*************************************;

retain IRW VARPY VARPYW SUMWI WMAX CRNUM CRDEN ;

if first.CINDPOV then do ;

IRW=0 ;
VARPYW=0 ;
VARPY=0 ;
SUMWI=0 ;
CRNUM=0 ;
CRDEN=0;
WMAX=0 ;
end ;
```

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

```
IRW=IRW + (W_I*IR_I) ;
VARPY=VARPY + ((W_I**2)*VARPY_I) ;
SUMWI=SUMWI + W_I ;
CRNUM=CRNUM+NMR ;
CRDEN=CRDEN+DNM ;
WMAX=MAX(WMAX,W_I/DNM) ;


if last.CINDPOV then do ;

VARPYW=VARPY/(SUMWI**2) ;

********************************

LOWER 95% GAMMA INTERVAL

LGAM gives the 95% gamma interval using the formula given by Fay and Feuer.
LGAM2 gives the 95% gamma interval using the formula given Anderson and Rosenberg.

***NOTE: FOR LGAM2 AND UGAM2, HAVE NOW PROGRAMMED OPTIONS FOR IRW=0 VARPYW=0 I.E. USE INVERSE
CHI SQUARE DISTRIBUTION CINV(0.975,2) AND DIVIDE BY DENOMINATOR TO GET UPPER LIMIT ON RATE

References:
Fay MP, Feuer EJ. Confidence intervals for directly standardized rates: a method based on the gamma
distribution. Statistics in Medicine 1997,16:791-801.

Anderson RN, Rosenberg HM. Age standardization of death rates: implementation of the year 2000 standard.
National Vital Statistics Reports: Vol 37, No. 3. Hyattsville, MD:
National Center for Health Statistics, 1998.

********************************;

LGAM=(VARPYW/(2*IRW)) * CINV(0.025,((2*(IRW**2))/VARPYW)) ;

IF IRW=0 AND VARPYW=0 THEN DO ;
LGAM2=0 ;
END ;
ELSE LGAM2=(VARPYW/IRW) * GAMINV(0.025,((IRW**2)/VARPYW)) ;

********************************

UPPER 95% GAMMA

UGAM gives the 95% gamma interval using the formula given by Fay and Feuer.
UGAM2 gives the 95% gamma interval using the formula given Anderson and Rosenberg.

********************************;

UGAM=((VARPYW + (WMAX**2))/2*(IRW+WMAX)) * CINV(0.975,((2*((IRW + WMAX)**2))/(VARPYW +
(WMAX**2)))) ;

IF IRW=0 AND VARPYW=0 THEN DO ;
UGAM2=(0.5 * CINV(0.975,2))/CRDEN ;
```

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

```
END ;
ELSE UGAM2=(VARPYW/IRW) * GAMINV(0.975,(((IRW**2)/VARPYW) + 1)) ;

********************************

REGULAR CONFIDENCE LIMITS

******************************** ;

LO95 = IRW - (1.96*SQRT(VARPYW)) ;
HI95 = IRW + (1.96*SQRT(VARPYW)) ;

OUTPUT ;
end ;

proc print ;
var CINDPOV IRW LGAM2 UGAM2 ;
run ;
```

## STEP 6. Estimate the age-standardized incidence rate ratio comparing the age standardized rates in each poverty stratum to the rate in the least impoverished poverty stratum (0-4.9%).

### ANALYTIC METHODS

*3. Confidence intervals for $IR_{st}$=0*

When the observed rate is zero (i.e. there were zero cases), the gamma method is unable to produce confidence limits for the direct standardized rates. In this situation, we adopt the following convention for the confidence limit. The lower limit is simply set to zero. For the upper limit, we assume that the number of cases (i.e. the count) follows a Poisson distribution, and use the formula for the "exact" upper confidence limit of a Poisson random variable[3]:

$$U(Y) = \frac{1}{2}\chi^{-1}_{2(y+1)df}\left(1-\frac{\alpha}{2}\right)$$

where y is the count, i.e. zero. When $\alpha$ =0.05 (i.e. for a 95% confidence limit) this simplifies to

$$U(Y) = \frac{\chi^{-1}_{2df}\left(1-\frac{\alpha}{2}\right)}{2} = 3.689\ .$$

We can then divide this upper limit on the count by the population denominator to give the upper limit on the rate.

*4. Age-standardized incidence rate ratio*

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

Two commonly used measures for comparing incidence rates from two different groups are the incidence rate difference (IRD) and the incidence rate ratio (IRR). The incidence rate difference compares the rates on the absolute scale, and summarizes the excess rate comparing the larger to the smaller rate. The incidence rate ratio compares the rates on a relative scale, summarizing the size of one rate relative to the other rate.

To compare two age-standardized incidence rates on the absolute scale, the age-standardized incidence rate difference ($IRD_{st}$) is the rate in one group minus the rate in the other, i.e. $IR_{st1}$ - $IR_{st0}$. The variance of this age-standardized incidence rate difference is simply the sum of the estimated variance of the two age-standardized rates[4],

$$Var(IRD) = Var(IR_{st1}) + Var(IR_{st0})$$

To compare age-standardized rates from two different groups or regions on the relative scale, the age-standardized incidence rate ratio ($IRR_{st}$) is simply $IR_{st1}/IR_{st0}$. Confidence intervals can be calculated using the variance estimator[4]:

$$Var[log(IRR_{st})] = \frac{Var(IR_{st1})}{IR_{st1}^2} + \frac{Var(IR_{st0})}{IR_{st0}^2}$$

## CASE EXAMPLE

Calculate the 95% confidence limits on the incidence rate ratio.

## SAS PROGRAMMING

```
DATA Step6 ;
SET Step5c ;

RETAIN IRREF VARPYREF ;

IF _N_=1 THEN DO ;
IRREF=IRW ;
VARPYREF=VARPYW ;
END ;

*************************************
Incidence rate difference
*************************************;

IRD=IRW - IRREF ;
VARIRD=VARPYW + VARPYREF ;

L_IRD=IRD - (1.96 * SQRT(VARIRD)) ;
U_IRD=IRD + (1.96 * SQRT(VARIRD)) ;

*************************************
Incidence rate ratio
*************************************;
```

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

```
IRR=IRW/IRREF ;
VARIRR=(VARPYW/(IRW**2)) + (VARPYREF/(IRREF**2)) ;

L_IRR=EXP(LOG(IRR) - (1.96 * SQRT(VARIRR))) ;
U_IRR=EXP(LOG(IRR) + (1.96 * SQRT(VARIRR))) ;

RUN ;

proc print ;
var Cindpov IRD L_IRD U_IRD IRR L_IRR U_IRR ;
RUN ;
```

## STEP 7. Estimate the relative index of inequality (RII) for CT level poverty in relation to all cause mortality.

### ANALYTIC METHODS

*5. Relative Index of Inequality (RII)*

Comparisons of socioeconomic gradients based on categorical ABSM may be complicated by differences in the population distributions of area-based socioeconomic measures. For example, it may be expected that the classifications producing smaller groups at the margins would lead to larger incidence rate ratios, comparing the most deprived to the most affluent, because finer discrimination of extremes of socioeconomic position is achieved. The relative index of inequality (RII) has been proposed as a measure which explicitly addresses this problem.[5-7] Assuming ordinality of the ABSM categories, the RII is calculated by regressing the incidence rate in each ABSM category on the total proportion of the population that is more deprived in the socioeconomic hierarchy.

In practice, this latter quantity is represented by the cumulative distribution function (cdf). We approximate the cdf for the jth level of a given ABSM by summing the proportion of the population represented by the categories $ABSM_1$, ..., $ABSM_{j-1}$, and adding one-half the proportion of the population represented by the category $ABSM_j$.

In order to compare RII meaningfully across groups with differing age composition, we developed an age-standardized RII, standardized to the year 2000 standard million, as follows. Let $observed_{ij}$ be the observed number of cases in the *ith* age group and the *jth* category of ABSM, and $pop_{ij}$ be the population at risk in the corresponding category. First, we calculate the age-standardized rate $IR_{st}$ in each stratum *j* defined by ABSM, as described above. For each stratum j, we estimate the expected number of cases in stratum j, $expected_j$, by multiplying the age-standardized rate $IR_{st}$ by the population denominator,

$$pop_j = \sum_i pop_{ij}$$

We determine the “marginal” cumulative distribution function, $cdf(ABSM_j)$, of the ABSM over the entire population, as noted above.

To calculate the age-standardized $RII_{st}$, we fit the following Poisson model for the expected cases:

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

$$expected_{ij} \sim Poisson(\lambda_{ij})$$

$$log(\lambda_{ij}) = log(pop_{ij}) + \beta_0 + \beta_1{*}cdf(ABSM_j)$$

Exponentiation of the ß1 yields the RII, which is interpretable as an incidence rate ratio comparing the rates in the bottom to the top of the socioeconomic hierarchy. A larger RII indicates a greater the degree of inequality across a socioeconomic hierarchy, which may be due to a steep socioeconomic gradient or large inequalities in the distribution of the ABSM itself.

## CASE EXAMPLE

a. Estimate the approximate cumulative distribution function for CT poverty, based on the population denominator for each poverty stratum (summed up over age).

b. Calculate the expected cases in each CT poverty stratum, based on the age-standardized incidence rate.

c. Fit a Poisson log linear model, modeling the expected number of cases as a function of the approximate cumulative distribution of CT poverty, using the population denominator as an offset.

d. Exponentiate the beta term from this model to get the relative index of inequality.

## SAS PROGRAMMING

```
data Step7a Step7b ;
set Step5c END=LASTOBS;
by CINDPOV ;

retain dxden ;

if _N_=1 then do ;
dxden=0 ;
end ;

dxnum=dxden + (CRDEN/2) ;
dxden=dxden+CRDEN ;

DUMMY=1 ;
output Step7a ;
IF LASTOBS then output Step7b ;

data Step7c ;
merge Step7a (drop=dxden) Step7b (keep=DUMMY dxden) ;
BY DUMMY ;


dxpct=dxnum/dxden ;
wght=CRDEN/dxden ;
```

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

```
CRCNT=CRDEN*IRW ;
LOGDEN=LOG(CRDEN) ;

ods output ParameterEstimates=param ;
PROC GENMOD DATA=Step7c ;
MODEL CRCNT = DXPCT /OFFSET=LOGDEN LINK=LOG DIST=POI ;
RUN ;

data Step7d;
set param (where=(Parameter not in ("Intercept")));

if stderr ne 0 then do;
riiest=exp(estimate);
riilo95=exp(estimate-1.96*stderr);
riihi95=exp(estimate+1.96*stderr);
end;

proc print ;
where Parameter="dxpct" ;
var Parameter riiest riilo95 riihi95 ;
run ;
```

## STEP 8. Calculate the population attributable fraction of all cause mortality due to CT poverty.

### ANALYTIC METHODS

To calculate the age-standardized RI$_{Ist}$, we fit the following Poisson model for the expected cases:

$$expected_{ij} \sim Poisson(\lambda_{ij})$$

$$log(\lambda_{ij}) = log(pop_{ij}) + \beta_0 + \beta_1 * cdf(ABSM_j)$$

Exponentiation of the $\beta_1$ yields the RII, which is interpretable as an incidence rate ratio comparing the rates in the bottom to the top of the socioeconomic hierarchy. A larger RII indicates a greater the degree of inequality across a socioeconomic hierarchy, which may be due to a steep socioeconomic gradient or large inequalities in the distribution of the ABSM itself.

*6. Population Attributable Fraction*

The population attributable fraction (PAF) is a useful summary measure for characterizing the public health impact of an exposure on population patterns of health and disease. It is defined as “the fraction of all cases (exposed and unexposed) that would not have occurred if exposure had not occurred.”[8] For a polytymous exposure, the population attributable fraction is a weighted sum of the attributable fractions for each level of the exposure, with the weights defined by the case fractions (number of exposed cases divided by overall number of cases):

$$PAF = CF_1\ x\ \frac{RR_1 - 1}{RR_1} + CF_2\ x\ \frac{RR_2 - 1}{RR_2} + \cdots + CF_j\ x\ \frac{RR_j - 1}{RR_j}$$

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

In order to aggregate multiple PAFs over several age strata i=1,…,I, note that

$$PAF_{agg} = \frac{\sum_i excess\ number\ of\ cases}{\sum_i number\ of\ cases}$$

$$= \frac{\sum_i number\ of\ cases\ x\ \dfrac{excess\ number\ of\ cases}{number\ of\ cases}}{\sum_1 number\ of\ cases}$$

$$= \frac{\sum_i number\ of\ cases\ x\ PAF_i}{\sum_i number\ of\ cases}$$

that is, a weighted average of stratum specific PAFs, with the number of cases in each age stratum as weights.

## CASE EXAMPLE

a. Starting with the data from Step 4, sum up over AREAKEY into strata defined by AGECAT and CT poverty.

b. Calculate (i) the total cases in each age stratum, over poverty; and (ii) the rate in the reference group of CT poverty.

c. Calculate stratum specific rates, rate ratios, and case fractions.

d. Calculate the age-stratum-specific population attributable fractions.

e. Calculate the grand total of all cases to use in calculating weights for all age strata.

f. Finally, calculate the aggregated population attributable fraction, using the age specific weights based on proportion of cases in each age stratum.

## SAS PROGRAMMING

Using the age-stratified numerators and denominators from Step 4, calculate the age-stratum specific population attributable risk fractions and aggregate population attributable risk fraction over all age strata following the method of Hanley8.

SOME NOTATION:

Assume that the dataset provided has stratum specific numbers of cases (NUMER) and denominator (DENOM).

Subscript i as age, j as covariate (in this case, CINDPOV)

1. Sort data by AGECAT and CINDPOV.

2. Calculate the quantities

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

```
NUMERi+ = SUMj(NUMERij)
RATEiREF = rate in the reference group and save them in a dataset.

3. Merge the quantities from (2) with the dataset and calculate

(a) rate RATEij=NUMERij/DENOMij
(b) case fractions CFij = NUMERij/NUMERi+
(c) rate ratio RRij = RATEij/RATEiREF

4. Calculate the (age)stratum-specific population attributable risk fraction

AFPi = CFi1 * (RRi1-1)/RRi1 + CFi2 * (RRi2-1)/RRi2 + ... CFij * (RRij-1)/RRij

5. Calculate the grand total of cases NUMER++ to use to calculate age-specific weights.

6. Determine age-stratum specific weights from the case distribution:

w_i = NUMERi+/NUMER++

and calculate the aggregated AFPagg = SUM(w_i * AFPi)

************************************************************;

**********************************

STEP 8a SUM OVER AREAKEY INTO STRATA BY AGECAT AND CINDPOV

**********************************;

PROC SORT DATA=Step4 ;
BY AGECAT CINDPOV AREAKEY ;
RUN ;

DATA Step8a ;
SET Step4 (WHERE=(CINDPOV^=.)) ;
BY AGECAT CINDPOV ;

RETAIN NNN DDD ;

IF FIRST.CINDPOV THEN DO ;
NNN=0 ;
DDD=0 ;
END ;

NNN=NNN+NUMER ;
DDD=DDD+DENOM ;

IF LAST.CINDPOV THEN DO ;
OUTPUT ;
END ;

KEEP AGECAT CINDPOV NNN DDD ;

RUN ;
```

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

```
**********************************

STEP 8b: CALCULATE THE QUANTITIES

NUMERI+
RATEIREF

**********************************;

DATA Step8b ;
SET Step8a ;
BY AGECAT ;

RETAIN NIPLUS RATEIREF ;

IF FIRST.AGECAT THEN DO ;
NIPLUS=0 ;
END ;

NIPLUS=NIPLUS + NNN ;

IF CINDPOV=1 THEN DO ;
RATEIREF=NNN/DDD ;
END ;

IF LAST.AGECAT THEN DO ;
OUTPUT ;
END ;

KEEP AGECAT NIPLUS RATEIREF;

**********************************

STEP 8c: CALCULATE STRATUM SPECIFIC RATES, RATE RATIOS, AND CASE FRACTIONS
(a) rate RATEij=NUMERij/DENOMij
(b) case fractions CFij = NUMERij/NUMERi+
(c) rate ratio RRij = RATEij/RATEiREF

**********************************;

DATA Step8c ;
MERGE Step8a Step8b ;
BY AGECAT ;

RATEij=NNN/DDD ;

*******************************

NOTE: IF NIPLUS=0 THEN THERE ARE NO CASES IN THIS AGE STRATUM AT ALL SO SET CASE FRACTION TO ZERO

*******************************;

IF NIPLUS=0 THEN CFij=0 ;
ELSE CFij = NNN/NIPLUS ;
```

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

```
RRij = RATEIJ/RATEIREF ;
RUN ;

***********************************

STEP 8d: Calculate the (age)stratum-specific population attributable risk fraction
AFPi = CFi1 * (RRi1-1)/RRi1 + CFi2 * (RRi2-1)/RRi2 + ... CFij * (RRij-1)/RRij

***********************************;

DATA Step8d ;
SET Step8c ;
BY AGECAT ;

DUMMY=1 ;
RETAIN AFPI ;

IF FIRST.AGECAT THEN DO ;
AFPI=0 ;
END ;

*************************

NOTE: IF RRij=. because of an infinite risk ratio, then (RR-1)/RR is basically equal to 1, so apply the whole case
fraction to the AFPI

*************************;

IF RRij in (.,0) THEN DO ;
AFPI=AFPI + CFij ;
END ;
ELSE DO ;
AFPI=AFPI + (CFij * (RRij-1)/RRij) ;
END ;

IF LAST.AGECAT THEN DO ;
OUTPUT ;
END ;

KEEP AGECAT NIPLUS AFPI DUMMY ;
RUN ;

***********************************

STEP 8e: CALCULATE THE GRAND TOTAL OF CASES TO USE IN CALCULATING WEIGHTS FOR EACH AGE STRATUM

***********************************;

DATA Step8e ;
SET Step8d END=LASTOBS ;

RETAIN NPLUSPLUS ;
IF _N_=1 THEN DO ;
NPLUSPLUS=0 ;
```

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

```
END ;
NPLUSPLUS=NPLUSPLUS+NIPLUS ;
IF LASTOBS THEN OUTPUT ;

KEEP DUMMY NPLUSPLUS ;

**********************************

STEP 8f: CALCULATE THE AGGREGATED POPULATION ATTRIBUTABLE RISK USING THE AGE SPECIFIC WEIGHTS.

Determine age-stratum specific weights from the case distribution:

w_i = NUMERi+/NUMER++

and alculate the aggregated AFPagg = SUM(w_i * AFPi)

**********************************;

DATA Step8f ;
MERGE Step8d END=LASTOBS Step8e ;
BY DUMMY ;

RETAIN AFPAGG ;

IF _N_=1 THEN DO ;
AFPAGG=0 ;
END ;

AFPAGG = AFPAGG + ((NIPLUS/NPLUSPLUS)*AFPi) ;

IF LASTOBS THEN DO ;

OUTPUT ;

END ;

RUN ;

proc print ;
var AFPAGG ;
run ;
```

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).

## REFERENCES

1. Breslow NE, Day NE (eds). Statistical Methods in Cancer Research, Vol. II: The Design and Analysis of Cohort Studies. Oxford, UK: Oxford University Press, 1987.

2. Anderson RN, Rosenberg HM. Age standardization of death rates: implementation of the year 2000 standard; National Vital Statistics Reports: Vol 37, No. 3. Hyattsville, MD: National Center for Health Statistics, 1998.

3. Fay MP, Feuer EJ. Confidence intervals for directly standardized rates: a method based on the gamma distribution. Statistics in Medicine 1997;16:791-801.

4. Rothman KJ, Greenland S. Modern Epidemiology. 2nd Edition. Philadelphia: Lippincott-Raven, 1998.

5. Pamuk ER. Social class inequality in mortality from 1921 to 1972 in England and Wales. Popul Stud 1985;39:17-31.

6. Wagstaff A, Paci P, van Doorslaer E. On the measurement of inequalities in health. Soc Sci Med 1991;33:545-57.

7. Davey Smith G, Hart C, Hole D, et al. Education and occupational social class: which is the more important indicator of mortality risk? J Epidemiol Community Health 1998;52:153-60.

8. JA Hanley, A heuristic approach to the formulas for population attributable fraction. J Epidemiol Community Health 2001;55:508-514.

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR). This training is supported by the National Institutes of Health (1 R25 HD047185-01).