Epi 265: Epi Methods 3/ Research Methods in Chronic Disease Epi

Longitudinal data: design tradeoffs

Why do we prefer longitudinal data?

- Substantive question entails longitudinal data
  - Long etiologic period between exposure and outcome
  - Disease incidence or disease recovery processes are substantive outcomes
- To strengthen causal inference
  - Establish temporal order
  - Help to avoid adjusting for mediating variables or variables downstream of outcome
  - Address time-constant confounders (even ones that you did not measure)
  - None of these is a slam dunk

Does longitudinal data establish temporal order?

- Often assumed that if exposure is measured prior to when the outcome is measured, temporal order is established and you have ruled out Y→X
- Sometimes yes:
  - Household income measured in 2000 and incident stroke measured from 2000-2008.
- Sometimes unclear:
  - Depressive symptoms measured in 2000 and incident stroke measured from 2000-2008.
- Sometimes, probably not:
  - Memory loss measured in 2000 and incident stroke measured from 2000-2008
- Sometimes *obviously* not:
  - Memory loss measured in 2000 and APOE allele status measured in 2008

Does longitudinal data establish temporal order?

- Often assumed that if exposure is measured prior to when the outcome is measured, temporal order is established and you have ruled out Y→X
- This assumption should be considered on case by case basis
- For many outcomes, including conditions we consider "acute onset", there is a long preceding period of developing pathophysiology that may influence exposure status
- Especially confusing is when exposure itself is defined as a long term experience

Does longitudinal data establish temporal order?

- Often assumed that if exposure is measured prior to when the outcome is measured, temporal order is established and you have ruled out Y→X
- This assumption should be considered on case by case basis
- For many outcomes, including conditions we consider "acute onset", there is a long preceding period of developing pathophysiology that may influence exposure status
- Especially confusing is when exposure itself is defined as a long term experience

Does longitudinal data establish temporal order? Education and smoking

- Farrell and Fuchs: Stanford Heart Disease Prevention Program, interviews in 1979 people aged 24+ (born 1947-1955), restricted to NH with 12-18 years of schooling
- Assumption is that all were still in school at age 17, and all had equivalent amounts of schooling as of age 17
- Retrospective question on lifetime history of smoking

Smoking at age 24

	Men	Women
Years of schooling	−0.352[c] (0.090)	−0.305[c] (0.094)
Father's years of schooling	−0.092 (0.049)	0.004 (0.049)
Intercept	6.065[c] (1.290)	3.706[c] (1.209)
N	178	195
$\bar{P}$	0.455	0.379

[a]Cohort defin 17 years c
[b]Significant at
[c]Significant at

More years of schooling predicts lower odds of smoking at age 24

Farrell and Fuchs, J Health Econ 1982

Does longitudinal data establish temporal order? Education and smoking

- Farrell and Fuchs: Stanford Heart Disease Prevention Program, interviews in 1979 people aged 24+ (born 1947-1955), restricted to NH with 12-18 years of schooling
- Assumption is that all were still in school at age 17, and all had equivalent amounts of schooling as of age 17
- Retrospective question on lifetime history of smoking

Smoking at age 24

	Men	Women
Years of schooling	−0.352[c] (0.090)	−0.305[c] (0.094)
Father's years of schooling	−0.092 (0.049)	0.004 (0.049)
Intercept	6.065[c] (1.290)	3.706[c] (1.209)
N	178	195
$\bar{P}$	0.455	0.379

[a]Cohort defin 17 years c
[b]Significant at
[c]Significant at

More years of schooling predicts lower odds of smoking at age 24

But the same pattern was apparent when all had the same years of schooling

Smoking at age 17

Men	Women
−0.317[c] (0.100)	−0.454 (0.122)
−0.083 (0.051)	−0.040 (0.057)
4.614[c] (1.362)	5.556[c] (1.546)
178	195
0.287	0.246

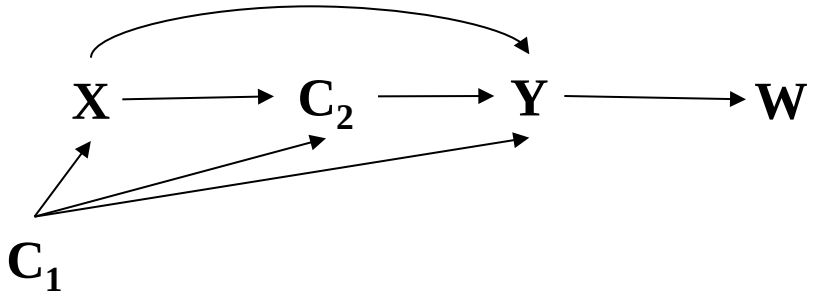
Farrell and Fuchs, J Health Econ 1982

Does longitudinal data establish temporal order? Education and smoking

- Despite these caveats, longitudinal data is obviously superior to cross-sectional data to establish temporal order.
- Key to ask whether the variables are truly "point in time" exposures and outcomes, or may have persisted across time in their current or some incipient form.

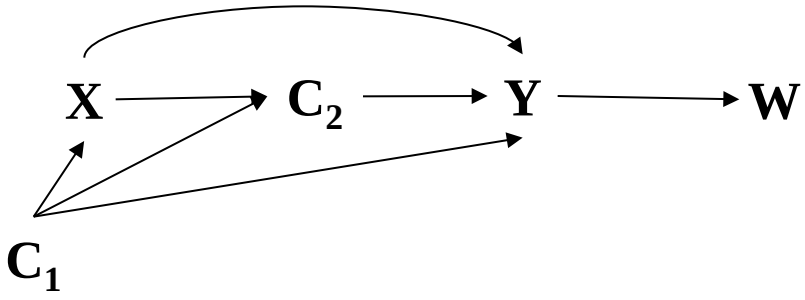
Longitudinal data helps to avoid adjusting for mediators or variables affected by the outcome

- Major conceptual error that turns out to be fairly common:
  - If you are interested in estimating the effect of X on Y, do not adjust for any variable potentially influenced by X or potentially influenced by Y.
  - The conceptual concern should be obvious, but consider how this relates to analyses of an RCT: *never* adjust for post-randomization variables, because they might be a consequence of randomization
- To estimate the effect of X on Y, do not adjust for A. Do not adjust for W. Do adjust for C.



Longitudinal data helps to avoid adjusting for mediators or variables affected by the outcome

- To estimate the effect of X on Y, do not adjust for A. Do not adjust for W.
- But in cross-sectional data, they are often measured all at the same time.
- You must impose the temporal order by assumption: might be possible for some variables (genetics, education, "ever smoked")
- If self-reported, potential for recall bias
- Regardless, consider possible selection/survival bias



W
Y
X
C

Longitudinal data help address time-constant confounders (even ones that you did not measure)

- Typical system of equations describing confounding:
- $Y=\beta_0+\beta_1*X+\beta_2*C+\beta_3*U+\varepsilon_Y$
- $X=\alpha_0+\alpha_1*C+\alpha_2*U+\varepsilon_X$
- C and U are both confounders of the X-Y association, because they influence both X and Y.
- Thus, the value of X received is not independent of the counterfactual value of Y
- One implication of this is that if you estimate a regression model omitting U:
- $Y=\hat{\beta}_0+\hat{\beta}_1*X+\hat{\beta}_2*C+\hat{\epsilon}_Y$
- $\hat{\beta}_1$ will be biased, and in a standardized linear model (paths=correlation coefficients), the bias will be: $\beta_{3*}\alpha_2$

Longitudinal data help address time-constant confounders (even ones that you did not measure)

- $Y=\beta_0+\beta_1*X+\beta_2*C+\beta_3*U+\varepsilon_Y$
- $X=\alpha_0+\alpha_1*C+\alpha_2*U+\varepsilon_X$
- If you estimate a regression model omitting U:
- $Y=\hat{\beta}_0+\hat{\beta}_1*X+\hat{\beta}_2*C+\hat{\epsilon}_Y$
- $\hat{\beta}_1$ will be biased, and in a standardized linear model (paths=correlation coefficients), the bias will be: $\beta_{3*}\alpha_2$
- If the coefficients are not standardized, the bias will be:
- $\beta_3*\alpha_2*\sigma_u^2/\sigma_x^2$
- (this is not that important: it's usually easier to think of correlation coefficients).

Longitudinal data help address time-constant confounders (even ones that you did not measure)

- Typical system of equations describing confounding:
- $Y=\beta_0+\beta_1*X+\beta_2*C+\beta_3*U+\varepsilon_Y$
- $X=\alpha_0+\alpha_1*C+\alpha_2*U+\varepsilon_X$
- Now let's make the above equations time specific:
- $Y_{i,t}=\beta_0+\beta_1*X_{i,t}+\beta_2*C_{i,t}+\beta_3*U_i+\varepsilon_Y$
- $X_{i,t}=\alpha_0+\alpha_1*C_{i,t}+\alpha_2*U_i+\varepsilon_X$
- Note that although C, X, and Y are time-specific, U is time constant
- As we have discussed, if you estimate:
- $Y_{i,t}=\hat{\beta}_0+\hat{\beta}_1*X_{i,t}+\hat{\beta}_2*C_{i,t}+\hat{\epsilon}_Y$
- It is biased as a function of: $\beta_{3*}\alpha_2$

Longitudinal data help address time-constant confounders (even ones that you did not measure)

- $Y_{i,t}=\beta_0+\beta_1*X_{i,t}+\beta_2*C_{i,t}+\beta_3*U_i+\varepsilon_{Yi,t}$
- $X_{i,t}=\alpha_0+\alpha_1*C_{i,t}+\alpha_2*U_i+\varepsilon_{Xi,t}$
- Note that although C, X, and Y are time-specific, U is time constant
- As we have discussed, if you estimate:
- $Y_{i,t}=\hat{\beta}_0+\hat{\beta}_1*X_{i,t}+\hat{\beta}_2*C_{i,t}+\hat{\epsilon}_{Yi,t}$
- It is biased as a function of: $\beta_{3*}\alpha_2$
- Consider instead the "true" structural equation for changes in Y:

$$Y_{i,t}-Y_{i,t-1}=(\beta_0+\beta_1*X_{i,t}+\beta_2*C_{i,t}+\beta_3*U_i+\varepsilon_{Yi,t})-(\beta_0+\beta_1*X_{i,t-1}+\beta_2*C_{i,t-1}+\beta_3*U_i+\varepsilon_{Yi,t-1})$$

$$Y_{i,t}-Y_{i,t-1}=\beta_1*(X_{i,t}-X_{i,t-1})+\beta_2*(C_{i,t}-C_{i,t-1})+\beta_3*(U_i-U_i)+\varepsilon_{Yi,t}-\varepsilon_{Yi,t-1}$$

The omitted time-constant confounder does not predict change in Y.

Longitudinal data help address time-constant confounders (even ones that you did not measure)

- Consider instead the "true" structural equation for changes in Y:

$$Y_{i,t}-Y_{i,t-1}=(\beta_0+\beta_1*X_{i,t}+\beta_2*C_{i,t}+\beta_3*U_i+\varepsilon_{Yi,t})-(\beta_0+\beta_1*X_{i,t-1}+\beta_2*C_{i,t-1}+\beta_3*U_i+\varepsilon_{Yi,t-1})$$

$$Y_{i,t}-Y_{i,t-1}=\beta_1*(X_{i,t}-X_{i,t-1})+\beta_2*(C_{i,t}-C_{i,t-1})+\beta_3*(U_i-U_i)+\varepsilon_{Yi,t}-\varepsilon_{Yi,t-1}$$

- The omitted time-constant confounder does not predict change in Y.
- Key assumptions:
  - The confounder does not change over time
  - The effect of the counfounder (β_3) does not change over time
  - (Equivalent): the confounder does not influence change in Y

- Given these assumptions, you can use longitudinal data to eliminate bias from confounders you did not even measure.

Example: Effect of TV viewing on caloric intake

- Two assessments of children in Boston (n=548)
- Many plausible confounders: some measured many unmeasured.

Table 2. Baseline (Fall 1995) and Follow-up (Spring 1997) Dietary and Activity Characteristics of the Cohort (N = 548)*

Characteristic	Baseline Value	Follow-up Value	Crude Change
Dietary			
Total daily intake, kcal	2141 ± 1076	2299 ± 1128	158 ± 1199
Physical activity and inactivity			
Television viewing, h/d	2.44 ± 1.41	2.29 ± 1.38	−0.16 ± 1.30

From: Wiecha et al., **When Children Eat What They Watch: Impact of Television Viewing on Dietary Intake in Youth** Arch Pediatr Adolesc Med. 2006

Example: Effect of TV viewing on caloric intake

Table 3. Adjusted Change in Daily Caloric Intake Associated With Baseline and Ea and Evidence for Mediation of This Relationship by Change in Intake of Foods Com

Variable	Additional Kilocalories at Follow-up Associated With Each Hour of Television Viewing at Baseline	
	Mean (95% Confidence Interval)	*P* Value
Basic model (no foods in model)*	127.0 (27.7 to 226.2)	.02

Example: Effect of TV viewing on caloric intake

Table 3. Adjusted Change in Daily h/d Increase in Television Viewing and Evidence for Mediation of Thisly Advertised on Television (N = 548)

Variable	Additional Kilocalories at Follow-up Associated With Each Hour Increase in Television Viewing	
	Mean (95% Confidence Interval)	*P* Value
Basic model (no foods in model)*	167.0 (135.9 to 197.9)	<.001

- What is the usefulness of regressing change in calories on *baseline* versus *changes* in TV viewing?
- In other words, should you use X_1 or ΔX?

Using baseline exposure versus changes in exposure as independent variables

- What is the usefulness of regressing change in calories on *baseline* versus *changes* in TV viewing?
- We've already shown the power of using ΔX
- Baseline X is appropriate if X is time-constant or highly correlated across time
- But note you are assuming time-constant X influences *changes* in Y but time-constant U does not. Is this credible?
- You may also hypothesize there is a delay from exposure to consequences, e.g., an etiologic period, so baseline X is most relevant
- This highlights a huge limitation with using ΔX: it is a disaster if X at time (t) does not influence Y at time (t).
- For example if X at time (t-1) influences Y at time (t) and you misspecify this, you can get the absolutely wrong answer.

Design tradeoffs

- Sample size
- Lack of selection bias (predictors of outcome do not predict participation)
- Diversity/representativeness of population of interest
- Measurements available
  - Substantive relevance of measures
  - Quality of measures
  - Number of measures

- These qualities are usually in direct opposition
- Larger sample size, more diversity, less selection entails lower quality of measures (assuming a limited budget)

Design tradeoffs: surveillance versus etiologic research

- Diversity/representativeness much more important if your task is surveillance or descriptive
  - Prevalence of disease in a population
  - Trends across time
  - Magnitude of inequalities between groups
- If your goal is to estimate the effect of an exposure on an outcome, you may be willing to compromise on representativeness or even diversity of sample
- This is controversial

The importance (or not) of representative samples

"statistical inference, the process of inferring from a sample to the source from which it was drawn, is …aided by…a representative sample. The mistake is to think that statistical inference is the same as scientific inference … representativeness can be counterproductive … the main road to general statements on nature is through studies that control skillfully for confounding variables and thereby advance our understanding of causal mechanisms. Representative sampling does not take us down that road." - Rothman IJE 2013 "Why representativeness should be avoided"

- Why *counterproductive*? Introduces extraneous variance in the outcome and may reduce measurement quality.
- NHS argues: nurses are really reliable reporters! And there's no confounding by SES. (note: this is not actually true)

The importance (or not) of representative samples

“the selection of the subjects can influence both internal validity and external validity; and further, can modify the hypothesis being tested… External validity … is clearly a secondary issue; if the study has very low internal validity, the conclusions are likely to be wrong, and so its generalizability is irrelevant..”

- Elwood IJE 2013 “On representativeness”

“representativeness should neither be avoided nor uncritically embraced, but adopted (or not) according to the particular questions that are being addressed. …many factors that are associated with the outcome of interest are also likely to be linked to self-selection into a study.

The importance (or not) of representative samples

"representativeness should neither be avoided nor uncritically embraced, but adopted (or not) according to the particular questions that are being addressed. …[re the ACS volunteer cohort] many factors that are associated with the outcome of interest are also likely to be linked to self-selection into a study…. In the pre-modern epidemiology world the focus was often on triangulating evidence from across as many sources as possible. Such evidence comes from a variety of sources, of which some are deliberately non-representative (for example the considerable value of twin studies for identifying potentially causal epigenetic influences on disease, or the follow-up of natural experiments) and some of which (including essentially 100% coverage population linkage studies) will be representative." Ebrahim and Davey Smith

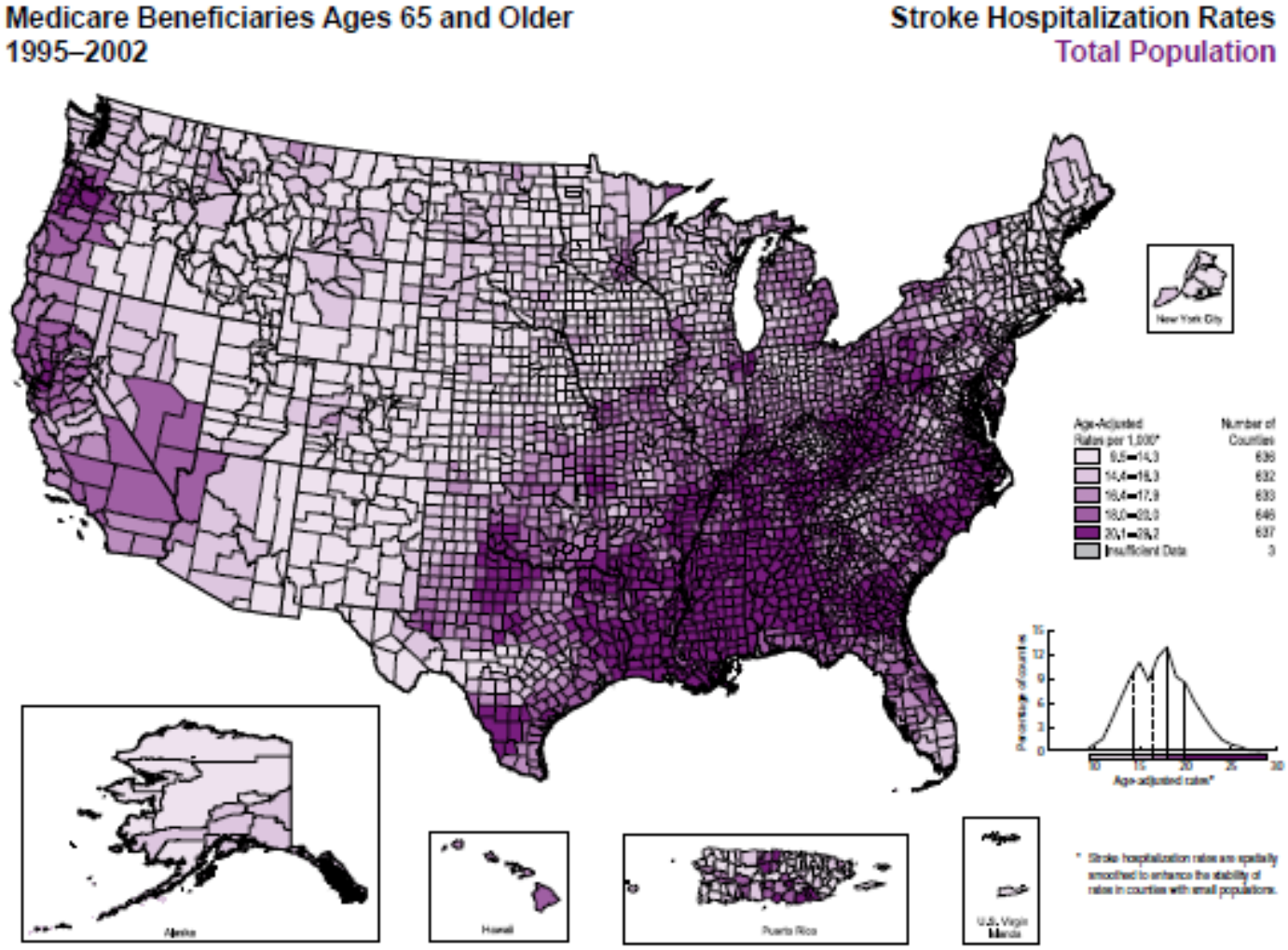
The importance (or not) of representative samples

- Not controversial: the use of complex sample designs
  - **Cluster** sampling: Separate the target population into non-overlapping clusters (neighborhoods, schools, states, any group such that everyone is included in one and only one such group). Choose a simple random sample of these clusters. Take a designated proportion of individuals from each of the selected clusters
  - **Stratified** sampling: take a different proportion of individuals with different characteristics, e.g., 1/100 80 year old women but 1/50 80 year old men.
- Generally need to correct for such design decisions in the analysis, but if sampling rules are known, you can reconstruct the original population.

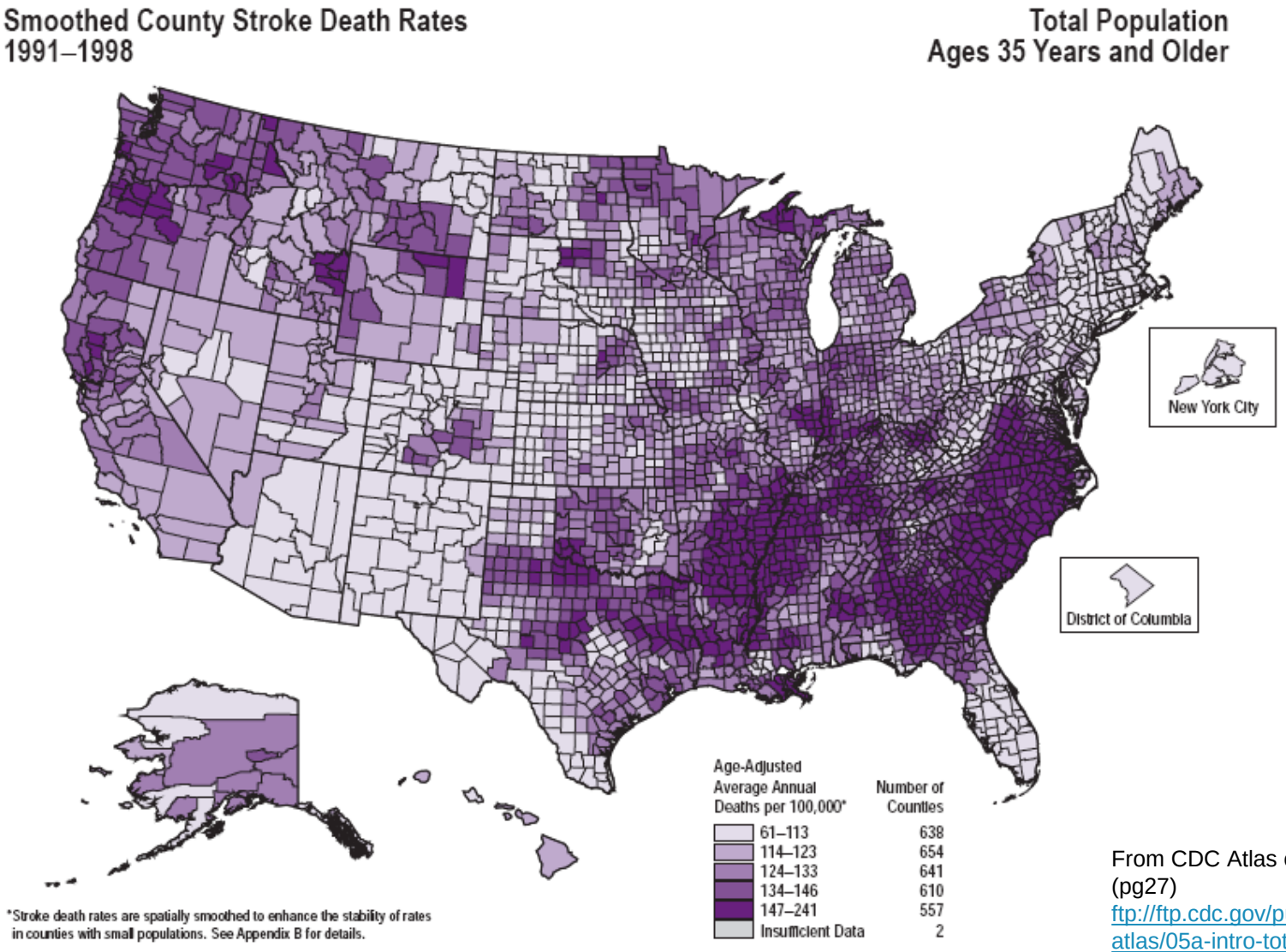
The importance (or not) of representative samples

- Possibly more critical: how does the sample constrain the research questions?
- Example of stroke
- Two important questions in stroke
  - Why such large geographic differences?
  - Why racial disparities?

Stroke Hospitalizations in the US: 1995-2002

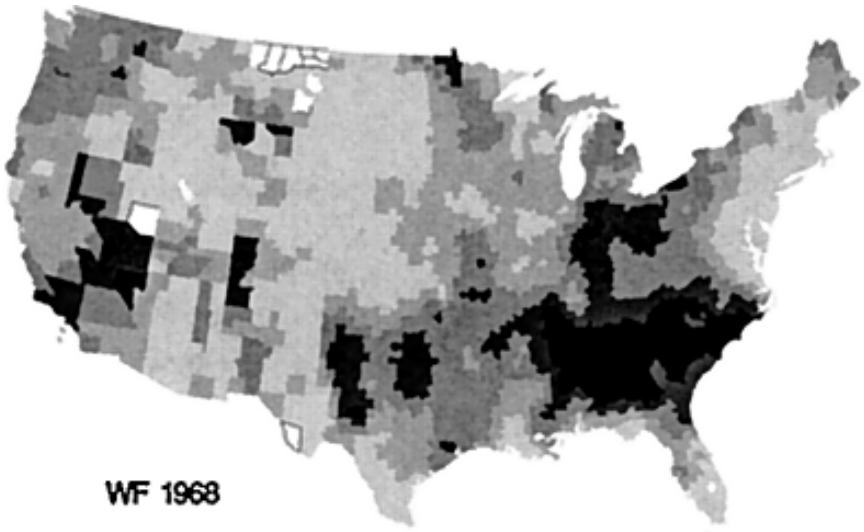


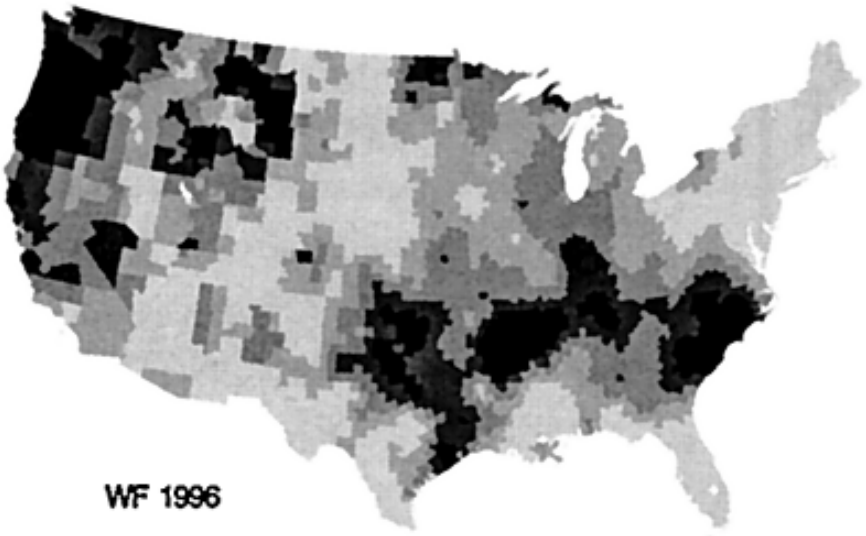
Stroke Mortality in the US: 1991-1998



From CDC Atlas of Stroke Mortality (pg27)
ftp://ftp.cdc.gov/pub/Publications/stroke_atlas/05a-intro-totalpop.pdf

Geographic Pattern of US Stroke Deaths: 1968 vs 1996





From Howard et al, *Stroke (*2001).

Excess Stroke in Blacks vs Whites

- "Risk factors for increased stroke mortality in blacks heretofore examined, primarily a higher prevalence of hypertension and diabetes and a lower socioeconomic status, explain only a fraction of this excess….

…It is possible that elevated blood pressure and insulin resistance/diabetes impact blacks more strongly, which perhaps suggests a genetic influence or susceptibility. …

… A more complete understanding of the mechanisms underlying the race/ethnic differences in stroke subtype will likely depend on advances in genomics. " – Wolf and Kannell, Circulation 2007

Competing Goals

- High quality stroke measures
  - Hospital based surveillance, with neuroimaging confirmation
- High quality exposure measures
  - Geographic variability
  - Measures prior to onset of stroke
- Long term follow-up

Major data sources for stroke surveillance (probably not comprehensive)

- Incidence:
  - Framingham Heart Study (one place, mostly white)
  - Rochester registry (one place, mostly white)
  - Brain attack surveillance in Corpus Christi (registry)
  - Atherosclerosis Risk in Communities Study (4 sites: NC, MS [only blacks], MN, MD)
  - Cardiovascular Health Study (4 sites: NC, CA, PA, MD)
  - Greater Cincinnati/Northern Kentucky Stroke Study (registry)
  - National Hospital Ambulatory Medical Care Survey (medical records)
  - National Hospital Discharge Survey
  - Northern Manhattan Stroke Study (one place)
  - Washington Heights/Inwood Community Aging Project (one place)
  - NHS: Nurses in 10 states
  - REGARDS (nationally representative)
  - HRS (nationally representative)
- Prevalence: add NHANES, BRFSS, NHIS

Single site sampling conflates race and migration

The overlap of race and place for older Americans

		African-Americans		Whites	
		Birthplace		Birthplace	
		Non-Southern	Southern	Non-Southern	Southern
Current Residence	Non-Southern	15%	29%	60%	5%
	Southern	2%	54%	10%	25%
		17%	83%	70%	30%

2000 Census 5% sample, age 65+, race black or white, southern= census region.

Place of birth predicts stroke risk, independently of race

	Socioeconomic model,[b] HR (95% CI)
Proportion of entire life in SB	1.20 (1.01–1.42)
Partial life course exposure to SB	
Born in SB	1.12 (0.96–1.31)
Proportion of first 12 years	1.19 (1.02–1.40)
Proportion of ages 13–18	1.20 (1.03–1.40)
Proportion of ages 19–30	1.16 (0.99–1.37)
Proportion of ages 31–45	1.13 (0.97–1.32)
Proportion of last 20 years	1.17 (1.00–1.36)
Current residence in the SB	1.12 (0.96–1.30)

n=24,544
141,736 P-Y
615 strokes

Howard et al., Neurology, 2013

Place of birth substantially confounds race effects on stroke

	Model 1		Model 4	
	HR	(95% CI)	HR	(95% CI)
Black race	1.46	(1.29 , 1.65)	1.30	(1.15 , 1.48)
Southern birthstate			1.27	(1.16 , 1.41)
Low Parental SES Index			1.27	(1.10 , 1.47)
Fair/poor child health			1.43	(1.21 , 1.71)
Adult SES				
Income*				
Wealth*				
Main Occupation				
Manual/Unskilled				
Cardiovascular risk factors				
Hypertension				
Current smoker				
Vigorous activity				
Obese				
Diabetes				
Heart disease				

Assessments

Health and Retirement Study

- "Has a doctor ever told you you had a stroke

Reasons for Geographic and Racial Disparities in Stroke (REGARDS)

- Self reported hospitalization
  - Medical record or death certificate review or proxy interview
  - Central adjudication

WHICAP

- Self report
- Medical records
- MRI: lesions 3mm or larger

Self Report vs Neuroimaging

Sensitivity: 32%
Specificity: 79%

Blacks:

Sensitivity: 37%
Specificity: 77%

Whites:

Sensitivity: 27%
Specificity: 84%

Reitz, Arch Neur, 2009

Self Report vs Neuroimaging

"In stroke research, sensitive neuroimaging techniques, rather than stroke self-report should be used to determine stroke history."

-Reitz, Arch Neur, 2009

Self Report vs CMS Medical Records

- Sensitivity 74%
- Specificity 93%

Blacks:

Sensitivity: 75%

Specificity: 90%

Whites:

Sensitivity: 75%

Specificity: 93%

- Glymour unpublished, from HRS-CMS linkage

Why do Glymour & Reitz disagree?

- Many (most?) cerebral ischemic events are not diagnosed
- Up to 5 times more silent infarcts than clinical strokes
- CHS:
  - Silent infarcts defined as focal lesions greater than 3 mm that were hyperintense on T2 images and, if subcortical, hypointense on T1 images
  - 28% of CHS participants (age 65+) had silent infarcts

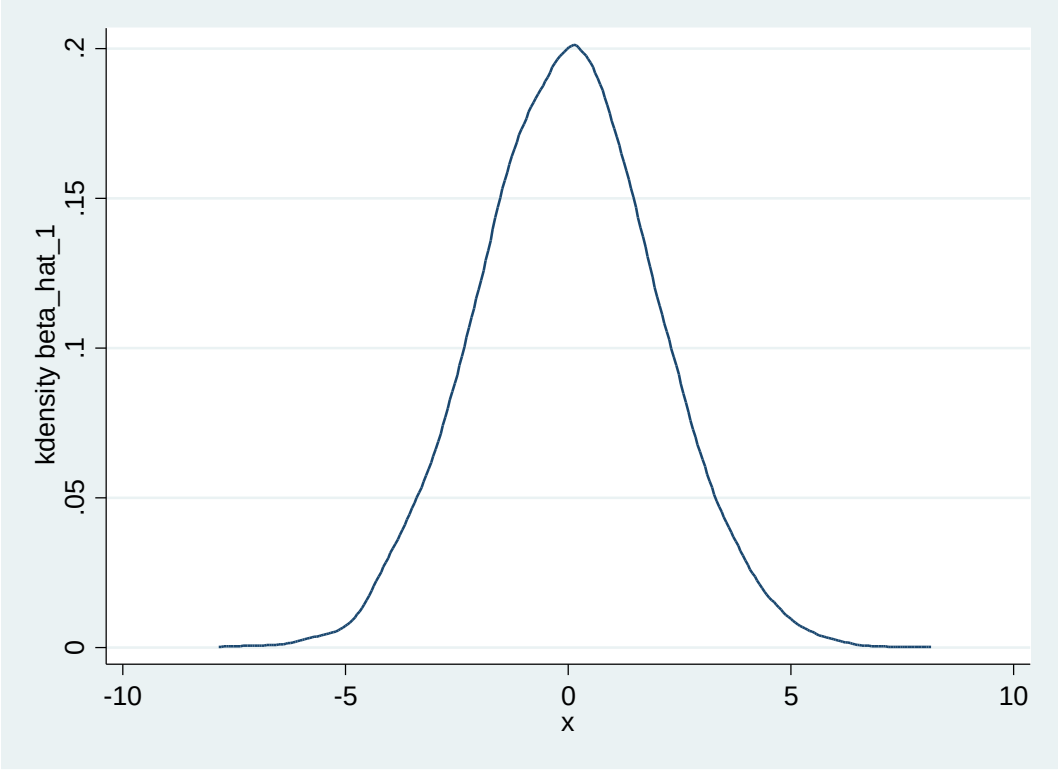
Implications

- Existing data sources very limited
- Drive research orientation away from social or contextual factors or even preventive factors
- Focus on stroke characteristics, stable personal characteristics, things that can be evaluated retrospectively from hospital based samples
- Need better validation built into ongoing cohort studies with self-report
- Because diagnosed stroke is only tip of the iceberg, need more than just medical record verification.
- Data availability fundamentally shapes the questions asked.

Sample size tradeoffs

- Mean squared error of an estimator:
- MSE=E[$(\hat{\beta}$-$\beta)^2$] = $\mathrm{Var}(\hat{\beta}) + (bias)^2$

An estimator with a small bias and small variance might be closer to the truth (on average) than an unbiased estimator with a large variance

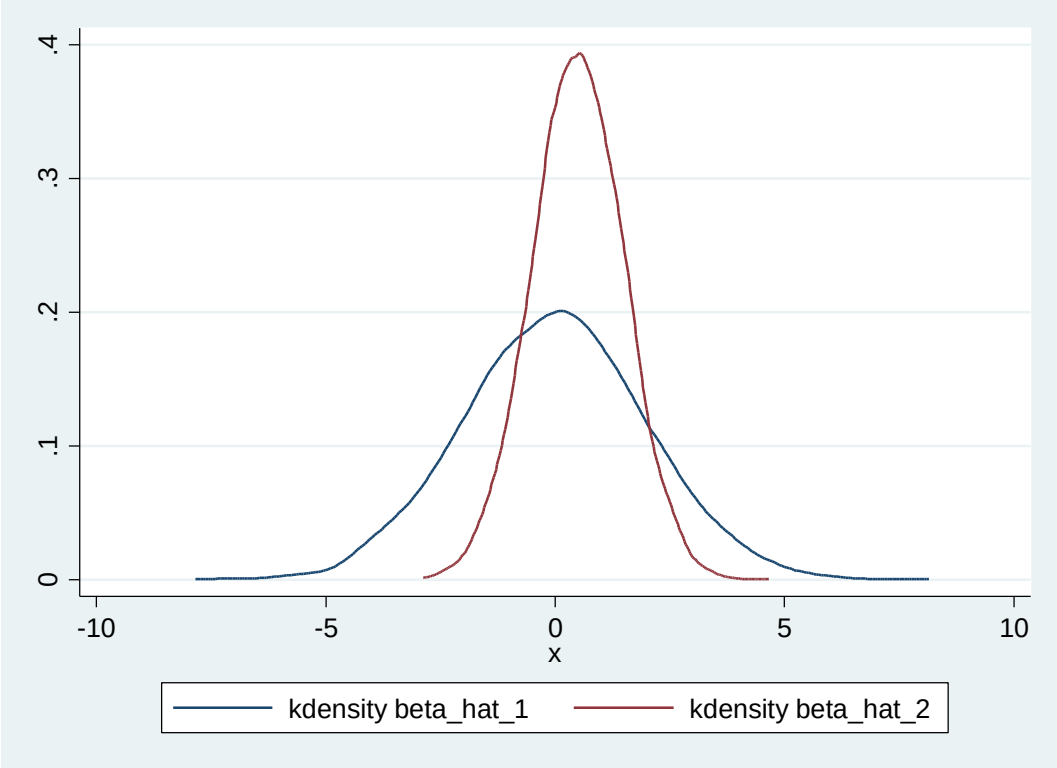


True beta is zero, but beta_hat_1 has some variance around zero.

Sample size tradeoffs

- Mean squared error of an estimator:
- MSE=E[$(\hat{\beta}\text{-}\beta)^2$] = $\mathrm{Var}(\hat{\beta}) + (bias)^2$

An estimator with a small bias and small variance might be closer to the truth (on average) than an unbiased estimator with a large variance

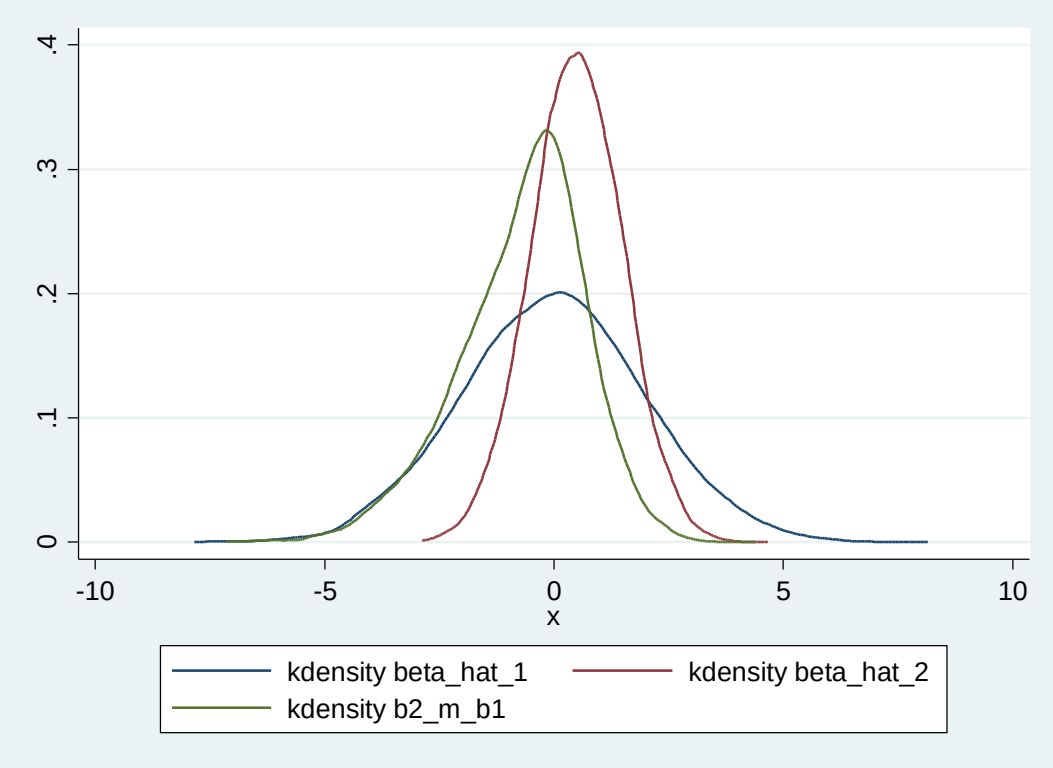


Beta_hat_2 is biased mean(=.5) but has less variance.

Sample size tradeoffs

- Mean squared error of an estimator:
- MSE=E[$(\hat{\beta}$-$\beta)^2$] = $\mathrm{Var}(\hat{\beta}) + (bias)^2$

An estimator with a small bias and small variance might be closer to the truth (on average) than an unbiased estimator with a large variance



Difference in absolute values (beta_hat_2 minus beta_hat_1) is negative

End Lecture 2