2021 SISMID Module 5
Lecture 8: Ecological Studies

Jon Wakefield and Lance Waller

Departments of Statistics and Biostatistics
University of Washington

Outline

Motivation

Confounding and Collapsibility

Ecological Bias

Other Issues

Motivation

Ecological studies

- Data are available on *groups of individuals*
  - rather that the individuals themselves
  - common grouping is based on geographic location; areas
- Examine associations at the *group*-level, rather than at the *individual*-level
- Ecological correlation studies compare group-level health outcome summaries to group-level predictor variables
  - Wakefield (2008)

Ecological Studies

Examples

- Cancer epidemiology
  - breast cancer vs dietary fat in different countries (*large* areas)
- Environmental epidemiology
  - water constituents, air pollution (*small* areas)
- Sociology
  - unemployment or crime and socioeconomic factors
- Political science
  - voter registration and race

Some history

- Durkheim (1897)
  - illustrated the ecological fallacy in the setting of a study of suicide rates and religion
- Robinson (1950)
  - showed how the correlation between race and literacy ranged from 0.95 to 0.2, depending on the level of aggregation
  - Subramanian et al. (2009b); Oakes (2009); Subramanian et al. (2009a); Wakefield (2009) provide recent analyses of these data.
- Selvin (1958)
  - coined the term *ecological fallacy*
  - "relationships between characteristics of individuals are wrongly inferred from data about groups"

Advantages

Scientific

- Suitable and appropriate for group-level associations
  - regional air/water pollution regulation

Data considerations

- Ecological data are often relatively easy and cheap to obtain
  - air/water pollution data
  - US Census online
- Ecological data may be the only available information
  - lack of high-quality individual-level information
  - confidentiality considerations; may be a researchers only recourse
  - rely on census information at the census tract level

Statistical considerations

- Exploit large between-group exposure variation
  - e.g. dietary fat intake studies conducted in different countries
  - increased power for exposure-response trends when within-area exposure variation is low
- Exposure subject to certain types of measurement error

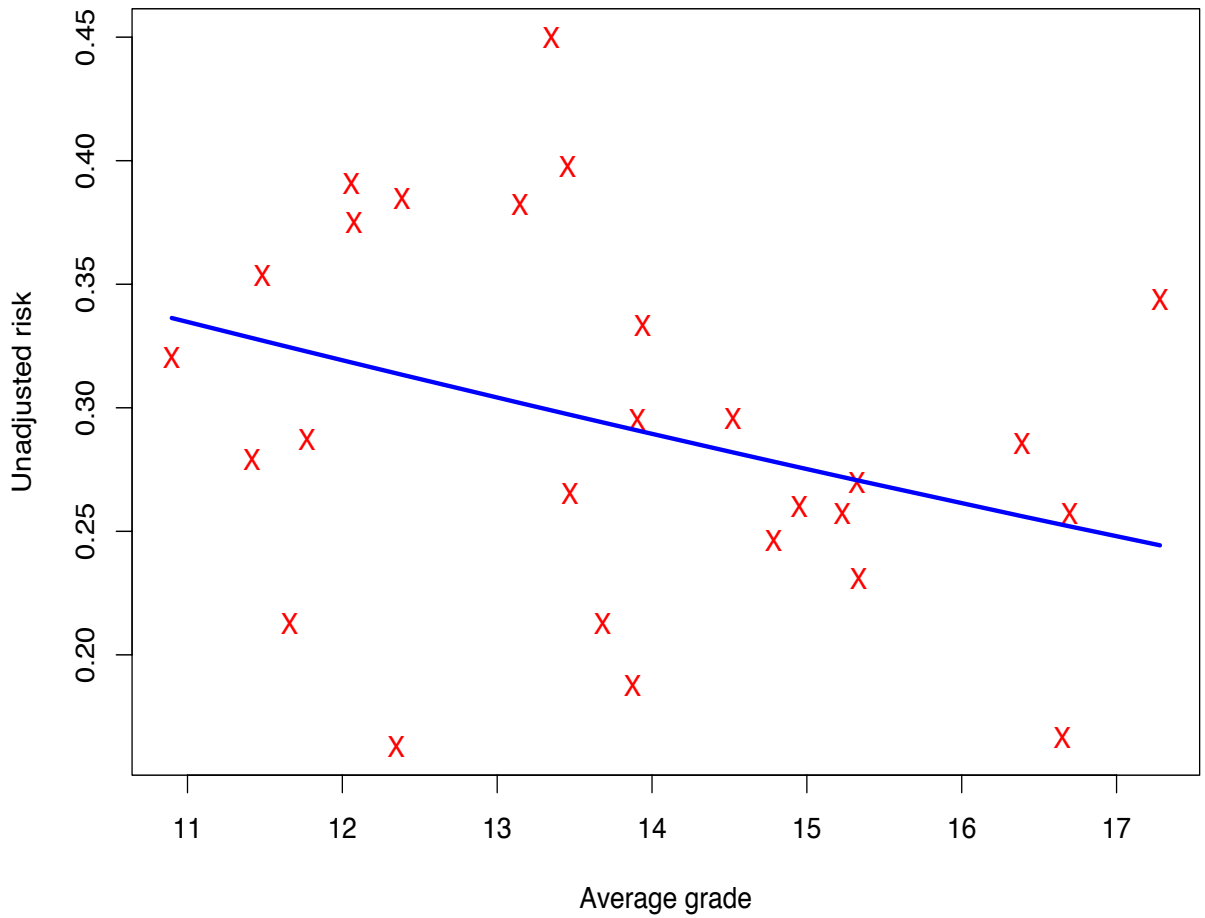
Disadvantages

- Extensive epidemiological literature: Greenland (1992); Greenland and Robins (1994); Richardson and Monfort (2000); Wakefield (2004, 2008)
- Observational study design
- Range of additional biases unique to the ecological study design
  - both within- and between-area confounding and effect modification
  - contextual effects
  - lack of mutual standardization
  - pure specification bias
- Collective impact is often referred to as *ecological bias*
- May lead to the phenomenon known as the *ecological fallacy*
  - conclusions drawn on the basis of group-level data are opposite to those based on an analysis which uses individual-level data

Ecological regression analysis

- Example from CHS: association between education and mortality
- Let $Y = 0/1$ be an indicator of death
  - within 11 years of follow-up
- Let X denote education (grade)
  - 0 = no schooling, . . ., 21 = graduate or professional
- Data from zipcodes with at least 30 study participants
  - 27 zipcodes with 2,347 people
- Consider a hypothetical ecological study where we observe:
  - N_k, the total number of individuals in area k
  - Y_k, the total number of diseased individuals in area k
  - $\overline{X}_k$, the average grade across individuals in area k
- Unadjusted risk as a function of average grade, across 27 zipcodes
  - Y_k/N_k vs $\overline{X}_k$, $k = 1, \ldots, 27$

Ecological regression analysis



Ecological regression analysis

- For a non-rare outcome, we might consider a logistic regression model:

$$Y_k | \, \overline{X}_k \;\sim\; \text{Binomial}(N_k, \; p_k)$$

where for $k = 1, \ldots, 27$

$$\log\left(\frac{p_k}{1-p_k}\right) \;=\; \beta_0^* \;+\; \beta_x^* \overline{X}_k$$

- This model yields $\exp\{\hat{\beta}_x^*\} = 0.93$, with a 95% CI of (0.89, 0.98)

Ecological regression analysis

Interpretation?

- p_k is an 'average risk' in area k
- Comparing the average risk between two groups of individuals (i.e., zipcodes) whose average education differ by one grade
- What is the mechanism here?
- What are the potential sources of bias?

Individual-level associations

- Distinguish between the *biological* (or sociological) effect at the individual level and the *ecological* effect at the group level
- Ecological effect may depend on
  - magnitude of the biological effect
  - degree and pattern of exposure within a group
- Ecological studies do not directly assess links between exposure and outcome at the level of the individual
  - interpretation, in terms of a biological effect, is difficult
- *Cross-level* inference.
  - transfer estimation/inference to the individual level
- Generally, we assess bias with respect to individual-level associations

Confounding and Collapsibility

Confounding and non-collapsability

Causal inference

- Suppose we wish to evaluate the impact of exposure X on outcome Y in some target population A
- Define

$$\begin{aligned} \mu_{A0} &= \mathrm{E}[Y|\text{population } A, X = 0] \\ \mu_{A1} &= \mathrm{E}[Y|\text{population } A, X = 1] \end{aligned}$$

- The *causal effect* is the difference between μ_{A0} and μ_{A1}
  - for convenience, consider the ratio: μ_{A1}/μ_{A0}
  - other choices; risk difference and odds ratio
- We cannot observe both of these, but we could observe μ_{A1} and, say,

$$\mu_{B0} = \mathrm{E}[Y|\text{population } B, X = 0]$$

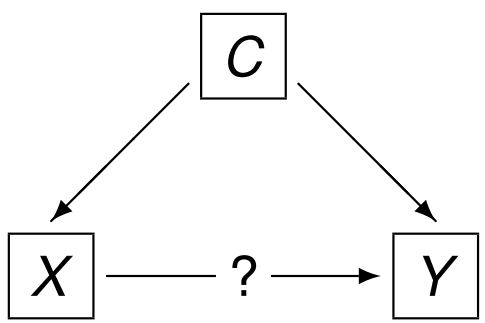
  - population B is the control or reference population

Confounding

- We estimate μ_{A1}/μ_{A0} via μ_{A1}/μ_{B0}
- Confounding occurs when $\mu_{B0} \neq \mu_{A0}$
  - due to differences between populations A and B
  - distorted view of the impact of X on Y
- Greenland et al. (1999)

Practical definition of Confounding

- A variable C which is associated to both X and Y, but not in the causal pathway and not caused by Y.
- Causal diagram:



Confounding

Example I

- Association between exposure X and outcome Y, controlling for a confounder (smoking):
  - Smokers:

	Y=1	Y=0	Total	P(Y=1$\mid X$)
X=1	8	2	10	0.80
X=0	18	12	30	0.60

  - Non-smokers:

	Y=1	Y=0	Total	P(Y=1$\mid X$)
X=1	9	21	30	0.30
X=0	2	8	10	0.20

- Exposure is harmful

Confounding

- Suppose we aggregate over the smoking variable:

	Y=1	Y=0	Total	P(Y=1$\mid X$)
X=1	17	23	40	0.43
X=0	20	20	40	0.50

- Exposure is protective
- Direction of the association is reversed:
  - imbalance of smoking among the exposure groups

	C=1	C=0	Total	P(C=1$\mid X$)
X=1	10	30	40	0.25
X=0	30	10	40	0.75

- Known as *Simpson's paradox* (Simpson, 1951).

Non-collapsibility

Example II

- Greenland et al. (1999)
  - Smokers:

	Y=1	Y=0	Total	P(Y=1$\mid X$)
X=1	80	20	100	0.80
X=0	60	40	100	0.60

  - Non-smokers:

	Y=1	Y=0	Total	P(Y=1$\mid X$)
X=1	40	60	100	0.40
X=0	20	80	100	0.20

- Exposure is harmful
  - odds ratio = 2.66 in both stratum
- Smoking is not a confounder here
  - exposure distribution is the same among smokers and non-smokers

Non-collapsibility

- Aggregating yields:

	Y=1	Y=0	Total	P(Y=1$\mid X$)
X=1	120	80	200	0.60
X=0	80	120	200	0.40

- Exposure is still harmful but now the odds ratio = 2.25
  - difference even though there is no confounding!
- Phenomenon known as *non-collapsability*
- Specific to the choice of 'association'
  - non-linearity of the odds ratio
  - risk difference is not affected

Bottom line

When we deal with aggregated data, both confounding and non-collapsability can result in a 'distorted' view of the impact of X on Y.

Ecological Bias

Ecological confounding

Between-area confounding

- In an ecological study the unit of analysis is a group or area
- Between-area confounding is analogous to conventional confounding
  - control for imbalances in group-level confounder distribution

Within-area confounding

- In an ecological study, only observe *marginal information*
  - binary exposure X and binary confounder C

	C=1	C=0	
X=1			N_1
X=0			N_0
	M_1	M_0	

  - observe the total number of smokers, M_1
  - observe the total number exposed, N_1
- Unfortunately, don't observe the number of exposed smokers
  - internal cells of the 2×2 table
- Need to be able to control for imbalances in the *within-area distribution* of exposures/confounders
  - in particular, across areas
- Many ways of filling in the table if we observe just the margins

	C=1	C=0	
X=1	??	??	N_1
X=0	??	??	N_0
	M_1	M_0	

We have 3 unknown probabilities but just two pieces of information (the marginal proportions).

Contextual effects

- A contextual effect is a characteristic of individuals in a shared group
- Intuitively, it is not just your own exposure that determines your own risk but also the exposures of those surrounding you.
  - in measurement of school test scores, IQ of classmates is an example
- Causal interpretation:
  - are they real?
  - perhaps reflect unmeasured confounders?
- In the social sciences, this aspect has been emphasized

Contextual effects

- In an aggregate study, one cannot distinguish between individual-level and contextual effects
  - in political science, Y is Republican/Democrat and X is White/Non-White (say): not possible to distinguish between individual and contextual race
  - Specifically, consider the individual-level models:

$$\begin{aligned} \mathrm{E}[Y_{ki} | \, X_{ki}] &= \beta_0 + \beta_I X_{ki} \\ \mathrm{E}[Y_{ki} | \, \overline{X}_k] &= \beta_0 + \beta_C \overline{X}_k \end{aligned}$$

    for individual's i in areas k
  - Under aggregation:

$$\mathrm{E}[\overline{Y}_k | \, \overline{X}_k] = \beta_0 + \beta_1 \overline{X}_k$$

    so we can't tell which individual model is appropriate from the ecological data alone

Contextual effects

Example

- Suppose interest lies in the association between income, X, and blood pressure, Y
- Consider the linear model

$$\mathrm{E}[Y_{ki} | \, X_{ki}, \overline{X}_k] \; = \; \beta_0 \; + \; \beta_W (X_{ki} - \overline{X}_k) \; + \; \beta_B \overline{X}_k$$

  - i^{th} individual in area k
- Parameter interpretation:
  - β_W is the effect of *within*-group income
  - β_B is the effect of *between*-group income, i.e. the contextual effect
- In the setting of an ecological study, we would aggregate to obtain the *induced* model:

$$\mathrm{E}[Y_k | \, \overline{X}_k] \; = \; \beta_0 \; + \; \beta_B \overline{X}_k$$

  - only estimate the contextual effect

Mutual standardization

- Standardization is a common technique to adjusting potential confounding.
- In some circumstances, disease rates are published after having been standardized to a particular population
  - care need to be taken to ensure that the ecological exposure data is also standardized to the same population
  - likely be an issue when data are obtained from different sources
- For example, suppose we wish to examine the association between poverty and disease risk, controlling for age (J age bands)
- Area-specific, age-standardized disease rate

$$R_k^* = \sum_{j=1}^{J} w_j R_{kj}$$

  - w_j is the proportion of a standard population in age band j
  - R_{kj} is the disease rate in the j^{th} age band in the k^{th} area

Mutual standardization

- Let X_k denote the (raw) proportion below the poverty line in area k
  - if we use this to estimate an association, bias will result
  - must standardize X to the same population
  - need to calculate

$$X_k^* = \sum_{j=1}^{J} w_j X_{kj}$$

  - X_{kj} is the proportion below the poverty line in the j^{th} age band in the k^{th} area

Pure specification bias

- Suppose the individual-level disease model is of the form:

$$\mathrm{P}(Y_{ki} = 1 | X_{ki}) = \exp\{\beta_0 + \beta_x X_{ki}\}$$

  - i^{th} individual in area k
  - $\exp\{\beta_x\}$ is the relative risk associated with a unit increase in X
- In an ecological study we may only observe

$$\text{Total diseased} : Y_k = \sum_{i=1}^{N_k} Y_{ki}$$

$$\text{Average exposure} : \overline{X}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} X_{ki}$$

  - N_k is the total number of individuals in area k
- Tempting to fit a log-linear model on the group-level data

$$\mathrm{E}[Y_k | \overline{X}_k] = N_k \exp\{\beta_0^* + \beta_x^* \overline{X}_k\}$$

Pure specification bias

Question When is $\beta_x^* = \beta_x$?

- Consider the aggregate risk model induced by the individual-level model
  - *average risk* in area k

$$\mathrm{P}(Y = 1 | \text{ area } k) \;=\; \frac{1}{N_k} \sum_{i=1}^{N_k} \exp\{\beta_0 \;+\; \beta_x X_{ki}\}$$

  - depends on all the X_{ki} within area k
- Investigate various scenarios by assuming within-area exposures follow a normal distribution; Normal($\overline{X}_k$, σ_k^2)
- For large N_k, the *induced* aggregate risk model is approximately

$$\mathrm{P}(Y = 1 | \text{ area } k) \;\approx\; \exp\left\{\beta_0 \;+\; \beta_x \overline{X}_k \;+\; \frac{\beta_x^2 \sigma_k^2}{2}\right\}$$

  - if $\sigma_k^2 = 0$ then we can fit the ecological regression model
  - we could fit this model if we observed information on the within-area exposure variability, σ_k^2
  - need more than just the area-specific mean exposure, $\overline{X}_k$

Pure specification bias

- Suppose the variance is a function of the mean (often $b > 0$ for environmental pollutants):

$$\sigma_k^2 = a + b\overline{X}_k$$

- The aggregate risk model reduces to

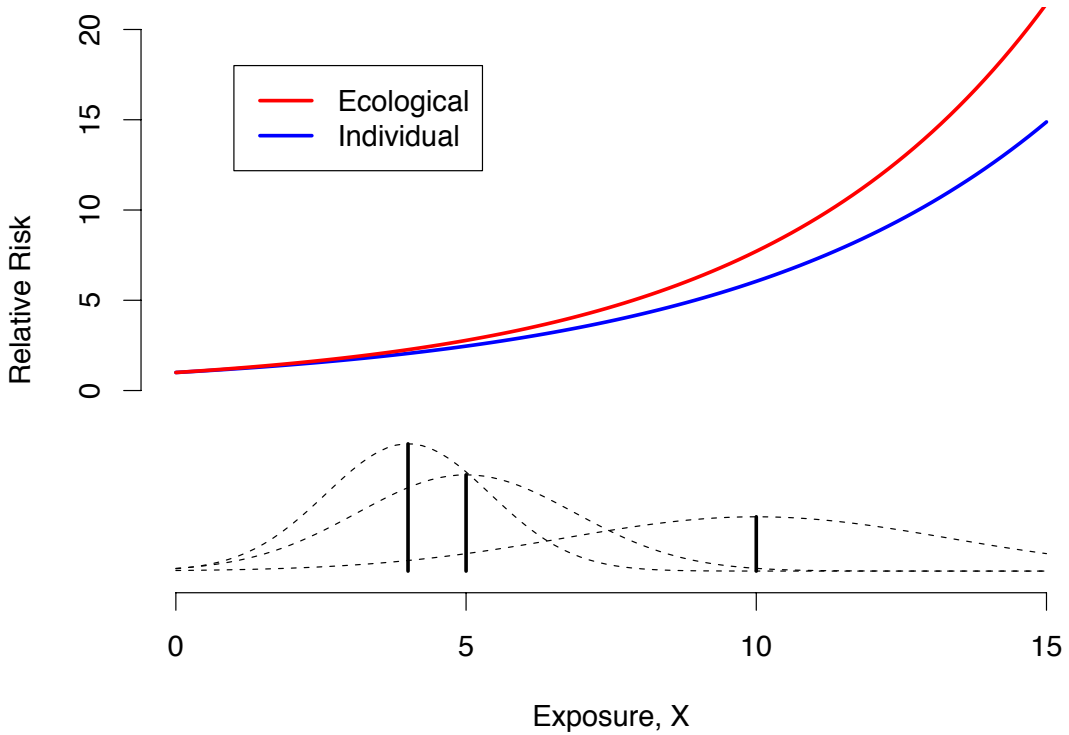
$$\mathrm{E}[Y_k | \overline{X}_k] = N_k \exp\{\beta_0^* + \beta_x^* \overline{X}_k\}$$

where

$$\beta_0^* = \beta_0 + \frac{a\beta_x^2}{2}, \quad \beta_x^* = \beta_x + \frac{b\beta_x^2}{2}$$

- If there is no mean-variance relationship (i.e. $b = 0$), there is *no bias*
- Suppose $\beta_x > 0$
  - if $b > 0$, the relative risk is *overestimated*
  - if $b < 0$, the relative risk is *decreased* and may change sign
- Suppose $\beta_x < 0$
  - if $b > 0$, the relative risk is *increased* and may change sign
  - if $b < 0$, the relative risk is *underestimated*

Illustration



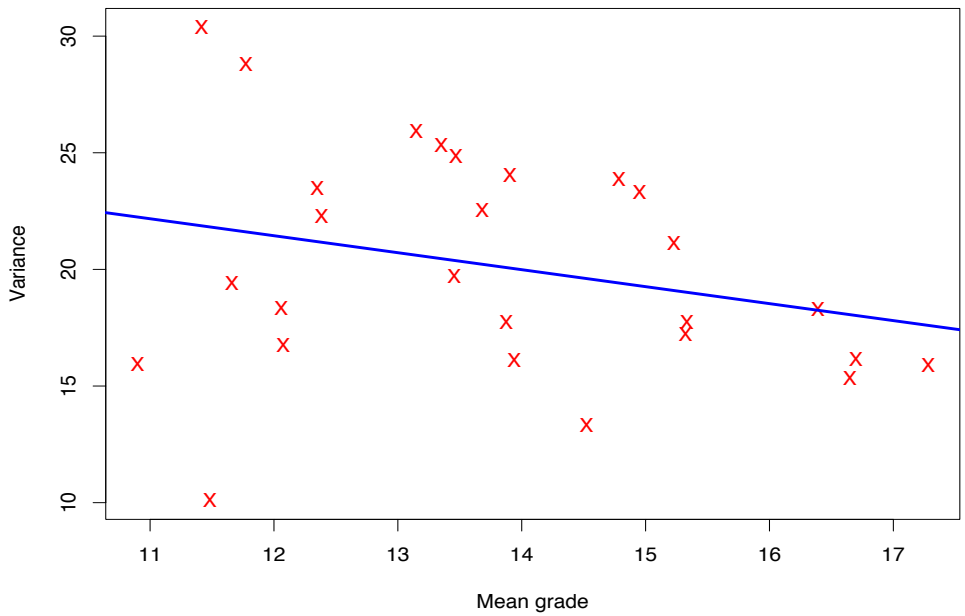
- Variance increases with the mean; $b > 0$
  - the values of $\overline{X}_k$ for 3 areas are shown

CHS example

Mean-variance relationship for grade within 27 zipcodes

Variance decreases with the mean; $b < 0$:

- less variation in 'more educated' zipcodes



CHS example

Ecological regression

- Recall that we earlier fitted the logistic model:

$$Y_k | \, \overline{X}_k \;\sim\; \text{Binomial}(N_k, \; p_k)$$

where

$$\log\left(\frac{p_k}{1-p_k}\right) \;=\; \beta_0^* \;+\; \beta_x^* \overline{X}_k$$

  - yields: $\exp\{\hat{\beta}_x^{\,*}\}$ = 0.93; 95% CI (0.89, 0.98)

CHS example

Individual-level analysis

- Now let's fit the individual-level logistic model:

$$\log\left(\frac{p_{ki}}{1-p_{ki}}\right) = \beta_0 + \beta_x X_{ki}$$

  - $p_{ki} = P(Y_{ki} = 1 | X_{ki})$; no longer modeling the 'average risk'
  - yields: $\exp\{\hat{\beta}_x\}$ = 0.97; 95% CI (0.95, 0.99)
  - ecological study overestimates the protective effect of education

Other Issues

Semi-ecological studies

- The fundamental challenge with ecological studies is characterizing within-area exposure/confounder distributions
- Often one might have access to individual-level outcomes and individual-level confounder information, but only group-level exposure information
  - e.g., air pollution studies
- Referred to as a *semi-ecological* study
  - Kunzli and Tager (1997)
  - they use the term *semi-individual* study
- Certainly superior to a fully ecological design
- Only having ecological data for the exposure of interest remains a drawback
  - little work has been carried out to understand the implications

Study design summary

- Very useful categorization:
  - at which level do we have information on disease and exposure?

		Exposure	
		Individual	Ecological
Disease	Individual	*Individual*	*Semi-ecological*
	Ecological	*Aggregate*	*Ecological*

  - Sheppard (2003)
- *Individual* encompasess all the usual designs
  - randomized clinical trial, cohort study, case-control study, etc
- The *Aggregate* study (Prentice and Sheppard, 1995; Sheppard et al., 1996) assumes knowledge on the joint exposure/confounder distribution
  - obtained via sample survey
  - tackle within-area confounding

Infectious Diseases

- With Leigh Fisher, I've looked at ecological bias in the context of infectious diseases (Fisher and Wakefield, 2020).
- Suppose we have weekly (say) incident counts Y_t, in an area with a proportion vaccinated x.
- A naive model is

$$Y_{t+1}|Y_t \sim \text{Poisson}(\lambda Y_t \exp(-\beta x) + \delta),$$

  or

$$Y_{t+1}|Y_t \sim \text{Poisson}(\lambda Y_t (1-x)^{\alpha} + \delta),$$

  see for example Herzog et al. (2011).
- A model that respects the aggregation is

$$Y_{t+1}|Y_t \sim \text{Poisson}(\lambda Y_t (1-\phi x) + \delta).$$

Summary/concluding comments

Cheap and practical design

- Data availability (or lack thereof) is the primary motivation

Individual-level associations

- Reconcile the level of scientific interest with the level of analysis
  - if the level of interest is the group level then the ecological design is appropriate
- Fundamental problem of not being able to characterize within-area exposure/confounder distributions
  - cannot control for within-area confounding
  - cannot assess contextual effects
  - cannot perform adequate model checking
- Loss of information is analogous to unmeasured confounding

Summary/concluding comments

Statistical Considerations

- Key is to collect additional information
  - use of two-phase methods, e.g. Breslow and Chatterjee (1999), Ross and Wakefield (2013)
  - use of case-control samples within areas (Haneuse and Wakefield, 2007, 2008a,b)
- Multi-level modeling allows dependence/correlation at difference levels of the data
  - towards getting valid standard errors
  - cannot sort out ecological bias

Summary/concluding comments

Statistical Considerations

- Much recent work has focused on accounting for 'spatial' correlation
- typically of secondary importance (Wakefield, 2007)
- Doesn't address critical problems of confounding
- Wakefield and Smith (2016) is a recent review of statistical issues for an epidemiological audience

Summary

- Ecological studies can add to the totality of evidence, but alone are susceptible to a broad range of biases

References

Breslow, N. and N. Chatterjee (1999). Design and analysis of two-phase studies with binary outcomes applied to Wilms' tumor prognosis. *Applied Statistics 48*, 457–468.

Durkheim, E. (1897). *Le Suicide*. Paris: Alcan.

Fisher, L. and J. Wakefield (2020). Ecological inference for infectious disease data, with application to vaccination strategies. *Statistics in Medicine 39*, 220–238.

Greenland, S. (1992). Divergent biases in ecologic and individual-level studies. *Statistics in Medicine 11*, 1209–1223.

Greenland, S. and J. Robins (1994). Ecologic studies – Biases, misconceptions, and counterexamples (Disc: p761-771). *American Journal of Epidemiology 139*, 747–760.

Greenland, S., J. Robins, and J. Pearl (1999). Confounding and collapsability in causal inference. *Statistical Science 14*, 29–46.

Haneuse, S. and J. Wakefield (2007). Hierarchical models for combining ecological and case-control data. *Biometrics 63*, 128–136.

Haneuse, S. and J. Wakefield (2008a). The combination of ecological and case-control data. *Journal of the Royal Statistical Society, Series B 70*, 73–93.

Haneuse, S. and J. Wakefield (2008b). Geographic-based ecological correlation studies using supplemental case-control data. *Statistics in Medicine 27*, 864–887.

Herzog, S., M. Paul, and L. Held (2011). Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data. *Epidemiology and Infection 139*, 505–515.

Künzli, N. and I. B. Tager (1997). The semi-individual study in air pollution epidemiology: A valid design as compared to ecological studies. *Environmetal Health Perspectives 105*(10), 1078–1083.

Oakes, J. (2009). Commentary: Individual, ecological and multilevel fallacies. *International Journal of Epidemiology 38*, 361–368.

Prentice, R. L. and L. Sheppard (1995). Aggregate data studies of disease risk factors. *Biometrika 82*, 113–125.

Richardson, S. and C. Monfort (2000). Ecological correlation studies. In P. Elliott, J. Wakefield, N. Best, and D. Briggs (Eds.), *Spatial Epidemiology: Methods and Applications*, pp. 205–220. Oxford: Oxford University Press.

Robinson, W. (1950). Ecological correlations and the behaviour of individuals. *Am. Sociol. Rev. 15*, 351–357.

Ross, M. and J. Wakefield (2013). Bayesian inference for two-phase studies with categorical covariates. *Biometrics 69*, 469–477.

Selvin, H. (1958). Durkheim's 'suicide' and problems of empirical research. *American Journal of Sociology 63*, 607–619.

Sheppard, L. (2003). Insights on bias and information in group-level studies. *Biostatistics 4*, 265–278.

Sheppard, L., R. L. Prentice, and M. A. Rossing (1996). Design considerations for estimation of exposure effects on disease risk, using aggregate data studies. *Statistics in Medicine 15*, 1849–1858.

Simpson, E. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B 13*, 238–241.

Subramanian, S., K. Jones, A. Kaddour, and N. Krieger (2009a). Response: The value of a historically informed multilevel analysis of robinson's data. *International Journal of Epidemiology 38*, 379–373.

Subramanian, S., K. Jones, A. Kaddour, and N. Krieger (2009b). Revisiting Robinson: the perils of individualistic and ecologic fallacy. *International Journal of Epidemiology 38*, 342–360.

Wakefield, J. (2004). A critique of statistical aspects of ecological studies in spatial epidemiology. *Environmental and Ecological Statistics 11*, 31–54.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics 8*, 158–183.

Wakefield, J. (2008). Ecologic studies revisited. *Annual Review of Public Health 29*, 75–90.

Wakefield, J. (2009). Multi-level modelling, the ecologic fallacy, and hybrid study designs. *International Journal of Epidemiology 38*, 330–336.

Wakefield, J. and T. Smith (2016). Ecological modeling: general issues. In A. Lawson, S. Banerjee, R. Haining, and L. Ugarte (Eds.), *Handbook of Spatial Epidemiology*, pp. 112–130. CRC Press.