Confounding and Interaction in Aggregate-Level Studies: A Practical Guide

Presenter: W. Douglas Thompson, Ph.D.
University of Southern Maine

Co-author: Daniel Wartenberg, Ph.D.
University of Medicine and Dentistry of New Jersey

August 10, 2006

What is the appropriate role for studies of exposure-disease associations within the context of EPHT?

- Tracking of an environmental indicator is unlikely to be useful unless that indicator has been shown to be associated with an important health outcome

- Multi-state and national linked data sets should permit the study of associations that individual states have not been able to examine with adequate power

- Such studies may be conducted by collaborating health departments, by CDC, by the APEXes, or by independent researchers

- Once an association has been established, the impact of of new or enhanced interventions may be monitored by tracking both the environmental indicator and the health outcome, but linking of the data is no longer so important

Motivation for aggregate-level studies of exposure-disease associations

- Exposure data ***is not available*** at the individual level

- Information ***is available*** on the distribution of exposures within each of a series of geographically defined units (e.g., census blocks, municipalities, counties, states)

- Characterizing the spatial distribution of disease ***is not*** the focus of these analyses

- Interest is in effects of exposure on disease in individuals

Motivation for considering confounding and interaction

- The environmental exposure of interest is virtually never the only risk factor for the outcome under study
- Failure to consider other risk factors can lead to seriously biased estimates of the effect of exposure
- Exposure may have different effects in various subgroups of the population
- Interactive effects may have important implications for prevention

Important general point about aggregate-level analysis

- Many of the well-known methodologic principles from epidemiologic research based on individuals ***do not carry over*** to aggregate-level studies

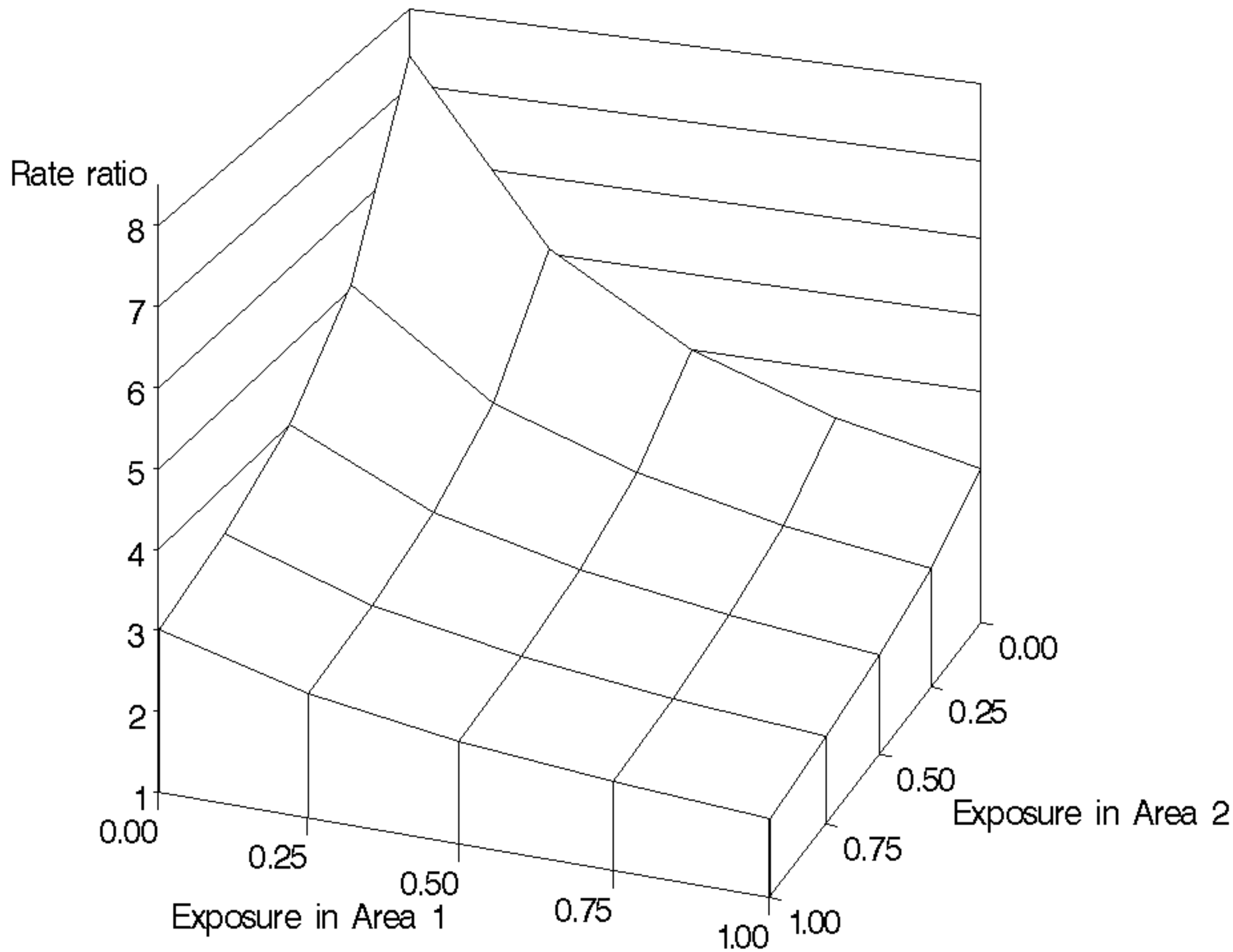
- These principles include the effects of misclassification and sampling error – i.e., the effects are quite different in aggregate-level studies

Yet another difference between analyses at the individual level and at the aggregate level

- At the individual level, both linear modeling of rates and log-linear modeling of rates can be used to estimate the pattern of rates for the various combinations of values for an environmental exposure and one or more covariates
- In aggregate-level analysis, log-linear analysis generally yields biased estimates

Example of bias when log-linear modeling is employed for aggregate-level analysis:

Two geographic areas with a disease rate of 30 per 1000 per year in the exposed and 10 per 1000 per year in the unexposed



Illustrative pattern of incidence rates* in 27 geographic units according to exposure and the presence/absence of a binary covariate: additivity for combined effects

Covariate +		Covariate -	
Exposure +	Exposure -	Exposure +	Exposure -
0.020	0.016	0.005	0.001

* Incidence rates are expressed per person per year

Illustrative data set at the individual level (n = 1000 in each unit)

Geographic unit	Covariate +				Covariate -			
	Exposure +		Exposure -		Exposure +		Exposure -	
	n	Expected number of events	n	Expected number of events	n	Expected number of events	n	Expected number of events
1	76	1.51	184	2.95	184	0.92	556	0.56
2	107	2.14	153	2.45	273	1.37	467	0.47
3	138	2.76	122	1.95	362	1.81	378	0.38
4	167	3.34	213	3.41	213	1.07	407	0.41
5	213	4.25	168	2.68	288	1.44	333	0.33
6	258	5.16	122	1.95	362	1.81	258	0.26
7	258	5.16	242	3.87	242	1.21	258	0.26
8	318	6.36	182	2.91	302	1.51	198	0.20
9	378	7.56	122	1.95	362	1.81	138	0.14
10	107	2.14	273	4.37	153	0.77	467	0.47
11	153	3.05	228	3.64	228	1.14	393	0.39
12	198	3.96	182	2.91	302	1.51	318	0.32
13	213	4.25	288	4.60	168	0.84	333	0.33
14	273	5.45	228	3.64	228	1.14	273	0.27
15	333	6.65	168	2.68	288	1.44	213	0.21
16	318	6.36	302	4.83	182	0.91	198	0.20
17	393	7.85	228	3.64	228	1.14	153	0.15
18	467	9.34	153	2.45	273	1.37	107	0.11
19	138	2.76	362	5.79	122	0.61	378	0.38
20	198	3.96	302	4.83	182	0.91	318	0.32
21	258	5.16	242	3.87	242	1.21	258	0.26
22	258	5.16	362	5.79	122	0.61	258	0.26
23	333	6.65	288	4.60	168	0.84	213	0.21
24	407	8.14	213	3.41	213	1.07	167	0.17
25	378	7.56	362	5.79	122	0.61	138	0.14
26	467	9.34	273	4.37	153	0.77	107	0.11
27	556	11.11	184	2.95	184	0.92	76	0.08

Running linear and log-linear individual-level Poisson regression models in SAS®

```
data individual;
input town covariate exposure n events;
product = covariate * exposure;
rate = events / n;
log_n = log(n);
datalines;
  1         1           1        76      1.51
  1         1           0       184      2.95
  1         0           1       184      0.92
  1         0           0       556      0.56
  2         1           1       107      2.14
  2         1           0       153      2.45
  2         0           1       273      1.37
                        0       467      0.47
 26         0                   138      2.76
 27         1           1
 27         1           0       184      2.95
 27         0           1       184      0.92
 27         0           0        76      0.08
run;
```

```
proc genmod data = individual;
title 'Individual-level: linear analysis';
model rate = covariate exposure product / d = poisson link = id;
weight n;
estimate 'cov = 1 exp = 1' intercept 1 covariate 1 exposure 1 product 1;
estimate 'cov = 1 exp = 0' intercept 1 covariate 1 exposure 0 product 0;
estimate 'cov = 0 exp = 1' intercept 1 covariate 0 exposure 1 product 0;
estimate 'cov = 0 exp = 0' intercept 1 covariate 0 exposure 0 product 0;
run;
```

```
proc genmod data = individual;
title 'Individual-level: log-linear analysis';
model events = covariate exposure product / d = poisson link = log offset = log_n;
estimate 'cov = 1 exp = 1' intercept 1 covariate 1 exposure 1 product 1 / exp;
estimate 'cov = 1 exp = 0' intercept 1 covariate 1 exposure 0 product 0 / exp;
estimate 'cov = 0 exp = 1' intercept 1 covariate 0 exposure 1 product 0 / exp;
estimate 'cov = 0 exp = 0' intercept 1 covariate 0 exposure 0 product 0 / exp;
run;
```

Results from linear analysis of illustrative individual-level data

The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	0.0010	0.0004	0.0003	0.0017	7.39	0.0066
covariate	1	0.0150	0.0017	0.0117	0.0182	82.03	► <.0001
exposure	1	0.0040	0.0010	0.0021	0.0059	16.84	► <.0001
product	1	-0.0000	0.0025	-0.0049	0.0049	0.00	► 0.9987
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Contrast Estimate Results

Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
cov = 1 exp = 1	► 0.0200	0.0016	0.05	0.0168	0.0232	147.13	<.0001
cov = 1 exp = 0	► 0.0160	0.0016	0.05	0.0128	0.0192	98.28	<.0001
cov = 0 exp = 1	► 0.0050	0.0009	0.05	0.0032	0.0068	30.76	<.0001
cov = 0 exp = 0	► 0.0010	0.0004	0.05	0.0003	0.0017	7.39	0.0066

Results from log-linear analysis of illustrative individual-level data

The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-6.9038	0.3679	-7.6248	-6.1828	352.23	<.0001
covariate	1	2.7681	0.3814	2.0205	3.5157	52.66 ►	<.0001
exposure	1	1.6065	0.4097	0.8035	2.4094	15.38 ►	<.0001
product	1	-1.3834	0.4299	-2.2259	-0.5408	10.36 ►	0.0013
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Contrast Estimate Results

Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
cov = 1 exp = 1	-3.9126	0.0824	0.05	-4.0742	-3.7511	2252.4	<.0001
Exp(cov = 1 exp = 1) ►	0.0200	0.0016	0.05	0.0170	0.0235		
cov = 1 exp = 0	-4.1357	0.1009	0.05	-4.3334	-3.9380	1681.0	<.0001
Exp(cov = 1 exp = 0) ►	0.0160	0.0016	0.05	0.0131	0.0195		
cov = 0 exp = 1	-5.2973	0.1803	0.05	-5.6507	-4.9440	863.18	<.0001
Exp(cov = 0 exp = 1) ►	0.0050	0.0009	0.05	0.0035	0.0071		
cov = 0 exp = 0	-6.9038	0.3679	0.05	-7.6248	-6.1828	352.23	<.0001
Exp(cov = 0 exp = 0) ►	0.0010	0.0004	0.05	0.0005	0.0021		

Estimated incidence rates according to type of analysis*

	Covariate +		Covariate -	
	Exposure		Exposure	
Type of analysis	+	-	+	-
Actual rates	0.020	0.016	0.005	0.001
Linear, individual level	0.020	0.016	0.005	0.001
Log-linear, individual level	0.020	0.016	0.005	0.001

* Incidence rates are expressed per person per year

Aggregate-level information for the same illustrative data set

Unit	Prevalence of covariate	Prevalence of exposure	Expected number of events	Incidence rate*
1	0.26	0.26	5.94	0.0059
2	0.26	0.38	6.42	0.0064
3	0.26	0.50	6.90	0.0069
4	0.38	0.38	8.22	0.0082
5	0.38	0.50	8.70	0.0087
6	0.38	0.62	9.18	0.0092
7	0.50	0.50	10.50	0.0105
8	0.50	0.62	10.98	0.0110
9	0.50	0.74	11.46	0.0115
10	0.38	0.26	7.74	0.0077
11	0.38	0.38	8.22	0.0082
12	0.38	0.50	8.70	0.0087
13	0.50	0.38	10.02	0.0100
14	0.50	0.50	10.50	0.0105
15	0.50	0.62	10.98	0.0110
16	0.62	0.50	12.30	0.0123
17	0.62	0.62	12.78	0.0128
18	0.62	0.74	13.26	0.0133
19	0.50	0.26	9.54	0.0095
20	0.50	0.38	10.02	0.0100
21	0.50	0.50	10.50	0.0105
22	0.62	0.38	11.82	0.0118
23	0.62	0.50	12.30	0.0123
24	0.62	0.62	12.78	0.0128
25	0.74	0.50	14.10	0.0141
26	0.74	0.62	14.58	0.0146
27	0.74	0.74	15.06	0.0151

* Expressed per person per year

Running linear and log-linear aggregate-level Poisson regression models in SAS®

```
data aggregate;
input town size covariate exposure events;
product = covariate * exposure;
rate = events / size;
log_size = log(size);
datalines;
  1     1000      0.26          0.26      5.94
  2     1000      0.26          0.38      6.42
  3     1000      0.26          0.50      6.90
  4     1000      0.38          0.38      8.22
                  0.38          0.50      8.70
                                0.62      9.18
 24     1000      0.62
 25     1000      0.74          0.50
 26     1000      0.74          0.62     14.58
 27     1000      0.74          0.74     15.06
run;
```

```
proc genmod data = aggregate;
title 'Aggregate-level: linear analysis';
model rate = covariate exposure product / d = poisson link = id;
weight size;
estimate 'cov = 1 exp = 1' intercept 1 covariate 1 exposure 1 product 1;
estimate 'cov = 1 exp = 0' intercept 1 covariate 1 exposure 0 product 0;
estimate 'cov = 0 exp = 1' intercept 1 covariate 0 exposure 1 product 0;
estimate 'cov = 0 exp = 0' intercept 1 covariate 0 exposure 0 product 0;
run;
```

```
proc genmod data = aggregate;
title 'Aggregate-level: log-linear analysis';
model events = covariate exposure product / d = poisson link = log offset = log_size;
estimate 'cov = 1 exp = 1' intercept 1 covariate 1 exposure 1 product 1 / exp;
estimate 'cov = 1 exp = 0' intercept 1 covariate 1 exposure 0 product 0 / exp;
estimate 'cov = 0 exp = 1' intercept 1 covariate 0 exposure 1 product 0 / exp;
estimate 'cov = 0 exp = 0' intercept 1 covariate 0 exposure 0 product 0 / exp;
run;
```

Results from linear analysis of illustrative aggregate-level data

The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	0.0010	0.0078	-0.0143	0.0163	0.02	0.8980
covariate	1	0.0150	0.0170	-0.0184	0.0484	0.77	0.3787
exposure	1	0.0040	0.0167	-0.0287	0.0367	0.06	0.8106
product	1	-0.0000	0.0334	-0.0655	0.0655	0.00	1.0000
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Contrast Estimate Results

Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
cov = 1 exp = 1	0.0200	0.0091	0.05	0.0022	0.0378	4.86	0.0275
cov = 1 exp = 0	0.0160	0.0099	0.05	-0.0035	0.0355	2.59	0.1074
cov = 0 exp = 1	0.0050	0.0096	0.05	-0.0137	0.0237	0.27	0.6009
cov = 0 exp = 0	0.0010	0.0078	0.05	-0.0143	0.0163	0.02	0.8980

Results from log-linear analysis of illustrative aggregate-level data

The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-5.8495	0.9205	-7.6536	-4.0453	40.38	<.0001
covariate	1	2.1433	1.8105	-1.4051	5.6918	1.40	0.2365
exposure	1	1.1025	1.8420	-2.5078	4.7127	0.36	0.5495
product	1	-1.3759	3.3943	-8.0286	5.2768	0.16	0.6852
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Contrast Estimate Results

Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
cov = 1 exp = 1	-3.9796	0.7980	0.05	-5.5436	-2.4155	24.87	<.0001
Exp(cov = 1 exp = 1)	0.0187	0.0149	0.05	0.0039	0.0893		
cov = 1 exp = 0	-3.7061	0.9618	0.05	-5.5913	-1.8210	14.85	0.0001
Exp(cov = 1 exp = 0	0.0246	0.0236	0.05	0.0037	0.1619		
cov = 0 exp = 1	-4.7470	0.9954	0.05	-6.6980	-2.7960	22.74	<.0001
Exp(cov = 0 exp = 1)	0.0087	0.0086	0.05	0.0012	0.0611		
cov = 0 exp = 0	-5.8495	0.9205	0.05	-7.6536	-4.0453	40.38	<.0001
Exp(cov = 0 exp = 0)	0.0029	0.0027	0.05	0.0005	0.0175		

Estimated incidence rates according to type of analysis*

	Covariate			
	+		-	
	Exposure		Exposure	
Type of analysis	+	-	+	-
Actual rates	0.020	0.016	0.005	0.001
Linear, individual level	0.020	0.016	0.005	0.001
Log-linear, individual level	0.020	0.016	0.005	0.001
Linear, aggregate level	0.020	0.016	0.005	0.001
Log-linear, aggregate level	0.019	0.025	0.009	0.003

* Incidence rates are expressed per person per year

Illustrative pattern of incidence rates* in 27 geographic units according to exposure and the presence/absence of a binary covariate: multiplicativity for combined effects

Covariate			
+		-	
Exposure		Exposure	
+	-	+	-
0.020	0.004	0.005	0.001

* Incidence rates are expressed per person per year

Estimated incidence rates according to type of analysis*

	Covariate			
	+		-	
	Exposure		Exposure	
Type of analysis	+	-	+	-
Actual rates	0.020	0.004	0.005	0.001
Linear, individual level	0.020	0.004	0.005	0.001
Log-linear, individual level	0.020	0.004	0.005	0.001
Linear, aggregate level	0.020	0.005	0.006	0.001
Log-linear, aggregate level	0.019	0.010	0.011	0.002

* Incidence rates are expressed per person per year

SUMMARY: Advice on conducting aggregate-level analysis with covariates

- Anticipate much lower power to detect effects than if individual-level data were available (particularly when there is little variability across areas in the prevalence of exposure, conditional on the covariates)

- Use linear rather than log-linear models to estimate rates of occurrence of disease events (Poisson regression performed on rates with the identity link function and using sample size as a weight variable)

- Anticipate some bias when the pattern of joint effects departs from additivity

- Evaluate interactions using product variables in the absence of information on the joint distribution of exposure and covariates within areas

e-mail contact for copy of presentation
and SAS code with data prototype:

dougt@usm.maine.edu