



UNIVERSITY of OULU
OULUN YLIOPISTO

Statistical issues related to use of aggregate level estimates

Esa Läärä

Department of Mathematical Sciences
University of Oulu, Finland

NOCCA Workshop, Oslo 3 to 5 December, 2007



Features of job-exposure matrix (JEM)

- ▶ Contains the estimated **proportion of exposed** P_{jk} and the **mean exposure level** L_{jk} among those exposed to agent k in occupational **group** j ($j = 1, \dots, J$; $k = 1, \dots, K$).
 - ▶ These **aggregate** measures are used as surrogates for the **true exposure level** x_{ijk} to agent k of an **individual** i working in occupation j ($i = 1, \dots, n_j$).
 - ▶ The measures are based on expert assessment & incomplete data and are thus prone to **measurement error**.
 - ▶ **Variability** and **covariation** in true exposure levels of different agents and confounders within any occupational group are ignored.
- ⇒ Estimates on the health effects (like hazard rate ratios, **RR**) of the true exposure at target level may be biased.



Special features in NOCCA

- ▶ Exceptionally big study population and long follow-up.
 - ▶ Exceptionally high observed numbers of cases.
 - ▶ Statistical precision or “power” exceptionally high.
- ⇒ Exceptionally narrow confidence intervals & a lot of “statistically significant” results on SIRs and other comparative measures.
- ▶ The high nominal precision completely ignores uncertainty concerning the magnitude of the bias due to various sources in the estimation of interesting hazard rate ratios (RR).
 - ▶ In relative terms this uncertainty is most likely essentially greater than the nominal standard errors capturing the “pure” random variation.



JEM and missing data methods in statistics

- ▶ From a statistical perspective, using JEM is like applying **mean imputation** to fill in the missing individual values.
- ▶ It is known that mean imputation
 - underestimates the true variability and uncertainty in the data, and thus leads to exaggerated precision in the estimates of interesting parameters,
 - in certain circumstances may introduce bias to these estimates, especially on associations and effects.
- ▶ An alternative approach – **multiple imputation (MI)**:
Missing individual values are substituted by several simulated values randomly sampled from an **imputation model**, based on relevant auxiliary information about the individuals and study variables.



JEM and missing data methods in statistics

JEM studies:

- ▶ Imputation is performed on everybody, because individual exposure data are typically missing from all subjects in an occupational cohort.
 - ▶ Relevant auxiliary data needed in an imputation model, provided by informative covariates, is typically unavailable (apart from occupational group).
- ⇒ Multiple imputation is only possible, if based on strong prior assumptions (expert assessment) on the variation and covariation of the relevant exposures and confounders.



Ecological studies in epidemiology

- ▶ In **ecological studies** the observational units are typically communes (*kommun*) or other regional entities.
- ▶ Types of explanatory variables:
 - ▶ **aggregate** measures (proportions, means) summarized from individual-level risk factor values
 - ▶ **environmental** measures: assessments of average exposures (proportions, means), like in JEM,
 - ▶ **global** features of the whole unit – **contextual variables** (like type of specific health care policy).
- ▶ JEM studies are **partially ecologic** or **semi-individual**:
 - ▶ some individual-level variables are available (like age and gender) and controllable
- ▶ Methodology for ecological studies is thus relevant for JEMs.



Famous ecological associations

- ▶ Abundance of storks in a commune and birth rate.
- ▶ Proportion of protestants in a province and suicide rate (Durkheim 1897).
- ▶ Sales of cigarettes per capita in a country and lung cancer incidence.
- ▶ Average fat consumption in a country and breast cancer incidence.
- ▶ *etc.*

How are these associations at the individual level?



Proportion protestants and suicide rate

In four groups of Prussian provinces 1883–90 the proportions of protestants P_j (%) in the population and the suicide mortality rates R_j (per 10^5 person-years) were as follows

	Group			
	1	2	3	4
Prop'n protestants (%)	30	45	78	95
Suicide rate (per 10^5 y)	10	16	22	26

Fitted linear regression line $R_j = 3.7 + 24.0 \times P_j$

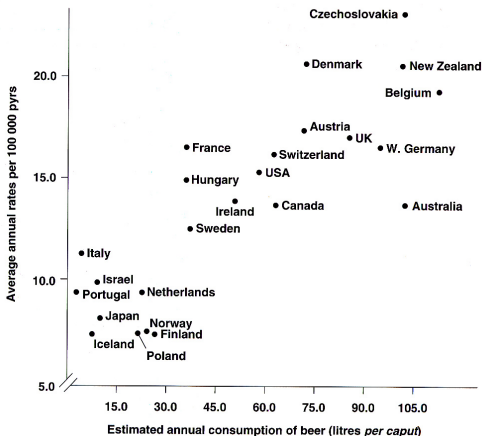
\Rightarrow ecological estimate of rate ratio $RR^* = 1 + 24.0/3.7 = 7.6$.

Interpretation?



Figure 11.17 (IS, p. 259)

Beer consumption and rectal cancer incidence in 24 countries
1960-62. (No latent period taken into account.)





Ecological bias or fallacy

The association estimated from commune level between an aggregate exposure measure and disease occurrence has

different strength and maybe opposite direction

than that estimated from individual data in a cohort or case-control study. – This bias may have various sources

- ▶ within-group variability of ind'l exposure levels about the mean value representing exposure for the whole group,
- ▶ non-linearity of the exposure effect on disease rate,
- ▶ effect modification and/or confounding at individual level (hidden within-group covariation between exposure factors),
- ▶ aggregate-level confounding, absent at individual level,
- ▶ measurement errors, latent periods, etc.



Binary exposure

- ▶ If individual data were available in each aggregate unit or group j , the association between exposure and disease at the individual level could be assessed.
- ▶ Occurrences (incidences or prevalences) could be obtained in “pure” exposure groups, and comparative measures, too.

	exposed	unexposed	total
cases	D_{1j}	D_{0j}	D_{+j}
group size	N_{1j}	N_{0j}	N_{1j}
occurrence	$R_{1j} = \frac{D_{1j}}{N_{1j}}$	$R_{0j} = \frac{D_{0j}}{N_{0j}}$	$R_{+j} = \frac{D_{+j}}{N_{+j}}$

- ▶ Proportion of exposed $P_j = N_{1j}/N_{+j}$.
- ▶ Total occurrence $R_{+j} = P_j \times R_{1j} + (1 - P_j) \times R_{0j}$
= mixture of occurrences in “pure” exposure groups.



Aggregated data & binary exposure

- ▶ Ecological study: Available are only overall occurrences R_j , and proportions exposed P_j in each aggregate unit j but not occurrences R_{1j} & R_{0j} in “pure” exposure groups.
- ▶ Analysis of ecological data:
Linear or exponential regression between P_j and R_j variables.
- ▶ Question:
Does the association observed at the cluster level reflect well that of exposure and outcome at the individual level?
- ▶ Answer: Sometimes “YES” but often “NO”!



Example: Ecological data in 3 groups

Exposure	Group 1		Group 2		Group 3	
	cases	p-years	cases	p-years	cases	p-years
yes ($E = 1$)	?	N_{11}	?	N_{12}	?	N_{13}
no ($E = 0$)	?	N_{01}	?	N_{02}	?	N_{03}
Total	D_{+1}	N_{+1}	D_{+2}	N_{+2}	D_{+3}	N_{+3}

In each group j ($j = 1, 2, 3$)

- ▶ proportion exposed to E is $P_j = N_{1j}/N_{+j}$, and
- ▶ the overall disease rate is $R_j = D_{+j}/N_{+j}$.

These are used in analysis at the aggregate level.



Example: Ecological association

In a study with 3 groups the following results were obtained.

Exposure	Group 1		Group 2		Group 3	
	cases	p-years	cases	p-years	cases	p-years
yes ($E = 1$)	?	2000	?	3000	?	4000
no ($E = 0$)	?	4000	?	3000	?	2000
Total	60	6000	100	6000	150	6000
Rate (/100 y)	1.0		1.7		2.5	
Exposed %	33		50		67	

- ▶ The association between the proportion exposed P_j and the disease rate R_j is positive, and the ecological rate ratio estimate is $RR^* = 3.7$, based on a multiplicative model.



Example (cont'd)

- ▶ Suppose the person-years were as before but the missing data on cases by exposure within groups were as follows:

Exposure	Group 1		Group 2		Group 3	
	cases	rate/10 ² y	cases	rate/10 ² y	cases	rate/10 ² y
yes ($E = 1$)	20	1.0	25	1.7	50	2.5
no ($E = 0$)	40	1.0	50	1.7	100	2.5
Total	60	RR = 1	100	RR = 1	150	RR = 1
Rate (/100 y)	1.0		1.7		2.5	
Exposed %	33		50		67	

- ▶ In each group the rate ratio between exposure and disease at the individual level is $RR = 1.0$ – no association.
- ▶ This is in clear contrast with the ecological estimate 3.7.



Example (cont'd)

The missing cells in the aggregate data could also have been according to the following table:

Exposure	Group 1		Group 2		Group 3	
	cases	rate/10 ² y	cases	rate/10 ² y	cases	rate/10 ² y
yes ($E = 1$)	30	1.5	67	2.2	120	3.0
no ($E = 0$)	30	0.75	33	1.1	30	1.5
Total	60	RR = 2	100	RR = 2	150	RR = 2
Rate (/100 y)	1.0		1.7		2.5	
Exposed %	33		50		67	

- ▶ Here, the rate ratio between exposure and disease at the individual level is $RR = 2.0$ – exposure seems harmful.
- ▶ The direction of the association is the same but the ecological estimate exaggerates the individual-level RR.



Example (cont'd)

Third possible within-group configuration:

Exposure	Group 1		Group 2		Group 3	
	cases	rate/10 ² y	cases	rate/10 ² y	cases	rate/10 ² y
yes ($E = 1$)	12	0.6	33	1.1	75	1.9
no ($E = 0$)	48	1.2	67	2.2	75	3.8
Total	60	RR = 0.5	100	RR = 0.5	150	RR = 0.5
Rate (/100 y)	1.0		1.7		2.5	
Exposed %	33		50		67	

- ▶ Now, within each group the rate ratio is $RR = 0.5$
– exposure beneficial?!
- ▶ A striking conflict between the individual and ecological estimates as to the direction of association.



Counfounding at aggregate level

- ▶ In each of the three instances the aggregate level association remained strongly positive.
 - ▶ Yet, the association at the individual level could be either positive, null, or even negative.
 - ▶ The baseline disease rate was associated with the proportion exposed across the aggregate units.
- ⇒ We have confounding at the aggregate level, even though no confounding at the individual level existed.
- ▶ The direction of association at the individual level could not be predicted from the aggregate results.



Binary exposure and binary covariate

- ▶ Consider now again disease and binary exposure E , but also a **covariate** C , also with values 1 = 'yes', and 0 = 'no'.
- ▶ In ecologic data each group j has its own 3-way table of the mutual associations between C , E and disease.

	$C = 1$		$C = 0$		Total	
	cases	p-years	cases	p-years	cases	p-years
$E = 1$?	?	?	?	?	N_{1j}
$E = 0$?	?	?	?	?	N_{0j}
Total	?	M_{1j}	?	M_{0j}	D_{+j}	N_{+j}

- ▶ Proportion exposed to E is $P_j = N_{1j}/T_j$ and proportion exposed to C is $Q_j = M_{1j}/T_j$.
- ▶ Overall disease rate is $R_j = D_{+j}/N_{+j}$.

These are used in an aggregate-level modelling and analysis.



Within-group confounding and modification

- ▶ Let C be a determinant of disease. At ind'l level C is
 - (a) **confounder**, if C and E are associated within groups,
 - (b) **effect-modifier**, if the effect of E differs between levels of C ,
- ▶ Ignoring these possibilities may induce bias at group level.
- ▶ Ecologic bias may also result, even if C were not confounder within groups, if its prevalence Q_j varies across groups. Then, namely, the *background rate* (in unexposed) R_{0j} of disease varies over $j \Rightarrow$ group becomes confounder.
- ▶ Group as a modifier may induce ecologic bias, too.
- ▶ Group-level adjustment by Q_j may increase bias!



Ecological regression and rate ratio

- ▶ Linear model for aggregate rates R_j on proportions exposed

$$R_j = a^* + b^* P_j$$

Approximation of the log-linear (multiplicative) model.

- ▶ Estimation of ecological rate ratio $RR^* = 1 + b^*/a^*$.
- ▶ Ecological adjustment for C : add term $c^* Q_j$ to the model.
Estimation of RR^* at given level Q_j of prop'n exposed to C :

$$RR^* = \frac{a^* + b^* + c^* Q_j}{a^* + c^* Q_j}$$

- ▶ These methods are used in subsequent examples taken from Greenland, S., Morgenstern, H. (1989) Ecological bias, confounding, and effect modification. *IJE* **18**: 269-274.



Ex 1: Smoking & oesophageal cancer

No confounding, but positive – and overall quite strong – rate ratio is modified by group.

Group		Smoking Yes	Smoking No	Rate ratio	Group rate	Exposed P_j (%)
1	Rate (/10 ⁵ y)	12	3	4.0	7.5	50
	P-years (1000s)	100	100			
2	Rate (/10 ⁵ y)	15	3	5.0	7.8	40
	P-years (1000s)	80	120			
3	Rate (/10 ⁵ y)	20	3	6.7	8.1	30
	P-years (1000s)	60	140			
Total	Rate (/10 ⁵ y)	15	3	5.0		
	Size (1000s)	240	360			

Ecological association negative:

$$R_j = 9 - 3 \times P_j \Rightarrow RR^* = 1 + (-3)/9 = 0.67!$$



Ex 2: Radon, smoking, and lung cancer

True $RR = 2$ for radon ('high' vs. 'low'); $RR = 10$ for smoking.

Three regions ("groups"), no association between exposures by group, but proportion of smokers varies.

Grp		Smokers		Non-smokers		Group rate	% exposed	
		High	Low	High	Low		radon	smok.
1	Rate	200	100	20	10	69.3	26	50
	$N (10^3y)$	26	74	26	74			
2	Rate	200	100	20	10	62.1	35	40
	$N (10^3y)$	28	52	42	78			
3	Rate	200	100	20	10	55.5	50	30
	$N (10^3y)$	30	30	70	70			

- Ecological regression on radon only $\rightarrow RR^* = 0.3$ - heavy bias.
- Ecological adjustment for smoking $\rightarrow RR^* = 1.0$ - still biased.



Ex 3: Alcohol, smoking & oesophageal ca.

True RR is 5 for alcohol, 'yes' vs. 'no', and 5 for smoking, too.

Alcohol and smoking associated within groups.

Group		Smokers		Non-smok.		Group rate	% exposed	
		Yes	No	Yes	No		alcohol	smoking
1	Rate	25	5	5	1	8.6	40	50
	Size	50	50	30	70			
2	Rate	25	5	5	1	8.9	40	40
	Size	59	21	21	99			
3	Rate	25	5	5	1	9.2	49	36
	Size	60	12	38	90			

Ecological association of cancer weakly positive with alcohol but negative with smoking!



Ex 3 (cont'd)

Analysis of individual level data:

- ▶ Crude RR for alcohol 9.2 – biased.
- ▶ Stratification by smoking yields unbiased $RR = 5$.

Analysis of aggregate data by ecological regression:

- ▶ on proportion exposed to alcohol only:

$$R_j = 6.8 + 5.0 \times P_j \Rightarrow RR^* = 1 + 5.0/6.8 = 1.7$$

– biased due to differential within-group confounding.

- ▶ adjusted for prop'n smokers: $R_j = 9.3 + 2.0 \times P_j - 3.0 \times Q_j$

$$\Rightarrow RR^* = \frac{9.3 + 2.0 - 3.0 \times 0.42}{9.3 - 3.0 \times 0.42} = 1.2$$

- ▶ Ecological adjustment increases bias!



Implications of aggregation to modelling

- ▶ In the special case of a simple additive linear model for rates (or risks) based on exposures at the individual level, the regression coefficients keep their interpretation even after aggregation.
 - ▶ Yet, most statistical models applied in epidemiology have a non-linear form, e.g. exponential or log-linear (like the Poisson and Cox regression).
 - ▶ Also, realistic models often contain interaction terms describing effect modification.
- ⇒ Seemingly similar group-level model based on aggregate exposure variables is not equal to the corresponding individual-level model, and the coefficients do not remain the same – **model misspecification**.



Binary exposure: model for rate ratio

Let x_{ij} indicate exposure (1/0) to agent X for ind'l i in group j .

Multiplicative model (Cox or Poisson) for the **hazard rate** of disease at the *individual-level*

$$r_{ij} = r_0 \times \exp(bx_{ij})$$

- ▶ r_0 = baseline rate (function of sex, age and period),
- ▶ $\exp(bx_{ij})$ = relative rate function of exposure,
- ▶ $\exp(b) = \text{RR}$; rate ratio associated with exposure.



Aggregated model for group rate

Model for *aggregate-level* or **average rate** in group j :

$$R_j = R_{0j} \times \exp(bP_j) \times \text{mean}[\exp(bd_{ij})],$$

- ▶ P_j is the proportion exposed in j ,
- ▶ $d_{ij} = x_{ij} - P_j$ are individual deviations from mean exposure,
- ▶ $\text{mean}[\exp(bd_{ij})]$ is a factor causing **specification bias** into the log-linear aggregate model.
- ▶ Taking logarithms and expanding $\log\{\text{mean}[\exp(bd_{ij})]\}$:

$$\begin{aligned} \log R_j &= \log R_{0j} + bP_j \\ &\quad + b^2 \text{var}(x_j)/2 + b^3 \text{skew}(x_j)/6 + b^4 \text{kurt}(x_j)/24 + \dots \end{aligned}$$



Aggregate-level rate

- ▶ Apart from mean exposure, the aggregate rate depends on the **variance, skewness, kurtosis**, etc. of the exposure distribution within group, and on $b = \log(\text{RR})$ *non-linearly*.
- ▶ Log-linear model for aggregate rates:

$$\log R_j = \log R_{0j}^* + b^* P_j$$

- ▶ Ecological-level $\text{RR}^* = \exp(b^*)$ IS NOT equal to the target $\text{RR} = \exp(b)$ unless $b = 0$, or x_{ij} is constant within groups.
- ▶ The bias in b^* depends on the size of b & variability of X .
- ▶ Approximation: linear “ecological regression model”:

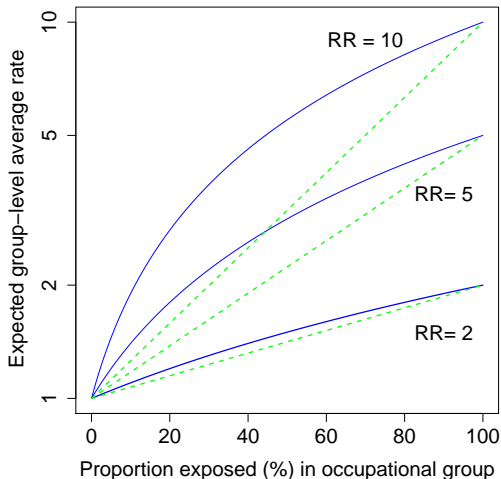
$$R_j = a^* + b^* P_j$$

Simple estimation of ecological rate ratio $\text{RR}^* = 1 + b^*/a^*$.



Example: Single binary exposure

Average rate as a function of proportion exposed in a group when true individual-level rate ratio is 2, 5 or 10.





Example: Single binary exposure

- ▶ Individual-level rate ratio for exposed: $RR = 2$, or $RR = 5$,
- ▶ 6 occupational groups,
- ▶ proportions of exposed (P_j , %) – four scenarios with different between-group and within-group variations:
 - 1: 0, 5, 10, 90, 95, 100 – wide btw, narrow within
 - 2: 0, 20, 40, 60, 80, 100 – uniform btw, variable withn
 - 3: 10, 20, 30, 30, 40, 50 – narrow btw, mostly L-skew withn
 - 4: 50, 60, 70, 70, 80, 90 – narrow btw, mostly R-skew withn

Expected ecological rate ratios $RR^* = e^{b^*}$:

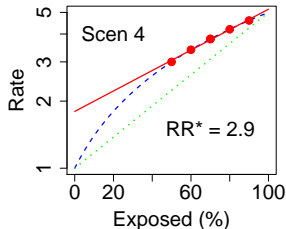
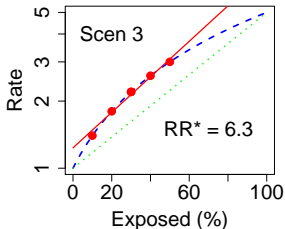
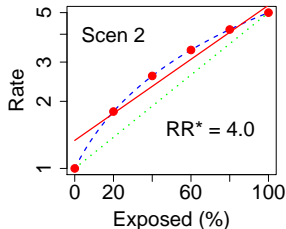
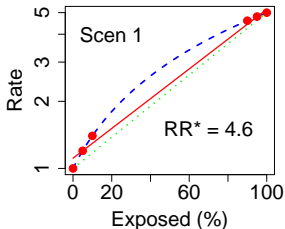
true RR	scen. 1	scen. 2	scen. 3	scen. 4
2	1.99	1.97	2.16	1.80
5	4.65	4.18	6.26	2.87

Influenced by between/within variability & skewness



Binary exposure, true $RR = 5$

Fitted rates (red lines) by prop'n exp'd (red dots) in 4 scenarios.





Dose-response models – single exposure

Let x_{ij} be the exposure level to agent X for individual i in job j .

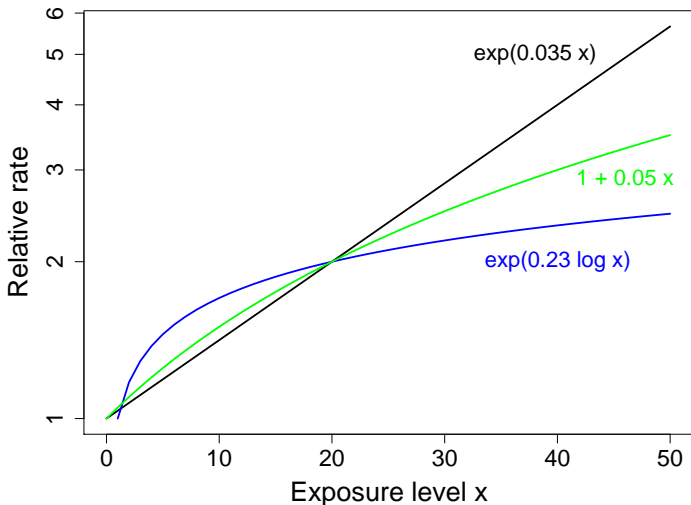
Class of multiplicative models for rates, where baseline rate r_0 (function of sex, age etc.) is combined with relative rate function

$$r_{ij} = r_0 \times rr(b, x_{ij}),$$

- ▶ **proportional hazards model:** $rr(b, x_{ij}) = \exp(bx_{ij})$, in which $\exp(b) = RR$ is the rate ratio associated with unit increase in X (not dependent on the level of X) – most popular model.
- ▶ **additive relative rate model:** $rr(b, x_{ij}) = (1 + bx_{ij})$ in which $RR(X_1, X_0) = (1 + bX_1)/(1 + bX_0)$ is the rate ratio associated with exposure contrast X_1 vs. X_0 .
- ▶ other, depending on the form of $rr(b, x_{ij})$.



Three relative rate functions





Aggregate-level rate

Model for *aggregate-level* or average rate R_j in group j :

$$R_j = r_0 \times rr(b, \bar{x}_j) \times \text{mean}[rr(b, d_{ij})],$$

where baseline rate r_0 and rr function as before, and

- ▶ $\bar{x}_j = P_j L_j$ is the mean exposure level,
- ▶ $d_{ij} = x_{ij} - \bar{x}_j$ are individual deviations from the mean, and
- ▶ $\text{mean}[rr(b, d_{ij})]$ is a factor causing *specification bias* into the aggregate model based on mean exposure \bar{x}_j only.
 - PH model: $\text{mean}[rr(b, d_{ij})] = \exp[K(d_{ij}, b)]$ where $K(d_{ij}, b)$ is the **cumulant-generating function** of d_{ij} s,
 - ARR model: $\text{mean}[rr(b, d_{ij})]$ cancels out!

(Often more realistic to assume group-specific baseline R_{0j} as a *random term*, to allow for nonspecific differences across groups.)



Proportional hazards model aggregated

- ▶ Taking logarithms and expanding $K(d_{ij}, b)$ we get

$$\log R_j = \log R_{0j} + b\bar{x}_j + b^2 \text{var}(x_j)/2 + b^3 \text{skew}(x_j)/6 + b^4 \text{kurt}(x_j)/24 + \dots$$

- ▶ Apart from mean exposure, the aggregate rate depends on the *variance*, *skewness*, *kurtosis*, etc. of the exposure distribution within group, and on $b = \log(\text{RR})$ *non-linearly*.
- ▶ Simple log-linear model for aggregate rates

$$\log R_j = \log R_{0j}^* + b^* \bar{x}_j$$

- ▶ Ecological $\text{RR}^* = \exp(b^*)$ IS NOT equal to individual-level $\text{RR} = \exp(b)$ unless $b = 0$, or x_{ij} is constant within groups.
- ▶ The bias in b^* depends on the size of b and the variance of X .



Quantitative exposure – Gaussian model

- ▶ Suppose the within-group distribution of exposure X were Gaussian with expectation L_j and standard deviation S_j .
- ▶ Covers in particular models with exposure effect specified as $b \log(z_{ij})$ for a log-normally distributed exposure Z .

- ▶ Realistic to have S_j positively dependent on L_j .
- ▶ True PH model for mean rate simplifies into

$$\log R_j = \log R_{0j} + bL_j + b^2 S_j^2 / 2$$

- ▶ Group-level $RR^* = e^{b^*}$ in the mean rate log-linear model would always be larger than individual-level $RR = e^b$. Sometimes $b^* > 0$ even when $b < 0$.
- ▶ The bias depends on the magnitude of S_j s and b .
- ▶ When S_j s are known, estimation of true RR is easy.



Quantitative exposure – skew distribution

- ▶ Many exposures have right-skewed distributions.
- ▶ Common model: **log-normal distribution** with geometric standard deviation $GSD \approx 2.5 = \sqrt{6.25}$. Coupled with mean L_j , we have

- ▶ geometric mean $G_j \approx 0.66 \times L_j$, and
- ▶ standard deviation $S_j \approx L_j \times \sqrt{e^{6.25} - 1}$.

⇒ 95% of values lie within range $G_j/6.25 - G_j \times 6.25$.

- ▶ Heavily skewed to the right, and positive kurtosis, too.
- ▶ *Mathematical anomaly*: $K(d_{ij}, b)$ is unbounded. Thus, specification bias cannot be evaluated analytically. Simulations become unstable, too.



Alternatives for pure log-normal model

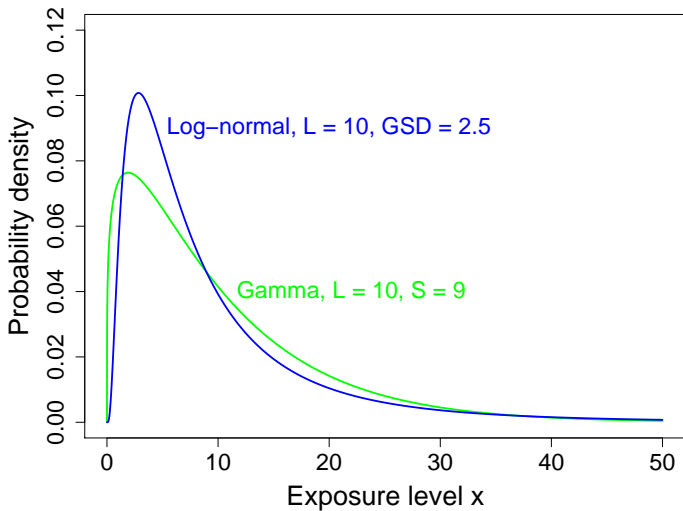
(A) **Trimmed** log-normal: The range of variation truncated at a suitably high percentile point. Simulations more reliable(?).

(B) **Gamma** distribution

- ▶ Right-skewed, but the tail is not so long and heavy.
- ▶ Characterized by two parameters: shape a and scale c .
- ▶ Mean $L = a/c$, standard deviation $S = \sqrt{ac^2}$, coefficient of variation $1/\sqrt{a}$.
- ▶ Finite expression for $K(d_{ij}, b)$ exists, hence specification bias may be evaluated analytically.
- ▶ Could be tried for approximating log-normal.
- ▶ However, geometric standard deviation does not so simply determine percentile points as in log-normal.
- ▶ More research is needed!



Log-normal and gamma with same mean





Zero-inflated skew distribution

- ▶ The proportion $1 - P_j$ of unexposed ($x_{ij} = 0$) may be substantial in many jobs j .
- ▶ Let X be distributed with mean L_j and $GSD \approx 2.5$, or suitable standard deviation S_j among those exposed, their proportion being P_j .
- ▶ The total exposure distribution is a *mixture* of constant zero component and variable non-zero component.
- ▶ Target dose-response model is not affected by the mixture.
- ▶ However, the aggregate rate will be dependent on the mixture proportions, apart from the form of the non-zero component, and its specification bias factor may be complicated.



Ex: Zero-inflated gamma in groups

Simulation with following assumptions

- ▶ Target model: PH with $rr(b, x_{ij}) = \exp(0.035x_{ij})$ implying $RR=2$ for $X = 20$ vs. 0 (see a previous figure)
- ▶ 6 job groups.
- ▶ Non-zero part: $L_j = 6, 8, \dots, 16$ and $S_j = 0.9L_j$.
- ▶ Proportions exposed P_j , 3 scenarios
 - 1: 0.0, 0.2, 0.4, 0.6, 0.8, 1.0 – proportional to L_j
 - 2: 1.0, 0.8, 0.6, 0.4, 0.2, 0.0 – negatively related to L_j
 - 3: 0.4, 0.8, 0.0, 1.0, 0.6, 0.2 – uncorrelated with L_j

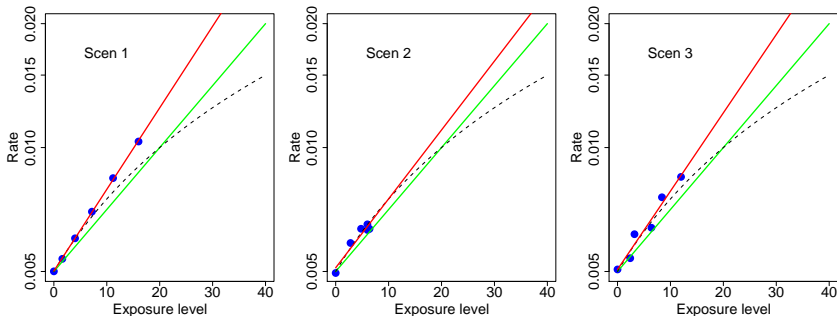
Results on ecologic RR^*

true RR	scen. 1	scen. 2	scen. 3
2	2.55	2.10	2.33



Simulation results

True model, green; fitted model, red; additive rr model, dashed



- ▶ In all scenarios the slope is exaggerated.
- ▶ The bias is largest when P_j and L_j positively correlated, and smallest with negative correlation.
- ▶ The range of $P_j L_j$ always limited compared to that of X
- ▶ Results most imprecise with small variation in $P_j L_j$



Example: radon, smoking, and lung cancer

(Greenland & Robins (1994) in AJE 139: 747-760)

Suppose true lung cancer hazard $r(x, z)$ (per 10^5 y) depends on

- ▶ x = radon exposure (pCi/l), and
- ▶ z = cigarettes smoked (packs/day, values 0, 1, 2)

according to the following multiplicative model

$$r(x, z) = 40 \times (1 + 0.2x) \times \exp(0.1z)$$

in which the rr functions are different for radon and smoking.

More concretely, following relative rates are implied:

- ▶ $RR = 2.0$ for radon 5 pCi/l vs. 0,
- ▶ $RR = 7.4$ for 1 pack smoked vs. 0.



Aggregate rates in regions

- ▶ Let Q_{zj} be prop'n smoking z packs/d in region j ; $z = 0, 1, 2$.
- ▶ Average lung cancer rate in region j by mean radon level \bar{x}_j :

$$R(\bar{x}_j) = 40 \times (1 + 0.2\bar{x}_j) \times (Q_{0j} + Q_{1j}e^2 + Q_{2j}e^4) \\ \times M(x_{ij}, z_{ij})$$

where the last factor is a complicated function of the joint distribution of radon and smoking within region.

- ▶ Suppose we have 41 regions $j = 0, 1, 2, \dots, 40$.
- ▶ Radon exposure x_{ij} is assumed constant x_j within any region j but varies across regions. Then factor $M(\cdot, \cdot)$ vanishes.
- ▶ Further, radon levels are determined simply by region index $x_j = 0.1 + 0.3\sqrt{j}$.



Scenario 1: no confounding across regions

- ▶ Mean numbers of packs/d smoked \bar{z}_j in regions would not be associated with regional radon levels x_j .
- ▶ Obtained by simulating smoking distributions such that the prop'ns (%) smoking 0, 1 or 2 packs/d in region j are generated from
 - ▶ $p_{0j} = 53 - 0.2j + 1.5u$ → expected p_{0j} 45 to 53 %
 - ▶ $p_{1j} = 34 + 0.4j - 1.0u$ → expected p_{1j} 34 to 50 %
 - ▶ $p_{2j} = 13 - 0.2j - 0.5u$ → expected p_{2j} 5 to 13 %

where u is std normal random variate.

- ▶ Aggregate level log-linear model estimated:

$$R_j = \exp(5.1 - 0.093x_j + 0.094\bar{z}_j)$$



Scenario 1: results

- ▶ Estimates of ecologic rate ratios
 - ▶ $\exp(0.093 \times 5) = 0.63$ for radon 5 pCi/l vs. 0,
 - ▶ $\exp(0.094 \times 20) = 6.6$ for smoking 1 pack vs. 0.
- ▶ For smoking the ecologic estimate is not far from truth.
- ▶ For radon, the direction of estimated effect changes. Why?
 - ▶ Strong ecologic association between radon and smoking distributions not adequately summarized by the relation of mean smoking level with radon level.
 - ▶ Strong non-linear effect of smoking.
- ▶ Conclusion: ecologic control of smoking based on only mean number of packs/d is ineffective at removing confounding.



Scenario 2: confounding across regions

- ▶ Mean numbers of packs/d smoked \bar{z}_j in regions would be positively correlated with regional radon levels x_j . Ecologic correlation $\approx +0.3$.
- ▶ Obtained by simulating smoking 0 and 1 packs/d as follows:
 - ▶ $p_{0j} = 53 - 0.25j + 1.5u$ → expected p_{0j} 43 to 53 %
 - ▶ $p_{1j} = 34 + 0.45j - 1.0u$ → expected p_{1j} 34 to 52 %
- ▶ Estimates of ecologic rate ratios for radon 5 pCi/l vs. 0,
 - ▶ 0.66 when unadjusted for smoking,
 - ▶ 0.58 when adjusted for mean smoking.
- ▶ Additional bias from adjustment occurs because mean smoking is positively correlated with radon and cancer, so it can only increase the already downward bias.



Conclusions

- ▶ Aggregate measures of occupational exposures are associated with an increased hazard of biased estimates of hazard ratios at individual level.
- ▶ The risk of bias is directly associated with the relative amount of variability of true exposure within groups to that between groups.
- ▶ Attempts to control for confounding by simple aggregate measures (like % smoking) may at worst increase the bias.
- ▶ In aggregate analysis even non-differential misclassification often leads to a bias away from the null.
- ▶ This problem is pronounced with surrogates of confounders (like SIRs/SMRs for lung cancer or liver diseases).



Conclusions (cont'd)

- ▶ More detailed data on the marginal and joint distributions of key exposures and confounders within-groups dispersion – even based on small samples – would be helpful in reducing the bias in aggregate analysis.
- ▶ Without reasonable information on variability within-groups due caution must be exercised when interpreting RR estimates based on aggregate data – especially in light of their misleadingly high nominal precision.



Selected references

- ▶ Gilks, W., Richardson, S. (1992). *Stat Med* **11**, 1443-
- ▶ Greenland, S., Robins, J. (1994). *Am J Epidemiol* **139**, 747-
- ▶ Jackson C., et al. (2006). *Stat Med* **25**, 2136-
- ▶ Morgenstern, H. (1995). *Annu Rev Public Health* **16**,61-
- ▶ Morgenstern, H. (2008). in *Modern Epidemiology* (Rothman et al., eds.), 511-
- ▶ Richardson, S., Best, N. (2003). *Environmetrics* **14**, 129-
- ▶ Wakefield, J., Salway, R. (2001). *J R Statist Soc* **164**, 119-
- ▶ Webster, T.F. (2007). *Environ Health* **6**, 17