Ottawa Hospital Research Institute
Institut de recherche de l'Hôpital d'Ottawa

uOttawa

CHEO RESEARCH INSTITUTE

Analyzing multiple exposures

Franco Momoli

Children's Hospital of Eastern Ontario Research Institute
Ottawa Hospital Research Institute
School of Epidemiology, Public Health, and Preventive Medicine
(University of Ottawa)

SPER workshop
Denver, June 2015

www.ohri.ca Affiliated with • Affilié à

Outline

- Examples, in brief:
  - Witte, 1994: Diet and breast cancer risk
  - Steenland, 2000: Occupations and cancer incidence
  - Momoli, 2010: Occupational chemicals and lung cancer risk
- Traditional teaching on modelling (and why it doesn't nicely apply)
- Hierarchical Modelling
- Types of research questions
- To Bayes or not to Bayes?
- The ideas behind hierarchical modelling
- Selection, shrinkage, exchangeability
- Three issues to overcome
  - Multiple inference and doing too much
  - Mutual confounding and "over-adjustment"
  - Small-sample bias and big "elephant-like" models
- Two-stage empirical Bayes and Semi-Bayes

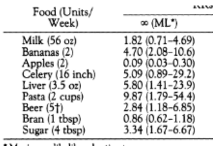
- The setting:
  - You have a study
  - In this study are multiple 'exposures' and you are interested in each one. You want, at the end of the day, an estimate of each exposure's effect on some outcome
  - You may need to identify some exposures for further study
  - You are considering a set of exposures
  - They are not merely a nuisance to you. This is not the problem of having one exposure of interest and too many candidate confounders
  - Exposures are often correlated. Sometimes highly correlated.

* I'll use 'exposures' generically throughout

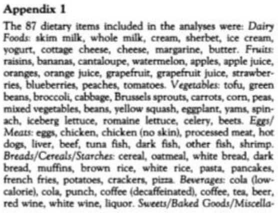
Example 1: Diet and breast cancer

Data for this application came from a case-control study of premenopausal bilateral breast cancer. Formal details of the study and data have been given elsewhere. We had complete dietary and hormonal information on 140 cases and 222 controls; controls were sisters of the cases.

140/10 ~ 14 parameters estimable



Food (Units/Week)	∞ (ML*)
Milk (56 oz)	1.82 (0.71–4.69)
Bananas (2)	4.70 (2.08–10.6)
Apples (2)	0.09 (0.03–0.30)
Celery (16 inch)	5.09 (0.89–29.2)
Liver (3.5 oz)	5.80 (1.41–23.9)
Pasta (2 cups)	9.87 (1.79–54.4)
Beer (5†)	2.84 (1.18–6.85)
Bran (1 tbsp)	0.86 (0.62–1.18)
Sugar (4 tbsp)	3.34 (1.67–6.67)

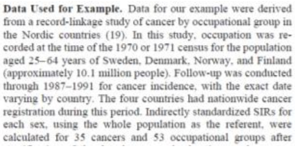


Appendix 1

The 87 dietary items included in the analyses were: *Dairy Foods:* skim milk, whole milk, cream, sherbet, ice cream, yogurt, cottage cheese, cheese, margarine, butter. *Fruits:* raisins, bananas, cantaloupe, watermelon, apples, apple juice, oranges, orange juice, grapefruit, grapefruit juice, strawberries, blueberries, peaches, tomatoes. *Vegetables:* tofu, green beans, broccoli, cabbage, Brussels sprouts, carrots, corn, peas, mixed vegetables, beans, yellow squash, eggplant, yams, spinach, iceberg lettuce, romaine lettuce, celery, beets. *Eggs/Meats:* eggs, chicken, chicken (no skin), processed meat, hot dogs, liver, beef, tuna fish, dark fish, other fish, shrimp. *Breads/Cereals/Starches:* cereal, oatmeal, white bread, dark bread, muffins, brown rice, white rice, pasta, pancakes, french fries, potatoes, crackers, pizza. *Beverages:* cola (low-calorie), cola, punch, coffee (decaffeinated), coffee, tea, beer, red wine, white wine, liquor. *Sweets/Baked Goods/Miscella-*

Witte et al., Hierarchical Regression Analysis Applied to a Study of Multiple Dietary Exposures and Breast Cancer. *Epidemiology,* Vol. 5, No. 6 (Nov., 1994), pp. 612-621.

Example 2: Occupations and cancer incidence



Data Used for Example. Data for our example were derived from a record-linkage study of cancer by occupational group in the Nordic countries (19). In this study, occupation was recorded at the time of the 1970 or 1971 census for the population aged 25–64 years of Sweden, Denmark, Norway, and Finland (approximately 10.1 million people). Follow-up was conducted through 1987–1991 for cancer incidence, with the exact date varying by country. The four countries had nationwide cancer registration during this period. Indirectly standardized SIRs for each sex, using the whole population as the referent, were calculated for 35 cancers and 53 occupational groups after

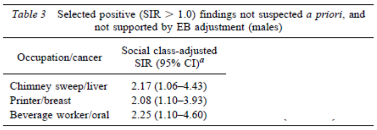


Table 3 Selected positive (SIR > 1.0) findings not suspected *a priori*, and not supported by EB adjustment (males)

Occupation/cancer	Social class-adjusted SIR (95% CI)[a]
Chimney sweep/liver	2.17 (1.06–4.43)
Printer/breast	2.08 (1.10–3.93)
Beverage worker/oral	2.25 (1.10–4.60)

Steenland et al., Empirical Bayes Adjustments for Multiple Results in Hypothesis-generating or Surveillance Studies. Cancer Epidemiology, Biomarkers & Prevention Vol. 9, 895–903, September 2000

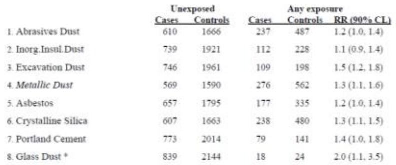
Example 3: Occupational exposures and lung cancer

Case-control study in Montreal 1979-1985

857 cases of lung cancer, 2172 controls

184 occupational chemicals

857/10 ~ 85 variables with rule-of-thumb??



	Unexposed		Any exposure		
	Cases	Controls	Cases	Controls	RR (90% CL)
1. Abrasives Dust	610	1666	237	487	1.2 (1.0, 1.4)
2. Inorg.Insul.Dust	739	1921	112	228	1.1 (0.9, 1.4)
3. Excavation Dust	746	1961	109	198	1.5 (1.2, 1.8)
4. *Metallic Dust*	569	1590	276	562	1.3 (1.1, 1.6)
5. Asbestos	657	1795	177	335	1.2 (1.0, 1.4)
6. Crystalline Silica	607	1663	238	480	1.3 (1.1, 1.5)
7. Portland Cement	773	2014	79	141	1.4 (1.0, 1.8)
8. Glass Dust *	839	2144	18	24	2.0 (1.1, 3.5)

Momoli et al., Analysis of Multiple Exposures: An Empirical Comparison of Results From Conventional and Semi-Bayes Modeling Strategies. Epidemiology 2010;21: 144–151).

Traditional teaching on modelling (and why it doesn't nicely apply)

- By "modelling" I refer narrowly to selecting which confounders to include in a model
- Traditionally are taught in epi to be concerned with one exposure
- When the objective is the assessment of the effects of multiple k exposures, analyses typically:

1. Fit a 'full' model, though this is rare because people are worried about putting too many variables in a model (and estimates can be very "unstable")
2. Use estimates from a 'reduced' model (usually by a testing algorithm)
3. An intermediate (most common?) approach that treats each exposure with its own model, usually with testing for confounders

** Great for publications if you do one exposure per manuscript!

The issues with conventional approaches and any kind of selection (especially via "P-value testing"):

- We are interested in each exposure's effect, yet we sometimes treat the exposures as nuisance variables (1-at-a-time approach)
- Standard errors of estimates for remaining variables are downwardly biased after selection
- Those variables removed have no confidence interval attached to their de facto 0
  - Each estimate within an arbitrary 1.96 standard deviations is set to 0
- It creates an almost philosophical contradiction
- Only a perfunctory consideration of mutual confounding (confounding can aggregate)

Freedman et al. Return to a note on screening regression equations. Am Stat 1989; 43: p.279-282.

Why consider hierarchical modelling?

- We are going to attempt to avoid selecting variables into and out of our model
- We are going to build a single model that reflects the complexity of the scientific/policy problem we are addressing
- Epidemiologic literature, however, also represents it as:
  - a way to model many effects simultaneously, accounting for mediating factors
  - a way of improving estimation
  - a way of dealing with problems of multiple inference
- This is based on ideas from Stein, empirical Bayes, Hierarchical Bayes

Robbins (1956). "An Empirical Bayes Approach to Statistics". Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics: 157–163.

Good (1980) Some history of the hierarchical Bayesian methodology. Trabajos de Estadistica Y de Investigacion Operativa 31 (1), pp 489-519

These are within the family of hierarchical modelling:

- Mixed models
- Multi-level models
- Random-effects models
- Random-coefficient models
- Variance components
- Two-stage models
- Random regression
- Penalized likelihood
- Ridge regression
- Stein estimation
- Bayesian, empirical-Bayes estimation
- Special cases (1 level)
  - Least squares estimation
  - Maximum likelihood estimation

When would hierarchical methods be useful?

1. We have a set of exposures of interest
2. Effects/exposures can be grouped by similarity
3. Random error/variability is an issue
4. There is a need to prioritize future work

Greenland et al., S., and Robins, J. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. Epidemiology, 2: 244–251, 1991.

Types of research questions

- Inference problems can be distinguished:
  - Single inference questions

  (is this exposure related to this illness?)
  - Multiple inference questions

  (are any exposures in this set related to this illness?)
    - The latter question evokes the universal null hypothesis (no exposures in this set are related to the illness) and an omnibus alternative hypothesis (at least one exposure in this set is related to the illness)
    - Criticized as a poor research strategy for most epidemiologic problems
  - The hierarchical (empirical Bayes) viewpoint maintains individual items but with the qualification that they belong to a set

  (which of this list of exposures is related to this illness?)

  (which are harmful?)

Rothman (1990) No adjustments are needed for multiple comparisons. Epidemiology 1: 43-46.

Modern Epidemiology, 3rd edition. Multiple Comparisons, p.234-237

To Bayes or not to Bayes?

- Showing your Bayesian colours
  - I suspect most of us are Bayesian in spirit but not in practice
  - You can discuss hierarchical models with Bayesian terminology (priors, posteriors) or with a more Frequentist terminology (random and fixed effects, multilevel)
  - Hierarchical models unify Frequentist and Bayesian methods, allowing us to incorporate external information (in a palatable way for Frequentists)
- Programming
  - Fully Bayesian programming (BUGS, etc.) using Markov-Chain Monte-Carlo algorithms is more accurate, and allows for more flexible solutions
  - However, sometimes accuracy is an illusion, especially in typical epi studies
  - Simplifications and approximations may suffice
  - Several teaching examples using SAS and R available (https://github.com/wittelab)

Greenland, Principles of multilevel modelling. *International Journal of Epidemiology* 2000; **29**:158–167.

Brief history of empirical Bayes models

- Recognizable form found in Gini (1911)
- Named EB by Robins ~ 1955
- Applications in other fields and in theoretical statistics throughout ’70s
- Empirical Bayes and parametric options, based on Stein estimator
  - ○ although the individual EB-adjusted estimates are not statistically unbiased, the average squared error of the adjusted estimates will be less than the average squared error of the original estimates
- Mainstream Epi literature with Thomas (1985), as a means of addressing problems of multiple inference
- Refined largely by Greenland in the 90s (semi Bayes)

Morris. Parametric empirical Bayes inference: theory and application. J. Am Stat Assoc 1983;78:47-65.

Thomas et al., The problem of multiple inference in studies designed to generate hypotheses. Am J Epidemiol 1985; 122:1080-1095

The ideas behind empirical Bayes

- The objective is to avoid selection and improve estimation
- We improve our model of all the exposures of interest by:
  - Pulling implausible estimates back to a more plausible range
  - Adding “external” information about properties that might mediate the effects of those exposures
- All first-level variables are modelled through fewer second-level variables

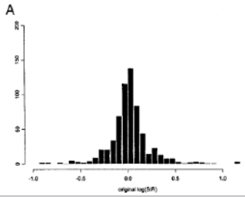
Logit R= B**X** + G**W** + A First stage (conventional)
B = P**Z** + D Second stage

EB estimate = **B**(1-W) + **P**ZW

- This is based on the idea by Stein (we weight our prior guesses at effect estimates with their agreement from observed data, resulting in lower squared error than MLE)

Selection or shrinkage?

- By pulling the estimates back toward the null, you are “anticipating regression to the mean”
- Given that an estimate is among the largest in the set, it is more likely to be an overestimate than an underestimate
  - P-value selection does not address these estimation issues



Modern Epidemiology, 3rd edition, Hierarchical (multilevel) regression, pg. 435-439

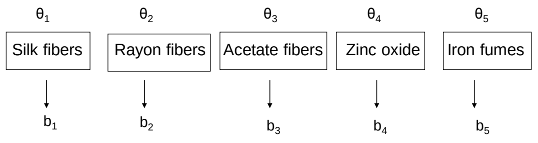
Steenland et al., Empirical Bayes Adjustments for Multiple Results in Hypothesis-generating or Surveillance Studies. Cancer Epidemiology, Biomarkers & Prevention Vol. 9, 895–903, September 2000

Exchangeability

- But we can go further
  - We can pull estimates – not to the null – but to the overall center of the distribution of estimates we are considering in our particular study
- We can also pull estimates closer to each other when the exposures are similar … this is the assumption of exchangeability
- Modelling similarities among the exposures of interest
  - Diet items and food constituents (if two diet items share the same nutrient quantities, perhaps we would believe that their effects on health should be similar … more similar than to another diet item with entirely different nutrient quantities)
  - Occupational agents and chemical properties
  - SNPs and function in genetic studies

- Given our lack of knowledge about the effects of many of these exposures, *exchangeability* refers to …
  - "… the *belief* that certain sets of exposures likely have similar effects on a health outcome."
- But being ignorant as to further distinctions within the set.
- This should (?) be backed up with some evidence.
- Correlated hypotheses:
  - This makes sense conceptually. If one of the exposures in a set is found to be causal, your suspicion of the effect of other exposures in that set might increase.
- It sounds very Bayesian. But it also appeals to Frequentists because you don't need to set prior distributions on parameters for this to work.

Illustrative diagram of exchangeability (Example 3: Momoli 2010)

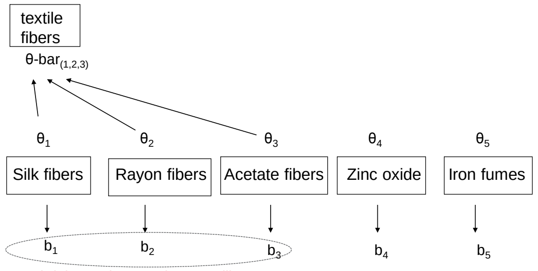


θ_1 θ_2 θ_3 θ_4 θ_5

Silk fibers | Rayon fibers | Acetate fibers | Zinc oxide | Iron fumes

b_1 b_2 b_3 b_4 b_5

Illustrative diagram of exchangeability (Example 3: Momoli 2010)



textile fibers

θ-bar$_{(1,2,3)}$

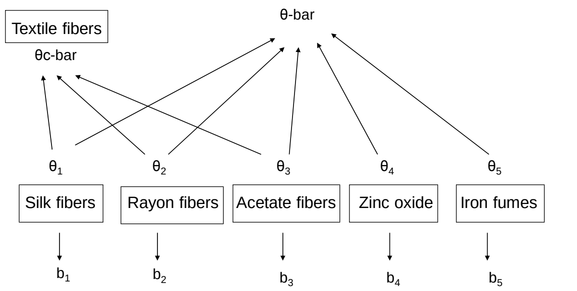
θ_1 θ_2 θ_3 θ_4 θ_5

Silk fibers | Rayon fibers | Acetate fibers | Zinc oxide | Iron fumes

b_1 b_2 b_3 b_4 b_5

Shrinkage: these estimates will be pulled closer to each other

Illustrative diagram of exchangeability (Example 3: Momoli 2010)



Shrinkage: And all estimates might be pulled closer to the overall mean (depending on modelling decisions)

Three issues to overcome: 1. Multiple inference and doing too much

- Is "multiple comparisons" an issue or not?
  - Some say yes
  - But if considering one exposure, looking at 99 others is immaterial
- It's a real problem, but not in the way it's usually presented
- Sometimes consideration of other exposures suggests confounder adjustment; sometimes it calls into question validity of the study.
- Normal corrections (e.g., Bonferroni) sacrifice power to maintain type-I error rate across the study
- "borrowing strength from the ensemble" – improving estimation

Greenland. Multiple comparisons and association selection in general epidemiology, Int J Epidemiology 2008, 37(3): 430-434.

From testing to estimation

- Bonferroni reasoning
  - 1-0.95^k false positives expected if global null hypothesis true
  - What is k? In a publication? Over a career? In an entire field of study?
- Hierarchical modelling approach eschews the testing approach, replacing it with a shrinkage approach
- Notice the shift from testing to estimation
  - Correction of P-values does nothing of use for us. Point estimates remain unchanged.
- The semi-Bayes "paradigm" is an effort to improve estimation.
  - It DOES NOT penalize you for doing too much in your study (which is nonsensical).
  - It penalizes individual estimates (pulling them closer to the null) if they have very wide confidence intervals or if they are very (implausibly) large

Three issues to overcome: 2. Mutual confounding and "over-adjustment"

- Selection-based methods do a poor job unless a conservative criterion is used (e.g., alpha 0.15)
- But with modern computing power and evidence to-date for hierarchical models, there may be no need for selection
- Shrinkage is often preferable both conceptually and empirically
- Some will object to models that "adjust for everything"
  - Misapply term "over-adjustment" when they likely mean "unnecessary adjustment"
  - But unnecessary adjustment affects precision, and hierarchical models address this directly

Schisterman et al. (2009) Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies. Epidemiology 20(4): 488-495.

Three issues to overcome:
3. Small-sample bias and big "elephant-like" models *

- Small sample bias is a well known bias in maximum likelihood estimates when the sample size to parameter ratio is small. This is typically an "away from the null" bias
- Another reason to consider pulling estimates back towards the null when there is much variability
- It may seem that hierarchical models are fitting too many parameters
  - Bayesian models can fit more parameters than data points
  - The second-level model in fact relaxes the assumptions
  - It builds dependency among first-level coefficients

Cordeiro et al., Bias correction in generalized linear models. Journal of the Royal Statistical Society, Series B (Methodological), 53 (3) 1991, p629-643.

* From quote attributed to Savage [Greenland (2000) When should epidemiologic regressions use random coefficients? Biometrics 56: 915-921].

- The dependency among parameters is the key
- It reduces the necessary degrees of freedom via exchangeability assumptions
  - e.g., you can estimate 86 diet item effects with only 73.5 model "effective degrees of freedom"

- Modeling dependency among parameters "frees up" degrees-of-freedom for more precise estimation of their variances, and it avoids over-fitting a model to sparse data when there are many effects being estimated simultaneously.

Behaviour of the empirical Bayes estimator

- EB estimate = **B**(1-W) + **P**Z(W)

- For betas that are imprecise, the EB estimate will move toward the estimated prior mean (borrowing strength from the ensemble)
- If the assumption of exchangeability was a good one (the estimated betas within a set are close together), then the betas will move toward the estimated prior mean
- Otherwise the estimates move toward the 1st stage (ML) estimate

---

The elements of a two-stage EB approach

- B, Vector of beta coefficients from conventional model
- V, covariance matrix from 1-stage model
- Z-matrix, prior or structural information <u>about</u> 1st stage variables

Steps:

1. Estimation of 1-stage parameters (logistic) — Logit R= B**X** + G**W** + A
2. Define 2-stage variables (categories of exchangeability)
3. Regress 1-stage estimates on 2-stage variables (with weighted least-squares regression — B = P**Z** + D
4. Model calculates prior means and variances
5. Model average 1st and 2nd stage estimates to derive EB "posterior" estimates — **B* = B**(1-W) + **P**ZW

Data augmentation: Z-matrix (aka, incomplete prior, prior structural information, categories of exchangeability) (Example 3: Momoli 2010)

	Chemical properties			
	Fibrous inorganic dusts	*Si-containing compounds*	*Metal oxide dusts*	*Previous evidence Log OR*
Alumina	0	0	1	0
Silica	0	1	0	0.18
Asbestos	1	1	0	0.1
Diesel exhaust	0	0	0	0.64
Titanium dioxide	0	0	1	0
Phosgene	0	0	0	0

Z-Matrix (Example 1: Diet and breast cancer)

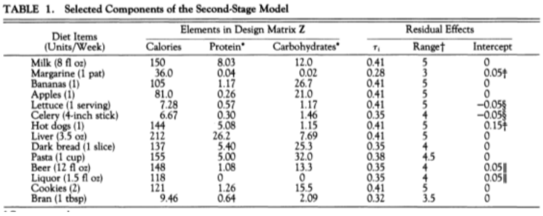


TABLE 1. Selected Components of the Second-Stage Model

Diet Items (Units/Week)	Elements in Design Matrix Z			Residual Effects		
	Calories	Protein*	Carbohydrates*	τ_i	Range†	Intercept
Milk (8 fl oz)	150	8.03	12.0	0.41	5	0
Margarine (1 pat)	36.0	0.04	0.02	0.28	3	0.05†
Bananas (1)	105	1.17	26.7	0.41	5	0
Apples (1)	81.0	0.26	21.0	0.41	5	0
Lettuce (1 serving)	7.28	0.57	1.17	0.41	5	−0.05§
Celery (4-inch stick)	6.67	0.30	1.46	0.35	4	−0.05§
Hot dogs (1)	144	5.08	1.15	0.41	5	0.15†
Liver (3.5 oz)	212	26.2	7.69	0.41	5	0
Dark bread (1 slice)	137	5.40	25.3	0.35	4	0
Pasta (1 cup)	155	5.00	32.0	0.38	4.5	0
Beer (12 fl oz)	148	1.08	13.3	0.35	4	0.05‖
Liquor (1.5 fl oz)	118	0	0	0.35	4	0.05‖
Cookies (2)	121	1.26	15.5	0.41	5	0
Bran (1 tbsp)	9.46	0.64	2.09	0.32	3.5	0

Witte et al., Hierarchical Regression Analysis Applied to a Study of Multiple Dietary Exposures and Breast Cancer. *Epidemiology,* Vol. 5, No. 6 (Nov., 1994), pp. 612-621.

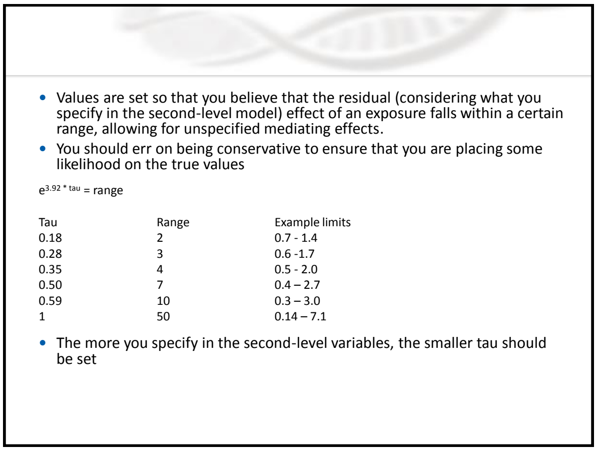
Semi-Bayes modelling

- A variant of empirical Bayes modelling
  - (naïve) Classical Bayes: Specify prior means and variances
  - Hierarchical Bayes: Fully specify hyperpriors on hyperparameters
  - Empirical Bayes: estimate prior means and variances from prior "structural" information
  - Semi-Bayes: estimate prior means, prior variances specified by investigator

SB models have been shown to typically outperform MLE and EB

Specifying tau, the prior standard deviation

- Tau is a smoothing parameter that affects the weighting when averaging the first- and second-stage estimates
- Residual effects: assumption is that any effect of first level variables on outcome NOT ACCOUNTED FOR in the second level model likely falls in a narrow range of values
  - e.g., conservatively assumed with 95% probability that residual effect of eating 4-inch celery stick per week would lead to a relative change in risk between 0.5 and 2. This is very conservative. Most would assign a much tighter range; that is, most would think one celery stick a week would not affect breast cancer risk at all.
- Making that range very big, leads us back to the MLE approach
  - But is this a plausible assumption? That any value is as likely as any other? So, an odds ratio of 1000 is as likely as an odds ratio of 1.5 for one celery stick a week?

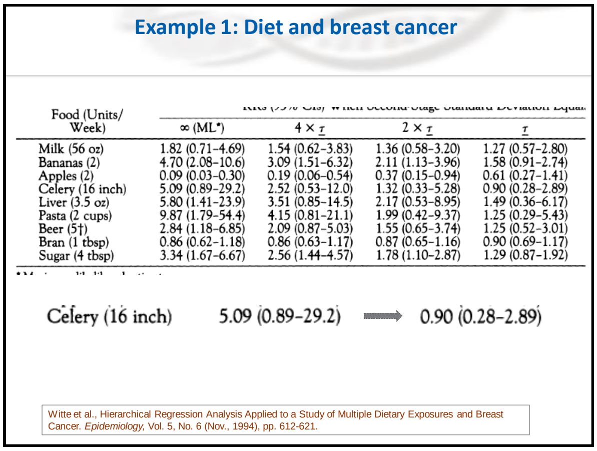


- Values are set so that you believe that the residual (considering what you specify in the second-level model) effect of an exposure falls within a certain range, allowing for unspecified mediating effects.
- You should err on being conservative to ensure that you are placing some likelihood on the true values

$e^{3.92 * tau}$ = range

Tau	Range	Example limits
0.18	2	0.7 - 1.4
0.28	3	0.6 -1.7
0.35	4	0.5 - 2.0
0.50	7	0.4 – 2.7
0.59	10	0.3 – 3.0
1	50	0.14 – 7.1

- The more you specify in the second-level variables, the smaller tau should be set

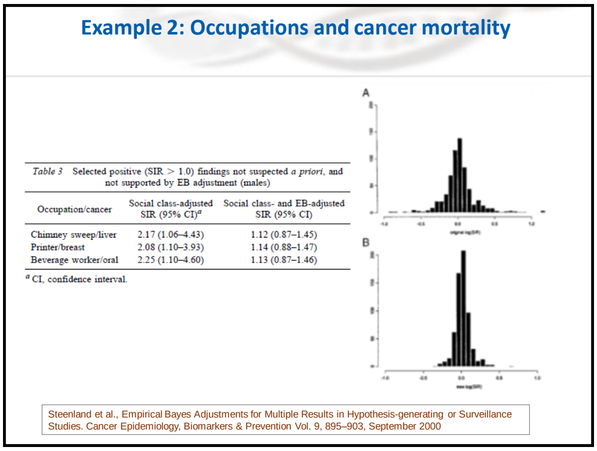


Example 1: Diet and breast cancer

Food (Units/ Week)	RRs (95% CIs) When Second-Stage Standard Deviation Equal			
	∞ (ML*)	4 × $\underline{\tau}$	2 × $\underline{\tau}$	$\underline{\tau}$
Milk (56 oz)	1.82 (0.71–4.69)	1.54 (0.62–3.83)	1.36 (0.58–3.20)	1.27 (0.57–2.80)
Bananas (2)	4.70 (2.08–10.6)	3.09 (1.51–6.32)	2.11 (1.13–3.96)	1.58 (0.91–2.74)
Apples (2)	0.09 (0.03–0.30)	0.19 (0.06–0.54)	0.37 (0.15–0.94)	0.61 (0.27–1.41)
Celery (16 inch)	5.09 (0.89–29.2)	2.52 (0.53–12.0)	1.32 (0.33–5.28)	0.90 (0.28–2.89)
Liver (3.5 oz)	5.80 (1.41–23.9)	3.51 (0.85–14.5)	2.17 (0.53–8.95)	1.49 (0.36–6.17)
Pasta (2 cups)	9.87 (1.79–54.4)	4.15 (0.81–21.1)	1.99 (0.42–9.37)	1.25 (0.29–5.43)
Beer (5†)	2.84 (1.18–6.85)	2.09 (0.87–5.03)	1.55 (0.65–3.74)	1.25 (0.52–3.01)
Bran (1 tbsp)	0.86 (0.62–1.18)	0.86 (0.63–1.17)	0.87 (0.65–1.16)	0.90 (0.69–1.17)
Sugar (4 tbsp)	3.34 (1.67–6.67)	2.56 (1.44–4.57)	1.78 (1.10–2.87)	1.29 (0.87–1.92)

Celery (16 inch) 5.09 (0.89–29.2) ⟹ 0.90 (0.28–2.89)

Witte et al., Hierarchical Regression Analysis Applied to a Study of Multiple Dietary Exposures and Breast Cancer. *Epidemiology,* Vol. 5, No. 6 (Nov., 1994), pp. 612-621.



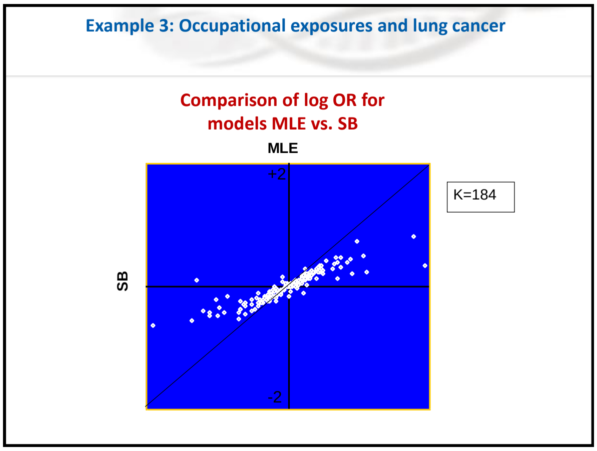
Example 2: Occupations and cancer mortality

Table 3 Selected positive (SIR > 1.0) findings not suspected *a priori*, and not supported by EB adjustment (males)

Occupation/cancer	Social class-adjusted SIR (95% CI)[a]	Social class- and EB-adjusted SIR (95% CI)
Chimney sweep/liver	2.17 (1.06–4.43)	1.12 (0.87–1.45)
Printer/breast	2.08 (1.10–3.93)	1.14 (0.88–1.47)
Beverage worker/oral	2.25 (1.10–4.60)	1.13 (0.87–1.46)

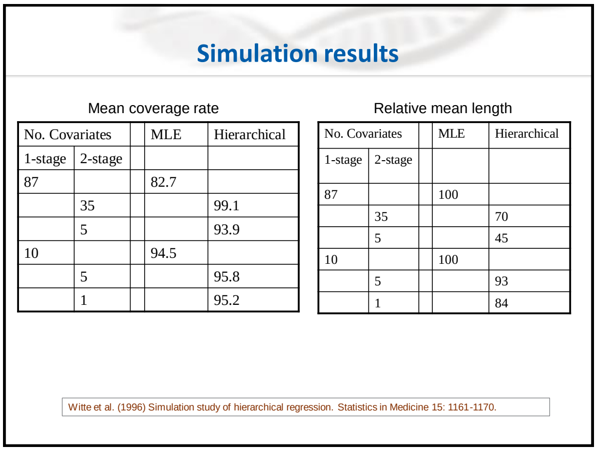
[a] CI, confidence interval.

Steenland et al., Empirical Bayes Adjustments for Multiple Results in Hypothesis-generating or Surveillance Studies. Cancer Epidemiology, Biomarkers & Prevention Vol. 9, 895–903, September 2000



Example 3: Occupational exposures and lung cancer

Comparison of log OR for models MLE vs. SB



Simulation results

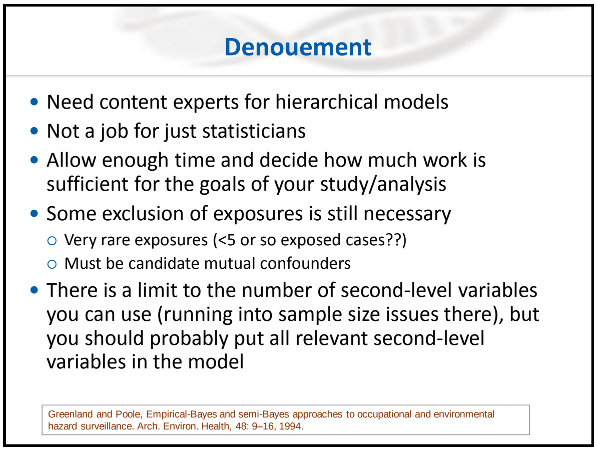
Mean coverage rate

No. Covariates			MLE	Hierarchical
1-stage	2-stage			
87			82.7	
	35			99.1
	5			93.9
10			94.5	
	5			95.8
	1			95.2

Relative mean length

No. Covariates			MLE	Hierarchical
1-stage	2-stage			
87			100	
	35			70
	5			45
10			100	
	5			93
	1			84

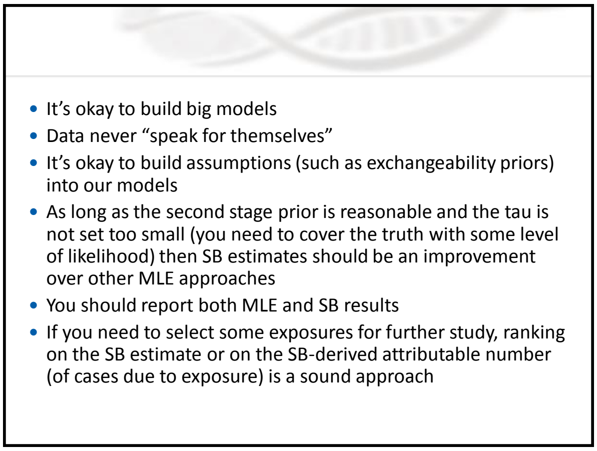
Witte et al. (1996) Simulation study of hierarchical regression. Statistics in Medicine 15: 1161-1170.



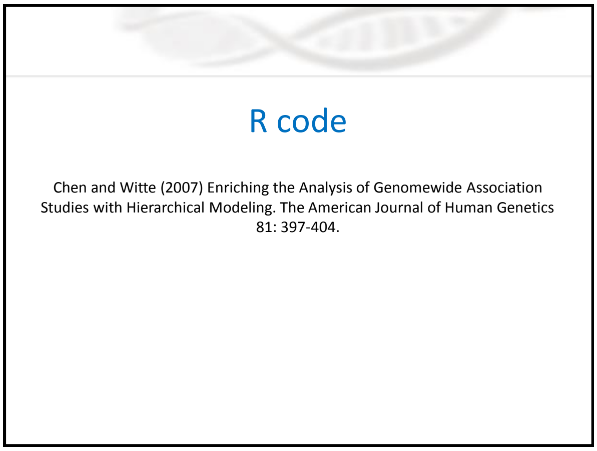
Denouement

- Need content experts for hierarchical models
- Not a job for just statisticians
- Allow enough time and decide how much work is sufficient for the goals of your study/analysis
- Some exclusion of exposures is still necessary
  - Very rare exposures (<5 or so exposed cases??)
  - Must be candidate mutual confounders
- There is a limit to the number of second-level variables you can use (running into sample size issues there), but you should probably put all relevant second-level variables in the model

Greenland and Poole, Empirical-Bayes and semi-Bayes approaches to occupational and environmental hazard surveillance. Arch. Environ. Health, 48: 9–16, 1994.



- It's okay to build big models
- Data never "speak for themselves"
- It's okay to build assumptions (such as exchangeability priors) into our models
- As long as the second stage prior is reasonable and the tau is not set too small (you need to cover the truth with some level of likelihood) then SB estimates should be an improvement over other MLE approaches
- You should report both MLE and SB results
- If you need to select some exposures for further study, ranking on the SB estimate or on the SB-derived attributable number (of cases due to exposure) is a sound approach



R code

Chen and Witte (2007) Enriching the Analysis of Genomewide Association Studies with Hierarchical Modeling. The American Journal of Human Genetics 81: 397-404.

Files

https://github.com/Wittelab/hm-gwa

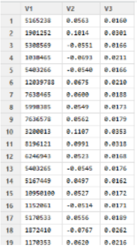
3 files:

first_stage.txt	(labels, first stage betas, standard errors)
covariates.txt	(second-stage “prior” covariate data)
hm.r	(r script)

- This is a two-stage approach that calculates the SB estimates as variance weighted averages of the first-stage and second-stage coefficients.
- There are penalized likelihood regression approaches that have better small-sample behaviour.

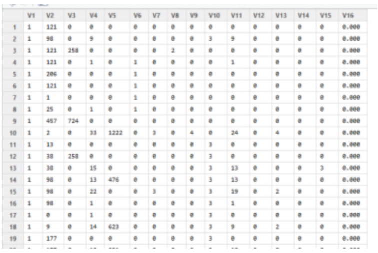
Matrices and vectors

Labels Betas SE



	V1	V2	V3
1	5165238	0.0563	0.0160
2	1901252	0.1014	0.0301
3	5308569	-0.0551	0.0166
4	1030465	-0.0693	0.0211
5	5403266	-0.0540	0.0166
6	12039788	0.0675	0.0210
7	7638465	0.0600	0.0188
8	5990385	0.0549	0.0173
9	7636578	0.0562	0.0179
10	3200013	0.1107	0.0353
11	8196121	0.0991	0.0318
12	6246943	0.0523	0.0168
13	5403265	-0.0545	0.0176
14	5167449	0.0497	0.0162
15	10950100	0.0527	0.0172
16	1152061	-0.0514	0.0171
17	5170533	0.0556	0.0189
18	1872410	-0.0767	0.0262
19	1170353	0.0620	0.0216

Z-matrix (second-level variables)



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16
1	1	121	0	0	0	0	0	0	0	0	0	0	0	0	0	0.000
2	1	98	0	9	0	0	0	0	0	3	9	0	0	0	0	0.000
3	1	121	258	0	0	0	0	2	0	0	0	0	0	0	0	0.000
4	1	121	0	1	0	1	0	0	0	0	1	0	0	0	0	0.000
5	1	206	0	0	0	1	0	0	0	0	0	0	0	0	0	0.000
6	1	121	0	0	0	1	0	0	0	0	0	0	0	0	0	0.000
7	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0.000
8	1	25	0	1	0	1	0	0	0	0	0	0	0	0	0	0.000
9	1	457	724	0	0	0	0	0	0	0	0	0	0	0	0	0.000
10	1	2	0	33	1222	0	3	0	4	0	24	0	4	0	0	0.000
11	1	13	0	0	0	0	0	0	0	3	0	0	0	0	0	0.000
12	1	38	258	0	0	0	0	0	0	3	0	0	0	0	0	0.000
13	1	38	0	15	0	0	0	0	0	3	13	0	0	0	3	0.000
14	1	98	0	13	476	0	0	0	0	3	13	0	0	0	0	0.000
15	1	98	0	22	0	0	3	0	0	3	19	0	2	0	0	0.000
16	1	98	0	1	0	0	0	0	0	3	1	0	0	0	0	0.000
17	1	0	0	1	0	0	0	0	0	3	0	0	0	0	0	0.000
18	1	9	0	14	623	0	0	0	0	3	9	0	2	0	0	0.000
19	1	177	0	0	0	0	0	0	0	3	0	0	0	0	0	0.000

Setup

```
firstStage<-read.table("first_stage.txt")
  markers<-firstStage[,1]
  betas<-abs(firstStage[,2])
  se<-firstStage[,3]

Z<-as.matrix(read.table("covariates.txt"))
Z<-replace(Z,is.na(Z),0)                              # replace NAs with 0
  rowLength<-length(markers)

  tau <- 0.1
  rho <- 0.01
  weightingCol <- 16
```

Estimation

```
  prior_scaling<-(log(tau^2)-log(rho^2))/max(Z[,weightingCol])
  shrinkage<-1/exp(prior_scaling*Z[,weightingCol])                    # Compute the weights
  wt<-(se[]^2+tau^2*shrinkage)^-1                                     # for now defined the weight as the linkage column
  lmsummary<-summary(lm(betas~Z,,,weights=wt))
  outputVector<-c(lmsummary$coefficients[,1],lmsummary$coefficients[,2])

write.table(matrix(outputVector,length(Z[1,])),file=paste("hm_tau",tau,"_rho",rho,"_coefficients.txt",sep=""),col.na
mes=FALSE,row.names=FALSE,quote=FALSE)
  second.betas<-Z %*% lmsummary$coefficients[,1]
  Z_transpose<-t(Z)
  # The following line of code assume that the second stage matrix T is
  # diagonal in order to speed up computations for vast numbers of SNPs
  zprimewmatrix<-apply(as.matrix(Z_transpose),1,function(x,weights) x*weights,weights=wt)
  secondVarFirstHalf<-Z %*% solve(t(zprimewmatrix) %*% Z)
  second.Var<-apply(t(secondVarFirstHalf)*Z_transpose,2,sum)
  second.SE<-sqrt(second.Var)
  B<-(se^2)*wt
  hm.betas<-B*second.betas + (1-B)*betas
  H<-second.Var * wt
  C<-se^2 * (1- (1-H) * B )
  hm.SE<-sqrt(C)
```