Why we should become Bayesians (and often already are without realising it)

Neil Pearce
Department of Medical Statistics
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine
Why we should become Bayesians (and often already are without realising it)

- Frequentist and Bayesian statistics
- Estimation of main effects
  - One exposure/outcome, one level of analysis
  - Many exposures/outcomes, one level of analysis
  - One/many exposures/outcomes, multilevel analysis
- Estimation of bias
  - Selection bias
  - Information bias
  - Confounding
Frequentist approaches

Our “standard” statistical methods are based on frequentist theory which (usually) assume that:

- The data have been sampled from an infinite population
- Generalizability depends on representativeness
- Exposure has been randomized (conditional on the confounders that have been controlled for)
- No information from outside the study is taken into account

These methods were developed for randomized controlled trials (and originally for agricultural experiments) where exposure is randomly assigned, and numbers are large.

Frequentist approaches depend on the model chosen; most methods of model selection ignore background information, and make questionable assumptions about interactions, dose-response relationships and lack of (uncontrolled) bias.
Frequentist approaches and causal inference

Causal criteria (Bradford-Hill) in epidemiology involve many more considerations including, plausibility, coherence, specificity,

Frequentist approaches attempt to make decisions solely on the basis of the data in the study being analysed

Almost no-one is stupid enough to do this in practice

• We take prior information into account when deciding to do a study
• We take prior information into account when interpreting the findings of a study

Thus, we write the methods and results sections of our papers as frequentists and the introduction and discussion as qualitative Bayesians
Bayesian approaches to randomized trials: [Pocock SJ. J Roy Statist Soc 1994; 157: 357-416.]

“One might adopt the term ‘closet Bayesian’ for statisticians and other scientists who adopt strategies in study design and data interpretation which include concepts of prior belief, but who do not explicitly express them in a formal Bayesian framework.”

– “Bayesian conceptualization is very useful in study design.”
– “It is useful to be a temporary Bayesian when faced with a surprising frequentist result.”
– “It is useful that every medical statistics unit should have a resident Bayesian.”

“Non-Bayesians need greater help in overcoming computational difficulties, in adapting Bayesian methods to more complex problems... and in learning how to communicate Bayesian findings to non-specialists.”
Bayesian approaches

Bayesian approaches formally take prior evidence into account.

- Classically, this involves having some belief (the prior distribution) about what the evidence was before we did the study. This belief could be based on:
  - Prior data from similar studies
  - More “subjective” beliefs based on information apart from the data being analysed
- This belief is then updated (the posterior distribution) on the basis of the new study
Some caveats

• I am using the term ‘Bayesian’ fairly loosely to refer to any data analysis which uses prior information
• I am focussing on observational studies, not randomised trials
• Many of the analyses/methods I will present can be interpreted within a frequentist framework
• I am (mostly) advocating supplementing rather than replacing frequentist with Bayesian methods
• I am not going to present sophisticated Bayesian methods but will mostly use ‘back of the envelope’ calculations
  • “We would welcome suggestions for articles that supply equally easy to implement procedures or SAS or Stata procs. But please spare us methods that can only be implemented through R, WinBUGS, or by spending £100,000 or more on statistical programmers.” (Pearce and Greenland)
Bayesian approaches do not require complex programming

- Our standard methods use approximations which are quite adequate given the other uncertainties with most epidemiologic data
- For example, most of our methods were developed at a time when exact tests were too complex computationally, but we still don’t use them now even though we could if we wanted to
- Bayesian methods have become popular at a time when more complex computational methods (e.g. MCMC – Winbugs) are available; but these are not necessary
- We can easily do Bayesian analyses using adaptations of our standard methods
Why we should become Bayesians (and often already are without realising it)

- Frequentist and Bayesian statistics
- Estimation of main effects
  - One exposure/outcome, one level of analysis
  - Many exposures/outcomes, one level of analysis
  - One/many exposures/outcomes, multilevel analysis
- Estimation of bias
  - Selection bias
  - Information bias
  - Confounding
Effect estimation for one exposure/outcome: dioxin exposure and cancer

- A New Zealand study found an increased risk of (total) cancer in phenoxy herbicide production workers exposed to dioxin: RR=1.24 (95% CI 0.90-1.67)
- A much larger international study had found similar findings
- There is also evidence of the likely aetiological mechanism
- As a result, the International Agency for Research on Cancer has classified dioxin as a Class 1 (sufficient evidence) human carcinogen
Effect estimation for one exposure/outcome: dioxin exposure and cancer

- The “prior data” indicates a RR=1.29 (95% CI 0.94-1.76)
- The new study shows a RR=1.24 (95% CI 0.90-1.67)
- The authors of the study concluded that the New Zealand findings were consistent with those of the larger international study
- The New Zealand study was repeated by different researchers who got the same findings but concluded that there was no statistically significant increase in risk
What if there is no “prior data”, but only “prior knowledge”? 

Relation of maternal antibiotic use during pregnancy ($X = 1$) to sudden infant death syndrome (SIDS, $Y = 1$), Kraus et al. 1989

Antibiotics might be associated with

- elevated risk (marking effects on the fetus of an infection, or via a direct effect)
- reduced risk (by reducing presence of infectious agents).

Source: Sander Greenland
These are weak speculations, but suppose strong effects seem unlikely.

A prior for ln(OR) that is normal with mean 0, standard deviation ln(4)/1.96 puts 95% probability on an OR between $\frac{1}{4}$ and 4 i.e. the prior assumption is that it is “95% likely” that the odds ratio is in the range 0.25-4.00 with the most likely value being 1.00 (i.e. OR=1.00, 95% CI 0.25-4.00)

Source: Sander Greenland
Question:

What is a “data equivalent” for the prior? That is, what data would give

• 0 as the conventional \( \ln(OR) \) point estimate
• \( \ln(4)/1.96 \) as its standard error?

Answers to such questions can be found by thought experiments.

Source: Sander Greenland
Augment the observed data with the prior data as a separate stratum:

<table>
<thead>
<tr>
<th></th>
<th>Prior $X=1$</th>
<th></th>
<th>Prior $X=0$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$X=1$</td>
<td>$X=0$</td>
<td>$X=1$</td>
</tr>
<tr>
<td>$Y=1$</td>
<td>4</td>
<td>4</td>
<td>173</td>
</tr>
<tr>
<td>$Y=0$</td>
<td>100,000</td>
<td>100,000</td>
<td>134</td>
</tr>
<tr>
<td>$RR_{prior}$</td>
<td>1.00</td>
<td></td>
<td>OR = 1.42</td>
</tr>
<tr>
<td>95% PL</td>
<td>0.25, 4.00</td>
<td></td>
<td>95% CL = 1.11, 1.83</td>
</tr>
</tbody>
</table>
Now combine the strata using any conventional method…

Approx. posterior median and 95% limits from Woolf, Mantel–Haenszel, and ML:

\[ 1.41 \ (1.10, \ 1.80) = \text{precision-weighted} \]

In a regression, add the stratum as a set of two prior records \((X=1,0)\) with

• an indicator \(\text{Prior}_X=1\) for the two prior records \((X=1,0)\), 0 for the remaining data
• all other regressors set to their means

Source: Sander Greenland
Effect estimation for one exposure/outcome

- One reasonable approach to the analysis of an exposure-disease association for which there is prior data is to:
  - Summarize the “prior data” as a summary RR and 95% CI (i.e., a meta-analysis)
  - Estimate the RR and 95% CI in the new study
  - Calculate an “updated” summary RR and 95% CI including the new data
Effect estimation for one exposure/outcome

• This is the “classic” Bayesian approach. It is probably also the least useful because:
  – If we don’t have any empirical prior data, then any sensible prior will not include much information and won’t change our estimates very much (and there is always concern about strong “subjective” priors)
  – If we have good empirical prior data, we can just do an updated meta-analysis
  – If we do a series of studies, a series of Bayesian analyses will arrive at the same conclusions as we would have reached with a single meta-analysis conducted at the end of the series
Point estimates and confidence intervals in five hypothetical studies
Effect estimation

- One exposure/outcome, one level of analysis
- Many exposures/outcomes, one level of analysis
- One/many exposures/outcomes, multilevel analysis
Effect estimation: many exposures/outcomes

- When we do a study involving many different exposures/outcomes, some of them will have strong “a priori” evidence and some will not.
- Standard (frequentist) methods for multiple comparisons (e.g. Bonferroni corrections) do not take prior knowledge into account, and in any case deal only with the statistical significance, not the magnitude of individual effect estimates.
- The strongest associations are likely to be due in part to chance, and to show “regression to the mean” if the study were repeated.
- Empirical Bayes (EB) or Semi-Bayes (SB) methods attempt to “correct for” extra variation and to anticipate “regression to the mean.”
A Large-Scale, Consortium-Based Genomewide Association Study of Asthma
(Moffatt, NEJM 2010)

- Brings together data from 23 studies; largest analysis ever conducted
- 10,365 asthma cases and 16,110 non-asthmatic controls
- 582,892 SNPs evaluated; 15 billion genotypes
- 10 loci identified; largest odds ratio was 1.27
- Estimated that these 10 loci together had a population attributable risk of about 30%
Regression to the mean

Regression to the mean pseudo-effect
High risk occupations for Non-Hodgkin’s Lymphoma in New Zealand
A case-control study

Andrea ’t Mannetje, Evan Dryson, Chris Walls, Dave McLean, Fiona McKenzie, Milena Maule, Soo Cheng, Chris Cunningham, Hans Kromhout, Paolo Boffetta, Aaron Blair, Neil Pearce

Centre for Public Health Research
Massey University
Wellington
New Zealand
Example of multiple exposures: New Zealand case-control study of occupation and non-Hodgkin’s lymphoma

- Many different job titles and exposures
- Some have strong prior evidence of an association with bladder cancer, but most do not
- In the latter case, we would expect the most “extreme” ORs to be in part due to chance, and to demonstrate “regression to the mean” if the study were repeated
### a priori high risk occupations and industries: farming

<table>
<thead>
<tr>
<th>OCCUPATION</th>
<th>cases/controls (n)</th>
<th>OR</th>
<th>95%CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>611-Market Farmers and Crop Growers</td>
<td>41/44</td>
<td>1.48</td>
<td>(0.92-2.37)</td>
</tr>
<tr>
<td>6111-Field Crop and Vegetable Growers</td>
<td>12/7</td>
<td>2.74**</td>
<td>(1.04-7.25)</td>
</tr>
<tr>
<td>6112-Fruit Growers</td>
<td>20/20</td>
<td>1.63</td>
<td>(0.84-3.16)</td>
</tr>
<tr>
<td>6113-Gardeners and Nursery Growers</td>
<td>17/18</td>
<td>1.27</td>
<td>(0.63-2.58)</td>
</tr>
<tr>
<td>61131-Nursery Grower, Nursery Worker</td>
<td>10/5</td>
<td>3.16**</td>
<td>(1.03-9.69)</td>
</tr>
<tr>
<td>612-Market Oriented Animal Producers</td>
<td>44/81</td>
<td>0.80</td>
<td>(0.52-1.21)</td>
</tr>
<tr>
<td>6121-Livestock Producers</td>
<td>19/43</td>
<td>0.65</td>
<td>(0.36-1.16)</td>
</tr>
<tr>
<td>6122-Mixed Livestock Producers</td>
<td>12/19</td>
<td>1.13</td>
<td>(0.53-2.40)</td>
</tr>
<tr>
<td>6125-Crop and Livestock Producers</td>
<td>14/29</td>
<td>0.71</td>
<td>(0.36-1.41)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>INDUSTRY</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A011-Horticulture and Fruit Growing</td>
<td>41/32</td>
<td>2.28**</td>
<td>(1.37-3.79)</td>
</tr>
<tr>
<td>A0111-Plant Nurseries</td>
<td>8/3</td>
<td>4.30**</td>
<td>(1.08-17.2)</td>
</tr>
<tr>
<td>A0113-vegetable growing</td>
<td>11/8</td>
<td>2.32*</td>
<td>(0.90-6.00)</td>
</tr>
<tr>
<td>A0115-Apple and Pear Growing</td>
<td>9/3</td>
<td>4.91**</td>
<td>(1.26-19.1)</td>
</tr>
<tr>
<td>A0117-Kiwi Fruit Growing</td>
<td>7/6</td>
<td>1.77</td>
<td>(0.55-5.63)</td>
</tr>
<tr>
<td>A012-Grain, Sheep, Beef Cattle Farming</td>
<td>14/39</td>
<td>0.56*</td>
<td>(0.29-1.08)</td>
</tr>
<tr>
<td>A013-Dairy Cattle Farming</td>
<td>14/36</td>
<td>0.55*</td>
<td>(0.29-1.07)</td>
</tr>
<tr>
<td>A015-Other Livestock Farming</td>
<td>9/2</td>
<td>9.75**</td>
<td>(2.04-46.5)</td>
</tr>
</tbody>
</table>

* *p<0.1
** p<0.05
Semi-Bayes estimates

Are these odds ratios elevated due to chance?

Particularly those odds ratio estimates based on small numbers could be elevated due to chance.

SB adjustment: shrinks the outlying relative risks towards the overall mean (of the relative risks of all occupations/industries).

The larger the individual variance of the relative risks, the stronger is the shrinkage, i.e. the shrinkage is stronger for less reliable estimates based on small numbers.
A priori high risk occupations and industries – effect of semi-Bayes adjustment

High risk occupations for Non-Hodgkin’s Lymphoma in New Zealand: case-control study
A posteriori high risk **occupations** and **industries**

<table>
<thead>
<tr>
<th>OCCUPATION</th>
<th>cases/controls (n)</th>
<th>OR</th>
<th>95%CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>31-Physical Science and Engineering Prof.</td>
<td>26/27</td>
<td>1.85**</td>
<td>(1.03-3.35)</td>
</tr>
<tr>
<td>91-Labourers and Rel. Elementary</td>
<td>86/84</td>
<td>1.76**</td>
<td>(1.22-2.56)</td>
</tr>
<tr>
<td>91111-Cleaner</td>
<td>34/32</td>
<td>2.11**</td>
<td>(1.21-3.65)</td>
</tr>
<tr>
<td>914-Packers and Freight Handlers</td>
<td>29/21</td>
<td>1.89**</td>
<td>(1.02-3.49)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>INDUSTRY</th>
<th>cases/controls (n)</th>
<th>OR</th>
<th>95%CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>H572-Pubs, Taverns and Bars</td>
<td>10/4</td>
<td>3.68**</td>
<td>(1.11-12.3)</td>
</tr>
<tr>
<td>O8729-Non-Residential Care Services nec</td>
<td>8/4</td>
<td>3.98**</td>
<td>(1.13-14.1)</td>
</tr>
</tbody>
</table>

**p<0.05
**A posteriori** high risk **occupations** and **industries** – effect of semi-Bayes adjustment

High risk occupations for Non-Hodgkin’s Lymphoma in New Zealand: case-control study

27/27                    86/84                   34/32                    29/21                    10/4
8/4
Effect estimation: many exposures/outcomes

- In some instances, “exposures” may be linked (e.g. adjacent geographical areas, or adjacent years when examining time trends)

TORINO

CUNEO

BIELLA

VERBANIA

NOVARA

VERCELLI

CASALE M.TO

ALESSANDRIA

ASTI

range: 0 - 17.03

range: 0.42 - 17.88

BMR Pleura (women)
Effect estimation

- One exposure/outcome, one level of analysis
- Many exposures/outcomes, one level of analysis
- One/many exposures/outcomes, multilevel analysis
Hierarchical Regression for multiple comparisons:
Application to a case-control study of occupational risks for lung cancer

Marine Corbin, Roel Vermeulen, Hans Kromhout, Susan Peters, Lorenzo Simonato,
Lorenzo Richiardi, Franco Merletti, Neil Pearce, Milena Maule

Cancer Epidemiology Unit
Department of Biomedical Sciences
and Human Oncology
University of Turin
Italy

Centre for Public Health Research
Massey University
Wellington
New Zealand
Existing methods for adjustment for multiple comparisons

• Traditional Bonferroni adjustment:
  ➢ Criticized because
    ➢ It treats all the associations equally
    ➢ It only affects the p-values

• Semi-Bayes adjustment towards the global mean
  ➢ Specification of an \textit{a priori} true standard deviation of the log ORs
  ➢ Shrinkage of the log ORs towards the global mean
Hierarchical Regression

- Inclusion of prior knowledge about carcinogenic exposures in a 2nd-stage model: Occupations with similar exposures to known carcinogens entail similar risks of disease

- Shrinkage of the log ORs towards each other when they have similar carcinogenic exposures
The first-stage models

- Estimation of the ORs of lung cancer through logistic regression for the 129 occupations represented by 3-digit ISCO codes and held by at least 10 subjects

- Adjustment for age, centre and cigarette smoking status

\[
\text{logit}\left[\Pr(Y = 1|\text{occ}_i, w)\right] = \alpha_i + \text{occ}_i \beta_i + w \gamma_i
\]
The 2nd-stage model (1)

- DOM-JEM: Job Exposure Matrix that classifies occupations in three categories of exposure (0=none, 1=low, 2=high) to several lung carcinogens

- Selection of 3 lung carcinogens (Agents classified by the IARC Monographs as *group-1 = carcinogenic to humans*): asbestos, chromium and silica to be included in the 2nd-stage model
The 2nd-stage model (2)

\[ \beta = Z\pi + U \]

<table>
<thead>
<tr>
<th>Occupation</th>
<th>Asb1</th>
<th>Asb2</th>
<th>Ch1</th>
<th>Ch2</th>
<th>Si1</th>
<th>Si2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nursery workers and gardeners</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Farm machin. operators</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Loggers</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fishermen</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Production supervisors</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Miners &amp; quarrymen</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
The 2nd-stage model (3)

$$\beta = Z\pi + U$$

\(U\) is a vector of the error terms representing the residual effect of being employed in each occupation after accounting for the exposure to asbestos, chromium and silica

$$U \sim N(0, \tau^2 T)$$

$$\tau^2 T$$ is specified a priori

$\Rightarrow$ $\tau$ controls the global strength of the shrinkage and is set successively to 0.76, 0.59, 0.41, 0.23

$\Rightarrow$ \(T\) is a Diagonal matrix where \(t_{ii}\) is inversely related to the sum of the categories of exposure of the occupation to asbestos, chromium and silica
Computation of the Hierarchical Regression estimates

- Estimation of the 2nd-stage coefficients $\tilde{\pi}$ through weighted least squares

- Computation of the prior means $Z\tilde{\pi}$ for the occupations coefficients

- Computation of the posterior Hierarchical Regression estimates

$$\tilde{\beta}_{HR} = BZ\tilde{\pi} + \left(I - B\right)\hat{\beta}$$

$B$ increases when the variance of $\hat{\beta}$ increases
$B$ decreases when $\tau^2 T$ increases
## Results (1)

<table>
<thead>
<tr>
<th></th>
<th>ML</th>
<th>SB</th>
<th>HR: ( \tau=0.76 )</th>
<th>HR: ( \tau=0.59 )</th>
<th>HR: ( \tau=0.41 )</th>
<th>HR: ( \tau=0.23 )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean of the log ORs</strong></td>
<td>-0.12</td>
<td>-0.07</td>
<td>-0.08</td>
<td>-0.07</td>
<td>-0.06</td>
<td>-0.04</td>
</tr>
<tr>
<td><strong>distribution</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Standard Deviation of the log ORs distribution</strong></td>
<td>0.63</td>
<td>0.31</td>
<td>0.41</td>
<td>0.35</td>
<td>0.28</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mean of the Standard Deviations of the log ORs</strong></td>
<td>0.45</td>
<td>0.32</td>
<td>0.37</td>
<td>0.34</td>
<td>0.28</td>
<td>0.20</td>
</tr>
</tbody>
</table>

The graph shows the distribution of log ORs for the 129 occupations, with three curves representing different methods: ML, SB, and HR (with \( \tau=0.59 \)).
## Results (3)

<table>
<thead>
<tr>
<th>Occupation</th>
<th>ML OR [95% CI]</th>
<th>SB OR [95% CI]</th>
<th>HR τ=0.76 [95% CI]</th>
<th>HR τ=0.59 [95% CI]</th>
<th>HR τ=0.23 [95% CI]</th>
<th>Carcinogenic exposure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tailors and dress makers</td>
<td>2.1 [0.9-5.0]</td>
<td>1.5 [0.8-3.0]</td>
<td>1.6 [0.8-3.5]</td>
<td>1.5 [0.7-3.0]</td>
<td>1.0 [0.7-1.6]</td>
<td>Asb 0  Ch 0  Si 0</td>
</tr>
<tr>
<td>Metal melters and reheaters</td>
<td>2.1 [0.7-5.7]</td>
<td>1.4 [0.7-3.0]</td>
<td>1.8 [0.9-3.9]</td>
<td>1.8 [0.9-3.5]</td>
<td>1.7 [1.1-2.6]</td>
<td>Asb 1  Ch 2  Si 0</td>
</tr>
<tr>
<td>Miners and quarrymen</td>
<td>1.2 [0.6-2.4]</td>
<td>1.1 [0.6-2.0]</td>
<td>1.2 [0.6-2.4]</td>
<td>1.3 [0.7-2.4]</td>
<td>1.3 [0.7-2.5]</td>
<td>Asb 1  Ch 0  Si 2</td>
</tr>
<tr>
<td>Painters, construction</td>
<td>1.8 [1.1-3.1]</td>
<td>1.6 [1.0-2.6]</td>
<td>1.7 [1.0-2.7]</td>
<td>1.6 [1.0-2.5]</td>
<td>1.2 [0.9-1.7]</td>
<td>Asb 1  Ch 0  Si 0</td>
</tr>
</tbody>
</table>
Conclusions

• Hierarchical Regression uses more specific priors than Semi-Bayes adjustment towards the global mean and performs then a more appropriate shrinkage if the information included in the 2nd-stage model is reliable.

• The choice of \( \tau \) depends on
  – The number and the reliability of the variables included in the 2nd-stage model.
  – The strength of their association with the outcome and the exposures of interest.
  – The specification of the first-stage model.

• Hierarchical regression is a valuable method to adjust for multiple comparisons in occupational studies when a reliable JEM is available.
Heirarchical regression: Equivalents and special cases


- Multilevel modelling
- Bayesian regression
- Empirical Bayes (EB) regression
- Bayes Empirical Bayes (BEB) regression
- Stein regression
- Penalized likelihood regression
- Best Linear Unbiased Prediction (BLUP)
- Mixed-model regression
- Ridge regression
- Random-coefficient regression
- Variance-components analysis

Source: Sander Greenland
Hierarchical regression:
Equivalents and special cases


- Hierarchical regression unifies frequentist and Bayesian methods of analysis
- It also effectively takes account of “clustering” – a one-level model produces confidence intervals that are too wide
- Conventional least squares and maximum likelihood are also special cases with just one level of analysis
- EB “adjustments” are a special case where the second-stage “prior” means are estimated from the data
- We can also add ‘second level’ information, e.g. on which occupations involve exposure to particular chemicals, or which chemicals have common properties

Source: Sander Greenland
Why we should become Bayesians (and often already are without realising it)

• Basics of frequentist and Bayesian statistics
• Estimation of main effects
  – One exposure/outcome, one level of analysis
  – Many exposures/outcomes, one level of analysis
  – One/many exposures/outcomes, multilevel analysis
• Estimation of bias
  – Selection bias
  – Information bias
  – Confounding
Approaches to bias modeling:
uncontrolled confounding

- We have no information on a potential confounder (e.g. smoking)
- If we can make assumptions about the associations between the potential confounder and the main exposure and between the potential confounder and the disease we can make “adjusted” estimates controlling for the potential confounder
- By varying these assumptions we can do a “sensitivity analysis” of the likely bias due to uncontrolled confounding
## Cause-specific Mortality in a New Zealand cohort study of meat workers

<table>
<thead>
<tr>
<th>Cause of death (ICD-9)</th>
<th>Obs</th>
<th>Exp</th>
<th>SMR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Causes</td>
<td>227</td>
<td>203.6</td>
<td>1.12</td>
<td>0.98 - 1.27</td>
</tr>
<tr>
<td>All Cancer (140-208)</td>
<td>69</td>
<td>61.4</td>
<td>1.12</td>
<td>0.88 - 1.42</td>
</tr>
<tr>
<td>Larynx (161)</td>
<td>1</td>
<td>0.4</td>
<td>2.63</td>
<td>0.07 – 14.62</td>
</tr>
<tr>
<td>Lung (162)</td>
<td>23</td>
<td>12.9</td>
<td>1.79*</td>
<td>1.13 – 2.68</td>
</tr>
<tr>
<td>Haematologic (200-208)</td>
<td>6</td>
<td>6.3</td>
<td>0.96</td>
<td>0.35 – 2.09</td>
</tr>
<tr>
<td>NHL (200, 202)</td>
<td>4</td>
<td>2.8</td>
<td>1.45</td>
<td>0.49 – 3.45</td>
</tr>
<tr>
<td>Leukaemia (204 – 208)</td>
<td>2</td>
<td>2.3</td>
<td>0.86</td>
<td>0.17 – 2.75</td>
</tr>
</tbody>
</table>
Systematic Error (bias): lung cancer in meat industry workers

- No information on smoking or ethnicity
- There was a small increase in mortality (SMR=1.31) but not incidence (SIR=0.90) for other smoking-related cancers
- Census information suggests that smoking rates in food and beverage workers are only slightly higher than in other workers, and that this could only account for a RR of 1.20
- Lung cancer mortality in Māori males is 1.4 times than in non-Māori males, and that in Māori females is 2.8 times that in non-Māori females; assuming that 40% of the cohort is Māori (compared with 15% of the general population) this would produce a RR of 1.09
Approaches to bias modeling

• Ordinary sensitivity analysis does not take account (easily) of uncertainties about the relationship between the confounder and disease, or random error in the “sampling” of the study population
• It also does not (easily) take account of multiple sources of bias
• It does not easily allow for control of (other) confounders
• It can sometimes be improved upon through the use of Monte Carlo Sensitivity Analysis or Bayesian methods
Approaches to bias modeling: misclassification

- Misclassification is a problem of missing data – we have data on a ‘measured’ variable, but not data on the ‘true’ variable
  - If there is a validation substudy, we can use multiple imputation
  - ‘Standard’ sensitivity analysis
  - Probabilistic/Monte Carlo sensitivity analysis
  - Bayesian sensitivity analysis using data augmentation
Approaches to bias modeling: misclassification of the main exposure

• We have no information on the true exposure (e.g. asbestos levels in the lung)
• We have imperfect information on a surrogate for the main exposure (e.g. cumulative asbestos exposure from a job-exposure-matrix)
• If we can make assumptions about the association between the true exposure and the surrogate measure (e.g. it’s sensitivity and specificity) then we can “impute” the missing values for the true exposure
• By varying these assumptions (about sensitivity and specificity) we can do a “sensitivity analysis” of the likely bias due to misclassification
Adjustment for misclassification of the smoking status in the association between smoking and lung cancer: *Multiple imputations and probabilistic sensitivity analysis*
The lung cancer case-control study

- Incident lung cancer cases notified 2007-2008 to the NZ Cancer Registry (aged 20-75)

- Population controls selected from the Electoral Roll

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>457</td>
<td>100.0%</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>227</td>
<td>49.7%</td>
</tr>
<tr>
<td>Women</td>
<td>230</td>
<td>50.3%</td>
</tr>
<tr>
<td><strong>Age at interview</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20-50 years</td>
<td>43</td>
<td>9.4%</td>
</tr>
<tr>
<td>51-60 years</td>
<td>118</td>
<td>25.8%</td>
</tr>
<tr>
<td>61-70 years</td>
<td>284</td>
<td>62.1%</td>
</tr>
<tr>
<td>70+ years</td>
<td>12</td>
<td>2.6%</td>
</tr>
<tr>
<td><strong>Smoking</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>49</td>
<td>10.7%</td>
</tr>
<tr>
<td>Ever</td>
<td>406</td>
<td>88.9%</td>
</tr>
<tr>
<td><strong>missing</strong></td>
<td>2</td>
<td>0.4%</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maori</td>
<td>79</td>
<td>17.3%</td>
</tr>
<tr>
<td>Non-Maori</td>
<td>378</td>
<td>82.7%</td>
</tr>
<tr>
<td><strong>NZSEI</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 1 (75-90) highest</td>
<td>7</td>
<td>1.5%</td>
</tr>
<tr>
<td>Class 2 (60-75)</td>
<td>52</td>
<td>11.4%</td>
</tr>
<tr>
<td>Class 3 (50-60)</td>
<td>67</td>
<td>14.7%</td>
</tr>
<tr>
<td>Class 4 (40-50)</td>
<td>114</td>
<td>24.9%</td>
</tr>
<tr>
<td>Class 5 (30-40)</td>
<td>116</td>
<td>25.4%</td>
</tr>
<tr>
<td>Class 6 (10-30) lowest</td>
<td>101</td>
<td>22.1%</td>
</tr>
</tbody>
</table>

We will use the lung cancer data set and assume that the measured smoking data is misclassified, and then estimate what the “true” odds ratio for smoking would have been if there had not been misclassification, first using standard sensitivity analysis, and then probabilistic sensitivity analysis.
Standard sensitivity analysis - Method

1. Determination of an *a priori* value for the bias parameters (sensitivity={0.6, 0.7, 0.8, 0.9, 1} and specificity ={0.6, 0.7, 0.8, 0.9, 1})

2. Calculation of the corrected frequencies according to the following formulas

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Corrected data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases</td>
<td>Controls</td>
</tr>
<tr>
<td>Ever smokers</td>
<td>a</td>
<td>c</td>
</tr>
<tr>
<td>Never smokers</td>
<td>b</td>
<td>d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>A+B</td>
<td>C+D</td>
</tr>
</tbody>
</table>

With

\[
A = \frac{a - N_{\text{Cases}}(1 - \text{SP})}{SE(1 - \text{SP})}
C = \frac{c - N_{\text{Controls}}(1 - \text{SP})}{SE(1 - \text{SP})}
B = N_{\text{Cases}} - A
D = N_{\text{Controls}} - C
\]

3. Estimation of the corrected odds ratio of the association between smoking and lung cancer using the corrected frequencies
## Standard sensitivity analysis - Results

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>1</td>
<td>NA</td>
</tr>
<tr>
<td>0.7</td>
<td>1</td>
<td>NA</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>NA</td>
</tr>
<tr>
<td>0.9</td>
<td>1</td>
<td>83.08</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>7.51</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6</td>
<td>17.47</td>
</tr>
<tr>
<td>1</td>
<td>0.7</td>
<td>11.66</td>
</tr>
<tr>
<td>1</td>
<td>0.8</td>
<td>9.42</td>
</tr>
<tr>
<td>1</td>
<td>0.9</td>
<td>8.24</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>7.51</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>0.6</td>
<td>NA</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7</td>
<td>NA</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>NA</td>
</tr>
<tr>
<td>0.9</td>
<td>0.9</td>
<td>91.15</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>7.51</td>
</tr>
</tbody>
</table>
Probabilistic sensitivity analysis

- Probabilistic sensitivity analysis and multiple imputation are useful methods to correct for misclassification.
- Both methods have the advantage, compared with “standard” sensitivity analysis, that they enable adjustment for covariates.
- A validation substudy must be available to use multiple imputation to correct for misclassification.
- Probabilistic sensitivity analysis can be regarded as Semi-Bayesian and it is easier to compute than fully Bayesian sensitivity analysis.
Standard sensitivity analysis

1) Determination of an \textit{a priori} distribution for the sensitivity and the specificity: Trapezoidal distribution (min=0.60, mode1=0.85, mode2=0.98, max=1)
Standard sensitivity analysis - Correction for the crude association

Use of an Excel spreadsheet related to Lash TL., Fox MP., Fink AK. *Applying quantitative bias analysis to epidemiologic data.*

https://sites.google.com/site/biasanalysis/

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Median (2.5th-97.5th percentile)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional</td>
<td>7.51 (5.41-10.43)</td>
</tr>
<tr>
<td>Corrected for misclassification</td>
<td>17.83 (8.9-288.51)</td>
</tr>
</tbody>
</table>
Probabilistic sensitivity analysis - Method

1) Determination of an \textit{a priori} distribution for the sensitivity and the specificity: Trapezoidal distribution (min=0.60, mode1=0.85, mode2=0.98, max=1)

2) Random draw of 500 combinations for the sensitivity and specificity from these distributions

3) Estimation of the corrected association for each combination of sensitivity and specificity to obtain a distribution of the corrected estimates
Probabilistic sensitivity analysis: Correction for misclassification

Use of a SAS macro program related to Fox MP., Lash TL., Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables

1) For each simulation, after calculating the corrected frequencies, the macro program calculates the positive predictive value and the negative predictive value.  
2) These values are applied to each individual to simulate the corrected smoking status.  
3) The association between smoking and lung cancer can then be estimated through a logistic regression model, adjusting for all the covariates.
Probabilistic sensitivity analysis: Correction for misclassification

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Median (2.5th-97.5th percentile)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional</td>
<td>7.80 (5.53-10.85)</td>
</tr>
<tr>
<td>Corrected for misclassification</td>
<td>19.78 (9.50-214.78)</td>
</tr>
</tbody>
</table>
## Frequentists and Bayesians (some stereotypes)

<table>
<thead>
<tr>
<th>Frequentist</th>
<th>Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Significance testing</strong> <em>(p-values, CIs)</em></td>
<td><strong>Effect estimation</strong> <em>(confidence intervals/distributions)</em></td>
</tr>
<tr>
<td>Methods/results focus on <em>random error</em> and <em>(measured) confounding</em></td>
<td>Methods/results focus on <em>systematic error</em> <em>(bias)</em> in addition to random error</td>
</tr>
<tr>
<td>Systematic error considered in <strong>discussion</strong></td>
<td>Systematic error also considered in methods/results</td>
</tr>
<tr>
<td>Prior knowledge only considered in <strong>introduction/discussion</strong></td>
<td>Prior knowledge formally incorporated into analysis</td>
</tr>
<tr>
<td>Generalisability/representativeness</td>
<td>Causal inference</td>
</tr>
</tbody>
</table>
Problems with p-values


‘Medical professionals and regulators act on the basis of evidence of causation that is not statistically significant’ [US Supreme Court, 2011]
Random error is not the main concern in epidemiological studies. Systematic error (bias) is: selection bias, misclassification, confounding.

We need to spend as much time/effort quantitatively assessing systematic error as we spend assessing random error.
Why we should all be Bayesians

- We write the methods and results sections of our papers as frequentists and the introduction and discussion as qualitative Bayesians.
- The Bayesian approach moves the consideration of prior evidence from the introduction and discussion sections to the methods and results sections.
- Prior evidence can be used both in effect estimation and in assessment of systematic error (bias).
- Subjective personal judgements (which are inherent in both frequentist and Bayesian methods) are made more explicit and the effects of changes in these assumptions can be assessed.
Why we should become Bayesians (and often already are without realising it)

Neil Pearce
Department of Medical Statistics
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine