Outline of talk

Usual measures of association (e.g. hazard ratios) do not have a direct interpretation in terms of aetiology / impact / causation

**Purpose:** To draw conclusions about clinical / public health impact

**Content:**
- Adjusting for measurement error
- Estimating life expectancy
- Estimating causal relationships using Mendelian randomization

**Need IPD for these analyses**
Adjusting for measurement error

Measurement error

‘Measurement error’ = technical / laboratory error
+ short-term within-person variation
+ long-term within-person variation

A single measurement of a risk factor is an imprecise estimate of long-term ‘usual’ level:
• Using error-prone exposure leads to underestimation of the aetiological association with usual levels
• Using error-prone confounders (usually) leads to exaggeration of the aetiological association

Require repeat measurements of exposure (and confounders) on (at least a subset of) individuals to make corrections
Measurement error in exposure

Extent of measurement error is often quantified by the regression dilution ratio (RDR)

\[ \text{RDR} = \text{coefficient from regression of one repeated measurement on another} \]

\[ \approx \text{correlation between repeated measurements} \]

**One method of correction:** Divide the naive risk regression coefficient (i.e. log hazard ratio) by the RDR

Regression dilution ratios for fibrinogen

27,000 individuals with repeat measurements of fibrinogen from 15 studies

Estimating regression dilution ratios

For repeat measurement \( (r) \) and baseline measurement \( (0) \) of exposure \( E \) on individual \( i \) in study \( s \):

\[
E_{i sr} = \alpha_{sr} + \rho_{sr}E_{i0} + \lambda X_{i0} + \epsilon_{i sr}
\]

\( \rho_{sr} = \) true RDR for repeat \( r \) in study \( s \)

An average RDR \( \rho \) can be estimated allowing for both within-study and between-study variability:

\[
\rho_{sr} = \rho + v_{sr} + u_s; \quad v_{sr} \sim N(0, \sigma^2_v); \quad u_s \sim N(0, \sigma^2_u)
\]

This model can be extended to:

– encompass time trends in the RDRs
– allow the RDR to depend on covariates
– allow the RDR to depend on the level of exposure

An RDR should be adjusted for the same covariates as used in the risk regression model

Overall hazard ratios for CHD per 1 g/L increase in fibrinogen, corrected for measurement error

<table>
<thead>
<tr>
<th>Measurement error correction</th>
<th>Hazard ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>1.38 (1.31 to 1.45)</td>
</tr>
<tr>
<td>In fibrinogen</td>
<td>1.96 (1.76 to 2.17)</td>
</tr>
</tbody>
</table>

(Results adjusted for age, sex, smoking, chol, SBP, BMI)

Equivalent results can be obtained using \textbf{conditional expectations}:

Replace the observed exposure in the risk regression model by its conditional expectation given observed values:

\[
E \left[ E_{sr} \mid E_{s0} \right] \text{ obtained by simple regression}
\]
Measurement error in exposure and confounders

Simultaneous models for repeat exposures and confounders in terms of baseline exposure and confounders
Expectations from these regression calibration models used as covariates in the disease risk model
Allow for between-study variability, and use empirical Bayes regression calibration coefficients
Appropriate (approximately) for linear terms in PH regression models

Assumptions
Errors in repeat measurements are independent of each other
Errors are non-differential and independent of the true value
Disease risk depends on usual (long-term average) risk factor levels

Overall hazard ratios for CHD per 1 g/L increase in fibrinogen, corrected for measurement error

<table>
<thead>
<tr>
<th>Measurement error correction</th>
<th>Hazard ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>1.38 (1.31 to 1.45)</td>
</tr>
<tr>
<td>In fibrinogen</td>
<td>1.96 (1.76 to 2.17)</td>
</tr>
<tr>
<td>In fibrinogen, smoking, chol, SBP, BMI</td>
<td>1.85 (1.66 to 2.06)</td>
</tr>
</tbody>
</table>

(Results adjusted for age, sex, smoking, chol, SBP, BMI)

Note: Because of residual confounding (e.g. from unmeasured confounders) the last estimate may still not represent a causal relationship.
Estimating life expectancy

Diabetes and mortality

- ERFC data from 97 prospective studies
- 821,000, participants with no known pre-existing CVD
- 123,000 deaths

Diabetes status on basis of self-report / medication use / fasting glucose > 7 mmol/l

Hazard ratios adjusted for age, sex, smoking status and BMI

ERFC, NEJM 2011
Hazard ratios for major causes of death associated with diabetes

Hazard ratios for major causes of death associated with diabetes, according to race, sex, and age
Estimating life expectancy according to diabetes status

Estimate cumulative survival from age 35 onwards:

- Calculate log hazard ratios specific to age-at-risk (5-year intervals) and sex for cause-specific mortality
- Smooth over age-at-risk categories using (quadratic / fractional) polynomials
- Apply to cause-specific rates of death at age 35 onwards from European Union

Estimated survival curves by sex and diabetes status
Estimated future years of life lost associated with diabetes, by sex, age, and cause

Note: Reduction in life-expectancy from long-term cigarette smoking is about 10 years

Multiple morbidities and life-expectancy: History of diabetes, stroke, myocardial infarction (MI)

Hazard ratios for all-cause mortality according to baseline diseases status, by sex

ERFC, JAMA 2015
Estimated future years of life lost associated with disease status at baseline

Estimated future years of life lost associated with diabetes and stroke / MI, attributable to vascular, cancer and other causes
Estimating causal relationships using Mendelian randomization

C-reactive protein (CRP) and CHD

CRP is an acute-phase protein, a marker of inflammation, strongly associated with CHD in observational prospective epidemiological studies.

IPD meta-analysis based on 54 prospective studies; 10,000 CHD events

<table>
<thead>
<tr>
<th>Adjustments (usual level of confounders)</th>
<th>Hazard ratio per 1 SD increase in usual log CRP (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, sex</td>
<td>1.68 (1.59 to 1.78)</td>
</tr>
<tr>
<td>+ SBP, smoking, diabetes, BMI, log TG, chol, HDL-C, alcohol</td>
<td>1.37 (1.27 to 1.48)</td>
</tr>
<tr>
<td>+ fibrinogen</td>
<td>1.23 (1.07 to 1.42)</td>
</tr>
</tbody>
</table>

ERFC, Lancet 2010

Is CRP causally related to CHD?
Genetic variants as instrumental variables  
= Mendelian Randomization (MR)

CRP CHD Genetics Collaboration (CCGC) collated individual participant data (IPD):

- 43 studies (cross-sectional, case-control, prospective)
- 160,000 participants of European descent
- 36,000 CHD events (MI, CHD death)
- Four pre-specified genetic variants (SNPs)
  on the CRP-regulatory gene on chromosome 1
- Blood CRP concentrations in most studies

**Aim:** To estimate the causal effect of CRP on CHD as precisely as possible

---

Diagram of causal effects

```
<table>
<thead>
<tr>
<th>Instrumental variable</th>
<th>Confounders (U)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRP gene variants (G)</td>
<td>CRP levels (X)</td>
</tr>
<tr>
<td></td>
<td>Outcome (Y)</td>
</tr>
</tbody>
</table>
```

**Three crucial assumptions:**
- G affects X
- G is not related to U
- Y is conditionally independent of G given X and U
Simplest instrumental variable analysis

2 genetic subgroups
Mean (95% CI) outcome and phenotype by genetic subgroup
Mean outcome = log odds of CHD

Ratio of coefficients method:
causal effect = \frac{\Delta \text{log odds of CHD}}{\Delta \text{mean phenotype}}

A modelling approach: one genetic marker in one study

Prospective study of new incident CHD with fixed follow-up

Individual i has: outcome y_i = 0/1
phenotype x_i
genetic variant g_i=0,1,2

Linear (per allele) model at individual level:
x_i \sim N(\xi_i, \sigma^2)
\xi_i = \alpha_0 + \alpha_1 g_i
y_i \sim \text{Bin}(1, \pi_i)
\logit(\pi_i) = \beta_0 + \beta_1 \xi_i

\beta_1 is the causal effect estimate (increase in log odds of event per unit increase in phenotype)

Two-stage or one-stage approach possible
Multiple genetic markers in one study

Individual i has:
outcome $y_i = 0/1$
phenotype $x_i$
genetic variants $g_{ik} = 0,1,2$ for $k=1...K$ SNPs

Additive linear (per allele) model at individual level:

$$x_i \sim N(\zeta_i, \sigma^2)$$

$$\zeta_i = \alpha_0 + \sum_k \alpha_k g_{ik}$$

$$y_i \sim \text{Bin}(1, \pi_i)$$

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \zeta_i$$

$\beta_1$ is the causal effect estimate

---

Multiple genetic markers in multiple studies

Individual i in study m has:
outcome $y_{im} = 0/1$, phenotype $x_{im}$
genetic variants $g_{ikm} = 0,1,2$ for $k=1...K_m$

Additive linear (per allele) model at individual level:

$$x_{im} \sim N(\bar{\zeta}_{im}, \sigma_m^2)$$

$$\bar{\zeta}_{im} = \alpha_{0m} + \sum_k \alpha_{km} g_{ikm}$$

$$y_{im} \sim \text{Bin}(1, \pi_{im})$$

$$\text{logit}(\pi_{im}) = \beta_{0m} + \beta_{1m} \bar{\zeta}_{im}$$

$\beta_{1m} = \beta_1$ fixed-effect meta-analysis

$\beta_{1m} \sim N(\beta_1, \tau^2)$ random-effects meta-analysis
Bayesian implementation

Vague priors:
- Wide normal $N(0,100^2)$ on regression parameters
- Wide uniform $U[0,20]$ on standard deviations

MCMC using WinBUGS

Propagates uncertainty from each stage
Allows feedback from each stage
Allows inclusion of studies with no blood CRP data

Are genetic variants instrumental variables?
Principal results of CCGC

Causal estimate = log odds ratio of CHD per unit increase in log CRP

<table>
<thead>
<tr>
<th>Studies / Cases</th>
<th>Causal est. (95%CI)</th>
<th>Heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-stage Bayesian analysis:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All studies</td>
<td>43 / 36463</td>
<td>–0.013 (–0.115 to 0.094)</td>
</tr>
</tbody>
</table>

Interpretation of principal result

Estimate of $\beta_1$  -0.013
95% CI (-0.115 to 0.094)
Estimate of $\tau$ 0.106

Overall OR per unit increase in log CRP:
0.99 (95%CI 0.89 to 1.10)

Overall OR per doubling in CRP:
0.99 (95%CI 0.92 to 1.07)

Predictive distribution for true OR in new study per doubling of CRP:
0.99 (95% range 0.84 to 1.16)

Not supportive of a causal role of CRP in CHD
Conclusions

IPD has enabled:

• Correction for measurement error
• Estimation of life expectancy
• Estimation of causal relationships in Mendelian randomization

References

Measurement errors


Life expectancy


Mendelian randomization