Advanced IPD meta-analysis methods for observational studies

Simon Thompson
University of Cambridge, UK

Part 4

UNIVERSITY OF CAMBRIDGE
Department of Public Health and Primary Care

IBC Victoria, July 2016

1

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

2

Adjusting for measurement error

3

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

4

Measurement error in exposure

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

RDR = coefficient from regression of one repeated measurement on another

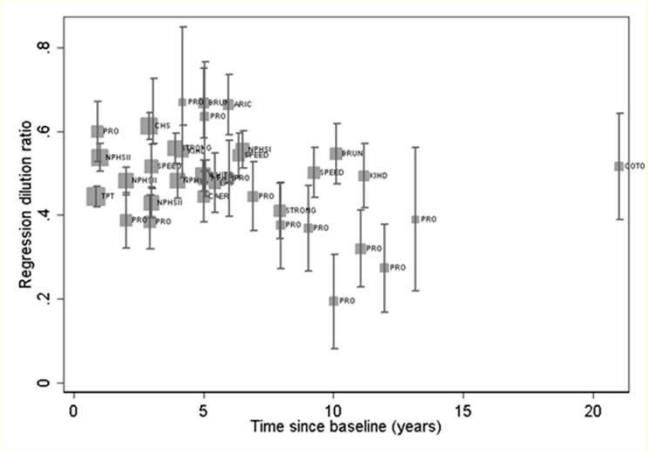
≈ correlation between repeated measurements

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

5

Regression dilution ratios for fibrinogen

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



Wood et al: Int J Epidemiol 2006

6

Estimating regression dilution ratios

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

$$E_{sir} = \alpha_{sr} + \rho_{sr} E_{si0} + \boldsymbol{\lambda X}_{si0} + \varepsilon_{sir}$$

ρ_{sr} = true RDR for repeat *r* in study *s*

An average RDR ρ can be estimated allowing for both within-study and between-study variability:

$$\rho_{sr} = \rho + v_{sr} + u_s;\ v_{sr} \sim N(0, \sigma_v^2); u_s \sim N(0, \sigma_u^2)$$

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

7

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

Measurement error correction	Hazard ratio (95% CI)
None	1.38 (1.31 to 1.45)
In fibrinogen	1.96 (1.76 to 2.17)

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

Equivalent results can be obtained using **conditional expectations**:

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

E [E_{sir} | E_{si0}] obtained by simple regression

8

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

Wood et al, Stat Med 2009

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

Measurement error correction	Hazard ratio (95% CI)
None	1.38 (1.31 to 1.45)
In fibrinogen	1.96 (1.76 to 2.17)
In fibrinogen, smoking, chol, SBP, BMI	1.85 (1.66 to 2.06)

(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

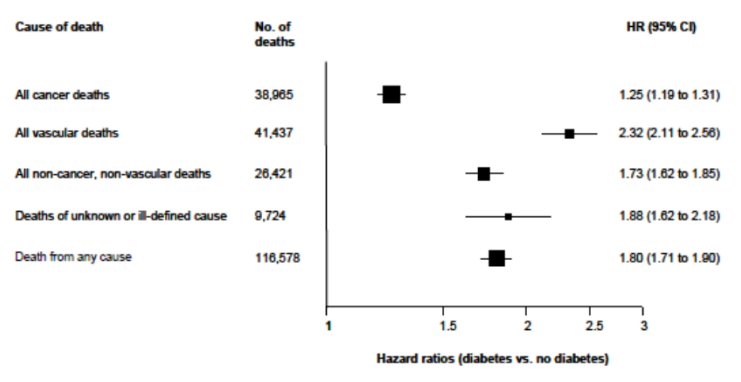
- ERFC data from 97 prospective studies
- 821,000, participants with no known pre-existing CVD
- 123,000 deaths

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

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

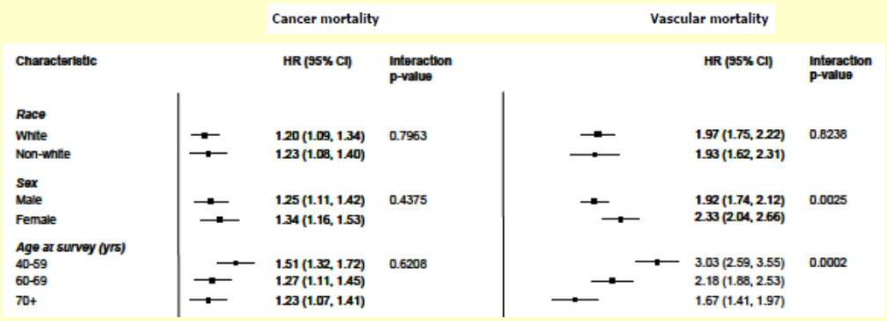
ERFC, NEJM 2011

Hazard ratios for major causes of death associated with diabetes



13

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



14

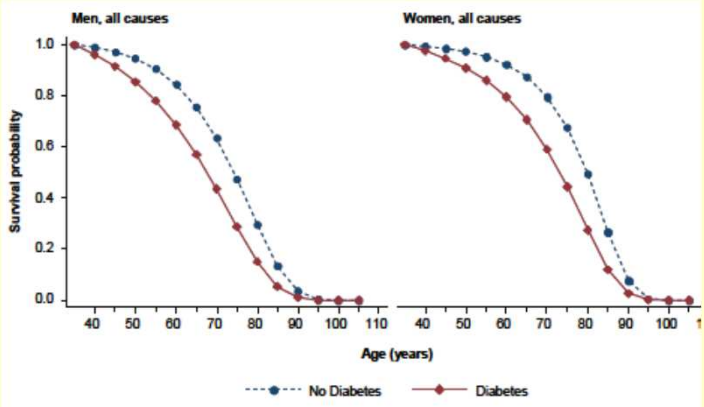
Estimating life expectancy according to diabetes status

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

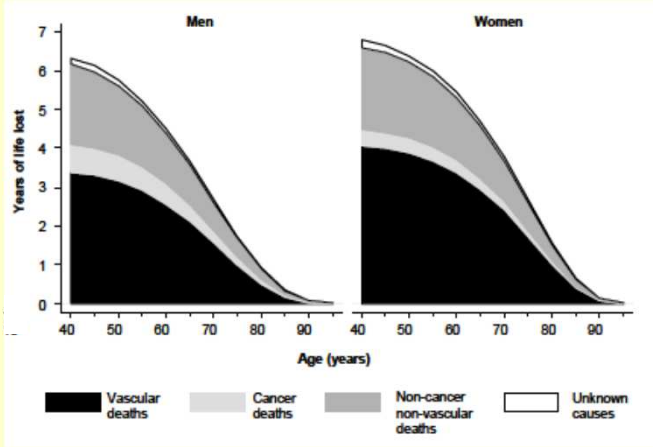
15

Estimated survival curves by sex and diabetes status



16

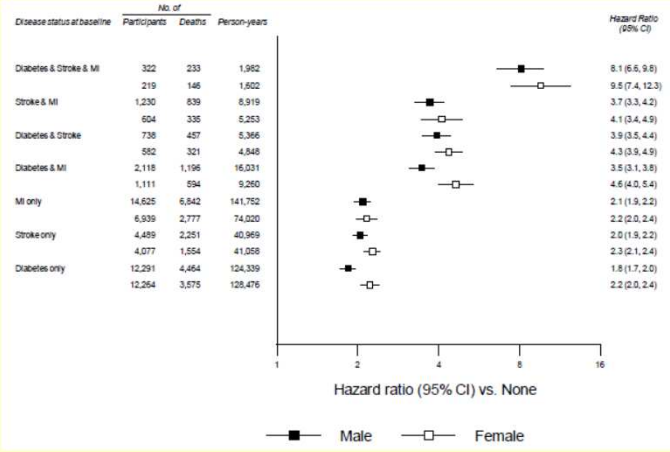
Estimated future years of life lost associated with diabetes, by sex, age, and cause



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

17

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

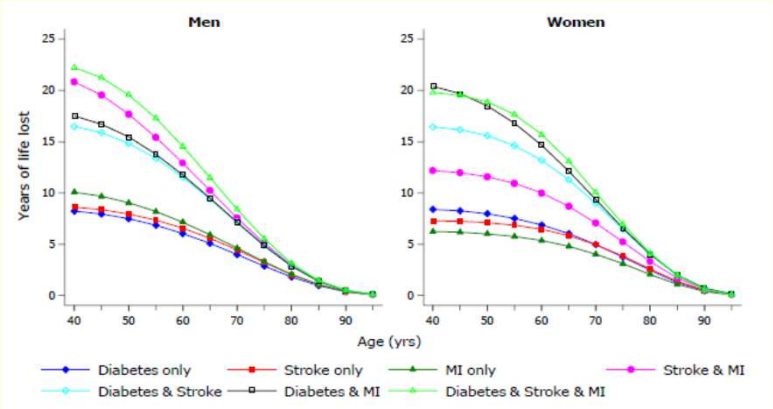


Hazard ratios for all-cause mortality according to baseline diseases status, by sex

ERFC, JAMA 2015

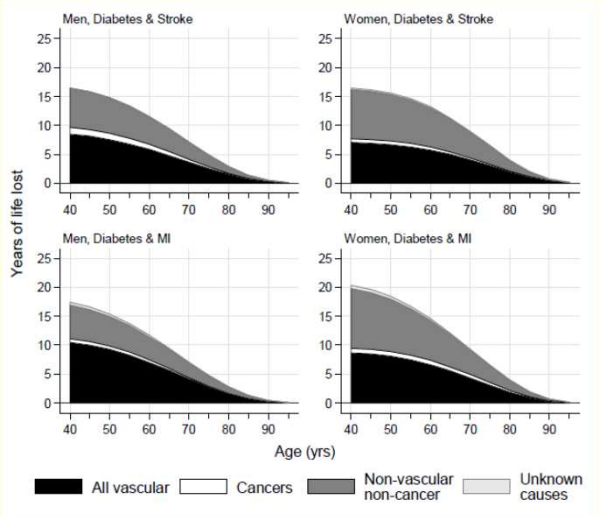
18

Estimated future years of life lost associated with disease status at baseline



19

Estimated future years of life lost associated with diabetes and stroke / MI, attributable to vascular, cancer and other causes



20

Estimating causal relationships using Mendelian randomization

21

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

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

ERFC, Lancet 2010

Is CRP causally related to CHD?

22

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)

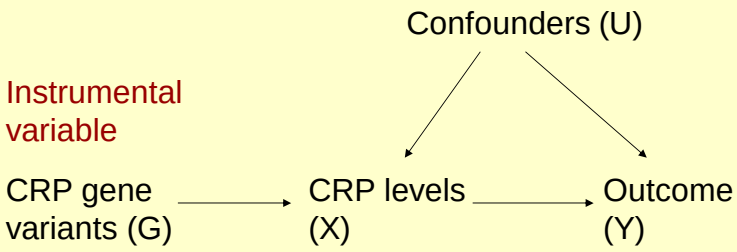
Four pre-specified genetic variants (SNPs)
on the CRP-regulatory gene on chromosome 1

Blood CRP concentrations in most studies

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

23

Diagram of causal effects



Three crucial assumptions:

G affects X

G is not related to U

Y is conditionally independent of G given X and U

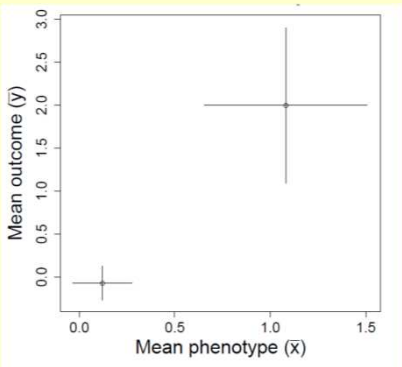
24

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:

$$\text{causal effect} = \frac{\Delta \text{ log odds of CHD}}{\Delta \text{ mean phenotype}}$$

25

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:

$x_i \sim N(\xi_i, \sigma^2)$
$\xi_i = \alpha_0 + \alpha_1 g_i$
$y_i \sim Bin(1, \pi_i)$
$\text{logit}(\pi_i) = \beta_0 + \beta_1 \xi_i$

β_1 is the causal effect estimate (increase in log odds of event per unit increase in phenotype)

Two-stage or one-stage approach possible

26

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 $k=1\ldots K$ SNPs

Additive linear (per allele) model at individual level:

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

$$\xi_i = \alpha_0 + \Sigma_k \alpha_k g_{ik}$$

$$y_i \sim Bin(1, \pi_i)$$

$$logit(\pi_i) = \beta_0 + \beta_1 \xi_i$$

β_1 is the causal effect estimate

27

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 $k=1\ldots K_m$

Additive linear (per allele) model at individual level:

$$x_{im} \sim N(\xi_{im}, \sigma_m^2)$$

$$\xi_{im} = \alpha_{0m} + \Sigma_k \alpha_{km} g_{ikm}$$

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

$$logit(\pi_{im}) = \beta_{0m} + \beta_{1m} \xi_{im}$$

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

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

28

Bayesian implementation

Vague priors:

Wide normal $N(0,100^2)$ on regression parameters

Wide uniform U[0,20] on standard deviations

MCMC using WinBUGS

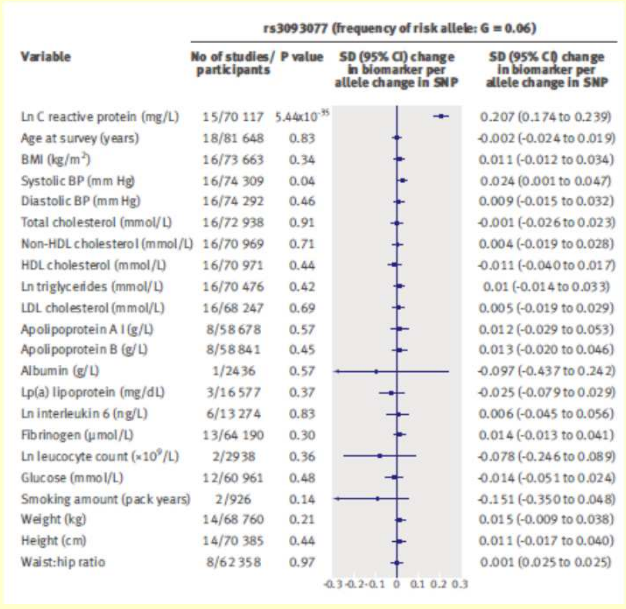
Propagates uncertainty from each stage

Allows feedback from each stage

Allows inclusion of studies with no blood CRP data

29

Are genetic variants instrumental variables?



Variable	rs3093077 (frequency of risk allele: G = 0.06)			
	No of studies/ participants	P value	SD (95% CI) change in biomarker per allele change in SNP	SD (95% CI) change in biomarker per allele change in SNP
Ln C reactive protein (mg/L)	15/70 117	$5.44x10^{-35}$		0.207 (0.174 to 0.239)
Age at survey (years)	18/81 648	0.83		-0.002 (-0.024 to 0.019)
BMI (kg/m^2)	16/73 663	0.34		0.011 (-0.012 to 0.034)
Systolic BP (mm Hg)	16/74 309	0.04		0.024 (0.001 to 0.047)
Diastolic BP (mm Hg)	16/74 292	0.46		0.009 (-0.015 to 0.032)
Total cholesterol (mmol/L)	16/72 938	0.91		-0.001 (-0.026 to 0.023)
Non-HDL cholesterol (mmol/L)	16/70 969	0.71		0.004 (-0.019 to 0.028)
HDL cholesterol (mmol/L)	16/70 971	0.44		-0.011 (-0.040 to 0.017)
Ln triglycerides (mmol/L)	16/70 476	0.42		0.01 (-0.014 to 0.033)
LDL cholesterol (mmol/L)	16/68 247	0.69		0.005 (-0.019 to 0.029)
Apolipoprotein A I (g/L)	8/58 678	0.57		0.012 (-0.029 to 0.053)
Apolipoprotein B (g/L)	8/58 841	0.45		0.013 (-0.020 to 0.046)
Albumin (g/L)	1/2436	0.57		-0.097 (-0.437 to 0.242)
Lp(a) lipoprotein (mg/dL)	3/16 577	0.37		-0.025 (-0.079 to 0.029)
Ln interleukin 6 (ng/L)	6/13 274	0.83		0.006 (-0.045 to 0.056)
Fibrinogen (µmol/L)	13/64 190	0.30		0.014 (-0.013 to 0.041)
Ln leucocyte count (×10^9/L)	2/2938	0.36		-0.078 (-0.246 to 0.089)
Glucose (mmol/L)	12/60 961	0.48		-0.014 (-0.051 to 0.024)
Smoking amount (pack years)	2/926	0.14		-0.151 (-0.350 to 0.048)
Weight (kg)	14/68 760	0.21		0.015 (-0.009 to 0.038)
Height (cm)	14/70 385	0.44		0.011 (-0.017 to 0.040)
Waist:hip ratio	8/62 358	0.97		0.001 (0.025 to 0.025)

-0.3 -0.2 -0.1 0 0.1 0.2 0.3

30

Principal results of CCGC

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

	Studies / Cases	Causal est. (95%CI)	Heterogeneity
One-stage Bayesian analysis:			
All studies	43 / 36463	–0.013 (–0.115 to 0.094)	τ=0.106

31

Interpretation of principal result

Estimate of β_1	-0.013
95% CI	(-0.115 to 0.094)
Estimate of τ	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

32

Conclusions

IPD has enabled:

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

33

References

Measurement errors

Wood AM et al. Regression dilution methods for meta-analysis: assessing long-term variability in plasma fibrinogen among 27,247 adults in 15 prospective studies. *Int J Epidemiol* 2006; **35**: 1570-1578.

Wood AM, White IR, Thompson SG. Correcting for multivariate measurement error by regression calibration in meta-analyses of epidemiological studies. *Stat Med* 2009; **28**: 1067-1092.

Life expectancy

Emerging Risk Factors Collaboration. Diabetes mellitus, fasting glucose and risk of cause-specific death. *NEJM* 2011; **364**: 829-841.

Emerging Risk Factors Collaboration. Association of cardiometabolic multimorbidity with mortality. *JAMA* 2015; **314**: 52-60.

Mendelian randomization

Burgess S, Thompson SG. Bayesian methods for meta-analysis of causal relationships estimated using genetic instrumental variables. *Stat Med* 2010; **29**: 1298-1311.

Burgess S, Thompson SG. Methods for meta-analysis of individual participant data from Mendelian randomisation studies with binary outcomes. *Stat Meth Med Res* 2016; **25**: 272-293.

34