Enhancing causal inference in observational research through the study of twin pairs and families using linked population-level data

Hopper J[1], Li S[1], Bui M[1], Lynch J[2], Callander E[3], Edwards B[4], Ferreira P[5], Ferreira L[1], Makalic E[1], Scurrah K[1], Nassar N[6], Preen D[7]

[1]Centre for Epidemiology and Biostatistics, The University of Melbourne, Parkville, Australia
[2]School of Public Health, The University of Adelaide, North Terrace, Australia
[3]School of Medicine, Griffith University, Gold Coast, Australia
[4]College of Arts and Social Sciences, The Australian National University, Canberra, Australia
[5]Charles Perkins Centre, The University of Sydney, Sydney, Australia
[6]The Children's Hospital at Westmead Clinical School, The University of Sydney, Sydney, Australia
[7]School of Population and Global Health, The University of Western Australia, Perth, Australia

Causality – the Holy Grail

- Bradford-Hill didn't invent "criteria" – just "Guidelines"

  Doll R. Fisher and Bradford Hill: their personal impact.
  Int J Epidemiol 2003;32:929–931

- If causal then … a set of Consequences

- Does not mean that the Consequences imply causality!

- Longitudinal data is also not sufficient to imply causation

Familial factors are not all genetic

- Vast majority of diseases and conditions are familial
- UK BioBank asked about family history for 12 major diseases
- Affected first-degree relative associated with risk
- For some, affected spouse/partner associated with risk
- On average ~40% of variability in ‘liability’ was familial
- At least one-third of this was not genetic

Familial factors are important

- 2-fold increased familial risk implies underlying familial risk factors (genetic or non-genetic)
- Must have >20-fold inter-quartile risk ratio
- Familial risk factors in total must be stronger than all known genetic, environmental and lifestyle-related risk factors
- Leaves open the prospect of uncontrolled familial confounding

Twin and family designs: higher quality evidence

Understanding Nature and Nurture

- Why are twins and relatives similar?
- Similarity implies there exists familial causes
- Comparing the extent to which MZ, DZ and non-twin pairs share genes and/or environment to estimate genetic and environmental causes of variation
- But these are not necessarily causes *per se*

Twin and family designs: higher evidence

Controlling for Nature and Nurture

- Why are twins and relatives different?
- Using twin pairs differing in exposure, or disease
- Derive risk estimates not confounded by familial factors
- Can even study sex differences this way using DZO pairs

Twin and family designs: higher quality evidence

Using Nature and Nurture to make inference about causation

- Does knowing about your twin's or relative's risk factor add more information about your future health than knowing about only your own risk factor?
- Tests whether associations between risk and disease are causal, or due to familial confounding

Administrative Linked Data Sets Contain Identifiable Twins and Families

Established linked data resources (researchers in brackets)	Participants	Twin pairs	Birth years
NSW perinatal linked data cohort: (Nassar, Falster)	1,704,000	23,780	1994-2017
SA Early Childhood Data Project (SA): (Lynch)	280,000	3,780	1999-2013
Victorian Data Linkage	1,574,000	24,600	1993-2018
Queensland perinatal linked data cohort: (Callander)	63,000	9,555	2007-2016
WA perinatal linked data cohort: (Preen, Pereira)	1,075,000	14,000	1974-2016

Zeltzer et al. *Acta Paediatrica* 2019 DOI:10:1111/apa.14966

Inference about Causation from Examination of FAmiliaL CONfounding (ICE FALCON)

- Regression model; cf. Mendelian Randomization (MR)
- Co-twin exposure acts as an ‘instrumental variable’ (IV)
- In MR, genetic variants for the exposure acts as an IV
- MR restricted to traits already genetically characterised
- Genetic characterisation must be conducted on all subjects
- MR makes assumptions that aren’t tested
- ICE FALCON is more broadly applicable, tests goodness-of-fit and assumptions, and has more power per subject

Atopic dermatitis and skin disease

Does eczema in infancy cause hay fever, asthma, or both in childhood? Insights from a novel regression model of sibling data

John L. Hopper, BSc, MSc, PhD, BA,[a] Quang M. Bui, BSc (Hons), PhD,[a] Bircan Erbas, BSc, MSc, PhD,[b] Melanie C. Matheson, BSc, MAppSc, PhD,[a] Lyle C. Gurrin, BSc (Hons), PhD, AStat, FRSS,[a] John A. Burgess, MBBS, FRACP, MEpid, PhD,[a] Adrian J. Lowe, BBSc, MPH, PhD,[a] Mark A. Jenkins, BSc (Hons), PhD,[a] Michael J. Abramson, MBBS, BMedSc (Hons), PhD, FRACP, FAFPHM,[c] E. Haydn Walters, BM BCh (Hons), MA, DM, FRCP, FRACP,[d] Graham G. Giles, BSc, MSc, PhD,[a,e] and Shyamali C. Dharmage, MBBS, MSc, MD, PhD[a] *Melbourne, Bundoora, Hobart, and Carlton, Australia*

International Journal of Obesity (2019) 43:243–252
https://doi.org/10.1038/s41366-018-0103-4

ARTICLE

Genetics and epigenetics

Inference about causation between body mass index and DNA methylation in blood from a twin family study

Shuai Li[1] · Ee Ming Wong[2,3] · Minh Bui[1] · Tuong L Nguyen[1] · Ji-Hoon Eric Joo[2,3] · Jennifer Stone[4] · Gillian S Dite[1] · Pierre-Antoine Dugué[1,5] · Roger L Milne[1,5] · Graham G Giles[1,5] · Richard Saffery[6,7] · Melissa C Southey[2,3] · John L Hopper[1]

Received: 14 November 2017 / Revised: 19 March 2018 / Accepted: 4 April 2018 / Published online: 17 May 2018

Li *et al. Clinical Epigenetics* (2018) 10:18
https://doi.org/10.1186/s13148-018-0452-9

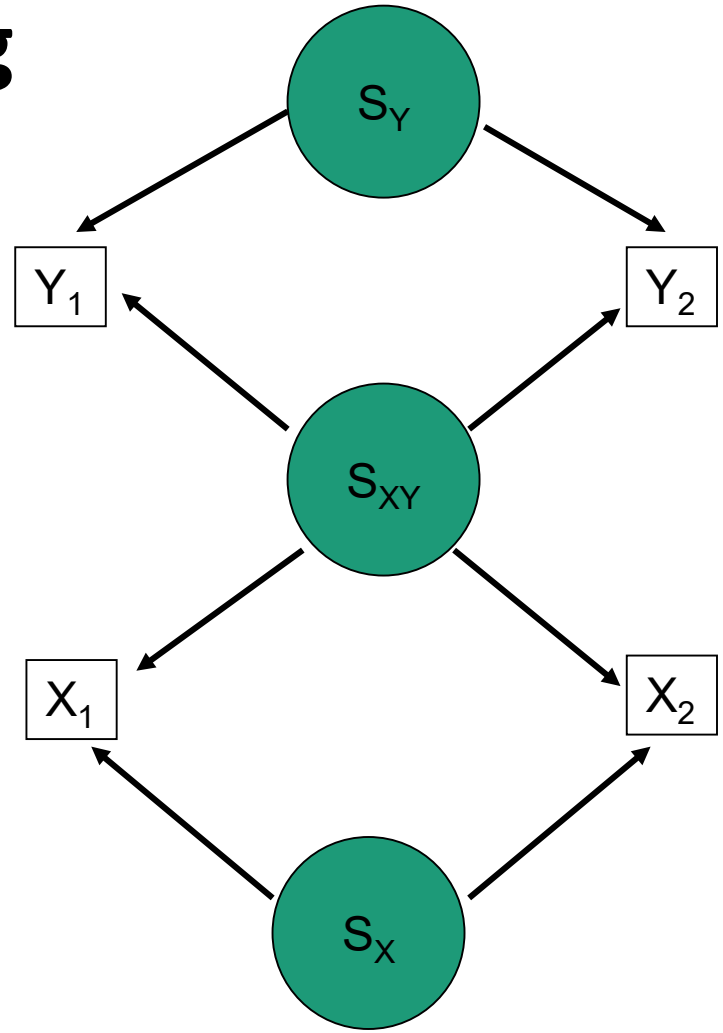
Clinical Epigenetics

RESEARCH Open Access

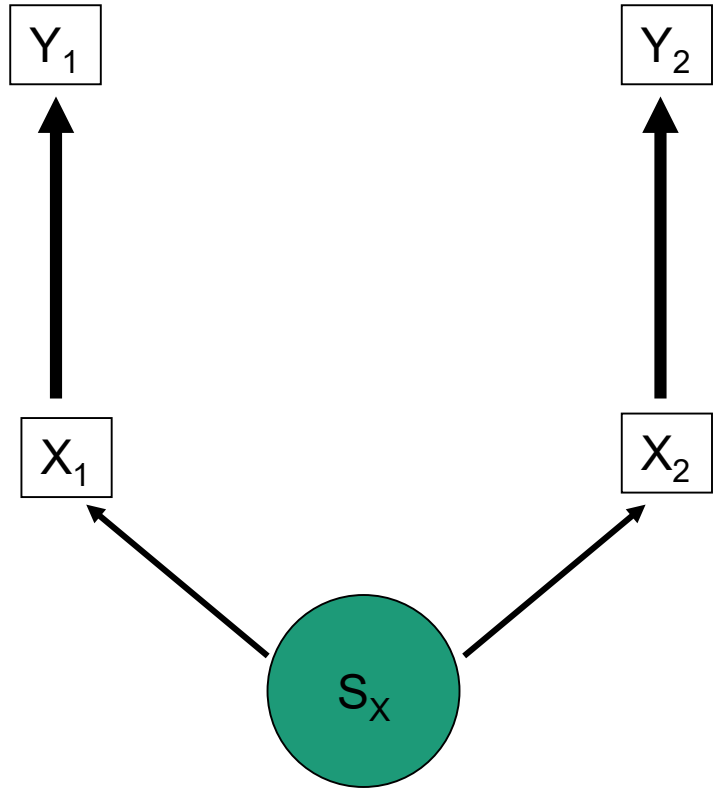
Causal effect of smoking on DNA methylation in peripheral blood: a twin and family study

Shuai Li[1], Ee Ming Wong[2,3], Minh Bui[1], Tuong L. Nguyen[1], Ji-Hoon Eric Joo[2,3], Jennifer Stone[4], Gillian S. Dite[1], Graham G. Giles[1,5], Richard Saffery[6,7], Melissa C. Southey[2,3] and John L. Hopper[1*]

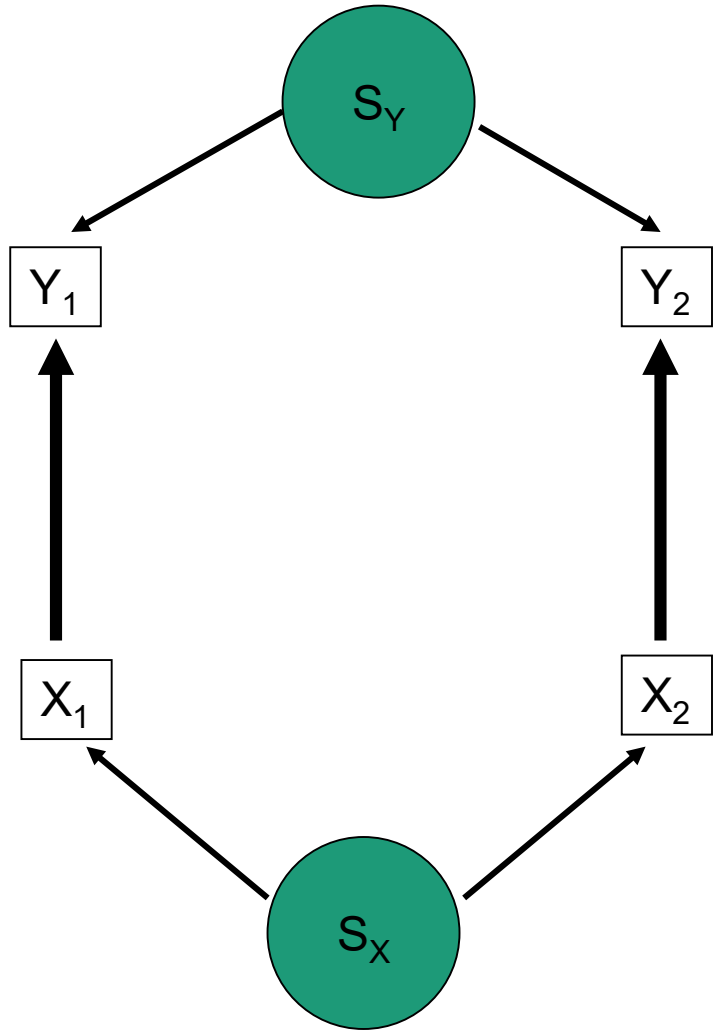
Familial confounding only



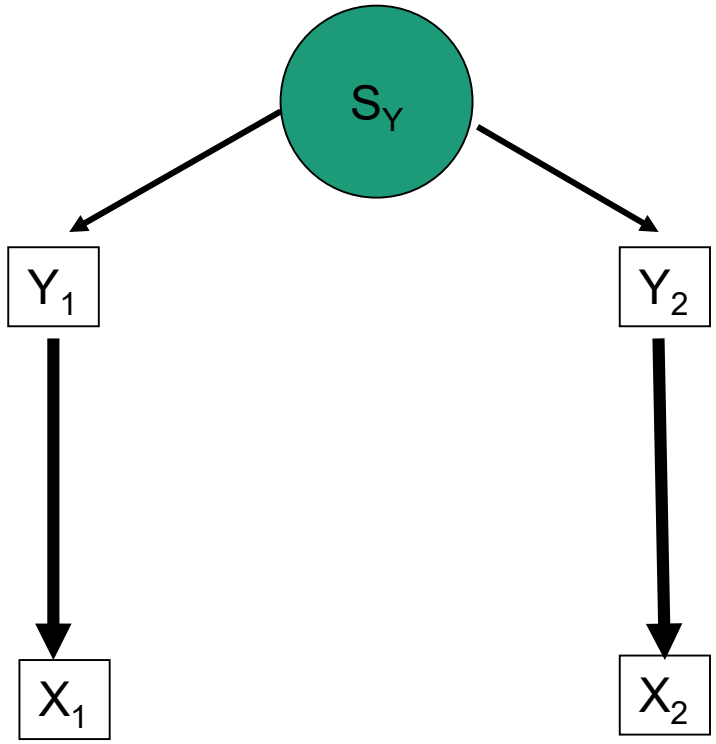
X causes Y



X causes Y



Y causes X



Expected results from ICE FALCON analysis of regressing Y on X for different causal scenarios

Model		Coefficient	Familial confounding	X causes Y	Y causes X
X as pre-dictor	Model 1	b_{self}	Association	Association	Association
	Model 2	$b_{co\text{-}twin}$	Association	Association	No association
	Model 3	b'_{self}	Association; attenuated compared with b_{self} of Model 1	Association; the same as b_{self} of Model 1	Association
		$b'_{co\text{-}twin}$	Association; attenuated compared with $b_{co\text{-}twin}$ of Model 2	No association; attenuated compared with $b_{co\text{-}twin}$ of Model 2	Association; negative if $r_X > r_Y$, otherwise positive

US World War II Veteran Twin Study

- Twins identified from induction into the armed forces (B)
- Surveyed 25-30 years later (F1), then 10-15 years later (F2), and again 10-15 years later (F3)
- 1320 twin pairs (720 MZ, 600 DZ) with complete F1 data
- BMI data at B, F1, F2 and F3
- Type II diabetes diagnosed at F2 and F3

BMI at induction as a predictor of BMI 20 years later

		Model I	Model II	Model III	% change	P
DZ	b_{self}	0.620 (0.017)		0.621 (0.018)	0	0.8
	$b_{co\text{-}twin}$		0.216 (0.019)	0.000 (0.015)	**100**	**10^{-64}**
MZ	b_{self}	0.543 (0.018)		0.452 (0.039)	17	10^{-16}
	$b_{co\text{-}twin}$		0.429 (0.019)	0.188 (0.031)	**56**	**10^{-46}**

- Consistent with BMI in early adulthood having causal effect on future BMI
- Familial confounding alone does not explain tracking in BMI for MZ pairs, but is unlikely to be due to their genes!

BMI as a predictor of Type II diabetes at a later age

		Model I	Model II	Model III	% change	P
DZ	b_{self}	0.225 (0.029)		0.217 (0.029)	4	0.2
	$b_{co-twin}$		0.091 (0.027)	0.028 (0.031)	**70**	**<.001**
MZ	b_{self}	0.218 (0.028)		0.195 (0.029)	11	0.001
	$b_{co-twin}$		0.157 (0.030)	0.042 (0.031)	**73**	**<.001**

- BMI at induction (B) and at F1 predicted diabetes at F2 and F3
- Not as well as did change in BMI from B to F1

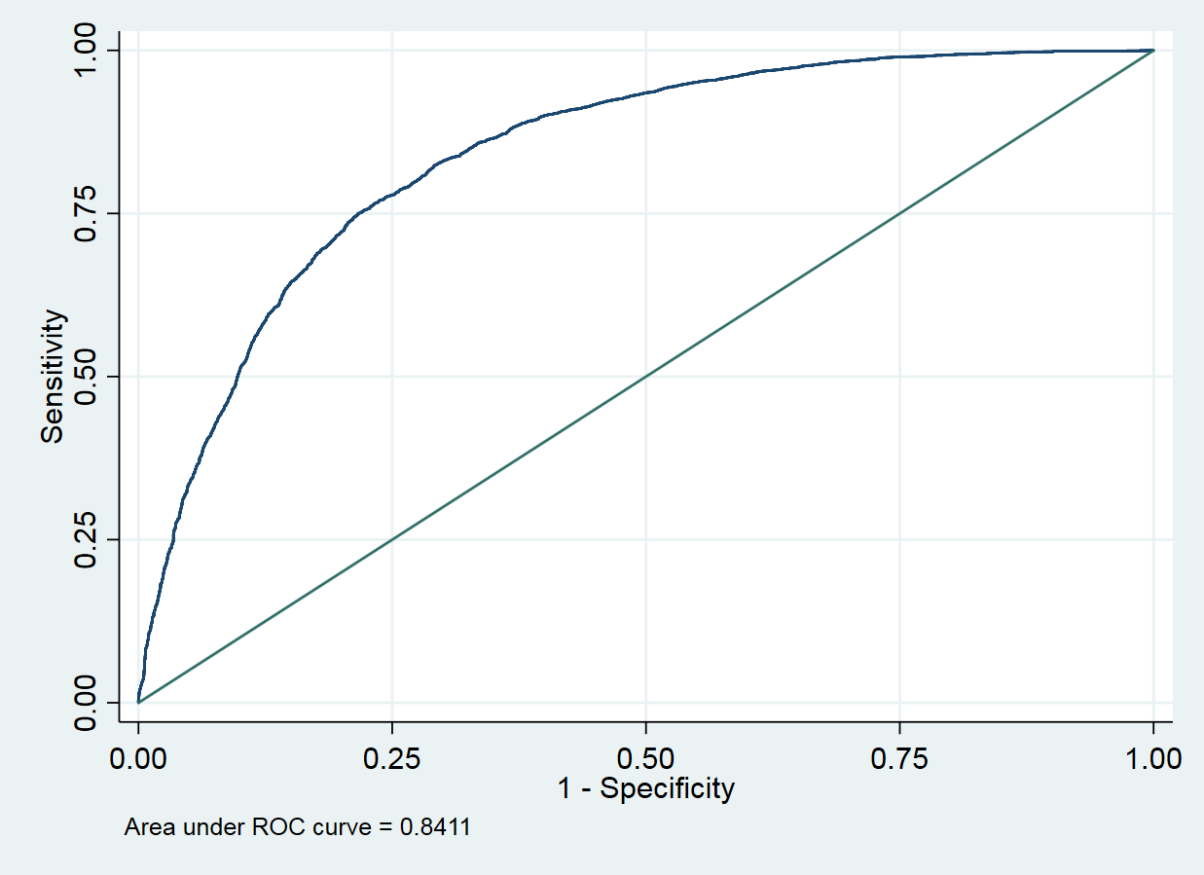
Conclusions

- Twin and family studies can be conducted within many linked population-complete administrative datasets
- Identify causal and non-causal exposure-outcome relationships
    - Existence and likely sources of familial causes of variation
    - Causal inference controlling for familial confounders
    - Inference on causation using ICE FALCON
- Potential interventions; exclude interventions doomed to failure
- Ensure translation of Big Data health research brings real benefits

UK BioBank

- 500,000 people
- Genome wide assays 500,000 SNPs
- 30,000 siblings identified

Zygosity prediction using WWI Twin Veterans data



Best predictors:

- Within-pair differences in height, weight, systolic and diastolic blood pressure, eye colour and mortality (at future follow ups)

Zygosity prediction using WWI Twin Veterans data

Logistic regression — Number of obs = 11,042; LR chi2(12) = 4374.00; Prob > chi2 = 0.0000; Log likelihood = -5408.6039; Pseudo R2 = 0.2879

zygosity	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
z_height_pdiff	.2950782	.010482	-34.36	0.000	.2752328	.3163545
z_weight_pdiff	.390008	.0135424	-27.12	0.000	.3643484	.4174747
z_dbp_pdiff	.9195671	.0243223	-3.17	0.002	.8731109	.9684951
z_sbp_pdiff	.8393104	.0222786	-6.60	0.000	.7967616	.8841314
z_height_pmean	1.125131	.0274339	4.84	0.000	1.072626	1.180206
z_weight_pmean	1.225275	.0316057	7.88	0.000	1.164869	1.288814
z_dbp_pmean	1.004675	.0279434	0.17	0.867	.9513726	1.060963
z_sbp_pmean	1.037375	.0286532	1.33	0.184	.9827092	1.095083
eye_color_pair_diff	.229078	.0140801	-23.98	0.000	.203079	.2584056
eye_color_pair_mean	.9455415	.0497928	-1.06	0.288	.852817	1.048348
final_status_pair_diff	.6958538	.0386978	-6.52	0.000	.6239949	.775988
final_status_mean	.7701029	.0588279	-3.42	0.001	.6630187	.8944823
_cons	1.166223	.0889541	2.02	0.044	1.004283	1.354277

Note: _cons estimates baseline odds.

Best predictors:

- Within-pair differences in height, weight, systolic and diastolic blood pressure, eye colour and mortality (at future follow ups)