Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Thornley et al., J Biomet Biostat 2013, 4:5
DOI: 10.4172/2155-6180.1000182

Research Article Open Access

Using Directed Acyclic Graphs for Investigating Causal Paths for Cardiovascular Disease

Simon Thornley*, Roger J Marshall, Susan Wells and Rod Jackson

Section of Epidemiology & Biostatistics, Level 4, School of Population Health, Tamaki Innovation Campus, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

Abstract

By testing for conditional dependence, algorithms can generate directed acyclic graphs (DAGs), which may help inform variable selection when building models for statistical risk prediction or for assessing causal influence. Here, we demonstrate how the method may help us understand the relationship between variables commonly used to predict cardiovascular disease (CVD) risk.

The sample included people who were aged 30 to 80 years old, free of CVD, who had a CVD risk assessment in primary care and had at least 2 years of follow-up. The endpoints were combined CVD events, and the other variables were age, sex, diabetes, smoking, ethnic group, preventive drug use (statins or antihypertensive), blood pressure, family history and cholesterol ratio. We used the 'grow shrink' algorithm, in the bnlearn library of R software to generate a DAG.

A total of 6256 individuals were included, and 101 CVD events occurred during follow-up. The accepted causal associations between tobacco smoking and age and CVD were identified in the DAG. Ethnic group also influenced risk of CVD events, but it did so indirectly mediated through the effect of smoking. Drug treatment at baseline was influenced by a wide range of other variables, such as family history of CVD, age and diabetes status, but drug treatment did not have a 'causal' association with CVD events.

Algorithms which generate DAGs are a useful adjunct to traditional statistical methods when deciding on the structure of a regression model to test causal hypotheses.

Keywords: Cardiovascular diseases; Causality; Epidemiology; Decision support techniques

Introduction

Most risk prediction and causation models in epidemiology are based on additive combinations of risk factors in a regression model framework, and the additive structure implies that variables typically act, unless interaction effects are introduced, without influence on the other variables, to yield a risk of developing disease. Since they are simply mathematical constructs, the models do not necessarily provide a plausible causal representation of how disease develops. One method to more explicitly consider causality is to attempt to describe the influence of variables on a particular disease outcome, accounting for causal pathways that are, at least, plausible, in the form of a directed acyclic graph (DAG). These can be built using learning Bayesian network algorithms [1].

Directed acyclic graphs (DAGs), also known as probabilistic networks, or Bayesian networks, encode a structure of conditional independence between variables, represented by nodes of a graph. Connections between nodes imply causal influence, observed in the data as statistical dependence. These connections are often directed, to indicate which variable influences the other (referred to as directed edges). In this way, DAGs represent a set of conditional dependence and independence properties associated with epidemiological variables [1].

In a DAG, no distinction is made between 'independent' and 'dependent' variables in the sense used in regression modelling. The idea underlying their use is to fuse domain knowledge with information from the collected data into a model which mimics a network of causal influences of how the observed data were generated.

DAGs are therefore useful for elucidating possible causal pathways and have been applied in epidemiology for this purpose [2]. However, they also have a role in forming sensible judgements about variables to be included in regression prediction models. For example, a key idea of Pearl, who has been a proponent of DAG ideas, is that variables may act as 'colliders' [3]. That is, on a causal path between exposure and outcome, another variable on the path is entered and exited through arrowheads, which indicate more than one influence (collision of influences) on the variable. Here, we interchangeably use the terms 'cause' and 'influence' to indicate directional conditional dependence, or a link between variables, generated by a computer algorithm.

This idea of including an explicit causal understanding is absent from much statistical analysis. Including colliders as regressors can result in unpredictable behaviour, biasing measures of association in a regression model. Pearl shows that bias may increase, by introducing dependence from unobserved or other variables, rather than reduce, after their inclusion. Further, in certain instances, adjusting for colliders, or their 'descendants', that is variables which are causally influenced by colliders, may indicate no causal influence between the variable of interest and the regression model's outcome variable, when in fact a causal relationship does exist [3]. DAGs, derived from data, may help identify such variables, so that they can be omitted, rather than be included in regression models.

***Corresponding author:** Simon Thornley, Section of Epidemiology & Biostatistics, Level 4, School of Population Health, Tamaki Innovation Campus, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand, Tel: 64 9 373 7599, 0212991752; E-mail: sithor@gmail.com

Received October 08, 2013; **Accepted** November 30, 2013; **Published** December 03, 2013

Citation: Thornley S, Marshall RJ, Wells S, Jackson R (2013) Using Directed Acyclic Graphs for Investigating Causal Paths for Cardiovascular Disease. J Biomet Biostat 4: 182. doi:10.4172/2155-6180.1000182

To develop prediction models, we believe that a causal understanding is likely to lead to more accurate and reliable predictions than those developed using standard statistical methods alone [4,5]. In this study, we explore a small database of CVD and associated risk factors using DAG techniques to inform variable selection for risk prediction models, and, it is hoped, to better explain the development of CVD.

Methods

The analysis is based on a cohort assembled by primary care practitioners in the Auckland and Northland regions of New Zealand using the PREDICT programme for CVD risk management that is integrated with patient electronic health records [6]. Cohort participants were patients attending their primary care practitioners who had their CVD risk formally assessed using a Framingham Heart Study risk prediction equation [7]. The information, from participating GPs, was stored on a secure project web-server and each patient was linked to national health databases via an encrypted version of the New Zealand national health index (the NHI) number. This unique number is allocated to all New Zealand residents and attached to their routine health records.

Databases that were linked included: hospital discharges, mortality, and drug dispensing. We selected a group of individuals enrolled between the 1st of Jan 2006 to the 31st of December 2007.Two years after the baseline assessment we determined if they had been admitted to hospital with CVD or died from CVD or other causes by consulting hospital diagnosis and cause-of-death information. Diagnosis codes that were used have been listed elsewhere [6].

Individuals under 30 or over 80 years of age at the time at the time of screening were excluded because CVD is uncommon under 30 years and hospital diagnoses are known to be less accurate in older people. We also excluded people with a history of prior CVD or heart failure, identified by a general practitioner diagnosis of CVD or hospitalisation with CVD in the last five years, or those dispensed a loop diuretic in the six months before assessment, who were assumed to have heart failure. The variables which were considered as candidates in the DAG were: age-at-enrolment, sex, diabetes, smoking, ethnic group, family history of premature CVD, statin use, antihypertensive drug use, systolic blood pressure, total: high-density-lipoprotein (HDL) cholesterol ratio and CVD events during follow-up. Continuous variables were categorised, mostly into deciles, as this format is required for the particular DAG algorithm (see below) that we selected. The categorical variables were included as dummy variables without an ordinal structure.

The R package bnlearn drew the DAG, using the 'growshrink' algorithm, first developed by Pearl [8]. An understandable summary of the algorithm has been documented elsewhere [1,9,10]. The algorithm effectively filters links out of a full skeletal DAG, in which all nodes are initially connected except those 'banned' (see below), based on tests of conditional independence between a pair of nodes given all possible subsets of the rest. We used the Monte Carlo permutation tests [11] option which has performed better in simulations in which the causal structure of the data is known, compared to standard chi-square tests [8]. Logical rules are applied to determine the direction of links (conditional dependence between variables), so that cycles are not introduced and patterns of conditional independence found in the data match the generated DAG.

We estimated link influence in the final DAG by estimating the beta-coefficient for a regression for each potential causal effect in which the variable at the base of the arrow ('cause') was considered a covariate, and the variable at the head of the arrow ('effect') was considered the outcome or dependent variable. Other variables which opened 'back door paths' (Pearl's terminology for confounding) between cause and effect variables were included as covariates in the regression. Either linear or logistic regression was used depending on whether the 'effect' variable was continuous or categorical.

For the link between ethnic group and family history of disease, we adjusted for age. For, although age directly causes CVD, it does not influence ethnic group, and is in fact 'banned' (see below), so does not qualify as a confounder. Age does, however, modify the risk of an individual reporting a positive family history of CVD and so we felt that it was sensible to adjust for age in this instance [12,13]. Other adjustments in the regressions are indicated in Table 1.

The bnlearn algorithm allows implausible causal influences to be 'banned'. The following rules generated the banned list:

- Sex, ethnic group and age must not be caused by any other variable.
- Family history must not be caused by drug treatment variables.
- he outcome, fatal and nonfatal CVD, must not cause any other variable.

Results

After the selection criteria were applied, 6256 subjects were available for analysis, 101 (1.6%) of whom experienced a CVD event during follow-up, and 35 (0.6%) of whom died of causes other than CVD. Table 2 shows that age-at-enrolment, ethnic group, smoking status, antihypertensive drug use, systolic blood pressure and diabetes status were significantly associated with event status. Among ethnic groups, Maori were at highest risk of a CVD event (estimated odds ratio: 1.87; 95% CI: 1.09 to 3.10). Those who used either statins or antihypertensive agents were at higher risk of CVD than non-users.

The derived DAG is depicted in Figure 1. Directed arrows indicate the direction of 'causal' influence between variables. Only two direct influences on cardiovascular disease are detected: age and cigarette smoking.

Ethnic group influences risk of cardiovascular disease, but it does so mediated through the effect of smoking. Age influences several other variables, such as family history of disease and the risk of taking preventive drug treatment. Ethnic group influences three variables: family history, smoking and diabetes status. The ratio of total: HDL-cholesterol concentration is influenced by two variables: sex and cigarette smoking.

There was no link between anti-hypertensive or statin therapy and cardiovascular disease. Also, we observed that commonly accepted causal associations, such as systolic blood pressure and total: HDL-cholesterol ratio did not show a causal link to CVD events. This contrasts with strong univariable associations between systolic blood pressure and CVD (Table 2). The analysis, also, did not causally link statin use with the cholesterol ratio variable.

Indices of link influence are given in Table 1. These are beta-coefficients derived from regressing the cause (tail of arrow) on the effect (arrowhead), using either linear or logistic regression, adjusting for other immediately adjacent influences on the effect variable. All links between age and other variables show strong evidence of association, along with ethnic group, male sex and diabetes and their causal links. Strong associations were noted between diabetes status and use of preventive drugs.

Cause	Effect	*Low*[+]	*High*[+]	*Beta-coeff.(95% CI)*	*Estimated odds ratio (95% CI)*
Age	CVD	43.4	65.2	1.54 (1.11, 1.97)	4.65 (3.03, 7.14)
Age	Statin use	43.4	65.2	0.84 (0.69, 0.99)	2.31 (1.99, 2.69)
Age	Anti-hypertensive	43.4	65.2	1.44 (1.31, 1.57)	4.23 (3.72, 4.82)
Age	Family history of CVD	43.4	65.2	-0.31(-0.43, -0.20)	0.73 (0.65, 0.82)
Age	Systolic blood pressure	43.4	65.2	10.42 (9.5, 11.34)	N/A
Age	Sex (men)	43.4	65.2	-0.73 (-0.83, -0.62)	0.48 (0.43, 0.54)
Ethnic group	Diabetes	Other	Indian	1.64(1.33, 1.95)	5.14 (3.78, 7.00)
		Other	Maori	1.03(0.84,1.23)	2.81 (2.31, 3.42)
		Other	Pacific	1.86(1.68, 2.04)	6.44(5.39, 7.70)
Ethnic group	Smoker	Other	Indian	-0.50 (-0.99, -0.01)	0.60(0.37, 0.99)
		Other	Maori	1.28(1.12, 1.45)	3.60(3.05, 4.25)
		Other	Pacific	0.72(0.54, 0.91)	2.06 (1.72, 2.48)
Ethnic group (adj. for age)	Family history of CVD	Other	Indian	0.02(-0.28, 0.32)	1.02(0.75, 1.37)
		Other	Maori	-0.24(-0.41, -0.07)	0.79 (0.67, 0.93)
		Other	Pacific	-1.03 (-1.24, -0.82)	0.36 (0.29, 0.44)
Diabetes (adj. for age)	Statin use	No	Yes	1.94 (1.77, 2.10)	6.94 (5.90, 8.16)
Diabetes (adj. for age)	Antihyper-tensive use	No	Yes	1.68 (1.53, 1.84)	5.38 (4.60, 6.28)
Statin use(adj. for age)	Antihyper-tensive use	No	Yes	1.70 (1.55, 1.86)	5.49 (4.69, 6.42)
Anti-hypertensive (adj. for age)	Systolic blood pressure	No	Yes	7.30(6.28, 8.33)	N/A
Smoker (no adj.)	CVD	No	Yes	0.59(0.15, 1.03)	1.80 (1.16, 2.79)
Smoker (no adj.)	Total: HDL-cholesterol ratio	No	Yes	0.51 (0.43, 0.59)	N/A
Family history (adj. for ethnic group and age)	Statin Use	No	Yes	0.42 (0.26, 0.58)	1.52 (1.30, 1.79)
Sex (no adj.)	Total: HDL-cholesterol ratio	Female	Male	0.61(0.55, 0.67)	N/A

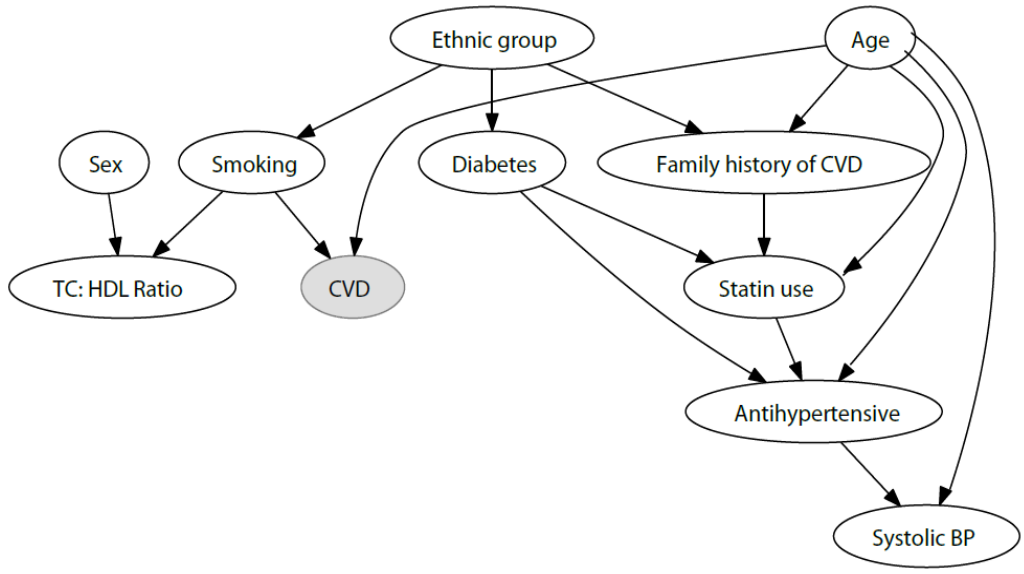
CVD: Cardiovascular disease. HDL: high density lipoprotein. adj.: adjustment. N/A: not applicable.
[+]for age, comparisons were made at the 84th and 16th centiles to allow comparison with measures of effects from binary variables [21].

Table 1: Lists of causal links andestimated beta-coefficients from linear or logistic regression, adjusted for variables, as indicated in the DAG). Also odds ratios for logistic regression models.

	CVD	No CVD	Total	Test stat.	P-value
Total	101	6155	6256		
Gender				Chisq. (1 df)	0.343
Men	61 (60.4)	3395 (55.2)	3456 (55.2)		
Age at enrolment				t-test (6254 df)	< 0.001
Mean(SD)	61.7 (10.2)	54.1 (10.5)	54.2 (10.5)		
Ethnic group				Fisher's exact	
test (3 df)	0.036				
Other	62 (61.4)	4348 (70.6)	4410 (70.5)		
Maori	22 (21.8)	826 (13.4)	848 (13.6)		
Pacific	16 (15.8)	773 (12.6)	789 (12.6)		
Indian	1 (1.0)	208 (3.4)	209 (3.3)		
Smoking status				Chisq. (1 df)	0.012
Yes	28 (27.7)	1082 (17.6)	1110 (17.7)		
Statin treatment at baseline?			Chisq. (1 df)	0.127	
Yes	20 (19.8)	860 (14.0)	880 (14.1)		
Antihypertensive treatment at baseline?			Chisq. (1 df)	<0.001	
Yes	48 (47.5)	1637 (26.6)	1685 (26.9)		
Systolic blood pressure (mmHg)			Ranksum test	<0.001	
Median(IQR)	140 (130,150)	130 (120,142)	130 (120,143)		
Diagnosis of diabetes?	Chisq. (1 df)	0.0143			
Yes	24 (23.8)	896 (14.6)	920 (14.7)		
Total to HDL-cholesterol ratio			Rank sum test	0.744	
Median(IQR)	3.7 (3.1, 4.8)	3.8 (3.1, 4.7)	3.8 (3.1, 4.7)		
Premature Family history?				Chisq. (1 df)	0.520
Yes	31 (30.7)	1681 (27.3)	1712 (27.4)		

IQR: Interquartile range. HDL: high density lipoprotein. CVD: cardiovascular disease. Stat: statistic. df: degrees of freedom. Chisq: chi-square test of independence

Table 2: Sample characteristics by cardiovascular disease status: numbers (% of column population unless otherwise stated).



CVD: Cardiovascular disease. HDL: high-density lipoprotein cholesterol concentration. TC: total cholesterol concentration.

Figure 1: DAG, derived from the grow-shrink algorithm. Grayed boxes indicate outcome variables.

From the logistic regression analyses, the greatest odds ratios were between ethnic group and diabetes status. Pacific people were4.4 times more likely than 'Others' to be diagnosed with diabetes (estimated OR: 6.44, 95% CI: 5.39, 7.70; prevalence of diabetes among Others: 8.6%) and Indian people were almost four times more likely than 'Others' to have the diagnosis in this cohort (estimated OR: 5.14, 95% CI: 3.78 to 7.00). For continuous outcome measures, those who used anti-hypertensive drugs had an average systolic blood pressure 7.30 mmHg (95% CI: 6.28 to 8.33) higher than people who did not use these drugs.

Discussion

In this exploratory analysis with a relatively small dataset, we have shown that a DAG learning algorithm generated a plausible graph explaining the occurrence of cardiovascular disease. The DAG captures the two known key causal influences of CVD: age and cigarette smoking. It also demonstrates the well-known influence of age on other variables, such as systolic blood pressure [14] and preventive drug use [15]. Positive or higher values of these variables increased with advancing age.

The DAG may help inform variable selection decisions for regression modelling to establish magnitude of effects. For example, from our data, ethnicity influences diabetes, cigarette smoking, and family history of premature CVD. These 'causal' relationships indicate that in trying to assess the effects of ethnic group on CVD, adjusting for any of these mediating variables will bias the association. It is equivalent to adjusting for blood pressure level when investigating if there is a causal relationship between body mass and CVD, as blood pressure is on the causal pathway. Similarly, the DAG simplifies assessing the influence of potential confounding factors on CVD incidence. It also suggests that none of the other baseline variables confounds the relationship between ethnicity and CVD, since no other variable directly influences ethnic group. Thus, when assessing the causal effect of ethnicity, it may only be necessary to adjust for age, since, as we argued before, it is a modifier of the effect of ethnic group on CVD.

An interesting feature of the DAG was the link between age and reported family history, showing a negative relationship. This may reflect the belief and reporting practice of the physician, who may only enquire about family history of CVD in younger patients, assuming that older patients will not have a family history. An alternative assumption is that genetic causes of CVD only manifest disease in younger patients, so that older patients, when risk assessed, are assumed not to have a genetic predisposition.

Again, if this DAG were a valid representation of causality it would suggest that very few of the variables that were measured actually cause CVD, so in assessing the effect of various exposures, some adjustment may cause more harm than good. It also counters the common practice in clinical research of reporting 'independent risk factors' after adjusting for a number of other variables by regression [16] and considering them as causal.

The DAG presented here also may help identify what Pearl terms 'barren proxies' when assessing causal influences. These are variables which have no direct influence on either the exposure or outcome variables, but are themselves causally influenced by factors that are either related to exposure or disease, or possibly both. In this sense, they could be considered as proxy measures of either exposure or disease. For example, consider a scenario in which one was to investigate the statistical evidence for a causal link between sex and CVD incidence. In this case, including the cholesterol ratio variable as a covariate, which, in this data set is influenced by sex, but does not show convincing evidence of influencing disease status, may increase (rather than reduce) bias in estimating the strength of association between sex and CVD in a regression model. Thus, in this dataset the cholesterol ratio would be termed a barren proxy. As with the ethnicity example above, the value of excluding the cholesterol ratio in a causal analysis is distinct from the value which the cholesterol ratio variable may play in predicting disease incidence.

Some known links emerged from the analysis, for example that between cigarette smoking and serum lipids has been long described [17]. The DAG did not, however, directly link serum lipids with CVD. In addition, the DAG and the effect estimates in Table 1 identified that

anti-hypertensive drug treatment increases systolic blood pressure. As anti-hypertensive drugs are known to lower blood pressure, this, at first, seems counter-intuitive. However, the drug use data is collected before the blood pressure information so this is the only sensible direction for the link to be oriented. The orientation of the link means that people were taking the drugs simply because their blood pressure was high, and that, on average, treated individuals had a higher blood pressure on average, than untreated individuals (7.3 mmHg, adjusted for age) when they were screened.

This analysis clearly has some limitations. These include the likelihood that some associations are not identified because of type-2 errors (only 101 CVD events occurred). There may also be information bias and unmeasured variables that could affect the nature of the DAG. However the main objective of this paper is to demonstrate the potential of the DAG learning algorithm rather than add to our knowledge of CVD risk. A further limitation of the DAG algorithm is that it does not deal with time-to-event data, commonly used in cohort studies, which may be censored. In these analyses we used a short, two year period of follow-up, and in this period there were few losses to follow up, mostly from non-CVD deaths.

There are a few other studies which have used learning Bayesian networks to explore similar datasets. Twardy et al. [18] used Bayesian network algorithms, based on minimisation of information metrics, to determine the causal structure of the data in two cohort studies of cardiovascular disease. The authors did not exclude, or ban, implausible relationships, as in our study. Also, their study was limited by a high proportion of cases in which some covariates were missing. In their 'final' model, several implausible relationships were present, such as diabetes and weight influencing age. Their model described age as the only influence on coronary heart disease and had some similar findings to our study, of age influencing many risk factors: total cholesterol, triglycerides, systolic blood pressure, smoking status and height. Unlike our study, some known causal links were included such as between diabetes and systolic blood pressure, which were not drawn in our DAG, even though it is well known that diabetes raises blood pressure [19].

To summarise the implications of our DAG for statistical modelling, we suggest that when using regression to assess the causal influences for cardiovascular disease, an analysis could be done to generate a DAG to estimate the conditional association between disease status and other variables. Only those variables which appear causally related – that is, with arrows that point to disease–should be included in the model. This means, for our data we would only include age and smoking status, along with the exposure of interest, in a regression. Other variables may be justified if they were thought to be important effect modifiers or confounders. Effect modification is not captured in the DAG, so inclusion of variables for this reason will not be informed by the DAG. If justified as confounders, researchers must think carefully about whether they are likely to act in such a way, that is, causally influence both the exposure of interest and the outcome, rather than act as 'barren proxies'. The Bradford-Hill criteria [20] may be used to guide these decisions. In contrast, for developing prediction algorithms, many variables can be used in statistical models that may be associated, but not necessarily causally related to disease.

Also, it is increasingly common practice for researchers to propose a DAG, drawn from informed scientific knowledge, which is then used to inform variable selection when testing causal relationships in observational studies. The algorithm used in this study provides a way of checking whether the assumptions encoded in the researcher-drawn DAG are actually observed in the study data.

In this exploratory study, we demonstrate how a simple DAG could shed light on the likely causal structure of risk factors for incident cardiovascular disease. The derived graph provides useful information to inform variable selection decisions when assessing causal relationships with the disease, and since they are related concepts [4], the DAG also usefully informs the development of models used for prediction.

Acknowledgements

We thank Tadd Clayton and Romana Pylypchuk for assistance with data cleaning and managing. The authors would like to thank the National Health Board Analytic Services and Diagnostic Medlab Ltd for supplying anonymized data. They would also like to thank affiliated general practitioners and practice nurses and patients belonging to the ProCare Network, Auckland PHO Ltd, Health West, East Health Services, National Māori Hauora Coalition, Alliance Health, Te Tai Tokerau and Manaia PHOs.

PREDICT was developed by a collaboration of epidemiologists at the University of Auckland, Information Technology specialists at Enigma Publishing Ltd (a private provider of online health knowledge systems), primary health care organizations, non-governmental organizations (New Zealand Guidelines Group, National Heart Foundation, Diabetes New Zealand, Diabetes Auckland), several district health boards and the Ministry of Health. The PREDICT software platform is owned by Enigma Publishing Ltd (PREDICT is a trademark of Enigma Publishing Ltd).

Funding

This work was supported by a Health Research Council Clinical Research Training Fellowship (grant number: 11/145), which supported the work of ST.

SW is partly funded by the Stevenson Foundation.

The PREDICT programme has received funding from the Health Research Council (HRC grants 03/183, 08/121, 11/800).

References

1. Kjaerulff UB, Madsen AL (2007) Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis, Springer, New York, USA.
2. Greenland S, Pearl J, Robins JM (1999) Causal diagrams for epidemiologic research. Epidemiology 10: 37-48.
3. Pearl J (2000) Causality: Models, Reasoning, and Inference. Cambridge University Press, UK.
4. Thornley S (2012) Causation and Statistical Prediction: Perfect Strangers or Bedfellows? J Biom Biostat 3: e115.
5. Kleinberg S, Hripcsak G (2011) A review of causal inference for biomedical informatics. J Biomed Inform 44: 1102-1112.
6. Riddell T, Wells S, Jackson R, Lee AW, Crengle S, et al. (2010) Performance of Framingham cardiovascular risk scores by ethnic groups in New Zealand: PREDICT CVD-10. N Z Med J 123: 50-61.
7. Bannink L, Wells S, Broad J, Riddell T, Jackson R, et al. (2006) Web-based assessment of cardiovascular disease risk in routine primary care practice in New Zealand: the first 18,000 patients (PREDICT CVD-1). N Z Med J 119: U2313.
8. Scutari M (2010) Learning Bayesian Networks with the bnlearn R Package. J Stat Soft 35: 1-22.
9. Korb KB, Nicholson AE (2011) Learning linear causal models. Bayesian Artificial Intelligence. Boca Raton. CRC Press 231-252.
10. Margaritis D (2003) Learning Bayesian Network Model Structure from Data. Carnegie-Mellon University, Pittsburgh, PA, USA.
11. Good P (2005) Permutation, Parametric and Bootstrap Tests of Hypotheses. (3rdedn), Springer-Verlag, New York, USA.
12. Rothman KJ, Greenland S, Walker AM (1980) Concepts of interaction. Am J Epidemiol 112: 467-470.
13. Ulfers MI (2008) Is age too automatically controlled for as a confounder in epidemiologic studies?: ProQuest.

14. Franklin SS, Gustin W 4th, Wong ND, Larson MG, Weber MA, et al. (1997) Hemodynamic Patterns of Age-Related Changes in Blood Pressure: The Framingham Heart Study. Circulation 96: 308-315.

15. Jeffrey JE, Steven RE, James GS, Steven JB, Renee AS, et al. (2004) Suboptimal Statin Adherence and Discontinuation in Primary and Secondary Prevention Populations. J Gen Intern Med 19: 638-645.

16. Brotman DJ, Walker E, Lauer MS, O'Brien RG (2005) In search of fewer independent risk factors. Arch Intern Med 165: 138-145.

17. Shaten BJ, Kuller LH, Neaton JD (1991) Association between baseline risk factors, cigarette smoking, and CHD mortality after 10.5 years. MRFIT Research Group. Prev Med 20: 655-659.

18. Twardy R, Nicholson AE, Korb KB, John M (2006) Epidemiological data mining of cardiovascular Bayesian networks. Electronic J Health Info 1.

19. Winocour PH, Durrington PN, Anderson DC, Cohen H (1987) Influence of proteinuria on vascular disease, blood pressure, and lipoproteins in insulin dependent diabetes mellitus. British medical journal (Clinical research ed) 294: 1648.

20. Hofler M (2005) The Bradford Hill considerations on causality: a counterfactual perspective. Emerging Themes in Epidemiology 2: 11.

21. Thornley S, Marshall RJ (2012) Measures of association in epidemiological studies: how best to compare discrete and continuous variables? J Biometrics Biostat 3: e111.