Identification of causal structures in simulated epidemiologic data

Zheng Sponsiello-Wang, Etienne Kaelin, Gerd Kallischnigg, Rolf Weitkunat
(Philip Morris International R&D, Quai Jeanrenaud 56, 2000 Neuchatel, Switzerland)

BACKGROUND

In epidemiology, quantification of potential cause-effect relations is currently dominated by regression methods. To minimize the chance of false-positive causal conclusions, the identified statistical associations are usually assessed by the pragmatic application of predefined criteria of causality. For a variety of reasons it might be more preferable to have a clearer separation of inductive (speculative) and deductive (confirmatory) steps in etiologic research. The methodology of probabilistic causal models, which has been developed in the last two decades, allows for quantitative predictions based on a priori formulated causal models.

APPROACH

The approach proposed in the present work is based on a top-down concept of scientific inquiry. It is assumed that at least one causal hypothesis in the form of a multifactor causal model is available to explain a set of empirical data. The (or each) pre-specified causal model is then confronted with the data in order to obtain an estimation of the likelihood of the model.

METHODS

Configuration Sequence Analysis

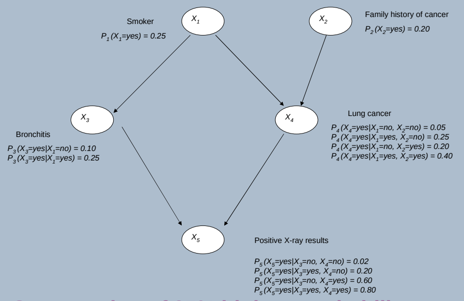
The term configuration Sequence Analysis (CSA) is proposed to denote the present conceptually-driven causal modeling approach. In order to support the application, a SAS macro was developed. The core tasks of the CSA procedure are firstly to compute unconditional and conditional probabilities for each configuration of model variables of the specified causal network based on the available data. Secondly, the conditional point estimates of the outcome variable are computed according to the model structure and are compared with the observed relative frequencies of all configurations of model variables.

Data simulation

Given the pictured causal structure, a data set was generated using the SAS (version 9.1) RANUNI routine of a randomized process without memory, being unidirectional in time. Random numbers were routinely generated from a uniform distribution on the interval of 0 to 1. The coherence of the simulated data with the preset parameters was validated. A simulated data set of 1000 data points was used to asses the fitting of the data with five concurrent causal models.

Computation of point estimates

In the context of Bayesian networks, the point estimate of the outcome variable (Pe) is defined by the hypothetical Bayesian network structure under investigation, according to the chain rule. For example, in the figure Pe=P(X_1)·P(X_2)·P(X_3|X_1)·P(X_4|X_1,X_2)·P(X_5|X_3,X_4)


Smoker
$P_1(X_1=yes) = 0.25$
X_1
X_2
Family history of cancer
$P_2(X_2=yes) = 0.20$
X_3
Bronchitis
$P_3(X_3=yes|X_1=no) = 0.10$
$P_3(X_3=yes|X_1=yes) = 0.25$
X_4
Lung cancer
$P_4(X_4=yes|X_1=no, X_2=no) = 0.05$
$P_4(X_4=yes|X_1=yes, X_2=no) = 0.25$
$P_4(X_4=yes|X_1=no, X_2=yes) = 0.20$
$P_4(X_4=yes|X_1=yes, X_2=yes) = 0.40$
X_5
Positive X-ray results
$P_5(X_5=yes|X_3=no, X_4=no) = 0.02$
$P_5(X_5=yes|X_3=yes, X_4=no) = 0.20$
$P_5(X_5=yes|X_3=no, X_4=yes) = 0.60$
$P_5(X_5=yes|X_3=yes, X_4=yes) = 0.80$

Computation of 95% highest probability density regions

Assuming that in the present analysis all the variables are dichotomous and have a binomial distribution function, denoted as X ~ B(n, π), and that the prior π has a beta distribution function, Beta(α, β). The posterior point estimate has then also a beta distribution, Beta(α+x, β+n-x), denoted as π' ~ Beta (α', β'), where α'=α+x and β'=β+n-x.

The precision of the point estimates Pe is estimated by computing the 95% highest probability density region of the posterior π'. According to Lee (2004), if π' ~ Beta(α', β'), then log(λ) is very near that of the logF distribution. Thus, a 95% highest probability density region for π', when π' ~ Beta(α', β'), can be determined by finding the values corresponding to the 95% highest probability density lower and upper values of log ($F_{2\alpha',2\beta'}$).

Model fit and model comparison

A χ^2 goodness-of-fit statistic was used to evaluate the overall model fit. In addition, deviance, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) were employed in the present analysis to select the most likely model. Assuming that the variables in the data set have a binomial distribution, the likelihood of the model, deviance, AIC and BIC are defined as follows:

$$L_m(Pe \mid n, x) = \sum_{i=1}^{K} \binom{n}{x_i} x_i^{Pe_i} (n - x_i)^{(1-Pe_i)}$$

$$Deviance: D(Pe, x) = -2\log(L_m(Pe \mid n, x))$$

$$AIC = -n \cdot \log(L_m(Pe \mid n, x)) + d$$

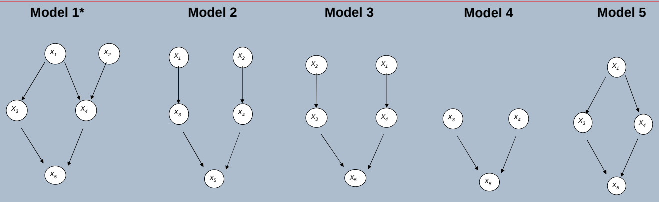
$$BIC = -n \cdot \log(L_m(Pe \mid n, x)) + \frac{d}{2} \cdot \log n$$

where d, referring to the dimension of the model, equals 2^N-1. N is the number of variables in the model. K= 2^N is number of model variable configurations.

CONCLUSIONS

1. The correct causal model could be identified with the CSA procedure.
2. The priors have substantial influence on 95% highest probability density regions only when the sample size is below about n=20.

RESULTS



	Model 1	Model 2	Model 3	Model 4	Model 5
χ^2	31.015	88.717	142.008	2619.624	14.583
df	31	31	31	7	15
P	0.4654	1.79E-7	<1.79E-7	<1.79E-7	0.4818
Deviance	-2.572	-2.538	-1.720	5.773	-0.344
AIC	-1255.168	-1237.858	-829.175	2893.403	-156.928
BIC	-1179.098	-1161.788	-753.105	2910.580	-120.119

* Underlying data-generation

						95% highest probability density regions		
X_1	X_2	X_3	X_4	X_5	Pe	Prior Beta (1,1)	Prior Beta (0,0)	Prior Beta (0.5, 0.5)
0	0	0	0	0	0.520	0.489 - 0.551	0.489 - 0.551	0.489 - 0.551
0	0	0	0	1	0.011	0.006 - 0.019	0.006 - 0.020	0.006 - 0.019
0	0	0	1	0	0.012	0.006 - 0.020	0.007 - 0.021	0.007 - 0.021
0	0	0	1	1	0.012	0.006 - 0.020	0.007 - 0.021	0.007 - 0.021
0	0	1	0	0	0.046	0.034 - 0.060	0.035 - 0.061	0.035 - 0.061
.	.	.	.	.	.	.	.	.
1	1	0	1	1	0.008	0.004 - 0.015	0.004 - 0.016	0.004 - 0.016
1	1	1	0	0	0.005	0.002 - 0.011	0.002 - 0.012	0.004 - 0.012
1	1	1	0	1	0.001	0.000 - 0.006	0.000 - 0.006	0.000 - 0.006
1	1	1	1	0	0.001	0.000 - 0.006	0.000 - 0.006	0.000 - 0.006
1	1	1	1	1	0.005	0.002 - 0.011	0.002 - 0.012	0.004 - 0.012

Choice of prior has small effects after n>20. The point estimate was set to 0.5.

	95% highest probability density regions		
Sample size (n)	Prior Beta (1,1)	Prior Beta (0,0)	Prior Beta (0.5, 0.5)
3	0.061 – 0.939	0.123 – 0.877	0.094 – 0.906
5	0.123 – 0.877	0.167 – 0.833	0.147 – 0.853
10	0.212 – 0.788	0.234 – 0.766	0.224 – 0.776
15	0.259 – 0.741	0.272 – 0.728	0.266 – 0.734
20	0.289 – 0.711	0.298 – 0.702	0.293 – 0.672
25	0.309 – 0.690	0.316 – 0.684	0.313 – 0.687
30	0.325 – 0.675	0.331 – 0.669	0.328 – 0.672
50	0.363 – 0.637	0.366 – 0.634	0.365 – 0.635
100	0.403 – 0.597	0.404 – 0.596	0.403 – 0.597
500	0.456 – 0.544	0.456 – 0.544	0.456 – 0.544
1000	0.469 – 0.531	0.469 – 0.531	0.469 – 0.531

Poster presented at the Objective Bayesian Methodology meeting on June 8-12, 2007 in Rome by Zheng Sponsiello-Wang