Identification of causal structures in epidemiologic data

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

BACKGROUND

Identification of causal relationships in observational data is one of the major challenges of epidemiology. The most common approach is to apply a list of criteria to derive causality from associations. Other than with statistical tests which evaluate associations among variables against chance, there is no generally accepted statistical procedure to directly test causal hypotheses. The methodology of probabilistic causal modeling, which has been developed in the last two decades, now allows for quantifying causality by calculating posterior model-based outcome probabilities. However, the question of how to validate the underlying causal models is not completely resolved. Various procedures to derive likely causal structures from associations among variables have been proposed, but such data-driven approaches are more related to exploratory data-mining methods than to a conceptually-driven scientific approach.

APPROACH

In order to support conceptually-driven causal modeling, a procedure to assess hypothetical causal structures on the basis of empirical categorical data was developed. After a causal model is specified a priori, it is compared against the data. Selecting the most likely model is supported by different model fit statistics. The term Configuration Sequence Analysis (CSA) is proposed to denote this approach.

METHODS

Configuration Sequence Analysis

A SAS macro to perform CSA was developed. The evaluation of the procedure was based on collections of causal models, which included the models on which the data simulations were based on. The first step of the CSA procedure is to calculate unconditional and conditional probabilities for each configuration of model variables of the specified causal network based on the 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

According to the causal structures denoted in the figures on the right hand-side, two data sets were generated using the SAS (version 9.1) RANUNI routine of a randomized process without memory, being unidirectional in time.

Computation of point estimates

In the context of Bayesian networks, the point estimate (Pe) of the outcome variable is defined by the hypothetical Bayesian network structure under investigation, according to the chain rule. In example 1,

$$Pe=P(X_1)\,I\,P(X_2)\,\square P(X_3|X_1)\,\square P(X_4|X_1,X_2)\,\square P(X_5|X_3,X_4)$$

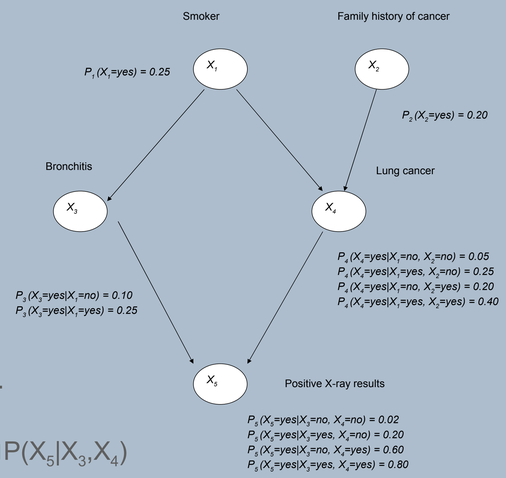
Model fit and model comparison

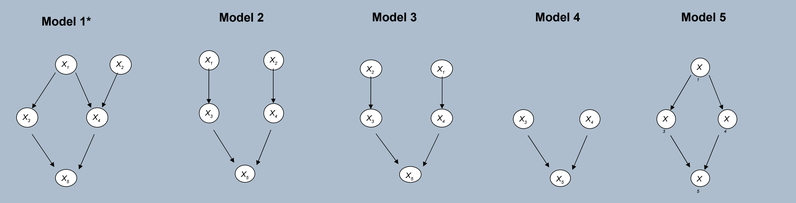
A I^2 goodness-of-fit statistic was used to evaluate the overall model fit. Deviance, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were also employed to select the most likely model.

CONCLUSIONS

In both instances, it was possible with the CSA process to identify the causal models on which the simulated data was based.

EXAMPLE 1

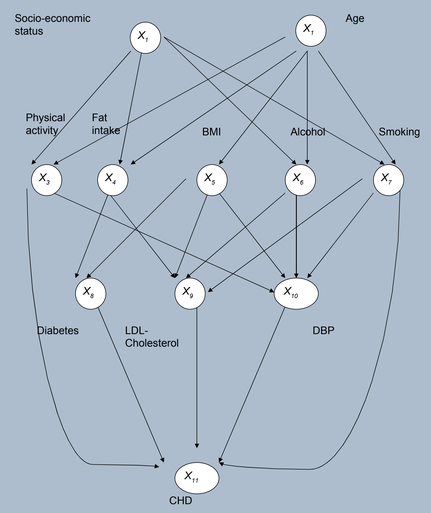


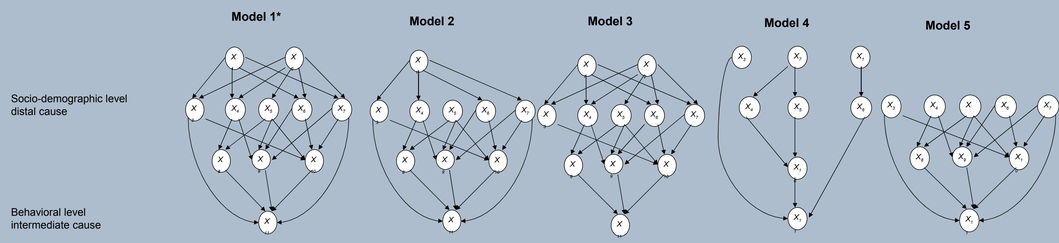


	Model 1*	Model 2	Model 3	Model 4	Model 5
I^2	31.015	88.717	142.008	2619.642	14.583
df	31	31	31	7	15
P	0.4654	1.79E-7	<1.79E-7	<1.79E-7	0.4818
Deviance	-2.572	-2.538	-1.720	5.773	-0.344
AIC	-1255.168	-1237.858	-829.175	2893.403	-156.928
BIC	-1179.122	-1161.098	-753.105	2910.580	-120.119

* Underlying data simulation

EXAMPLE 2





	Model 1*	Model 2	Model 3	Model 4	Model 5
$\square^2$	1824.424	6556.067	8478.250	8372.969	510.960
df	2047	1023	2047	255	512
P	0.9998	<1.79E-7	<1.79E-7	<1.79E-7	0.4921
Deviance	-10.804	-8.348	-10.478	-2.434	-6.621
AIC	-268052.786	-207681.782	-259897.304	-165278.119	-165022.119
BIC	-259025.743	-203170.465	-250870.261	-164153.597	-162768.665

* Underlying data simulation

Poster presented at the Society for Epidemiologic Research meeting on June 19-22, 2007 in Boston Massachusetts by Zheng Sponsiello-Wang and Rolf Weitkunat