Causal Discovery on Health-related Quality of Life of cancer patients

Maria Ganopoulou[1], Efstratios Kontopoulos[2], Konstantinos Fokianos[3], Lefteris Angelis[4], Ioannis Kotsianidis[1] and Theodoros Moysiadis[1,*]

[1]Department of Hematology, University Hospital of Alexandroupolis, Democritus University of Thrace Medical School, Alexandroupolis, Greece; [2]Foodpairing NV, Oktrooiplein 1, Box 401, 9000 Gent, Belgium [3]Department of Mathematics & Statistics, University of Cyprus, Nicosia, Cyprus, [4]School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

Abstract: The management of cancer patients increasingly includes Health-related Quality of Life (HRQoL) as a crucial endpoint. Physical, psychological, lifestyle, and social aspects expressed via responses to HRQoL questionnaires offer valuable insights for patient care. However, a still unexplored field is the identification and understanding of causal relationships among the questions involved. This study evaluates the capability of detecting cause-effect relationships in this context, applying causal structure learning algorithms to simulated data. Different data configurations are examined, encompassing the number of hypothetical questions in an HRQoL questionnaire, the quantity of cause-effect relationships, and the number of participants involved. Exploring this issue holds potential significance in shaping the design and/or selection of HRQoL questionnaires, accounting for limitations in sample size and intuition regarding the underlying causal structure. Uncovering cause-effect relationships can contribute to enhanced management and improved HRQoL for cancer patients.

Causal Discovery Aim: Investigate cause-effect relationships among questions within HRQoL questionnaires. HRQoL questions typically involve 3-5 answers representing increasing health burden. Data setups considered in the simulations involved increasing:

- Total number of hypothetical questions (4 answers each)
- Number/complexity of cause-effect relationships
- Number of simulated participants

Method: 6 directed acyclic graphs (DAGs) were specified with increasing complexity based on the number of hypothetical questions/nodes, and the number of directed edges (**Figure 1**).

Based on each DAG, 1000 samples were generated for each number of simulated participants (*n*= 50, 100, 200, 500, 1000, 2000 and 5000).

For each sample, two constraint-based structure learning algorithms were used to estimate the equivalence class of each DAG from the simulated data:

- (A) PC algorithm
- (B) Interleaved Incremental Association (Inter-IA)

The estimation was assessed based on:

- The structural Hamming distance (SHD) between the true and the estimated DAG
- The relative structural Hamming distance (rSHD), which was defined as the SHD divided by the number of the true DAG edges (to adjust for DAG complexity)

The mean value of these metrics was computed across the 1000 samples, for each DAG and each *n* (**Table 1**).

Figure 1: The **6** specified DAGs. A directed edge represents a cause-effect relationship.

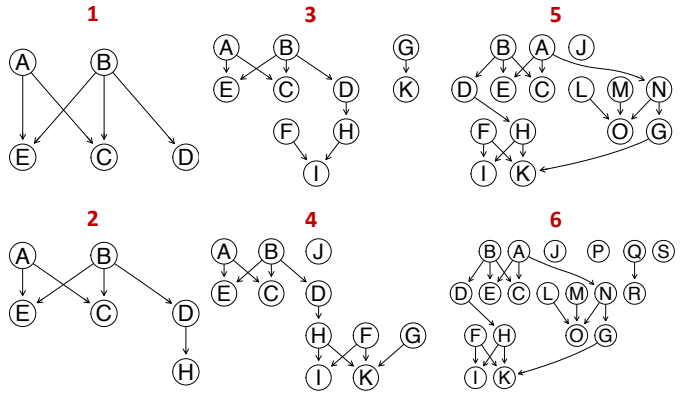
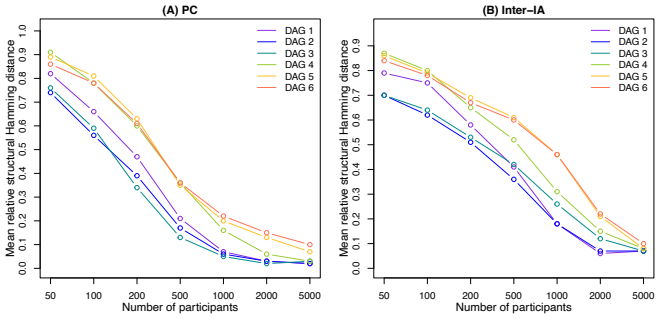


Table 1: The simulation results are displayed for the algorithms (A) PC, and (B) Inter-Iamb. The number of nodes/edges in each DAG is recorded. The 1st line for each *n* (50, 100, 200, 500, 1000, 2000, and 5000) corresponds to the mean SHD between the true and the estimated DAG, and the 2nd line to the corresponding mean rSHD.

(A) PC	DAG #	1	2	3	4	5	6
	# of nodes	5	6	10	11	15	19
	# of edges	5	6	9	11	16	17
number of simulated participants (*n*)	50	4.08	4.41	6.8	9.98	14.27	14.65
		0.82	0.74	0.76	0.91	0.89	0.86
	100	3.29	3.35	5.29	8.62	12.90	13.20
		0.66	0.56	0.59	0.78	0.81	0.78
	200	2.34	2.34	3.10	6.65	10.04	10.37
		0.47	0.39	0.34	0.60	0.63	0.61
	500	1.05	1.03	1.13	3.96	5.62	6.05
		0.21	0.17	0.13	0.36	0.35	0.36
	1000	0.34	0.35	0.42	1.74	3.14	3.74
		0.07	0.06	0.05	0.16	0.20	0.22
	2000	0.13	0.15	0.22	0.71	2.09	2.50
		0.03	0.03	0.02	0.06	0.13	0.15
	5000	0.11	0.13	0.23	0.37	1.12	1.65
		0.02	0.02	0.03	0.03	0.07	0.10

(B) Inter-IA	DAG #	1	2	3	4	5	6
	# of nodes	5	6	10	11	15	19
	# of edges	5	6	9	11	16	17
number of simulated participants (*n*)	50	3.96	4.21	6.30	9.54	13.81	14.22
		0.79	0.70	0.70	0.87	0.86	0.84
	100	3.74	3.74	5.74	8.76	12.66	13.24
		0.75	0.62	0.64	0.80	0.79	0.78
	200	2.92	3.05	4.78	7.18	11.11	11.38
		0.58	0.51	0.53	0.65	0.69	0.67
	500	2.05	2.19	3.79	5.69	9.74	10.27
		0.41	0.36	0.42	0.52	0.61	0.60
	1000	0.91	1.07	2.38	3.37	7.41	7.84
		0.18	0.18	0.26	0.31	0.46	0.46
	2000	0.31	0.43	1.10	1.62	3.29	3.82
		0.06	0.07	0.12	0.15	0.21	0.22
	5000	0.33	0.43	0.59	0.89	1.29	1.65
		0.07	0.07	0.07	0.08	0.08	0.10

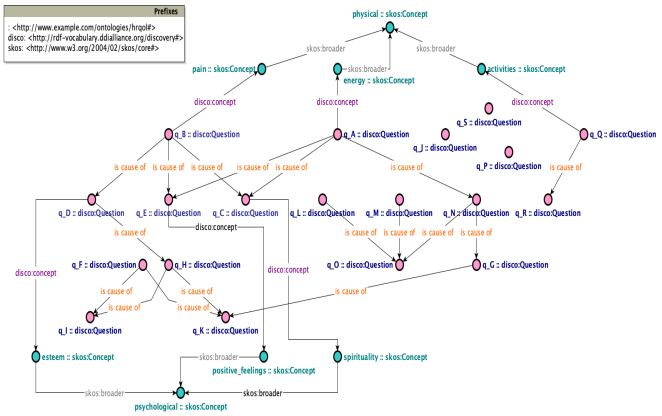
Figure 2: The mean rSHD is displayed for the (A) PC and the (B) Inter-IA algorithms across the 6 DAGs and the number of simulated participants.



Semantics Aim: Illustrate a sample instantiation of a DAG representing a set of cause-effect relationships among questions in a hypothetical HRQoL questionnaire. The underlying knowledge representation formalism is based on W3C-endorsed semantic standards:

- The Resource Description Framework (RDF) as the representation schema
- The Simple Knowledge Organization System (SKOS) for representing facets and domains of questions and their narrower-broader interrelationships
- DDI-RDF for representing research and survey data, such as questionnaires

Figure 3: The diagram representing a sample instantiation of a DAG corresponding to a hypothetical HRQoL questionnaire .



Conclusion: The results have shown that:

- Both algorithms were inefficient for *n*<200
- Larger values of *n* (≥ 100) resulted in a satisfactory performance
- The algorithms performed better for the simpler DAGs 1-3
- PC slightly outperformed Inter-IA

The added value of deploying semantic technologies lies in:

- Wider shareability of findings with interested stakeholders
- Improved interoperability with third-party applications and AI agents
- Extended collaboration with other research parties

Funding: The research project was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "2nd Call for H.F.R.I. Research Projects to support Post-Doctoral Researchers" (Project Number: 553).

personaCLLized

GENERAL SECRETARIAT FOR RESEARCH AND INNOVATION

H.F.R.I. Hellenic Foundation for Research & Innovation

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ DEMOCRITUS UNIVERSITY OF THRACE