Causal Structure assessment in Health-Related Quality of Life questionnaires

Maria Ganopoulou[1], Dimitris Koparanis[1], Konstantinos Liapis[1], Eleftheria Lamprianidou[1], Stavros Papadakis[1], Konstantinos Fokianos[2], Ioannis Kotsianidis[1], Lefteris Angelis[3], Theodoros Moysiadis[1,*]

[1]Department of Hematology, University Hospital of Alexandroupolis, Democritus University of Thrace Medical School, Alexandroupolis, Greece; [2]Department of Mathematics & Statistics, University of Cyprus, Nicosia, Cyprus [3]School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

Abstract: Health-related quality of life (HRQoL) is emerging as an important endpoint in managing cancer patients. Physical, psychological, lifestyle and social parameters, such as perceived family and social support, may provide guidance on how to approach and manage the individual patient. Several HRQoL questionnaires are available, and/or can be developed that assess the quality of life of the patients. An aspect that has not yet been investigated, however, is the existence and detection of causal relationships among the questions within such questionnaires. This study aims to assess the ability to detect cause-effect relationships within this context, by employing different causal structure-learning algorithms, based on simulated data. To this end, different data setups are considered, involving the total number of hypothetical questions within a HRQoL questionnaire, the number and complexity of cause-effect relationships, and the number of participants. Addressing this issue may be of potential merit when considering the design and/or selection of a HRQoL questionnaire, taking into account sample size limitations, and scientific intuition regarding the underlying causal structure. Since the aforementioned questions may concern, among else, physical, psychological, lifestyle and social aspects, related to the individual patient, unveiling cause-effect relationships among these questions may aid to improve the management, and the health-related quality of life of cancer patients.

Aim: The focus in this study was to investigate the ability to detect cause-effect relationships among questions within health-related quality of life (HRQoL) questionnaires. HRQoL questions typically involve 4 or 5 answers representing increasing health burden related to the question at hand. The data setups considered in the simulations involved different/increasing:

- total number of hypothetical questions (4 answers each)
- number and complexity of cause-effect relationships
- number of simulated participants

Method: 8 networks were specified with increasing complexity based on the number of hypothetical questions (each question was represented by a node), and the number of directed edges (see **Figure 1**). The specific networks' parameters were custom fitted (not shown here).

For each network, based on its' specific parameters, 1000 samples were generated for each number of simulated participants (n= 20, 30, 50, 100, 200, 300, 500, 1000, and 2000). Thus, each sample included n sets of simulated answers corresponding to the n participants. Then, for each sample, two different constraint-based structure learning algorithms (A) the standard PC, and (B) the Max-Min Parents and Children (MMPC) algorithms, were independently used to estimate the equivalence class of a directed acyclic graph (DAG) from the simulated data. To assess the ability of the algorithms to detect cause-effect relationships in each case (network, n), two metrics were used,

- the Hamming distance between the true and the estimated network,
- a relative Hamming distance, which was defined as the hamming distance divided by the number of the true network edges. This metric was used in order to adjust the complexity (number of edges) of each network.

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

Figure 1: The 8 specified networks.

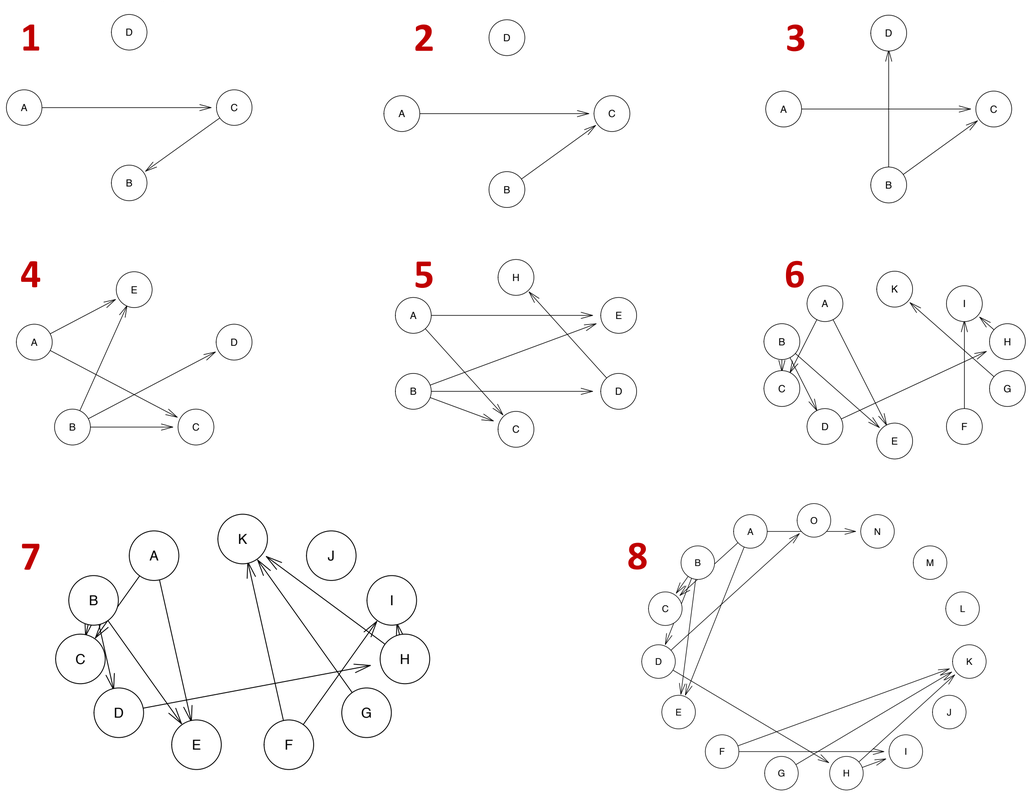

Table 1: The simulation results are displayed for the algorithms (A) PC, and (B) MMPC. The number of edges in each network is recorded. The 1st line for each n (20, 30, 50, 100, 200, 300, 500, 1000, and 2000) corresponds to the mean Hamming distance between the true and the estimated network, and the 2nd line to the mean relative Hamming distance.

(A) PC

number of simulated participants (n)	# of edges	2	2	3	5	6	9	11	13
	Net #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
20		0.86	0.95	1.61	3.97	4.79	6.6	9.68	11.29
		0.43	0.47	0.54	0.79	0.8	0.73	0.88	0.87
30		0.42	0.50	1.13	3.57	4.08	5.53	8.59	10.35
		0.21	0.25	0.38	0.71	0.68	0.61	0.78	0.80
50		0.18	0.19	0.75	3.03	3.25	4.26	6.93	9.05
		0.09	0.09	0.25	0.61	0.54	0.47	0.63	0.7
100		0.07	0.06	0.39	2.12	2.15	2.68	4.84	6.86
		0.03	0.03	0.13	0.42	0.36	0.3	0.44	0.53
200		0.07	0.08	0.13	1.25	1.30	1.51	2.95	4.53
		0.04	0.04	0.04	0.25	0.22	0.17	0.27	0.35
300		0.08	0.09	0.06	0.93	0.95	1.08	2.11	3.25
		0.04	0.05	0.02	0.19	0.16	0.12	0.19	0.25
500		0.06	0.08	0.04	0.57	0.54	0.59	1.25	1.88
		0.03	0.04	0.01	0.11	0.09	0.07	0.11	0.14
1000		0.05	0.09	0.03	0.15	0.16	0.21	0.46	0.74
		0.03	0.05	0.01	0.03	0.03	0.02	0.04	0.06
2000		0.05	0.09	0.02	0.07	0.08	0.12	0.19	0.44
		0.03	0.04	0.01	0.01	0.01	0.01	0.02	0.03

(B) MMPC

number of simulated participants (n)	# of edges	2	2	3	5	6	9	11	13
	Net #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
20		0.72	0.93	1.43	3.64	4.41	6.15	9.15	10.88
		0.36	0.47	0.48	0.73	0.74	0.68	0.83	0.84
30		0.45	0.51	1.08	3.26	3.86	5.28	8.11	9.85
		0.22	0.26	0.36	0.65	0.64	0.59	0.74	0.76
50		0.18	0.16	0.73	2.76	3.04	4.08	6.60	8.46
		0.09	0.08	0.24	0.55	0.51	0.45	0.60	0.65
100		0.09	0.09	0.41	1.78	1.84	2.25	4.24	6.08
		0.05	0.04	0.14	0.36	0.31	0.25	0.39	0.47
200		0.08	0.10	0.14	1.03	1.05	1.26	2.58	4.12
		0.04	0.05	0.05	0.21	0.17	0.14	0.23	0.32
300		0.07	0.09	0.05	0.86	0.85	0.93	1.85	3.11
		0.04	0.05	0.02	0.17	0.14	0.10	0.17	0.24
500		0.07	0.09	0.03	0.55	0.54	0.57	1.16	1.95
		0.04	0.04	0.01	0.11	0.09	0.06	0.11	0.15
1000		0.07	0.10	0.03	0.17	0.18	0.19	0.44	0.86
		0.03	0.05	0.01	0.03	0.03	0.02	0.04	0.07
2000		0.06	0.09	0.03	0.10	0.09	0.12	0.22	0.6
		0.03	0.04	0.01	0.02	0.01	0.01	0.02	0.05

Figure 2: The mean relative Hamming distance is displayed for the PC (A1, B1), and MMPC (A2, B2) algorithms across the 8 networks and the number of simulated participants .

Conclusion: It is shown that both algorithms are not capable to efficiently detect the cause-effect relationships among HRQoL questions when the number of simulated participants is small, particularly for n<200. On the other hand, large values of n (≥500), ensure that both metrics indicate a satisfactory performance. In addition, as expected, the algorithms performed better for the more simple networks 1-3. The performance was similar for networks 4-8, indicating that increasing complexity did not pronouncedly inflict the performance from network 4 and on. The performance of the PC and MMPC algorithm was very similar in terms of both the Hamming distance, and the mean relative Hamming distance.

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).