Bayesian Networks and Causal Inference for the Interpretation of Patients' Symptom Experience

Inference in Bayesian Belief Networks

The structure of the edges in BNs is based on two analytical assumptions. Based on the Causal Markov Assumption, given the values of a variable's immediate causes, this variable is independent of its earlier causes. The second assumption is that no latent or hidden variables exist that affect the observable variables.

Additionally, BNs are DAGs comprised of qualitative and quantitative parts.[1,2] The qualitative part is the network structure that we have explored in the main of this paper. The qualitative part is captured by the Conditional Probabilities Table (CPT) of the nodes (i.e., symptoms) included in the aforementioned BN. More specifically, the CPT displays the conditional probabilities of each symptom with respect to the others (i.e., the probability of each possible value of one symptom if we know the values taken on by the other symptoms). Based on demographic and clinical characteristics, as well as biological processes, a patient's experience, can be modeled as a network of causal influences.[3]

Consider a two-node network where A and B are binary variables with two states (N or Y). The marginal table at node A would contain the marginal probability P(A = Y). For simplicity, we will use "A" to mean A = Y and "a" to mean A = N so that P(A = Y) and P(A = N) can be written more briefly as P(A) and P(a), respectively. By complementarity, P(a) = 1 – P(A). At node B we would have the conditional probabilities P(B|A) and P(B|a) that define how the state of B depends on the state of A. The CPT can be completed using complementarity: P(b|A) = 1 – P(B|A) and P(b|a) = 1 – P(B|a). Thus, the marginal table for A lists all possible states for A, and the CBT lists all possible state combinations of A and B. Once the network is constructed and the probabilities are specified, Bayes' theorem is used to propagate probability through the model.

Figure S1 illustrates a simple BN to demonstrate its application in probabilistic and causal inference. As mentioned above, we assume that no latent or hidden variables exist in this particular BN. An edge occurs from N to W because being "nervous" has a direct influence on the occurrence of the symptom "worry". In addition, being "nervous" has a direct influence on the occurrence of the symptom of "difficulty concentrating". "Difficult sleeping" may be due to feeling "worry" or having "difficulty concentrating" or to both of the symptoms. o direct edge occurs from W to DC. In probabilistic terms, knowing W and DC renders N and DS independent.

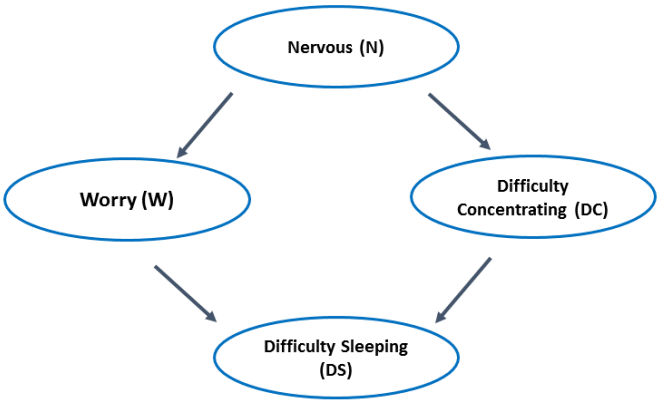

Figure S1. A simple BN representing dependencies among four cancer symptoms

For example, let's say that we know a patient has "difficulty sleeping" and we want to know if this symptom was due to feeling "worry" or "difficulty concentrating". We can use Bayes' Theorem to compute the posterior probability of each case. The

probabilities of W and DC conditional on DS will be computed from the following equation:

$$P(W|DS) = \frac{P(DS|W) \cdot P(W)}{P(DS)} \tag{1}$$

$$P(DC|DS) = \frac{P(DC|DS) \cdot P(DC)}{P(DS)} \tag{2}$$

By comparing these two probabilities, we can infer which symptom this patient is more likely to report. Increased knowledge of the Markov structure and conditional probabilities of all the symptoms a cancer patient can experience, will provide additional information to answer this question.

Causal Reasoning in Bayesian Networks of Cancer Symptoms

For this example, we will focus on the network structure learned with the Max-Min Hill Climbing algorithm and optimal threshold according to Scutari et al.[4] (see Appendix A, Figure S2). Fitting our data on this network structure with the Expectation Maximization (EM) algorithm we learn the local probability distributions of all 38 nodes (i.e., symptoms) in our study (Figure S2).

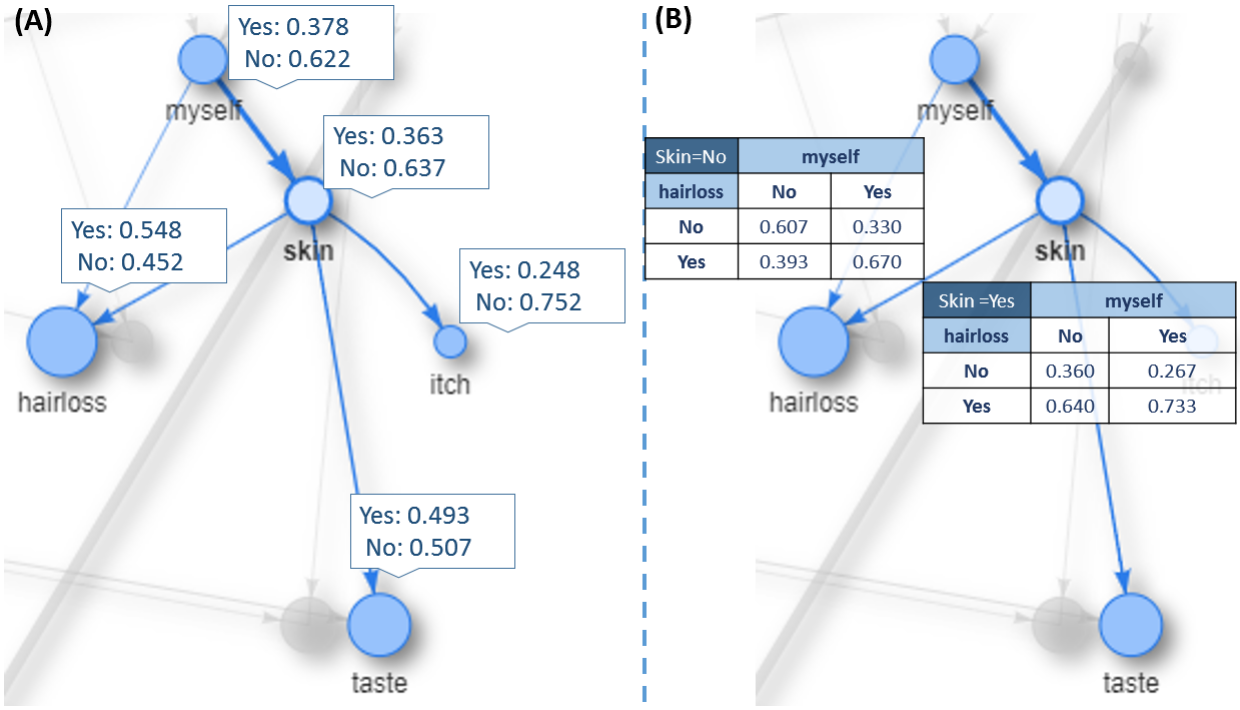

Figure S3. Causational model of cancer symptom experience. The cancer symptoms represented in the graph are coded in the following fashion: myself: I Do Not Look Like Myself, skin: Changes in Skin, itch: Itching, taste: Change in the Way Food Tastes, hairloss: Hair Loss

Focusing on the symptom "Changes in Skin" (node: "skin") we get the causational model in Figure S3. Figure S3A depicts the Markov Blanket of the symptom "skin" with the local probability distributions of each symptom (i.e., "myself", "skin", "hairloss", "itch", "taste") inside it. Figure S3B depicts the conditional probabilities of the occurence of symptom "Hair Loss" (node: hairloss) based on the occurence (i.e., "Yes") of its parents symptoms "I Do Not Look Like Myself" (node: myself) and "Changes in Skin" (node: skin). The higher the probability, the more likely it will be that a specific set of combined observations will happen. For example, having "Changes in Skin", "I Do Not Look Like Myself" and "Hair Loss" can occur

together only with a probability of 0.733. Whereas, ”I Do Not Look Like Myself” can occur without ”Changes in Skin” or ”Hairloss” with a probability of only 0.330.

Querying the aforementioned probability distributions, helps us to understand how symptoms interact with each other. More precisely, we can infer which symptom(s) a patient is more likely to report or which symptom(s) are more probable to cause a subsequent one. On the occasion of a specific set of observations, the closer the probability gets to 1, the higher the chances are that these combined observations will occur. Based on this structure and the parameters learnt for each node, we can predict the probability of occurrence (i.e., Yes) of ”Change in the Way Food Tastes” and ”Itching” for a given set of observations for ”Changes in Skin” (e.g., Yes), ”Hair Loss” (e.g., No) and ”I Do Not Look Like Myself” (e.g., Yes),

$$P(taste{=}Yes, itch{=}Yes | skin{=}Yes, hairloss{=}No, myself{=}Yes){=}0.253 \tag{3}$$

$$P(taste{=}Yes | skin{=}Yes, hairloss{=}No, myself{=}Yes){=}0.635 \tag{4}$$

$$P(itch{=}Yes | skin{=}Yes, hairloss{=}No, myself{=}Yes){=}0.390 \tag{5}$$

or we can enter a required state of ”Hair Loss” (e.g. No), ”Itching” (e.g., No), ”Change in the Way Food Tastes” (e.g., Yes), ”I Do Not Look Myself” (e.g. Yes) as observations to examine how the occurrence of ”Changes in Skin” can explain these specific observations.[5]

$$P(skin{=}Yes | hairloss{=}No, itch{=}No, taste{=}Yes, myself{=}Yes){=}0.647 \tag{6}$$

This procedure can be done iteratively for all the cancer symptoms included in our BN (Figure S2). This application may help clinicians intervene with patients who have specific symptom profiles, based on select demographic and clinical characteristics (e.g., the BNs learned based on gender and age). As the use of BNAs becomes more common and findings regarding causal relationships between and among symptoms are confirmed, clinicians can use this information to identify high risk patients and initiate pre-emptive and/or targeted interventions.[6–9]

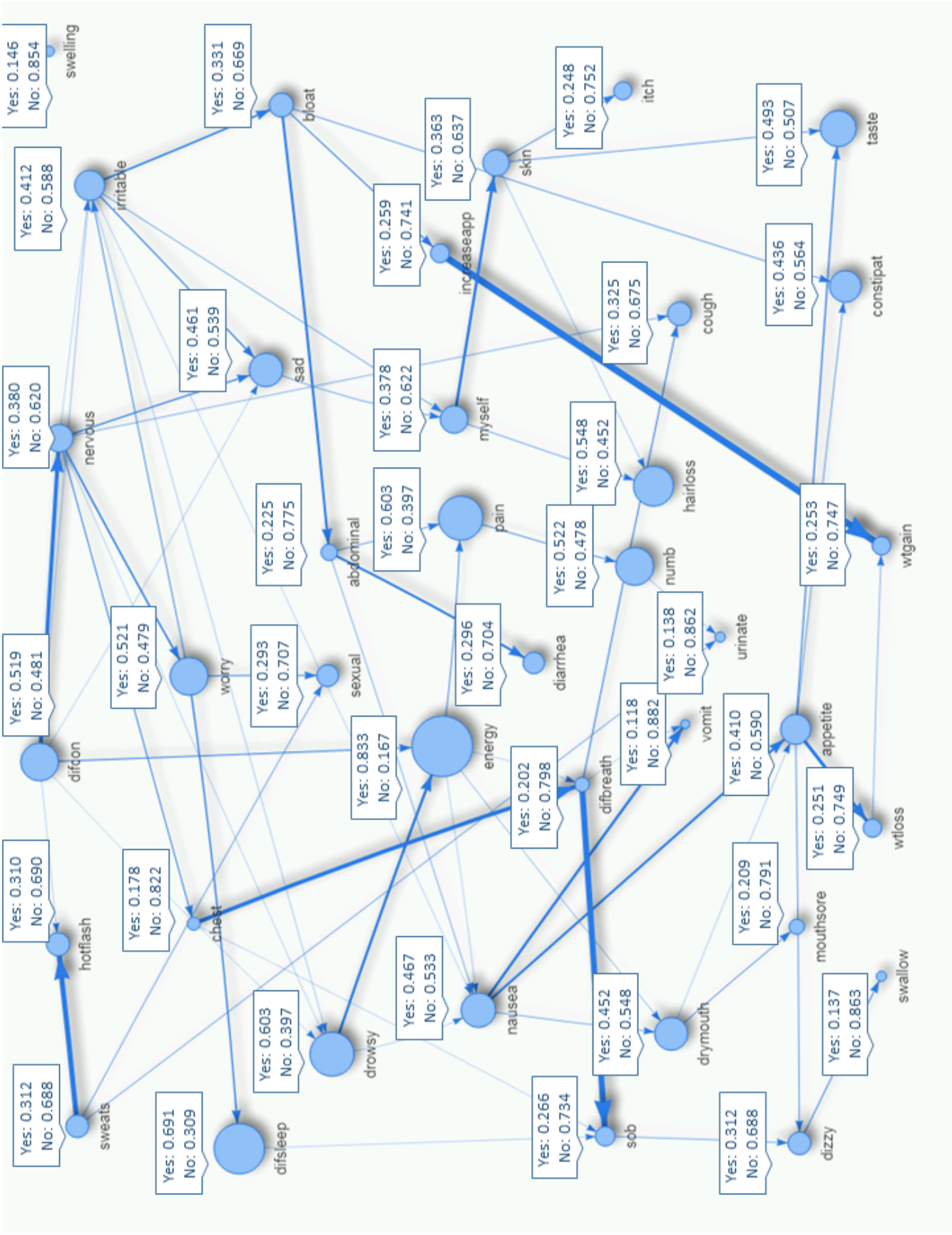

Figure S2. The marginal probabilities for the complete set of 38 symptoms

References

1. Pearl, J. [bayesian analysis in expert systems]: comment: graphical models, causality and intervention. *Stat. Sci.* **8**, 266–269 (1993).

2. Kyrimi, E. *et al.* A comprehensive scoping review of bayesian networks in healthcare: Past, present and future. *arXiv preprint arXiv:2002.08627* (2020).

3. Puga, J. L., Krzywinski, M. & Altman, N. Points of significance: Bayesian networks (2015).

4. Scutari, M. & Nagarajan, R. Identifying significant edges in graphical models of molecular networks. *Artif. intelligence medicine* **57**, 207–217 (2013).

5. Constantinou, A. C. & Fenton, N. Things to know about bayesian networks: decisions under uncertainty, part 2. *Significance* **15**, 19–23 (2018).

6. Jaffee, E. M. *et al.* Future cancer research priorities in the usa: a lancet oncology commission. *The Lancet Oncol.* **18**, e653–e706 (2017).

7. Penedo, F. J. *et al.* The increasing value of ehealth in the delivery of patient-centred cancer care. *The Lancet Oncol.* **21**, e240–e251 (2020).

8. Lindner, O. C., Velikova, G. & Stark, D. P. Digitally enabled patient-reported outcome measures in cancer care. *The Lancet Oncol.* **20**, e2 (2019).

9. Heathcote, L. C. *et al.* Digitally enabled patient-reported outcome measures in cancer care–authors' reply. *The Lancet Oncol.* **20**, e3 (2019).