Explaining Bayesian networks: a use case in endometrial cancer

Casper Verhoeve[1](✉), Marcos L. P. Bueno[2], and Casper Reijnen[3]

[1] iCIS, Radboud University, the Netherlands
[2] School of AI, Radboud University, the Netherlands
[3] Department of Radiotherapy, RadboudUMC, the Netherlands
casper.verhoeve@ru.nl

Abstract. Despite the tremendous accuracy Artificial Intelligence (AI) achieves in the medical domain, there is a clear need for explainable AI (XAI) for increased adoption. In this work we implement and evaluate several XAI methods on a Bayesian network for endometrial cancer.

Keywords: Bayesian networks · explainable AI · healthcare · causality

1 Introduction

Unprecedented accuracy has been achieved by AI models in healthcare, notably in the diagnosis of cancers using deep learning models [1, 9, 3]. However, to achieve increased adoption and trust in AI by clinicians, as well as adherence to legislation, additional requirements arise, such as explainability. This makes white-box AI models such as Bayesian networks (BNs) very relevant. Even though Bayesian networks are interpretable, clinicians often still report having trouble understanding the BN's predictions [5]. To improve upon this, methods for deriving explanations from a BN have been introduced [6, 2, 10, 4].

In this work we aim to implement and evaluate explanation methods for Bayesian networks in healthcare. Most implementations of BN explanations use either fictional situations or target other domains (e.g. legal [10]), and very few were evaluated (e.g. analytically or with human participants). To demonstrate this, we use the recently developed Endorisk Bayesian network [7] for prognostication of endometrial cancer patients. Patient data in this study included clinical and biomarker variables, which makes BN a suitable model for this scenario.

2 Preliminary Results and Next Steps

Table explanations. The "Table" method [6] generates a table with textual explanations with the most important factors contributing to the output. It generates explanations in three different levels, giving the reader the choice of how much in-depth they would like to have an explanation, which can be useful in cases where factors such as time apply to a decision-making process.

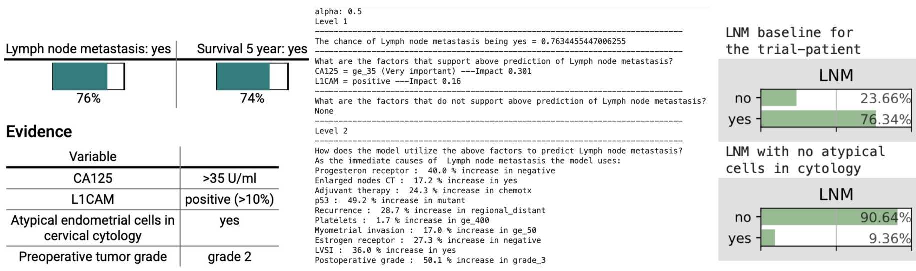

Fig. 1. Patient evidence (left), Table explanation (middle) and Counterfactual explanation (right-top: before intervention; right-bottom: after intervention on citology).

Counterfactual explanations. The second method is a counterfactual explanation of the model output. In a counterfactual explanation, a "what if" scenario is created to advance the understanding of the model reasoning. For example, a counterfactual can be calculated for one or more of the evidence variables to see what the output would be if those variables would have been different, showing their importance to the predicted outcome. For this calculation, we assume that the model is a causal Bayesian network as indicated in the original study [7].

Results. The Endorisk Bayesian network [7] has 18 nodes, including clinical, histopathological and biomarker data. The main targets are lymph node metastasis (LNM) and 5-year survival (DSS5). For this experiment, two patients were simulated with different prognoses, based on expert consultation. In this abstract however, we will only show one trial patient. As a baseline, according to the Endorisk model, a patient without any evidence noted has a chance of LNM of around 9%, and a chance of DSS5 of 93%.

The *trial patient* has evidence shown in Figure 1. These markers give the patient a chance of LNM of 76% according to the model, so we consider this a high-risk patient. In Figure 1, Table output and Counterfactual output for atypical cells in cytology (Cytology) are also shown. For Table, level 1 and 2 are shown, which outline the significant evidence for the LNM prediction, their impact and the changes in the nodes that have a direct influence on LNM.

As for counterfactuals, small changes are considered to explore alternative scenarios. In this case, Cytology is considered, as the result of the counterfactual shows that if there would have been no atypical cells in cytology, LNM would have a 90% chance of being not present. This is a big change from the original outcome, which had a 70% chance of LNM being present.

Next steps. This research will provide a thorough evaluation of explanations for a medical Bayesian network, which can hopefully also serve as inspiration for explaining other healthcare cases. Our next steps include: add more fictional patients (yet realistic); implement other BN explanation methods (e.g. [10]) as well as model-agnostic explainable AI methods [8]; finally, evaluate the explanations with experts/non-experts using (semi-structured) interviews and questionnaires.

References

1. Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al.: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature medicine **25**(6), 954–961 (2019)
2. Balke, A., Pearl, J.: Probabilistic evaluation of counterfactual queries. In: Probabilistic and Causal Inference: The Works of Judea Pearl, pp. 237–254 (2022)
3. Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., Litjens, G.: Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. The Lancet Oncology **21**(2), 233–241 (2020)
4. Butz, R., Hommersom, A., van Eekelen, M.: Explaining the most probable explanation. In: Scalable Uncertainty Management: 12th International Conference, SUM 2018, Milan, Italy, October 3-5, 2018, Proceedings 12. pp. 50–63. Springer (2018)
5. Butz, R., Schulz, R., Hommersom, A., van Eekelen, M.: Investigating the understandability of xai methods for enhanced user experience: When bayesian network users became detectives. Artificial Intelligence in Medicine **134**, 102438 (2022)
6. Kyrimi, E., Mossadegh, S., Tai, N., Marsh, W.: An incremental explanation of inference in bayesian networks for increasing model trustworthiness and supporting clinical decision making. Artificial Intelligence in Medicine **103** (3 2020). https://doi.org/10.1016/j.artmed.2020.101812
7. Reijnen, C., Gogou, E., Visser, N.C., Engerud, H., Ramjith, J., Van Der Putten, L.J., Van de Vijver, K., Santacana, M., Bronsert, P., Bulten, J., et al.: Preoperative risk stratification in endometrial cancer (endorisk) by a bayesian network model: A development and validation study. PLoS medicine **17**(5), e1003111 (2020)
8. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
9. Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Broeders, M., Gennaro, G., Clauser, P., Helbich, T.H., Chevalier, M., Tan, T., Mertelmeier, T., et al.: Standalone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. JNCI: Journal of the National Cancer Institute **111**(9), 916–922 (2019)
10. Vlek, C.S., Prakken, H., Renooij, S., Verheij, B.: A method for explaining bayesian networks for legal evidence with scenarios. Artificial Intelligence and Law **24**, 285–324 (2016)