Bayesian Network Reasoner for Breast Cancer risk factors in the UK

No Author Given

Vrije Universiteit, Amsterdam

1 Performance Evaluation

In this research project, we measured the performance of our BNR implementation by looking at the average time taken to complete a random MAP *and* MPE query for a randomly generated Bayesian Network, comparing three different elimination order heuristics (Random, *Min. Degree* & *Min. Fill*). By varying the size (amount of variables present) of these networks, we get to observe how our BNR handles increasingly larger networks and compare the effects of the different heuristics on performance.

Randomly generated Bayesian Networks & Queries For even more robust results, we decided to generate networks and related queries on the fly, allowing us to get more generalized results. Even so, for a given the amount of variables wanted, we wrote a function that creates random edges and CPTs, while still complying to the rules for Bayesian Networks. For this network, a number of queries were generated, though not as random: a map and evidence was made up out of two random, non-intersecting variables for each, which were used in the MPE and MAP algorithms.

1.1 Experimental Setup

While our implementation allowed for lengthy experiments, time constraints limited the amount of runs we managed to do. We ended up experimenting on networks of three different sizes: 5, 10 or 15 variables. For each of these size, we generated five different networks with the same size and making a single MPE and MAP query for each.

1.2 Results

Figure 1 showcases the results, where even with our limited dataset, clear differences can be noticed.

Focusing on the ordering heuristics at first, their effects are negligible on the smaller networks, but as they grow, so does the influence of the heuristics on performance. Both the *Min. Degree* and *Min. Fill* heuristics distance themselves from the random ordering and significantly decrease the time taken for larger

networks. While the distribution of network-widths ended up the same for these two heuristics, *Min. Fill* still managed to edge ahead over *Min. Degree.*
As for the effects of network size on performance, it does seem that the time it takes to complete queries grows exponentially (notice the log-scale) with network size. Fortunately, quality heuristics significantly reduce the time needed.

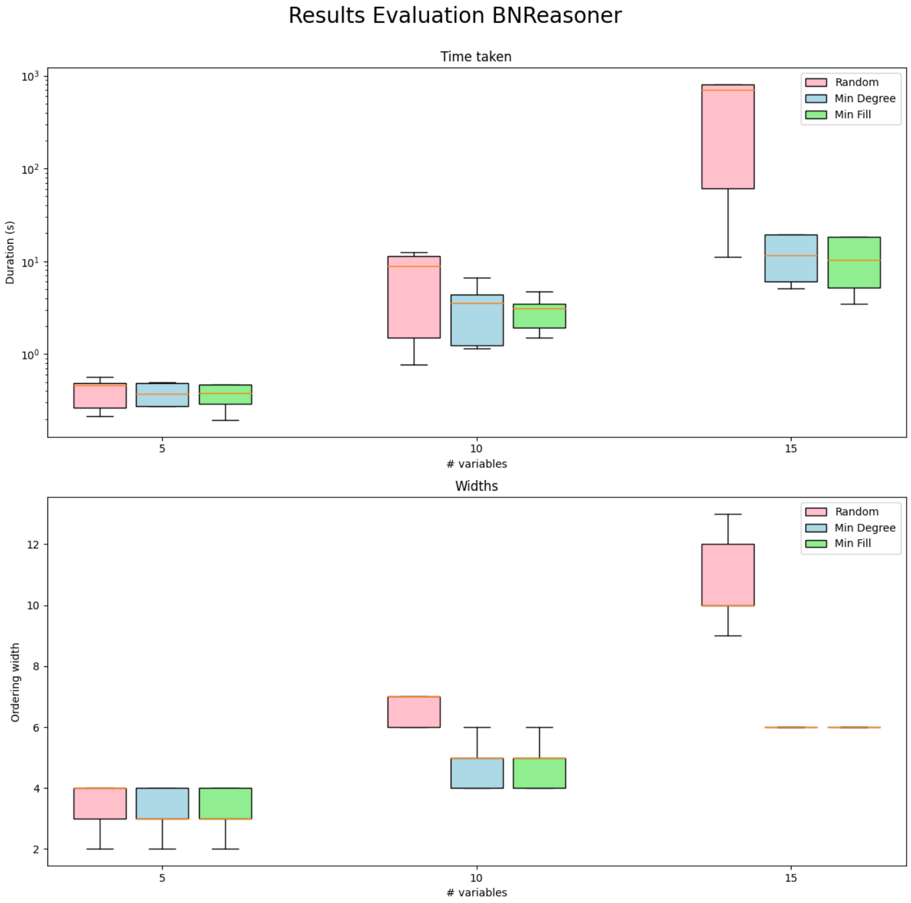

Fig. 1: The results of our evaluation, per network size and per heuristic. Upper plot shows the time taken in log-scale. Lower plot shows the distribution of widths.

1.3 Performance discussion

Ordering heuristics are clearly beneficial to implement, but there are most definitely other ordering heuristics or algorithmic methods out there that can cut down the time needed, which ought to be further explored. Moreover, there are still more improvements to be done, specifically implementation-wise. We did not get to spend a lot of time optimizing performance, which made gathering more and better results rather difficult unfortunately.

2 Modelling a Use-Case

2.1 Introduction

Breast cancer is the most prevalent type of cancer among women; around 11.500 women and 85 men die from it yearly in the UK [10].

The next subsections discuss Bayesian Network Model for Breast Cancer risk Factors in the UK. The choice of country was mainly based on the availability of data. Furthermore, the definition of the model, as well as investigating different queries using Bayesian Network Reasoning techniques based on the resulting model, will be discussed.

2.2 Defining the Model

The choice of variables was based on the work of Momenimovahed and Salehiniya (2019) [1], who identified risk factors for breast cancer worldwide. They divided risk factors into three categories: protective, predisposing and controversial. For the sake of simplicity, only protective and predisposing factors were chosen, resulting in 9 risk factor variables in total: Sex, Family History, Smoking, Having Kids, oral contraceptives (from now on referred to as Hormone Therapy or HT), Alcohol Consumption and Lack of Physical activity.

The connectivity between the nodes was based on accessibility of statistics for a specific category i.e $P(Smoker? \cap Alcohol?)$, common sense i.e Hormone Therapy? is a child of Female?, since being female has a major influence on whether or not one takes oral contraceptives and attempting to limit the amount of nodes directly leading to cancer (for the sake of simplicity).

Data for the Table 5, sub-tables a-d, defining the probability distributions for root nodes, were retrieved from [3-6].

Data for Table 6(a) was retrieved from [2]. The probability of taking hormone therapy and being male was set to 1.0, since only oral contraception pill was taken into consideration.

Data for Table 6(b) was partially retrieved from [7], which provides the prevalence rates of misuse of alcohol in UK. Specifically, the estimated number of people misusing alcohol was taken into consideration as long as finding that in treatment for alcohol dependency were male. It, therefore, resulted in 0.006 and 0.02. The probability distribution in the table was counted based on conditional probability formula $P(A|B) = P(A \cap B)/P(B)$.

Data for tables 7-8 were made up by the authors based on their subjective beliefs about the probability distributions of each variable.

2.3 Implementing Bayesian Network Reasoner

As mentioned above, among the most prominent and recognised risk factors for breast cancer are family history, alcohol abuse and whether or not one have had kids (for women). All the online assessment tools [8-9] base their decision mostly on these factors. Nevertheless, smoking, even though mostly advised to be avoided in order to prevent cancer in general and breast cancer specifically, and even though listed in predisposing factors in the work of Momenimovahed and Salehiniya, is not used to calculate one's risk of having breast cancer in the most of accessible risk factor calculators. It was, therefore, decided to look into the relationship of breast cancer and smoking using different Bayesian Network Reasoning techniques.

Prior Marginal distribution of Cancer and Smoking First, we computed prior Marginal distribution of Cancer, Smoking and both to see if there is any visible relationship between them before any evidence is found. Table 1 represents the findings.

As can be seen from the prior marginal distribution tables, no visible interaction is happening between the two variables. Marginal probabilities of both cancer and smoking are very low, and the marginal probability of both is very lower (since the intersection of the two is always smaller). Therefore, posterior marginal probability of having cancer given being a smoker and vice versa should be investigated.

Table 1: Prior Marginal distribution of Cancer and Smoking

(a)

Cancer?	Pr(c)
True	0.088053
False	0.911947

(b)

Smoker?	Pr(s)
True	0.078245
False	0.921755

(c)

Smoker?	Cancer?	Pr(s,c)
True	True	0.007629
True	False	0.070616
False	True	0.080425
False	False	0.841331

Posterior Marginal distribution of Cancer and Smoking As mentioned above, since prior marginal distribution did not provide any insight into the relationship between smoking and cancer, the posterior distribution for Smoking given that Cancer was found and for Cancer given that person smokes, was calculated. Table 2 represents the findings.

As can be seen, there is a small increase in a chance of having cancer given smoking 0.097 (as compared to prior probability of having cancer 0.088) and increase in probability of being a smoker 0.078 compared to probability of being a smoker given that you have cancer 0.087. Though these differences are relatively

Table 2: Posterior Marginal distribution of Cancer and Smoking

(a)

Cancer?	Pr(c)
True	0.097497
False	0.902503

(b)

Smoker?	Pr(s)
True	0.086637
False	0.913363

small, they still pose and issue, since in medicine it is of vital importance to pay attention to the smallest of trends. These findings go in line with the base article, yet it is still unclear, why smoking is not taken into consideration in online assessment tools. On of possible explanation for this could be low reliability: most of these assessment tools only return high, low or medium risk of breast cancer, and for such computation adding smoking as an assessment criteria would add very little to the final decision but would still take up computational power.

MAP Maximum A Posteriori technique was also applied to the model, which allows to see most likely instantiation of a given variable given some evidence. The most likely instantiation of Hormone Therapy given that Cancer is true was of particular interest, since oral contraception is the most prevalent choice of contraception in the UK (excluding women taking it for other reasons, such as normalising cycle duration and hormone levels). The most likely instantiation of cancer given that Hormone Therapy is true was also investigated. The results can be found in Table 3.

Table 3: MAP Queries

Hormone Therapy?	Pr(c)
False	0.063927

(a) Hormone Therapy given, when Cancer is present

Cancer?	Pr(s)
False	0.150837

(b) Cancer present, when Hormone Therapy is given

As can be retrieved from the tables above, both queries did result in very low probabilities. The most likely instantiation of Hormone therapy given that cancer is true is false with probability 0.063927. The most likely instantiation of cancer given that Hormone therapy is true is also false with probability 0.150837. These findings suggest that the effect of oral contraceptive on breast cancer is indirect and is amplified by other variables.

MPE The last analysis performed on the given model was Most Probable explanations analysis, which allows to look at the most likely instantiation of all the variables given some evidence. Here, the most likely instantiation of all variables was investigated, given that cancer is true. Table 4 represents the findings.

Table 4: MPE Query, most likely instantiation when cancer is present

Pr(M)	NW?	Pollution?	LPA?	Alcohol?	FH?	Cancer?	Female?	HT?	Smoker?	Have Kids?
0.223603	False	False	False	False	False	False	True	False	False	True

As can be retrieved from the table 4, the most likely instantiation of the whole model given that cancer is true has probability of 0.22 and most of the risk factors in it are false. Interestingly enough, whether or not someone had kids is returned as true in this model, even though having kids is actually considered as a protective factor. It is possible though, that given that in our model most of predisposing factors for breast cancer influence having kids negatively, which therefore makes having kids more probable, if they are set to False.

References

1. Momenimovahed Z., Salehiniya H.. Epidemiological characteristics of and risk factors for breast cancer in the world. Breast Cancer (Dove Med Press). 11 (2019): 151–164. doi: 10.2147/BCTT.S176070
2. Statista (2021, March). Contraception in the United Kingdom (UK) - Statistics. https://www.statista.com
3. TUC (2018, october). Number of people working night shifts up by more than 150,000 in 5 years. Retrieved from: https://www.tuc.org.uk/news/number-people-working-night-shifts-more-150000-5-years
4. statista (2021, July). Population of the United Kingdom from 1953 to 2020, by gender. retrieved from: https://www.statista.com
5. Winton Centre for Risk and Evidence Communication. Does air pollution kill 40,000 people each year in the UK? Retrieved from: https://wintoncentre.maths.cam.ac.uk
6. Cancer Research UK. Family history of breast cancer and inherited genes. https://www.cancerresearchuk.org
7. Alcohol Change(2021). Alcohol in the UK. Retrieved from: https://alcoholchange.org.uk
8. The Breast Cancer Risk Assessment Tool. Retrieved from: https://bcrisktool.cancer.gov/
9. Breast cancer. Retrieved from: https://www.mycanceriq.ca/Cancers/Risk
10. Cancer Research UK. Breast cancer statistics. Retrieved from: https://www.cancerresearchuk.org/

3 Appendix

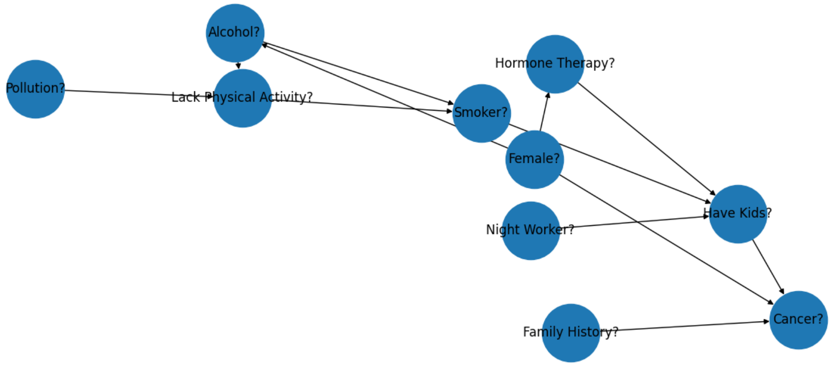

Fig. 2: Diagram for our Use-Case

CPTs describing our use-case.

Table 5

(a)

Night Worker?	Pr(nw)
True	0.11
False	0.89

(b)

Female?	Pr(f)
True	0.62
False	0.38

(c)

Pollution?	Pr(po)
True	0.05
False	0.95

(d)

Family History?	Pr(fh)
True	0.14
False	0.86

Table 6

(a)

Female?	Hormone Therapy?	Pr(ht—f)
True	True	0.28
True	False	0.72
False	True	0.0
False	False	1.0

(b)

Female?	Alcohol?	Pr(a—f)
True	True	0.006
True	False	0.994
False	True	0.02
False	False	0.98

Table 7

(a)

Lack of PA?	Alcohol?	Smoker?	Pr(s—lpa,a)
True	True	True	0.05
True	True	False	0.95
True	False	True	0.06
True	False	False	0.94
False	True	True	0.093
False	True	False	0.907
False	False	True	0.08
False	False	False	0.92

(b)

Pollution?	Alcohol?	Lack of PA?	Pr(lpa—p,a)
True	True	True	0.92
True	True	False	0.08
True	False	True	0.16
True	False	False	0.84
False	True	True	0.88
False	True	False	0.12
False	False	True	0.07
False	False	False	0.93

Table 8

(a)

S?	HT?	NW?	HK?	Pr(hk—s,ht,nw)
True	True	True	True	0.08
True	True	True	False	0.92
True	True	False	True	0.12
True	True	False	False	0.88
True	False	True	True	0.33
True	False	True	False	0.67
True	False	False	True	0.87
True	False	False	False	0.13
False	True	True	True	0.26
False	True	True	False	0.74
False	True	False	True	0.93
False	True	False	False	0.07
False	False	True	True	0.37
False	False	True	False	0.63
False	False	False	True	0.9
False	False	False	False	0.1

(b)

F?	FH?	HK?	C?	Pr(c—f,fh,hk)
True	True	True	True	0.22
True	True	True	False	0.78
True	True	False	True	0.42
True	True	False	False	0.58
True	False	True	True	0.1
True	False	True	False	0.9
True	False	False	True	0.15
True	False	False	False	0.85
False	True	True	True	0.09
False	True	True	False	0.91
False	True	False	True	0.09
False	True	False	False	0.91
False	False	True	True	0.01
False	False	True	False	0.99
False	False	False	True	0.01
False	False	False	False	0.99