Artificial Intelligence-Driven Breast Cancer Risk Assessment

Jacub Mateusz Szalacha
200535773
Evangelia Kyrimi
MSc Artificial Intelligence

Abstract*— Breast cancer remains the most commonly diagnosed cancer among women in the UK, posing substantial challenges on both diagnosed individuals and the healthcare systems. Despite many advancements in treatment that have greatly improved patients' treatment and survival rates, early-stage detection remains a critical factor in effective treatment and improved patient outcomes. Traditional screening methods such as mammography face accessibility limitations; this project aims to design and develop a Bayesian Network Model for breast cancer risk assessments. The model can be utilised by healthcare professionals during routine check-ups in order to recommend further testing and additional screening. Previous models such as the Gail model have provided the foundation for breast cancer risk assessment models however, they lack the integration of key factors such as genetic mutations. This project incorporates risk factors identified from seminal papers and official statistics to create a personalised risk stratification tool for breast cancer risk assessments. Preliminary testing of the model on three case studies has shown that the model is capable of distinguishing between low, medium, and high-risk individuals. The results display the model's potential ability to enhance early-stage detection rates and therefore contribute to better patient outcomes.

Keywords: (breast cancer, risk stratification, Bayesian networks, early-stage)

I. Introduction

Breast cancer is the abnormal growth of breast cells which then can form malignant tumours with the capability of metastasising to the rest of the body. Breast cancer is the most commonly diagnosed cancer among women in the UK, with one in eight women developing it during their lifetime (Cancer Research, 2011). Despite numerous medical technology advancements the successful treatment of breast cancer relies on early-stage diagnosis (Cancer Research, 2024b). Therefore, early-stage detection is an essential component of successful treatment.

Traditional screening methods, such as mammography, have several limitations and accessibility issues resulting in the need for alternative approaches that can assess individual risk factors and isolate those most at risk. A study by Tian et al. in 2012 concluded that access to mammography screening facilities does not necessarily correlate with an increased screening utilisation. This is due to factors such as financial restrictions. Despite having easier physical access, areas with racial disparities in breast cancer mortalities require greater resources. (Tian et al., 2012) The usage of a risk stratification model has the potential to reduce mortality rates by providing a more financially available screening technique that can recommend further testing based on the obtained results, thus potentially increasing the early-stage detection rates whilst reducing mortality.

The primary objective of this project is the development of a Bayesian Network model that integrates known risk factors to provide a comprehensive risk assessment for the development of breast cancer. The model aims to assist early detection of breast cancer with the intention to potentially improve treatment success, due to the early-stage detection and identification of relationships between various risk factors to offer a more precise and personalised tool.

This paper aims to answer the following research questions:

- What risk factors are the most significant in breast cancer development and how do they interact with one another?
- What are the practical implications of implementing a Bayesian Network model in a clinical setting?

II. Related Works

Breast cancer risk assessments have developed significantly over the past few decades and have begun incorporating various techniques and methodologies that improve prediction accuracy and clinical availability. This section will review key models and seminal works to highlight their contributions and limitations.

The Gail Model, introduced by Gail et al. in 1989, is one of the earliest documented and most influential breast cancer risk prediction tools (Stevanato et al., 2022). It is still used worldwide to predict the risk of developing breast cancer. Gail's model estimates the risk of developing breast cancer based on a variety of risk factors, such as family medical history. The significance of the model is the result of its widespread adoption and its role in public health initiatives such as the Breast Cancer Risk Assessment Tool (BCRAT) used by the National Cancer Institute (NCI). However, the model has several key limitations such as the lack of genetic factors and benign disease inclusion, which are key risk factors in breast cancer development. Important lifestyle and environmental risk factors like medical radiation exposure are also not considered.

The BOADICEA model developed by Antonia Antoniou et al. in 2002 integrates genetic, family history and epidemiological risk factors together to predict both the development of breast and ovarian cancer (Antoniou et al., 2004). The model can accurately predict women with BRCA1 or BRCA2 mutations. The model's ability to incorporate genetic factors provides a more personalised risk assessment with hereditary predispositions to breast cancer now capable of being included in the risk assessment. However, the model requires detailed genetic information to successfully perform risk assessments resulting in limitations to its practical application. This is evident in environments that do not have widespread access to genetic testing.

Tyrer-Cuzick's IBIS model, which is also known as the IBIS risk evaluator, was introduced in 2004, and combines personal, family history and genetic information risk factors to predict breast cancer risk (Kurian et al., 2021). The model's strength lies in the detailed inclusion of hormonal and reproductive factors and its adaptability to include genetic test results. However, similarly to the BOADICEA model its complexity results in the need for extensive patient data, making clinical implementation extremely difficult.

Artificial Neural Networks (ANNs) have been repeatedly explored for breast cancer prediction due to their complex analysis capabilities, which enable the discovery of patterns in large datasets. The study performed by Burke et al., highlights the potential of ANNs in medical environments, especially in cancer prediction, through their ability to learn from data (Burke et al., 1997). ANNs ability to handle numerous variables simultaneously provides potential for higher accuracy compared to other statistical models. However, ANNs require substantial amounts of data and computational resources limiting usage in clinical settings where these resources are not freely available.

Machine learning algorithms, such as decision trees and random forests, have been utilised to predict breast cancer development risk by analysing data patterns (Zhang et al., 2023). Machine learning algorithms offer high accuracy and adaptability as they can integrate many types of data, however similarly to the previous models they require large high-quality datasets for training and implementation.

Bayesian Networks are probabilistic graphical models, that provide a formalism for reasoning about partial beliefs under conditions of uncertainty, numerical parameters signifying the degree of belief accorded under a body of knowledge. These parameters are then combined and manipulated according to the rules of probability theory and later graphically displayed. (Pearl, 2014) Bayesian Networks have been widely applied in a variety of medical diagnostic scenarios such as breast cancer risk assessments where they have achieved results similar to radiologists. (Burnside et al., 2006) The ability to handle complex data and perform pattern recognition makes them particularly useful in understanding cause-effect dynamics. As the model is solely based on seminal papers and official statistics the results are explainable and interpretable.

Several comparative studies have evaluated these models, offering comprehensive assessments of their strengths and weaknesses across a wide range of environments and populations. (Amir et al., 2010) These studies have provided important insights into the accuracy and applicability of various models, allowing for appropriate tool selection for breast cancer risk assessments. However, as there are several testing method variations, factors such as data quality can influence the results, highlighting the need for a standardised testing criterion.

Several key themes emerged from the review of previously designed breast cancer risk assessment models. One of the themes is the increasing incorporation of genetic factors into models such as BOADICEA and Tyrer-Cuzick's IBIS, which allows for more accurate risk stratification. Personalisation of risk assessments is another prominent theme, as the inclusion of a variety of risk factors enables more tailored healthcare approaches. Additionally, computational methods such as Bayesian Networks and Artificial Neural Networks are gaining prominence due to their data handling and pattern recognition capabilities. However, the key limitation that is common across all these models is the requirement for high-quality, extensive datasets that contain the necessary data in a suitable format for the model to function effectively.

Breast cancer risk assessment models such as the Gail, BOADICEA and IBIS have contributed greatly to the development of future models. However, each model has notable limitations. The Gail model omits a number of key risk factors such as genetic mutations and benign disease history, limiting the accuracy of the model. While the BOADICEA and IBIS models incorporate genetic mutations, they rely heavily on detailed patient data limiting their accessibility. Emerging breast cancer risk assessment development methods such as Bayesian Networks offer greater accuracy and flexibility but share the same data limitations. The critical gap identified between various breast cancer risk assessment models is the lack of a standardised testing criterion which results in a variation of testing methods being used with different data qualities across studies. The varying testing methods and data quality make comparisons of performance and applicability difficult, thus showing the need for a standardised validation process that will allow for clinical adoption.

III. Methodology

This study aims to develop a Bayesian Network model capable of performing breast cancer risk assessments, through leveraging a comprehensive understanding of known risk factors. The methodology was divided into four stages:

- Secondary Research
- Bayesian Network Structure
- Bayesian Network Parameters
- Bayesian Network Reasoning

A. Secondary Research

Developing an understanding of risk factors associated with breast cancer development is essential for the development of effective and efficient assessment tools. Secondary research and an extensive literature review was conducted to collect existing knowledge on breast cancer risk factors. A systematic search of various seminal paper databases such as Google Scholar and Scopus was conducted to identify studies related and contributing to breast cancer risk factors. Four main categories of risk factors were identified and categorised into:

- Lifestyle risks

- Medical and Personal history risks
- Reproductive risks
- Genetic and hormonal risks

Official statistics from reputable sources such as the National Cancer Institute and Cancer Research UK were then gathered to extract quantitative data for the model's parameters. The identified statistics provided a basis for the Bayesian Network parameters. The significance of each risk factor in relation to breast cancer development was then evaluated based upon its prevalence in literature and its documented impact on breast cancer risk. This evaluation, enabled the assignment of appropriate weights to each risk factor.

B. Bayesian Network Structure

After the risk factors were identified, the Bayesian Network's structure was developed using Agena.ai, a probabilistic modelling and decision analysis platform, https://www.agena.ai/. Bayesian Networks are graphical models that represent probabilistic relationships amongst the identified risk factors. The structure of the model was derived from the relationships and interactions between variables identified in the secondary research.

The model consists of fourteen key risk factors categorised into four key groups:

- Genetic and Hormonal Risks – Includes factors such as BRCA1 and BRCA2 mutations as well as hormone replacement therapies
- Personal and Family History Risks – Includes factors such as the presence of benign diseases and family history of breast cancer
- Lifestyle Risks – Includes factors such as alcohol consumption, smoking and physical activity
- Reproductive Risks – Includes factors such as parity and breastfeeding history

The structure of the model was developed using the synthesis idiom proposed by Kyrimi et al. (Kyrimi et al.,2020). This idiom allows for the limiting of the number of learned parameters through grouping into intermediate nodes, which both simplify the network and reduce the number of direct parameters that need to be learned. The grouping of risk factors using the synthesis idiom allowed the model to capture casual relationships between risk factors and breast cancer whilst managing the model's complexity.

In addition to the synthesis idiom the risk factor idiom was also utilised. This idiom was employed to connect the categorised intermediate nodes to the target outcome, the risk of developing breast cancer. This approach enabled the model to represent the cause-effect relationship between the risk factors and their combined impact on the likelihood of developing breast cancer.

C. Bayesian Network Parameters

As the structure of the Bayesian Network was established the model's parameters were developed to quantify the relationships between risk factors. The probability of each risk factor affecting the development of breast cancer was learnt from secondary research and official statistics. For example, BRCA1 and BRCA2 genetic mutations are associated with a significantly higher risk of breast cancer development (Vaidyanathan & Kaklamani, 2021), whilst HRT (Hormone Replacement Therapy) was associated with only a slight increase in risk. (NHS Choices, 2023).

Each parameter was assigned a probability derived from the secondary research and was then incorporated into the network's Node Probability Tables (NPTs). To ensure the accurate and appropriate representation of the severity of each risk factor, the weighted mean equation was used to reflect the varying impacts of numerous risk factors. The weighted mean equation aggregated the impact of multiple risk factors within the same category to calculate the categories overall risk, the overall risk of the four categories was then combined into a further weighted mean equation to calculate the overall risk.

Equation 1 Weighted Mean Equation

$$x = \frac{\sum(w_i \cdot x_i)}{\sum w_i}$$

Equation 1 shows the weighted mean equation, where the sum of the weighted values is divided by the total weight to determine the overall contribution to the risk of breast cancer.

D. Bayesian Network Reasoning

As no data was available, the model's performance (accuracy, calibration, discrimination) could not be tested using classical re-sampling techniques, such as cross-validation or bootstrapping. Available breast cancer datasets such as the Breast Cancer Wisconsin Diagnostic Data Set (Kaggle, 2016) contained little to no risk factor information. Furthermore, the combining of datasets is not plausible. The model's reasoning was evaluated using a case-study approach, instead. Three fictional patients, each of which represents a different risk group, were used to test whether the model could accurately identify the low, medium, and high-risk individuals.

IV. Results

A. Bayesian Network Structure

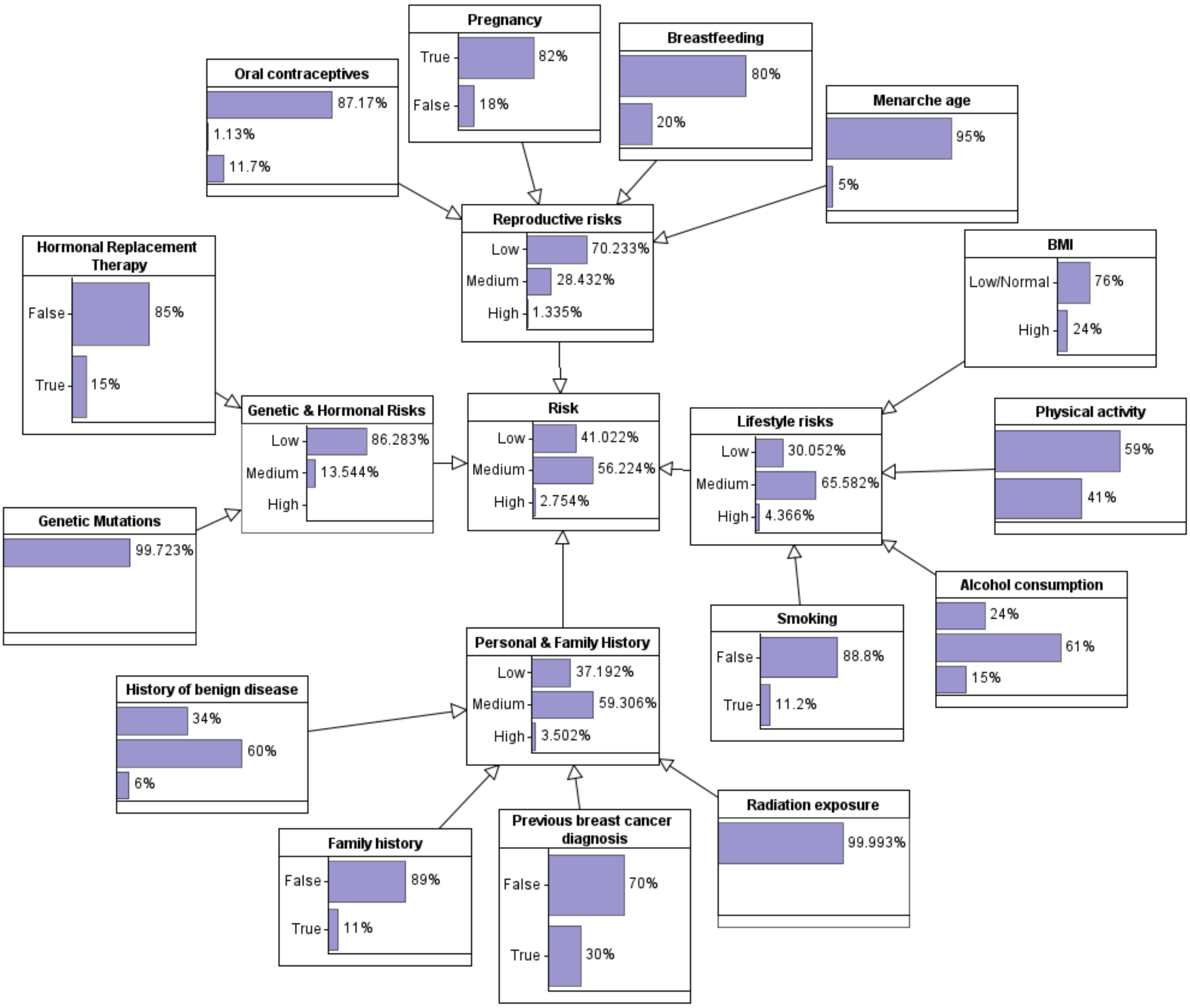
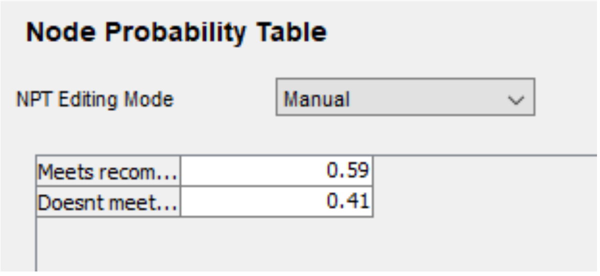


Figure 1 Bayesian Network Layout

Figure 1 displays the layout of the Bayesian Network, and the official statistics gathered during secondary research. The risk factors were categorised into the above mentioned four categories (Lifestyle risks, Medical and Personal History risks, Reproductive risks and Genetic and Hormonal risks), the overall risk of each category is then used to calculate the overall risk of the patient.

B. *Bayesian Network Parameters*

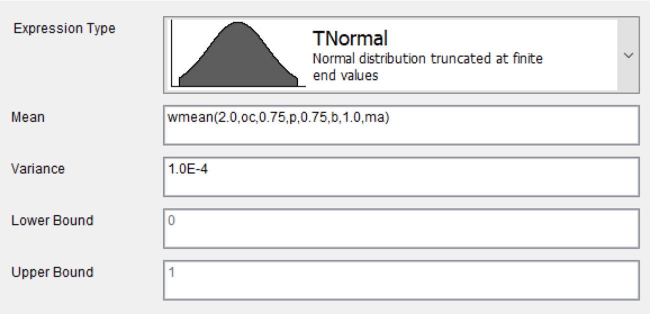
Statistical probabilities for each risk factor were sourced from reputable official statistics and seminal studies and were then inputted into the Bayesian Network model as seen in Figure 1. These probabilities form the model's Node Probability Tables (NPTs), which display the effect of each risk factor on the development of breast cancer. Figure 2 displays the Physical Activity node NPT that states that 59% of people meet the recommended amount of physical activity recommended by the World Health Organisation. The probabilities were then inputted for each of the fourteen risk factor nodes. The weighted mean equation seen in Figures 3 to 6 was then used to ensure that each risk factor has an appropriate effect on the risk based upon the secondary research. The weighted mean equation allows for each risk factor to have varying degrees of impact enabling an accurate display of how the risk factors affect the development of breast cancer. Tables 5 to 8 display the weight value assigned for each node within the model, each weight was assigned based upon the highest risk percentage change the risk factor can induce. For example, radiation exposure can increase the risk of breast cancer by approximately 35% and a family history of breast cancer can increase the risk up to 60%, thus radiation exposure is assigned a weight of 0.5 whilst family history is assigned a weight of 1, as it has a more significant impact. Once all the weights were assigned to the nodes, the weighted mean equation for the risk categories was calculated. The percentage changes for each node in a category were summed together and the total of each category was divided by the largest value, this value was then multiplied by 10 to scale all the results and was then rounded to one decimal place, the category weights are displayed in Table 9. The weighted mean and scaling approach ensures that the results of the model are explainable as each nodes weight value is based upon secondary research and therefore, improves the model's accuracy as well as allows for accurate risk stratification of low, medium, and high-risk individuals.


Node Probability Table

NPT Editing Mode: Manual

Meets recom...	0.59
Doesnt meet...	0.41

Figure 2 Physical Activity NPT


Expression Type: TNormal — Normal distribution truncated at finite end values

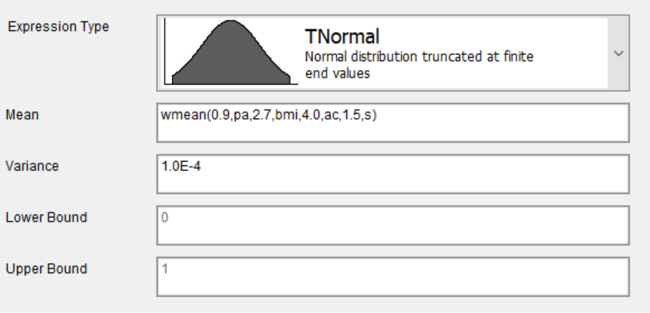
Mean: wmean(2.0,oc,0.75,p,0.75,b,1.0,ma)

Variance: 1.0E-4

Lower Bound: 0

Upper Bound: 1

Figure 3 Reproductive Risks Weighted Mean


Expression Type: TNormal — Normal distribution truncated at finite end values

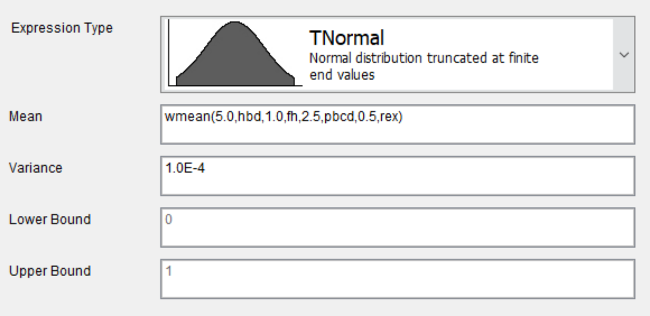
Mean: wmean(0.9,pa,2.7,bmi,4.0,ac,1.5,s)

Variance: 1.0E-4

Lower Bound: 0

Upper Bound: 1

Figure 4 Lifestyle Risks Weighted Mean


Expression Type: TNormal — Normal distribution truncated at finite end values

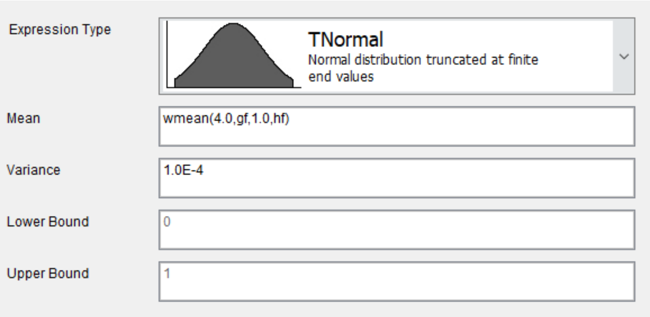
Mean: wmean(5.0,hbd,1.0,fh,2.5,pbcd,0.5,rex)

Variance: 1.0E-4

Lower Bound: 0

Upper Bound: 1

Figure 5 Personal & Family History Risks Weighted Mean


Expression Type: TNormal — Normal distribution truncated at finite end values

Mean: wmean(4.0,gf,1.0,hf)

Variance: 1.0E-4

Lower Bound: 0

Upper Bound: 1

Figure 6 Genetic and Hormonal Risks Weighted Mean

V. DISCUSSION

The development of a breast cancer risk assessment Bayesian Network model has proved to provide promising results and insights into the usage of probabilistic graphical models capable of enhancing early-stage detection and providing personalised risk evaluations. The study integrated a variety of known risk factors and aimed to offer an alternative risk assessment tool to typical methods. The results of the model can be seen in Figures 7 to 9.

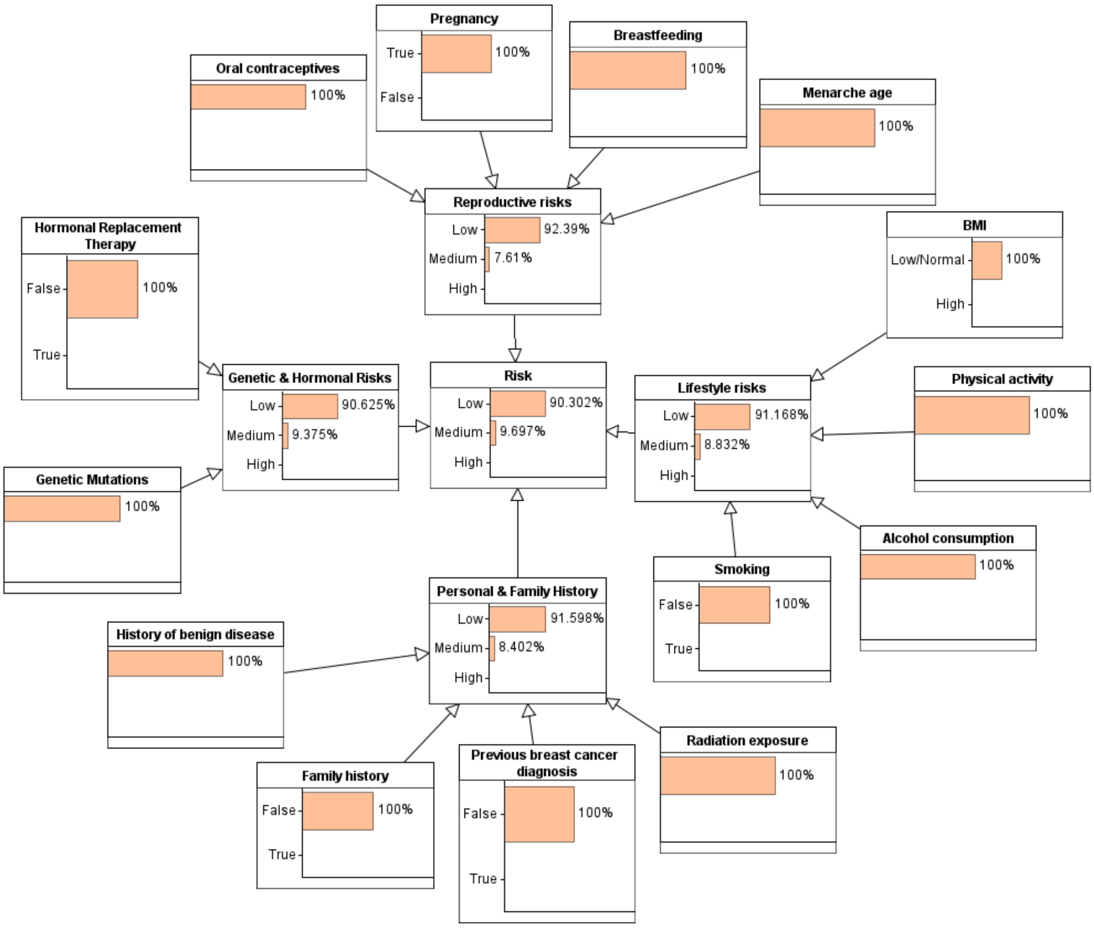


Figure 7 Low Risk Individual Results

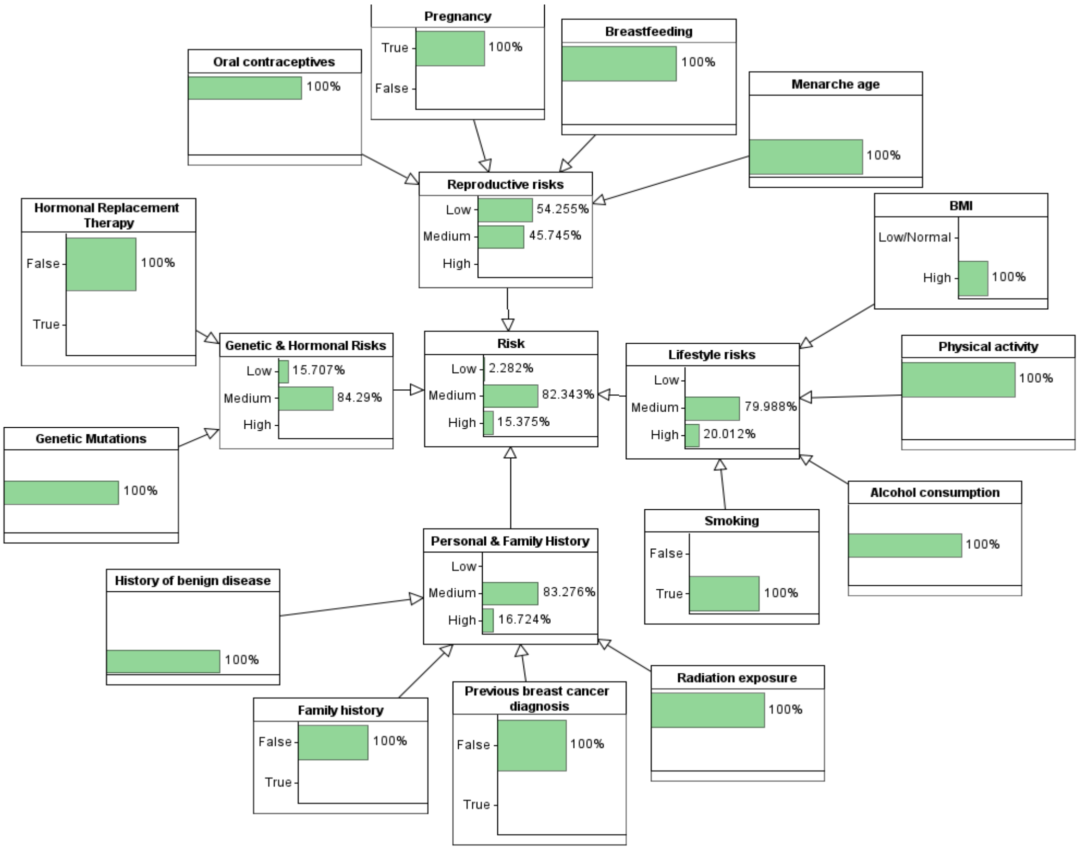


Figure 8 Medium Risk Individual Results

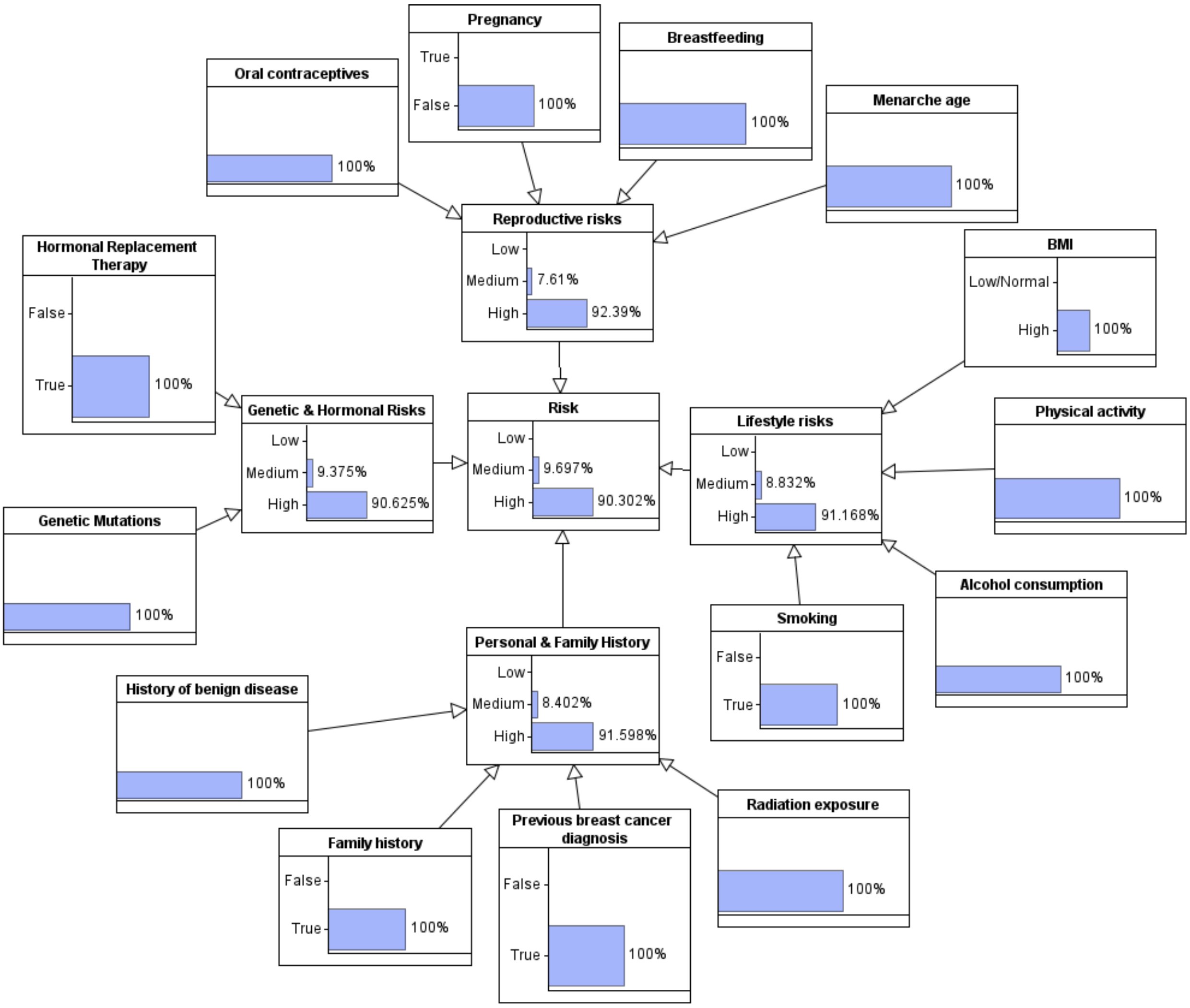


Figure 9 High Risk Individual Results

The developed model has shown potential to improve early-stage detection through the leveraging of key risk factors identified in both literature and official statistics. The model's design allows for a personalised approach due to the variety of customisable options available to the patient, thus this integrated approach enhances the ability to predict the patients risk of developing breast cancer.

Compared to other previously established models such as the Gail or BOADICEA models the Bayesian Network is far more flexible and adaptable. The model can incorporate all the known risk factors providing a much broader risk assessment compared to the previous models. An example of this is the Gail model's primary focus on family medical history whilst other key factors such as genetic mutations are omitted. The ability to function with varying types of data potentially makes the tool more practical in varied clinical settings.

Practical implications of implementing the model in clinical settings are substantial, the model's ability to offer a risk stratification tool that incorporates most known risk factors facilitates earlier detection and more timely screening and further testing recommendations. The identification of those most at risk can allow for the recommendation of targeted screening plans, contributing to early-stage detection and thus contributing to an increase in survival and recovery rates.

The model is limited by the unavailability of an appropriate dataset, the model relies solely on literature and statistics, thus potentially not capturing all relevant factors especially within specific populations. Furthermore, due to the limited data availability the model was evaluated through a series of fictional case studies, potentially affecting the model's results. If an appropriate dataset was compiled ethical issues arise, the amount of personal data collected and its usage may not be approved by all patients, furthermore data storage and ethical usage is affected by General Data Protection Regulation (GDPR) laws.

VI. Future Work

The future work of this project will focus on two key factors:

- Usage of real data

- Inclusion of manifestation variables such as signs and symptoms

The usage of real data for both training and validation of the model will allow for the refinement of the model to achieve enhanced predictive accuracy. Incorporating data from various ethnic backgrounds will allow for the model to learn from actual cases improving its reliability and applicability in clinical settings as it will be capable of providing accurate risk stratification and analysis for all ethnicities.

Inclusion of manifestation variables such as early signs and symptoms of breast cancer will further increase the model's predictive accuracy. Variables such as lumps or changes in breast shape and size as well as other physical symptoms can be included; the integration of manifestation variables will allow the model to account for both risk factors and clinical indicators. This approach will enable faster identification of those at greater risk of developing breast cancer and therefore will enable faster screening and early-stage diagnosis decreasing the mortality rate.

VII. Conclusion

This project has successfully developed a Bayesian Network model for breast cancer risk assessment that incorporates a variety of known key risk factors. The model provides a personalised and data-driven approach to early-stage detection and is capable of accurately classifying patients into low, medium, and high-risk categories. Successful classification of patients can potentially enhance early-stage detection for high-risk patients through increased screening and additional testing, improving detection, treatment and survival rates. Although the model currently displays promising results, it is currently limited by the absence of real-world datasets and was therefore validated on simulated case studies. Future work on the model will focus on obtaining real patient data to refine and optimise the model based upon real data to boost its predictive power. The model has the potential to offer a practical, adaptable tool for healthcare providers, contributing to targeted screening and early-detection allowing for a greater treatment and recovery rate whilst reducing mortality rates.

References

1. Vaidyanathan, A. and Kaklamani, V. (2021) 'Understanding the clinical implications of low penetrant genes and breast cancer risk', *Current Treatment Options in Oncology*, 22(10). doi:10.1007/s11864-021-00887-4.
2. *NHS choices*. (2024) Available at: https://www.nhs.uk/conditions/predictive-genetic-tests-cancer/#:~:text=BRCA1%20and%20BRCA2%20are%20examples,breast%20cance%20and%20prostate%20cancer
3. Best, S. (2024) *Hereditary cancers - Li Fraumeni syndrome*, *Manchester Cancer Research Centre*. Available at: https://www.mcrc.manchester.ac.uk/hereditarycancers-li-fraumeni-syndrome/
4. *Male and female populations* (2023) *GOV.UK Ethnicity facts and figures*. Available at: https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/demographics/male-and-female-populations/latest/
5. *NHS choices (2023)*. Available at: https://www.nhs.uk/medicines/hormone-replacementtherapy-hrt/benefits-and-risks-of-hormone-replacement-therapy-hrt/
6. *Hundreds of thousands of women experiencing menopause symptoms to get cheaper HRT GOV.UK.*(2024) Available at: https://www.gov.uk/government/news/hundreds-of-thousands-of-women-experiencing-menopause-symptoms-to-get-cheaper-hormone-replacement-therapy
7. Dyrstad, S.W. *et al.* (2015) 'Breast cancer risk associated with benign breast disease: Systematic review and meta-analysis', *Breast Cancer Research and Treatment*, 149(3), pp. 569–575. doi:10.1007/s10549-014-3254-6.
8. Oyelowo, T. (2007) *Mosby's Guide to Women's Health: A Handbook for Health Professionals*. St. Louis, Mo: Mosby Elsevier.
9. Myers, D.J. (2023) *Atypical breast hyperplasia*, *StatPearls [Internet]*. Available at: https://www.ncbi.nlm.nih.gov/books/NBK470258/
10. Shiyanbola, O.O. *et al.* (2017) 'Emerging trends in family history of breast cancer and associated risk', *Cancer Epidemiology, Biomarkers &amp; Prevention*, 26(12), pp. 1753–1760. doi:10.1158/1055-9965.epi-17-0531.
11. Durham, D.D. *et al.* (2022) 'Breast cancer incidence among women with a family history of breast cancer by relative's age at diagnosis', *Cancer*, 128(24), pp. 4232–4240. doi:10.1002/cncr.34365.
12. Ramin, C. *et al.* (2021) 'Risk of contralateral breast cancer according to first breast cancer characteristics among women in the USA, 1992–2016', *Breast Cancer Research*, 23(1). doi:10.1186/s13058-021-01400-3.
13. Courtney, D. *et al.* (2022) 'Breast cancer recurrence: Factors impacting occurrence and survival', *Irish Journal of Medical Science (1971 -)*, 191(6), pp. 2501–2510. doi:10.1007/s11845-022-02926-x.
14. *Breast cancer prevention (PDQ®) Breast Cancer Prevention (PDQ®) – NCI (2024)*. Available at: https://www.cancer.gov/types/breast/hp/breast-prevention-pdq#_149_toc
15. *What is hodgkin lymphoma?* (2024) *Cancer Research UK*. Available at: https://www.cancerresearchuk.org/about-cancer/hodgkin-lymphoma/about
16. *Oral contraceptives (birth control pills) and cancer risk Oral Contraceptives (Birth Control Pills) and Cancer Risk - NCI.* (2018) Available at: https://www.cancer.gov/about-cancer/causes-prevention/risk/hormones/oral-contraceptives-fact-sheet
17. *Reproductive Health Profiles: Statistical Commentary GOV.UK.*(2024b) Available at: https://www.gov.uk/government/statistics/reproductive-health-2023-update/reproductive-health-profiles-statistical-commentary
18. *Reproductive history and cancer risk NCI.* (2016) Available at: https://www.cancer.gov/about-cancer/causes-prevention/risk/hormones/reproductive-history-fact-sheet
19. Sharfman, A. (2022) *Childbearing for women born in different years, England and Wales: 2020*, *Childbearing for women born in different years, England and Wales - Office for National Statistics*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/conceptionandfertilityrates/bulletins/childbearingforwomenbornindifferentyearsenglandandwales/2020 (
20. Brinton, L.A. *et al.* (1995) 'Breastfeeding and breast cancer risk', *Cancer Causes and Control*, 6(3), pp. 199–208. doi:10.1007/bf00051791.
21. *Breastfeeding in the UK - position statement RCPCH (2022)*. Available at: https://www.rcpch.ac.uk/resources/breastfeeding-uk-position-statement
22. Goldberg, M. *et al.* (2020) 'Pubertal timing and breast cancer risk in the sister study cohort', *Breast Cancer Research*, 22(1). doi:10.1186/s13058-020-01326-2.
23. *New study reveals factors behind age of girls' first period The Institute of Cancer Research (2010).* Available at: https://www.icr.ac.uk/news-archive/new-study-reveals-factors-behind-age-of-girls-first-period
24. Dehesh, T. *et al.* (2023) 'The relation between obesity and breast cancer risk in women by considering menstruation status and geographical variations: A systematic review and meta-analysis', *BMC Women's Health*, 23(1). doi:10.1186/s12905-023-02543-5.

25. *Obesity statistics*, Carl Baker (2023) Available at: https://researchbriefings.files.parliament.uk/documents/SN03336/SN03336.pdf
26. Xu, Y. and Rogers, C.J. (2020) ‘Physical activity and breast cancer prevention: Possible role of immune mediators’, *Frontiers in Nutrition*, 7. doi:10.3389/fnut.2020.557997.
27. *NHS choices(2023b)*. Available at: https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/2021-part-2/physical-activity
28. Boyle, P. and Boffetta, P. (2009) ‘Alcohol consumption and breast cancer risk’, *Breast Cancer Research*, 11(S3). doi:10.1186/bcr2422.
29. *NHS choices (2022)*. Available at: https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/2021/part-3-drinking-alcohol#:~:text=Among%20those%20adults%20that%20drank,at%20lower%20levels%20than%20men
30. Jones, M.E. *et al.* (2017) ‘Smoking and risk of breast cancer in the generations study cohort’, *Breast Cancer Research*, 19(1). doi:10.1186/s13058-017-0908-4.
31. Lauren Revie, D.M. (2023) *Adult smoking habits in the UK: 2022*, *Adult smoking habits in the UK - Office for National Statistics*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/bulletins/adultsmokinghabitsingreatbritain/2022
32. UK, C.R. (2011) *One woman in eight will get breast cancer*, *Cancer Research UK - Cancer News*. Available at: https://news.cancerresearchuk.org/2011/02/04/one-woman-in-eight-will-get-breast-cancer/
33. *Early detection and diagnosis of cancer roadmap* (2024b) *Cancer Research UK*. Available at: https://www.cancerresearchuk.org/funding-for-researchers/research-opportunities-in-early-detectionand-diagnosis/early-detection-and-diagnosis-roadmap
34. Tian, N. *et al.* (2012) ‘Identifying risk factors for disparities in breast cancer mortality among AfricanAmerican and Hispanic women’, *Women’s Health Issues*, 22(3). doi:10.1016/j.whi.2011.11.007.
35. Stevanato, K. *et al.* (2022) ‘Use and applicability of the Gail model to Calculate Breast Cancer Risk: A scoping review’, *Asian Pacific Journal of Cancer Prevention*, 23(4), pp. 1117–1123.doi:10.31557/apjcp.2022.23.4.1117.
36. Antoniou, A.C. *et al.* (2004) ‘The boadicea model of genetic susceptibility to breast and ovarian cancer’,*British Journal of Cancer*, 91(8), pp. 1580–1590. doi:10.1038/sj.bjc.6602175.
37. Kurian, A.W. *et al.* (2021) ‘Performance of the Ibis/tyrer-cuzick model of breast cancer risk by race and ethnicity in the Women’s Health initiative’, *Cancer*, 127(20), pp. 3742–3750. doi:10.1002/cncr.33767.
38. Burke, H.B. *et al.* (1997) ‘Artificial Neural Networks improve the accuracy of cancer survival prediction’, *Cancer*, 79(4), pp. 857–862. doi:10.1002/(sici)1097-0142(19970215)79:4&lt;857::aidcncr24&gt;3.0.co;2-y.
39. Zhang, B., Shi, H. and Wang, H. (2023) ‘Machine learning and AI in cancer prognosis, prediction, and treatment selection: A critical approach’, *Journal of Multidisciplinary Healthcare*, Volume 16, pp. 17791791. doi:10.2147/jmdh.s410301.
40. Pearl, J. (2014) *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco: Morgan Kaufmann.
41. Burnside, E.S. *et al.* (2006) ‘Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: Initial experience’, *Radiology*, 240(3),pp. 666–673. doi:10.1148/radiol.2403051096.
42. Amir, E. *et al.* (2010) ‘Assessing women at high risk of breast cancer: A review of risk assessment models’, *JNCI Journal of the National Cancer Institute*, 102(10), pp. 680–691. doi:10.1093/jnci/djq088.
43. Kyrimi, E. *et al.* (2020) ‘Medical idioms for Clinical Bayesian Network Development’, *Journal of Biomedical Informatics*, 108, p. 103495. doi:10.1016/j.jbi.2020.103495.
44. Learning, U.M. (2016) *Breast cancer wisconsin (diagnostic) data set*, *Kaggle*. Available at: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data (Accessed: 14 August 2024).

APPENDIX

Table 1 Genetic and Hormonal Risks

Variable Name	Variable State	Literature	Knowledge/Statistics
Genetic Mutations	High penetrance Medium penetrance Low/No Penetrance	(Vaidyanathan & Kaklamani, 2021), (NHS Choices, 2024), (Best,2024), (Gov.uk, 2023)	- High penetrance increases risk by 60% -Medium penetrance by 15-60% -1 in 400 women have BRCA (High)mutations, using population 0.25% of women have it -1 in 5000 people have medium penetrance mutations, using population 0.02% of women have it. Approximately 30 million women according to 2021 census.
Hormone Replacement Therapy	True False	(NHS Choices, 2023), (Gov.uk, 2024)	- HRT increases risks very slightly - 15% of women are prescribed HRT in the UK

Table 2 Personal and Family History Risks

Variable Name	Variable State	Literature	Knowledge/Statistics
History of benign disease	No disease Benign disease without atypia Benign disease with atypia	(Dyrstad et al., 2015), (Oyelowo, 2007), (Myers, 2023)	-Benign disease with atypia increases risk by four to five times -Benign disease without atypia increases risk by one and a half to two times - 60% of women experience benign breast diseases
Family history	True False	(Shivanbola et al.,2017), (Durham et al., 2022)	-Family history of breast cancer increases risk by 60% -Approximately 11% of women have a first degree family member with a history of breast cancer
Previous breast cancer diagnosis	True False	(Ramin et al., 2021), (Courtney et al., 2022)	-A previous breast cancer diagnosis increases risk twofold -25-30% of patients develop recurrence of breast cancer
Radiation exposure	High medical Normal medical	(Gov.uk, 2023), (NCI, 2024), (Cancer Research UK, 2024)	-High radiation medical treatments such as for Hodgkin's Lymphoma can cause a 35% increase in breast cancer development risk - Approximately 2100 people develop it each year in the UK therefore using the 30 million population it can be

			inferred that approximately 0.007% women will be affected

Table 3 Reproductive Risks

Variable name	Variable state	Literature	Knowledge/Statistics
Oral contraceptives	Not used Previously used Currently using	(Gov.uk, 2023), (NCI, 2018), (Gov.uk, 2024b)	-117 per 1000 women use short term oral contraceptives, approximately 3510000 women, 11.7% in 2022-2023 -128.3 per 1000 in 2021-2022 approximately 3850000 women, 12.83% -12.83-11.7 = 1.13 therefore 1.13% of women previously used -Previous use increases risk by 7% -Current use increases risk by 24%
Pregnancy	True False	(NCI, 2016), (Sharfinan, 2022)	-Women with 5 children have their risk halved therefore a 10% decrease per child can be inferred -18% of women were childless according to the statistic therefore it can be inferred 82% of women have at least one child and pregnancy.
Breastfeeding	Breastfed Never breastfed	(Brinton et al., 1995), (RCPCH, 2022)	-Breastfeeding reduces the risk to 0.87 and the number of children doesn't affect this value -81% of women initiated breastfeeding therefore it can be assumed 80% of women breastfed
Menarche age	Twelve or later Eleven or before	(Goldberg et al., 2020), (ICR, 2010)	-First periods before the age of 12 increase the risk by 10% -5% of menarches are before the age of eleven

Table 4 Lifestyle Risks

Variable name	Variable state	Literature	Knowledge/Statistics
BMI	High Normal	(Dehesh et al., 2023), (Baker, 2023)	- Pre menopausal women with a high BMI decreases risk to 0.93 whilst in postmenopausal women it increases to 1.26 - Using the values provided by the government a value of 24% of women can be

			classified as having a high BMI.
Physical Activity	Equal to or more than WHO recommendation Less than WHO recommendation	(Xu & Rogers, 2020), (NHS Choices, 2023b)	-If the 2.5 hours recommended a week are fulfilled the risk is reduced to 0.88. -Inadequate physical activity leads to increased BMI and therefore depending on the menopausal status an increase or decrease in risk. -59% of women meet the recommended 150 minutes of aerobic activity.
Alcohol consumption	High Medium Low	(Boyle & Boffeta, 2009), (NHS Choices 2022)	-1-2 drinks per day increase the risk by approximately 20% -3 drinks a day or more increase the risk by 40% -61% of women drink at medium/low levels of harm -15% of women drinks at higher risk levels
Smoking	True False	(Jones et al., 2017), (Revie, 2023)	- Women who have ever smoked have an increased risk of 14% -11.2% of women in the UK smoke

Weight tables

Table 5 Genetic and Hormonal Weights

Risk	Highest risk percentage	Weight
Genetic Mutation	60%	4
Hormone Replacement therapy	5%	1

Table 6 Personal and Family History Weights

Risk	Highest risk percentage	Weight
History of benign disease	500% increase	5
Family history	60% increase	1
Previous diagnosis	200% increase	2.5
Radiation exposure	35% increase	0.5

Table 7 Reproductive Weights

Risk	Highest risk percentage	Weight
Oral contraceptives	24% increase	2
Pregnancy	10% decrease	0.75
Breastfeeding	13% decrease	0.75

Menarche age	10% increase	1

Table 8 Lifestyle Weights

Risk	Highest risk percentage	Weight
BMI	Increases by 26%	2.7
Physical activity	Reduced by 12%	0.9
Alcohol consumption	40% increase	4
Smoking	14% increase	1.5

Table 9 Overall Risk Weights

Risk Category	Overall risk percentage	Weight
Genetic risks	65% increase	0.8
Personal and family history risks	795% increase	10
Reproductive risks	11% increase	0.1
Lifestyle risks	68% increase	0.9

Weights were calculated through dividing by the largest percentage increase and then scaling the results by multiplying by 10 to achieve single digits rounded to one decimal place.