doi:10.20944/preprints202112.0018.v1

Article

Machine-learning to discern interactive clusters of risk factors for late recurrence of metastatic breast cancer

Juan Luis Gomez Marti[1,4], Adam Brufsky[2,5], *Alan Wells[1,4,5], **Xia Jiang[3]

[1] Department of Pathology, University of Pittsburgh and [5] Hillman Cancer Center
[2] Department of Medicine, University of Pittsburgh and [5] Hillman Cancer Center
[3] Department of Biomedical Informatics, University of Pittsburgh
[4] R&D Service, Pittsburgh VA HealthSystem

* Correspondence: A.W wellsa@upmc.edu; **Contact for questions about the paper, methods and data: X.J. xij6@pitt.edu

Simple Summary: Breast cancer is the most frequently diagnosed cancer and second leading cause of cancer-related death among women worldwide. After initial tumor resection, breast cancer may recur locally and/or in distant organs within several months to years or even decades. Multiple methods exist to prognosticate disease progression in the early months and years after diagnosis. However, further efforts are needed to identify risk factors that relate to recurrence beyond the initial 5-year window. In this study, we applied machine-learning to retrieve single and interactive clinical and pathological risk factors of 5-, 10- and 15-year metastasis.

Abstract: Background: Risk of metastatic recurrence of breast cancer after initial diagnosis and treatment depends on the presence of a number of risk factors. Although most univariate risk factors have been identified using classical methods, machine-learning methods are also being conducted to tease out non-obvious contributors to a patient's individual risk of developing late distant metastasis. Bayesian-network algorithms may predict not only risk factors but also interactions among these risks, which consequently lead to metastatic breast cancer. We proposed to apply a previously developed machine-learning method to predict risk factors of 5-, 10- and 15-year metastasis. Methods: We applied a previously validated algorithm named the Markov Blanket and Interactive risk factor Learner (MBIL) on the electronic health record (EHR)-based Lynn Sage database (LSDB) from the Lynn Sage Comprehensive Breast Cancer at Northwestern Memorial Hospital. This algorithm provided an output of both single and interactive risk factors of 5-, 10-, and 15-year metastasis from LSDB. We individually examined and interpreted the clinical relevance of these interactions based on years to metastasis and the reliance on interactivity between risk factors. Results: We found that with lower alpha values (low interactivity score), the prevalence of variables with an independent influence on long term metastasis was higher (i.e., HER2, TNEG). As the value of alpha increased to 480, stronger interactions were needed to define clusters of factors that increased the risk of metastasis (i.e., ER, smoking, race, alcohol usage). Conclusion: MBIL identified single and interacting risk factors of metastatic breast cancer, many of which were supported by clinical evidence. These results strongly recommend the development of further large data studies with different databases to validate the degree to which some of these variables impact metastatic breast cancer in the long term.

Keywords: metastatic breast cancer, metastasis, causal learning, machine learning, Markov Blanket and Interactive risk factor Learner (MBIL), risk factors

1. Introduction

Women who are diagnosed with invasive breast cancer will likely present with distant recurrence in the years after diagnosis [1]. Patterns and time to recurrence varies depending on tumor subtypes and the presence of concomitant biomarkers, as well as other clinical risk factors. In this regard, while recurrence occurs most frequently within the first 5 years after diagnosis in estrogen receptor (ER)-negative breast cancer, ER-positive tumors remain at higher risk for recurrence at later times including decades later. Tamoxifen use greatly reduces the 5-year recurrence risk of ER-positive tumors, but still the annual increase in risk of recurrence is 2% for at least 15 years[2]. Remarkably, prolonged tamoxifen use seems to further reduce the onset of metastasis [3]. Once distant recurrence occurs, patients usually have a poor prognosis [1].

Original lymph node (LN) presence and tumor diameter are essential clinical predictors of late (5 to 20 years after diagnosis) recurrence in ER-positive tumors. Tumor grade, Ki-67 positivity, and progesterone receptor (PR) status have also been found as predictors of recurrence, but only in the first 5 years after diagnosis. Importantly, a considerable risk of late recurrence is still present even among women with T1N0 disease [4]. These clinical predictors were also identified in another study [5]. Gene signatures have also been characterized to predict recurrence and need for adjuvant chemotherapy, although these apply mainly to the first decade after diagnosis [5,6].

To date, factors that predict breast cancer recurrence in the long term have not been fully characterized. Given the need for more data to predict breast cancer recurrence, so as to guide our clinical follow-up and approaches in these patients, artificial intelligence is being implemented. We previously validated a method that used Bayesian networks and information theory to identify key risk factors for breast cancer metastasis more accurately than other known Bayesian network learning algorithms [7]. This algorithm, named as Markov Blanket and Interactive risk factor Learner (MBIL), learns single and interactive risk factors that have a direct influence on a patient's outcome. Risk factors that are dependent on other variables to have a predictive effect are called interactions [8,9]. In the present study, we applied MBIL to learn both a set of direct risk factors and a set of interactive risk factors for 5-, 10- and 15-year recurrence. This algorithm extracted risk factors from the Lynn Sage Database (LSDS) at Northwestern Memorial Hospital, as previously described [7,10]. Direct and interactive risk factors are presented herein.

2. Materials and Methods

Bayesian networks (BNs) have become a leading architecture for modeling uncertain reasoning in artificial intelligence and machine learning. A Medline search reveals that 3,910 papers contained the term "Bayesian network" from 2003 to 2017, while only 252 contained that term from 1993 to 2002. A Bayesian network (BN) consists of a directed acyclic graph (DAG), whose node set contains random variables, and the conditional probability distribution of every variable in the network given each set of values of its parents [11,12].

In general, the Markov blanket of a node T in a Bayesian network model consists of all parents of T, children of T, and parents of children of T [13]. Figure 1 [7] shows a Bayesian network DAG structure in which the node T is a leaf node because it has no children. So, the Markov blanket of T only consists of its parents, namely nodes X11 through X15. If we run a machine learning algorithm without knowing the BN DAG structure, nodes X1 through X10, X16, and X17 would all be learned as risk factors of T because these nodes can pass information to T through the parent nodes, i.e., the direct risk factors of T, even though they don't have a direct influence on T. Hence, when learning a BN DAG, we will be able to identify the direct risk factors of a node T via Markov blanket. This will help get rid of the background noise which often affects the prediction performance. By incorporating our previous work concerning learning interaction from data [8,9], we developed the Markov Blanket and Interactive risk factor Learner (MBIL) method which not only identifies the Markov Blanket of a node like T but also detects interactive risk factors of a note like T [7]. Interactive risk factors work together to have a nonadditive joint-effect on a target node such as T. In Figure 1, there are two groups of interactive risk factors of T, nodes X13 and X14, and nodes X8 and X9. MBIL detects all direct risk factors included in the Markov Blanket of a target node like T, both a single one like node X11, X12, or X15 and interactive ones like nodes X13 and X14. MBIL also detects all other interactive risk factors such as nodes X8 and X9, regardless whether they are included in the Markov Blanket.

In this research we applied MBIL to learn the direct and interactive risk factors of 5-, 10-, and 15-year breast cancer metastasis. MBIL takes a score-based structure-learning approach to learn the Markov Blanket of a node. The Bayesian score is the probability of the data given the BN DAG [14]. It measures how well a BN DAG represents the data. We used the Bayesian Dirichlet equivalent uniform (BDeu) score [15] as our score criterion, which is a variant of the Bayesian score. Ideally, we would like to learn a model that represents the reality perfectly. But due to various reasons such as the complexity of a real-world problem and the limitations of data collected, it is often impossible to learn such a perfect model. Instead, a major task of machine learning is to adjust parameters to learn model(s) that represents the data most closely. As to MBIL we can adjust alpha, also called the Prior Equivalent Sample Size (PESS), which is a parameter built in the BDeu score [15]. Adjusting alpha can affect the complexity of BN models that are learned from data. This resembles somewhat the fishing activity, in which we can adjust the size of the fishnet holes to govern the sizes of seafoods we catch. In this study, we ran MBIL using three different values of alpha, namely, alpha =1, 120, and 480 each respectively when learning risk factors for 5-year, 10-year, or 15-year metastasis. MBIL reported both a set of interactive risk factors and a set of direct risk factors for each of the alpha values. Note that the set of direct risk factors learned by MBIL includes both single and interactive ones.

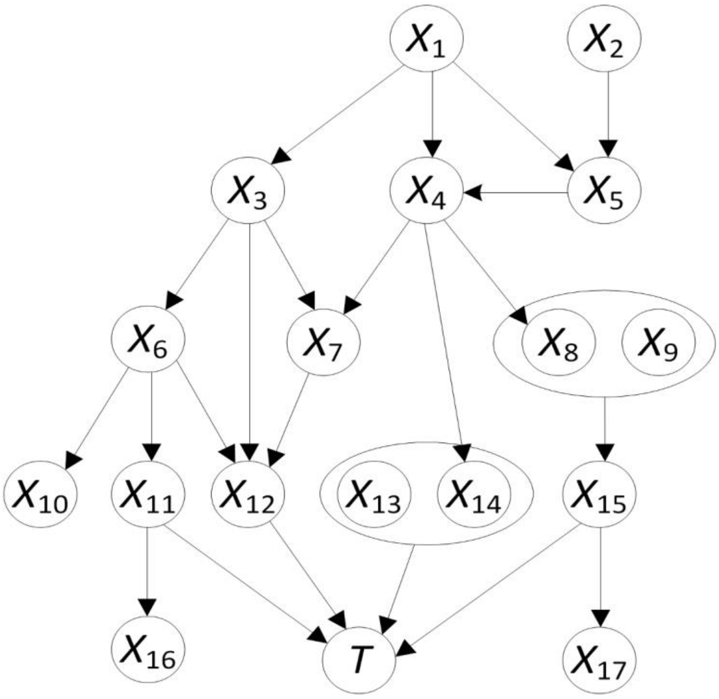

Figure 1. A BN DAG model illustrating Markov Blanket. The Markov Blanket of T consists of nodes X11, X12, X13, X14 and X15. These nodes are the direct risk factors of T and separate T from the influence of the noisy predictors X1-X10, X16, and X17 (adapted from [7]).

3. Results

Using the MBIL algorithm, we previously described the superior effectiveness of this method in learning direct risk factors in the context of 5-year breast cancer metastasis, which were also present in the literature [7]. In the present study, we first used MBIL to search causal sets of 5-, 10- and 15-year breast cancer metastasis. MBIL produced an output with three alpha values: 1, 120, and 480; and a list of interacting risk factors for each alpha value and time to metastasis. These learned interactions are ranged by its Bayesian score from the highest to lowest, which is the probability of data given the Bayesian Network model [7]. Additionally, an output was originated with direct causal sets of metastases. These are a few risk factors that necessarily interact together to produce the studied outcome [7].

We found that MBIL predicted that variables such as HER2 were frequent with low alpha values for all 5-, 10- and 15-year outcomes. As alpha became stronger, the presence of HER2 as a predictor of metastatic breast cancer (mBC) declined while ER became a stronger predictor, meaning that while HER2 was identified by MBIL as an independent predictor of mBC, stronger interactions between ER and other variables were necessary to predict the future occurrence of mBC.

Causal sets of metastases

When observing causal sets directly related to metastatic breast cancer, MBIL found that at 5 years, direct causal sets of metastases on alpha 1 were lymph node positivity and the interaction of TNEG (triple negative breast cancer) with HER2 assessments (Figure 2); at a more stringent alpha 120, direct causal sets at 5 years were the interaction of ER, n-TNM and surgical margins; and on the highly interacting alpha 480 the causal sets were stage, TNEG, and ER on interaction with n-TNM and surgical margins (Figure 2).

At 10 years, disease stage was a sole causal set of metastases with an alpha of 1; with an alpha of 120, having MRI evaluations within 60 days of surgery was a direct causal risk factor for metastases, and ER, n-TNM and surgical margins interacted to be direct causal sets; 15-year metastases causal sets were stage, MRIs_60_surgery, and the interaction of ER, n-TNM and surgical margins. With an alpha of 480, 5-year causal sets were the age at diagnosis, menopausal status, and lymph node status; at 10-year the causal sets were invasiveness of the tumor, and the interaction of age at diagnosis, menopausal status and lymph node status (Figure 2).

To predict mBC after 15 years, the causal sets were the interaction of lymph node status, histology and invasive tumor location, and the interaction of n-TNM, histology and invasive tumor location (Figure 2).

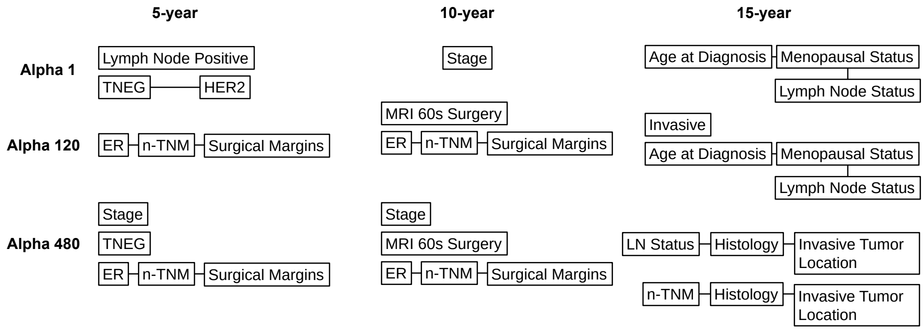

Figure 2. MBIL generated causal sets of 5-, 10- and 15-year breast cancer metastasis.

Learned interactions

Using the MBIL algorithm, we searched for direct interactions that were predictive of metastatic breast cancer. We first investigated the absolute frequency of known variables impacting breast cancer prognosis. Mainly, we calculated the frequency of ER, HER2, TNEG, and tumor grade. We found that ER was most frequent when looking at risk factors of 5-year metastases at an alpha of 120 (Figure 3). The presence of HER2 among predictive interactions was highest at alpha 1 with decreasing abundance as the time to late relapse became longer, suggesting that HER2 scored higher as an independent structure likely to predict 5-year metastasis rather than 10- or 15-year. TNEG was most predictive of tumor metastasis at 15-years when observing interactions using an alpha of 1, supporting its strength as a likely independent predictor of metastasis [16]. It was present to predict 5-year metastasis at an alpha 1 when in conjunction with HER2. At an alpha of 120, TNEG was found to predict 5-year metastasis when in conjunction with smoking and n-TNM on one interaction, and with n-TNM and invasive_tumor_location on another interaction. TNEG was found on one interaction to predict 10- and 15-year metastasis; for both predictions TNEG interacted with n-TNM and surgical_margins. Finally, at an alpha of 480 TNEG was only found on one interaction with n-TNM and surgical margins to predict 10-year metastasis.

We next found tumor grade to be a strong predictor of 5-year metastasis with an alpha of 1, whereas at higher alpha its frequency among interactions predicting metastasis was reduced. Herein, tumor grade appeared to be a strong independent predictor of 5-year metastasis given its frequent presence among interactions predicting this outcome.

We then calculated the absolute frequency of smoking/alcohol and race/ethnicity as more evidence is needed regarding these interactions to predict late recurrence. We found that smoking and/or alcohol were only present in one interaction for 15-year metastases with an alpha of 1. However, these were present in one interaction at 5- and 15-year metastases with an alpha of 120, and in up to 4 interactions to predict 5-year metastases at an alpha of 480 (Figure 3). The presence of the risk factors “race” and “ethnicity” were higher as time increased from 5- to 15-year recurrence, particularly at an alpha of 480, suggesting that race as a dependent variable may in fact favor strongly late recurrence (Figure 3).

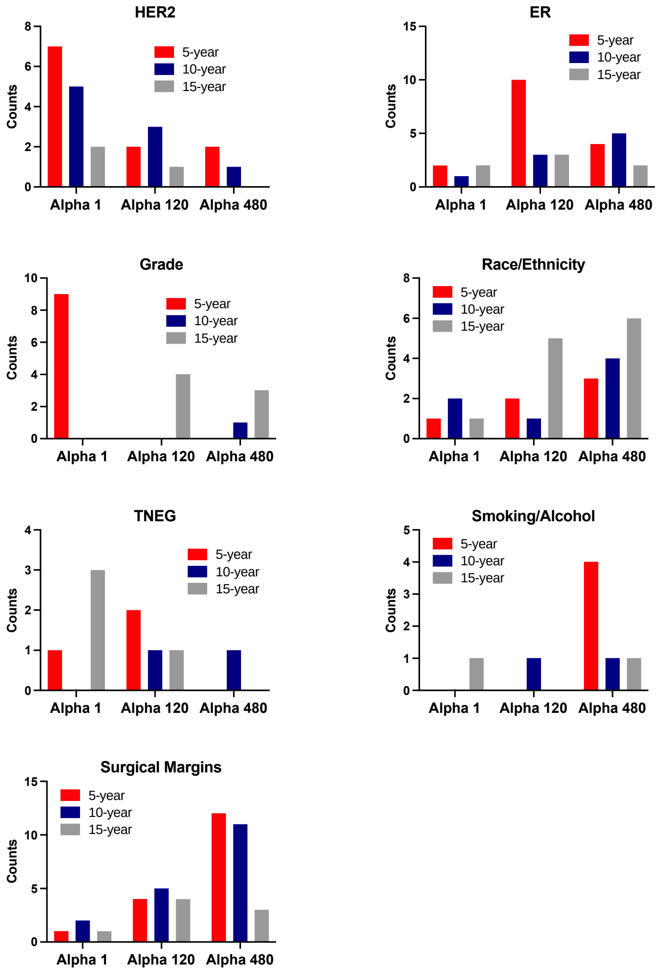

Figure 3. MBIL generated an output of clinical interactions predictive of 5-, 10- and 15-year breast cancer metastasis. HER2, ER, Grade, Race/ethnicity, TNEG, Smoking/Alcohol and Surgical Margins are represented. Each bar-plot indicates the number of counts in which each of these variables was identified as a predictor of metastasis at different values of alpha.

We next calculated the frequency in which these variables interact with any other variable to constitute direct risk factors of metastases (Tables 1-3). With an alpha 1, ER interacted with only n-TNM, HER2, and LN Positive 33% of the times each, suggesting that the presence of ER may only be a predictor of metastases when it is present in conjunction with these three variables (Table 1). ER was seen to have more interactions with an alpha of 120, but most frequently with n-TNM (33%), surgical margins (13%) and lymph node positive (13%) (Table 2). The variables that interacted most frequently with ER at an alpha of 480 were n-TNM (32%), surgical margins (16%) and race (16%) (Table 3).

With an alpha 1, TNEG was found to interact with LN positive/status (29%), HER2 (14%), Age at diagnosis (14%), ethnicity (14%), stage (14%) and re_excision (14%). The most frequent interaction was LN positive (28%), which related to 15-year recurrences (Table 1). With an alpha of 120, the most interactive variable of TNEG was n-TNM (50%), followed by surgical margins (25%) and smoking (12.5%) and invasive tumor location (12.5%) (Table 2). With an alpha of 480, TNEG was found to only interact with n-TNM (50%) and surgical margins (50%) to predict 10-year recurrence (Table 3).

With an alpha 1, HER2 was found to interact most frequently with stage (26%), MRIs 60 surgery (43%), ER percent (43%) and TNEG (43%) (Table 1). The influence of HER2 in predicting late recurrence was reduced with alpha values of 120 and 480. With an alpha of 120, HER2 interacted most frequently with stage (42%) and surgical margins (17%) (Table 2). With an alpha of 480, HER2 only interacted with surgical margins (57%), stage (29%) and t-TNM (14%) (Table 3).

Race and ethnicity, at an alpha of 1, interacted once with histology (12.5%), grade (12.5%), ER_percent (12.5%), n-TNM (12.5%), side (12.5%), LN positive/status (12.5%), TNEG (12.5%) and stage (12.5%) (Table 1). With an alpha of 120, interactions of race/ethnicity were more frequent with ER (18%), n-TNM (18%), stage (18%) and LN positive/status (18%) (Table 2). At an alpha of 480, race and ethnicity interacted more frequently with stage (27%), ER (17%) and n-TNM (17%) (Table 3).

Lastly, smoking and alcohol appeared to interact, at an alpha of 1, with LN positive/status to predict 15-year metastasis (Table 1). At an alpha of 120, smoking and alcohol interacted with TNEG (25%), n-TNM (25%), stage (25%) and histology (25%) (Table 2). Finally, at an alpha of 480, smoking and alcohol were found to interact most frequently with n-TNM (25%) and stage (25%), followed by t-TNM (16.7%), surgical margins (16.7%), ER (8.3%) and race (8.3%) (Table 3).

Table 1. Alpha 1-interacting variables. ER, TNEG, HER2, Race/Ethnicity and Alcohol/Smoking are represented with their interacting variables, times they interacted, years after diagnosis when these interactions predicted metastases, total number of times the variable interacts, and frequency of interaction.

Alpha 1					
Variable	**Interacts with**	**n times**	**Years after DG**	**total**	**%**
ER	n-TNM	2	5, 10	6	33.33%
ER	HER2	2	5, 15	6	33.33%
ER	LN Positive	2	15, 15	6	33.33%
TNEG	HER2	1	5	7	14.29%
TNEG	Age at DG	1	15	7	14.29%
TNEG	LN Positive/Status	2	15	7	28.57%
TNEG	Ethnicity	1	15	7	14.29%
TNEG	Stage	1	15	7	14.29%
TNEG	Re_excision	1	15	7	14.29%
HER2	Stage	6	5, 5, 10, 10, 10, 15	23	26.09%
HER2	MRIs_60_surgery	1	5	23	4.35%
HER2	ER_Percent	1	5	23	4.35%
HER2	TNEG	1	5	23	4.35%
HER2	Histology	3	5, 10, 10	23	13.04%
HER2	Grade	2	5, 5	23	8.70%
HER2	Invasive tumor location	2	5, 10	23	8.70%
HER2	ER	1	5, 15	23	4.35%
HER2	PR	2	5, 10	23	8.70%
HER2	LN Positive/Status	3	10, 10, 15	23	13.04%
HER2	Surgical Margins	1	10	23	4.35%
Race/Ethnicity	Histology	1	5	8	12.50%
Race/Ethnicity	Grade	1	5	8	12.50%
Race/Ethnicity	ER_Percent	1	10	8	12.50%
Race/Ethnicity	n-TNM	1	10	8	12.50%
Race/Ethnicity	Side	1	10	8	12.50%
Race/Ethnicity	LN Positive/Status	1	10	8	12.50%
Race/Ethnicity	TNEG	1	15	8	12.50%
Race/Ethnicity	Stage	1	15	8	12.50%
Alcohol/Smoking	LN Positive/Status	1	15	1	100.00%

Table 2. Alpha 120-interacting variables. ER, TNEG, HER2, Race/Ethnicity and Alcohol/Smoking are represented with their interacting variables, times they interacted, years after diagnosis when these interactions predicted metastases, total number of times the variable interacts, and frequency of interaction.

Alpha 120					
Variable	**Interacts with**	**n times**	**Years after DG**	**total**	**%**
ER	n-TNM	9	5, 5, 5, 5, 5, 5, 5, 10, 15	30	30.00%
ER	Surgical Margins	4	5, 5, 10, 15	30	13.33%

ER	Family History	2	5, 10	30	6.67%
ER	LN Positive/Status	4	5, 10, 15, 15	30	13.33%
ER	HER2	1	5	30	3.33%
ER	MRIs_60_surgery	1	5	30	3.33%
ER	Race/Ethnicity	3	5, 15, 15	30	10.00%
ER	Histology	1	5	30	3.33%
ER	Invasive tumor location	1	5	30	3.33%
ER	Size	1	5	30	3.33%
ER	Side	1	5	30	3.33%
ER	DCIS_level	2	5, 10	30	6.67%
TNEG	n-TNM	4	5, 5, 10, 15	8	50.00%
TNEG	Surgical Margins	2	10, 15	8	25.00%
TNEG	Smoking	1	5	8	12.50%
TNEG	Invasive tumor location	1	5	8	12.50%
HER2	ER	1	5	12	8.33%
HER2	n-TNM	1	5	12	8.33%
HER2	Stage	5	5, 10, 10, 10, 15	12	41.67%
HER2	Surgical Margins	2	5, 10	12	16.67%
HER2	Histology	1	10	12	8.33%
HER2	Grade	1	15	12	8.33%
HER2	PR	1	10	12	8.33%
Race/Ethnicity	ER	3	5, 15, 15	17	17.65%
Race/Ethnicity	n-TNM	3	5, 10, 15	17	17.65%
Race/Ethnicity	Stage	3	5, 15, 15	17	17.65%
Race/Ethnicity	Surgical Margins	1	5	17	5.88%
Race/Ethnicity	ER_Percent	1	10	17	5.88%
Race/Ethnicity	Grade	1	15	17	5.88%
Race/Ethnicity	LN Positive/Status	3	15	17	17.65%
Race/Ethnicity	Re_Excision	1	15	17	5.88%
Race/Ethnicity	PR_percent	1	15	17	5.88%
Smoking/Alcohol	TNEG	1	5	4	25.00%
Smoking/Alcohol	n-TNM	1	5	4	25.00%
Smoking/Alcohol	Stage	1	10	4	25.00%
Smoking/Alcohol	Histology	1	10	4	25.00%

Table 3. Alpha 480-interacting variables. ER, TNEG, HER2, Race/Ethnicity and Alcohol/Smoking are represented with their interacting variables, times they interacted, years after diagnosis when these interactions predicted metastases, total number of times the variable interacts, and frequency of interaction.

Alpha 480					
Variable	**Interacts with**	**n times**	**Years after DG**	**total**	**%**
ER	n-TNM	6	5, 5, 5, 10, 10, 10	19	31.58%
ER	Surgical Margins	3	5, 5, 10	19	15.79%
ER	Race	3	5, 10, 15	19	15.79%

ER	Size	1	5	19	5.26%
ER	Smoking	1	5	19	5.26%
ER	Family History	1	10	19	5.26%
ER	LN Positive/Status	1	10	19	5.26%
ER	Stage	1	10	19	5.26%
ER	DCIS_level	1	10	19	5.26%
ER	Age at DG	1	10	19	5.26%
TNEG	n-TNM	1	10	2	50.00%
TNEG	Surgical Margins	1	10	2	50.00%
HER2	Stage	2	5, 10	7	28.57%
HER2	Surgical Margins	4	5, 5, 10	7	57.14%
HER2	t-TNM	1	5	7	14.29%
Race/Ethnicity	Stage	6	5, 15, 15, 15, 15, 15	22	27.27%
Race/Ethnicity	Surgical Margins	2	5, 5	22	9.09%
Race/Ethnicity	ER	3	5, 10, 15	22	13.64%
Race/Ethnicity	n-TNM	3	5, 10, 10	22	13.64%
Race/Ethnicity	Family History	1	10	22	4.55%
Race/Ethnicity	LN Positive/Status	1	10	22	4.55%
Race/Ethnicity	ER_Percent	1	10	22	4.55%
Race/Ethnicity	Grade	1	15	22	4.55%
Race/Ethnicity	Invasive tumor location	1	15	22	4.55%
Race/Ethnicity	Re-excision	1	15	22	4.55%
Race/Ethnicity	Alcohol	1	15	22	4.55%
Race/Ethnicity	histology2	1	15	22	4.55%
Smoking/Alcohol	t-TNM	2	5, 5	12	16.67%
Smoking/Alcohol	n-TNM	3	5, 5, 5	12	25.00%
Smoking/Alcohol	Stage	3	5, 10, 15	12	25.00%
Smoking/Alcohol	Surgical Margins	2	5, 10	12	16.67%
Smoking/Alcohol	ER	1	5	12	8.33%
Smoking/Alcohol	Race	1	15	12	8.33%

4. Discussion

Advances in treatment of breast after surgical resection of the primary lesion has altered our approach to those women who are at risk for recurrences. As we now can see mBC two to three decades after the primary lesion was removed without evidence of dissemination [17], there is a need to personalize follow up care based on the likelihood of finding cancer recurrence. For this reason, much effort has gone into determining which clinic-pathological features can predict longer term outcomes. Classical methods have found a number of characteristics that predict recurrence, but mainly these are singular parameters best at alerting the oncologist to rapid recurrence, usually in the setting of HER2-positivity or TNBC. To broaden the coverage provided by transcriptomic predictors of therapy response, we used machine learning to find not just independent factors for later recurrences, but also sets of risk factors that in aggregate would be prognostic.

The MBIL algorithm escalates the degree of interactivity between parameters. With the alpha set at the bottom level of 1, known predictors of recurrence were found including TNEG, HER2 positivity, and TNM stage. However, as the alpha was elevated to define interacting sets of parameters, race, age, the alcohol and smoking were scored as part of the prognostic sets. Interestingly, alcohol and smoking were more often linked to recurrences at 5 years than at 15 years; this suggests that pathobiologic effects are either short-term or reversible on the scale of years to a decade [18,19]. Race and ethnicity have also been implicated in higher recurrence risks, as shown by the 21-gene recurrence score [20]. In our study, race and ethnicity were more often linked to later recurrences, providing for partial personalization of follow-up as these are non-modifiable parameters.

These findings need to be validated in additional cancer databases and with other machine-learning methods. This is particularly true of some confusing denotations. For instance, TNEG was related to later recurrences, which goes against the well documented clinical course of TNBC recurring usually within three years, and if not by five years, with the disease considered cured. However, this prognostic found herein may simply reflect a statistical fluke that absence of TNEG means less likelihood of rapid recurrence and therefore any recurrence that happens is more likely to occur after 10 or 15 years. This and other prognostic situations need to be refined in further studies. Still, the work herein does point to the value of these machine-learning algorithms in discerning prognostic sets at a level of resolution (as to years out from primary cancer diagnosis) than classical methods of biomarker development.

5. Conclusions

MBIL may guide on the identification of direct causal sets and interactive risk factors of late breast cancer recurrence. Application of this and similar machine-learning methods are encouraged in further databases to help interpret risk of late mBC.

6. Patents

Not applicable.

Supplementary Materials: There are no supplementary materials

Author Contributions: Conceptualization: AW, AB, and XJ; methodology, XJ; resources: XJ; data curation: JLGM; Data analyses: JLGM; manuscript writing: JLGM, AW, and XJ; review and editing: JLGM, AW, XJ, AB; supervision: AW, AB.

Funding: This study was supported by the U.S. Department of Defense through the Breast Cancer Research Program under Award No. W81XWH1910495 (to X.J). Other than supplying funds, the funding agencies played no role in the research.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data used in this study are available at datadryad.org [10].

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare having no conflicts of interest.

References

1. Sopik, V.; Sun, P.; Narod, S.A. Predictors of time to death after distant recurrence in breast cancer patients. *Breast Cancer Res Treat* **2019**, *173*, 465-474, doi:10.1007/s10549-018-5002-9.
2. Sestak, I.; Cuzick, J. Markers for the identification of late breast cancer recurrence. *Breast Cancer Res* **2015**, *17*, 10, doi:10.1186/s13058-015-0516-0.
3. Davies, C.; Pan, H.; Godwin, J.; Gray, R.; Arriagada, R.; Raina, V.; Abraham, M.; Medeiros Alencar, V.H.; Badran, A.; Bonfill, X.; et al. Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years after diagnosis of oestrogen receptor-positive breast cancer: ATLAS, a randomised trial. *Lancet* **2013**, *381*, 805-816, doi:10.1016/S0140-6736(12)61963-1.
4. Pan, H.; Gray, R.; Braybrooke, J.; Davies, C.; Taylor, C.; McGale, P.; Peto, R.; Pritchard, K.I.; Bergh, J.; Dowsett, M.; et al. 20-Year Risks of Breast-Cancer Recurrence after Stopping Endocrine Therapy at 5 Years. *N Engl J Med* **2017**, *377*, 1836-1846, doi:10.1056/NEJMoa1701830.
5. Bhutiani, N.; Egger, M.E.; Ajkay, N.; Scoggins, C.R.; Martin, R.C., 2nd; McMasters, K.M. Multigene Signature Panels and Breast Cancer Therapy: Patterns of Use and Impact on Clinical Decision Making. *J Am Coll Surg* **2018**, *226*, 406-412 e401, doi:10.1016/j.jamcollsurg.2017.12.043.
6. Markopoulos, C.; Hyams, D.M.; Gomez, H.L.; Harries, M.; Nakamura, S.; Traina, T.; Katz, A. Multigene assays in early breast cancer: Insights from recent phase 3 studies. *Eur J Surg Oncol* **2020**, *46*, 656-666, doi:10.1016/j.ejso.2019.10.019.
7. Jiang, X.; Wells, A.; Brufsky, A.; Shetty, D.; Shajihan, K.; Neapolitan, R.E. Leveraging Bayesian networks and information theory to learn risk factors for breast cancer metastasis. *BMC Bioinformatics* **2020**, *21*, 298, doi:10.1186/s12859-020-03638-8.
8. Jiang, X.; Jao, J.; Neapolitan, R. Learning Predictive Interactions Using Information Gain and Bayesian Network Scoring. *PLoS One* **2015**, *10*, e0143247, doi:10.1371/journal.pone.0143247.
9. Zeng, Z.; Jiang, X.; Neapolitan, R. Discovering causal interactions using Bayesian network scoring and information gain. *BMC Bioinformatics* **2016**, *17*, 221, doi:10.1186/s12859-016-1084-8.
10. Jiang, X.; Wells, A.; Brufsky, A.; Neapolitan, R. A clinical decision support system learned from data to personalize treatment recommendations towards preventing breast cancer metastasis. *PLoS One* **2019**, *14*, e0213292, doi:10.1371/journal.pone.0213292.
11. Neapolitan, R.E. *Probabilistic reasoning in expert systems : theory and algorithms*; Wiley: New York, 1990; pp. xiii, 433 p.
12. Neapolitan, R.E.; Jiang, X. *Contemporary artificial intelligence*, 1st edition ed.; p. 1 online resource (508 pages).
13. Neapolitan, R.E. *Learning Bayesian networks*; Pearson Prentice Hall: Upper Saddle River, NJ, 2004; pp. xv, 674 p.
14. Cooper, G.F.; Herskovits, E. A Bayesian method for the induction of probabilistic networks from data. *Machine learning* **1992**, *9*, 309-347.
15. Heckerman, D.; Geiger, D.; Chickering, D.M. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning* **1995**, *20*, 197-243, doi:10.1023/A:1022623210503.
16. O'Brien, K.M.; Cole, S.R.; Tse, C.K.; Perou, C.M.; Carey, L.A.; Foulkes, W.D.; Dressler, L.G.; Geradts, J.; Millikan, R.C. Intrinsic breast tumor subtypes, race, and long-term survival in the Carolina Breast Cancer Study. *Clin Cancer Res* **2010**, *16*, 6100-6110, doi:10.1158/1078-0432.CCR-10-1533.
17. Ekholm, M.; Bendahl, P.O.; Ferno, M.; Nordenskjold, B.; Stal, O.; Ryden, L.; South, S.; South-East Swedish Breast Cancer, G. Effects of adjuvant tamoxifen over three decades on breast cancer-free and distant recurrence-free interval among premenopausal women with oestrogen receptor-positive breast cancer randomised in the Swedish SBII:2pre trial. *Eur J Cancer* **2019**, *110*, 53-61, doi:10.1016/j.ejca.2018.12.034.

18. Heitz, A.E.; Baumgartner, R.N.; Baumgartner, K.B.; Boone, S.D. Healthy lifestyle impact on breast cancer-specific and all-cause mortality. *Breast Cancer Res Treat* **2018**, *167*, 171-181, doi:10.1007/s10549-017-4467-2.
19. Kwan, M.L.; Kushi, L.H.; Weltzien, E.; Tam, E.K.; Castillo, A.; Sweeney, C.; Caan, B.J. Alcohol consumption and breast cancer recurrence and survival among women with early-stage breast cancer: the life after cancer epidemiology study. *J Clin Oncol* **2010**, *28*, 4410-4416, doi:10.1200/JCO.2010.29.2730.
20. Hoskins, K.F.; Danciu, O.C.; Ko, N.Y.; Calip, G.S. Association of Race/Ethnicity and the 21-Gene Recurrence Score With Breast Cancer-Specific Mortality Among US Women. *JAMA Oncol* **2021**, *7*, 370-378, doi:10.1001/jamaoncol.2020.7320.