Comparison of Logistic Regression and Bayesian Networks for Risk Prediction of Breast Cancer Recurrence

Witteveen, Annemieke; Nane, Gabriela F.; Vliegen, Ingrid M.H.; Siesling, Sabine; IJzerman, Maarten J.

DOI
10.1177/0272989X18790963

Publication date
2018

Document Version
Final published version

Published in
Medical Decision Making: an international journal

Citation (APA)
Witteveen, A., Nane, G. F., Vliegen, I. M. H., Siesling, S., & IJzerman, M. J. (2018). Comparison of Logistic Regression and Bayesian Networks for Risk Prediction of Breast Cancer Recurrence. *Medical Decision Making: an international journal*, *38*(7), 822-833. https://doi.org/10.1177/0272989X18790963

Original Article

MDM Medical Decision Making

Medical Decision Making
2018, Vol. 38(7) 822–833

DOI: 10.1177/0272989X18790963
journals.sagepub.com/home/mdm
SAGE

Comparison of Logistic Regression and Bayesian Networks for Risk Prediction of Breast Cancer Recurrence

Annemieke Witteveen, Gabriela F. Nane, Ingrid M.H. Vliegen, Sabine Siesling, and Maarten J. IJzerman

Purpose. For individualized follow-up, accurate prediction of locoregional recurrence (LRR) and second primary (SP) breast cancer risk is required. Current prediction models employ regression, but with large data sets, machine-learning techniques such as Bayesian Networks (BNs) may be better alternatives. In this study, logistic regression was compared with different BNs, built with network classifiers and constraint- and score-based algorithms. **Methods**. Women diagnosed with early breast cancer between 2003 and 2006 were selected from the Netherlands Cancer Registry (NCR) (N = 37,320). BN structures were developed using 1) Bayesian network classifiers, 2) correlation coefficients with different cutoffs, 3) constraint-based learning algorithms, and 4) score-based learning algorithms. The different models were compared with logistic regression using the area under the receiver operating characteristic curve, an external validation set obtained from the NCR from 2007 and 2008 (N = 12,308), and subgroup analyses for a high- and low-risk group. **Results**. The BNs with the most links showed the best performance in both LRR and SP prediction (c-statistic of 0.76 for LRR and 0.69 for SP). In the external validation, logistic regression generally outperformed the BNs in both SP and LRR (c-statistic of 0.71 for LRR and 0.64 for SP). The differences were nonetheless small. Although logistic regression performed best on most parts of the subgroup analysis, BNs outperformed regression with respect to average risk for SP prediction in low- and high-risk groups. **Conclusions**. Although estimates of regression coefficients depend on other independent variables, there is no assumed dependence relationship between coefficient estimators and the change in value of other variables as in the case of BNs. Nonetheless, this analysis suggests that regression is still more accurate or at least as accurate as BNs for risk estimation for both LRRs and SP tumors.

Keywords
Bayesian network, breast cancer, locoregional recurrence, logistic regression, machine learning, risk prediction, second primary

Date received: April 19, 2017; accepted: June 11, 2018

Risk prediction models can be used to support clinical decisions for various conditions. Although many prediction models are developed and available, the uptake in clinical practice is slow. Two important challenges associated with conventional yet most popular (regression) prediction models are the difficulty to incorporate dependencies among all variables and the presence of numerous risk factors with only a small effect.[1]

Department of Health Technology and Services Research (HTSR), Technical Medical Centre, University of Twente, Enschede, the Netherlands (AW, SS, MJIJ); Delft Institute of Applied Mathematics (DIAM), Delft University of Technology, Delft, the Netherlands (GFN); Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Eindhoven, the Netherlands (IMHV); and Department of Research, Netherlands Comprehensive Cancer Organisation (IKNL), Utrecht, the Netherlands (SS). Presented at the 38th Annual North American Meeting, Vancouver, Canada (Session 3I: Oral Abstracts: Improving Modeling Research).

Corresponding Author
Annemieke Witteveen, Department of Health Technology and Services Research, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands (a.witteveen@utwente.nl).

These challenges are addressed by Bayesian networks (BNs), also known as Bayes nets or probabilistic causal networks. BNs are flexible probabilistic graphical models that capture the dependence relationships between selected variables. The variables are represented with nodes and can be continuous, as well as discrete. In case of discrete variables, the nodes are connected with links to present the dependence relations, and for each discrete node, a probabilistic table provides the probability of the possible values, conditional on the nodes that influence this node. Advantages of using BNs are the ease of interpretation due to the graphical representation, simple validation, the possibility to include prior information, their flexibility of including both observational and causal inference, flexibility in outcome parameter within the model, and how they handle missing data.[2–5]

A very high number of BNs can be modeled given a set of variables, and machine-learning methods have been successfully employed to learn the structure of BNs in an automated fashion. As Bouhamed et al.[6] state, "Currently, Bayesian Networks have become one of the most complete, self-sustained and coherent formalisms used for knowledge acquisition, representation and application through computer systems." Machine learning is a collection of methods for systems that can learn and automatically improve with experience.[7] In the past decades, there has been increasing interest in the application of machine learning, mainly because of the availability of the required computational power and the emergence of big data. Machine learning can be subdivided in supervised learning (for classification), unsupervised learning (for clustering), and reinforcement learning (for decision making). If the aim is to predict an outcome measure based on several input variables, supervised learning is used.[2] An example of such supervised learning for classification are BNs.

There are a few reasons why BNs may perform better than standard regression. Ng and Jordan[8] made a theoretical and empirical comparison between a naive BN and a logistic regression model. With naive BNs, the "naive" assumption of conditional independence between variables is made. This assumption is often violated, but the algorithm can still perform well.[9] In comparison with logistic regression, the BN had a higher asymptotic error, but the naive BN converged faster to approach its higher error.[8] This means that with an infinite training data set, logistic regression is expected to outperform naive BNs as it has a lower error. But with limited data, BNs can outperform regression as it needs less data to reach its best performance. And if the naive conditional independence assumption does not hold, the error could be lower than with logistic regression, even with more data. Although the value of the coefficients included in logistic regression is conditional on the other variables that are included, there is no dependence relationship between the values of the coefficients and the *change* in value of one of the influencing variables, as is the case with BNs. Also, if the number of events is very low, there is a risk of overfitting when using regression.[10]

Most models for cancer risk prediction are based on regression.[11,12] With an accurate insight in the risk of breast cancer recurrence, patients with a high risk can be identified who might benefit from a more intensive follow-up after breast cancer and to aid clinical decision making. Recently, our study group developed a nomogram based on logistic regression to give insight in the time-dependent risk of locoregional recurrence (LRR) of breast cancer.[13] The model satisfied the employed assumptions and showed good discriminative ability in external validation. Besides early detection of LRRs, the aim of clinical follow-up after curative treatment of breast cancer is also the detection of asymptomatic second primary (SP) breast cancer.[14] SP breast cancer is defined as a new manifestation of breast cancer in the contralateral breast.[15] As SPs are a separate entity from the primary tumor,[16] they are hard to predict using clinical data that contain mostly information about the primary tumor. Since BNs also take into account the dependence relationships between the influencing factors, it may result in better estimates of the risk. Furthermore, as it is of interest to predict the risk for a new patient, given what we know of previous patients, it may be more appropriate to formulate the problem within the Bayesian paradigm. In this study, we developed different BNs and assessed whether they outperformed logistic regression with regard to the prediction of LRR or SP breast cancer at the patient level using a large population-based data set with clinical risk factors.

Methods

Study Population

Patients were selected from the Netherlands Cancer Registry (NCR), a nationwide population-based registry, which has registered almost all newly diagnosed tumors since 1989. The information on patient, tumor, and treatment characteristics, as well as data concerning recurrences within the first 5 years following primary breast cancer, was recorded directly from the patient files by specially trained registration clerks.

Women who had primary invasive breast cancer without distant metastasis (DM) or previous or synchronous

Table 1 Overview of Constraint- and Score-Based Algorithms That Were Used, with Their Corresponding Tests and Scores

Constraint-based algorithms	**Tests**			
Grow-shrink (GS) Incremental association (IA) Fast incremental association (fIA) Interleaved incremental association (iIA)	Mutual information	Shrinkage estimator for the mutual information	Pearson's χ^2	
Score-based algorithms	**Scores**			
Hill climbing (HC) Tabu search (TS)	Bayesian information criterion (BIC)	Log-likelihood	Akaike information criterion (AIC)	Bayesian Dirichlet equivalent	Modified Bayesian Dirichlet equivalent	K2

tumors (diagnosed within 3 months after the first tumor[17]), were diagnosed between 2003 and 2006, and were treated with curative intent were selected from the registry as the training or index cohort (N = 37,230). Curative intent was defined as surgical removal of the primary tumor without macroscopic residual disease. Adjuvant treatment should have been received in case of microscopic residue. Of the included patients, 205 (0.6%) had incomplete 5-year follow-up; they were censored in the logistic regression models and treated as event free in the BNs. In the first 5 years following primary breast cancer treatment, 926 of the selected patients developed a LRR (2.5%) and 896 a SP tumor (2.4%) as a first event. Patient, tumor, and treatment characteristics can be found in Supplemental Table S1. For external validation, data from a selection of Dutch hospitals from 2007 to 2008 were used as validation cohort (43 of 91 hospitals, N = 12,308). Recurrence rates were slightly lower in the validation cohort: 2.2% developed a LRR and 2.3% a SP tumor.

Logistic Regression Model

The details of our logistic regression model for LRR have been reported elsewhere[13] but are summarized here for the convenience of the reader. Variables with an expected influence on LRR and SP risk were selected using the literature and availability in the NCR. Because of the nonlinear effect of age on risk, age was discretized into 4 groups. The other factors were already categorical. As missing values were believed to be random, they were imputed using chained equations.[18,19] A first logistic regression model was made including all the variables of interest. A second model only included variables with an effect of at least 10% (based on the odds ratios [ORs]).

Bayesian Networks

A network structure was defined for the BNs with all the variables represented as nodes. The structures were determined in 2 ways. The first method was with a specific focus on the outcome of interest (Bayesian network classifiers, correlation coefficients), while the second (constraint- and score-based learning algorithms) was data driven, without a focus on the outcome. For the Bayesian network classifiers, a naive network, assuming all variables are only connected to the variable of interest, was created, as well as a tree-augmented naive (TAN) network, which built on the naive network using minimal description length scoring.[20] The structures, based on Spearman's rank correlation, also started with a naive network, and with the cutoffs 0.3 (moderate to high) and 0.1 (low to high)[21] for the rank correlation coefficients, links were added to gain more insight into the difference in performance of different levels of correlation.

Constraint-based algorithms test the conditional independence to find the direct connections of a node and their direct connections (Markov blanket). With score-based algorithms, goodness-of-fit scores are used for optimization.[22] The constraint- and score-based algorithms we used and their corresponding tests and scores can be found in Table 1. For more information on the specific algorithms, tests, and scores, the reader is referred to the study by Scutari.[23] For all methods, we represented the joint probability distributions for both LRR and SP tumor as outcome variables using conditional probability tables (CPTs), which were learned via maximum likelihood estimation by assuming uniform Dirichlet prior distributions over all variables. Cases with missing values were not excluded but included with the information of the variables that were not missing.

Table 2 Example Risk Groups Based on 3 Risk Factors

	Age, y	Grade	Endocrine Therapy	No. (%) of Patients
Low risk	50-60	I	Yes	864 (1.7)
High risk	<50	II	No	2098 (4.2)

Comparison of the Models

We assessed several aspects of the validity and performance of the models: 1) the ability to distinguish between high- and low-risk patients (discrimination), 2) the agreement between observed and predicted risks (calibration), and 3) the performance in an external data set (generalizability). Besides the overall performance, we also assessed the performance of BNs and logistic regression to estimate recurrence risk in much smaller subgroups (an example high- and low-risk group) as we are interested in making more individualized risk predictions. The groups were based on age, primary tumor grade, and treatment with endocrine therapy, as they are established risk factors for LRR,[13] and age and endocrine treatment also for SP.[24,25] For low risk, we used patients aged 50 to 60 years with grade I primary tumors who received endocrine therapy, and for high risk, we used patients aged <50 years with grade II primary tumors without endocrine therapy (Table 2).

The discrimination of the different models was compared by using the Harrell c-statistic for area under the receiver operating characteristic (ROC) curve. A c-statistic of 1.0 indicates a perfect predictive ability, whereas 0.5 represents no predictive discrimination. As an example, we chose a high-risk profile for a patient aged <50 years, with a primary tumor size 2 to 5 cm, >3 positive nodes, grade III, ductal morphology, positive hormone status, no multifocality, mastectomy, with axillary lymph node dissection without radiation therapy, and with chemotherapy and endocrine therapy. Information on risk factors was added in this order, and the risks were plotted against the actual events from matching patients in the data set. The differences in observed and predicted probabilities were quantified with the Brier score, which captures both calibration and discrimination.[26]

For calibration, the error rate was determined by comparing the actual events with the predicted events. As a permutation test to look at the performance in a random data set, 10 data sets with randomly assigned labels for the outcome variable were made and the results were pooled. The estimates from the models were compared with the percentage of events in the patients with the corresponding characteristics. To make ranges around the estimates, the risk for patients with the best and worst possible characteristics in the risk groups was determined. For the performance in the subgroups, the c-statistic was estimated. To see if the patients who had the corresponding characteristics from the risk groups and were diagnosed with an event were in fact assigned as high risk, the average assigned risks by the models were compared. For checking the generalizability, an external validation using the validation cohort was performed. The regression analyses were performed using STATA 14.0 (StataCorp, College Station, TX), and the Netica software package from Norsys (Vancouver, BC, Canada) was used for the BNs.

Results

The patients in the index and validation cohort had small differences for the included variables of age, grade, size, lymph node status, hormone status, and treatments (all <3%). The variables included as influencing factors in the original regression models were age, primary tumor size, involved lymph nodes, grade of differentiation, hormone status, multifocality, and whether or not patients were treated with radiation, chemotherapy, or endocrine therapy. When only selecting variables with an OR of at least 1.1, no variables were omitted in the LRR model and 3 in the SP model (hormone status, axillary lymph node dissection, and grade of differentiation). Healthy convergence was achieved with the multiple imputations. From the correlation (Suppl. Table S2), 14 links were identified with a coefficient $>|0.3|$ and 41 with a coefficient $>|0.1|$, which were added to the naive BNs. The number of links to LRR or SP and the total number of links for each network can be found in Table 3.

Comparison of the Models

The performance of the best-performing score- and constraint-based algorithms is summarized in Table 3. There was a clear association between the number of links and discriminative performance for the constraint- and score-based algorithms in the index cohort; more

Table 3 Characteristics and Performance of the Best Performing Models per Method[a]

		No. of Links		Index Cohort		Validation Cohort (External Validation)		Permutation Test	
		To LRR	In Total	C-statistic	Error Rate, %	C-statistic	Error Rate, %	C-statistic[b]	Error Rate, %[b]
LRR									
Logistic regression	All variables	NA	NA	0.71	2.5	**0.71**	2.6	0.55	3.0
BN									
Network classifiers	Naive	12	12	0.67	3.0	0.67	2.6	0.50	3.0
	TAN	12	22	0.69	3.0	0.67	2.6	0.50	3.0
Correlation	Coefficient ≥0.3	12	26	0.69	3.0	0.66	2.6	0.50	3.1
	Coefficient ≥0.1	12	53	0.73	3.0	0.64	2.7	0.51	3.1
Constrained based	iIA (χ^2)	3	31	0.61	3.0	0.60	2.6	0.50	3.0
Score based	HC/TS[c] (log likelihood)	12	78	**0.76**	2.9	0.62	2.6	0.50	3.1
	HC/TS[c] (AIC)	4	43	0.68	3.0	0.69	2.6	0.50	3.0
SP									
Logistic regression	All variables	NA	NA	0.66	2.8	0.63	2.6	0.52	3.0
	Selection	NA	NA	0.64	2.8	**0.64**	2.6	0.52	3.0
BN									
Network classifiers	Naive	12	12	0.61	2.8	**0.64**	2.6	0.49	2.9
	TAN	12	22	0.62	2.8	0.63	2.6	0.50	2.9
Correlation	Coefficient ≥0.3	12	26	0.61	2.8	0.61	2.6	0.50	2.9
	Coefficient ≥0.1	12	53	0.66	2.8	0.61	2.7	0.49	2.9
Constrained based	iIA (χ^2/mutual information)[c]	1	30/31	0.59	2.8	0.53	2.6	0.50	2.9
Score based	HC/TS[c] (log likelihood)	12	78	**0.69**	2.8	0.57	2.7	0.49	3.1
	HC/TS[c] (BIC/BDE/K2)[c]	1/2/1	31/37/37	0.59	2.8	0.62	2.6	0.50	2.9

AIC, Akaike information criterion; BDE, Bayesian Dirichlet equivalent; BIC, Bayesian information criterion; BN, Bayesian network; HC, Hill climbing; iIA, interleaved incremental association; LRR, locoregional recurrence; NA, not applicable; SP, second primary; TAN, tree-augmented naive; TS, Tabu search.

[a]Bold indicates the best estimate.

[b]Pooled results.

[c]Similar performance.

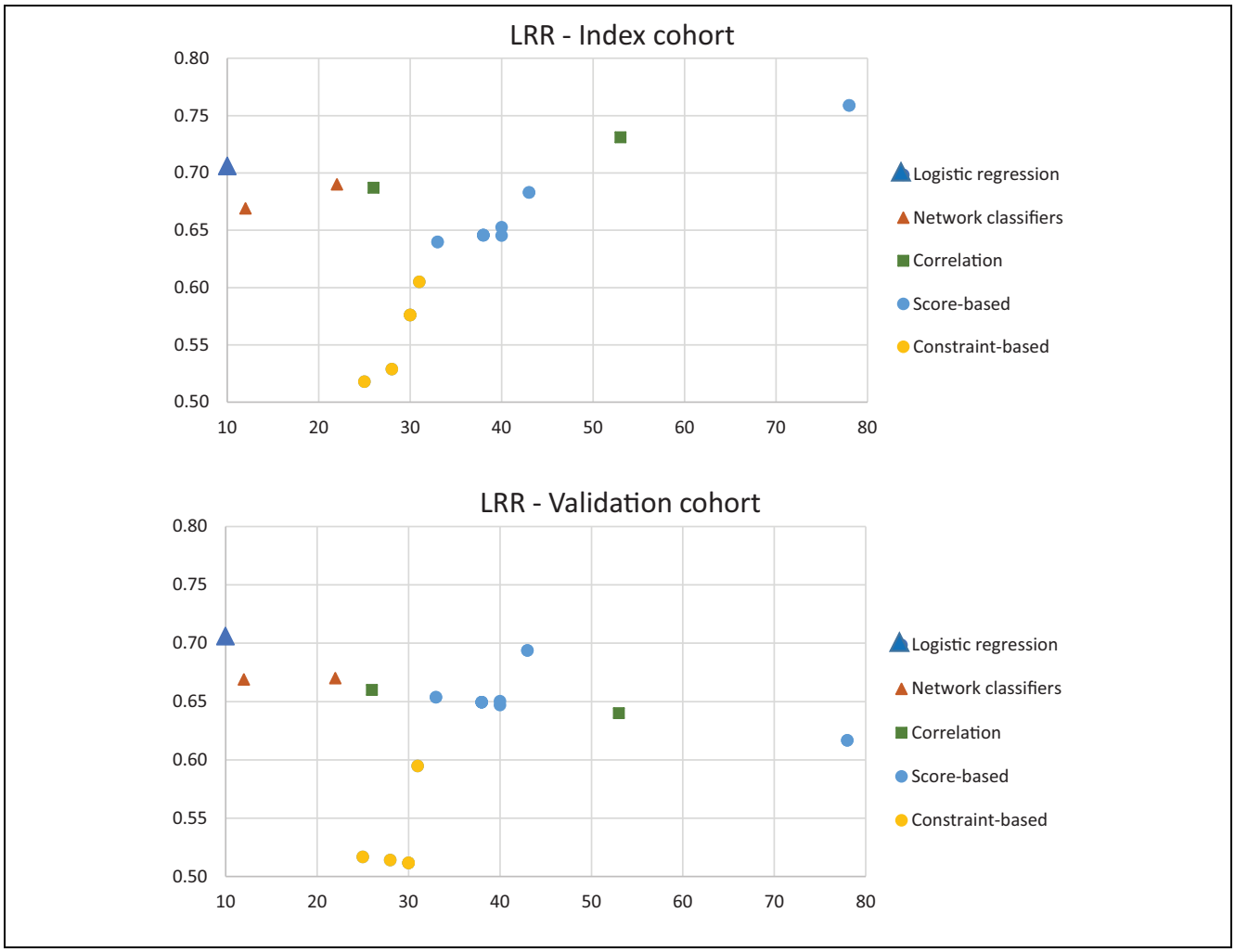

Figure 1 Performance of the models for LRR in (A) the index cohort (2003–2006) and (B) the validation cohort (2007–2008). LRR, locoregional recurrence; ROC, receiver operating characteristic.

links resulted in a higher c-statistic and therefore higher prognostic validity (Figure 1A). As a consequence, the constraint-based BNs were outperformed by all score-based algorithms, as they consisted of more links. The Tabu search (TS) and Hill-climbing (HC) algorithms with log-likelihood score contained all possible links and had a c-statistic of 0.76 for LRR and 0.69 for SP. Logistic regression scored lower with 0.71 and 0.65 for LRR and SP, respectively.

In contrast to the performance of the BNs in the index cohort, the number of links in the BNs was not related to the performance in the validation cohort (Figure 1B). Logistic regression outperformed the BNs, with c-statistics of 0.71 and 0.64 for LRR and SP, respectively, compared to 0.69 for LRR (TS/HC with AIC score) and 0.62 for SP (TS/HC with Bayesian information criterion [BIC], Bayesian Dirichlet equivalent [BDE], or K2 score). A notable exception was the naive network for SP, for which the c-statistic was equal to logistic regression (c-statistic of 0.64). Error rates did not differ much between the models (all 2.5%–3.0%). Despite the overall improved performance of logistic regression, note the small differences with BNs.

Subgroup Analysis

The predictions of the risks for the high- and low-risk groups can be found in Table 4. For LRR risk, estimates from BNs (TAN for low risk and TS for high risk) were closer to the actual percentage of LRRs in the data. Note

Table 4 Performance for the Low- and High-Risk Groups in the Index Cohort[a]

	Risk Group	Logistic Regression: Imputed Data	BN				% in Data
			Network Classifier: TAN	Correlation: Cutoff 0.1	Constraint-Based: iIA (χ^2)	Score-Based: TS (Log Likelihood)	
Risk estimate (range), %							
LRR	Low	0.70 (0.1–12.0)	**0.90** (0.4–96.3)	1.60 (0.1–99.1)	3.30 (1.3–49.1)	2.80 (0.1–47.8)	1.00
	High	4.40 (0.7–49.7)	4.70 (0.7–95.9)	5.40 (0.2–84.4)	3.60 (1.2–49.7)	**5.90** (0.2–65.1)	6.00
SP	Low	**1.50** (0.6–4.1)	2.10 (1.2–22.9)	2.40 (0.3–93.0)	2.60 (1.1–22.6)	9.50 (0.1–57.1)	1.50
	High	**4.20** (1.5–10.2)	3.80 (1.9–16.9)	4.00 (0.9–79.6)	3.30 (1.5–19.3)	6.90 (0.7–52.6)	4.20
C-statistic							
LRR	Low	0.59	**0.62**	0.59	0.52	0.59	
	High	**0.62**	0.61	0.57	0.49	0.61	
SP	Low	**0.94**	0.57	0.57	0.59	0.55	
	High	**0.59**	0.58	0.5	0.45	0.5	
Average risk in cases (≥5%), %							
LRR	Low	**23.3 (100)**	1.4 (0)	4.1 (33)	10.6 (33)	11.3 (33)	
	High	**13.6 (57)**	5.7 (41)	7.7 (64)	3.5 (9)	9.0 (55)	
SP	Low	10.8 (70)	3.2 (0)	16.8 (70)	3.7 (20)	**70.8 (100)**	
	High	4.9 (21)	5.3 (69)	4.2 (19)	2.6 (0)	**14.6 (50)**	

BN, Bayesian network; iIA, interleaved incremental association; LRR, locoregional recurrence; SP, second primary; TAN, tree-augmented naive; TS, Tabu search.
a. Bold indicates the best estimate.

the very wide ranges for BN risk intervals. The discrimination was poor (c-statistic 0.49 for interleaved incremental association [iIA] BN) to moderate (0.62 with logistic regression and TAN BN). For the prediction of SP, logistic regression performed very well in the risk subgroups, with spot-on estimates and a c-statistic of 0.94 in the low-risk group, whereas the BNs all overestimated the risk in the low-risk group and again showed wide ranges in estimates. When comparing the risks that were assigned to LRR cases (women actually diagnosed with recurrence), logistic regression assigned on average the highest risk for LRR in both the low- and high-risk groups (Table 4). For SP cases, the HC/TS BN assigned higher risk. In the validation cohort, the BNs still provided higher risk in cases than the logistic regression model, but the differences were much smaller (4.1% in the low-risk group and 2.8% in the high-risk group). If a threshold of 5% was used to define high risk, 100% of the low-risk cases and 57% of the high-risk cases of LRR would have been identified with logistic regression compared to 33% and 55% for the HC/TS BN.

The change in predicted risk for an example risk profile was assessed for logistic regression and the BNs by adding risk factors one by one and comparing with the events in the data set. For most parts, the predictions from logistic regression followed the true values more closely, as can also be seen from the lower Brier score (Figure 2).

Discussion

In this study, logistic regression estimates were compared with estimates obtained from BNs for the prediction of both LRR and SP breast cancer risk. BN structures were developed using constraint- and score-based learning algorithms and Bayesian network classifiers. Although the score-based algorithms showed the highest performance with the index cohort data, in the external validation, logistic regression outperformed the BNs for both LRR and SP risk prediction.

As the c-statistic is an average performance measure across all possible cutoffs and may not accurately represent the predictions at the individual level, it is hard to draw firm conclusions based on the improvement in the c-statistic alone.[27] Consequently, we also reviewed the error rate and change in risk by adding information on risk factors and a subgroup analysis in an example high- and low-risk group. Error rates were slightly lower for the external validation, most likely because the event rates were even lower in the validation cohort: with a lower number of events, less patients are incorrectly specified as not getting a recurrence. With the exception of

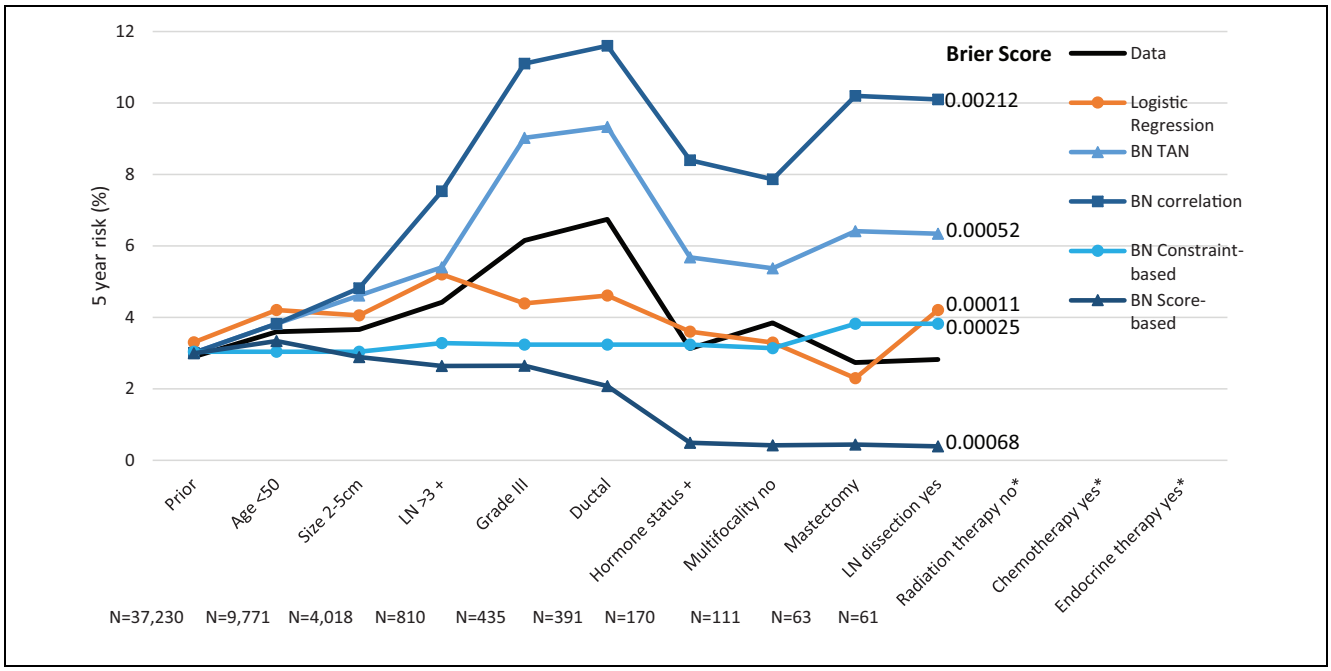

Figure 2 Change in risk by added information on risk factors for an example high-risk profile. BN, Bayesian network; LN, lymph node; TAN, tree-augmented naive. *No cases left to compare with that correspond to patient profile.

the performance in the risk groups, the overall performance of the predictions for SP was slightly worse than for LRR, as there are less influencing factors to take into account.

While we are interested in more individual risk estimates and BNs were shown to perform well in smaller data sets in the literature, we assessed the performance of the models in a subgroup analysis with 2 risk groups. Even though the number of patients in the subgroups was relatively low compared to the index cohort, BNs just take into account the data that are available per patient and require no minimum sample size.[28] However, these results need to be interpreted with caution, since the risk groups were chosen as an example and results for different subgroups might differ. The subgroup analysis showed good performance for logistic regression in SP risk prediction. The performance of the logistic regression model was not as good for the LRR prediction. Nonetheless, the results show an overall better performance than the prediction with the best-performing BNs algorithms, which also showed huge ranges. In the low-risk group, significantly higher risk was assigned to cases of SP by the score-based HC/TS algorithm. However, the difference was smaller in the validation data. Moreover, as the c-statistic of this BN showed no discriminatory accuracy, it means that noncases were also (needlessly) assigned with high risks. Not all the cases have a high risk, as there are more people who have a low risk. This is described as the prevention paradox of Rose[29]: "A large number of people at a small risk may give rise to more cases of disease than the small number who are at a high risk." Another seemingly contradicting result is the low risk for SP for women with a high risk of LRR. This is caused by the competing risks: if a woman experiences a LRR or a DM, she cannot be diagnosed with a SP as a first event anymore. But as follow-up decisions should be made by taking into account both SP and LRR risk, the low risk assigned for SP will not result in undertreatment. Further research needs to point out relevant risk thresholds for follow-up decisions. Then it only matters whether the risk meets this threshold in actual cases, not exactly how high it is (e.g., there is no difference in decision if the assigned risk is 11% or 90% if a threshold is set at 10%).

Bayesian networks are graphical tools to explore the dependence structure of the data. The variables are assumed to be independent, conditionally independent, or dependent. The Pearson correlation coefficient is one of the most well-known dependency measures. However, it is only able to capture linear associations between 2 variables.[30,31] An alternative is to consider Spearman rank correlation, which accounts for monotone association between 2 variables. Spearman rank correlation can be used to specify the structure of a BN. Alternatively,

constraint-based algorithms employ conditional independence tests, whereas score-based algorithms find structure with best networks scores, either in terms of AIC or log likelihood. These approaches led to different network structures, which were evaluated from a fitting and, more important, from a predictive point of view, with the c-statistic, as well as the error rate. BNs consistently provided a higher c-statistic in the training data, suggesting a better fitting model than logistic regression. Nevertheless, the lower c-statistic in the validation cohort suggests a lower predictive performance. It should be emphasized, however, that the difference was overall negligible. In general, we noticed that the more links BNs had, the smaller the c-statistic for the test set compared to the training set. This might suggest that the performance of the BNs in the test set was relatively sensitive to the structure of the network. In this respect, it is also worth mentioning that the structure obtained using a score-based algorithm based on AIC (for LRR) and BIC (for SP) resulted in a higher c-statistic for the validation cohort compared to the index cohort. As expected, the c-statistic of all models on the pooled permutation data sets with randomly assigned labels was around 0.5. Finally, it is of note that the error rate was consistently lower in the validation set compared to the training set.

The actual performance of a model is unrelated to the methods used for evaluation and exists objectively. This real performance can only be estimated using performance measures. There is no single measure that is able to describe all aspects of the performance of a model. Consequently, it is important to look at several and make a comparison, also keeping in mind the aim of the model. Different aims could lead to a different importance of the performance of the measures used and subsequently also different conclusions on which model to use for a specific application. This is exemplified in our study with the good performance of the TS BN in the subgroup analysis for SP risk prediction (Table 4). Although this model was best in subscribing a high risk to actual cases, from the c-statistic of 0.5, it could be seen that the model had no discriminative ability, which means that in this application, also noncases would needlessly be assigned with high risks. In addition, the use of a validation data set for assessing the performance of models was also shown to be of great importance, as there was a decline in the performance found for all the models.

Several approaches can be taken to extend the methodology used in this study. One option could have been to explore a combination of the logistic regression model and BNs, for example, by using the Markov blanket from a BN as input selection for the regression model or estimating the conditional probabilities with logistic regression as input for a BN. Rijmen[32] used an approach where all conditional probability tables were restricted according to a regression model but found worse performance compared to an unrestricted BN. However, the differences became smaller for larger sample sizes and more missing values. Rijmen[32] proposes to develop a BN starting with an unrestricted model and use a learning scheme to gradually remove links starting with the highest order interaction. Although we did not use this specific approach, we did use several different structure learning approaches, ending up comparing 28 different BNs for each of the events of interest with logistic regression models. Alternatively, a targeted maximum likelihood estimation (MLE) approach might be considered in an attempt to improve the performance of logistic regression or BNs. Targeted MLE has appealing theoretical properties and has been compared to logistic regression on several occasions.[33,34] A comparison between target MLE and BNs would be interesting to explore if targeted MLE would improve the performance of logistic regression and BNs. This is, however, beyond the scope of our study. Also, it would be rather difficult to incorporate into a decision aid.

In our study, we had the advantage of using the large cohort from the nationwide population-based NCR, including almost all early staged breast cancer patients diagnosed in the Netherlands between 2003 and 2006. We were, however, limited in the amount of variables from the patients. For example, regular testing and registration of HER2-neu status started after 2005. Furthermore, the number of patients who are diagnosed with a recurrence is very low. For a higher ratio of events v. nonevents, results with BNs can become better. Kim et al.[35] found a c-statistic of 0.81 for predicting DM after breast cancer with a training set of only 458 patients. When looking at patients with DM as a first event in our data (9%), we found higher performance (c-statistics 0.73–0.76), but logistic regression still outperformed the BNs (data not shown). The data set was quite complete, with only 0.6% of the patients having an incomplete 5-year follow-up. With this low number, combined with the fact that light censoring ($<20\%$) does not influence the development of the BNs,[36] it is not expected that the overall results were influenced by the missing follow-up.

As we had a relatively large data set, the overall better performance of logistic regression could have been expected. However, although the literature is not consistent, in the subgroups and for prediction of SP, BNs might have had an advantage. Several studies showed good performance of BNs in prediction.[5,37–46] However,

there the model relied (partly) on expert opinion by lack of other data[5,44]; the comparison was made between BNs and clinician performance,[37,38,40,42] as well as between BNs and another machine-learning technique[41]; or there was no comparison made.[39,43,45] In 2 studies, BNs were outperformed by logistic regression as well, but they only contained small data samples (<190 patients).[47,48] As more and more information on individual patients will become available in clinical practice, models that are able to incorporate numerous variables are expected to outperform conventional models. This can be seen, for example, in the study by Gevaert et al.,[46] in which clinical and microarray data were combined in a BN. When data are high dimensional and there are not many training data, logistic regression will lead to overfitting.[9] So if there is an abundance of variables or a lack of data (which could instigate the need for implementation of expert opinion), BNs could be a better option. And in contrast with other machine-learning techniques such as artificial neural networks, BNs do allow for easy interpretation. BNs can enable risk estimates rapidly via conditionalization, whereas for logistic regression, further steps are necessary. Another advantage of using a BN is the flexible handling of missing data, as BNs use all available information, without excluding entries with missing data, like with logistic regression. For logistic regression, it is possible to use imputed data sets, but this requires an extra step in the analysis. An alternative for using Netica is the *bnlearn* package in R. However, this package is not compatible with data sets that have missing values. A downside of using the Netica program is that discretization is needed.[49] However, in our case, all the variables except age were categorical. It is difficult to quantify when which technique is best because it is not just dependent on size of the data set but also on event rate and number of included explaining variables. A simulation study to find thresholds by which BNs would outperform logistic regression as a function of the number of patients in the training set or the number of explaining variables for our specific case falls outside the scope of this study.

Current prediction models are largely based on conventional clinical factors. The maximum predictive value that can be attained with those is limited. A growing effort is put into prediction using multigene prognostic tests.[50] Examples include Mammaprint,[51] PAM50,[52] and Oncotype DX.[53] However, comparative studies found that individual risk predictions were often discordant.[54–56] As such, we aimed to improve the risk prediction using an alternative modeling strategy. Still, LRRs and SP tumors proved difficult to predict. In the absence of new clinically available (genetic) risk factors, another option might be to make optimal use of all the available data. Going from an aggregate level (e.g., chemotherapy yes/no) to the individual level (e.g., timing of chemotherapy, which regimens, and how long) could result in improved estimates.

Summarizing, an accurate breast cancer recurrence risk prediction is required to identify higher or lower risk patients and develop individualized follow-up schemes. Although there is no dependence relationship between the values of the coefficients and the *change* in value of one of the influencing variables in logistic regression, this analysis suggests that it is still more accurate for risk estimation for both LRRs and SP tumors using clinical risk factors than BNs. Despite the modest performance results in terms of prediction, differences were not very large and BNs remain an attractive graphical alternative that can clearly depict existing influences.

Acknowledgments

We thank the registrars of the Netherlands Cancer Registry for their effort in gathering the data essential to this study. Also, we acknowledge the reviewers for their helpful suggestions.

ORCID iD

Annemieke Witteveen https://orcid.org/0000-0001-5581-6478

Supplementary Material

Supplementary material for this article is available on the *Medical Decision Making* Web site at http://journals.sagepub.com/home/mdm.

References

1. Thrift AP, Whiteman DC. Can we really predict risk of cancer? *Cancer Epidemiol*. 2013;37:349–52.
2. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. 2nd ed. New York: Springer; 2009.
3. James G, Daniela W, Trevor H, Robert T. *An Introduction to Statistical Learning*. New York: Springer; 2013.
4. Cheng J, Greiner R. Learning Bayesian belief network classifiers: algorithms and system. *Adv Artif Intell*. 2001; 2056:141–51.
5. Sesen MB, Nicholson AE, Banares-Alcantara R, Kadir T, Brady M. Bayesian networks for clinical decision support in lung cancer care. *PLoS One*. 2013;8:e82349.
6. Bouhamed H, Masmoudi A, Lecroq T, Rebaï A. Structure space of Bayesian networks is dramatically reduced by subdividing it in sub-networks. *J Comput Appl Math*. 2015;287:48–62.
7. Mitchell TM. The discipline of machine learning. *Mach Learn*. 2006;17:1–7.

8. Ng A, Jordan MI.On generative vs. discriminative classifiers: a comparison of logistic regression and naive Bayes. *Proc Adv Neural Inf Process*. 2002;28:169–87.
9. Mitchell TM. Generative and discriminative classifiers: naive Bayes and logistic regression. Available from: http://www.cs.cmu.edu/~tom/NewChapters.html
10. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. *BMJ*. 2015;351:h3868.
11. Chen H-C, Kodell RL, Cheng KF, Chen JJ. Assessment of performance of survival prediction models for cancer prognosis. *BMC Med Res Methodol*. 2012;12:102.
12. Anothaisintawee T, Teerawattananon Y, Wiratkapun C, Kasamesup V, Thakkinstian A. Risk prediction models of breast cancer: a systematic review of model performances. *Breast Cancer Res Treat*. 2012;133:1–10.
13. Witteveen A, Vliegen IMH, Sonke GS, Klaase JM, IJzerman MJ, Siesling S. Personalisation of breast cancer follow-up: a time-dependent prognostic nomogram for the estimation of annual risk of locoregional recurrence in early breast cancer patients. *Breast Cancer Res Treat*. 2015;152:627–36.
14. Netherlands Comprehensive Cancer Organisation (IKNL). Dutch breast cancer guideline. 2012. Available from: http://www.oncoline.nl/breastcancer
15. Moossdorff M, Van Roozendaal LM, Strobbe LJ, et al. Maastricht Delphi consensus on event definitions for classification of recurrence in breast cancer research. *J Natl Cancer Inst* 2014;106:dju288.
16. Witteveen A, Kwast ABG, Sonke GS, IJzerman MJ, Siesling S. Survival after locoregional recurrence or second primary breast cancer: impact of the disease-free interval. *PLoS One*. 2015;10:e0120832.
17. Nederlandse Kankerregistratie. *Codeerhandleiding Nederlandse Kankerregistratie*. Utrecht (Netherlands): Netherlands Comprehensive Cancer Organisation (IKNL); 2013.
18. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30:377–99.
19. Spratt M, Carpenter J, Sterne JC, et al. Strategies for multiple imputation in longitudinal studies. *Am J Epidemiol*. 2010;172:478–87.
20. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. *Mach Learn*. 1997;29:131–63.
21. Rice ME, Harris GT. Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law Hum Behav*. 2005;29:615–20.
22. Pearl J, Verma T. A theory of inferred causation. *Logic Methodol Philos Sci*. 1994;9:789–811.
23. Scutari M. Learning Bayesian networks with the bnlearn R package. *J Stat Softw*. 2010;35:1–22.
24. Aalders KC, van Bommel ACM, van Dalen T, et al. Contemporary risks of local and regional recurrence and contralateral breast cancer in patients treated for primary breast cancer. *Eur J Cancer*. 2016;63:118–26.
25. Buist DSM, Abraham LA, Barlow WE, et al. Diagnosis of second breast cancer events after initial diagnosis of early stage breast cancer. *Breast Cancer Res Treat*. 2010;124:863–73.
26. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21:128–38.
27. Newcombe PJ, Reck BH, Sun J, et al. A comparison of Bayesian and frequentist approaches to incorporating external information for the prediction of prostate cancer risk. *Genet Epidemiol*. 2012;36:1442–8.
28. Myllymaki P, Silander T, Tirri H, Uronen P. B-course: a web-based tool for Bayesian and causal data analysis. *Int J Artif Intell Tools*. 2002;11:369–87.
29. Rose G. Sick individuals and sick populations. *Int J Epidemiol*. 1985;14:32–8.
30. Speed T. A correlation for the 21st century. *Science*. 2011;334:1502–3.
31. Reimherr M, Nicolae DL. On quantifying dependence: a framework for developing interpretable measures. *Stat Sci*. 2013;28:116–30.
32. Rijmen F. Bayesian networks with a logistic regression model for the conditional probabilities. *Int J Approx Reason*. 2008;48:659–66.
33. Lendle SD, Fireman B, Van Der Laan MJ. Targeted maximum likelihood estimation in safety analysis. *J Clin Epidemiol*. 2013;66(Suppl):S91–8.
34. Rosenblum M, van der Laan MJ. Targeted maximum likelihood estimation of the parameter of a marginal structural model. *Int J Biostat*. 2010;6:19.
35. Kim W, Kim KS, Park RW. Nomogram of naive Bayesian model for recurrence prediction of breast cancer. *Healthc Inform Res*. 2016;22:89.
36. Štajduhar I, Dalbelo-Bašić B, Bogunović N. Impact of censoring on learning Bayesian networks in survival modelling. *Artif Intell Med*. 2009;47:199–217.
37. Burnside E, Rubin D, Fine J. Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: initial experience. *Radiology*. 2006;240:666–73.
38. Cruz-Ramírez N, Acosta-Mesa HG, Carrillo-Calvet H, Nava-Fernández LA, Barrientos-Martínez RE. Diagnosis of breast cancer using Bayesian networks: a case study. *Comput Biol Med*. 2007;37:1553–64.
39. Forsberg JA, Eberhardt J, Boland PJ, Wedin R, Healey JH. Estimating survival in patients with operable skeletal metastases: an application of a Bayesian belief network. *PLoS One*. 2011;6:19956.
40. Burnside E, Davis J, Chhatwal J. Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings. *Radiology*. 2009;251:663–72.
41. Giskeødegård G, Grinde M. Multivariate modeling and prediction of breast cancer prognostic factors using MR metabolomics. *J Proteome Res*. 2010;9:972–9.

42. Kahn CE, Roberts LM, Shaffer KA, Haddawy P. Construction of a Bayesian network for mammographic diagnosis of breast cancer. *Comput Biol Med*. 1997;27: 19–29.
43. Wang XH, Zheng B, Good WF, King JL, Chang YH. Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network. *Int J Med Inform*. 1999;54:115–26.
44. Watt EW, Watt E, Bui AAT, Bui AA. Evaluation of a dynamic Bayesian belief network to predict osteoarthritic knee pain using data from the osteoarthritis initiative. *Proc Annu AMIA Symp*. 2008;2008:788–92.
45. Zheng B, Ramalingam P, Hariharan H, Leader JK, Gur D. Prediction of near-term breast cancer risk using a Bayesian belief network. *Proc SPIE*. 2013;8673:1–7.
46. Gevaert O, De Smet F, Timmerman D, Moreau Y, De Moor B. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*. 2006;22:184–90.
47. Mazzocco T, Hussain A. Novel logistic regression models to aid the diagnosis of dementia. *Expert Syst Appl*. 2012;39:3356–61.
48. Forsberg JA, Sjoberg D, Chen Q-R, et al. Treating metastatic disease: which survival model is best suited for the clinic? *Clin Orthop Relat Res*. 2013;471:843–50.
49. Kuhnert PM, Hayes KR. How believable is your BBN? *Proc 18th World IMACS*. 2009;13–7.
50. Gybrffy B, Hatzis C, Sanft T, Hofstatter E, Aktas B, Pusztai L. Multigene prognostic tests in breast cancer: past, present, future. *Breast Cancer Res*. 2015;17:11.
51. Mook S, Schmidt MK, Viale G, et al. The 70-gene prognosis-signature predicts disease outcome in breast cancer patients with 1–3 positive lymph nodes in an independent validation study. *Breast Cancer Res Treat*. 2009;116:295–302.
52. Dowsett M, Sestak I, Lopez-Knowles E, et al. Comparison of PAM50 risk of recurrence score with oncotype DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. *J Clin Oncol*. 2013;31:2783–90.
53. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med*. 2004;351:2817–26.
54. Prat A, Parker JS, Fan C, et al. Concordance among gene expression-based predictors for ER-positive breast cancer treated with adjuvant tamoxifen. *Ann Oncol*. 2012;23:2866–73.
55. Kelly CM, Bernard PS, Krishnamurthy S, et al. Agreement in risk prediction between the 21-gene recurrence score assay (Oncotype DX®) and the PAM50 breast cancer intrinsic Classifier^{TM} in early-stage estrogen receptor-positive breast cancer. *Oncologist*. 2012;17:492–8.
56. Iwamoto T, Lee JS, Bianchini G, et al. First generation prognostic gene signatures for breast cancer predict both survival and chemotherapy sensitivity and identify overlapping patient populations. *Breast Cancer Res Treat*. 2011; 130:155–64.