Send Orders for Reprints to reprints@benthamscience.ae

16 *The Open Bioinformatics Journal,* **2014,** *8,* **16-21**

Open Access

Predicting Neutropenia Risk in Breast Cancer Patients from Pre-Chemotherapy Characteristics

Sodiq Lawal[1,*], Michael J. Korenberg[1,*], Natalia Pittman[2] and Mihaela Mates[2]

[1]*Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada*

[2]*Kingston General Hospital Cancer Centre of Southeastern Ontario, Canada*

Abstract: A previous study (Pittman, Hopman, Mates) of breast cancer patients undergoing curative chemotherapy (CT) found that the third most common reason for emergency department (ER) visits and hospital admission (HA) was febrile neutropenia. Factors associated with ER visits and HA included (1) stage of the cancer, (2) size of tumor, (3) adjuvant versus neo-adjuvant CT ("adjuvance"), and (4) number of CT cycles. We hypothesized that a statistically-significant predictor of neutropenia could be built based on some of these factors, so that risk of neutropenia predicted for a patient feeling unwell during CT could be used in weighing need to visit the ER. The number of CT cycles was not used as a factor so that the predictor could calculate the neutropenia risk for a patient before the first CT cycle. Different models were built corresponding to different pre-chemotherapy factors or combinations of factors. The single factor yielding the best classification accuracy was tumor size (Mathews' correlation coefficient $\varphi = +0.18$, Fisher's exact two-tailed probability $P < 0.0374$). The odds ratio of developing febrile neutropenia for the predicted high-risk group compared to the predicted low-risk group was 5.1875. Combining tumor size with adjuvance yielded a slightly more accurate predictor (Mathews' correlation coefficient $\varphi = +0.19$, Fisher's exact two-tailed probability $P < 0.0331$, odds ratio = 5.5093). Based on the observed odds ratios, we conclude that a simple predictor of neutropenia may have value in deciding whether to recommend an ER visit. The predictor is sufficiently fast that it can run conveniently as an Applet on a mobile computing device.

Keywords: Breast cancer, chemotherapy, classifier, nearest neighbor, neutropenia.

1. INTRODUCTION

The present paper introduces a fast and efficient method to predict whether a breast cancer patient undergoing chemotherapy (CT) is at high risk of developing febrile neutropenia, using predictive factors available before the first chemotherapy cycle. The motivation for this work is a recent study [1] into factors associated with emergency room visits and hospital admissions in patients undergoing curative chemotherapy for breast cancer in the Southeast Ontario Local Health Integration Network (LHIN). These patients have a higher risk of emergency room visits and hospital admission rates compared to other LHINs in Ontario, Canada. The study found that febrile neutropenia was the third most common cause of emergency room (ER) visits [1]. It also found that the only statistically significant factor associated with ER visits was the stage of the cancer, while factors with statistically significant associations with hospital admissions (HA) were tumor size, chemotherapy type (namely adjuvant versus neoadjuvant), and the number of CT cycles [1]. A natural follow-up of this work is to develop a statistically-significant predictor of neutropenia risk, which could preferentially earmark high-risk patients for increased surveillance. Previously, another study [2] of neutropenia prediction using first cycle blood cell counts produced a FOS-3NN classifier that was extremely accurate in predicting neutropenic events. It had a Fisher's exact 2-tailed probability of $P < 0.00023$ and a Mathew's correlation coefficient of +0.83.

Building on the results from [1], we hypothesized that a statistically-significant predictor can be developed to calculate the neutropenia risk of a patient undergoing chemotherapy, based on predictive factors available before the first CT cycle. Since the predictor was developed assuming that the patient had yet to undergo the first CT cycle, we did not use the number of CT cycles as one of the factors in the predictor because that information will not be available before the patient's CT.

The predictive model was developed in MATLAB. The goal is to predict if a patient is at high risk of developing neutropenia based on the above mentioned factors: Stage of Cancer, Tumor Size and CT Type ("Adjuvance") using Nearest Neighbour Classifiers. Therefore, using some of these factors, we built a predictor of neutropenia risk based on a nearest neighbor classifier. A Leave-One-Out test protocol was employed, omitting the data of a patient under test from the training data for the predictive model. The model first found the mean value of each factor used for the training neutropenic and non-neutropenic patient data, then it classified the patient left out of the model as neutropenic or non-neutropenic depending on which mean its factor value was closest to in terms of standard deviations. For example,

*Address correspondence to these authors at the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada;
Tel: 1-613-533-2931, 1-343-363-3776; Fax: 1-613-533-6615;
E-mails: Korenber@queensu.ca, 0soil@queensu.ca

if tumor size was the only factor used, the mean size m_1 and standard deviation s_1 for the neutropenia patients, and the mean size m_2 and standard deviation s_2 for the non-neutropenia patients, were calculated, without including the data for the test patient. The test patient was then classified according to whether that patient's tumor size was closer, in the number of standard deviations, to m_1 or m_2. If two or more factors were used, we summed the number of standard deviations for the factors, then chose the class for which the sum was smallest. This data analysis algorithm is easy on a computer's processor and accurately creates the weighted sums of the data needed for classifying a test patient.

2. METHOD

The program operates by first importing the Breast Cancer Patient Database from an EXCEL spreadsheet into MATLAB. The database contains data for 149 patients, of which 9 patients are neutropenic, and 140 patients are non-neutropenic. We could not use one non-neutropenic patient from the database because of missing data. Each patient's data consists of information such as age, gender, date of CT Cycles and more importantly, the three characteristics (factors) the predictor uses to predict risk of developing neutropenia, namely Tumor Size, Chemotherapy Type, and Stage of Cancer. In the database, each type and size are assigned an integer value as in Table **1** below.

The program then separates the training data into two classes. One class contains the Neutropenic Patients, the other class contains the Non-Neutropenic Patients.

2.1. Nearest Neighbour Classifier Algorithm and Leave One-Out Protocol

Once the data have been separated, each patient's classification is determined using a nearest neighbour classifier algorithm and a Leave-One-Out protocol. The Leave-One-Out protocol means that the data of a patient being classified is not used when the mean and standard deviation for each factor is calculated. The algorithm finds the arithmetic mean value for all three factors for both classes using the equation:

$$m(x) = \frac{1}{N}\sum_{i=1}^{N} a_i\,(x) \tag{1}$$

Here $m(x)$ is the mean of a factor *x* for a class, *N* is the number of patients in the class (not including the test patient being classified) and $a_i(x)$ is the value of the *i* patient's factor *x*. For example, if *x* denotes tumor size, then the mean tumor size $m(x)$ is calculated for the neutropenic patients class and again for the non-neutropenic patients class, without including the test patient in either calculation.

Once the arithmetic mean is found, the magnitude of one standard deviation for each factor is calculated using the equation,

$$S(x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i(x) - m(x))^2} \tag{2}$$

Here $S(x)$ is the magnitude of one standard deviation for a factor *x* for a class, *N* is the number of patients in the class (not including the patient being classified), $m(x)$ is the mean for the factor *x* in the class, and $a_i(x)$ is the value of the factor *x* for patient *i*. Note that $S(x)$ is calculated for the neutropenic patients class and again for the non-neutropenic patients class.

The program then calculates the distance from the mean in standard deviations for the test patient *k* under classification, for each factor *x* in both classes using the equation

$$d(x) = \left|\frac{\left(a_k(x) - m(x)\right)}{S(x)}\right|, S(x) \neq 0$$

$$= 0, S(x) = 0, a_k(x) = m(x)$$

$$= \infty, S(x) = 0, a_k(x) \neq m(x) \tag{3}$$

Here $a_k(x)$ is the value of the factor *x* for the test patient *k* under classification, $m(x)$ is the arithmetic mean of the values of factor *x* in one of the classes, $S(x)$ is the magnitude of one standard deviation for factor *x* in that class, and $d(x)$ is the distance from the mean for the factor *x* in that class. Thus, for the test patient *k*, Eq. (3) was used to calculate the distance $d(x)$ from the mean for factor *x* in the neutropenia patients class and again in the non-neutropenia patients class, without including patient *k* in calculating $m(x)$ and $S(x)$. The test patient *k* was then predicted to belong to the class for which $d(x)$ was smaller.

2.2. Classification Based on Distance From Mean

As noted, once $d(x)$ is determined for each factor for both classes for a test patient, that patient will be classified as neutropenic or non-neutropenic depending on which mean its factor value is closest to in terms of standard deviations. There were different combinations of factors used for classifying the patient. A single factor was used, or the corresponding values of $d(x)$ for two factors were added or multiplied together, or the corresponding values of $d(x)$ for three factors were added or multiplied together.

Since the earlier study [1] had identified tumor size, chemotherapy type (adjuvant vs neoadjuvant), and number of CT cycles as statistically associated with HA, the first two factors were tried both alone and together as predictors of neutropenia risk. In our nearest neighbor classifiers, tumor size was the best single predictor of neutropenia risk (Matthews' correlation coefficient $\varphi = 0.18$, Fisher's exact two-tailed probability $P < 0.0374$). The odds ratio of developing febrile neutropenia in the predicted high-risk group was 5.1875 relative to the predicted low-risk group and, significant statistically, the 95% confidence interval [1.0394, 25.8907] does not contain the point 1. Combining tumor size with adjuvance slightly improved accuracy (Matthews' correlation coefficient $\varphi = 0.19$, Fisher's exact two-tailed probability $P < 0.0331$). The odds ratio of developing febrile neutropenia in the predicted high-risk group rose to 5.5093, relative to the predicted low-risk group, with 95% confidence interval [1.1033, 27.509] not containing the point 1. For completeness, we also explored all three factors and factor combinations, as summarized in Table **2**.

Once the predictor classifies each patient as seen in Table **3** below, the results were entered into a 2x2 contingency table

Table 1. Table of Tumor Size, CT Type and Stage, their respective subtypes and the corresponding number value in the database.

Number Value in Database	Tumor Size *Size of Primary Tumor*	CT Type *Pre-Surgical Chemotherapy vs. Post-Surgical Chemotherapy*	Stage *Extent of Disease in Terms of Spread and/or Size*
1	TX *Primary Tumor Cannot Be Evaluated*	Adjuvant *Chemotherapy given post-surgery to eliminate left-over cancer cells*	0 *Carcinoma in Situ;* *Abnormal cells are present but no spread*
2	T0 *No evidence of tumor*	NeoAdjuvant *Chemotherapy given pre-surgery to reduce size of tumor*	IA *Tumor is less than 2 cm.*
3	Tis *Carcinoma in situ; Abnormal cells are present but no spread*	-	IB *Clusters of cancer cells in lymph nodes and/or tumor is less than 2 cm.*
4	T1*	-	IIA *Cancer in 1 to 3 lymph nodes in axilla or near breastbone and/or tumor is smaller than 2cm OR tumor is less than 5cm and larger than 2cm*
5	T2*	-	IIB *Tumor is less than 5 cm and larger than 2 cm and clusters of cancer cells in lymph nodes OR tumor greater than 5 cm*
6	T3*	-	IIIA *Cancer in 4 to 9 lymph nodes in axilla near the breastbone and tumor can be nonexistent or any size OR tumor larger than 5 cm and clusters of cancer cells in lymph nodes or cancer cells in lymph nodes near axilla or near the breastbone.*
7	T4*	-	IIIB *Tumor may be any size and cancer has spread to the chest wall and/or to the skin of the breast causing inflammation, swelling or an ulcer OR cancer may have spread to 9 axillary lymph nodes or lymph nodes near breast bone.*
8	-	-	IIIC *Tumor is any size and cancer in 10 or more lymph nodes in the axilla, cancer or lymph nodes above collarbone or breastbone.*
9	-	-	IV *Cancer has spread to other parts of body*

*T1, T2, T3, T4 indicate size and/or extent of tumor, the higher the number, the larger the tumor.
Note: Definitions of Cancer Pathology and Treatment notations and terms are summarized from American Cancer Society and National Cancer Institute Website [3-5].

Table 2. Table of all the factors and factor combinations used to determine the classification of a patient.

Tumor Size	Stage
CT Type	Tumor Size + Stage
Tumor Size × Stage	Stage + CT Type
Stage × CT Type	CT Type + Tumor Size
CT Type × Tumor Size	Tumor Size + Stage + CT Type
Tumor Size × Stage × CT Type	

Table 3. Results of the predictors' classification for all factor combinations.

Factor Combination	Non-Neutropenic Patients Classified as Non-Neutropenic	Non-Neutropenic Patients Classified as Neutropenic	Neutropenic Patients Classified as Neutropenic	Neutropenic Patients Classified as Non-Neutropenic
Tumor Size	83	56	7	2
Stage	78	61	6	3
CT Type	22	117	9	0
Tumor Size + Stage	104	35	2	7
Tumor Size × Stage	75	64	7	2
Stage + CT Type	85	54	6	3
Stage × CT Type	22	117	9	0
CT Type + Tumor Size	85	54	7	2
CT Type × Tumor Size	22	117	9	0
CT Type + Tumor Size + Stage	85	54	6	3
CT Type x Tumor Size x Stage	22	117	9	0

(Vasserstats [6]) to find Fisher's exact two-tailed probability, Mathews' correlation coefficient and the odds ratio. From there we determined the feasibility of using the factor combinations to predict the risk of a neutropenic event.

3. SIGNIFICANCE OF RESULTS

As shown in Table **4**, the most statistically significant single factor predicting neutropenia is tumor size, and the best factor combination is CT Type + Tumor Size. They both have a Fisher's Exact test two-tailed Probability P less than 0.04 and an odds ratio above 5.1. Tumor size, and the combination CT Type + Tumor Size, have the highest Mathews' correlation coefficient of +0.18 and +0.19 respectively. This result is in alignment with the conclusion from ref. [1]**.** As mentioned above, the latter study concluded that Tumor Size and CT Type were the 2 most significant factors associated with hospital admissions (of factors available before CT). It should be noted that every statistically significant result had exactly the same 2 (and only two) neutropenia patients misclassified.

A significant pattern was found when using the factor CT Type by itself, or multiplied by Tumor Stage, Tumor Size, or both. All patients with neutropenia were taking Adjuvant Therapy (or, equivalently, have a 1 in their CT Type column). Therefore by Eq. (1), the mean of the CT Type factor was 1 for the Neutropenic class, and since there was no variation, by Eq. (2) the Length of a Standard Deviation was 0 for that class. Therefore any patient that took adjuvant therapy was classified as Neutropenic and any patient undertaking Neo-Adjuvant Therapy was classified as Non-Neutropenic. Therefore the classification results for CT Type, CT Type X Tumor Stage, CT Type X Tumor Size, and CT Type X Tumor Stage X Tumor Size all reflect how many patients took adjuvant therapy and how many took neo-adjuvant therapy. Also the odds ratio for those factor and factor combinations is infinite. Again that is because every patient with neutropenia is taking Adjuvant Therapy.

3.1. Speed of Results

The predictor takes 25-30 seconds to run on a first generation Core i3-370M processor. Table **5** below shows MATLAB's profile data, which clocks how long the CPU takes to process each line of code. The specific profile shown is for when the program was classifiying patients using all the additive factor combinations. The important

Table 4. Results of 2x2 Contingency Table for all Factor Combinations.

Factor Combination	Mathews' Correlation Coefficient (φ)	Fisher's Exact Two-Tailed Probability (P)	Odd Ratio
Tumor Size	+.18	.0374	5.1875
Stage	+.11	.3004	2.5574
CT Type	+.11	.3567	Infinity
Tumor Size + Stage	-.02	1	.849
Tumor Size × Stage	+.15	.0808	4.1016
Stage + CT Type	+.14	.1589	3.1481
Stage × CT Type	+.11	.3567	Infinity
CT Type + Tumor Size	+.19	.0331	5.5093
CT Type × Tumor Size	+.11	.3567	Infinity
CT Type + Tumor Size + Stage	+.14	.1589	3.1481
CT Type x Tumor Size x Stage	+.11	.3567	Infinity

Table 5. MATLAB's Profiler displaying time spent on each line of code. Note that 95.1% of the time spent running the code is on lines associated with importing and exporting data to an excel file.

Parents (calling functions) No Parent				
Lines Where the Most Time was Spent				
Line Number	**Code**	**Calls**	**Total Time**	**% Time**
296	xlswrite(' PatientClassifier.xlsx…	1	1.873 s	7.0%
4	Master_Data = xlsread('Breast_...	1	1.853 s	6.9%
590	xlswrite(' PatientClassifier.xls…	1	1.840 s	6.9%
582	xlswrite(' PatientClassifier.xls…	1	1.836 s	6.8%
322	xlswrite(' DualCharacteristicCl.xls…	1	1.805 s	6.7%
All other lines			17.609 s	65.7%
Totals			26.816 s	100%
Children (called functions)				
Lines where the most time was spent				
Function Name	Function Type	Calls	Total Time	% Time
xlsxwrite	function	14	23.646 s	88.2%
xlsxread	function	1	1.850 s	6.9%
Self time (built-ins, overhead, etc.)		1	1.320 s	4.9%
Totals			26.816s	100%

thing to note is that MATLAB spends 88.2% of the time in the function "xslswrite", which is MATLAB's function that writes the output of the predictor to an excel file, while 6.9% of the CPU time is spent in the xlsread function. The latter function loads the Breast Cancer Patient Data from an excel file into MATLAB's private workspace. The actual time spent creating the model and classifying all the patients in the database is 4.9% or 1.32 seconds. If this model was to be used in practical applications, a more efficient means of retrieving and writing to a database would be implemented as MATLAB is known to be slow when importing and exporting data.

4. DISCUSSION

This paper demonstrates that a Nearest Neighbour Classifier can be used to achieve statistically-significant prediction

of risk of developing neutropenia based on factors available before the first chemotherapy cycle. The aim of this pre-assessment of a patient's vulnerability is to enable quick implementation of preventative measures for the patient. Wingard and Elmongy [7] have examined the use of myeloid growth factors and antibiotics to reduce infectious complications, and report that both strategies can be beneficial in selected patients. Altwairgi, Hopman, and Mates [8] report that over 50% of patients treated for early-stage breast cancer in the Southeast Ontario LHIN receive primary prophylaxis with granulocyte colony-stimulating factors (G-CSFs), leading to less febrile neutropenia, but risk is still high for older-age patients, taxane-based CT, and use of the G-CSF filgrastim. In [9], Aarts, Grutters, Peters, et al. compared two treatment strategies of G-CSF pegfilgrastim prophylaxis, showing that G-CSF throughout all 6 CT cycles, though more costly, was associated with 10% incidence of febrile neutropenia as opposed to 36% incidence for G-CSF during only the first 2 CT cycles.

The results in the present study have shown that the most statistically significant factor combination associated with neutropenia is CT Type combined with Tumor Size, followed closely by Tumor Size alone. It also corroborates the finding of the earlier study [1] that a patient with a tumor size of T2 was more likely to be admitted to hospital. Finally it highlights that only patients in this particular group undertaking adjuvant therapy were at risk of developing neutropenia. No patient receiving neo-adjuvant therapy had been diagnosed with neutropenia but this is limited by the small number of patients in this group. Additionally patient's comorbidities were not significant predictors of febrile neutropenia in this small cohort but merits further research.

CONCLUSION

Based on our observed odds ratios over 5 of developing neutropenia in the predicted high-risk group relative to the predicted low-risk group, we conclude that a simple predictor of neutropenia may have value in deciding whether to recommend an ER visit. A possible use of the predictor is that when it places a patient in the non-neutropenia group, the probability that the patient will develop neutropenia is apparently small (about 2/85 in the present study), so the predictor appears to have good negative predictive value. The success of this neutropenia prediction must be confirmed in a larger independent data set. In addition, in view of ref. [2], first cycle blood count data should be combined with the predictive factors used in the present study to increase the sensitivity and specificity of recognizing high-risk patients.

The processing time for the classifier algorithm is less than 1.5 seconds when classifying all 148 patients. In a practical setting only one patient will be classified at a time, therefore the processing time will be approximately a hundredth of a second.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

We thank Grier Owen for establishing the contact between the physicians and the ECE researchers in this study.

REFERENCES

[1] N. M. Pittman, W. M. Hopman and M. Mates, "Emergency room visits and hospital admission rates after curative chemotherapy for breast cancer. A retrospective single centre experience", in San Antonio Breast Cancer Symposium, San Antonio, TX, 2013, Poster P6-06-56.

[2] E. A. Shirdel, M. J. Korenberg, and Y. Madarnas, "Neutropenia prediction based on first-cycle blood counts using a fos-3nn classifier", *Advances in Bioinformatics*, vol. 2011, pp. 1-8, 2011.

[3] National Cancer Institute, "Adjuvant and neoadjuvant therapy for breast cancer". (National Cancer Institute at the National Institutes of Health), [online] 2009, http://www.cancer.gov/cancertopics/factsheet/Therapy/adjuvant-breast [Accessed: 26 July, 2014]

[4] National Cancer Institute, "Stages of Breast Cancer". (National Cancer Institute at the National Institutes of Health), [online] 2014, http://www.cancer.gov/cancertopics/pdq/treatment/breast/Patient/page2#Keypoint15 [Accessed: 26 July, 2014]

[5] American Cancer Society, "Staging". (American Cancer Society), [online] 2012, http://www.cancer.org/treatment/understandingyourdiagnosis/staging [Accessed: 26 July, 2014]

[6] R. Lowry, "Clinical Research Calculators. For a 2x2 Contingency Table". (VassarStats: Website for Statistical Computation), [online] 2001, http://vasserstats.net [Accessed: 21 Feb 2014]

[7] J. R. Wingard and M. Elmongy, "Strategies for minimizing complications of neutropenia: prophylactic myeloid growth factors or antibiotics", *Critical Review Oncology/Hematology*, vol. 72, no. 2, pp. 144-154, 2009.

[8] A. K. Altwairgi, W. M. Hopman, and M. Mates, "Real-world impact of granulocyte-colony stimulating factor on febrile neutropenia", *Current Oncology*, vol. 20, pp. e171-179, 2013.

[9] M. J. Aarts, J. P. Grutters, F. P. Peters, C. M. Mandigers, M. Wouter Dercksen, J. M. Stouthard, H. J. Nortier, H. W. van Laarhoven, L. J. van Warmerdam, A. J. van de Wouw, E. M. Jacobs, V. Mattijssen, C. C. van der Rijt, T. J. Smilde, A. W. van der Velden, M. Temizkan, E. Batman, E. W. Muller, S. M. van Gastel, M. A. Joore, G. F. Borm, and V. C. Tjan-Heijnen, "Cost effectiveness of primary pegfilgrastim prophylaxis in patients with breast cancer at risk of febrile neutropenia", *Journal of Clinical Oncology*, vol. 31, no. 34, pp. 4283-4289, 2013.

Received: June 04, 2014 Revised: September 02, 2014 Accepted: October 09, 2014