# Data Mining Techniques for Predicting Response to Palliative Chemotherapy

Grisselle Centeno, Ludwig Kuznia, Bo Zeng, Brian Decker, Veronica Decker, and David Decker

***Abstract*—The purpose of this work is to develop a model to predict a stage IV breast cancer patient's response to first-line chemotherapy using data mining techniques. We discuss the process of extracting and processing electronic medical record (EMR) data from a private oncology practice and the method for developing a logistic regression model based on commonly collected laboratory data. There were approximately 1200 patients from a large medical oncology practice in the mid-west initially identified for participation in our study. A $k$-fold cross validation was utilized to train and test the model with accuracy, specificity, and sensitivity being used for evaluation purposes. Three consensus models (CM1-3) were constructed with accuracy being the primary measure of model performance. The accuracies, as a percent, for the three models were $71.96 \pm 0.22$, $71.87 \pm 0.22$, and $71.04 \pm 0.15$. The difference in accuracies was found to be significantly different for each pair of consensus models (CM1 vs CM2, $p = 0.02$; CM1 vs CM3, $p < 0.001$; CM2 vs CM3, $p < 0.001$).**

***Index Terms*—Data mining, logistic regression, chemotherapy, cancer treatment prognosis.**

## I. Introduction

ACCORDING to the National Cancer Institute, approximately 1.5 million individuals were diagnosed with cancer in 2010 and over 200,000 of these cases were breast cancer [1]. Breast cancer is categorized into one of four stages, which are, in increasing level of severity, I, II, III, and IV, with each stage having subcategories. There are five types of treatment for breast cancer patients: surgery, radiation therapy, chemotherapy, hormone therapy, and targeted therapy [2]. The treatment selected is dependent on the stage of the disease. Stages I, II, III, and operable IIIC are treated with some combination of surgery and radiation therapy followed by adjuvant therapy. Adjuvant therapy is the additional treatment given after primary treatment to reduce the risk of recurrence of cancer [3]. It can consist of some combination of radiation therapy, hormone therapy, and chemotherapy. In the case of adjuvant chemotherapy, a patient will finish a protocol completely unless they experience toxicity, e.g., nausea, anemia, mouth sores, etc., to the drugs. Stage IV and inoperable IIIC are often treated with systemic chemotherapy and/or hormone therapy [2]. For stage III and IV breast cancer patients, which comprise 38% of the breast cancer population [1], chemotherapy may in fact be the only cancer fighting component of the long term treatment plan. This indicates that advancements in chemotherapy treatment administration can impact a significant portion of breast cancer patients.

Chemotherapy most commonly refers to the use of antineoplastic drugs to treat cancer. Since chemotherapy works by killing cells that divide rapidly, healthy cells that grow rapidly under normal circumstances, including bone marrow cells, digestive tract cells, and hair follicles, are adversely affected by chemotherapy treatment. As such, it is vital that chemotherapy be administered in manner which minimizes the risk to the patient. Chemotherapy can be classified into three categories: i) neoadjuvant – designed to shrink the primary tumor to aid in primary treatment; ii) adjuvant – given after primary treatment designed to reduce the risk of recurrence when little or no evidence of cancer is present; iii) palliative – systemic treatment designed to decrease tumor load and extend life expectancy often without curative intent. These treatment options are dependent on the stage of the disease. An example of neoadjuvant care would be a stage II patient receiving chemotherapy prior to surgery to reduce the tumor load and decrease the chance of relapse. Adjuvant therapy on the other hand could consist of a stage I patient receiving chemotherapy after a tumor is removed and there is no evidence of disease. Treatment is given in this case only to reduce the possibility of relapse. Palliative chemotherapy is reserved for patients with a terminal diagnosis, this is usually only stage IV. The purpose is to improve the patient's quality of life by reducing tumor load. Palliative chemotherapy will rarely cure a stage IV patient and is given as long as the

L. Kuznia, G. Centeno, and B. Zeng are with Dept. of Industrial and Management Systems Engineering, Tampa, FL 33620 email: lkuznia@mail.usf.edu, {gcenteno, bzeng}@usf.edu.

B.Decker is the CEO of EMOL Health, Clawson, MI email: brian.decker@emolhealth.com.

V. Decker and D.Decker are with Florida Hospital, Orlando, FL.

benefits to the patient out weight the side effects.

A line of chemotherapy (also known as a protocol) is defined by a combination of drugs and a time line to administer the drugs. It is generally broken down into cycles which dictate the timing of treatment. Ideally, these cycles are timed to attack cancer cells when they are most vulnerable. There are periods in between cycles that allow patients to recover from treatment. At the beginning of each cycle, the patient must be evaluated to determine the best course of action among three options: continue current line, switch to a new line, or stop treatment altogether.

Response to treatment is defined through the RECIST guideline [4] which involves measuring present tumors and determining the relative change in size. This is a disease oriented definition of response and does not capture any information about patient response to treatment. Specifically, due to the nature of chemotherapy, a patient may suffer from toxicity and would also be classified as not responding to treatment in this case. Therefore, we defined response to treatment from a clinical standpoint to incorporate both the disease and patient responses. This was done by noting that treatment would be stopped for one of two reasons: i) cancer is progressing with current line of treatment (i.e., negative disease response); ii) the patient is experiencing significant side effects from the current line of treatment (i.e., negative patient response). According to a leading oncologist, if a patient remains on a line of treatment for at least two months, neither of the above items occurred [5]. Based on this, we adopt the convention that if first-line chemotherapy lasts at least 50 days, then this would be classified as a positive response to treatment. From this point, patient response to treatment is based on this definition.

The aim of this work is to develop a model that predicts a patient's response to first-line stage IV chemotherapy treatment, as opposed to neoadjuvant or adjuvant treatment, based on a subset of commonly collected vital signs and laboratory results (collectively referred to as labs). Basically, patients are classified into one of two categories, responds well to treatment or not. This is accomplished through the use of logistic regression. Logistic regression is a powerful tool for predicting dichotomous outcomes as a function with multiple inputs [6]. It can be particularly useful for predicting the disease state of a patient as well as determination of yes/no decisions [7]. Logistic regression has been a successful tool for classifications relating to cancer. In [8], Chhatwal et al. developed two models for predicting breast cancer risk based on the descriptors of the National Mammography Database. Moreover, the factors impacting patient mortality and transferring were explored in [9] by Zhang et al. through the use of logistic regression. This work will provide insight to how late-stage breast cancer patients react to long term treatment. Due to the low 5 year survival rate for stage IV breast cancer patients, approximately 15% [10], adding to this body of knowledge is of substantial importance. Moreover, the findings of this research serves as the groundwork for ongoing research for developing tools to improve the quality of palliative care through the use of stochastic modeling.

This paper is organized as follows. Section II describes the process by which patients were selected for study as well as which labs were considered as variables in our models. Next, the procedure by which the models were generated is explained in detail. In Section III, the results of the study are presented along with an interpretation of each model's performance. Finally, we provide concluding remarks and directions for future research in Section IV.

## II. METHODS

For this research, data were provided by a large medical oncology practice in the mid-west. EMOL Health [11] manages the database that houses this practice's electronic medical records (EMRs) as well as text files of dictations prepared by the physicians. The company has developed a number of tools to extract information from these dictations; using these tools and data from the EMRs a total of 1253 potential stage IV breast cancer patients were selected for our study. This exploratory research focuses on stage IV patients receiving chemotherapy. Additionally, at the practice supplying data, stage IV patients received a standard line of chemotherapy (i.e., a standard dose on a standard schedule). We note here that the data used were de-identified and HIPAA compliant. Patient selection was done in a two step process described in Table I. This assumes the patient has only one type of cancer at a time.

Patients receiving any chemotherapy were selected from the group of 1253 potential stage IV patients; this resulted in 471 patients for our study. From this group, we needed to identify the true first-line of chemotherapy treatment (at stage IV). Since the treatment of stage IV breast cancer does not include adjuvant therapy, any adjuvant therapy had to be identified and removed. This is due to the fact that, as noted in Section I, patients nearly always finish adjuvant therapy. Since there is no field in the EMR to identify if a line a

TABLE I: Selection Process

| | |
|---|---|
| **Step 1** | Select patients with any history of breast cancer by checking for an ICD-9 code of 174.X in the EMR |
| **Step 2** | Of those patients selected in Step 1, select those meeting any of the following criteria:<br>1) Maximum stage entered by the doctor in the EMR was "IV"<br>2) An ICD-9 code indicating cancer has metastasized to another location is present<br>3) Information in the doctor's dictations indicated the cancer had metastasized to another location |

therapy is adjuvant, all commonly used adjuvant lines of therapy were removed. A list of the commonly used adjuvant therapy lines at the practice that provided the data for this study is given in Table II. A total of 209 patients remained after removing all adjuvant lines of therapy from the stage IV patients that received any chemotherapy. The first treatment a patient in this group received is considered to be a true first-line of treatment since possible lines of neoadjuvant and adjuvant treatment were removed.

TABLE II: List of Common Adjuvant Therapy Lines

| |
|---|
| Adriamycin, Cytoxan, Taxol |
| Adriamycin, Cytoxan, Taxotere |
| Adriamycin, Cytoxan |
| Taxotere, Carboplatin, Herceptin |

The number of labs reported in varying frequencies for each patient totaled 47. Due to the sparseness of data for certain labs, some were eliminated from consideration. Specifically, labs reported less than 50% of the time were excluded. After this exclusion, 29 labs were considered in the analysis, see the Appendix for a list. Before describing the process of model creation, we digress to discuss the format of the data. EMRs are designed to store large amounts of data effectively and this storage method may not be conducive to statistical analysis and modeling. Therefore, it is vital that an efficient method for re-formatting EMR data be available for those working with real EMR data. By working closely with a company that understands oncology, oncology data, and are experts on EMR data extraction (EMOL Health in our case), we were able to understand the format that is used to store data and reasons for using said format. Specifically, lab results are stored in the following format (`Patient ID`, `Lab Date`, `Lab Value`, `Lab Name`). After re-formatting the data, model construction was performed.

A variation of $k$-fold cross validation was used to create three consensus models for predicting a patient's responsiveness to therapy based on a subset of commonly recorded lab results. This variation is designed to exploit as much of the data as possible since our dataset was reduced significantly. Table III outlines this procedure.

TABLE III: Procedure for Model Construction

Let $N$ be the initial set of patients
Let $L$ be the initial set of predictors (labs)
**Step 0**
  Select the subset of patients, $N_0$, from $N$ with all predictors in $L$ present
**Step 1**
  Divide $N_0$ into $K$ subsets (folds) of equal size, $N_0^1, \ldots, N_0^K$
  For $i = 1, \ldots, K$
    For each $\ell \in L$
      Run a univariate regression with $\ell$ as the predictor and $N_0 - N_0^i$ the training set.
        If the p-value for $\ell$ is less than 0.05, add $\ell$ to $L_0$
**Step 2**
  Select the subset of patients, $N_1$, from $N$ with all predictors in $L_0$ present
**Step 3**
  Divide $N_1$ into $K$ subsets (folds) of equal size, $N_1^1, \ldots, N_1^K$
  For $i = 1, \ldots, K$
    For each $\ell \in L$
      Run a univariate regression with $\ell$ as the predictor, and $N_1 - N_1^i$ the training set.
        If the p-value for $\ell$ is less than 0.05, add $\ell$ to $L_1^i$
    Run a multivariate regression with $L_1^i$ as the set of predictors, $N_1 - N_i^i$ the training set, and $N_1^i$ as the test set.
    Call the resulting model $M_i$
    Record the accuracy of $M_i$ when applied to $N_1^i$, call this $A_i$
**Step 4**
  Create a consensus model $M$

The impacts of each step in the process of model construction in one sample run, which was coded in R [12], are now discussed. After Step 0, there were 96 patients with all 29 lab results present. Six folds were used in Step 1 ($K = 6$ and $|N_0^i| = 16$). After Step 1, $L_0$ consisted of 4 labs; `AST`, `PLT`, `Temp (F)`, and `Total Protein`. Step 3 resulted in 104 patients with all labs from $L_0$ present. In Step 3, 8 folds were used with size 13 ($K = 8$ and $|N_1^i| = 13$). The predictors used in the eight multivariate regressions are given in Table IV with the 95% confidence interval for the model coefficients and the accuracy[1] of each model against the corresponding test set.

[1]Let $y^j$ be the observed outcomes and $\hat{y}^j$ the predicted outcomes, then the accuracy is given by $\frac{1}{N}\sum_{j=1}^{N}|y^i - \hat{y}^j|$.

TABLE IV: Sample Results from Multivariate Regressions in Step 3 of Table III

| Model/ Accuracy | Variables | Coefficient | p-Value |
|---|---|---|---|
| 1 | *Intercept* | $1.898 \pm 1.384$ | 0.0067 |
| 0.6 | PLT | $-0.00504 \pm 0.00457$ | 0.0289 |
| 2 | *Intercept* | $2.323 \pm 1.459$ | 0.0016 |
| 0.5 | PLT | $-0.00578 \pm 0.00484$ | 0.0180 |
| 3 | *Intercept* | $-7.310 \pm 6.185$ | 0.0188 |
| 0.3 | PLT | $-0.00722 \pm 0.00512$ | 0.0053 |
| | Total Protein | $1.454 \pm 0.937$ | 0.0021 |
| 4 | *Intercept* | $-4.084 \pm 5.231$ | 0.1209 |
| 0.8 | PLT | $-0.00595 \pm 0.00476$ | 0.0132 |
| | Total Protein | $0.874 \pm 0.758$ | 0.0224 |
| 5 | *Intercept* | $66.436 \pm 65.841$ | 0.0459 |
| 0.3 | Temp (F) | $-0.670 \pm 0.671$ | 0.0482 |
| 6 | *Intercept* | $-3.942 \pm 5.312$ | 0.1401 |
| 0.5 | PLT | $-0.00625 \pm 0.00484$ | 0.0107 |
| | Total Protein | $0.895 \pm 0.775$ | 0.0221 |
| 7 | *Intercept* | $-4.360 \pm 4.966$ | 0.0806 |
| 0.5 | Total Protein | $0.701 \pm 0.706$ | 0.0483 |
| 8 | *Intercept* | $77.049 \pm 71.915$ | 0.0340 |
| 0.3 | PLT | $-0.00629 \pm 0.00509$ | 0.0143 |
| | Temp (F) | $-0.873 \pm 0.752$ | 0.0216 |
| | Total Protein | $1.525 \pm 0.913$ | 0.0009 |

Three methods for combining the $k$ models generated by the cross-validation were considered. These consensus models, created in Step 4, were obtained by averaging the regression models found in Step 3 in various ways. Recall the classifier for a logistic regression is obtained by the following formula

$$y = R\left(\frac{1}{1+e^{-z}}\right),$$

where $z = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$ and $R(-)$ is the rounding function. The predictors are $x_1, \ldots, x_n$, and the model coefficients $\beta_0, \ldots, \beta_n$ are estimated by maximizing the log-likelihood for a given set of observations.

Consensus model 1 (CM1) was created by averaging the coefficients of the logistic regression models found in Step 3. Specifically, letting $(\beta_0^i, \ldots, \beta_n^i)$ be the coefficients from model $i$ in Step 3, define $(\beta_0, \ldots, \beta_n)$ for CM1 by

$$\beta_k = \sum_{i=1}^{8} \frac{\beta_k^i}{8} \quad k = 0, \ldots, n.$$

Letting $(x_1^j, \ldots, x_n^j)$ be the set of observed predictors for patient $j$, then the predicted outcome from CM1 for patient $j$ is given by

$$\hat{y}_1^j = R\left(\frac{1}{1+e^{-z^j}}\right),$$

$$\text{where } z^j = \beta_0 + \beta_1 x_1^j + \cdots + \beta_n x_n^j.$$

Consensus model 2 (CM2) was generated by averaging the probabilistic outcome of the logistic regression models then rounding the result. That is,

$$\hat{y}_2^j = R\left(\sum_{i=1}^{8} \frac{1}{8\left(1+e^{-z_i^j}\right)}\right),$$

$$\text{where } z_i^j = \beta_0^i + \beta_1^i x_1^j + \cdots + \beta_n^i x_n^j.$$

Finally, consensus model 3 (CM3) we generated by averaging the classifiers from the logistic regression models:

$$\hat{y}_3^j = R\left(\sum_{i=1}^{8} \frac{1}{8} R\left(\frac{1}{1+e^{-z_i^j}}\right)\right),$$

$$\text{where } z_i^j = \beta_0^i + \beta_1^i x_1^j + \cdots + \beta_n^i x_n^j.$$

## III. Results

Steps 3 and 4 were iterated 100 times. At the end of each iteration, the consensus models were used to predict the outcome for the set of 104 patients found in Step 3 of model construction. The accuracy for each was recorded along with the sensitivity (true positive prediction rate) and specificity (true negative prediction rate). The average and half-width of the $95\%$ confidence interval for each performance measure are given in Table V for each of the consensus models. Since the accuracy of CM1 was superior to that of CM2 and CM3 ($p = 0.02$ and $p < 0.001$, respectively), the remainder of our discussion uses only this model. Although the model had an average sensitivity above 90%, the specificity performance was low. This indicates conservative model performance in the sense that a patient is classified as responding to treatment only if there truly is a high chance for success. Thus, we can be confident in treating a patient if the model indicates they will respond well to treatment. On the other hand, if the model indicates the patient will not respond well to treatment, then the patient should be more carefully examined before beginning treatment. These results are promising given the size of the dataset and restricted number of covariates available for consideration.

TABLE V: Summary of Performance of Consensus Models

| Model | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|---|---|---|---|
| CM1 | $71.96 \pm 0.22$ | $91.79 \pm 0.51$ | $36.05 \pm 0.59$ |
| CM2 | $71.87 \pm 0.22$ | $91.60 \pm 0.50$ | $36.14 \pm 0.53$ |
| CM3 | $71.03 \pm 0.15$ | $88.70 \pm 0.25$ | $39.05 \pm 0.36$ |

## IV. Conclusions

In this paper, the process of extracting data from EMRs at a private oncology practice in order to create a model for predicting a stage IV breast cancer patient's response to chemotherapy was presented. In particular, three models were constructed to predict a stage IV breast cancer patient's response to first-line treatment. Based on the results given in Table IV, we can see how changes in `PLT`, `Temp (F)`, and `Total Protein` influence a patient's probability of successful first-line treatment. Specifically, as `PLT` or `Temp (F)` decrease, probability of success increases. On the other hand, an increase in `Total Protein` causes an increased probability of success. This means that within the range of observed values, it is desirable for a patient to have high `Total Protein` and low `PLT` and `Temp (F)`.

Although the accuracy of our best model was approximately 72%, this is in fact a promising result. We were able to extract valuable information from a condensed dataset, indicating that further study is merited. This will include exploring more characteristics of patients as predictors in our model. For instance, it is indicated in [9] and [13] that various comorbities, diseases or disorders present in addition to the primary disease, impact a patient's response to chemotherapy. This data is often available in the doctor's dictations, but tools must be developed to accurately extract it. Additionally, increasing the population size would allow for more robust modeling techniques to be used, thereby improving model performance. It would also be beneficial to validate our model's performance with tumor size data and using RECIST criteria as the measure of a patient's response to treatment.

## References


[1] National Cancer Institute, "Surveillance epidemiology and end results," http://seer.cancer.gov/statfacts/.

[2] ——, http://www.cancer.gov/cancertopics/pdq/treatment/breast/Patient/.

[3] ——, http://www.cancer.gov/dictionary/?CdrID=45587.

[4] E. Eisenhauer, P. Therasse, J. Bogaerts, L. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij, "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)," *European Journal of Cancer*, vol. 45, pp. 228–247, 2009.

[5] D. Decker, Personal Interview, May 2010.

[6] D. W. Homer and S. Lemeshow, *Applied Logistic Regression*. John Wiley and Sons, Inc., 2000.

[7] S. Bagley, H. White, and B. Golomb, "Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain," *Journal of Clinical Epidemiology*, vol. 54, no. 10, pp. 979–985, 2001.

[8] J. Chhatwal, O. Alagoz, M. J. Lindstrom, C. E. Kahn, K. A. Shaffer, and E. S. Burnside, "A logistic regression model based on the national mammography database format to aid breast cancer diagnosis," *American Journal of Roentgenology*, vol. 192, no. 4, pp. 1117–1127, 2009.

[9] S. Zhang, J. S. Ivy, F. C. Payton, and K. M. Diehl, "Modeling the impact of comorbidity on breast cancer patient outcomes," *Health Care Management Science*, vol. 13, no. 2, pp. 137–154, 2010.

[10] National Cancer Institute, http://www.cancer.gov.

[11] EMOL Health, https://www.emolhealth.com/.

[12] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[13] L. Lee, W. Cheung, E. Atkinson, and M. Krzyzanowska, "Impact of comorbidity on chemotherapy use and outcomes in solid tumors: A systematic review," *Journal of Clinical Oncology*, vol. 29, no. 1, pp. 106–117, 2011.


## V. Appendix

| Name | Description |
|---|---|
| `ALP` | Alkaline Phosphatase |
| `ALT` | Alanine Aminotransferase |
| `AMGFR` | A Multiple of Glomerular Filtration Rate |
| `AST` | Aspartate aminotransferase |
| `BSA` | Body Surface Area |
| `BUN` | Blood Urea Nitrogen |
| `Calcium` | Calcium |
| `Chloride` | Chloride |
| `CO2` | Bicarbonate |
| `Creatinine` | Creatinine |
| `Diastolic` | Diastolic Blood Pressure |
| `Glucose` | Blood Glucose Level |
| `HCT` | Hematocrit |
| `Height (in)` | Height in Inches |
| `HGB` | Hemoglobin |
| `MCH` | Mean Corpuscular Hemoglobin |
| `MCHC` | Mean Corpuscular Hemoglobin Concentration |
| `MCV` | Mean Corpuscular Volume |
| `PLT` | Platelet |
| `Potassium` | Blood Serum Potassium |
| `RBC` | Red Blood Cell Count |
| `Sodium` | Blood Serum Sodium |
| `Systolic` | Systolic Blood Pressure |
| `TBILI` | Total Bilirubin |
| `Temp (F)` | Temperature F |
| `Total Protein` | Total Blood Serum Protein |
| `WBC` | White Blood Cell Count |
| `Weight (lb)` | Weight in Pounds |
| `AGE` | Age in Years |