National Surgical Adjuvant Breast and Bowel Project (NSABP) Foundation

Annual Progress Report: 2008 Formula Grant

Reporting Period

July 1, 2011 – December 31, 2011

Formula Grant Overview

The National Surgical Adjuvant Breast and Bowel Project (NSABP) Foundation, Inc. received $1,288,794 in formula funds for the grant award period January 1, 2009 through December 31, 2011. Accomplishments for the reporting period are described below.

Research Project 1: Project Title and Purpose

*Development of Prognostic Index for Colon Cancer Patients Using Gene Expression Profiling* – Currently about a half of patients with a colon cancer diagnosis receive very toxic chemotherapy regimens containing oxaliplatin. However, we know that much milder chemotherapy that does not contain oxaliplatin may be effective in preventing recurrence in a considerable proportion of colon cancer patients. A test that can identify those who do not need oxaliplatin will help many patients avoid unnecessary, toxic therapy. We will utilize a whole genome expression analyses method to screen tumor specimens collected from already finished, large clinical trials conducted by the National Surgical Adjuvant Breast and Bowel Project to develop a prognostic test to achieve this goal. Statistical rigor will be exercised in the project to assure that the test we develop will be highly reliable for actual clinical use.

Duration of Project

1/1/2009 - 12/31/2011

Project Overview

The broad research objective of the project is to improve the clinical care of patients with colon cancer diagnoses. The specific aim is to develop a prognostic test using gene expression profiling to identify patients who may not require more than 5-fluorouracil plus leucovorin (FULV) adjuvant chemotherapy. The current standard of care for Stage II or III colon cancer is adjuvant chemotherapy with an oxaliplatin-containing regimen (FULV plus oxaliplatin [FLOX or FOLFOX, based on schedule]). This standard was established by the NSABP trial C-07, which showed the superiority of FLOX over FULV. However, the FOLFOX regimen is very toxic with numerous side effects, including neurotoxicity. Data from NSABP C-07 and preceding trials conducted by our group and others make it clear that many patients may not require oxaliplatin. Yet, there are no reliable prognostic markers in clinical use to identify patients whose prognoses are good enough after treatment with only FULV such that oxaliplatin...
is not required. Developing such a prognostic test will relieve suffering from unnecessary toxic therapy and lead the way to personalized care of colon cancer patients.

Our goal is achievable because 1) we have collected tumor biopsy specimens in paraffin blocks from a majority of patients enrolled in the NSABP C-07 trial that tested oxaliplatin, and 2) we have methods to use these blocks for whole genome expression profiling at relatively low cost to allow screening of a large number of cases with statistical rigor.

We will use the Whole-Genome DASL® assay (Illumina®) to examine expression levels of 25,000 genes in tumor biopsy samples to identify genes that are prognostic for FULV-treated patients in NSABP C-07 (N=1,100). The cohort will be divided into two sub cohorts of 550 each; one will be used for gene discovery and the other for refinement of discovered genes in order to minimize false discovery. Linearity and dynamic range of the expression data for each prognostic gene and correlation between the genes will be used to further select candidate genes. We will then develop the nCounter™ assay (NanoString) to rescreen the selected prognostic genes in the same samples to develop a prognostic algorithm based on the nCounter assay. The algorithm will be prospectively tested in completely independent FULV-treated cohorts from the NSABP C-05 and C-06 trials (N=1000) not used for building of the prognostic algorithm. The statistical rigor built into the project will ensure that the developed prognostic algorithm will be a reliable clinical test for use in routine clinical practice.

Principal Investigator

Soonmyung Paik, MD
Director, Division of Pathology
NSABP Foundation, Inc.
1307 Federal St, Suite 303
Pittsburgh, PA 15212

Other Participating Researchers

Joseph P. Costantino, DrPH – employed by University of Pittsburgh

Expected Research Outcomes and Benefits

Treatment of colon cancer is evolving rapidly and many new agents are being added to what used to be considered a standard chemotherapy (FULV). However, the addition of new agents comes at the expense of increased toxicity to patients and increased health care costs. Since not all patients need such toxic treatments, identifying those who do not need these treatments is a high clinical research priority.

The prognostic index for colon cancer treated with an FULV adjuvant chemotherapy regimen developed from this project is expected to help identify patients with Stage II or III colon cancer diagnoses, who may not need more than FULV treatment and still enjoy good long-term survival. This prognostic test will result in sparing these patients from unnecessary treatment
and toxicity associated with additional therapy such as oxaliplatin, bevacizumab, or cetuximab. It will also reduce the social burden resulting from less discriminate use of the expensive agents.

**Summary of Research Completed**

During the last 6 months of this proposal, we

- developed an nCounter Colon code set3 composed of 282 prognostic and oxaliplatin-predictive genes;
- completed gene expression profiles of discovery (N=860; control and treatment arms) and validation cohorts (N=918; control and treatment arms) from C-07 tumors using the nCounter Colon code set3;
- identified statistical methods for the analysis of the discovery cohort of patients for the development and evaluation of models, with the purpose of improving the accuracy of colon cancer prognosis and providing a tool that would better define patients who would receive benefit from oxaliplatin; and
- evaluated models in the discovery cohort.

*Developed an nCounter (NanoString) code set composed of prognostic and oxaliplatin-predictive genes:*

In the last annual report, we reported on the analyses used to identify genes that were prognostic and predictive for interaction with oxaliplatin. We also performed additional analyses using the C-07 DASL data to identify genes of interest by pathway analysis within Biometric Research Branch (BRB)-Array tools. The final NanoString code set was composed of a total of 310 genes of which 14 were included as positive (6) and negative (8) controls for the NanoString assays and another 14 as housekeeping genes. The rest of the genes were included because they were identified as prognostic or oxaliplatin-predictive genes. We have not included a list of genes because there may be an interest in patenting a subset of these genes for clinical use.

*Completed gene expression profiling of discovery (N=866) and validation cohorts (N=918) from patients from C-07 using the nCounter Colon code set3:*

We completed gene expression profiling with the nCounter system using our custom nCounter Colon code set3 for 1778 cases, which consisted of all of the available and appropriate tumor blocks from clinical trial C-07. Appropriate tumor blocks included only those blocks with proper consent for use of the tissue and for which there was follow-up information available. A discovery cohort of N=860 cases was profiled.

*Identified statistical methods for the analysis of the discovery cohort of patients for the development and evaluation of models, with the purpose of improving the accuracy of colon cancer prognosis and providing a tool that would better define patients who would receive benefit from oxaliplatin:*

This data set represents a very large number of samples and the prognostic and predictive genes that have been selected for the nCounter colon code set3 have been carefully chosen. Analysis of the discovery cohort will involve a through exploratory analysis in order to identify the best
possible model that has the best chance of being validated in the validation cohort. Therefore, considerable care and time has been invested in developing a detailed protocol describing the statistical methods that will be used to identify prognostic and predictive models. An abbreviated version of the protocol describing these statistical methods is given below.

**Statistical Methods for the Building of Prognostic Models**

Because we had previously observed that in C-07 DASL data prognostic genes were distinctly different between the control arm (FULV) and the experimental (FLOX) arms, we propose to analyze the treatment arms separately and together. For the initial analysis we have built models using the combined data set.

Model 0 will be built as a prognostic model using only the information of treatment and TNM staging. Model 1-1 will be a prognostic model based on univariate selection. Each gene will be evaluated by applying the univariate Cox regression model, and the top $k$ genes with p-values less than 0.05 will be selected. Furthermore, a multivariate Cox regression will be used to select by applying forward selection with clinical covariates and $k$ genes. Model 1-2 will be built using Supervised Principal Component (SuperPC) analysis using the information of genes and other clinical covariates (treatment, sex, age, TNM staging). Genes identified as meeting the false discovery rate of 0.1 or p-value from the Cox model of less than 0.05 in the primary analysis will be used in this model. The criteria to screen genes are dependent upon the number of genes that are selected in the next stage. We have applied (SuperPC) as part of BRB-Array Tools to C-07 samples using nCounter assay data for the genes meeting entry criteria. The first 3 principle components will be used in the models. Model 1-3 which will be built with Cox univariate shrinkage (CUS). The R package uniCox will be used to build the model using CUS. Clinical covariates will be included in the model. Model 1-4 will be built with Random Forests™ (RF). R package randomSurvivalForest will be used to build the model using RF. Clinical covariates will be included in the model.

Model 2 will be built using biological information. Previous analysis of C-07 whole genome expression utilizing DASL data identified genes associated with immune reaction and cytotoxic T cells as prognostic genes almost exclusively in the control (FULV) arm (along with genes associated with apoptosis and cell division). In the oxaliplatin arm, prognostic genes were mostly related to cell division, apoptosis, membrane transport, and metabolism. Thus, we anticipate that prognostic genes in the C-07 control arm will be related to immune response, including a cytotoxic T-cell response, apoptosis, and cell division and will be preferentially selected for model building. For the oxaliplatin arm, we anticipate that prognostic genes will play a role in cell division, absorption, oxaliplatin metabolism, and apoptosis and will be preferentially selected for model building.

1. Patient subgroups will be defined using k-means cluster analysis.
2. Stepwise discriminant analysis will be used to select a subset of the quantitative variables for use in discriminating among the subgroups.
3. The canonical discriminant function will be estimated to predict each patient's subgroup. The number of the canonical discriminant function will be defined based on the amount of a given value that explains >70% of the data or eigen values larger than 1. We will build the
prognostic model using the Cox regression model with canonical correlation score calculated with the genes selected in step 2.

4. For the validation set, we will apply both canonical discriminant function and parameters in the Cox regression model estimated from the discovery set to calculate the deviance of the Cox model.

**Method Assessment of Prognostic Models**

We will apply 10-fold cross-validation to build each model. To evaluate the models, averaged deviance will be calculated for each model and the model with the smallest averaged deviance will be chosen for the validation. If two methods have comparably small average squared deviance, we can use other factors to select one.

**Plans for Building Oxaliplatin-Predictive Models**

The goal is to provide a model that will improve on the current clinical guidelines for oxaliplatin treatment recommendations. In order to understand whether the prediction models that are built with gene expression information can improve upon how patients are selected for oxaliplatin treatment, we will build a model using only clinical co-variates which will include only Stage III patients and Stage II patients if just one of the following: lymphovascular invasion or perforation or T4 lesion or grade 3-4 lesions or bowel obstruction at presentation.

Model 0 refers to the model based strictly on the above clinical variables. Based on this definition of subgrouping, we will estimate the parameter from the Cox linear regression model with an interaction term of the above indicator variable with treatment in the training set and will calculate averaged deviance and the interaction effect from the test set.

Several different prediction models for oxaliplatin benefit will also be built and evaluated using the nCounter data. We will apply 10-fold cross-validation to build each model. All statistics calculated to build predictive model will be adjusted for clinical covariates (sex, age, TNM stage).

Model 1-1 will be built based on univariate p-value assuming linear interaction effect. Genes will be discovered within the training set with p-value below 0.05 for interaction of genes and treatment in a linear model, or the top 21 (target 282 genes *0.05) genes with smaller p-values will be regarded as candidate genes if there are more than 21 genes with p-values <0.05. Then we will pick out the top k ranked selected genes using backward elimination.

Model 1-2 will be built based on univariate p-value calculated by STEPP (Subpopulation Treatment Effect Pattern Plot) analysis and categorical analysis assuming the existence of non-linear interaction effects. Discovery genes within the training set with p-value below 0.05 for interaction of genes and treatment calculated with the STEPP method and the Cox regression model with categorized expression values, or the top 21 genes with smaller p-values will be regarded as candidate genes if there are more than 21 genes with p-values <0.05. As well as model 1-1, the k top ranked selected genes will all be included in a multivariate Cox regression model using backward elimination, but the interaction effect for each gene will be assessed with optimal transformation based on STEPP analysis for each candidate gene.
Additional models for both prognosis and prediction are currently being evaluated and developed. Methods for model assessment have been detailed in our protocol here but are too long to include here. All the relevant investigators will agree on the model that will be used to validate the model by analyzing the data in the validation cohort, which consists of cases from C-07 that were not used in the development of the model. Because only one model can be tested in the validation cohort, it is essential that careful analysis of the different models is conducted so that we make the best choice possible for validation.