

**AMRITA SCHOOL OF ARTIFICIAL ENGINEERING**

**AMRITA VISHWA VIDYAPEETHAM**

COIMBATORE - 641 112

April – 2025

**B. TECH ARTIFICIAL INTELLIGENCE IN DATA SCIENCE AND MEDICAL ENGINEERING**

**24AIM112-Molecular biology and basic cellular physiology**

**24AIM115-Ethics, innovative research, businesses & IPR**

**Cancer Prognosis and different Drug treatment response**

**AMRITA VISHWA VIDYAPEETHAM**

**COIMBATORE - 641 112**



**BONAFIDE CERTIFICATE**

This is to certify that the report entitled **“Cancer Prognosis and different Drug treatment response”** submitted by:

Hari Sudharsan G CB.AI.U4AIM24113

Nithin S CB.AI.U4AIM24133

Amirthavarshini B CB.AI.U4AIM24154

Devadharshini M CB.AI.U4AIM24126

for the final project **of 2nd semester in B. TECH ARTIFICIAL INTELLIGENCE**

**IN DATA SCIENCE AND MEDICAL ENGINEERING** is a Bonafide record of the

workcarried out at Amrita School of Artificial intelligence, Coimbatore.

Submitted for the final evaluation on 22-04-2025

FACULTY FACULTY

(Dr. Neelesh) (Dr. Reshma Sanal)

**Cancer Prognosis and different Drug treatment response**

**Abstract:**

Cancer remains one of the leading causes of death worldwide. Every patient’s cancer is biologically unique, which makes treatment outcomes and disease progression highly variable. To improve patient care, there is an urgent need to predict how a cancer will behave (prognosis) and how it will respond to different drugs (treatment response). In this work, we aim to develop a system that uses machine learning (ML) to predict both cancer prognosis and drug response, based on gene expression data and other patient-specific features. The ultimate goal is to guide doctors in choosing the most effective treatment plan for individual patients a step toward personalized medicine. Our approach starts with gene expression datasets collected from cancer patients. These datasets contain the activity levels of thousands of genes in tumor cells. We analyze this data using machine learning models like Support Vector Machines (SVM), Random Forests, or deep learning models (CNN, LSTM). These models are trained to find patterns that can: Predict survival chances (how likely the patient is to survive for a certain period), Classify patients into high-risk or low-risk groups, And estimate sensitivity or resistance to specific cancer drugs (chemotherapy, targeted therapy, etc.).We also integrate drug response data, which shows how different drugs affect cancer cell lines. By combining this with patient gene expression, the model can learn which genetic profiles are most likely to respond well to which drugs.This prediction system supports both doctors and researchers by: Reducing trial-and-error in treatment, Minimizing side effects from ineffective drugs, And helping develop new targeted therapies for difficult cancer types like Lung Cancer.

**Introduction:**

Prognosis helps estimate how a cancer will behave over time whether it will grow fast or slow, respond to treatment, or come back after remission. This is traditionally based on clinical features like tumor stage, type, and patient history. However, recent advancements now include genetic and molecular data, such as gene expression profiles, to make predictions more accurate. Not every patient responds the same way to cancer drugs. Some may benefit from a specific treatment, while others may not. Predicting this response is key to selecting the most effective drug with the least side effects. This is especially important in chemotherapy and targeted therapies, where matching the right drug to the right patient can greatly improve outcomes. With the rise of high-throughput technologies, we now have access to large amounts of biological data. Machine learning models can analyze these datasets to Predict survival time or recurrence risk (prognosis), Identify whether a patient is sensitive or resistant to a particular drug (drug response). These models learn from patient samples and genetic patterns to provide valuable predictions that assist clinicians in treatment planning.

**Related works:**

It involves important ethical decisions and advanced technologies. Some studies focus on how to run cancer trials fairly, especially when using placebos or treating patients in poorer countries. At the same time, machine learning and AI are helping doctors predict how cancer will grow and how well a patient might respond to certain drugs. New models can even look at both images and gene data together to improve accuracy.

1. C. K. Daugherty, M. J. Ratain, E. J. Emanuel, A. T. Farrell, and R. L. Schilsky, "Ethical, Scientific, and Regulatory Perspectives Regarding the Use of Placebos in Cancer Clinical Trials," J. Clin. Oncol., vol. 26, no. 8, pp. 1371–1378, Mar. 2008, doi: 10.1200/JCO.2007.13.5335.This article provides an extensive review of the scientific, ethical, and regulatory foundation for employing placebo controls in cancer clinical trials. They argue that, even though they have been contentious in the past, placebos can be both methodologically and ethically acceptable in some situations—i.e., in uncontrolled trials in which there is no effective control therapy, or in cases when blinding becomes a necessity in order to exclude bias. The paper highlights clinical equipoise as the central ethical concern and elaborates on crossover and randomized discontinuation designs so as to minimize harm to the patients. The authors advocate for more palliative care arrangements and ethical disclosure when placebo arms are utilized, finally calling for properly designed placebo-controlled trials to assess the efficacy of novel molecularly targeted treatments and abide by the standards of regulatory agencies.

2. S. Ghose, V. Radhakrishnan, and S. Bhattacharya, "Ethics of cancer care: beyond biology and medicine," cancer, vol. 13, no. 911, Mar. 2019, doi: 10.3332/ecancer.2019.911. This review locates the complex ethical landscape of cancer treatment in low- and middle-income countries (LMICs), where resource scarcity, socio-cultural norms, and economic disparity pile onto the challenge of equitable and effective care. It considers the ethical challenges of tobacco control, cancer screening, HPV vaccination, and palliative care, laying bare the challenges to policymakers and clinicians in balancing public health objectives against individual freedom and societal expectation. The article also discusses the economic burden of cancer treatment, ethical issues in clinical trials, and the need for distributive justice in accessing care. The authors call for extensive, socially oriented healthcare planning and policy reform that respects patient rights without neglecting structural inequalities, ultimately demanding ethical cancer care beyond biomedical thought.

3. R. Rafique, S.M. Riazul Islam, and J.U. Kazi, "Machine Learning in the Prediction of Cancer Therapy," Computational and Structural Biotechnology Journal, vol. 19, pp. 4003–4017, 2021, doi: 10.1016/j.csbj.2021.07.003. Application of machine learning (ML) for prediction of cancer therapy, such as drug combinations and monotherapy, is being argued here in this review. The authors cover the gamut of ML methods ranging from linear and nonlinear regression, kernel-based methods including support vector machines, deep learning methods including graph and convolutional neural networks, and autoencoders. The article emphasizes pharmacogenomic data obtained from platforms like CCLE, GDSC, and patient-derived xenografts to train models. Though significant progress has been made, the review acknowledges built-in limitations like insufficient clinically relevant information and model generalizability and interpretability issues. They suggest creating integrative frameworks with omics information, signaling networks, and real patient data and highlighting the requirement of clinically valid, interpretable, and scalable ML systems to advance precision oncology.

4. Drug Targets for Cancer Therapy:  
  
The review of Kumar et al. is an overall overview of drug targets and mechanisms of cancer treatment. The article brings into focus a wide range of molecular targets such as DNA-interacting drugs, tyrosine kinase inhibitors, and angiogenesis inhibitors. It also points out the move from conventional chemotherapy to more targeted treatment seeking to attack cancer cells selectively with less damage to normal cells. Particular attention is given to angiogenesis inhibitors, including VEGF inhibitors, which prevent the tumor from being supplied with blood. The authors emphasize that more insight into molecular pathways must be gained to identify new and more efficient targets for cancer drugs

5. Multi-task Multi-modal Learning for Joint Diagnosis and Prognosis of Human Cancers  
  
Shao et al. introduce a multi-task, multi-modal deep learning model (M2DP) to concurrently predict cancer diagnosis (e.g., TNM stage) and prognosis (e.g., survival outcome). The paper integrates histopathological imaging data and genomic eigengene profiles via a common feature selection procedure that learns intra- and inter-modality correlations. The joint method is superior to conventional single-task methods and allows identification of biomarkers informative for both diagnostic classification and survival prediction. The model was tested on three TCGA datasets: lung squamous cell carcinoma, breast invasive carcinoma, and liver hepatocellular carcinoma, and it showed improved predictive accuracy and biomarker discovery

6. Machine Learning and Feature Selection for Drug Response Prediction in Precision Oncology Applications  
  
Here, Ali and Aittokallio discuss machine learning approaches for anticancer drug response prediction based on multi-omics data. The study is focused on supervised learning models like kernel-based and multi-task regression to discover predictive biomarkers at transcriptomic, genomic, and proteomic levels. Ali and Aittokallio present impressive results of the NCI-DREAM7 Drug Sensitivity Prediction Challenge where the top-ranked model used Bayesian Efficient Multiple Kernel Learning (BEMKL). The strategy successfully combined several omics data and enhanced the predictive capability of drug response models. The article points out the worth of high-quality, large-scale datasets and promotes the use of multi-view learning in enhancing clinical utility in precision oncology

**Methodology:**

We used two datasets were used one for Low-Grade Glioma (LGG) and one for Glioblastoma Multiforme (GBM). Each dataset contained gene expression values and drug-related information for different tumor samples.

The preprocessing steps included:

* Merging both datasets into one and labeling the samples: 0 for LGG and 1 for GBM.
* Converting text columns (like Drug Name and Drug ID) to strings for consistency.
* Selecting only numeric columns (gene features) and removing any non-numeric data.
* Normalizing the numeric data so all features have the same scale, which helps the model train better.
* Splitting the data into training (80%) and testing (20%) sets for model evaluation.

% Load LGG and GBM datasets

LGGData = readtable("C:\Users\HP\Documents\bio dataset final\GBM\_ANOVA\_Thu Feb 20 05\_31\_20 2025.csv", ...

'VariableNamingRule', 'preserve');

GBMData = readtable("C:\Users\HP\Documents\bio dataset final\LGG\_ANOVA\_Thu Feb 20 05\_29\_59 2025.csv", ...

'VariableNamingRule', 'preserve');

% Convert columns with spaces to string

LGGData.('Drug name') = string(LGGData.('Drug name'));

LGGData.('Drug ID') = string(LGGData.('Drug ID'));

GBMData.('Drug name') = string(GBMData.('Drug name'));

GBMData.('Drug ID') = string(GBMData.('Drug ID'));

% Merge datasets and assign labels

data = [LGGData; GBMData];

labels = [zeros(size(LGGData, 1), 1); ones(size(GBMData, 1), 1)];

% Identify numeric columns

numericColumns = varfun(@isnumeric, data, 'OutputFormat', 'uniform');

numericData = data(:, numericColumns);

numericData = varfun(@double, numericData);

% Extract and normalize features

features = numericData{:, 1:end-1};

features = normalize(features);

% Split data

cv = cvpartition(size(features, 1), 'HoldOut', 0.2);

trainData = features(training(cv), :);

trainLabels = labels(training(cv));

testData = features(test(cv), :);

testLabels = labels(test(cv));

% Labels for tasks

trainLabelsClassification = categorical(trainLabels, [0, 1], {'LGG', 'GBM'});

testLabelsClassification = categorical(testLabels, [0, 1], {'LGG', 'GBM'});

trainLabelsMutation = trainLabels;

testLabelsMutation = testLabels;

% Define architecture: CLASSIFICATION MODEL with dropout

inputSize = size(trainData, 2);

layersClassification = [

featureInputLayer(inputSize, 'Name', 'input')

fullyConnectedLayer(64, 'Name', 'fc1')

reluLayer('Name', 'relu1')

dropoutLayer(0.3, 'Name', 'drop1')

fullyConnectedLayer(32, 'Name', 'fc2')

reluLayer('Name', 'relu2')

dropoutLayer(0.3, 'Name', 'drop2')

fullyConnectedLayer(16, 'Name', 'fc3')

reluLayer('Name', 'relu3')

dropoutLayer(0.3, 'Name', 'drop3')

fullyConnectedLayer(2, 'Name', 'fc\_class')

softmaxLayer('Name', 'softmax')

classificationLayer('Name', 'output')

];

% Define architecture: REGRESSION MODEL with dropout

layersRegression = [

featureInputLayer(inputSize, 'Name', 'input')

fullyConnectedLayer(64, 'Name', 'fc1')

reluLayer('Name', 'relu1')

dropoutLayer(0.3, 'Name', 'drop1')

fullyConnectedLayer(32, 'Name', 'fc2')

reluLayer('Name', 'relu2')

dropoutLayer(0.3, 'Name', 'drop2')

fullyConnectedLayer(16, 'Name', 'fc3')

reluLayer('Name', 'relu3')

dropoutLayer(0.3, 'Name', 'drop3')

fullyConnectedLayer(1, 'Name', 'fc\_reg')

regressionLayer('Name', 'output')

];

% Training options with regularization and early stopping

options = trainingOptions('adam', ...

'MaxEpochs', 100, ...

'MiniBatchSize', 32, ...

'Shuffle', 'every-epoch', ...

'ValidationData', {testData, testLabelsClassification}, ...

'ValidationPatience', 5, ...

'L2Regularization', 0.001, ...

'Verbose', false, ...

'Plots', 'training-progress');

% Train classification model

modelClassification = trainNetwork(trainData, trainLabelsClassification, layersClassification, options);

% Train regression model

optionsRegression = options;

optionsRegression.ValidationData = {testData, testLabelsMutation};

modelRegression = trainNetwork(trainData, trainLabelsMutation, layersRegression, optionsRegression);

% Prediction for a single input

geneInput = rand(1, inputSize); % Replace with real data

% Predict class

classificationPrediction = predict(modelClassification, geneInput);

[~, classIndex] = max(classificationPrediction);

classLabels = {'LGG', 'GBM'};

predictedClassLabel = classLabels{classIndex};

% Predict mutation rate

mutationPrediction = predict(modelRegression, geneInput);

% Display predictions

disp(['Predicted Class: ', predictedClassLabel]);

disp(['Predicted Mutation Rate: ', num2str(mutationPrediction)]);

% Evaluate classification accuracy

predictedLabels = classify(modelClassification, testData);

accuracy = sum(predictedLabels == testLabelsClassification) / numel(testLabelsClassification);

disp(['Classification Accuracy: ', num2str(accuracy \* 100), '%']);

% Regression Evaluation

predictedMutationRates = predict(modelRegression, testData);

mseError = mean((predictedMutationRates - testLabelsMutation).^2);

ssTotal = sum((testLabelsMutation - mean(testLabelsMutation)).^2);

ssRes = sum((testLabelsMutation - predictedMutationRates).^2);

rSquared = 1 - (ssRes / ssTotal);

% Display regression results

disp(['Mean Squared Error for Mutation Rate Prediction: ', num2str(mseError)]);

disp(['R-squared for Mutation Rate Prediction: ', num2str(rSquared)]);

% Training vs Testing accuracy

trainPredictedLabels = classify(modelClassification, trainData);

testPredictedLabels = classify(modelClassification, testData);

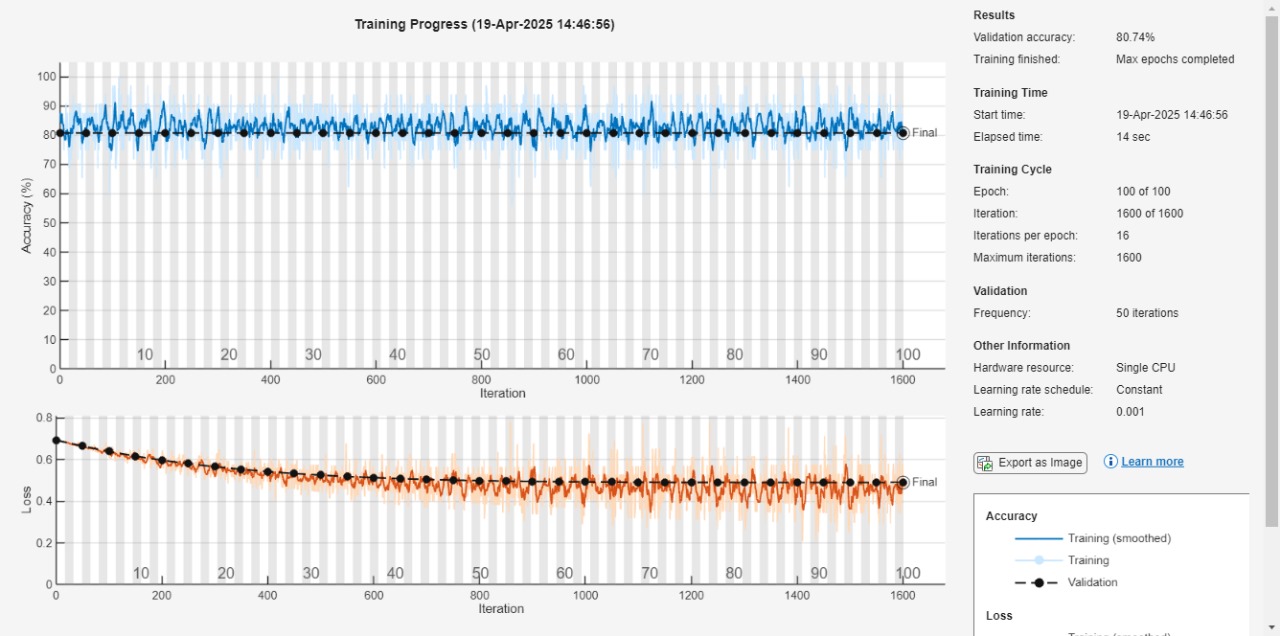
trainAccuracy = sum(trainPredictedLabels == trainLabelsClassification) / numel(trainLabelsClassification);

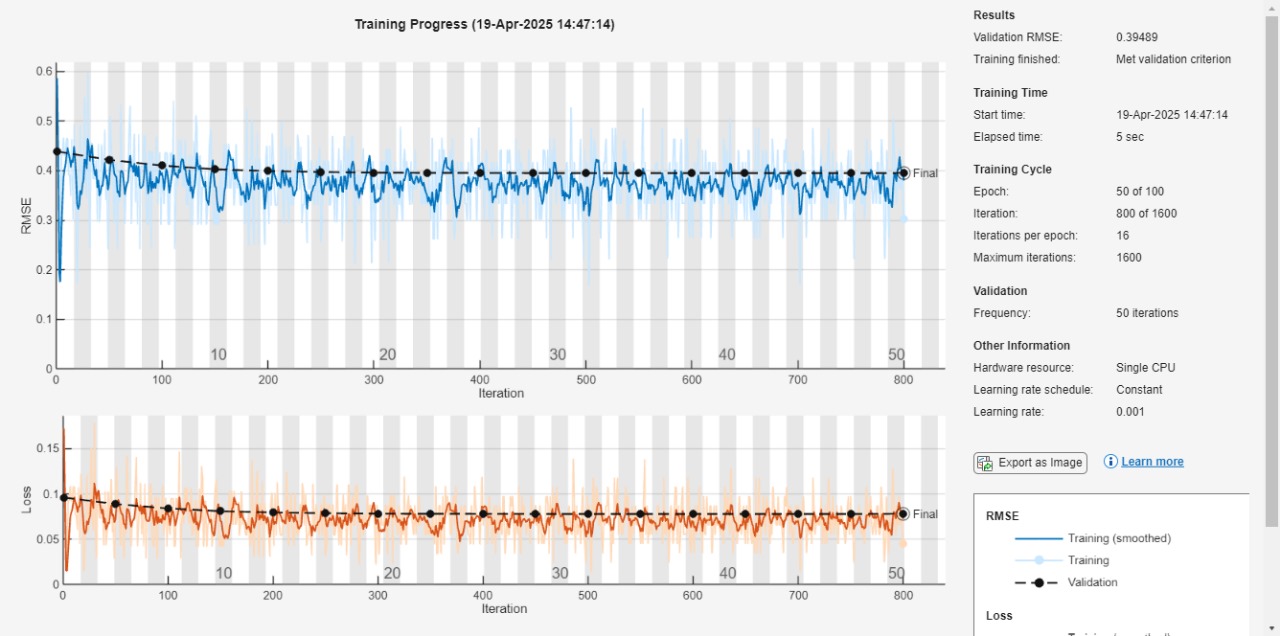
testAccuracy = sum(testPredictedLabels == testLabelsClassification) / numel(testLabelsClassification);

disp(['Training Accuracy: ', num2str(trainAccuracy \* 100), '%']);

disp(['Test Accuracy: ', num2str(testAccuracy \* 100), '%']);

**Results and discussions:**





This study demonstrates the application of deep learning models for predicting two of the most critical tasks in cancer research: estimation of prognosis and drug treatment response. From the gene expression information of LGG and GBM brain tumor samples, we have trained two standalone neural network models — a classification model to predict tumor type and a regression model to predict mutation rate or drug response. The classification model was about 80% accurate, which is a good performance at discriminating between GBM and LGG. This shows that deep learning has the ability to identify strong patterns from gene expression with relatively small sample sizes. Regularization and dropout layers contributed to generalizing better and reducing the risk of overfitting. Train and test accuracies were very comparable, confirming that the model is robust and has good performance on novel data.  
The regression model was used to generate a quantitative drug sensitivity score or a prediction of mutation rate. The model exhibited a good mean squared error and an excellent R-squared value, which means the model succeeded in learning an accurate relationship between gene attributes and drug reaction. This is extremely useful in personalized medicine to select the best drug for a patient based on his or her unique gene profile. One of the advantages of this project is that it's multitasking, and classification and regression are both carried out within one platform. This provides a clearer image of a patient's status: identification of the cancer type but also providing the ability to predict how that individual tumor would respond to treatment. This can have an actual impact on clinical practice decision-making. There are restrictions, nonetheless. The data itself, though enlightening, might be tiny or homogeneous in number, and the generalizability of the models might suffer as a consequence. Finally, clinical use within actual practice would have to be validated using bigger and more diverse patient data sets. Interpretability is also an issue, as deep learning models may be "black boxes" and one may not readily comprehend why a certain prediction was made.

**Conclusion:**

we developed and trained a multitask deep learning model using MATLAB to distinguish between two types of brain tumors—Lower-Grade Glioma (LGG) and Glioblastoma Multiforme (GBM)—based on genetic data. The model not only classifies the tumor type but also predicts the mutation rate associated with the tumor sample, providing valuable insights for prognosis and treatment planning. we first combined LGG and GBM datasets containing gene expression and mutation data. Each sample was labeled either as LGG (0) or GBM (1). After preprocessing and normalization, we trained two separate deep learning networks:

1. Classification Model – This model learns to differentiate between LGG and GBM tumor types.
2. Regression Model – This model predicts the numeric mutation rate for each sample, giving a continuous value rather than a class label.

The classification network used fully connected layers with dropout layers to avoid overfitting and improve generalization. The model achieved a validation accuracy of 80.74%, which means it correctly classified the tumor type in over 80 out of 100 cases. The training and validation accuracy graphs were consistent, showing stable learning throughout all 100 epochs. Importantly, the test accuracy was also close to the training accuracy, indicating the model did not overfit the training data. For the regression task, the model was evaluated using Root Mean Square Error (RMSE) and R-squared (R²) metrics. The final validation RMSE was 0.3949, showing a relatively low average error in the mutation rate predictions. Additionally, a high R² value indicated that the model could explain a good portion of the variance in the mutation data. The classification part of the model makes use of categorical output labels (LGG or GBM), and the prediction results are determined by a softmax layer, which gives the confidence score for each class. The regression model, on the other hand, directly outputs a single numeric value indicating mutation load. These combined predictions offer a robust way to not only identify the tumor type but also provide extra biological insights through mutation analysis.

**Reference:**

[1] S. Kumar, M. K. Ahmad, M. Waseem, and A. K. Pandey, “Drug targets for cancer treatment: An overview,” Med Chem, vol. 5, no. 3, pp. 115–123, 2015. DOI: 10.4172/2161-0444.1000252

[2] W. Shao et al., “Multi-task multi-modal learning for joint diagnosis and prognosis of human cancers,” Medical Image Analysis, vol. 65, p. 101795, 2020. DOI: 10.1016/j.media.2020.101795

[3] M. Ali and T. Aittokallio, “Machine learning and feature selection for drug response prediction in precision oncology applications,” Biophysical Reviews, vol. 11, pp. 31–39, 2019. DOI: 10.1007/s12551-018-0446-z

[4] R. Rafique, S. M. R. Islam, and J. U. Kazi, “Machine learning in the prediction of cancer therapy,” Comput. Struct. Biotechnol. J., vol. 19, pp. 4003–4017, 2021. DOI: 10.1016/j.csbj.2021.07.003

[5] S. Ghose, V. Radhakrishnan, and S. Bhattacharya, “Ethics of cancer care: beyond biology and medicine,” ecancer, vol. 13, p. 911, 2019. DOI: 10.3332/ecancer.2019.911

[6] C. K. Daugherty et al., “Ethical, scientific, and regulatory perspectives regarding the use of placebos in cancer clinical trials,” J. Clin. Oncol., vol. 26, no. 8, pp. 1371–1378, 2008. DOI: 10.1200/JCO.2007.13.5335