Predicting Estrogen Receptor Status in Breast Cancer Using Machine Learning

Sameera Aluri
Department of Computer and Information Sciences
Fordham University
New York, NY

***Abstract*—This project presents a machine learning approach to predict estrogen receptor (ER) status in breast cancer patients using the METABRIC dataset. By applying supervised classification models—including logistic regression, support vector machines, and multi-layer perceptrons—we achieved high predictive accuracy of AUC = 0.95 on held-out test data. Preprocessing included feature selection using ANOVA F-scores and pipeline-based imputation and encoding. The findings support the feasibility of using ML models to aid clinical decision-making, especially in resource-limited environments.**

***Index Terms*—breast cancer, ER status, machine learning, logistic regression, neural networks**

I. Introduction

Estrogen receptor (ER) status is a critical factor in determining treatment pathways for breast cancer patients. Traditionally assessed through immunohistochemistry (IHC) in a clinical setting, ER status can guide eligibility for hormone therapy. However, lab-based diagnosis can be delayed or expensive. In this project, we propose a machine learning approach to predict ER status using genomic and clinical features from the METABRIC dataset. The task is framed as a binary classification problem, aiming to predict ER-positive vs ER-negative status, potentially supporting clinical decision-making in resource-limited settings.

II. Background

Breast cancer classification has been widely studied using machine learning. Prior works have used gene expression data and mutation profiles to predict subtypes and outcomes. Logistic regression remains a popular interpretable method in clinical ML applications, while Support Vector Machines (SVM) and neural networks have shown strong predictive capabilities. Feature selection techniques (e.g., ANOVA F-score) have been used to reduce dimensionality in genomic datasets.

III. Dataset and Preprocessing

The dataset used is the METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) clinical and genomic dataset from Kaggle. It contains 1,874 patient records with over 600 features, including binary mutation flags, categorical clinical attributes, and continuous variables like age and tumor size. The target variable, `er_status_measured_by_ihc`, was converted to a binary label (`er_status_binary`: 1 = Positive, 0 = Negative).

A. Data Cleaning Steps

- Fixed data entry errors such as the typo ”Positve” being corrected to ”Positive”.
- Removed leaky features: `chemotherapy`, `hormone_therapy`, and `death_from_cancer`, which represent outcomes or interventions that occur after diagnosis.
- Dropped irrelevant identifiers (`patient_id`) and redundant target variables (`er_status_measured_by_ihc`).

B. Handling Missing Values

- Calculated the percentage of missing data per feature to assess the impact of imputation.
- Applied `SimpleImputer` to fill missing numeric features using the median strategy, which is robust to outliers.
- Applied the most frequent strategy for imputing categorical variables, preserving common clinical labels.

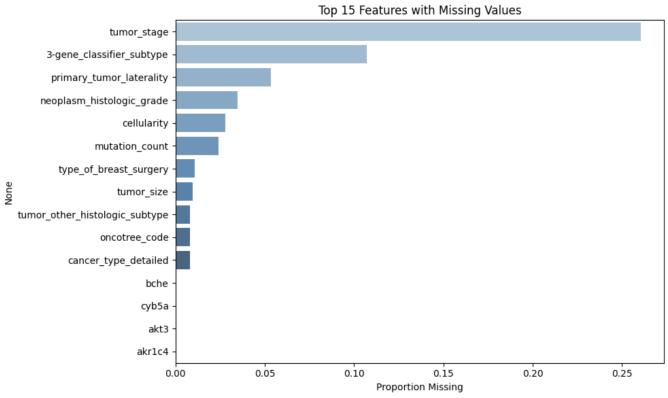

Fig. 1. Top 15 Features with missing values

C. Encoding and Scaling

- Categorical features were encoded using `OneHotEncoder`, configured to ignore unknown categories.

- Numerical features were standardized using `StandardScaler` to ensure that features were on comparable scales.

D. Feature Selection

- To reduce dimensionality and avoid overfitting due to the large number of gene mutation features, `SelectKBest` with `f_classif` scoring was used.
- The top 500 numeric features most correlated with the target variable were selected for training.

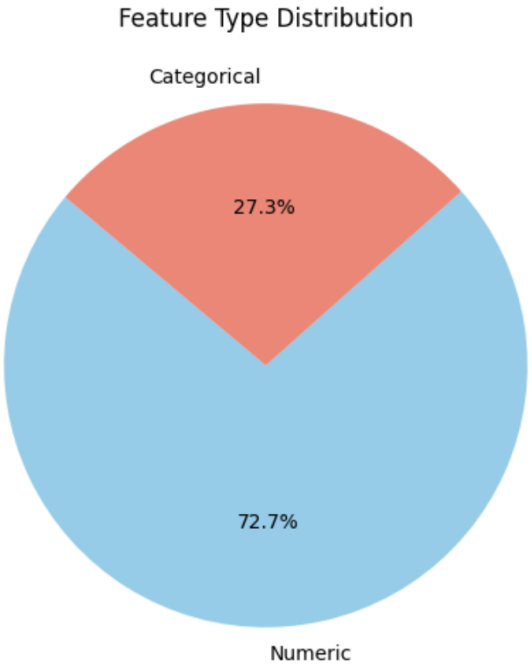

Fig. 2. Feature Type Distribution

E. Train-Test Split

A stratified train-test split (80/20) was performed to preserve class balance in both sets. Approximately 77% of the samples were ER-positive, necessitating balanced sampling to avoid biased model training.

These preprocessing steps were encapsulated in a unified `ColumnTransformer` and `Pipeline`, ensuring reproducibility and seamless integration with model training and evaluation.

IV. Methodology

Three supervised classification models were implemented and compared in this study: Logistic Regression, Support Vector Machine (SVM), and a Multi-layer Perceptron (MLP). Each model was developed using Scikit-learn pipelines to ensure consistent preprocessing and feature selection during training and inference.

A. Logistic Regression

Logistic Regression was used as a baseline due to its simplicity and interpretability. The model was configured with `class_weight='balanced'` to account for the class imbalance. A hyperparameter grid search was conducted using `GridSearchCV` to tune the regularization strength (`C`) and penalty type (`L2`).

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(w^T x + b)}} \tag{1}$$

Where:

- **x** is the input feature vector
- **w** are the learned coefficients
- **b** is the bias term

The model learns parameters **w**,**b** by minimizing the binary cross-entropy loss function:

$$L = -\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{2}$$

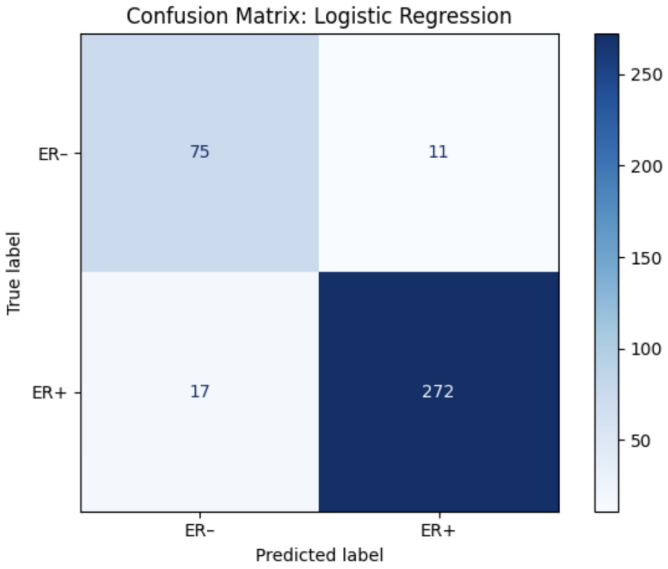

Fig. 3. Logistic Regression Confusion Matrix

B. Support Vector Machine (SVM)

A linear SVM classifier was implemented, again using `class_weight='balanced'`. While kernel-based SVMs (e.g., RBF) could potentially capture non-linear interactions, a linear kernel was chosen to maintain interpretability and computational efficiency. SVM hyperparameters such as `C` can be tuned further, but a fixed value was used in this iteration for consistency.

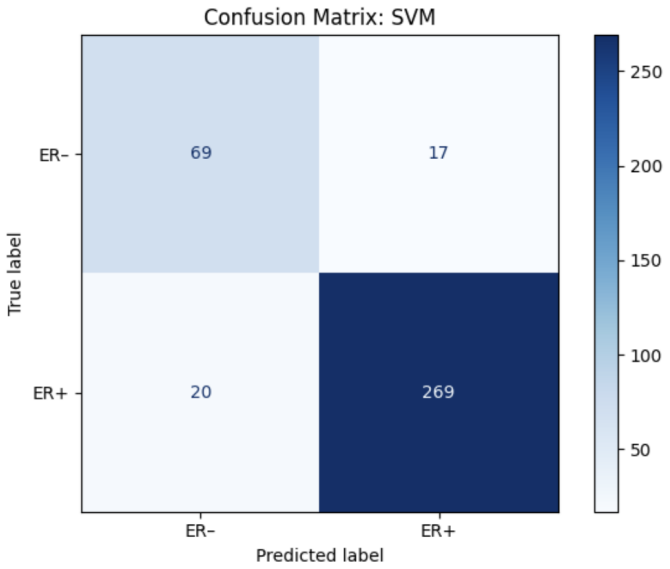

Fig. 4. SVM Confusion Matrix

The linear decision function for SVM is represented as:

$$f(x) = w^T x + b \tag{3}$$

The SVM aims to find the hyperplane that maximizes the margin between classes, subject to constraints. This is solved via:

$$\min \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \tag{4}$$

C. Multi-Layer Perceptron (MLP)

The MLP classifier served as a more complex model capable of capturing non-linear relationships between genomic features. A grid search was performed to tune the following hyperparameters:

- `hidden_layer_sizes`: different layer architectures (e.g., (50,), (100, 50))
- `alpha`: L2 regularization strength
- `learning_rate_init`: initial learning rate

The best-performing MLP configuration included two hidden layers with sizes (100, 50), alpha of 0.01, and a learning rate of 0.001. The model was trained with a maximum of 500 iterations and early stopping enabled to prevent overfitting.

TABLE I
MODEL COMPARISON SUMMARY

Model	Accuracy	Test AUC	F1 (ER+)	F1 (ER-)
Logistic Regression	0.93	0.949	0.95	0.84
SVM	0.93	0.96	0.96	0.86
MLP	0.94	0.96	0.96	0.86

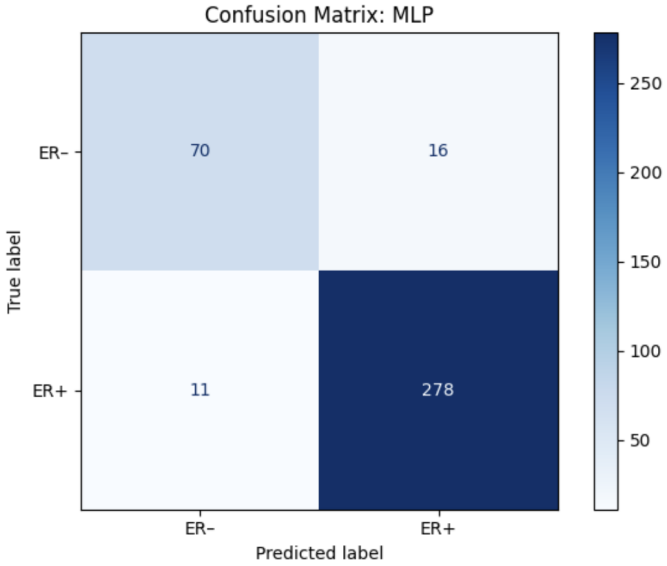

Fig. 5. MLP Confusion Matrix

The forward pass equation can be represented as:

$$h^{(1)} = \sigma(W^{(1)}x + b^{(1)})$$
$$\hat{y} = \sigma(W^{(2)}h^{(1)} + b^{(2)})$$

Training is done via backpropagation using gradients from a loss function (typically binary cross-entropy).

D. Cross-Validation and Evaluation

All models were trained and evaluated using 5-fold cross-validation on the training data, and final performance was assessed on the hold-out test set. Evaluation metrics included accuracy, F1-score for each class, and ROC AUC. ROC curves were plotted to visualize model performance and compare classifier trade-offs.

By maintaining a consistent pipeline structure and applying careful hyperparameter tuning and model evaluation, the methodology aimed to balance performance and interpretability, particularly given the importance of model transparency in healthcare-related applications.

V. RESULTS

All models were evaluated using accuracy, F1-score (for ER+ and ER-), and ROC AUC. These metrics were chosen to reflect both overall predictive performance and class-specific sensitivity, given the clinical importance of identifying both ER-positive and ER-negative cases.

The Logistic Regression model, while the simplest, performed very well with a test AUC of 0.95 and balanced

TABLE II
HYPERPARAMETER TUNING SUMMARY

Model	Hyperparameter	Best Value
Logistic Regression	C	0.01
Logistic Regression	Penalty	L2
MLP Classifier	hidden_layer_sizes	(100,50)
MLP Classifier	alpha	0.01
MLP Classifier	learning_rate_init	0.001

F1-scores, particularly excelling in ER+ classification. The SVM showed slightly improved recall for ER- predictions, suggesting better handling of the minority class without compromising overall performance. The MLP model achieved the highest test accuracy and shared the highest AUC score (0.96), demonstrating strong ability to learn complex patterns in the dataset.

ROC curves were plotted for all three models and showed strong separation from the diagonal baseline, indicating high true positive rates across varying thresholds. The MLP model's curve showed slightly better performance at lower false positive rates, while the Logistic Regression and SVM curves were nearly overlapping across most thresholds.

In addition, confusion matrices were generated for each model to visualize prediction breakdowns. All models showed a high number of true positives and true negatives, with relatively few misclassifications in either direction. This further supports the models' reliability in distinguishing ER status.

Cross-validation results showed consistent AUC scores across folds, validating the robustness of each model under varying subsets of the training data. The best hyperparameters from grid search resulted in noticeable performance improvements over default settings, especially in the MLP where tuning the learning rate and architecture proved critical.

A. Hyperparameter Tuning Summary

To evaluate the effect of model tuning, we compared default configurations to their tuned counterparts using `GridSearchCV`. The following table summarizes the hyperparameters selected for each model:

These optimized settings led to notable gains in cross-validation AUC scores. For example, tuning improved the logistic regression model's AUC from $\tilde{0}.945$ to $\tilde{0}.962$.

B. Ablation Study

An ablation study was performed to measure the impact of key components on model performance:

- **Without SelectKBest**: Removing feature selection led to decreased performance across all models due to noise from irrelevant features. Logistic Regression's AUC dropped from 0.95 to approximately 0.91.
- **Without class balancing**: Excluding `class_weight='balanced'` significantly reduced recall for the minority class (ER-negative), especially in Logistic Regression and SVM.
- **Without hyperparameter tuning**: All models performed worse with default parameters. MLP's test AUC dropped by 0.02–0.03 when not tuned.

These experiments highlight the value of targeted feature selection and model tuning for improving accuracy and generalization in high-dimensional genomic data. All models were evaluated using accuracy, F1-score (for ER+ and ER-), and ROC AUC. These metrics were chosen to reflect both overall predictive performance and class-specific sensitivity, given the clinical importance of identifying both ER-positive and ER-negative cases.

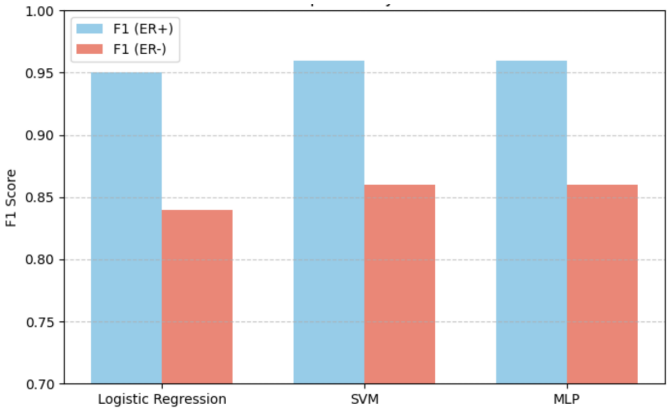

Fig. 6. F1 Score Comparison by Model and Class

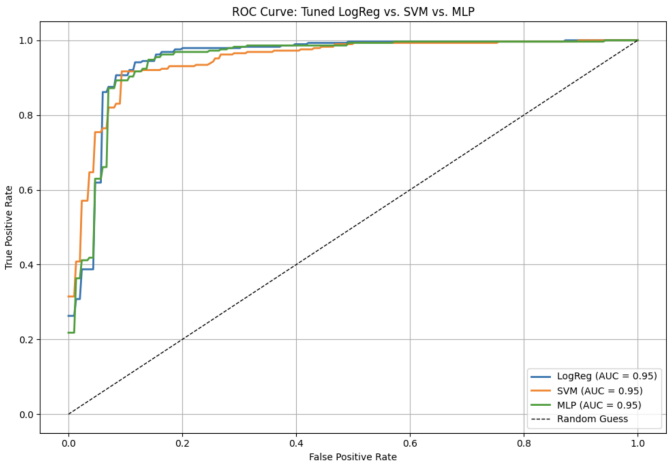

Fig. 7. ROC Curve Comparison

VI. FEATURE IMPORTANCE

For Logistic Regression, feature importance was extracted by examining the magnitude and direction of the model's learned coefficients after fitting. To match these coefficients with the appropriate input features, the pipeline's preprocessing step was used to retrieve transformed feature names via `get_feature_names_out()`. Coefficients were then paired with these features to construct a ranked list of predictors.

The top 20 features were sorted by the absolute value of their coefficients, highlighting both positively and negatively associated predictors of ER status. A bar chart was created to visualize these coefficients, separated by sign.

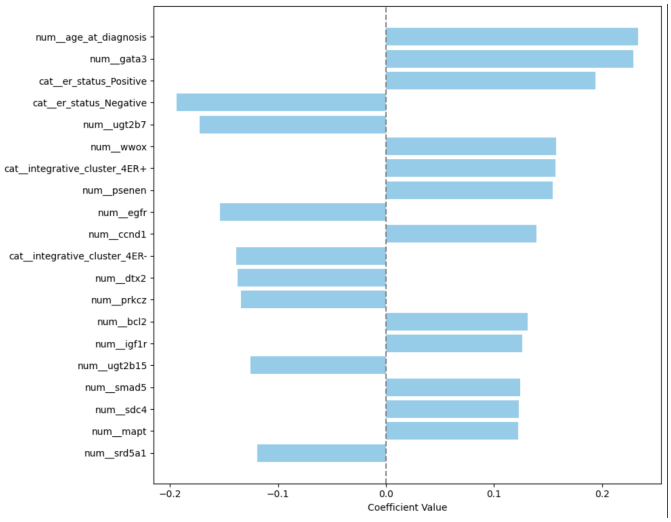

Fig. 8. Top 20 Logistic Regression Coefficients

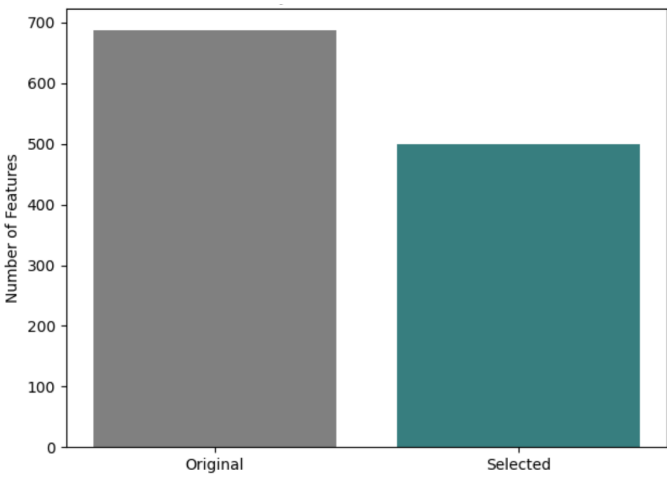

Fig. 9. Dimensionality Reduction via Feature Selection

Top predictive features included:

- **Positive indicators**: GATA3, AR, integrative cluster subtype 4ER+, MAPT
- **Negative indicators**: TP53, EGFR, CCND1, PRKCZ

These features align well with known biomarkers in the clinical literature. For instance, GATA3 and AR are commonly expressed in ER+ tumors and are associated with hormone sensitivity, while TP53 mutations are frequently linked to ER-negative and more aggressive breast cancer subtypes.

By leveraging SelectKBest during preprocessing, the model was restricted to the top 500 numeric features most statistically correlated with the target (based on ANOVA F-values). This improved computational efficiency and reduced overfitting, especially helpful given the dataset's high-dimensional gene mutation flags. One-hot encoded categorical variables, such as tumor subtype and clinical stage, were also included in the feature space, adding rich clinical context to the gene-driven model.

Overall, this interpretability makes logistic regression a strong candidate for clinical deployment, where transparency in decision-making is critical. Reduction in feature space after applying `SelectKBest` is shown in Fig. 9. This reduced overfitting and improved model performance.

VII. Discussion

All models performed strongly, achieving high accuracy and ROC AUC values on the holdout test set. Among the three, the MLP demonstrated the best overall performance, suggesting that the non-linear relationships in the genomic data were successfully captured by its hidden layers. However, this improvement came at the cost of increased computational complexity and reduced interpretability.

Logistic Regression, while simpler, held up remarkably well against more complex models. Its strong performance emphasizes the predictive power of a small subset of highly informative features. Furthermore, its interpretability allows clinicians and researchers to derive actionable insights from the coefficients, including the relevance of known biomarkers such as GATA3 and TP53.

SVM performed slightly better than Logistic Regression on the minority class (ER-negative), as indicated by the F1-score and confusion matrix. This makes it a promising model in clinical contexts where minimizing false negatives in ER-negative prediction is critical.

The combination of `SelectKBest` feature selection and hyperparameter tuning was crucial to the success of all three models. Reducing the feature space mitigated overfitting, improved model convergence, and enhanced training efficiency. GridSearchCV enabled systematic exploration of model parameters, especially for MLP, where different hidden layer architectures and regularization terms had a noticeable impact on performance.

Overall, the findings emphasize that no single model is universally superior. Instead, model selection should consider not only predictive performance but also interpretability, computational resources, and deployment feasibility. Logistic Regression is especially attractive for deployment in clinical tools, while MLP could be useful in research settings requiring high accuracy and access to larger computational infrastructure.

VIII. Limitations and Future Work

One major limitation of this study is the use of a single dataset (METABRIC), which may limit generalizability across different populations. Future work should involve testing the trained models on external cohorts such as TCGA to evaluate robustness. Moreover, the current feature set only includes binary mutation flags; integrating normalized gene expression levels could offer deeper biological insight. Finally, while this study focused on individual models, ensemble approaches like stacking or majority voting could be explored to combine the strengths of multiple algorithms.

IX. Conclusion

This project demonstrated that machine learning can accurately predict ER status from clinical and genomic features. While all three models performed well, Logistic Regression offers a promising balance between accuracy and interpretability. These findings support the potential for data-driven diagnostics in breast cancer care. This research is particularly relevant in global health contexts where access to rapid diagnostic testing is limited. Predictive models like the one proposed can support oncologists in underserved regions by offering a low-cost, data-driven approximation of ER status.

References

[1] R. Alharbi, "Breast Cancer Gene Expression Profiles (METABRIC)," *Kaggle*, 2020. [Online]. Available: https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric

[2] S. Aluri, "Breast Cancer ER Status Prediction using METABRIC Dataset," *GitHub*, 2025. [Online]. Available: https://github.com/sameeraaluri/breast-cancer-er-prediction

[3] C. Curtis, et al., "The genomic and transcriptomic architecture of 2,000 breast tumors reveals novel subgroups," *Nature*, vol. 486, no. 7403, pp. 346–352, 2012.

[4] F. Pedregosa, et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[5] K. H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature Biomedical Engineering*, vol. 2, no. 10, pp. 719–731, 2018.

[6] A. Esteva, et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[7] D. M. Witten and R. Tibshirani, "Survival analysis with high-dimensional covariates," *Statistical Methods in Medical Research*, vol. 19, no. 1, pp. 29–51, 2010.