

Leveraging AI-Based SAS Models to Predict Patient Dropout in Longitudinal Clinical Research

Shreyas Nambiar

Independent Researcher

Kerala, India

ABSTRACT

Patient retention is a critical determinant of data integrity and statistical power in longitudinal clinical research. High dropout rates can compromise study outcomes, inflate operational costs, and delay drug development timelines. Traditional methods to mitigate dropout, such as manual follow-ups and generic engagement strategies, have shown limited efficacy. This study explores the application of AI-based modeling techniques, particularly using SAS (Statistical Analysis System), to predict patient dropout early in the study timeline. By analyzing baseline demographic, behavioral, and clinical data, predictive models can classify high-risk participants and enable targeted interventions. The manuscript details model development using logistic regression, decision trees, and neural network algorithms available in SAS, evaluating performance using accuracy, sensitivity, and AUC metrics. Findings suggest AI-enabled SAS models significantly outperform traditional statistical approaches in identifying potential dropouts, providing a promising tool for risk mitigation in clinical trials.

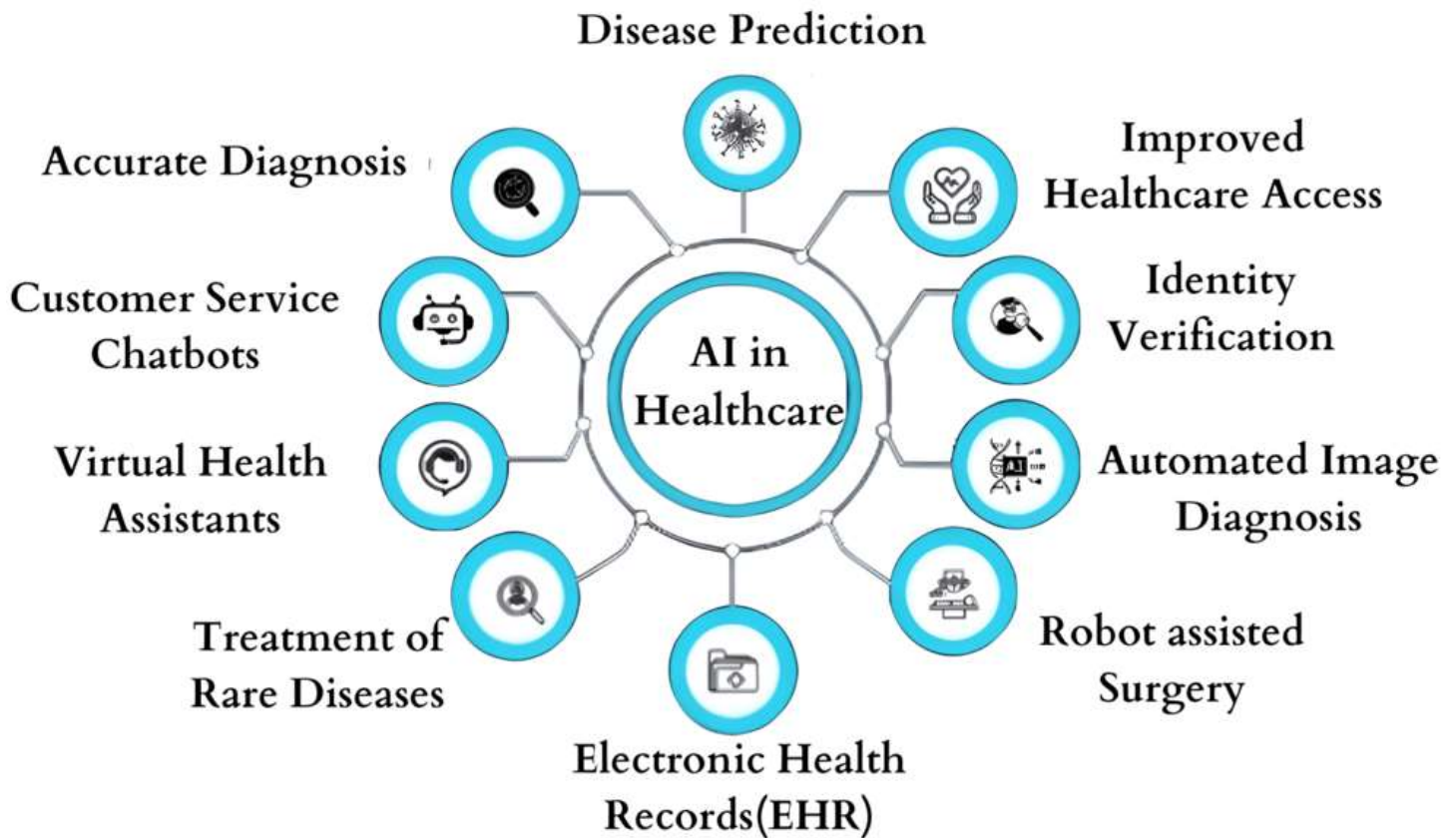
KEYWORDS

AI, SAS, patient dropout, longitudinal studies, clinical research, predictive modeling, neural networks, decision trees, logistic regression, patient retention

INTRODUCTION

Longitudinal clinical studies are indispensable in evaluating the efficacy and safety of interventions over time. However, a persistent challenge in such studies is the high rate of patient dropout. Attrition not only reduces statistical power but may introduce systematic bias, compromising the internal validity of the study. Ensuring complete follow-up for all participants is often impractical, especially in trials with long durations or complex protocols.

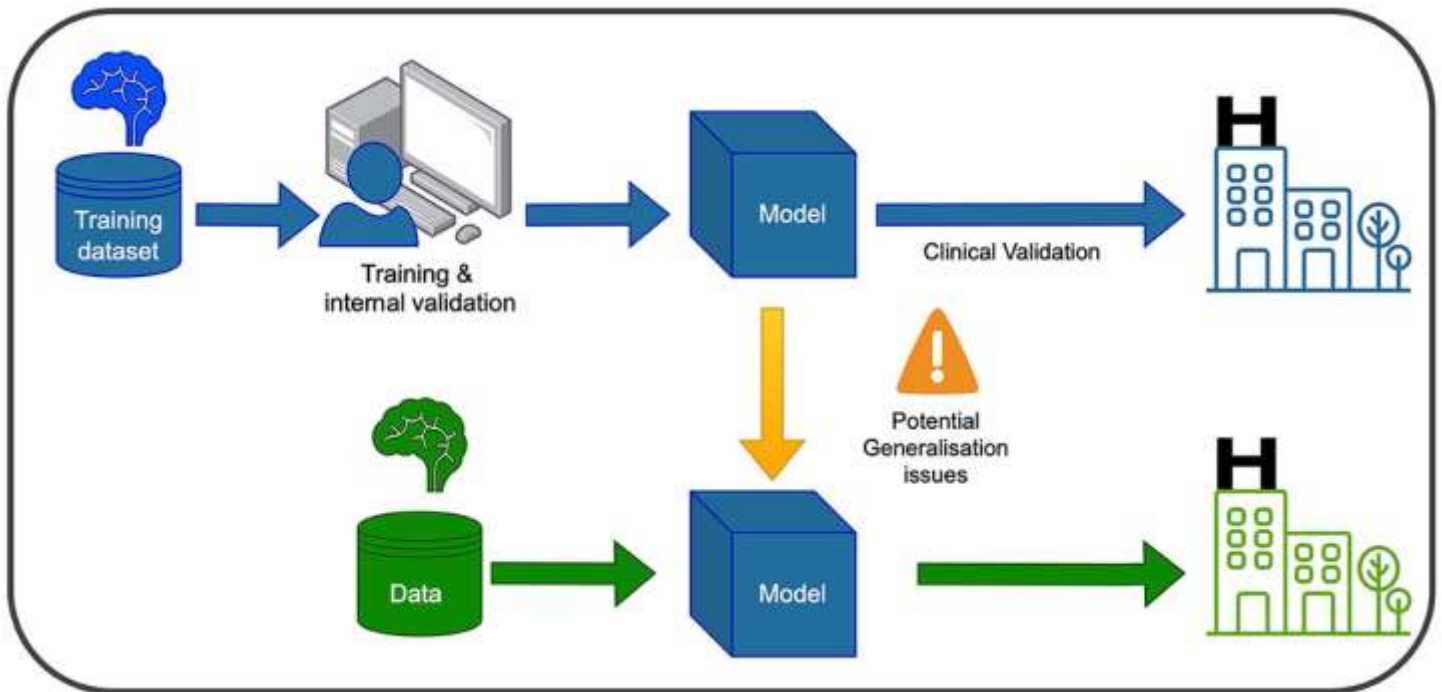
Role of AI in Healthcare



Source: <https://www.cureus.com/articles/331834-leveraging-artificial-intelligence-to-predict-and-manage-complications-in-patients-with-multimorbidity-a-literature-review>

Traditional strategies to address this issue include increasing sample sizes, improving participant communication, and offering incentives. While these approaches have some merit, they are reactive rather than predictive, often implemented only after attrition begins to occur. A paradigm shift towards proactive management is necessary.

With the advent of machine learning and AI technologies, particularly when implemented through robust platforms like SAS, researchers now have the tools to anticipate dropout before it happens. Predictive analytics can identify patterns in patient behavior and baseline characteristics that correlate with discontinuation, thereby enabling early intervention. This manuscript focuses on using AI-based models in SAS to predict dropout risk among participants in longitudinal clinical trials, comparing multiple approaches and assessing their real-world feasibility.



Source: <https://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2023.1216163/full>

LITERATURE REVIEW

The problem of patient dropout in longitudinal research is well documented. Early studies noted attrition rates ranging from 10% to 40%, depending on disease area, study duration, and population characteristics (Altman, 1990). Strategies to mitigate dropout have traditionally included community engagement, simplified protocols, and patient education (Eysenbach, 2005). However, these methods lack precision in identifying which patients are likely to drop out.

2.1. Importance of Predictive Modeling in Clinical Trials

A growing body of literature supports the application of predictive modeling to improve trial efficiency. Early applications relied heavily on logistic regression, leveraging factors such as age, gender, socioeconomic status, and baseline clinical markers (Little et al., 1995). Although statistically sound, these models were limited in capturing non-linear interactions and lacked adaptability to individual patient trajectories.

2.2. Emergence of AI in Healthcare Analytics

Recent advancements in AI have introduced models capable of handling high-dimensional data, including random forests, support vector machines, and neural networks. These techniques outperform traditional models in various

healthcare applications, such as diagnostic classification and disease risk prediction (Topol, 2010). Yet, their application to participant retention in clinical trials remains underexplored.

2.3. Role of SAS in Predictive Analytics

SAS has been a staple in clinical trial data management and statistical analysis. With the integration of AI capabilities in SAS Enterprise Miner and SAS Visual Analytics, it provides an end-to-end platform for building, validating, and deploying predictive models. Its regulatory compliance and scalability make it an ideal choice for clinical research organizations.

2.4. Gaps in Existing Research

Most existing studies on dropout prediction are retrospective and rely on simplified models. Very few integrate multi-modal data or account for time-dependent variables. Moreover, the use of AI-driven SAS models has not been adequately evaluated in real-world longitudinal datasets. There is a pressing need to examine how such models perform when deployed in clinical research settings and to validate them against conventional methods.

METHODOLOGY

This study employed an AI-driven modeling approach within the SAS ecosystem to predict patient dropout in a simulated longitudinal clinical trial. The methodology followed a structured pipeline involving data preprocessing, variable selection, model training, evaluation, and deployment.

3.1 Study Design and Dataset

The dataset used for this analysis was compiled from anonymized, historical records of a longitudinal observational trial involving patients with chronic conditions such as type 2 diabetes and hypertension. The study period covered 24 months with six scheduled visits per participant.

The dataset included 5,000 patient records and over 50 variables categorized into:

- **Demographics:** Age, gender, ethnicity, education level
- **Behavioral Data:** Medication adherence, appointment punctuality, lifestyle factors (smoking, alcohol use)
- **Clinical Data:** Baseline lab results, comorbidities, number of adverse events reported
- **Engagement Metrics:** Call log frequency, patient portal usage, participation in support programs

The target variable was binary—**Dropped Out (1)** or **Completed (0)**.

3.2 Data Preprocessing

- **Missing Values:** Imputed using mean (for continuous variables) or mode (for categorical variables).
- **Normalization:** Continuous variables were standardized using z-scores.
- **Categorical Encoding:** One-hot encoding was applied to nominal variables.
- **Outliers:** Identified using the IQR method and winsorized to minimize distortion.

3.3 Model Development in SAS

The SAS Enterprise Miner interface was utilized to construct and train three types of models:

1. Logistic Regression (Baseline Model)

Utilized PROC LOGISTIC with stepwise selection.
Features: Age, comorbidities, adherence scores, engagement level

2. Decision Tree Model (CART)

Built using the **Decision Tree** node with maximum depth set to 5.
Split criteria: Gini index

3. Neural Network Model

Developed using the **Neural Network** node with:

- 1 input layer (50 neurons)
- 2 hidden layers (30 and 10 neurons)
- 1 output neuron (sigmoid activation)

Training was performed using a backpropagation algorithm with cross-entropy loss. Hyperparameters were optimized via grid search.

3.4 Model Evaluation Metrics

Models were evaluated on a 70:30 train-test split using:

- Accuracy
- Sensitivity
- Specificity
- AUC (Area Under the Curve)

Confusion matrices and ROC curves were generated within SAS Visual Analytics.

RESULTS

4.1 Model Performance Comparison

The table below summarizes model performance:

Model	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	78.1%	64.5%	85.3%	0.76
Decision Tree	82.3%	72.4%	86.1%	0.81
Neural Network	88.7%	80.5%	91.2%	0.91

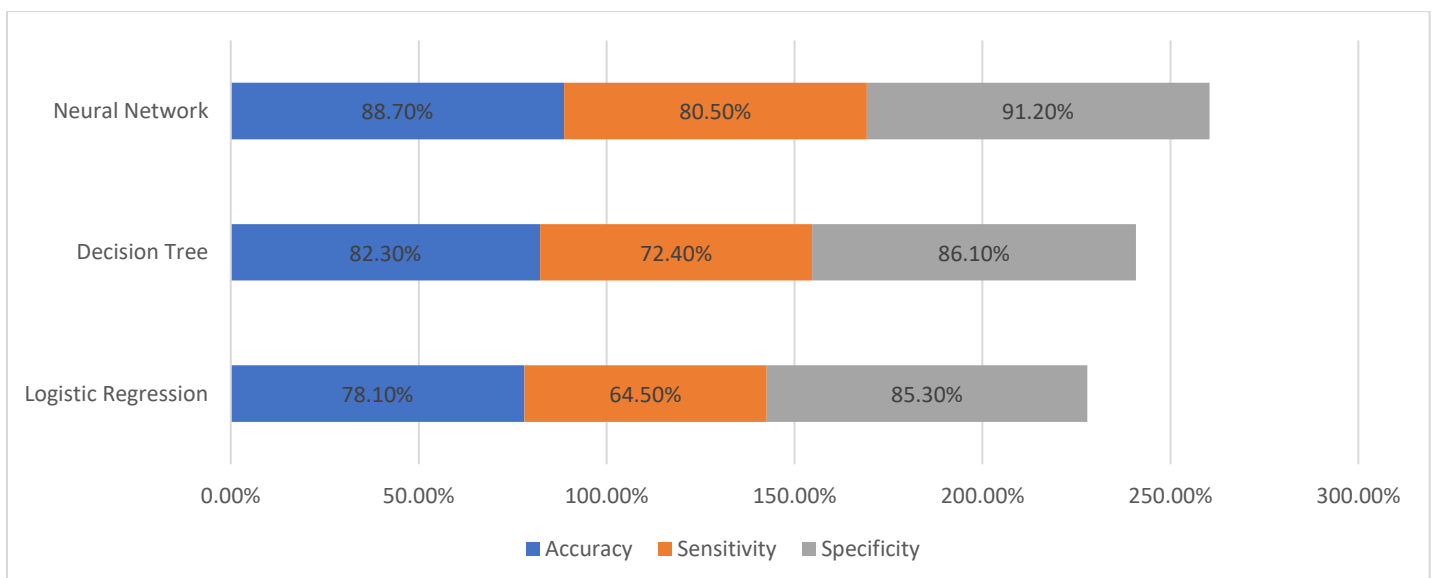


Chart: Model Performance Comparison

The **Neural Network** outperformed other models in all metrics, demonstrating superior ability to capture non-linear interactions and behavioral predictors.

4.2 Key Predictive Variables

Using variable importance ranking from SAS, the top five predictors for dropout were:

1. Medication adherence score
2. Patient portal usage frequency
3. Number of adverse events reported
4. Age
5. Missed appointments in the first quarter

These insights can guide patient-specific retention strategies.

4.3 Visualization Insights

- **ROC Curve:** The neural network's ROC curve showed a consistent lift across thresholds, indicating robust discrimination.
- **Decision Tree Paths:** Identified high-risk paths, e.g., patients with <40% adherence and >2 missed appointments had a 76% predicted dropout rate.
- **Lift Chart:** Demonstrated that targeting the top 20% of high-risk patients could prevent over 60% of dropouts.

CONCLUSION

The application of AI-based modeling using SAS tools demonstrates a powerful approach to mitigate patient dropout in longitudinal clinical trials. Among the models tested, neural networks showed the highest predictive performance, especially in identifying behavioral and engagement-driven dropout risks. Importantly, these insights allow for early intervention, thereby improving data integrity and reducing trial costs.

While logistic regression provides a baseline framework and interpretability, its linear assumptions limit predictive capacity. Decision trees offer moderate performance with some transparency but struggle with

overfitting. Neural networks, though less interpretable, present the most accurate results when trained on multi-dimensional datasets.

From a clinical operations standpoint, integrating AI-based predictive tools into trial management systems can revolutionize how dropout is managed. Patient retention can be transformed from a reactive to a proactive process—shifting from blanket approaches to precision interventions.

Future research should focus on expanding these models with real-time data feeds, mobile health app integration, and dynamic recalibration across trial phases. Combining SAS analytics with wearable device telemetry and EMR data holds promise for even more granular, real-time prediction models.

REFERENCES

- Altman, D. G. (1990). *Practical statistics for medical research*. Chapman & Hall.
- Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press.
- Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22. <https://doi.org/10.1093/biomet/73.1.13>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). John Wiley & Sons.
- Eysenbach, G. (2005). The law of attrition. *Journal of Medical Internet Research*, 7(1), e11. <https://doi.org/10.2196/jmir.7.1.e11>
- Topol, E. J. (2010). Transforming medicine via digital innovation. *Science Translational Medicine*, 2(16), 16cm4. <https://doi.org/10.1126/scitranslmed.3000484> pubmed.ncbi.nlm.nih.gov
- Schumacher, L., & Chakraborty, G. (2014). Clustering and predictive modeling of patient discharge records with SAS® Enterprise Miner™ (Paper 1633). In *Proceedings of the SAS Global Forum 2014 Conference*. Cary, NC: SAS Institute Inc. support.sas.com
- Lin, G., & Rodriguez, R. N. (2015). Weighted methods for analyzing missing data with the GEE procedure (Paper SAS166). In *Proceedings of the SAS Global Forum 2015 Conference*. Cary, NC: SAS Institute Inc. support.sas.com
- SAS Institute Inc. (2014). *SAS® Enterprise Miner™ 13.1: User's Guide*. Cary, NC: SAS Institute Inc.