IFSCC 35TH
CONGRESS
FRANCE_CANNES 15-18 SEPT 25
THE FUTURE IS SCIENCE
ifscc2025.com
IFSCC
SOCIÉTÉ FRANÇAISE DE COSMÉTOLOGIE

IFSCC 2025 full paper (IFSCC2025-1599)

Machine Learning Models to Forecast Results in Cosmetic Clinical Research

Thais Pontes, MD. phD[1]; Ester Maia[1], Ana Luiza Decotelli[1]; Amanda Ornellas[1]; Ursula Cabral[1]; Bianca Azevedo[1]; Samuel Costa[1]; Pedro Fernando-Silva[1]; Guilherme Vanzan[1]; Ingrid Adão[1]; Isadora Marcondes[1]; Barbara Fonseca[1]

[1] ARTHA research

1. Introduction

Clinical research is fundamental for determining the efficacy and safety of new treatments, procedures and products. A critical component of trials success relies on the robustness of the study design, once a well-designed trial minimizes errors and biases, and ensures realibility of results [1]. Inadequate study design can lead to profound financial repercussions and missed therapeutic opportunities [2,3]. These challenges are explored in pharmaceutical studies [4], however neglected for cosmetical clinical research. In this niche, accelerated innovation cycles, intense consumer demand and an increasingly sophisticated regulatory environment converge to impose heightened demand for precise and efficient trials. In this fast-paced field, inadequate study design not only risks resource wastage but also delays the introduction of promising cosmetic products to the market. Within this landscape, optimizing study methodologies by adopting dynamic, predictive, and individualized strategies for trial planning and execution has become imperative to meet the unique demands of this sector.

In recent years, the advent of machine learning (ML) techniques offers transformative potential in this context. ML enables the processing of large and heterogeneous datasets, uncovering latent non-linear relationships and constructing predictive models that can forecast clinical trial outcomes [5]. Advanced artificial intelligence (AI)-driven methods have already demonstrated their value in clinical pharmacology and translational science, where they have improved dosing strategies [6]. Despite these advances, the integration of ML into clinical research methodologies for optimizing trial design and predicting study outcomes has yet to be broadly applied beyond domains such as diagnostic imaging and patient care pathways.

In this study, we introduce a data-driven framework that leverages a proprietary repository of completed cosmetic clinical trials curated in a Contract Research Organization (CRO). Employing advanced ML algorithms, we develop predictive models to forecast trial success and to guide the formulation of personalized trial design strategies. Our approach seeks to accommodate the distinct features of cosmetic research such as the diversity of data modalities and the inherent subjectivity of outcomes thereby enhancing methodological precision, operational efficiency, and ethical rigor. By addressing the multifaceted complexities intrinsic to cosmetic trials, this work aspires to accelerate the development pipeline for safe and efficacious cosmetic products, while promoting sustainable market integration.

2. Materials and Methods

Data Source: The dataset analyzed in this study was derived from clinical trials conducted at a CRO between 2022 and 2025. Data were extracted from Case Report Forms (CRFs) and instrumental equipment outputs.

Variables and Structure: The compiled dataset includes three core domains: Study-level characteristics – product type, application routine, and study code; Participant demographics – gender, age, phototype, skin type, ethnicity, and sensitive skin; Outcome variables – parameter evaluated, timepoint, and recorded value.Three types of measurements were included: clinical grading scales (evaluated by medical professionals), instrumental readings (e.g., skin pH, hydration, transepidermal water loss, oiliness, skin color), and lesion or acne mark counts.

Data Preparation and Selection: The raw data originated from multiple sources: CRFs were exported in .xlsx format and instrumental data in .xls. After extraction, datasets were cleaned, standardized, and merged. Data validation and processing were conducted using Microsoft Excel and Python 3. Missing values were identified in the Skin Type, Ethnicity, and Sensitive Skin variables. As these are categorical, records containing missing values were excluded from the analysis from statistical models if the absence of this parameters did not compromise the models. Categorical variables were properly formatted, and no duplicate entries were found. Unique identifiers were maintained for each study and participant to ensure accurate tracking of repeated measures over time. Only participants who completed the studies in accordance with protocol were retained in the dataset. To ensure consistency, only studies evaluating skin oiliness at standardized timepoints (Days 0, 7, 14 and 28) were selected. The final dataset contained 1216 rows and 28 columns. Since observations are recorded by timepoint, multiple entries may correspond to the same subject. A total of 4 studies were included. Data from 2 studies was not used due to missing demographic data.

Ethics and Confidentiality: Data used in this study were obtained from clinical studies where participants had already been anonymized. Ethical approval for all clinical studies was obtained from the corresponding institutional review boards. To ensure reproducibility, a random seed was applied to all applicable models, partitionings and hyperparameter tunings.

Overall design and Model selection: The analysis was conducted by evaluating performance metrics obtained from four Machine Learning models. The primary objective of the study was to predict the value of a variable at any given timepoint, based on subject- and study-level characteristics. As the primary objective of the study is to predict the value of a continuous outcome variable at any given timepoint, based on subject- and study-level characteristics, appropriate regression models were selected. Considering the numeric nature of the response variable, Sebum Measurement (SM), and the structure of the data, the models selected for evaluation included Linear Model, Generalized Linear Model (GLM), Random Forest, XGBoost and Linear Model (LM) is commonly used when a linear relationship between the independent variables and the response variable is assumed. However, when one or more of its assumptions are violated, an extension of LM known as the Generalized Linear Model (GLM) can be applied, as it provides greater flexibility by allowing different distributions for the response variable through the use of link functions. Despite their advantages, both LM and GLM may struggle to model complex, high-dimensional, or nonlinear data. To address these limitations, Random Forests (RF) were employed. This algorithm builds multiple independent decision trees on different random subsets of the data and is useful in handling complex datasets with nonlinear relationships. Finally, XGBoost (XGB), a gradient boosting algorithm, was used. XGBoost constructs decision trees sequentially, improving performance by correcting errors from previous iterations. To optimize the performance machine learning algorithms, hyperparameter tuning was performed using K-Fold cross-validation (n=5) in combination with Grid Search. This approach was adopted to improve predictive performance, enhance model robustness, and minimize overfitting.The performance of each predictive model was assessed using a combination of error-based and goodness-of-fit metrics: R-squared (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). R-squared (R^2) ranges from 0 (indicating poor model fit) to 1 (indicating excellent model fit). For all other metrics (MAE, RMSE, and MAPE), lower values indicate better model performance.

Train/test datasets: For model training, using a random seed to ensure reproducibility, one observation per subject in a study was selected for testing, while the remaining observations were used for training.

3. Results

The database contained a total of 307 subjects, although some individuals may appear in multiple studies. The majority of participants were aged between 20 and 29 years (57.98%), with no records of individuals aged 50 years or older. Women constituted the majority of the sample (76.87%), with a significant proportion presenting with oily skin (76.22%) and sensitive skin (59.28%).

Ethnic distribution showed good diversity, with 21.50% identifying as Latin/Hispanic, 20.85% as Black/Afro-American, and 20.20% as White/Caucasian. However, 35.83% of participants had no recorded ethnicity data. Regarding skin phototypes, a balanced distribution was observed across phototypes II to V, ranging from 19.87% to 23.13%. The overall data is summarized in the table 1.

Table 1. Demographic data of selected datasets

Parameter	Variable	N	%
Age Group	Less than 20	27	8.79%
	20-29	178	57.98%
	30-39	90	29.32%
	40-49	12	3.91%
Ethnicity	Asian	5	1.63%
	White/Caucasian	62	20.20%
	Latin/Hispanic	66	21.50%
	Black/Afro-American	64	20.85%
	NA	110	35.83%

Gender	Female	236	76.87%
	Male	71	23.13%
Sensitive Skin	Yes	182	59.28%
	No	125	40.72%
Skin Type	Dry	1	0.33%
	Mixed/Combination	62	20.20%
	Normal	10	3.26%
	Oily	234	76.22%
Phototype	I	17	5.54%
	II	61	19.87%
	III	66	21.50%
	IV	71	23.13%
	V	60	19.54%
	VI	32	10.42%

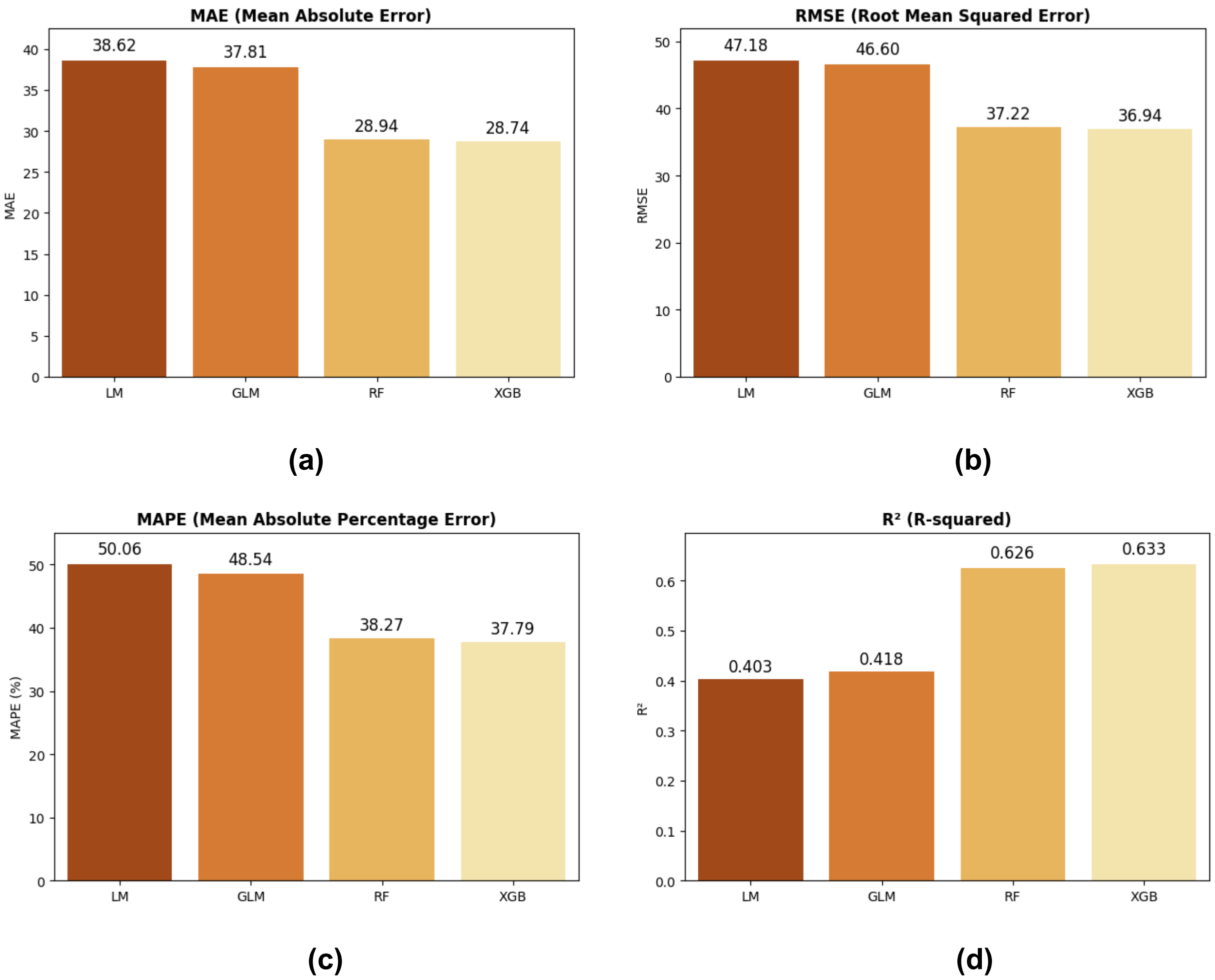
Following data cleaning and selection process, six studies were initially retained in the database. However, three of these studies contained missing values (NAs) in various demographic columns. Specifically, two studies had missing data in the Skin Type and Sensitive Skin columns, while one study exhibited missing values in both of these columns as well as in the Ethnicity column. Concerning the IP aspect, only two studies presented groups with similar IP aspects, while the others were unique.

Given that LM and GLM methods cannot be effectively applied to datasets with missing values, two studies were excluded from the final database. However, the study with missing values in the Ethnicity column was retained. This decision was based on the fact that including the study offered greater benefits to model performance than excluding it. As a result, the developed LM and GLM models included all demographic characteristics except Ethnicity, along with the days of product usage and value of last timepoint (D0 was considered as 0) and product aspect. The Sebum Measurement (SM), value measured by an equipment, was defined as the response variable.

The development of Random Forest and XGBoost models was conducted on the same dataset (containing only four studies). However, since these models can handle missing values, the Ethnicity variable was retained in both models. The metrics obtained from all developed models are presented in table 2 and figure 1.

Table 2. Error-based and goodness-of-fit metrics of tested Machine learning models

Model	MAE	RMSE	MAPE	R²
Linear Model (LM)	38.62	47.18	50.06%	0.403
Generalized Linear Model (GLM)	37.81	46.60	48.54%	0.418
Random Forest (RF)	28.94	37.22	38.27%	0.626
XGBoost (XGB)	28.74	36.94	37.79%	0.633



(a)

(b)

(c)

(d)

Figure 1. Machine learning error-based and goodness-of-fit metrics (a) MAE, (b) RMSE, (c) MAPE and (d) R^2

The statistical models, LM and GLM, produced similar results, with GLM showing a slight advantage across all four evaluation metrics. However, both models underperformed compared to the machine learning algorithms, RF and XGB, which achieved comparable results. Among them, XGB consistently outperformed RF by a small margin across all metrics. Despite this, the differences between RF and XGB remained within approximately one point for all evaluation criteria. As the machine learning models demonstrated superior performance, we continued with their analysis to identify the most suitable model. Figure 2 illustrates the feature importance for each model:

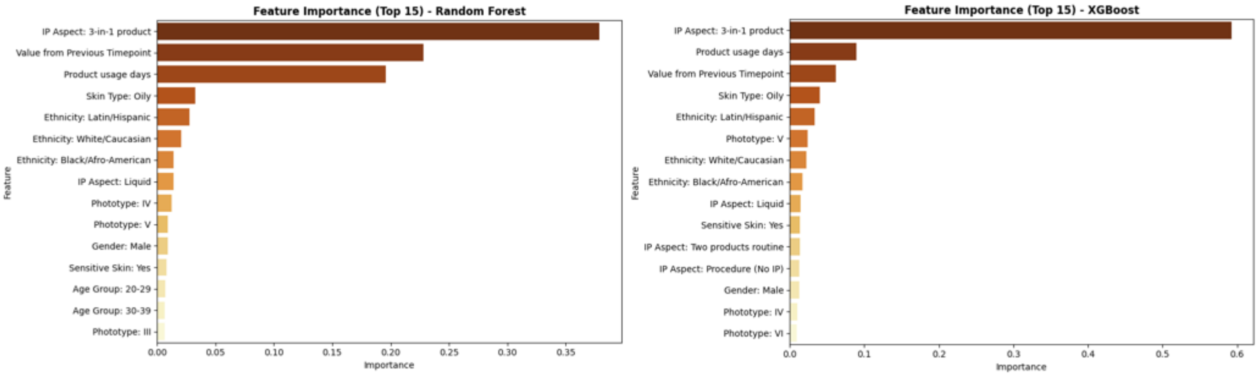


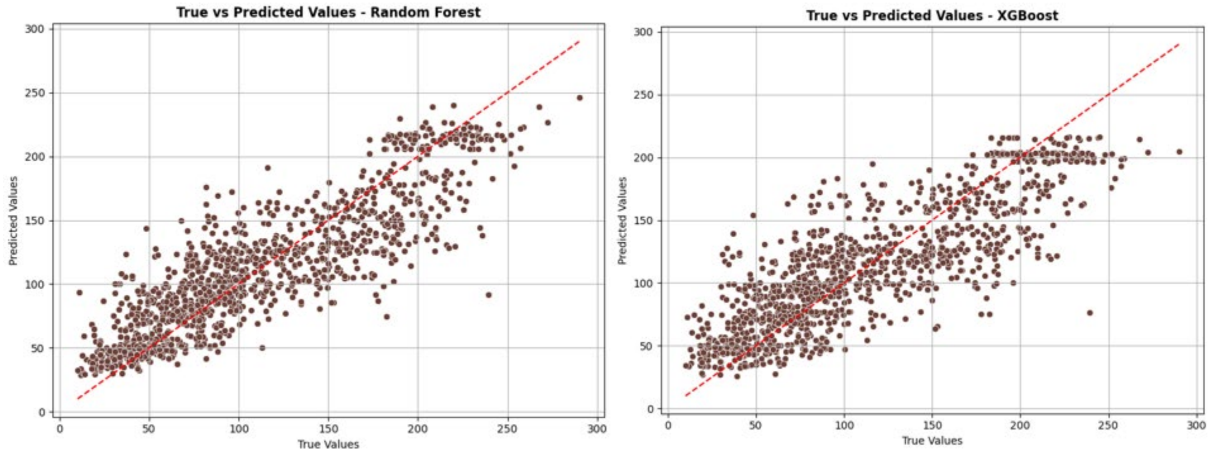
Figure 2. Features and their impact on the (a) Random Forest and (b) XGBoost model

The top four contributing variables were consistent across both models: an IP with a 3-in-1 use, the Value from the Previous Timepoint, Days of Product Usage and Oily Skin Type. In the Random Forest model, the first three features showed a relatively balanced distribution of importance. In contrast, the XGBoost model heavily emphasized the IP with a 3-in-1 aspect from the Previous Timepoint, with an importance score exceeding 0.5, while the remaining variables contributed less than 0.15 each. The table 3 presents the performance metrics of RF and XGB applied to the test dataset:

Table 3. Performance metrics of RF and XGB applied to the test dataset

Model	MAE	RMSE	MAPE	R²
Random Forest (RF)	22.97	29.78	29.70%	0.764
XGBoost (XGB)	26.34	33.79	34.83%	0.696

Although both models performed significantly better than the statistical approaches, the Random Forest model showed a clear advantage across all four metrics on the test set, with metrics of RF and XGB (Figure 3). Therefore, in this case, Random Forest is the most appropriate model to be applied.



(a) **(b)**

Figure 3. Comparison of the predicted values generated by the (a) RF model and (b) XGBoost model with the actual values from the test dataset

The model achieved reasonable performance metrics overall, which is also visually supported by the plot. Although several predictions are close to the real values, in some cases the differences are considerable, specially in higher SM values. These deviations may derivate from various factors, including participant non-compliance with the protocol, acclimatization issues, external conditions, or even poor model adjustment.

4. Discussion

Although the development and application of regression algorithms including statistical models, ML, and deep learning techniques are well established in other fields, their use in clinical research remains limited. This scarcity of studies creates uncertainty regarding which data transformations, modeling strategies, and evaluation metrics are most appropriate for achieving reliable predictions in clinical contexts.

In this study, four different algorithms were developed, each encountered specific challenges. These included missing data, limited dataset size, and a lack of comparable research using similar parameters.

Despite these limitations, the machine learning models demonstrated an acceptable fit to the data, with the Random Forest model achieving an R^2 of 0.764. This suggests a moderate to good predictive capability, even when working with relatively small and imperfect datasets. Increasing the volume and quality of available data would likely improve model performance and yield more accurate predictions. The models were constructed with a careful separation of training and testing data, and their robustness was enhanced through the combined use of K-Fold cross-validation and hyperparameter tuning. Increasing the volume and quality of available data would likely improve model performance and yield more accurate predictions.

It is also important to recognize that no predictive model can achieve perfect accuracy, especially in clinical trials where unmeasured or unknown external factors such as lifestyle, environment, or individual biological variability may significantly influence outcomes. These elements introduce inherent randomness that cannot be fully captured by observed variables.

In this context, the intentional introduction of controlled random noise into the data or even development of Neural Network models may help improve model robustness and generalizability. By simulating real-world variability, this approach can prevent overfitting and enable the model to better adapt to new, unseen data.

One promising application of this approach lies in the early stages of clinical trials or longitudinal studies. Predictive models could be used to estimate individual trajectories and project expected improvements based on baseline characteristics. Even with moderate predictive power, these forecasts can support personalized treatment planning, better resource allocation, and the early identification of subjects who are most likely to benefit from a given intervention.

5. Conclusion

This study demonstrates that ML models, particularly RF and XGB, offer a substantial improvement over traditional methods in predicting outcomes of cosmetic clinical trials. These findings pave the way for more accurate, efficient, and personalized trial designs in the cosmetic industry, but further refinement of the models and expansion of dataset diversity are essential next steps.

6. References

1. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. Contemp Clin Trials Commun. 2018 Aug 7;11:156-164. doi: 10.1016/j.conctc.2018.08.001.

2. Getz K. Improving protocol design feasibility to drive drug development economics and performance. Int J Environ Res Public Health. 2014 May 12;11(5):5069-80. doi: 10.3390/ijerph110505069.

3. Getz KA, Wenger J, Campo RA, Seguine ES, Kaitin KI. Assessing the impact of protocol design changes on clinical trial performance. Am J Ther. 2008 Sep-Oct;15(5):450-7. doi: 10.1097/MJT.0b013e31816b9027.

4. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. Contemp Clin Trials Commun. 2018 Aug 7;11:156-164. doi: 10.1016/j.conctc.2018.08.001.

5. Kavalci E, Hartshorn A. Improving clinical trial design using interpretable machine learning based prediction of early trial termination. Sci Rep. 2023 Jan 4;13(1):121. doi: 10.1038/s41598-023-27416-7.

6. van Gelder T, Vinks AA. Machine Learning as a Novel Method to Support Therapeutic Drug Management and Precision Dosing. Clin Pharmacol Ther. 2021 Aug;110(2):273-276. doi: 10.1002/cpt.2326.