frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 06 November 2025
DOI 10.3389/fdgth.2025.1645233

OPEN ACCESS

EDITED BY
Gururaj H. L.,
Manipal Institute of Technology, India

REVIEWED BY
Petar Ozretić,
Rudjer Boskovic Institute, Croatia
Sreeram Vallabhaneni,
Harvard Medical School, United States

*CORRESPONDENCE
Melania Prete
melania.prete@istitutotumori.na.it

RECEIVED 11 June 2025
ACCEPTED 15 October 2025
PUBLISHED 06 November 2025

CITATION
Crispo A, Pagnano ME, Bonfigli A, Pecchia L, Luongo A, Porciello G, Coluccia S, Prete M, Bacco L, Vitale S, Palumbo E, Giaccone P, Pica R, Grimaldi M, Cascella M, Cavalcanti E, Minopoli A, De Laurentiis M, Libra M, Polesel J, Massarut S, Celentano E and Augustin LSA (2025) Adopting machine learning to predict breast cancer patients adherence with lifestyle recommendations and quality of life outcomes.
Front. Digit. Health 7:1645233.
doi: 10.3389/fdgth.2025.1645233

Adopting machine learning to predict breast cancer patients adherence with lifestyle recommendations and quality of life outcomes

Anna Crispo[1], Maria Elisabetta Pagnano[2], Agnese Bonfigli[2], Leandro Pecchia[2,3], Assunta Luongo[1], Giuseppe Porciello[1], Sergio Coluccia[4], Melania Prete[1]*, Luca Bacco[5], Sara Vitale[1], Elvira Palumbo[1], Paolo Giaccone[2], Rosa Pica[1], Maria Grimaldi[1], Marco Cascella[6], Ernesta Cavalcanti[7], Anita Minopoli[7], Michelino De Laurentiis[8], Massimo Libra[9], Jerry Polesel[10], Samuele Massarut[11], Egidio Celentano[1] and Livia S. A. Augustin[1]

[1]Epidemiology and Biostatistics Unit, Istituto Nazionale Tumori - IRCCS, "Fondazione G. Pascale", Naples, Italy, [2]Research Unit of Intelligent Health-Technologies, Department of Engineering, Università Campus Bio-Medico di Roma, Rome, Italy, [3]Fondazione Policlinico Universitario Campus Bio-Medico, Rome, Italy, [4]Branch of Medical Statistics, Biometry and Epidemiology "G. A. Maccacaro", Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Milan, Italy, [5]Research Unit of Computer System and Bioinformatics, Department of Engineering, Università Campus Bio-Medico di Roma, Rome, Italy, [6]Anesthesia and Pain Medicine, Department of Medicine, Surgery and Dentistry "Scuola Medica Salernitana", University of Salerno, Baronissi, Italy, [7]Division of Laboratory Medicine, Istituto Nazionale Tumori - IRCCS, "Fondazione G. Pascale", Naples, Italy, [8]Division of Breast Medical Oncology, Department of Breast and Thoracic Oncology Director, Istituto Nazionale Tumori - IRCCS, "Fondazione G. Pascale", Naples, Italy, [9]Department of Biomedical and Biotechnological Sciences, University of Catania, Catania, Italy, [10]Cancer Epidemiology Unit, Centro di Riferimento Oncologico di Aviano (CRO), IRCCS, Aviano, Italy, [11]Breast Surgical Oncology Unit, Centro di Riferimento Oncologico di Aviano (CRO), IRCCS, Aviano, Italy

Introduction: Healthy lifestyle behaviors and improved quality of life have been associated with better prognoses in breast cancer survivors. However, sustaining behavioral changes remains challenging; therefore, identifying effective components of lifestyle education programs is essential to enhance adherence, improve quality of life, and facilitate their integration into clinical practice. This study aimed to predict patient adherence to a lifestyle intervention of diet, physical activity, and vitamin D supplementation and to forecast the most frequent Health-Related Quality of Life over the subsequent three measurements.

Methods: A total of 316 breast cancer survivors were included in the analysis. Adherence was modeled as a multi-label time series classification task, with compliance recorded on a three-point scale for each treatment component at quarterly intervals over one year. Health-Related Quality of Life was predicted by evaluating first-year adherence data to estimate the mean score over the subsequent three measurements.

Results: The dataset was split into 70% for training and 30% for evaluation. Random forest classifiers were employed for adherence prediction, achieving accuracy of up to 81%. An XGBoost regressor was used for Health-Related quality of life prediction, and it was compared to a baseline linear regression model. XGBoost demonstrated superior predictive performance, achieving an R-squared value of 0.62.

Discussion: Our findings highlight the promise of machine learning techniques in supporting personalized medicine. Advanced predictive models may aid in identifying patients at risk of non-adherence, enabling early interventions, and improving long-term outcomes through tailored lifestyle strategies for breast cancer survivors.

KEYWORDS

breast cancer, machine learning, missing data, diet, health-related quality of life

1 Introduction

According to the latest data from Global Cancer Observatory (GLOBOCAN 2022) breast cancer (BC) is the second most common cancer worldwide and the fourth leading cause of cancer-related death.

In 2022, there were 2.3 million new cases and 666,000 deaths (1), and the number of BC survivors continues to increase, with 5-years survival rates at around 80% (2). Therefore, there is a growing interest in clarifying how cancer, treatment, and lifestyle factors affect BC survivors (3). In this regard, evidence indicates that modifiable risk factors such as weight gain and physical inactivity, both prior to and following diagnosis and treatment could negatively affect BC prognosis (4, 5). Furthermore, strong evidence suggests that intervention studies aimed at increasing physical activity may improve the quality of life in BC patients (6, 7). In this context, the assessment of the adherence to a lifestyle modification program plays a central role. Dietary intervention adherence is evaluated using validated tools (food records, food frequency questionnaires, blood tests), and goal attainment scales. Physical activity adherence is evaluated through validated tools that include digital technology (i.e., steps count), and goal attainment scales. These latter can be used to quickly obtain information on adherence to a lifestyle program over time and to understand the impact on patient-reported outcomes (PRO), specifically Health-Related Quality of Life (HRQoL) in conjunction with validates tools (4, 8). According to the World Health Organization (WHO), HRQoL is an important self-perceived parameter of patients' general health, providing information on physical, psychological and emotional characteristics, and social appearance (9). HRQoL assessment in cancer patients provides important information to clinicians, representing a crucial endpoint in health and clinical research. In this regard, evidence indicates that BC survivors may experience several physical and mental disorders including pain, fatigue and anxiety (10–15). Moreover, the long-term effects of cancer and its treatment could negatively influence cognitive function, including symptoms such as anxiety, depression, fear of recurrence, psycho-physical stress, lack of concentration, memory loss, disease-related cognitive fog ("chemobrain") and sleep disturbances (16–18). A growing number of studies evaluated the role of healthy dietary patterns on HRQoL in BC survivors. Evidence from prospective cohort studies showed that higher consumption of a vegetables and fruits-based dietary pattern is associated with better scores in global health status/quality of life, physical functioning, emotional functioning and cognitive functioning, as well as fewer symptoms of nausea and vomiting, dyspnea, insomnia, loss of appetite, constipation and diarrhea (19).

In Italy, two cross-sectional investigations from DEDiCa study indicate that higher adherence to the Mediterranean diet in a subgroup of BC survivors is associated with better aspects of quality of life, specifically higher physical functioning, better sleep, lower pain, and generally higher well-being (20) as well as higher overall quality of life (21). As the impact of diet and lifestyle on HRQoL in BC survivors becomes increasingly evident, integrating innovative tools such as Machine learning (ML) may further enhance our ability to monitor, predict, and personalize these interventions.

Within this context, artificial intelligence is increasingly gaining importance in patient-reported clinical outcomes evaluation and adherence to lifestyle interventions, as well as in other areas of research (22). ML approaches, and related predictive analytics are now used to enhance cancer diagnosis, forecast treatment outcomes, and inform therapy plans (23). One of the key objectives of oncological research is the identification of reliable and validated methodologies for predicting risk, enabling early diagnosis, assessing clinical prognosis, and understanding disease-related behaviors in cancer patients (24).

This study aims to apply advanced ML techniques, to model and predict health-related behaviors in women diagnosed with BC who are enrolled in the DEDiCa study (25). The patients attended study visits every three months from the baseline (BL) visit and were evaluated on their adherence to the treatment. We aimed to address two primary research questions (RQ): the RQ1 was to predict the most frequent pattern of patient compliance with recommendations regarding diet, exercise, and vitamin D supplementation across the follow-up points (M3, M6, M9, M12) over the subsequent 9 months. This prediction was based on compliance data collected from third (M3) up to 12th month of follow-up (M12) and on patients' BL clinical and demographic characteristics. RQ2 was to forecast quality of life outcomes by using HRQoL measures assessed from BL to M12 in addition with BL patients' characteristics.

The target variable was the average HRQoL across the subsequent 9 months of follow-up. By analyzing the trajectories of HRQoL measures from the study's early phase, these approaches seek to gain deeper insights into how these behaviors evolve, helping to tailor interventions that support

sustained lifestyle changes and ultimately improve patient outcomes.

2 Materials and methods

2.1 The trial

DEDiCa study is an Italian multicenter randomized controlled trial, started in 2016 and approved by the Ministry of Health, Italian Drugs Agency-AIFA (EudraCT 2015-005147-14), and the Ethics Committees of the participating centers (ClinicalTrials.gov identifier https://clinicaltrials.gov/ct2/show/NCT02786875). The primary endpoint of DEDiCa study is to evaluate the effect of an intervention combining diet, physical activity (PA), and vitamin D supplementation on BC recurrence and disease-free survival. While the secondary endpoint includes improvements in cardio metabolic health and HRQoL (25). The patients observed in this study were recruited in cancer units of research hospitals in Italy: Istituto Nazionale Tumori IRCCS Fondazione "G. Pascale" (Naples), Azienda Ospedaliera per l'emergenza Cannizzaro (Catania), Ospedale San Vincenzo di Taormina (Taormina), Centro Riferimento Oncologico—CRO (Aviano). Eligible participants were women aged ≥30 <75 years with a primary diagnosis of histologically confirmed BC (stages I-III, without metastasis), within 12 months from diagnosis (25), who can understand and sign informed consent, as well as adhere to the study protocol. Patients with other malignancies, severe renal failure, hypercalcemia, kidney stones, granulomatous diseases, or sarcoidosis are excluded.

2.2 Data collection

Data on anthropometric measurements, dietary intake, PA, HRQoL, and blood parameters [including serum 25(OH)D] were collected at BL and during follow-up visits (M3, M6, M9, M12).

At the BL visit, anamnesis, clinical data, and information on vitamin D supplementation were also recorded. At each follow-up, trained nutritionists collected data on ongoing pharmacological treatments, clinical notes, and adherence to lifestyle modifications. Dietary intake was assessed using a 7-day food diary and were processed using a professional software WinFood© (version 3.9.0; Medimatica Srl Italy), which utilized two Italian nutrition databases, CREA (Council for Agricultural Research and Economics), and BDA (Food Composition Database for Epidemiological Studies in Italy). While PA was monitored via an electronic pedometer (Omron Walking Style IV, HJ-325-EB—OMRON Healthcare Customer Service Europe© 2025) and a structured questionnaire. Serum 25(OH)D concentrations were measured using the chemiluminescent immunoassay (CLIA) method with DiaSorin kits on the Liaison XL analyzer (DiaSorin S.p.A., Italy). Samples, collected in anticoagulant-free Vacutainer tubes (Becton, Dickinson and Co., Franklin Lakes, NJ, USA), were analyzed within 2 h of blood collection or thawing. All blood samples were processed in the reference laboratory (Istituto Nazionale Tumori IRCCS Fondazione "G. Pascale" Naples) under standard quality control procedures (25). Vitamin D dosage was monitored and adjusted at follow-ups based on 25(OH)D levels to meet group targets.

2.3 Adherence to dietary intervention in Bc patients

Daily foods intake and portion sizes were assessed using food diaries. If necessary, nutritionists supplemented the information with targeted questions. Foods were classified as recommended or discouraged according to the principles of the Mediterranean diet, with specific adaptations for breast cancer survivors, in line with the World Cancer Research Fund (WCRF) recommendations (https://www.wcrf.org/cancer-prevention-recommendations/). Diet adherence (AD_DIET) was classified on a three-point scale: 1 point for poor compliance (i.e., adherence to <50% of dietary advice), 2 points for moderate compliance (i.e., adherence from 50% to <80% of dietary advice), and 3 points for higher adherence (ranging from 80% to 100%).

2.4 Adherence to physical intervention in Bc patients

Adherence to PA recommendations (AD_PA) was assessed quarterly by calculating the average number of steps taken by participants during the week prior to the visit. Patients were encouraged to walk briskly for half an hour per day. PA adherence was classified on a scale from 1 to 3: if patients increased the average number of steps by 4,000−5,000 compared to BL, a score of 3 was assigned; if the average number of steps was half of the BL, a score of 2 was assigned. Otherwise, a score of 1 was assigned.

2.5 Adherence to vitamin D intake recommendations

Vitamin D dosage was monitored every 3 months via serum 25(OH)D and adjusted to achieve sufficiency (between 30 ng/mL and 60 ng/mL). Adherence to vitamin D supplementation (AD_VITD) was evaluated on a 3-point scale: 3 points were assigned if the patient achieved sufficient vitamin D levels when initially deficient; 2 points were given if the patient frequently forgot to take the supplement, resulting in inconsistent intake. While 1 point was assigned if the patient did not report improvement in vitamin D status.

2.6 Quality of life assessment

HRQoL was assessed through a validate questionnaire, the European Quality of Life 5 Dimensions 3 Level (EQ-5D-3l) (26). It gives a non-cancer-specific measure of generic health status

that includes a descriptive system comprising five dimensions (mobility, self-care, usual activities, pain or discomfort, and anxiety or depression) and three levels of perceived problems (1 for no problems, 2 for some problems, and 3 for extreme problems). A unique health index score is calculated by applying an algorithm that sues coefficients (called weights) to each value of the levels for each dimension; the Italian Model was used to estimate EQ-5D-3l index score (27). Using this model, the EQ-5D-3l health index spans from 1.00 for the best possible health state to −0.38 for the worst possible health state.

2.7 Features and outcomes

The analyzed dataset includes adherence to three intervention categories (diet, PA, and vitamin D Supplementation) recorded at three time points during the first year of the program.

In addition to the temporal adherence measures, the dataset includes at BL a comprehensive set of clinical characteristics, which serves as key variables for understanding patient profiles. These characteristics are shown in Supplementary Table S1. Compliance with diet, exercise, and vitamin D supplementation was assessed based on data collected up to M12. We then predicted the most common compliance behavior across the three follow-up points in the next 9 months. For HRQoL prediction, we used scores from BL to M12, considering baseline characteristics, and then predicted the average HRQoL score for the following 9 months.

3 Statistical analysis

All the computational and statistical analysis were performed using Python 3.10.6 [https://www.python.org/downloads/release/python-3106]. A complete list of the Python packages and libraries employed, along with their respective versions, is provided in Supplementary Table S1.

3.1 Data preprocessing

The initial analysis focused on the entire trajectories of HRQoL scores among patients, revealing a significant amount of missing data across the 12 quarterly measurements. Initially, 492 subjects were included in the study. Among them, 176 were excluded because they had more than 4 missing values in their HRQoL trajectory, a criterion adopted to ensure the reliability of the imputations and to reduce potential bias. Most subjects had no missing data, followed by those with 1, 2, or 3 missing values.

A more restrictive threshold than the one adopted (e.g., 3 missing values) would have led to the exclusion of a substantial number of subjects, significantly reducing the amount of data available for model training. To impute missing values, we used the 93 fully observed cases to train and test various imputation models. The model that achieved the best performance on the test set (shown in Table 1) was then applied during the inference phase to estimate the missing values in the remaining 223 subjects. Following imputation, 316 complete cases were obtained. While it is well known in the literature that imputation may introduce some degree of bias, in real-world scenarios where ideal, (gold standard) data are not available, it remains a valuable approach to enable the development and validation of predictive ML and Deep Learning models (28). Specifically, we employed the Iterative Imputer with an Extra Trees Regressor estimator (29) due to its ability to handle non-linear relationships within the data. The method iteratively predicts missing values by modeling each incomplete feature as a function of the other features. This process repeats until the imputations stabilize, ensuring consistent and reliable estimates. To assess the performance of this approach, we compared it with several other imputation methods, including Iterative Imputer with a Bayesian Ridge estimator (30), K-Nearest Neighbors (KNN) Imputer (31), and Simple Imputer (mean strategy) (32). Each method was implemented and evaluated using the Root Mean Square Error (RMSE) and the Coefficient of Determination (R^2), which are explained more in detail in the paragraph 3.3. These metrics calculated between the imputed values and the observed values in complete cases.

TABLE 1 Summary of the imputation methods.

Imputation Methods	RMSE	R^2
Iterative Imputer (Bayesian Ridge)	0.55	0.66
K-Nearest Neighbors (KNN) Imputer	0.61	0.55
Simple Imputer (Mean Strategy)	0.58	0.56
Iterative Imputer (Extra Trees Regressor)	**0.41**	**0.75**

RMSE, Root Mean Squared Error; R^2, coefficient of determination.
The best algorithm is highlighted in bold. RMSE measures prediction accuracy, with smaller values indicating better performance, the coefficient of determination (R^2) indicates the proportion of variance explained, with larger values indicating better performance.

3.2 Clustering—QoL group identification

Following the imputation process, hierarchical clustering was performed on the imputed HRQoL trajectories to explore potential differences over the follow-up period (33). Specifically, a divisive hierarchical clustering approach was used. The clustering analysis identified a majority group, which represented the most common pattern in the HRQoL trajectories. The Mann–Whitney *U*-test was used to test the null hypothesis of non-different means between groups with a 95% significance level. Results revealed significant differences in the slopes of the HRQoL trajectories between this majority cluster and the other identified groups. These findings highlighted the majority cluster as the primary focus for subsequent analysis, enabling a more targeted exploration of the dominant HRQoL trends. A *post-hoc* analysis was assessed to quantify the proportion of patients with imputed data within each cluster to evaluate whether the clustering solution might have been influenced by data completeness.

3.3 Forecasting models

To answer RQ1, we approached the adherence prediction as a multi-label time series classification task, with patient compliance for each advice category (diet, PA, and supplemental vitamin D) recorded at each trimester on a discrete scale of 1–3. Specifically, three separate Random Forest classifier (RF) models were developed, each one for a single specific outcome: future adherence to one of the three recommendations. These models utilized a comprehensive input dataset, which included temporal adherence patterns for diet, PA, and vitamin D supplementation recorded during the first year of the program (M3, M6, M9, and M12), combined with BL characteristics (age, cancer stage, molecular subtypes, radio-, chemo-, neoadjuvant therapy status, comorbidities). The target label for each classifier represented the most frequently recorded adherence value in its respective category during the subsequent nine-month period and was treated as a classification problem. Although the input data remained constant across models, each classifier was tasked with predicting a different target variable, specifically the future adherence to one of the three recommendations. To answer RQ2, we employed an XGBoost, a gradient boosting regressor, chosen for its effectiveness with continuous outcomes (34). This algorithm is part of supervised regression ensemble models adopting a Decision Tree (DT) sequence: the algorithm works by adding a new DT to the former ones to minimize the regression error, which was the residue of that series. XGBoost incorporated BL values and QoL time series data from the program's first year to forecast the average QoL score by the one-year treatment. A linear regression (LR) model was fitted as a reference algorithm to compare the performances of the former one. We performed a 70/30% train/test split on our dataset to prepare it for model training and evaluation. This split allowed us to use 70% of the data for training the models and 30% for testing their performance on unseen data. All samples were uniquely assigned to either the training or the test set. The split was carried out randomly to avoid selection bias, ensuring that the distribution of the target variable was preserved across training and test sets. We adopted a k-fold cross validation on the training data to assess the best tuning of algorithms, namely that set of hyperparameters, chose from a wide possibility of combinations (often set in a grid), which are associated with the highest performances. For each combination of hyperparameters, data are split into random k equal folds of observations: each part is further divided into $k-1$ parts dedicated to fit the algorithm while the kth part is used to measure its performances. The best combination of hyperparameters is the one to be finally chose for the final algorithm. We used $k=5$ as default value from the sci-kit learn method in Python. A fixed random seed was applied to ensure the reproducibility of the experiments.

After training and optimizing our models, we evaluated their performance on the test set that was held out during the initial split. Main metrics were used for performance assessment. Precision, Recall, F1-score and Accuracy were calculated for classification problems. Precision measures the accuracy of correctly labeled predictions, indicating how many of the predicted labels were correct. Recall represents the ratio of correctly detected labels out of all the true labels. The F1-score provides a balance between precision and recall, calculated as the harmonic mean of these two metrics. It is defined as the ratio of true positives to the sum of true positives and the average number of misclassified labels. These measures are typically used for classification problems. RMSE, mean absolute error (MAE), mean squared error (MSE), R^2 and The Bland-Altman plots were performed to analyze the quality from regression models (35). For assessing prediction quality on continuous data, MAE is a common metric used in regression tasks, providing the mean of absolute differences between predictions and actual values. However, it is an absolute measure, meaning it does not allow direct comparison between models with different units of measurement or different value ranges (36, 37). To address this issue, coefficient of determination (R^2), ranging from zero to one, was also considered. Moreover, we adopted the RMSE which gives a measure of the standard deviation of errors and penalizes larger errors more heavily than MAE, which treats errors linearly, making it less sensitive to outliers (36–38). The choice of evaluation metric depends on the application context and the data: to penalize large errors, RMSE is preferable when large errors need to be penalized, while MAE is better for a more robust approach to outliers' detection.

3.4 Feature importance

Computing the Feature Importance (FI) is a convenient technique to rank the included features in terms of association with the outcome of interest (39). Both the adopted algorithms allowed such a function, so importance scores were extracted from the RF classifiers and the XGBoost regressor to identify the key variables contributing to the predictions. For the RFs, these scores were calculated using the mean decrease in Gini impurity, allowing us to rank the features based on their influence on the classification outcomes. In the XGBoost regression model, used for HRQoL predictions, FI was evaluated based on the gain, which measures the improvement in accuracy brought by a feature to the branches it appears in. Plots assessing the importance scores of each feature were shown and the most important variable were listed and commented. Additionally, R^2 provides a measure of the model's goodness of fit, though it may not always be sufficient for a complete evaluation of performance (40, 41). All the computational and statistical analyses were performed using Python 3.10.6 [https://www.python.org/downloads/release/python-3106].

4 Results

4.1 Missing imputation

A total of 316 BC patients were included in the analysis. Table 1 summarizes the performance of four different

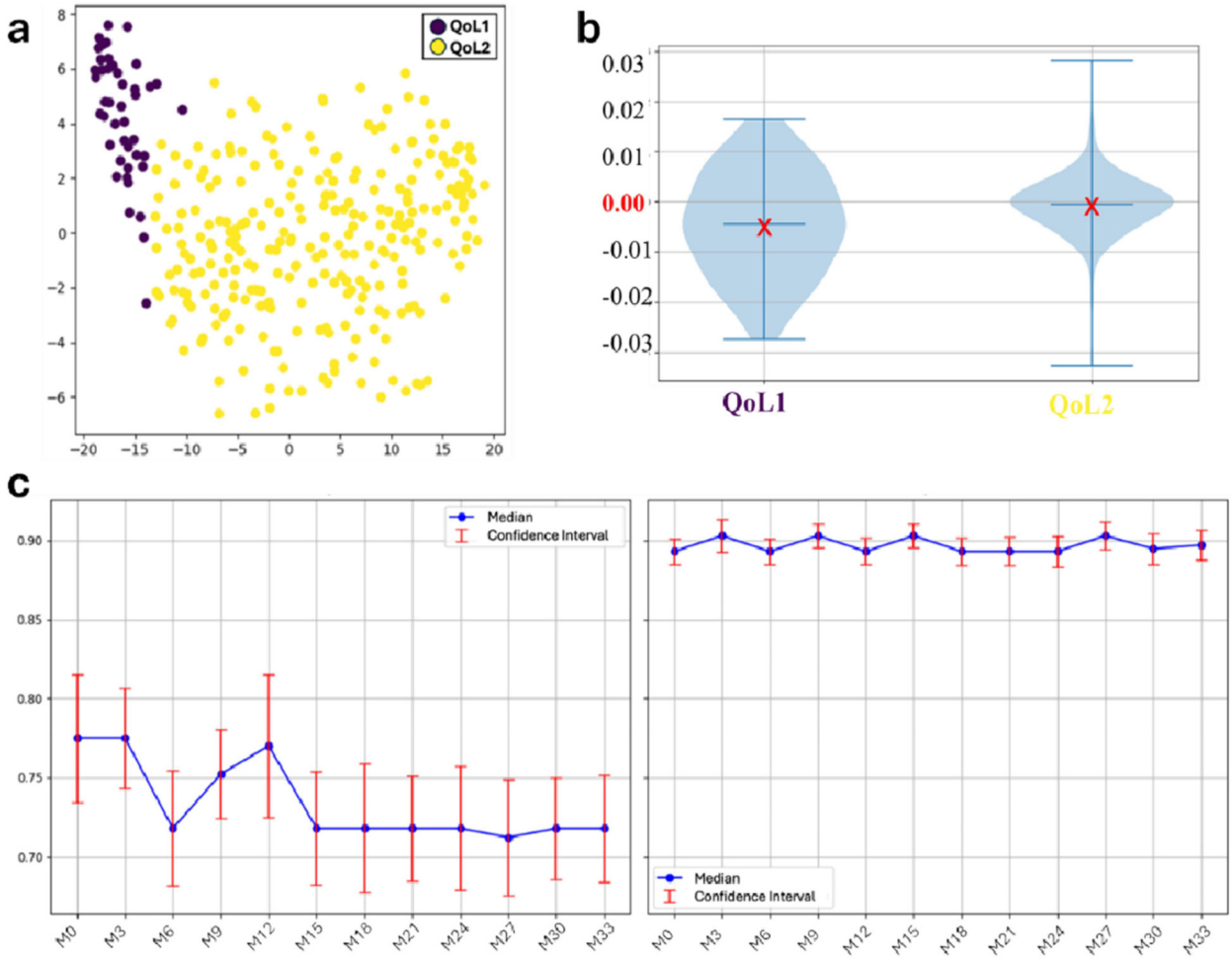

FIGURE 1
Results from clustering algorithm and trend analysis among the two HRQoL groups throughout the 9 months period: (a) scatter plot; (b) HRQoL slope distribution by groups. The red X marks the mean slope within each group: negative mean values indicate a decreasing trend over time, whereas values close to zero suggest overall stability. (c) QoL trends among the groups. The red bounded bars represent the 95% confidence interval for each single mean point. After imputing missing values, all the HrQoL measures were used for clusering. QoL, Quality of Life; HRQoL, Health Related Quality of Life.

imputation techniques applied to the longitudinal trajectories of HRQoL scores, aimed at handling missing data (Iterative Imputer with Extra Trees Regressor, Iterative Imputer with Bayesian Ridge Regressor, KNN Imputer, Simple Imputer). The method based on the Extra Trees Regressor showed the best performance (RMSE of 0.41 and the R^2 of 0.75), demonstrating higher accuracy compared to the alternative approaches.

4.2 Clustering

Following imputation, the hierarchical clustering algorithm automatically detected two main QoL-oriented groups (QoL1 and QoL2) (Figure 1a). QoL1 ($n = 45$) showed more variability compared to QoL2 ($n = 271$) (Figure 1b); regarding trends, the QoL2 group showed greater stability and higher scores (Figure 1c) and it was adopted for our further analysis. The Mann–Whitney *U*-test confirmed statistically significant differences between the two clusters ($p < 0.001$). The *post-hoc* analysis revealed that 80% of patients in QoL1% and 69% in QoL2 had imputed values. The relatively modest difference between the two groups supported the validity of the clustering approach, demonstrating that the algorithm identified patient groups based on HRQoL patterns rather than data completeness. Table 2 provides a detailed comparison between the two QoL clusters in terms of clinical profiles, highlighting distinctive patterns among patients with different HRQoL trajectories.

4.3 Forecasting the compliance to the treatment

The optimal hyperparameters identified for the three RFs and the XGBoost model used for the prediction task are summarized

10.3389/fdgth.2025.1645233

TABLE 2 Distribution of BL characteristics between quality of life (QoL) clusters.

Feature	QoL1 (*n* = 45)	QoL2 (*n* = 271)
Age (years)		
Median (Q1–Q3)	50 (46–57)	55 (49–62)
Smoking status		
No smoking	20 (44.6%)	115 (42.2%)
Smoker	10 (22.9%)	72 (26.7%)
Former Smoker	15 (32.5%)	84 (31.1%)
Number of comorbidities		
0	26 (59%)	187 (68.9%)
1	12 (25.8%)	36 (13.3%)
2	5 (10.0%)	36 (13.3%)
3	2 (4.1%)	6 (2.2%)
4+	1 (1.1%)	6 (2.2%)
Cancer Stage		
I	13 (28.4%)	72 (26.7%)
IIA	20 (45.0%)	102 (37.8%)
IIB	5 (11.8%)	54 (20%)
IIIA	4 (10.3%)	42 (15.6%)
IIIC	2 (4.4%)	0 (0.0%)
Lymph node status		
N0	21 (46.9%)	114 (42.2%)
N1	17 (38.4%)	114 (42.2%)
N2	5 (10.3%)	42 (15.6%)
N3	2 (4.4%)	0 (0.0%)
Tumor size		
T1	28 (61.6%)	157 (57.8%)
T2	17 (38.0%)	108 (40.0%)
T3	1 (0.4%)	6 (2.2%)
Molecular subtypes		
Luminal A	15 (34.3%)	78 (28.9%)
Luminal B	21 (46.1%)	138 (51.1%)
HER2+	2 (4.4%)	12 (4.4%)
TN	7 (15.1%)	42 (15.6%)
Radiotherapy status		
Never	16 (34.6%)	123 (45.5%)
Ongoing	4 (9.8%)	18 (6.8%)
Finished	25 (55.6%)	129 (47.7%)
Chemotherapy status		
Never	16 (36.2%)	102 (37.8%)
Completed (<2 months)	7 (15.9%)	42 (15.6%)
Completed (≥2 months)	14 (30.6%)	78 (28.9%)
Ongoing	8 (17.3%)	48 (17.8%)
Neoadjuvant therapy status		
Never	40 (88.9%)	253 (93.3%)
Finished	5 (11.1%)	18 (6.7%)
Type of surgery		
Quadrantectomy	35 (77.5%)	203 (75.0%)
Mastectomy	10 (22.5%)	68 (25.0%)
Time from surgical intervention (months)		
Median (Q1–Q3)	7 (5–10)	8 (5–10)

Numbers may not sum up to the *n*'s from QoL clusters due to the presence of missing values, subsequently treated. Data are presented as median (interquartile range, 25th–75th percentile) for numerical variables, and as absolute numbers for categorical variables. The number of comorbidities is a cumulative measure of pre-existing health conditions that reflects the overall burden of comorbidities. It includes type 1 and type 2 diabetes, hypertension, hypertriglyceridemia, hypercholesterolemia, and hyperglycemia. Additionally, it accounts for cancer stage, which represents the extent of BC at baseline. Chemotherapy status (chemo status) is classified as completed (<2 months or ≥2 months), ongoing, or never initiated. Molecular subtypes, derived from receptor status (ER, PgR, HER2), are used to classify the intrinsic molecular subtypes of BC. Finally, neoadjuvant chemotherapy status is categorized as completed or never initiated, indicating whether neoadjuvant chemotherapy was finished prior to surgery.

QoL, Quality of Life; ER, estrogen receptors; PgR, progesterone receptors; HER2, Human Epidermal Growth Factor Receptor 2; TN, Triple-Negative.

TABLE 3 Performance of RFs in predicting future adherence to the three recommendations: each row corresponds to a single RF trained to predict adherence to one of the three recommendations.

Adherence	Precision	Recall	F1-Score	Accuracy
Diet	0.79	0.80	0.80	0.81
Physical Activity	0.81	0.77	0.79	0.71
Vitamin D	0.77	0.78	0.77	0.79

RF, Random Forest.
The reported metrics are macro-averages, ensuring robustness to class imbalance.

in Supplementary Tables S2a,b. The evaluation metrics for the three RFs trained to predict future AD_DIET, AD_PA, and AD_VITD are summarized in Table 3. The highest accuracy was achieved by the model predicting AD_DIET (0.81), followed by the model for AD_VITD (0.79) and AD_PA (0.71).

On the right side of Figure 2, the confusion matrices are shown. In all three models, the BL characteristics were found to be less related to the outcomes. A comparison between the XGBoost model and the multivariable LR on predicting the mean HRQoL score over the following 9 months is presented in Table 4. The XGBoost model demonstrated superior performance across all metrics, with an R^2 of 0.62 compared to 0.42 for the LR model, underscoring the enhanced predictive capability of non-linear methods in modeling complex health outcomes. Notably, the Bland-Altman analysis (Figures 3b,c) revealed a strong difference in predictions between XGBoost and LR. Specifically, plot Figure 3b shows a good agreement without systematic drift, with minimal differences randomly distributed around the mean difference line; plot Figure 3c exhibits a negative drift, indicating that the model tends to underestimate predictions as mean of the measurements increases.

4.4 Important features

The important score reflects the impact of each feature on enhancing the model accuracy in predicting outcomes (Figure 2, left hand). Diet adherence was mainly influenced by its M9 and M3 components, respectively, and by M6 adherence to PA (Figures 2a,a1). Adherence to PA was principally predicted by its closer components, i.e., M9 and M12 adherence to PA, and by M12 adherence to vitamin D (Figures 2b,b1). All the M12 adherence measures were found as the features, which mostly were related to the adherence to vitamin D (Figures 2c,c1). Most important features contributing to the HRQoL average prediction were the 5 measures throughout the year of HRQoL; the number of comorbidities was found as the most ranked first BL characteristic out of HRQoL lags (Figure 3a).

5 Discussion

Our goal was to apply ML algorithms, to model and predict health-related behaviors in women diagnosed with BC who were enrolled in the DEDiCa study, a lifestyle modification clinical trial that followed participants for three years during treatment involving diet, PA, and vitamin D supplementation. Clinical studies have provided substantial evidence supporting the effectiveness of targeted lifestyle interventions. A systematic review published in 2022 assessed the impact of physical and nutritional interventions in BC patients, concluding that an integrated program combining physical activity and diet can reduce the risk of relapse and enhance quality of life (42).

Regular physical exercise contributes to reducing the risk of recurrence, improving survival, energy, sleep quality, mental health, and reducing anxiety, depression, and fatigue (43–46). Similarly, nutritional interventions play a key role before, during, and after cancer treatment. A balanced and adequate diet, rich in essential nutrients and antioxidants, such as the Mediterranean diet, can contribute to improving immune function and reducing the risk of comorbidities, such as diabetes or cardiovascular diseases, frequently observed in patients with BC (47). Promoting a Mediterranean diet can improve metabolic health and reduce chronic inflammation in these patients, as well as improving the response to oncological treatments, and is also associated with a better QoL and a lower incidence of BC recurrence (48, 49). Porciello et al. in two studies showed that adherence to the Mediterranean diet and a high-quality diet (according to the Healthy Eating Index (HEI-2015 index) are associated with significant improvements in QoL particularly in terms of physical functioning, pain, and overall wellbeing (20, 21). The inclusion of PA and therapeutic phase in the analysis enhances the understanding of the factors influencing QoL in this context. These results showed the importance of integrated, multidisciplinary interventions combining nutritional strategies and PA within follow-up programs for BC survivors (20, 21). Nevertheless, while these studies provide robust evidence, traditional statistical approaches may fail to capture the complexity and individual variability of QoL trajectories over time.

In this context, our study introduces an innovative contribution through the application of ML models, offering a dynamic, predictive, and interpretive dimension to the analysis of lifestyle interventions in oncology. From these multiple investigations, conducted on a cohort of 316 patients the FI framework detected the presence of underlying associations between diet quality, PA, and HRQoL, while providing new insights into the temporal patterns of these relationships.

Through the integration of ML methodologies, this research complements existing evidence and advances the personalization of supportive care strategies in BC survivorship. In particular, the RF and XGBoost models demonstrated superior performance compared to traditional approaches, showing greater accuracy in predicting average HRQoL levels. Moreover, these models effectively identified the behavioral factors that most strongly influence HRQoL trajectories during follow-up. Our analyses revealed that repeated measurements over time of diet, PA, and vitamin D supplementation are significantly more relevant predictors than baseline variables, thus highlighting the importance of continuous and personalized monitoring throughout the care pathway. A major challenge in longitudinal clinical studies is managing missing data due to patient

10.3389/fdgth.2025.1645233

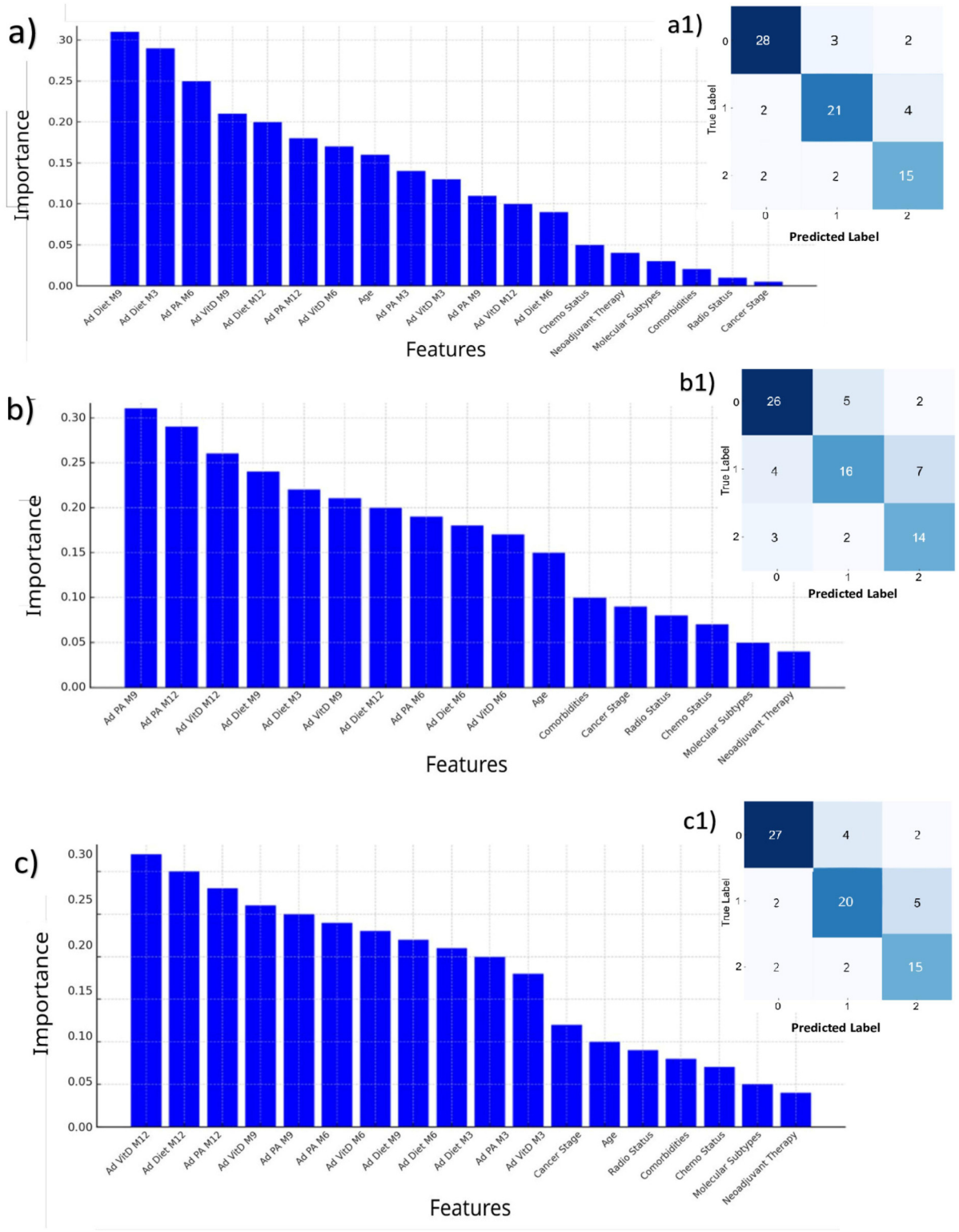

FIGURE 2
Feature importances (left) and prediction results (right) of the three RFs, for predicting the most frequent value over the next 9 months in adherence to dietary advice **(a,a1)**; physical activity **(b,b1)**; supplemental vitamin D **(c,c1)**. In each confusion matrix, the classes labeled as 0, 1, and 2 correspond to adherence levels 1, 2, and 3, respectively. Each class represents the adherence level most frequently observed over the subsequent 9 months. RF, Random Forest.

dropouts (50): in this regard, we found that only 93 patients reported complete data, at least in HRQoL measure. By employing four ML algorithms for missing data imputation, we found that the Iterative Imputer outperformed the other methods, including the Simple Imputer, despite this is the most commonly used technique for imputing missing data in diet-related studies. Furthermore, this analysis identified the key factors influencing patient adherence to the lifestyle program and future HRQoL outcomes, offering valuable insights into the complex interactions of variables across the different predictive models. This approach allowed for a comprehensive analysis of both adherence to lifestyle program and HRQoL trajectories during the initial phase of the intervention.

TABLE 4 Performances measures from XGboost and LR.

Model	RMSE	MAE	MSE	R^2
XGBoost	0.04	**0.03**	**0.002**	**0.62**
LR	0.04	0.06	0.005	0.42

LR, linear regression; MAE, Mean Absolute Error; MSE, Mean Squared Error; RMSE, Root Mean Squared Error.
The best model is highlighted in bold, indicating the one with better prediction accuracy, as smaller values mean having a better performance. The coefficient of determination (R^2) indicates the proportion of variance explained, with larger values indicating better performance.

We selected a more stable subset of patients in terms of HRQoL behavior throughout the follow-up period, we focused on the patient's compliance to the treatment. We aimed to predict the most frequent value over the next 9 months after one year of intervention in adherence to diet, PA, and vitamin D supplementation: PA was harder to depict based on baseline characteristics and its four quarter lags (accuracy = 0.71) than vitamin D (accuracy = 0.79) or diet (accuracy = 0.81). Among

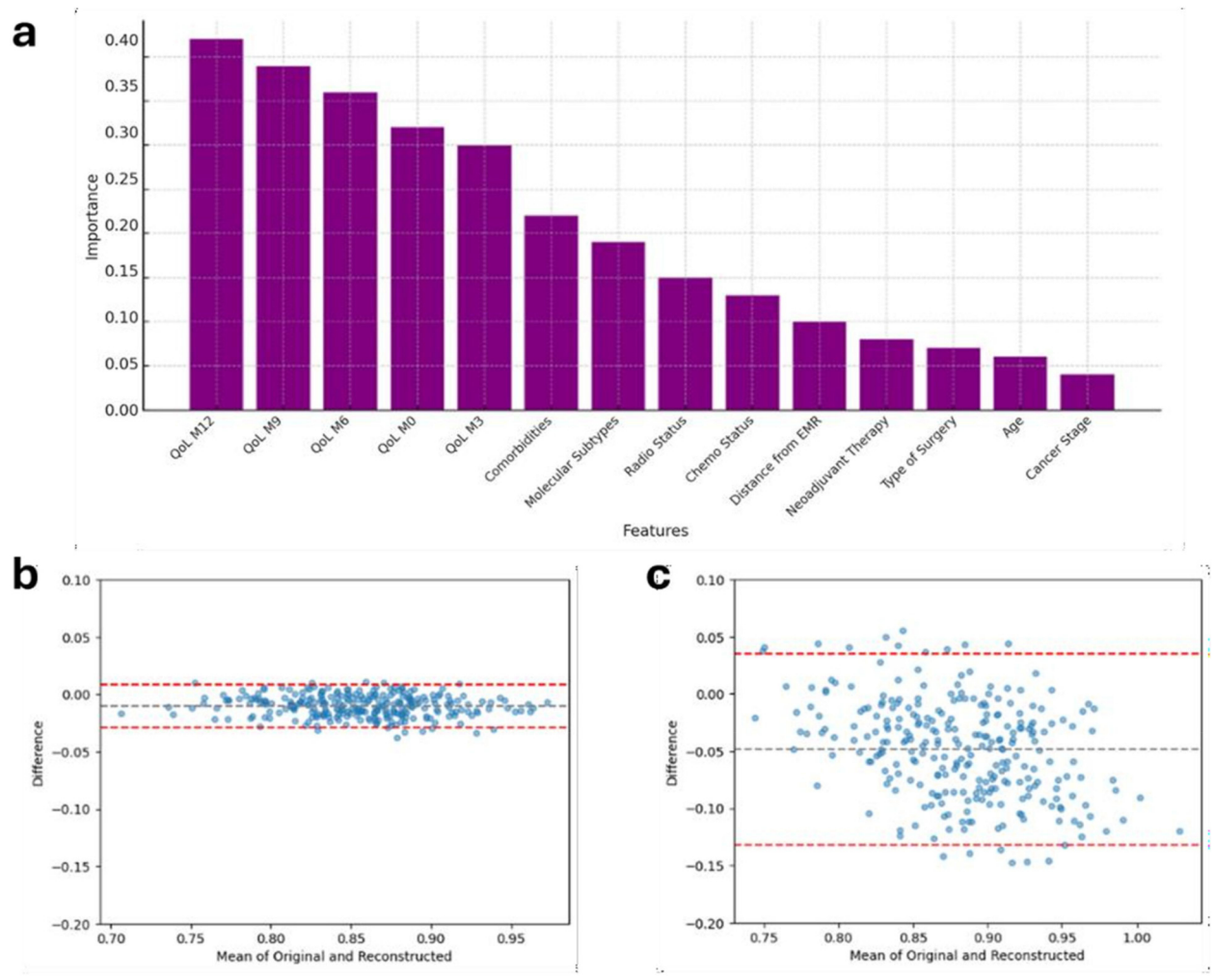

FIGURE 3
Results from the XGBoost model. **(a)** Feature importance plot illustrating the most influential predictors for estimating mean QoL scores over the subsequent 9 months of clinical follow-up. **(b)** Bland–Altman plot for the XGBoost model, and **(c)** Bland–Altman plot for LR model, provide a comparative assessment of agreement between predicted and observed HRQoL scores. Each dot represents a subject; the black dashed line indicates the mean difference (bias), and the red dashed lines the 95% limits of agreement, providing establish if accordance between predicted and observed values (subjects) statistically exists. LR, linear regression; QoL, Quality of Life; HRQoL, Health releated QoL.

our analyses, we found out that baseline characteristics played a secondary role on predictions, with the age almost the first ranked feature among them. In particular, the most frequent value of vitamin D was quite mostly associated with the M12 measures compared to the advice regarding diet and PA. Secondly, the HRQoL score was analyzed in terms of mean value over the subsequent 9 months of follow-up after one year of intervention by adopting the XGBoost regressor and a multivariable LR model.

In this context, the boosting method revealed a much higher performance compared to the linear model, with a plus 20 points-percent in R^2 statistics. This finding suggested a likely higher power of such method to predict the outcome on well-known methods. Indeed, it was confirmed from the Bland-Altman plot that the latter model showed an underestimation of HRQoL mean value. This interesting result may confirm the non-linearity relation between the outcome and the set of features.

6 Limitations

This predictive analysis has several limitations. First, since the dataset used for model training and evaluation had a relatively small sample size, especially after excluding patients with significant missing data, this gap may limit the generalizability of the findings. Second, despite advance imputation techniques being employed to handle missing data, imputation methods may introduce biases. It is more evident when data are not missing completely at random.

Alternative methods like Multiple Imputation by Chained Equations could have been considered. However, this approach was not implemented because it assumes parametric relationships between variables, which may not adequately capture the non-linear interactions present in this dataset. Third, the algorithms used for ML modelling (i.e., RF and XGBoost), although effective, may not fully capture complex temporal dependencies or interactions between features over time. For multi-label classification, algorithms like gradient boosting machines might have delivered better performance with a larger dataset. A similar challenge was observed in the cluster analysis. With a larger dataset, we could have improved the temporal analysis by utilizing techniques like K-Shape or density-based clustering (DBSCAN) to better handle groups with non-uniform distributions. Additionally, another key limitation of this study is the lack of a thorough explainability analysis. The use of more advanced explainability methods, such as SHapley Additive exPlanations (known as the SHAP method), could have enhanced the interpretability of the models' predictions.

Lastly, the analysis relied on a 3-level only adherence scale based on operators' perception and on a single cohort from a specific geographic and clinical context limiting the applicability of the results to other populations. Future studies should aim to validate these findings on larger, more diverse datasets and to better evaluate the models' performances in real world scenarios.

7 Conclusions

The findings of our study underscore the significant potential of ML algorithms in advancing personalized healthcare for BC survivors. By predicting adherence to lifestyle recommendations and forecasting QoL outcomes, these tools provide critical insights into patients' health behaviors and trajectories. The study itself revealed that adherence to dietary advice and vitamin D supplementation could be predicted with higher accuracy compared to physical activity, emphasizing the complexity of the latter. Additionally, the superior performance of XGBoost in QoL prediction highlights the value of employing advanced regression techniques for non-linear relationships in health data. Key insights include the pivotal role of adherence measures at specific time points (e.g., M12) in influencing predictions, as well as the limited but notable influence of baseline characteristics. These findings suggest that interventions focusing on consistent and measurable adherence behaviors are essential for optimizing long-term outcomes.

Future applications of this work could expand beyond BC to other chronic conditions where lifestyle modifications play a crucial role and integrating ML-driven insights into clinical practice to support healthcare providers in developing targeted and effective intervention strategies, ultimately improving patient QoL. Clinical prediction has gained greater importance in modern healthcare, involving the use of medical data to forecast future health outcomes. This process is applied in various fields of disease, from prevention to diagnosis and treatment resulting in better patient outcomes and improved efficiency of healthcare systems. ML algorithms can rapidly analyze large, complex medical data with high precision, detecting patterns and correlations that might be beyond the scope of human analysis. When feed with temporal data, they can be designed to continuously learn from new data, improving their predictive accuracy over time (51). To the best of our knowledge, few studies have aimed to predict future patient compliance with lifestyle programs using ML approaches. Mousavi et al. (52) implemented, a hybrid model combining artificial neural network and genetic algorithm to predict adherence to diet among patients referred to a private clinic in Iran, leading to high accuracy in predicting diet adherence (93.5%) and proper performance. Regarding feature importance, a genetic algorithm selected some patients-related factors that could affect diet adherence, including weight, weight satisfaction, and body mass index, lunch, dinner and sleep time. The implementation of this model in the clinical practice could be useful to identify patients with low chance of diet adherence, supporting dietitians to employ the proper nutritional strategy (52). Similarly, in the study of Kim et al. (53), seven ML algorithms were implemented to predict QoL in middle-aged South Korean adults. The RF method showed the highest performance in predicting QoL deterioration and the highest performance. Regarding feature importance, the authors

showed that sleep quality and stress were identified as the most important predictors of QoL by the model.

Data availability statement

The original data presented in this study will be available upon request. Requests to access the datasets should be directed to the last author (LA), l.augustin@istitutotumori.na.it.

Ethics statement

DEDiCa study was conducted according to the guidelines of the Declaration of Helsinki, and approved by Ethics Committee of each recruiting hospital (ClinicalTrials.gov NCT02786875; 17 March 751 2016). Informed consent was obtained from all subjects involved in the study.

Author contributions

AC: Validation, Conceptualization, Project administration, Writing – original draft. MPa: Writing – original draft, Formal analysis, Software, Methodology. AB: Methodology, Writing – review & editing, Software, Formal analysis. LP: Methodology, Writing – review & editing, Software, Formal analysis, Conceptualization. AL: Writing – review & editing, Visualization, Project administration. GP: Project administration, Writing – original draft. SC: Formal analysis, Writing – review & editing, Methodology, Software. MPr: Visualization, Writing – original draft, Writing – review & editing, Supervision. LB: Methodology, Formal analysis, Software, Writing – review & editing. SV: Writing – review & editing, Project administration, Visualization. EP: Visualization, Project administration, Writing – review & editing. PG: Formal analysis, Software, Writing – review & editing, Methodology. RP: Project administration, Writing – review & editing. MG: Project administration, Writing – review & editing. MC: Visualization, Writing – review & editing, Supervision. ECa: Supervision, Visualization, Writing – review & editing. AM: Visualization, Supervision, Writing – review & editing. MLa: Writing – review & editing, Visualization. MLi: Funding acquisition, Project administration, Writing – review & editing. JP: Project administration, Writing – review & editing. SM: Writing – review & editing, Visualization. ECe: Writing – review & editing, Visualization, Project administration. LA: Project administration, Funding acquisition, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. DEDiCa study is funded by a grant of the Italian Ministry of Health (Grant no. PE-2013-02358099) and Lega Italiana per la Lotta Contro i Tumori (LILT Nazionale). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

This work was (partially) supported by the Italian Ministry of Health Ricerca 5XMILLE_2021_2.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2025.1645233/full#supplementary-material

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancerstatistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. (2024) 74(3):229–63. doi: 10.3322/caac.21834

2. Sharma R. Breast cancer incidence, mortality and mortality-to-incidence ratio (MIR) are associated with human development, 1990–2016: evidence from global burden of disease study 2016. *Breast Cancer*. (2019) 6:428–45. doi: 10.1007/s12282-018-00941-4

3. Heidary Z, Ghaemi M, Hossein Rashidi B, Kohandel Gargari O, Montazeri A. Quality of life in breast cancer patients: a systematic review of the qualitative studies. *Cancer Control*. (2023) 30:10732748231168318. doi: 10.1177/10732748231168318

4. Reeves MM, Terranova CO, Erickson JM, Job JR, Brookes DS, McCarthy N, et al. Living well after breast cancer randomized controlled trial protocol: evaluating a telephone-delivered weight loss intervention versus usual care in women following treatment for breast cancer. *BMC Cancer*. (2016) 16(1):830. doi: 10.1186/s12885-016-2858-0

5. Zemlin C, Schleicher JT, Altmayer L, Stuhlert C, Wormann C, Lang M, et al. Improved awareness of physical activities is associated with a gain of fitness and a stable body weight in breast cancer patients during the first year of antineoplastic therapy: the BEGYN-1 study. *Front Oncol*. (2023) 13:1198157. doi: 10.3389/fonc.2023.1198157

6. Agussalim NQ, Ahmad M, Prihantono P, Usman AN, Rafiah S, Agustin DI. Physical activity and quality of life in breast cancer survivors. *Breast Dis*. (2024) 43(1):161–71. doi: 10.3233/BD-249005

7. Nie X, Yang T, Nie X, Yuan J. Comparative effects of different types of physical activity on health-related quality of life in breast cancer survivors: a systematic review, network meta-analysis, and meta-regression. *Heliyon*. (2024) 10(10):e31555. doi: 10.1016/j.heliyon.2024.e31555

8. De Cicco P, Catani MV, Gasperi V, Sibilano M, Quaglietta M, Savini I Nutrition and breast cancer: a literature review on prevention, treatment and recurrence. *Nutrients* 2019, 11(7), 1514. doi: 10.3390/nu11071514

9. Cella D, Nowinski CJ. Measuring quality of life in chronic illness: the functional assessment of chronic illness therapy measurement system. *Arch Phys Med Rehabil*. (2002) 83(12 Suppl 2):S10–17. doi: 10.1053/apmr.2002.36959

10. Zarghani EH, Geraily G, Hadisinia T. Comparison of different TBI techniques in terms of dose homogeneity - review study. *Cancer Radiotherapie*. (2021) 25(4):380–9. doi: 10.1016/j.canrad.2020.12.004

11. Carreira H, Williams R, Muller M, Harewood R, Stanway S, Bhaskaran K. Associations between breast cancer survivorship and adverse mental health outcomes: a systematic review. *J Natl Cancer Inst*. (2018) 110(12):1311–27. doi: 10.1093/jnci/djy177

12. Fiszer C, Dolbeault S, Sultan S, Bredart A. Prevalence, intensity, and predictors of the supportive care needs of women diagnosed with breast cancer: a systematic review. *Psychooncology*. (2014) 23(4):361–74. doi: 10.1002/pon.3432

13. Berger AM, Gerber LH, Mayer DK. Cancer-related fatigue: implications for breast cancer survivors. *Cancer*. (2012) 118(8 Suppl):2261–9. doi: 10.1002/cncr.27475

14. Sato K, Fukumori S, Matsusaki T, Maruo T, Ishikawa S, Nishie H, et al. Non-immersive virtual reality mirror visual feedback therapy and its application for the treatment of complex regional pain syndrome: an open-label pilot study. *Pain Med*. (2010) 11(4):622–9. doi: 10.1111/j.1526-4637.2010.00819.x

15. Yi JC, Syrjala KL. Anxiety and depression in cancer survivors. *Med Clin North Am*. (2017) 101(6):1099–113. doi: 10.1016/j.mcna.2017.06.005

16. Wang X, Wang N, Zhong L, Wang S, Zheng Y, Yang B, et al. Prognostic value of depression and anxiety on breast cancer recurrence and mortality: a systematic review and meta-analysis of 282,203 patients. *Mol Psychiatry*. (2020) 25(12):3186–97. doi: 10.1038/s41380-020-00865-6

17. Izci F, Ilgun AS, Findikli E, Ozmen V. Psychiatric symptoms and psychosocial problems in patients with breast cancer. *J Breast Health*. (2016) 12(3):94–101. doi: 10.5152/tjbh.2016.3041

18. Boquiren VM, Esplen MJ, Wong J, Toner B, Warner E, Malik N. Sexual functioning in breast cancer survivors experiencing body image disturbance. *Psychooncology*. (2016) 25(1):66–76. doi: 10.1002/pon.3819

19. Lei YY, Ho SC, Kwok C, Cheng A, Cheung KL, Lee R, et al. Association of high adherence to vegetables and fruits dietary pattern with quality of life among Chinese women with early- stage breast cancer. *Qual Life Rese*. (2022) 31(5):1371–84. doi: 10.1007/s11136-021-02985-0

20. Porciello G, Montagnese C, Crispo A, Grimaldi M, Libra M, Vitale S, et al. Correction: mediterranean diet and quality of life in women treated for breast cancer: a baseline analysis of DEDiCa multicentre trial. *PLoS One*. (2021) 16(8):e0256944. doi: 10.1371/journal.pone.0256944

21. Porciello G, Coluccia S, Vitale S, Palumbo E, Luongo A, Grimaldi M, et al. Baseline association between healthy eating Index-2015 and health- related quality of life in breast cancer patients enrolled in a randomized trial. *Cancers (Basel)*. (2024) 16(14):2576. doi: 10.3390/cancers16142576

22. Weerarathna IN, Kamble AR, Luharia A. Artificial intelligence applications for biomedical cancer research: a review. *Cureus*. (2023) 15(11):e48307. doi: 10.7759/cureus.48307

23. Fan P, Zhang J, Gao L, Wang M, Kong H, He S. Exploring the frontier of plant phase separation: current insights and future prospects. *New Crops*. (2024) 1:100026. doi: 10.1016/j.ncrops.2024.100026

24. Gaur K, Jagtap MM. Role of artificial intelligence and machine learning in prediction, diagnosis, and prognosis of cancer. *Cureus*. (2022) 14(11):e31008. doi: 10.7759/cureus.31008

25. Augustin LS, Libra M, Crispo A, Grimaldi M, De Laurentiis M, Rinaldo M, et al. Low glycemic index diet, exercise and vitamin D to reduce breast cancer recurrence (DEDiCa): design of a clinical trial. *BMC Cancer*. (2017) 17(1):69. doi: 10.1186/s12885-017-3064-4

26. Devlin NJ, Brooks R. EQ-5D and the EuroQol group: past, present and future. *Appl Health Econ Health Policy*. (2017) 15(2):127–37. doi: 10.1007/s40258-017-0310-5

27. Scalone L, Cortesi PA, Ciampichini R, Belisari A, D'Angiolella LS, Cesana G, et al. Italian population-based values of EQ-5D health states. *Value Health*. (2013) 16(5):814–22. doi: 10.1016/j.jval.2013.04.008

28. Adhikari D, Jiang W, Zhan J, Rawat DB, Aickelin U, Khorshidi HA. A comprehensive survey on imputation of missing data in internet of things. *ACM Comput Surv*. (2022) 55(7):1–38. doi: 10.1145/3533381

29. Ashfaq KA. Missing data estimation with extremely randomized tree regressor and dimensionality reduction. *Microarray Gene Expression Data Adv Mech*. (2021) 9(3):1680–91. doi: 10.14704/WEB/V19I1/WEB19271

30. Mostafa S, Abdelrahman SE, Hamad S, Amano H. CBRG: a novel algorithm for handling missing data using Bayesian ridge regression and feature selection based on gain ratio. *IEEE Access*. (2020) 8:216969–85. doi: 10.1109/ACCESS.2020.3042119

31. Batista GE, Monard MC. An analysis of four missing data treatment methods for supervised learning. *Appl Artif Intell*. (2003) 17(5-6):519–33. doi: 10.1080/713827181

32. Jager S, Allhorn A. Biessmann F: a benchmark for data imputation methods. *Front Big Data*. (2021) 4:693674. doi: 10.3389/fdata.2021.693674

33. Łuczak M. Hierarchical clustering of time series data with parametric derivative dynamic time warping. *Expert Syst Appl*. (2016) 62:116–30. doi: 10.1016/j.eswa.2016.06.012

34. Cai-Xia L, Shu-Yi A, Bao-Jun Q, Wei W. Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model. *BMC Infect Dis*. (2021) 21(1):839. doi: 10.1186/s12879-021-06503-y

35. Gillespie BM, Chaboyer W, Longbottom P, Wallis M. The impact of organisational and individual factors on team communication in surgery: a qualitative study. *Int J Nurs Stud*. (2010) 47(6):732–41. doi: 10.1016/j.ijnurstu.2009.11.001

36. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: With Applications in R*. 2nd ed. New York, NY: Springer (2021).

37. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell*. (2013) 35(8):1798–828. doi: 10.1109/TPAMI.2013.50

38. Gareth J, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: With Applications in R*. 2nd ed. New York: Springer (2013).

39. Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim Res*. (2005) 30:79–82. doi: 10.3354/cr030079

40. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J Mach Learn Res*. (2019) 20(177):1–81. doi: 10.48550/arXiv.1801.01489

41. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geosci Model Dev*. (2014) 7(3):1247–50. doi: 10.5194/gmdd-7-1525-2014

42. Joaquim A, Leao I, Antunes P, Capela A, Viamonte S, Alves AJ, et al. Impact of physical exercise programs in breast cancer survivors on health-related quality of life, physical fitness, and body composition: evidence from systematic reviews and meta-analyses. *Front Oncol*. (2022) 12:955505. doi: 10.3389/fonc.2022.955505

43. De Nys L, Anderson K, Ofosu EF, Ryde GC, Connelly J, Whittaker AC. The effects of physical activity on cortisol and sleep: a systematic review and meta-analysis. *Psychoneuroendocrinology*. (2022) 143:105843. doi: 10.1016/j.psyneuen.2022.105843

44. Aydin M, Kose E, Odabas I, Meric Bingul B, Demirci D, Aydin Z. The effect of exercise on life quality and depression levels of breast cancer patients. *Asian Pac J Cancer Prev*. (2021) 22(3):725–32. doi: 10.31557/APJCP.2021.22.3.725

45. Lee KJ, An KO. Impact of high-intensity circuit resistance exercise on physical fitness, inflammation, and immune cells in female breast cancer survivors: a randomized control trial. *Int J Environ Res Public Health*. (2022) 19(9):5463. doi: 10.3390/ijerph19095463

46. Ibrahim EM, Al-Homaidh A. Physical activity and survival after breast cancer diagnosis: meta-analysis of published studies. *Med Oncol*. (2011) 28(3):753–65. doi: 10.1007/s12032-010-9536-x

47. Rock CL, Doyle C, Demark-Wahnefried W, Meyerhardt J, Courneya KS, Schwartz AL, et al. Nutrition and physical activity guidelines for cancer survivors. *CA Cancer J Clin*. (2012) 62(4):243–74. doi: 10.3322/caac.21142

48. Tsigalou C, Konstantinidis T, Paraschaki A, Stavropoulou E, Voidarou C, Bezirtzoglou E. Mediterranean diet as a tool to combat inflammation and chronic diseases. An overview. *Biomedicines*. (2020) 8(7):201. doi: 10.3390/biomedicines8070201

49. Gonzalez-Palacios Torres C, Barrios-Rodriguez R, Munoz-Bravo C, Toledo E, Dierssen T, Jimenez-Moleon JJ. Mediterranean Diet and risk of breast cancer: an umbrella review. *Clin Nutr*. (2023) 42(4):600–8. doi: 10.1016/j.clnu.2023.02.012

50. Bell ML, Kenward MG, Fairclough DL, Horton NJ. Differential dropout and bias in randomised controlled trials: when it matters and when it may not. *Br Med J*. (2013) 346:e8668. doi: 10.1136/bmj.e8668

51. Khalifa M, Albadawy M. Artificial intelligence for clinical prediction: exploring key domains and essential functions. *Comput Methods Programs Biomed Update*. (2024) 5:100148. doi: 10.1016/j.cmpbup.2024.100148

52. Mousavi H, Karandish M, Jamshidnezhad A, Hadianfard AM. Determining the effective factors in predicting diet adherence using an intelligent model. *Sci Rep*. (2022) 12(1):12340. doi: 10.1038/s41598-022-16680-8

53. Kim J, Jeong K, Lee S, Baek Y. Machine-learning model predicting quality of life using multifaceted lifestyles in middle-aged south Korean adults: a cross-sectional study. *BMC Public Health*. (2024) 24(1):159. doi: 10.1186/s12889-023-17457-y