npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01299-y

A machine-learned model for predicting weight loss success using weight change features early in treatment

Farzad Shahabi[1,2] ✉, Samuel L. Battalio[1], Angela Fidler Pfammatter[3], Donald Hedeker[4], Bonnie Spring[1] & Nabil Alshurafa[1,2]

Stepped-care obesity treatments aim to improve efficiency by early identification of non-responders and adjusting interventions but lack validated models. We trained a random forest classifier to improve the predictive utility of a clinical decision rule (>0.5 lb weight loss/week) that identifies non-responders in the first 2 weeks of a stepped-care weight loss trial (SMART). From 2009 to 2021, 1058 individuals with obesity participated in three studies: SMART, Opt-IN, and ENGAGED. The model was trained on 80% of the SMART data (224 participants), and its in-distribution generalizability was tested on the remaining 20% (remaining 57 participants). The out-of-distribution generalizability was tested on the ENGAGED and Opt-IN studies (472 participants). The model predicted weight loss at month 6 with an 84.5% AUROC and an 86.3% AUPRC. SHAP identified predictive features: weight loss at week 2, ranges/means and ranges of weight loss, slope, and age. The SMART-trained model showed generalizable performance with no substantial difference across studies.

Obesity is a growing population health concern and is among the leading preventable causes of premature death globally[1]. Gold-standard treatment for obesity involves intensive behavioral treatment that can be costly and burdensome to administer, impeding the scalability needed to redress obesity on a global scale[2]. Stepped-care treatment models may be a viable solution. A rational resource allocation strategy, stepped-care, involves starting with a less resource-intensive intervention, and, for those who respond suboptimally, stepping them up by adding more vigorous, expensive intervention components. A major challenge facing obesity stepped-care intervention, however, is the lack of a predictive treatment algorithm (policy) that accurately identifies non-responders early in treatment. Improved early detection of non-response to weight loss treatment will guide stepped-care decisions for treatment that can be scaled for population-level impact. This would result in a treatment allocation algorithm (policy) that gives people the treatment resources they need—not more and not less[3].

Prior stepped-care obesity treatment approaches have had two main limitations: (a) poor accuracy of the algorithm predicting non-response, or (b) prolonged delay (2–3 months) before augmenting treatment[4], a problematic design feature because initial weight loss predicts long-term weight loss. There is disagreement as to whether weight loss variability and other statistical measures increase the predictive power of weight loss. Moreover, it is unknown how we operationalize initial weight loss, weight loss variability, trends, and other statistical measures. Given the recent adoption of mobile health tools, such as wireless weight scales, we can further identify, with greater granularity regarding daily changes, critical predictors for stepped-care treatment interventions. Creating adequate tools for early prediction of weight loss treatment success prevents researchers from resorting to inaccurate weight loss prediction algorithms or subjective judgment. SHapley Additive exPlanations (SHAP)[5] is an explainable AI tool that helps us understand the direction and magnitude of the marginal contribution of each feature to the weight loss outcome. Explainable AI is increasingly enhancing clinicians' understanding of machine learning (ML) models and augmenting their decision-making processes. By improving transparency in model predictions, it increases clinicians' trust in the outcome of ML models and advances their integration into clinical practice.

Although conventional statistical methods remain a powerful predictive tool, recent weight loss research has developed ML-based prediction models using longitudinal data that uncover new knowledge that can improve the models' predictive performance[6,7]. Our study had two primary aims. First, we aimed to build a supervised ML model that incorporates both dynamic features obtained during the initial treatment period combined with baseline features to improve the predictive utility of a clinical decision rule that identifies non-responders in the first 2 weeks of a stepped-care

[1]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA. [2]Department of Computer Science, McCormick School of Engineering, Northwestern University, Evanston, IL, USA. [3]College of Education, Health, and Human Sciences, University of Tennessee, Knoxville, TN, USA. [4]Department of Public Health Sciences, The University of Chicago, Chicago, IL, USA. ✉e-mail: farzad.shahabi@northwestern.edu

weight loss trial. Second, we aimed to evaluate the generalizability of our model by testing it on populations from other weight loss trials.

Results

Data from 400 participants with obesity were obtained from the SMART study, of which 281 participants were eligible for inclusion in the model development dataset (15 withdrew, 43 missed 6-month follow-up, and 61 did not produce sufficient weight data from the Fitbit Aria wireless weight scale during the first 2 weeks of the study). We allocated ~60% (169/281) of the data in the training dataset, 20% (55/281) in the validation set, and 20% (57/281) in the holdout test set. Approximately 64% (179 participants) of the whole development dataset reached suboptimal weight loss (encoded as zero), and 36% (102 participants) achieved sufficient weight loss (encoded as one) based on the 5% weight outcome at month 6. Table 1 shows the baseline characteristics of patients in the training, validation, and test sets. The profiles for the SMART, Opt-IN, and ENGAGED studies are provided (Supplementary Figs. 1–3).

Proposed model compared to reference methods

Results for the three reference methods show imbalanced sensitivity and specificity as indicated in Table 2. The average of the three models yields a low sensitivity of 26.1% (95% CI, 14.1–38.1) and high specificity of 86.2% (95% CI, 82.9–89.4) suggesting a tendency towards higher false negatives (misidentifying actual sufficient weight loss as suboptimal) and lower false positives (misidentifying actual inadequate weight loss as sufficient), implying that the models predominantly predict a single outcome, suboptimal weight loss, akin to a negatively biased decision maker. The baseline ML models, both discriminative and generative, which were trained solely using static features (Supplementary Table 2), achieved an average weighted F1 score of 59.0% (95% CI, 57.6–60.4), and AUROC of 62.8% (95% CI, 59.7–65.9). The generative and discriminative ML models, when trained on dynamic and static features at week 2, achieved an average F1 weighted score of 68.9% (95% CI, 66.2–71.6) and a AUROC 76.0% (95% CI, 72.2–79.9), demonstrating the value of dynamic features in enhancing prediction accuracy (Supplementary Table 3). By incorporating dynamic features in addition to static features in the model, we observed a 9.9% improvement in the F1 weighted score and a 13.2% increase in the AUROC. Consequently, we selected the best-performing ML model, the proposed random forest, which uses the complete feature set. This model outperformed all reference methods yielding a weighted F1 score of 75.7%, AUROC of 84.5%, AUPRC of 86.3%, and Brier score loss of 0.173, as indicated in Table 2. To ensure an unbiased approach to modeling, further analysis was performed on 15 other random train/validation/test splits, yielding similar performance and explainability outcomes (Supplementary Note 1).

Explainability

Through SHAP, we generated a ranking of feature importance and corresponding explanations using SHAP values and feature magnitudes. Our analysis highlights the significance of certain features. For instance, SHAP magnitudes and directionality support the expectation that a higher mag-

Table 1 | Baseline characteristics of the individuals who participated in the SMART study (development dataset), Opt-IN and ENGAGED studies (generalizability datasets)

	Development dataset [SMART] (*n* = 281)			Generalizability datasets [Opt-IN & ENGAGED] (*n* = 472)	
Sample size	**Train (*n* = 169)**	**Validation (*n* = 55)**	**Test (*n* = 57)**	**Opt-IN (*n* = 442)**	**ENGAGED (*n* = 30)**
Age range, years	20–60	23–60	18–60	18–60	22–57
Sex, *n* (%)					
Female	123 (72.8)	40 (72.7)	43 (75.4)	357 (80.8)	27 (90)
Male	43 (25.5)	15 (27.3)	14 (24.6)	85 (19.2)	3 (10)
Other	3 (1.7)	0 (0)	0 (0)	0 (0)	0 (0)
Race/ethnicity, *n* (%)					
Non-Hispanic White	103 (60.9)	34 (61.8)	35 (61.4)	335 (75.8)	14 (47.7)
Non-Hispanic Black	32 (18.9)	12 (21.8)	11 (19.3)	51 (11.5)	12 (40)
Hispanic	21 (12.5)	5 (9.2)	6 (10.5)	40 (9.1)	4 (13.3)
Non-Hispanic Other	13 (7.7)	4 (7.2)	5 (8.7)	16 (3.6)	0 (0)
Weight loss success, *n* (%)					
No	104 (61.5)	34 (61.8)	34 (59.6)	209 (47.3)	19 (63.3)
Yes	65 (38.5)	21 (38.2)	23 (40.4)	233 (52.7)	11 (36.7)

Data are *n* (%) and (range) for baseline characteristics.

Table 2 | Comparison of the proposed random forest model with three reference methods

Model	Sensitivity, %	Specificity, %	F1 weighted, %	AUROC, %	AUPRC, %	Brier score loss	Time endpoint
Historic clinical decision rule (0.5 lb/week)	5.5	81.4	44.4	NA	NA	NA	Two weeks
Logistic regression (weight loss on day 14)	40.9	85.7	66.5	76.1	70.7	0.217	Two weeks
Logistic regression (Static features)	31.8	91.4	64.8	68.4	70.0	0.239	Baseline
Proposed random forest (full feature set)	82.2	71.1	75.7	84.5	86.3	0.173	Two weeks

The clinical decision rule, which selects responders based on weight loss exceeding 0.5 lb/week, and the logistic regression model trained solely on weight loss data from day 14 on SMART data. The third reference model was trained on static baseline features. Logistic regression outperformed all other ML models, both generative and discriminative, that were trained using static features at the baseline. The fourth model is the proposed random forest model trained on the full feature set by the two-week endpoint.

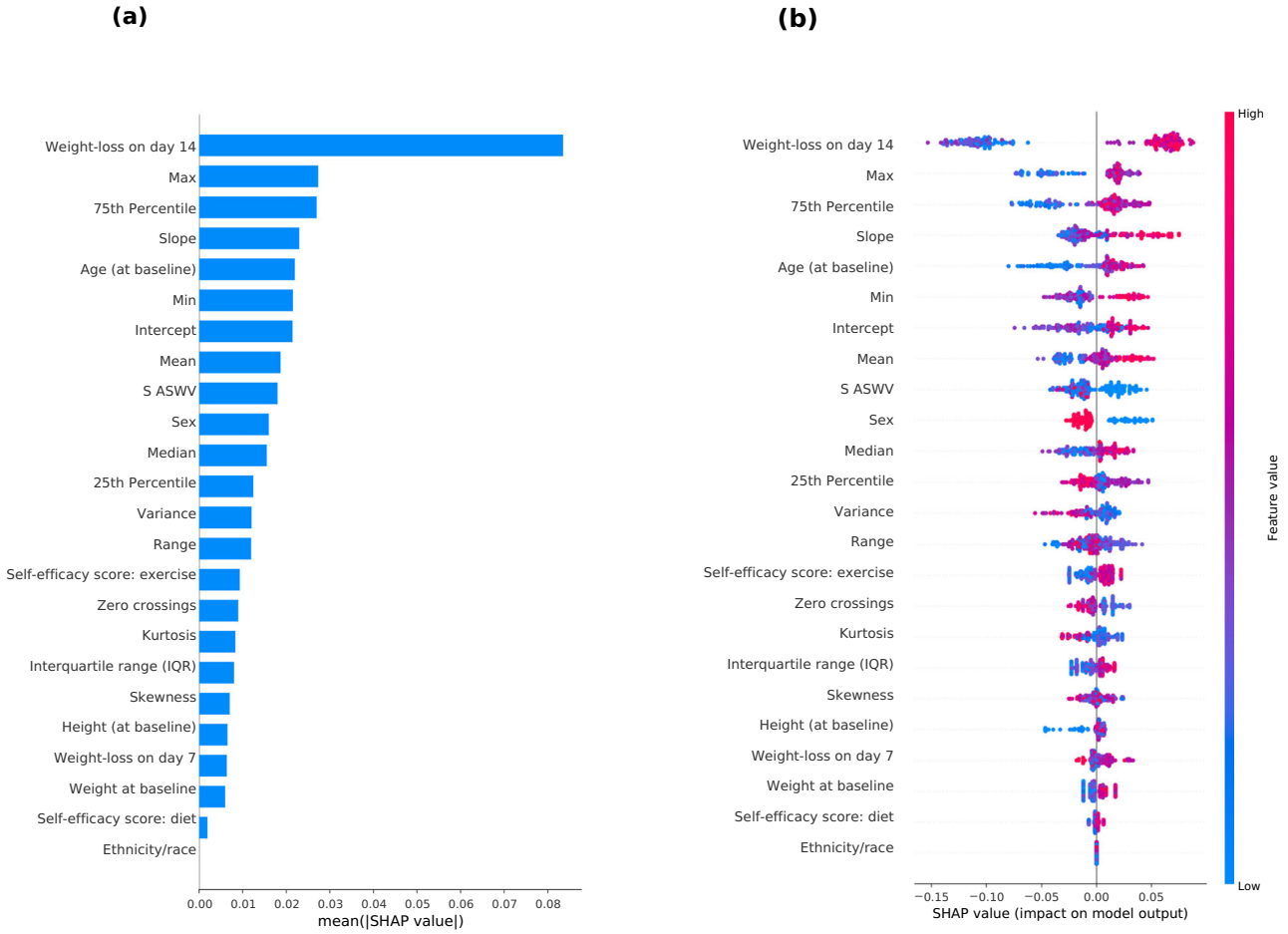


Fig. 1 | Explainability plots by SHAP for the features used in the study. a Feature importance of the variables used in the random forest model trained on the SMART dataset. **b** Correlation between the feature magnitude and SHAP values, indicating the direction and impact of features on the model's predictions.

nitude of weight loss on the 14th day increases the likelihood of successful weight loss shown in Fig. 1. Furthermore, we found that features derived from the weight trajectory over a two-week period, such as high minimum, maximum, mean, 75th percentile, IQR, and slope, played a crucial role in weight loss prediction based on SHAP plots. Additionally, by examining the SASWV feature, we observed that higher SASWV was indicative of weight gain. Our examination also revealed multiple baseline features such as age (being older), sex (being male), height (being tall), and high self-efficacy in exercise are predictive of weight loss.

Generalizability

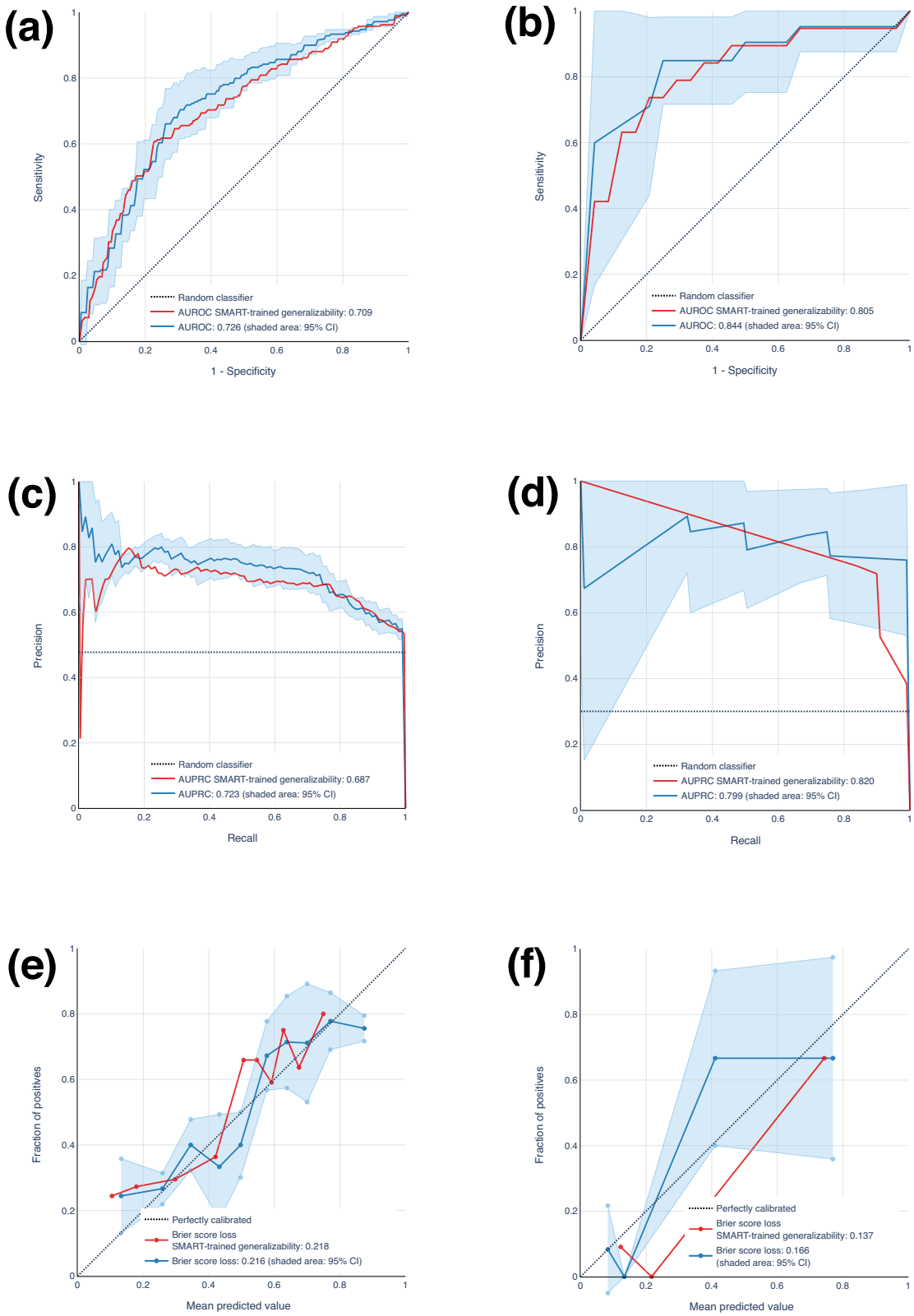
The results of our generalizability analysis are indicated in Fig. 2, and Table 3. The results for Opt-IN revealed that the 5-fold cross-validation procedure yielded an average AUROC of 72.6% (95% CI, 70.3–74.9), AUPRC of 72.3% (95% CI, 68.3–76.3), and a Brier score loss of 0.216 (95% CI, 0.207–0.225). For generalizability of Opt-IN, AUROC, AUPRC, and Brier score loss were 70.9% 68.7%, and 0.218, respectively. Applying a similar methodology to the ENGAGED dataset, for 3-fold cross-validation, we achieved an AUROC of 84.4% (95% CI, 75.6–93.2), an AUPRC of 79.9% (95% CI, 70.9–88.9), and a Brier score loss of 0.166 (95% CI, 0.126–0.206). For ENGAGED generalizability, AUROC, AUPRC, and Brier score loss were 80.5%, 82.0%, and 0.137%, respectively.

Discussion

In this study, we developed a machine-learned model that leverages recent advances in explainability and generalizability to increase validity, reliability, and trust in our ability to detect non-responders to weight loss programs with a 6-month endpoint and trigger adjustment in treatment. Importantly, our algorithm outperformed the clinical decision rule that has been used in past stepped-care trials. The algorithm was also generalizable, given that its effectiveness was replicated in two other distinct weight loss trials. For these reasons, our model holds promise as a valuable tool for personalized treatment allocation and improved outcomes in stepped-care weight management treatments.

The healthcare sector is witnessing a growing concern surrounding the adoption of ML algorithms that suffer from low interpretability[8]. Clinicians and healthcare professionals hesitate to rely on decisions rendered by these models, given the lack of insight and explanations driving the outcome prediction and the high stakes of clinical decision-making. Although computer scientists and researchers often report the most highly performant model, clinicians and experts will likely not use a tool they do not understand[9]. Fortunately, recent advances in interpreting complex models make it more feasible to incorporate explainability into the decision-making process. Explainability methods are known to enhance trust in the model by allowing clinicians to understand that the model's accurate predictions are grounded in valid reasoning and that any errors made by the model are based on understandable and justifiable factors found in the data[10].

The SMART-trained model exhibited stable performance with no substantial difference compared to the models trained and tested on the Opt-IN and ENGAGED datasets separately, as indicated in Table 3 and Fig. 2. This finding suggests features derived from objectively measured weights from wireless weight scales were discriminative and generalized effectively across external data from Opt-IN and ENGAGED studies that used self-reported weights. Furthermore, the SMART-trained model outperformed the model trained and tested on ENGAGED in terms of AUPRC



and Brier score loss. This might be due to the smaller sample size of 30 participants in ENGAGED, which can make it challenging to train a well-calibrated classifier capable of delivering stable predictions. Therefore, a generalizable model trained on a larger sample size with objective measurement is better equipped to withstand minor changes in study design and measurement procedures.

Simple rule-based approaches, such as the clinical decision rule (responders identified as losing ≥1 pound in 2 weeks) result in low sensitivity (5.5%) and an F1 weighted score of 44.4%. The simple model generates a high number of false negatives, predicting non-responders when they were actually responders. Such a model would result in an increase of treatment components or care intensity for participants who do not need them,

Fig. 2 | Generalizability assessment of a random forest model for weight loss prediction using external datasets (Opt-IN and ENGAGED). **a**, **b** ROC curves for both models, with the blue curve (and a shaded 95% CI) representing the model trained/tested on Opt-In and ENGAGED. The red curve represents the ROC curve for the generalizability. The generalizability curve consistently fell within the 95% CI, achieving an AUROC of 0.709 and 0.805, which closely aligns with the AUROC of the model trained/tested on Opt-IN and ENGAGED, respectively (mean: Opt-IN, 0.726; ENGAGED, 0.844). **c**, **d** Correspondingly, the Precision-Recall curves for the Opt-IN and ENGAGED, based on models trained and tested on these cohorts individually, were compared with those from the SMART-trained model's generalizability performance. The generalizability curves consistently fell within the 95% confidence intervals, achieving AUPRC values of 0.687 and 0.820, respectively. These values are in close alignment with the AUPRC of the models trained and tested on Opt-IN and ENGAGED (mean: 0.723 and 0.799, respectively). **e**, **f** In a similar analysis for Brier score loss, the total losses recorded were 0.218 and 0.137 for the generalizability assessment, closely corresponding to the results from models trained and tested on Opt-IN and ENGAGED, with mean values of 0.216 and 0.166, respectively. Therefore, the SMART model generalized effectively to the external datasets without substantial loss in performance on Opt-IN and ENGAGED.

Table 3 | This table presents the results of generalizability for the model trained using the SMART dataset when applied to the external cohorts, Opt-IN, and ENGAGED

SMART-trained model generalizability results	Opt-IN	ENGAGED
Sensitivity (%)	77.7	90.9
Specificity (%)	60.3	78.9
F1 weighted (%)	69.2	83.6
AUROC (%)	70.9	80.5
AUPRC (%)	68.7	82.0
Brier score loss	0.218	0.137

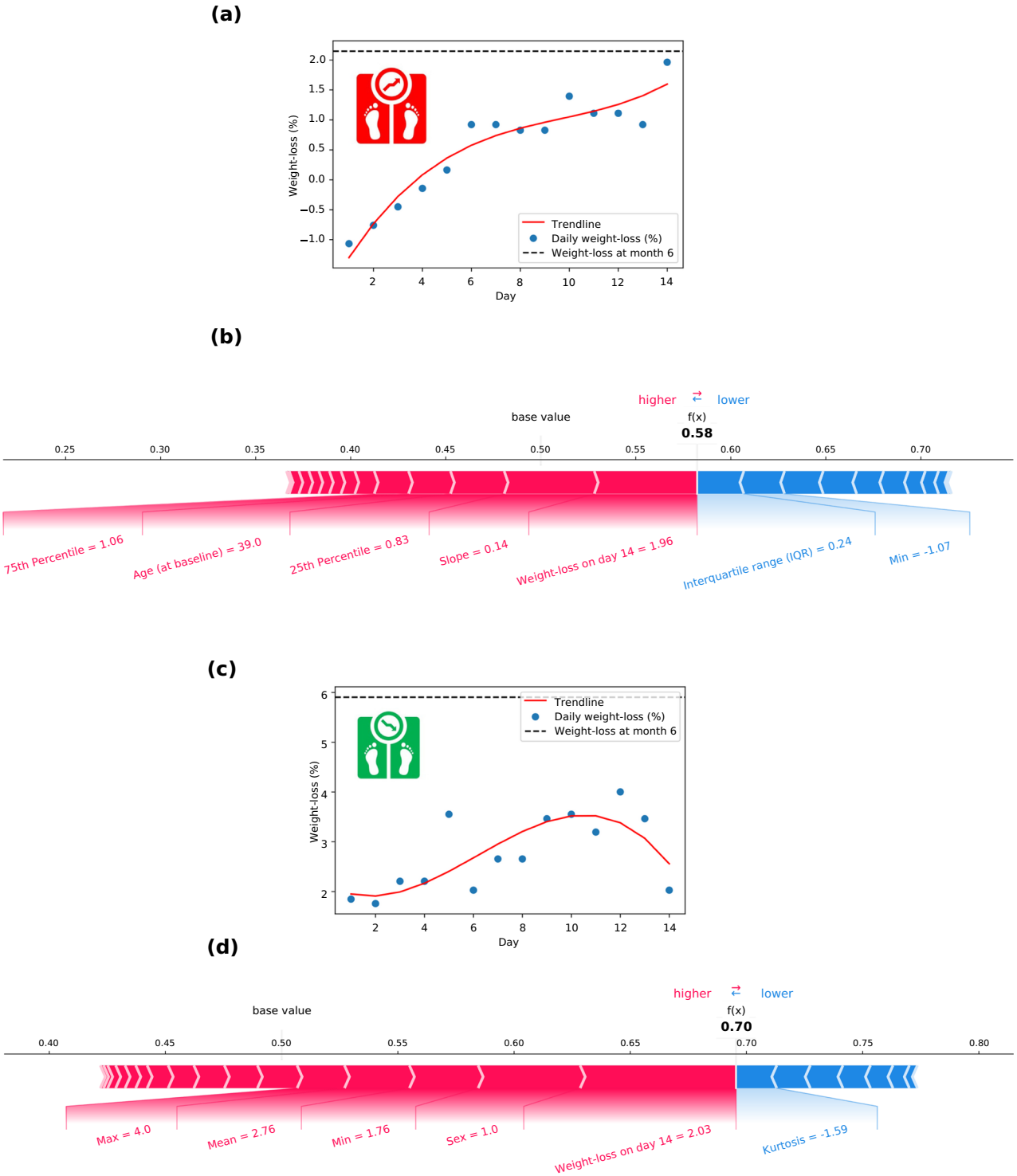
resulting in increased cost and care team burden. Given that initial weight loss has been shown to predict long-term weight loss[11], we designed a logistic regression model that uses weight loss at the end of the second week as the sole predictor. Although this model improved the clinical decision rule (F1 weighted score, 66.5%) and provided greater interpretability, it comes at the expense of low accuracy and low sensitivity (40.9%). We then opted for building a model using the random forest classifier, which is known to prevent overfitting compared with typical decision tree models, generalizes well, and has shown success in various domains including finance and healthcare[12,13]. Fig. 3 shows a comparative analysis between the selected random forest model and the logistic regression model.

We then set out to build a model using data only collected at baseline to test prediction before treatment initiation. The results were promising and uncovered some important features; however, the model had low accuracy, comparable sensitivity (40.9%), and a slightly lower F1 weighted score (59.0%). Our findings suggest that the statistical features extracted from the 2-week weight loss trajectory possess discriminative power, leading to a notable 16.7% increase in the F1 weighted score when compared with a random forest model trained solely on demographic and psychological features. Notably, certain demographic attributes such as age, gender, and height exhibit considerable discriminatory potential, implying that older and taller individuals, as well as males, are more likely to succeed in losing weight. These observations are in line with findings from prior work describing the reasons females and younger adults have greater difficulty losing weight[14,15]. Increased body weight variability is a result of fluctuations in weight loss and gain and is influenced by a diverse range of factors, encompassing variations in dietary, physical activity, and other behavioral patterns (e.g., time of daily weighing)[16]. The consequential impact of weight variability extends to numerous health outcomes, including conditions like depression[17], non-alcoholic fatty liver[18], and cardiovascular disease[19].

Measuring weight variability is challenging due to outliers in weight data, inconsistent recording of weights by individuals, and non-linear trends in weight over time. These factors make accurate quantification of weight variability complex. Researchers vary in how they operationalize or measure weight variability, adopting distinct definitions tailored to their specific research objectives. However, the rationale driving the selection of these definitions often remains unexplained, contributing to lack of consensus about the optimal characterization of weight variability to inform the investigation of weight loss prediction. Researchers have shown that short-term elevated weight variability is predictive of poor long-term outcome[20]. Additionally, variation between weekdays and weekends, such that high compensatory weight loss during weekdays counteracts the upward trajectory of weight gain during weekends, predicts long-term weight loss and maintenance over time[21]. This may be attributed to the operationalization of weekly weight variability, which focuses on comparing non-consecutive days of the week rather than emphasizing variations in consecutive day-to-day changes. Although weekly weight variability is an interesting area of inquiry, our desire to predict weight loss as early in the intervention as possible renders it unfeasible to estimate weekday/weekend weight variability. The prevailing approach for long-term weight loss prediction has typically relied on the use of RMSE through linear regression methods. However, this method is not without limitations, primarily due to its inability to capture non-linearities in weight trajectory fluctuations. A different approach proposed employs LOESS regression, yet like other least squares methods, this technique is sensitive to outliers[22]. The estimation of body weight variability, regardless of the chosen definition, frequently encounters challenges stemming from the presence of missing data[23]. As the frequency of missing weight records increases, weight variability can decrease depending on the imputation method employed. Whether using self-report or technology-based smart devices to more objectively operationalize the measure of weight, there is a need to address missing data. We therefore scale our definition of a successive weight variability feature by the count of missing weight records spanning the 2-week interval. This underscores that higher successive weight variability and/or suboptimal utilization of technology-based tools (higher missing rate) are associated with reduced success in weight loss.

Our findings suggest another important aspect to consider: an elevated level of missing data likely indicates a decrease in adherence to the study protocol, perhaps reflecting faltering motivation that would be expected to result in failing to achieve weight loss in 6 months. However, isolated total missingness failed to demonstrate predictive capacity. This becomes more predictive when coupled with successive weight variability and suggests that the predictive factor for long-term weight gain may not solely stem from elevated weight variability but rather arises from the interplay between heightened weight variability and diminished adherence to prescribed study protocols.

Misclassification in behavioral weight loss trials can have significant implications, particularly concerning the risks associated with both under- and over-treatment. Incorrectly classifying an individual as a responder may result in the failure to provide necessary interventions, leading to suboptimal weight loss outcomes and potentially exacerbating obesity-related health conditions[24]. Conversely, the consequences of misclassifying an individual as a non-responder, leading to more intensive treatment than necessary, are not as well understood and have not been extensively studied. Although previous studies have found that the risks associated with intensive behavioral interventions, including weight loss maintenance programs for adults with obesity, are minimal[2], over-treatment may introduce risks at multiple levels, at both the individual and population levels. At the population level, a prominent adverse consequence of over-treatment is the inefficient allocation of resources to superfluous care, which can result in the depletion of critical resources[25], subsequently limiting their availability for those who



really need it. On an individual level, providing support that exceeds what is necessary can compromise autonomy. Individuals who perceive their weight loss as being reliant on external assistance may fail to cultivate the self-confidence and self-efficacy that could be gleaned by losing weight of their own accord, without external support. Achieving the right balance between providing support and encouraging autonomy is essential in weight loss interventions[26].

Weight loss is a multifaceted process influenced by diverse physiological, psychological, and environmental factors. Our analysis is limited in that our feature set does not encompass elements capturing dietary patterns or variations in physical activity, which could significantly enhance the study's comprehensiveness. However, our choice of static/dynamic features aligns with those commonly found in most standard clinical settings. Furthermore, it is essential to acknowledge that ML models are inherently opaque. Although SHAP aids in mitigating this opacity, it does not address potential algorithmic biases embedded in the model, often due to underlying assumptions in the algorithm's design. It is important to acknowledge that excessive reliance on SHAP interpretations may lead to overinterpretation, potentially resulting in the emergence of narrative fallacies[27].

Our primary outcome variable is measured at 6 months, which presents a limitation in predicting the long-term sustainability of weight

Fig. 3 | Comparative analysis of weight loss predictions: random forest vs. logistic regression models across initial two-week trajectory. The plots **a** and **c** delineate the trajectory of weight change over a 14-day period from the baseline for two participants, one achieving suboptimal weight loss and the other achieving sufficient weight loss at the 6-month mark. Accompanying the weight change data are SHAP force plots (**b** and **d**), illustrating the features contributing to the predictive models' outcomes for these individuals. Additionally, daily weight loss percentages are depicted using blue dots, while a red line signifies the trend in weight loss trajectories through orthogonal distance regression (ODR), for both participants. A comparative analysis is conducted between the predictive capabilities of the proposed random forest model (trained on the full set of features), and a logistic regression model (trained exclusively on the data pertaining to weight loss on day 14). **a** The logistic regression model functions as a tailored decision rule, categorizing participants who exhibit a weight loss of ~2.44% by the end of the second week as likely to achieve weight loss success at the 6-month mark. The weight loss trajectory over the initial two-week period indicates a potential for weight loss, as suggested by the trend's slope. The random forest model considers a broader set of variables, including the trajectory of weight loss as determined by the slope, the participant's age (**b**), and the weight loss observed on the 14th day. However, the logistic regression model, relying solely on the singular metric of weight loss on day 14—approximately 2%—predicts suboptimal weight loss by the six-month marker. The accuracy of this prediction may be attributed to random chance rather than serving as a reliable indicator of future outcomes. **c** In contrast, for this participant, the random forest model leverages a broader array of variables (**d**), including maximum, minimum, mean weights, and gender, enhancing the precision of its prediction. Conversely, the logistic regression model's reliance on a single variable—weight loss on day 14, estimated at 2%—renders it susceptible to misclassification of this participant's outcome.

loss. Future research should aim to predict weight loss maintenance by extending the analysis to 12-month outcomes and beyond. Given the complexity and uncertainty of human behavior, predicting weight loss is an inherently challenging problem. Although initial weight loss is known as a strong predictor of weight loss maintenance, we shed light on the utility of low weight loss variability in predicting weight loss success through the predictive lens of machine-learned models. Our findings show that we can reliably predict weight loss as early as 2 weeks into a weight loss attempt, laying the foundation for future dynamic stepped-care weight loss models. However, the model is based on a limited universally measured set of features in common in clinical settings and a limited sample size from 3 studies; as such, it is meant as a proof-of-concept analysis. Before this kind of model can be used as a decision-support tool, the effectiveness of early-prediction models needs to be confirmed in a prospective study encompassing a diverse array of populations, a broad spectrum of treatment methodologies, and a range of different temporal intervals.

Methods

Overview

Our data analysis pipeline aimed to demonstrate whether dynamic features (weight loss information obtained early in the intervention) improve the early prediction of weight loss in an interventional weight management program, select the best-performing predictive model, and test its explainability and generalizability. Participants' weights in the SMART study were objectively measured using the Fitbit Aria wireless smart scale. They were instructed to weigh themselves every morning immediately after waking up, post-urination, and without any clothing. The weight loss outcome, operationalized as a dichotomous variable, is assessed at month 6, with success defined as participants achieving a clinically meaningful weight loss of 5% or more from their baseline weight by the 6-month endpoint. We compared our machine-learned model with several reference methods. The first was the clinical decision rule in which non-responders were defined as participants who lost <0.5 pounds/week (0.227 kg/week). The clinical decision rule is operationalized by evaluating participants at the end of week 2 and flagging them as non-responders if they did not lose at least 1 pound from their baseline weight at the end of the 2-week period. Because most prior work suggests the best predictor of weight loss success is the degree of initial (early) weight loss, our second reference method included a logistic regression based on weight loss at week 2. This allowed us to move towards a more tailored personalized approach to assess whether collecting a feature such as weight loss from baseline at week 2 (expressed as a percentage)–temporally closer to the 6-month endpoint–increased the predictive utility of our model. Our third reference method was the *baseline machine-learned model*, which tests the predictive power of static features obtained from participants (e.g., demographics, psychological measures, etc.) at baseline using a supervised machine-learned model. This was essential to assess whether data prior to beginning the intervention was sufficient for the prediction of weight loss. Our final proposed weight loss machine-learned model combined baseline static variables with dynamic statistical variables obtained early (first 2 weeks) during the weight loss treatment period to highlight the predictive power of data-driven statistical features. To evaluate the in-distribution generalizability of the final model, we utilized the randomly selected holdout test set from the SMART study. With the goal of building a trustworthy machine-learned model that could reliably augment clinical practice, we used explainability methods to uncover not only the most predictive variables but also the magnitude and direction of a variable's effect on weight loss. A clinician needs to know the direction and impact by which a model's features predict weight loss to understand how the model "works," and thereby to decide whether it is credible and trustworthy. To evaluate the out-of-distribution generalizability of the final model (SMART-trained), we assessed its performance on the Opt-IN and ENGAGED datasets and compared it to models trained separately on these datasets, allowing us to gain insights into its ability to generalize despite differences in study design and measurement procedures. Moreover, given our interest in understanding behavior change and the existing uncertainty regarding weight variability and its association with long-term weight loss, we propose a new definition for body weight variability incorporating the factor missing rate and gauge its discriminative power and relationship with weight loss at the 6-month mark.

SMART study design

The SMART study[28] was a sequential multiple assignment randomized trial that aimed to optimize an mHealth-intensive lifestyle obesity stepped-care treatment package by identifying the optimal starting condition and the optimal augmentation ("step-up") strategies for treatment non-responders. The study included a 3-month weight loss treatment program with a weight loss outcome measure at 6 months. The sample included 400 adults who were overweight or obese at baseline ([BMI], 27–45 kg/m^2). Participant recruitment spanned 2016–2021 in the Chicagoland area.

Opt-IN study design

The Opt-IN study[29] was a clinical trial that used a full factorial design to identify the combination of remotely delivered, technology-supported weight loss treatment components that maximized weight loss over a 6-month period. The study included 562 adult participants with overweight or obesity. All participants received a base treatment intervention including a custom-built smartphone app that facilitated self-management of diet and activity behaviors, online lessons, and a coach. In addition to the base package, participants were randomly assigned to 32 experimental conditions, covering all possible permutations of five treatment components: moderate vs intense (12 vs 24) number of coaching call sessions, text messaging, buddy (social support) training, primary care provider engagement, and meal replacement. This allowed researchers to identify the optimal combination of these treatment components in terms of the primary outcome, weight loss achieved at 6 months. Participant recruitment spanned 2013–2017 in the Chicagoland area, with the study's protocol and design previously documented.

ENGAGED study design

The ENGAGED study[30], conducted between 2009 and 2013, was a randomized controlled trial involving 96 adults with obesity, comparing three different 6-month weight loss treatments: Self-Guided (SELF), Standard (STND), and Technology-Supported (TECH). The STND and TECH groups attended eight in-person group sessions, while the SELF and STND groups used paper diaries for self-monitoring. In contrast, the TECH group utilized a smartphone app with social networking features and a wireless accelerometer for self-monitoring. The primary goal of the study was to evaluate the effectiveness of each treatment, measured by the amount of weight loss achieved after 6 months.

All studies were conducted in compliance with ethical standards and received approval and oversight from the Northwestern Institutional Review Board under the following study identification numbers: SMART (STU00202075), Opt-IN (STU00066546), and ENGAGED (STU00017350).

Development and external cohorts

We selected the SMART study as the development cohort due to its accurate and objective weight measurement, facilitated by the Fitbit Aria wireless smart scale, compared to the self-reported weight measurements in the other studies. Additionally, the SMART study was the most recent among the weight loss studies, ensuring more up-to-date methodologies and technologies. Since part of the SMART study was run during the COVID-19 pandemic, a post hoc sensitivity analysis was conducted, showing no significant difference in weight loss between groups assessed before and during the lockdown[31]. Opt-IN and ENGAGED were chosen as external cohorts to deploy the SMART-trained model and test its generalizability.

Outlier detection architecture

We removed duplicate records and eliminated erroneous observations with invalid formats in each study separately. In the first step, we leveraged a traditional statistical approach to remove explicit outliers (i.e., unreasonably high/low weight records) by applying a within-person threshold (lower bound = Q1 − 1.5 * interquartile range [IQR]; upper bound = Q3 + 1.5 * IQR, where Q1 = 25th percentile and Q3 = 75th percentile). We supplemented this approach with another validated method, time-windowed geometric path analysis to identify within-person errors in weight measurements in time series data[32]. We measured the time interval and weight difference between three consecutive weight measurements for each participant. Subsequently, for every three contiguous weight measurements, we computed the path ratio. We created a personalized distribution of path ratios for each individual and excluded weight values that had a path ratio that exceeded five z-scores, empirically set, and verified through visual confirmation by two authors (FS and NA) to ensure accuracy and reliability in the analysis. This method was applied uniformly to all datasets (SMART, Opt-IN, and ENGAGED). Consequently, all baseline models and ML methods utilized the same cleaned dataset, ensuring consistency and comparability in the analysis.

Measures and features

We extracted data-driven features to assist the ML models in discriminating between individuals who achieved weight loss success and those with suboptimal weight loss results. Features were added to create generalizable and robust machine-learned models that identify early non-response in the first 2 weeks of a weight loss intervention. Two baseline psychological features (self-efficacy for diet and for exercise) and four baseline demographic variables (height, age, gender, race/ethnicity) were included in the model. The data-driven features included daily weight deviation ratio from baseline weight for days 7 and 14, weight variability, ranges and means of weight loss during the first 2 weeks, and slope/intercept (obtained from linear regression), zero crossing, kurtosis, and skewness, which reveal weight distribution characteristics over the 14-day course. Formal definitions of the features are provided (Supplementary Table 4). To handle missing data, we used linear regression as a curve-fitting method[33] to estimate and impute missing weight data from the known records over 2 weeks. For instance, if a weight on day 14 was missing or removed due to the outlier detection architecture, we used the imputed value.

Weight variability

Previous research has used different methods to define/measure weight variability; however, their impact on weight/health outcomes remains unknown. Some researchers have derived weight variability as the coefficient of variation known as the relative standard deviation[34]. Others have defined weight variability as the mean successive weight change[35]. Another weight variability metric is derived from non-linear mean deviation estimated from locally weighted scatterplot smoother (LOESS) regression[36]. However, weight variability is also widely used as root mean square error (RMSE) calculated from the estimated weight records' distance from the best-fitted linear regression line[37,38]. All derivations are provided (Supplementary Table 1).

We propose a sparsity-adjusted successive weight loss variability estimated by Equation (1), where we measure the within-person successive (day-to-day) weight loss variability and adjust by the count of missingness over 2 weeks, mathematically formalized as follows:

$$\sigma_i^2 = \frac{\sum_{j=1}^{M-1}(y_{i,j} - y_{i,j+1} - \mu_i)^2}{M}, \mu_i = \frac{\sum_{j=1}^{M-1}(y_{i,j} - y_{i,j+1})}{M}, SASWV_i = \sigma_i^2 * M_i \quad (1)$$

where $M = 14$ is the number of observations, $j \in \{1, 2, \ldots, \mathrm{M}\}$, and N is the number of participants, $i \in \{1, 2, \ldots, \mathrm{N}\}$, μ_i is the within-person mean of the successive weight loss differences, σ_i^2 is the variability of the within-person successive weight loss difference and M_i is the count of missingness (i.e., the number of imputed observations) over 2 weeks (Supplementary Fig. 4).

Model preparation

We tested supervised ML algorithms to identify early non-response and predict suboptimal weight loss at month 6 in the SMART study at baseline and by week 2. To identify the optimal model, we ran discriminative (e.g., XGBoost, random forest, support-vector machines, logistic regression, and k-nearest neighbors) and generative (e.g., naive Bayes) supervised ML classifiers. Figure 4 displays the analytic pipeline, with each component described in detail in the following sections.

Model generalizability and explainability

To test the in-distribution generalizability of the machine-learned models, we created a train/validation/test split (60%/20%/20%), ensuring that the data was randomly partitioned without cross-contamination between sets. In the model training and refinement phase, we used Bayesian optimization[39] to select the optimal set of parameters for each algorithm on the validation set in an efficient and informed manner. To address sample diversity and class imbalance, we used sample stratification (based on gender, race/ethnicity, and outcome variable) to eliminate bias and model overfitting. Finally, we evaluated in-distribution generalizability by assessing the performance of the fine-tuned model (trained on the combined train and validation sets) on the holdout test set. Based on these evaluations, we selected the best-performing machine-learned model and proceeded to the model deployment phase. To assess the generalizability of this final model (the SMART-trained model), we evaluated its performance on external datasets, specifically ENGAGED and Opt-IN. Additionally, we independently trained models on each of these external datasets using k-fold cross-validation. We then compared the cross-validation metrics from these independently trained models to the generalizability metrics obtained from applying the SMART-trained model to the same datasets. We used SHAP to improve the explainability of our model by uncovering the magnitude, directionality (positive or negative), and predictive order of the features.

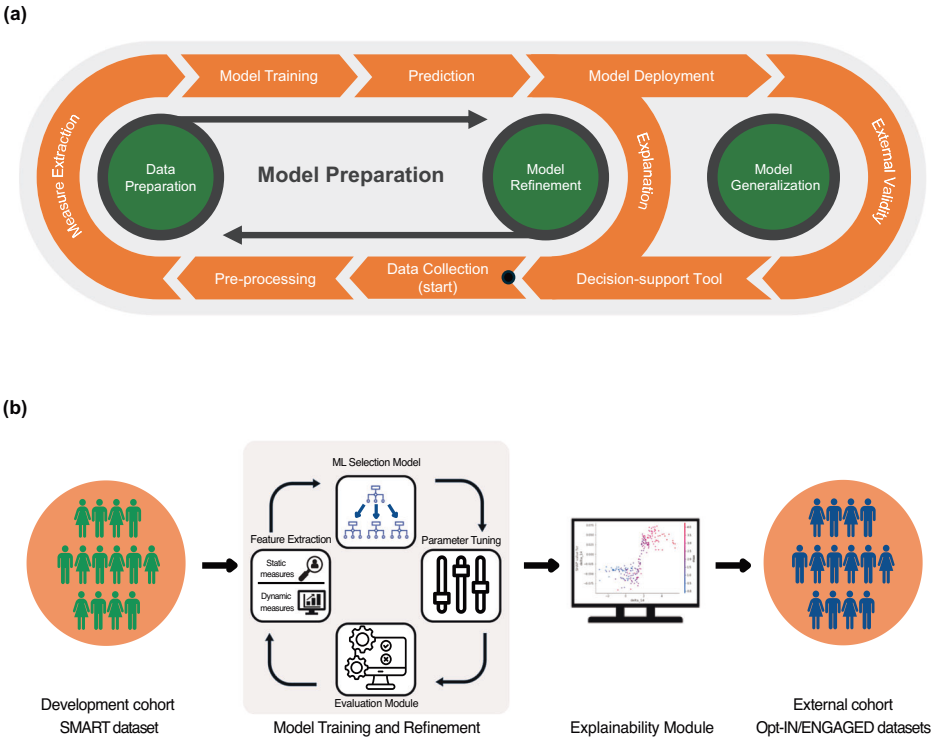


Fig. 4 | Developing a generalizable and explainable weight loss prediction machine learning model for healthcare providers. a The SMART weight loss trial data was cleaned before extracting measures, which were used to train an ML model for accurate and explainable predictions. The model was continually refined, and its performance was assessed on external datasets, such as Opt-IN and ENGAGED to evaluate generalizability. The goal is to develop a reliable decision-support tool for healthcare providers through ongoing model refinement and data preparation. **b** Raw weight data were obtained from the SMART study, then cleaned and pre-processed. Static measures, including demographics and baseline clinical characteristics, as well as dynamic measures, consisting of statistical features, were extracted for use in training the models. Following this, the machine-learned model was trained on the extracted features and tuned its hyperparameters. The prediction model was then evaluated using well-validated evaluation metrics. This iterative cycle ensures continuous improvement of the model by re-evaluating and tuning until optimal performance is achieved. Using an explainability tool, e.g., SHAP, the refined model is analyzed to identify the features that drive its predictions, enhancing the transparency and interpretability of the model's decisions. The final model was deployed on other datasets, Opt-IN and ENGAGED, to evaluate the model's generalizability.

Evaluation

We evaluated our model using various metrics, including sensitivity, specificity, F1 scores for weight loss success, F1 scores for suboptimal weight loss, weighted F1 scores, the area under the receiver operating curve (AUROC), the area under the precision-recall curve (AUPRC), and Brier score loss. F1 score is a precise measure of performance used in ML to capture the precision of the algorithm and recall of both weight loss success and suboptimal weight loss or as a combined weighted F1 score that assigns weights based on each class's support.

Data availability

The data underpinning the results of this article will be accessible for academic use through a reasonable written request to the corresponding author. Requests will be considered on a case-by-case basis and evaluated in compliance with ethical and regulatory guidelines governing clinical research.

Code availability

The specifics of the implementation of the machine learning models are available on the GitHub page: https://github.com/HAbitsLab/WeightlossPredictionModel.

Received: 14 May 2024; Accepted: 12 October 2024;
Published online: 29 November 2024

References

1. Organization, W. H. *Obesity and Overweight* https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight (2021).
2. Curry, S. J. et al. Behavioral weight loss interventions to prevent obesity-related morbidity and mortality in adults: US Preventive Services Task Force recommendation statement. *JAMA* **320**, 1163–1171 (2018).
3. Spring, B. Sound health care economics: provide the treatment needed (not less, not more). *Health Psychol.* **38**, 701–704 (2019).
4. Jakicic, J. M. et al. Effect of a stepped-care intervention approach on weight loss in adults: a randomized clinical trial. *JAMA* **307**, 2617–2626 (2012).
5. Lundberg, S. & Lee, S. *A Unified Approach to Interpreting Model Predictions* (Curran Associates, Inc, 2017).
6. Fujihara, K. et al. Machine learning approach to predict body weight in adults. *Front. Public Health* **11**, 1090146 (2023).
7. Babajide, O. et al. A machine learning approach to short-term body weight prediction in a dietary intervention program. *Comput. Sci. ICCS* **12140**, 441–455 (2020).
8. Kolyshkina, I. & Simoff, S. Interpretability of machine learning solutions in public healthcare: the CRISP-ML approach. *Front. Big Data* **4**, 660206 (2021).
9. Moreno-Sánchez, P. A. Improvement of a prediction model for heart failure survival through explainable artificial intelligence. *Front. Cardiovasc. Med.* **10**, 1219586 (2023).
10. Elshawi, R., Al-Mallah, M. H. & Sakr, S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med. Inf. Decis. Mak.* **19**, 146 (2019).
11. Nackers, L. M., Ross, K. M. & Perri, M. G. The association between rate of initial weight loss and long-term success in obesity treatment: does slow and steady win the race? *Int. J. Behav. Med.* **17**, 161–167 (2010).
12. Iwendi, C. et al. COVID-19 patient health prediction using boosted random forest algorithm. *Front. Public Health* **8**, 357 (2020).
13. Zhu, L., Qiu, D., Ergu, D., Ying, C. & Liu, K. A study on predicting loan default based on the random forest algorithm. *Procedia comput. Sci.* **162**, 503–513 (2019).
14. Elliott, M., Gillison, F. & Barnett, J. Exploring the influences on men's engagement with weight loss services: a qualitative study. *BMC Public Health* **20**, 249 (2020).
15. Svetkey, L. P. et al. Greater weight loss with increasing age in the weight loss maintenance trial. *Obesity* **22**, 39–44 (2014).
16. Lissner, L. et al. Variability of body weight and health outcomes in the Framingham population. *N. Engl. J. Med.* **324**, 1839–1844 (1991).

17. Park, M. J. et al. High body weight variability is associated with increased risk of depression: a nationwide cohort study in South Korea. *Psychol. Med.* **53**, 3719–3727 (2023).
18. Jung, I. et al. Increased risk of nonalcoholic fatty liver disease in individuals with high weight variability. *Endocrinol. Metab.* **36**, 845–854 (2021).
19. Kaze, A. D. et al. Body weight variability and risk of cardiovascular outcomes and death in the context of weight loss intervention among patients with type 2 diabetes. *JAMA Netw. Open* **5**, e220055 (2022).
20. Feig, E. H. & Lowe, M. R. Variability in weight change early in behavioral weight loss treatment: theoretical and clinical implications. *Obesity* **25**, 1509–1515 (2017).
21. Orsama, A. L. et al. Weight rhythms: weight increases during weekends and decreases during weekdays. *Obes. Facts* **7**, 36–47 (2014).
22. Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* **74**, 829–836 (1979).
23. Turicchi, J. et al. Data imputation and body weight variability calculation using linear and nonlinear methods in data collected from digital smart scales: simulation and validation study. *JMIR Mhealth Uhealth* **8**, e17977 (2020).
24. Binsaeed, B. et al. Barriers and motivators to weight loss in people with obesity. *Cureus* **15**, e49040 (2023).
25. LaRose, J. G., Lanoye, A., Ferrell, D., Lu, J. & Mosavel, M. Translating evidence-based behavioral weight loss into a multi-level, community intervention within a community-based participatory research framework: the Wellness Engagement (WE) Project. *Transl. Behav. Med.* **11**, 1235–1243 (2021).
26. Gorin, A. A., Powers, T. A., Koestner, R., Wing, R. R. & Raynor, H. A. Autonomy support, self-regulation, and weight loss. *Health Psychol.* **33**, 332–339 (2014).
27. Petch, J., Di, S. & Nelson, W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. *Can. J. Cardiol.* **38**, 204–213 (2022).
28. Pfammatter, A. F. et al. SMART: Study protocol for a sequential multiple assignment randomized controlled trial to optimize weight loss management. *Contemp. Clin. Trials* **82**, 36–45 (2019).
29. Pellegrini, C. A., Hoffman, S. A., Collins, L. M. & Spring, B. Optimization of remotely delivered intensive lifestyle treatment for obesity using the Multiphase Optimization Strategy: Opt-IN study protocol. *Contemp. Clin. Trials* **38**, 251–259 (2014).
30. Spring, B. et al. Effects of an abbreviated obesity intervention supported by mobile technology: the ENGAGED randomized clinical trial. *Obesity.* **25**, 1191–1198 (2017).
31. Spring, B. et al. An adaptive behavioral intervention for weight loss management: a randomized clinical trial. *JAMA* **332**, 21–30 (2024).
32. Chen, S. et al. Identifying and categorizing spurious weight data in electronic medical records. *Am. J. Clin. Nutr.* **107**, 420–426 (2018).
33. Kim, H. Y. Statistical notes for clinical researchers: simple linear regression 3 - residual analysis. *Restor. Dent. Endod.* **44**, e11 (2019).
34. Nam, G. E. et al. Impact of body mass index and body weight variabilities on mortality: a nationwide cohort study. *Int. J. Obes.* **43**, 412–423 (2019).
35. Bangalore, S. et al. Body-weight fluctuations and outcomes in coronary disease. *N. Engl. J. Med.* **376**, 1332–1340 (2017).
36. Turicchi, J. et al. Body weight variability is not associated with changes in risk factors for cardiometabolic disease. *Int. J. Cardiol. Hypertens.* **6**, 100045 (2020).
37. Cologne, J. et al. Association of weight fluctuation with mortality in Japanese adults. *JAMA Netw. Open* **2**, e190731 (2019).
38. Benson, L., Zhang, F., Espel-Huynh, H., Wilkinson, L. & Lowe, M. R. Weight variability during self-monitored weight loss predicts future weight loss outcome. *Int. J. Obes.* **44**, 1360–1367 (2020).
39. Snoek, J., Larochelle, H. & Adams, R. P. *Practical Bayesian Optimization of Machine Learning Algorithms* (Curran Associates, Inc, 2012).

Acknowledgements

This work has been funded by the US National Institute of Diabetes and Digestive and Kidney Diseases R01DK125414 and NIH - National Heart, Lung, and Blood Institute F31HL162555. The funders had no role in the analysis, interpretation of data, or preparation of the manuscript. The authors would like to thank Dr. Juned Siddique, Elyse Daly, Harvey Gene McFadden, Charles Olvera, Chris Romano, Rowan McCloskey, and Boyang Wei for their support during this project.

Author contributions

F.S., S.B., A.P., B.S., and N.A. participated in conceptualization. F.S., S.B., and N.A. were involved data curation, investigation, and formal analysis. B.S., D.H., N.A., and A.P. were involved in funding acquisition, supervision, and project administration. F.S., N.A., A.P., D.H., and B.S. were involved in validation. F.S. and N.A. were involved in visualization, methodology, and writing the original draft. All the authors edited and reviewed the drafted manuscript. Every author had complete access to all the data, and the decision to submit the manuscript for publication was made collectively.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01299-y.

Correspondence and requests for materials should be addressed to Farzad Shahabi.