2017

Using data mining to predict success in a weight loss trial

Marijka Batterham
University of Wollongong, marijka@uow.edu.au

Linda C. Tapsell
University of Wollongong, ltapsell@uow.edu.au

Karen E. Charlton
University of Wollongong, karenc@uow.edu.au

Jane E. O'Shea
University of Wollongong, janeo@uow.edu.au

Rebecca L. Thorne
University of Wollongong, beck@uow.edu.au

Publication Details

Batterham, M., Tapsell, L., Charlton, K., O'Shea, J. & Thorne, R. (2017). Using data mining to predict success in a weight loss trial. Journal of Human Nutrition and Dietetics, 30 (4), 471-478.

Using data mining to predict success in a weight loss trial

Abstract

Background: Traditional methods for predicting weight loss success use regression approaches, which make the assumption that the relationships between the independent and dependent (or logit of the dependent) variable are linear. The aim of the present study was to investigate the relationship between common demographic and early weight loss variables to predict weight loss success at 12 months without making this assumption.

Methods: Data mining methods (decision trees, generalised additive models and multivariate adaptive regression splines), in addition to logistic regression, were employed to predict: (i) weight loss success (defined as ≥5%) at the end of a 12-month dietary intervention using demographic variables [body mass index (BMI), sex and age]; percentage weight loss at 1 month; and (iii) the difference between actual and predicted weight loss using an energy balance model. The methods were compared by assessing model parsimony and the area under the curve (AUC).

Results: The decision tree provided the most clinically useful model and had a good accuracy (AUC 0.720 95% confidence interval = 0.600-0.840). Percentage weight loss at 1 month (≥0.75%) was the strongest predictor for successful weight loss. Within those individuals losing ≥0.75%, individuals with a BMI (≥27 kg m-2) were more likely to be successful than those with a BMI between 25 and 27 kg m-2.

Conclusions: Data mining methods can provide a more accurate way of assessing relationships when conventional assumptions are not met. In the present study, a decision tree provided the most parsimonious model. Given that early weight loss cannot be predicted before randomisation, incorporating this information into a post randomisation trial design may give better weight loss results.

Disciplines

Medicine and Health Sciences | Social and Behavioral Sciences

Publication Details

Batterham, M., Tapsell, L., Charlton, K., O'Shea, J. & Thorne, R. (2017). Using data mining to predict success in a weight loss trial. Journal of Human Nutrition and Dietetics, 30 (4), 471-478.

Using data mining to predict success in a weight loss trial

Marijka Batterham, MMedStat PhD AdvAPD AStat

Statistical Consulting Centre, National Institute for Applied Statistical Research Australia

University of Wollongong, Northfields Ave

Wollongong, NSW, Australia 2522. Phone 61 2 4221 8190, email: marijka@uow.edu.au

Linda Tapsell, AM FDAA PhD

Professor and Discipline Leader

Nutrition and Dietetics/School of Medicine/ Faculty of Science Medicine and Health

University of Wollongong NSW 2522 Australia

Karen Charlton, PhD, APD, RPHNutr

Associate Professor

School of Medicine| Faculty of Science, Medicine and Health

University of Wollongong NSW 2522 Australia

Jane O'Shea, MND APD

School of Medicine| Faculty of Science, Medicine and Health

University of Wollongong NSW 2522 Australia

Rebecca Thorne BND APD

School of Medicine| Faculty of Science, Medicine and Health

University of Wollongong NSW 2522 Australia

Abstract

Background

Traditional methods for predicting weight loss success use regression approaches which make the assumption that the relationships between the independent and dependent (or logit of the dependent) variable is linear. The aim of this research was to investigate the relationship between common demographic and early weight loss variables to predict weight loss success at 12 months without making this assumption.

Methods

Data mining methods (Decision Trees, generalised additive models and multivariate adaptive regression splines) in addition to logistic regression were used to predict weight loss success (defined as ≥5%) at the end of a 12 month dietary intervention using demographic variables (BMI, sex and age), percent weight loss at one month and the difference between actual and predicted weight loss using an energy balance model. Methods were compared by assessing model parsimony and the area under the curve (AUC).

Results

The Decision Tree provided the most clinically useful model and had a good accuracy (AUC 0.720 95%CI 0.600,0.840). Percent weight loss at one month (≥0.75 of a percent) was the strongest predictor for successful weight loss. Within those losing ≥0.75 of a percent those with BMI (≥27kg/m^2) were more likely to be successful than those with a BMI between 25-27 kg/m^2.

Conclusion

Data mining methods can provide a more accurate way to assess relationships when conventional assumptions are not met. Here a Decision tree provided the most parsimonious model. Given that early weight loss cannot be predicted before randomisation incorporating this information into post randomization trial design may give better weight loss results.

Greater early weight loss has consistently been shown to predict long term success in weight loss trials(1; 2; 3; 4). From a clinical practice perspective there is a lack of an easy to use guideline that would help identify early success and the measurement values that could be used to trigger a decision making process to intervene and change therapy for those unlikely to lose a clinically beneficial amount of weight. Recently a dynamic energy balance model has been developed and used as part of an algorithm (5) to predict whether a subject will lose weight in a weight loss trial (6). A limitation of this method, acknowledged by the authors and the accompanying editorial (7) is the complexity of applying such a model quickly in clinical practice as it relies on the input of several variables into a web or locally based algorithm. However, in this initial paper the algorithm was shown to outperform the simple process of examining early weight loss as a predictor of success. Early weight loss was not shown to be significant in predicting weight loss success, a finding in conflict with previous research (1; 2; 3; 8).

Most previous research investigating predictors of longer term weight loss have used linear or logistic regression (1; 2; 3; 4; 8). Linear and logistic regression make the assumption that the relationship between the dependent (or logit of the dependent variable for logistic regression) and the independent variable is linear and this may not always be the case. Data mining methods are increasing in use as an alternative to traditional methods when assumptions are not met[9]. In addition, although coefficients in linear regression and the odds in a logistic regression are relatively easy to understand and communicate they are not easy to apply in clinical decision making. Some data mining methods can be used to establish cut-off criteria using different variables to predict an outcome or establish the importance of difference variables in predicting an outcome. For example Batterham et al [10]showed, using a decision tree, that those losing < than 2% of their initial weight in a weight loss trial were significantly more likely to drop out than those losing greater than this amount. The cut-off of 2% was not selected a priori but was determined by the data mining procedure and provides a guide for researchers and clinicians to target participants likely to discontinue a program. If at one month the participant has lost less than 2% additional interventions or follow ups may be initiated to prevent attrition. The relationship here is not linear in that the response depends on whether the weight loss is less than 2%. Santos et al[11] used recursive partitioning, where the data is repeatedly split into partitions containing similar observations[12] to predict long-term weight loss maintenance, however only behavioural and psychological factors (self determination, exercise motivation, difference between perceived self and ideal body image, self esteem, social support for exercise, depression, quality of life and dietary restraint) measured using an extensive battery of questionnaires, which are not collected in our research practice, were considered. They identified that better body image and exercise motivation were associated with weight loss maintenance. More readily available demographic and clinical measures such as early weight loss was not used as a predictive variables in Santos and colleagues research.

Data mining is broadly defined as “the study of collecting, cleaning, processing, analyzing and gaining useful insights from data”[13] or “the process of discovering insightful, interesting, and novel patterns as well as descriptive, understandable and predictive models”[14]. There are some examples of the use of data mining in nutrition related research for example decision trees have been used to examine the relationship between diet and lifestyle factors associated with oesphageal and gastric cancer[15] and dietary patterns and their association with childhood obesity[16]. However these methods are not widely used in the nutrition domain and offer an opportunity to gain additional insights from data compared with more traditional methods.

The aim of this research is to closely examine the relationship between early weight loss and weight loss success(defined as greater than or equal to 5% weight loss [17] to determine a clinical cutoff to use in practice). Initial weight loss was considered as it has previously been shown to be a predictor of weight loss success using linear[2] and logistic regression. Other variables, such as demographic factors[1; 2], weight loss history[1; 2], psychological factors (for example depression, stress, anxiety[1; 11], quality of life[2; 11]), eating disorders[1; 2], physiological measures(blood pressure, glucose and lipids)[4] and attendance[1] have been investigated for their role in predicting weight loss success however initial weight loss is the only variable consistently shown to predict successful weight loss in several studies[1; 2; 3; 4]. We investigated whether more sophisticated decision making processes using data mining methods will have better accuracy than traditional approaches. This analysis will show whether data mining procedures, which do not make the same assumptions of the traditional methods, can be used to develop an easy to use decision process for predicting weight loss success. We further propose that this will outperform a more complicated algorithm[6] previously published in this area.

Methods

Data for this analysis was made available from a previously published weight loss trial investigating the effectiveness of high vegetable consumption in the context of an energy reduction diet for weight loss where the treatment effect (the difference in weight loss between the prescribed vegetable consumption and control group) was not significant [18]. For the analysis reported here, the demographic variables BMI, sex and age were considered in addition to percent weight loss at one month and the difference between actual and predicted weight loss using the energy balance model developed by Thomas et al[5; 6]. This algorithm for predicted weight loss includes the weight, height, age, sex and target caloric intake of the subject. The variables included were ones considered to be easily collected in research or clinical practice. . Data for the 93 participants who completed the trial were considered for this analysis. For the GAM and MARS models only data on the 76 subjects with complete data for all the considered variables were analysed. Summary statistics of the study sample and weight loss variables are shown in Table 1. Predicting weight loss success is a secondary analysis not considered in the initial study publication. The outcome or response variable in this analysis was a binary variable determining whether or not each participant had been successful in losing weight (defined as greater than or equal to 5% weight loss [17])Models were constructed using several data mining methods: Decision trees or classification and regression trees, generalized additive models (GAM), and multivariate adaptive regression splines (MARS). The reason for using these methods was that we made no a priori assumptions that the relationships were linear. The models chosen are popular data mining methods and all covered in detail in the core texts on these techniques [9; 19; 20]. Decision trees are based on linear regression and partition significant independent variables in a binary (two-way) split based on a function minimizing the sum of squared errors[21]. MARS can be regarded as a modification of the decision (or classification and regression) tree method. MARS uses piecewise functions (functions with a kink in them to model the non linearity[22]) instead of step functions which are used in decision tree models to perform an adaptive regression [20] MARS is a non-parametric regression method. GAM is an extension of regression where the linear function (the beta coefficients) are replaced by a more general non parametric functions [20]. The non-linear relationships between the response and significant independent variables is usually visualized by using a scatterplot of the partial residuals (where the effect of all the other independent variables is removed) and the independent variable which is smoothed (the random noise has been reduced)[23; 24]. The more traditional methods of logistic regression and a Receiver Operating Characteristic (ROC) or area under the curve (AUC) analysis were also used. These models werebased on predicted weight loss at one month or the probability of weight loss success at one month. Methods were compared using the AUC, which ranges from 0-1 with values closer to 1 indicating a better model and when the lower CI is > 0.5 the model is statistically significant. Model parsimony and use in clinical practice was also considered in deciding on the best model. For this purpose we wished to have the least number of significant predictors which would be easy to calculate in a research or clinical context. Prediction model validation [25] was established for the decision tree by using k-fold cross validation and the complexity parameter to fine-tune the tree based on the cross-validated error to achieve a model which is a balance between complexity and interpretability. Further verification of the variable importance was confirmed by generating 1000 trees using a random forest procedure [19]. In the random forest procedure bootstrap resampling is used to generate independent trees which are combined to determine the variable importance. Data were analysed in R Studio (Version 0.99.489 – © 2009-2015 RStudio, Inc. incorporating R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree" The R Foundation for Statistical Computing) [26]. The main packages used were 'rpart' and 'rattle' for the decision tree, 'gam' for the generalised additive model, 'earth' for the binary multivariate additive regression splines, and 'stats' ('glm' for the logistic regression).

Results

The difference between the predicted and actual weight loss at 1 month was statistically significant and the difference between the predicted probability, (calculated using the algorithm of Thomas and colleagues[6]) of meeting the 5% criteria and the percentage actually meeting the 5% weight loss criteria showed poor agreement using the kappa statistic (kappa=-0.024, P=0.807), the predicted probability correctly classified 49 of 63 who met the 5% criteria and only 6 of 30 who did not meet the criteria. The results of the different models are shown in Table 1. The Decision tree (Figure 1) shows that percent weight loss is the main predictor of weight loss success. The cutoffs determined by the partitioning algorithm suggested those losing ≥0.75 of a percent in the first month being the most likely to succeed. Within those who have lost more than this amount those with a BMI ≥ 27kg/m^2 are more likely to succeed than those with a BMI< 27kg/m^2 . The AUC is above 0.7 which is considered good [27]. The random forest procedure confirmed that the percent weight loss at one month and BMI were the most important predictors. The mean decrease in accuracy related to percent weight loss at one month was 21, and BMI 11, compared with 6, 4, and 3 for the difference between actual and predicted weight, gender and age respectively, this value reflects the decrease in classification accuracy if the variable is removed with higher values reflecting more impact. Using percent weight loss at 1 month alone also gives a good AUC, however this model only considers one variable and does not consider the relationships with the other predictors. The logistic regression(GLM) has only a moderate AUC identifying only weight loss at 1 month as a predictor. The GAM and MARS models both show non linear relationships. The GAM model shows the relationships with percent 1 month weight loss and the actual versus predicted weight are non linear. Figure 2 shows the non linear spline fit for Percent weight loss at 1 month. The GAM gives the best accuracy for prediction with the highest AUC and the MARS model also has a good AUC. The MARS model selects predicted weight loss at 1 month and BMI as the only predictors with the former being of more importance than the later. Both the GAM and MARS models can be influenced by collinearity [28]. The correlation between percent weight loss at 1 month and the difference between the actual and predicted weight loss at 1 month was 0.810 P<0.001. Generally, correlations >0.9 can be problematic although it is recommended that correlations >0.8 should be investigated [29; 30]. In this analysis the GAM was affected by this relationship giving inconsistent estimates when both variables were included. When including each separately the percent weight loss at 1 month variable was a stronger predictor and so the model containing this predictor was considered (the coefficient and P value for the model including difference between actual and predicted weight loss is also included in Table 1). When percent weight loss was removed from the MARS model only BMI was included indicating that collinearity was not affecting this model. The MARS model was unaffected by this relationship giving identical results with and without the difference between actual and predicted variable.

Discussion

Using data mining methods this research demonstrates that the relationship of common demographic variables and weight loss success is non linear and developing models to predict weight loss success should account for this. A newly developed dynamic energy balance model shown to have good accuracy in a lifestyle based intervention was not able to improve on simple measures for prediction in the present sample of participants in a dietary weight loss trial. A simple decision tree approach incorporating the percent weight loss and baseline BMI provided good predictive accuracy in this sample.

Current research investigating predictors of weight loss success (1; 2; 3; 4; 8) relies heavily on the use of linear or logistic regression which assume relationships with continuous predictor variables are linear. If the relationship between the log odds and a continuous predictor variable in a logistic regression or the independent and dependent variable in linear regression is non linear the strength of relationships may be underestimated. Sometimes a polynomial or other power term can be fitted however when the relationship does not fit one of these defined terms. GAM can improve the model fit by using splines (or other methods) to more accurately fit the relationship between the independent and dependent variables. Although the logistic regression in this analysis still showed that percent weight loss at 1 month was a significant predictor and this is a reasonable representation of the data, the logistic regression does not clearly define the relationship between this variable and BMI the way the Decision tree does. The use of data mining methods in this analysis clearly shows these relationships are non linear and incorporating this non linearity improves the models. Despite their ability to model non linear relationships these models have other considerations and the GAM model in particular was influenced by the correlation of the percentage weight loss at 1 month and the difference between the actual and predicted weight loss. The Decision trees are robust to collinearity, and this again, with the greater ease of interpretation, suggests they are the preferred model in this analysis.

MARS and GAM were included in this analysis as it is increasingly recognised that some relationships in health research are non-linear and methods which accommodate this non linearity are growing in use in lifestyle and health research (31; 32; 33; 34; 35; 36). These models are also included as they are related to the tree models in the seminal data mining text (20). While they can be used to predict weight loss success, the complexity of the algorithms make the Decision tree provide a more accurate approach for model parsimony.

Although BMI is easily assessed prior to commencing a weight loss intervention, early weight loss, the strongest predictor of success, percent weight loss at one month, can only be established AFTER the trial is commenced. Thus, a rethink of trial design is necessary to incorporate this knowledge even though it cannot be included in the initial randomisation.

New trial designs such as adaptive randomisation or sequential multiple assignment can be used to maintain all subjects in the study and at the same time target effects [37; 38]. From a statistical point of view, this design element could lead to reduced variability in study arm outcomes thus improving the effect size and subsequent power, and reducing the required initial sample size. Most importantly, however, targeting non responders and implementing a treatment change may result in better individual outcomes for the participants and more successful therapy for roll out to the general community to treat the obesity epidemic.

Dynamic energy balance models rely on physical laws to predict the amount of weight loss which should occur given the subject characteristics and dietary prescription. Two publically available energy balance calculators are widely reported in the literature http://www.pbrc.edu/research-and-faculty/calculators/weight-loss-predictor/ [5] and https://supertracker.usda.gov/bwp/index.html [39]. The former was chosen for this anlaysis as this model has been further developed to predict weight loss success. Both models suffer from the limitation that access to a computer or smart device is required [7] as is input on weight goals and demographic information. The' Supertracker' can also incorporate information on activity levels, macronutrient intake and input measures of body fat and resting metabolic rate. Both are useful in the research setting and potentially for goal setting in clinical practice. In this case the use of the predicted probability of success at one month gave a prediction accuracy that was not better than chance alone using the AUC. Incorporating the difference between the actual and predicted weight loss [5] (which can be used as a marker of compliance to the dietary prescription) did not come out as a primary predictor in most of the models however by using a GAM model it is clear this relationship was not linear and it could be investigated further in other samples.

There are some limitations to this analysis. Many variables have been associated with predicting weight loss[40] and the current analysis was limited to the variables collected for the study considered. Dietary prescription was defined in 500kJ increments in order to make the recommendations on vegetable intake easier to implement. The rate of attrition was moderate (22.5%) in this study compared to others conducted by our research group and other facilities [41; 42]. This may reflect differences in this study population and the results require replication in other populations. Nevertheless, given the ease of calculation of a decision tree in both the free package R Studio (R Studio® Inc, Boston MA) and commercial packages such as SPSS (IBM Corporation, Armonk NY) it is possible that researchers and in fact clinicians with their own unique populations should be investigating the use of these tools. Even without the use of an algorithm the real clinical message here is that in clients seeking weight reduction low early weight loss and lower baseline BMI (25-27) should be targeted for more intense approaches or combinatorial approaches (exercise, psychological counselling or potentially pharmacotherapy).

In summary this analysis demonstrates the potential utility of data mining methods over more traditional analyses to produce better models for prediction of weight loss. It also demonstrates that conventional assumptions such as the linearity of relationships may not be valid. For example, the results of the decision tree show that weight loss success is determined by a cutpoint in weight loss at 1 month of 0.75% and a BMI of 27. The scatterplot for the GAM also shows the non-linear relationship between weight loss success and percent weight loss at one month. This information is lost if the relationships are considered to be linear. Some of the limitations of these methods, in this case with respect to collinearity, are also highlighted. When modeling data there is often a trade off between accuracy and model parsimony. In this analysis the simpler decision tree approach although slightly less accurate than the GAM and MARS is easy to interpret and not susceptible to the collinearity issue observed in the GAM. The models all suggest that percent weight loss at 1 month is the strongest predictor of weight loss success at the end of the one year study and that within those with greater initial loss, baseline BMI is also important. Alternative trial designs and clinical strategies are recommended where this information is incorporated to improve weight loss outcomes.

Conflict of interest

The authors have no conflict of interest to declare

Funding Acknowledgement

The initial published trial (not the current analysis) was funding by Horticulture Australia Limited using the vegetable levy and matched funding from the Australian Government. The study and secondary analysis was approved by the University of Wollongong Human Research Ethics Committee and registered with ANZCTR (www.anzctr.org.au) (ACTRN12610000784011).

1. Fabricatore AN, TA W, Moore RH *et al.* (2009) Predictors of attrition and weight loss success: Results from a randomized controlled trial. *Behav Res Ther* **47**, 685-691.

2. Elfhag K, Rossner S (2010) Initial weight loss is the best predictor for success in obesity treatment and sociodemographic liabilities increase risk for drop-out. *Patient education and counseling* **79**, 361-366.
3. Ortner Hadziabdic M, Mucalo I, Hrabac P *et al.* (2015) Factors predictive of drop-out and weight loss success in weight management of obese patients. *Journal of human nutrition and dietetics : the official journal of the British Dietetic Association* **28 Suppl 2**, 24-32.
4. Packianathan I, Sheikh M, Boniface D *et al.* (2005) Predictors of programme adherence and weight loss in women in an obesity programme using meal replacements. *Diabetes, obesity & metabolism* **7**, 439-447.
5. Thomas DM, Martin CK, Heymsfield S *et al.* (2011) A Simple Model Predicting Individual Weight Change in Humans. *Journal of biological dynamics* **5**, 579-599.
6. Thomas DM, Ivanescu AE, Martin CK *et al.* (2015) Predicting successful long-term weight loss from short-term weight-loss outcomes: new insights from a dynamic energy balance model (The POUNDS Lost study). *The American Journal of Clinical Nutrition*.
7. Finer N (2015) Predicting therapeutic weight loss. *Am J Clin Nutr* **101**, 419-420.
8. Unick JL, Hogan PE, Neiberg RH *et al.* (2014) Evaluation of early weight loss thresholds for identifying nonresponders to an intensive lifestyle intervention. *Obesity (Silver Spring, Md)* **22**, 1608-1616.
9. James G, Witten D, Hastie T *et al.* (2013) *An Introduction to Statistical Learning with applications in R*.
10. Batterham M, Tapsell LC, Charlton KE (2015) Predicting dropout in dietary weight loss trials using demographic and early weight change characteristics: Implications for trial design. *Obesity research & clinical practice*.
11. Santos I, Mata J, Silva MN *et al.* (2015) Predicting long-term weight loss maintenance in previously overweight women: a signal detection approach. *Obesity (Silver Spring, Md)* **23**, 957-964.
12. Strobl C, Malley J, Tutz G (2009) An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychological methods* **14**, 323-348.
13. Aggarwal C (2015) *Data Mining: The textbook.* Switzerland: Springer.
14. Zaki MJ, Meira W (2014) *Data mining and analysis: Fundamental concepts and algorithms.* New York, NY: Cambridge University Press.
15. Navarro Silvera SA, Mayne ST, Gammon MD *et al.* (2014) Diet and lifestyle factors and risk of subtypes of esophageal and gastric cancers: classification tree analysis. *Annals of epidemiology* **24**, 50-57.
16. Lazarou C, Karaolis M, Matalas AL *et al.* (2012) Dietary patterns analysis using data mining method. An application to data from the CYKIDS study. *Computer methods and programs in biomedicine* **108**, 706-714.
17. Williamson DA, Bray GA, Ryan DH (2015) Is 5% weight loss a satisfactory criterion to define clinically significant weight loss? *Obesity (Silver Spring, Md)* **23**, 2319-2320.
18. Tapsell LC, Batterham MJ, Thorne RL *et al.* (2014) Weight loss effects from vegetable intake: a 12-month randomised controlled trial. *Eur J Clin Nutr* **68**, 778-785.
19. Williams GJ (2011) *Data Mining with Rattle and R: The art of excavating data for knowledge discovery*: Springer.
20. Hastie T, Tishirani R, Friedman J (2009) *The elements of statistical learning: Data mining, inference, and prediction.* New York, NY: Springer.

21. Brown MS (2014) *Data Mining for Dummies.* Hoboken, NJ: John Wiley & Sons, Inc.
22. Brownlowe J (2014) Jump-Start Machine Learning in R: MachineLearningMastery.com.
23. Hill T, Lewicki P (2007) *Statistics: Methods and Applications. .* Tulsa, OK: Dell Inc.
24. Hastie T, Tibshirani R (1990) *Generalized Additive Models.* Baton Roca, FL: CRC Press.
25. Ivanescu AE, Li P, George B *et al.* (2015) The importance of prediction model validation and assessment in obesity and nutrition research. *International journal of obesity (2005)*, .
26. R Core Team (2013) R: A Language and Environment for Statistical Computing. Vienna, Austria.
27. Pines JM, Carpenter CR, Raja AS *et al.* (2013) *Evidence-based emergency care: Diagnostic testing and clinical decision rules.* West Sussex, UK: Wiley-Blackwell.
28. Morlini I (2002) *Facing multicollinearity in data mining.* Milano-Bicocca, June 9-11: XLI meeting of the Italian Statistical Association.
29. Tuffery S (2011) *Data mining and statistics for decision making.* West Sussex, UK: Wiley.
30. Dormann CF, Elith J, Bacher S *et al.* (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36**, 27-46.
31. Butte NF, Wong WW, Adolph AL *et al.* (2010) Validation of cross-sectional time series and multivariate adaptive regression splines models for the prediction of energy expenditure in children and adolescents using doubly labeled water. *J Nutr* **140**, 1516-1523.
32. Zakeri IF, Adolph AL, Puyau MR *et al.* (2010) Multivariate adaptive regression splines models for the prediction of energy expenditure in children and adolescents. *J Appl Physiol* **108**, 128-136.
33. Butte NF, Wong WW, Lee JS *et al.* (2014) Prediction of energy expenditure and physical activity in preschoolers. *Med Sci Sports Exerc* **46**, 1216-1226.
34. Zakeri IF, Adolph AL, Puyau MR *et al.* (2013) Cross-sectional time series and multivariate adaptive regression splines models using accelerometry and heart rate predict energy expenditure of preschoolers. *J Nutr* **143**, 114-122.
35. Jiang M, Foster EM (2013) Duration of breastfeeding and childhood obesity: a generalized propensity score approach. *Health Serv Res* **48**, 628-651.
36. Kwate NO, Yau CY, Loh JM *et al.* (2009) Inequality in obesigenic environments: fast food density in New York City. *Health Place* **15**, 364-373.
37. Almirall D, Nahum-Shani I, Sherwood NE *et al.* (2014) Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research. *Translational behavioral medicine* **4**, 260-274.
38. Chow SC, Chang M (2008) Adaptive design methods in clinical trials - a review. *Orphanet J Rare Dis* **3**, 11.
39. Hall KD, Sacks G, Chandramohan D *et al.* (2011) Quantification of the effect of energy imbalance on bodyweight. *Lancet* **378**, 826-837.
40. Moroshko I, Brennan L, O'Brien P (2011) Predictors of dropout in weight loss interventions: a systematic review of the literature. *Obesity Reviews* **12**, 912-934.
41. Batterham MJ, Tapsell LC, Charlton KE (2013) Analyzing weight loss intervention studies with missing data: which methods should be used? *Nutrition* **29**, 1024-1029.
42. Elobeid MA, Padilla MA, McVie T *et al.* (2009) Missing data in randomized clinical trials for weight loss: scope of the problem, state of the field, and performance of statistical methods. *PLoS ONE* **4**, e6624.

Table 1 Descriptive statistics of study sample and model summaries.

Variable				**Mean(95%CI) and P value**
BMI at baseline				29.94kg/m^2 (29.38,30.50)
Weight loss at 1 year				7.46% (6.29,8.63)
Age				49years (47,51)
Percentage of women				91% (n=85)
Weight loss at 1 month				2.64% (2.23,3.06%)
Actual weight at 1 month				82.22kg (79.99,84.44)
Predicted weight at 1 month				80.03kg (77.89,82.18)
Difference between actual and predicted weight				2.18kg (1.78,2.59) $P<0.001$
Percentage meeting ≥5% weight loss criteria				68% (n=63)
Percentage predicted to meet 5% criteria at 1 month				79% (n=73), K=-0.024, P=0.807
Model				**AUC**
Decision Tree (classification and regression tree)				0.720(0.600,0.840) P=0.001
ROC (probability of success at 1 month)				0.489(0.363,0.614) P=0.863
ROC (% weight loss at 1 month) (cutpoint -1.68%)				0.740(0.635,0.845) P=0.001
GLM	Coefficient	Z value	P	0.670(0.545,0.795) P=0.008
Age	-0.042	-1.464	0.143	
Gender	-0.349	-1.160	0.246	
BMI	0.134	1.378	0.168	
Weight loss at 1 month	-0.974	-2.550	0.011	
Difference between actual and predicted weight	0.578	1.412	0.158	
GAM				0.777(0.638,0.915) $P<0.001$
	Coefficient	F	P	
Age	-0.004	0.014	0.908	
Gender	-1.003	0.823	0.368	
BMI	0.143	1.366	0.246	
Weight loss at 1 month	-0.533	6.744	0.011	
Difference between actual and predicted weight*	-0.503	3.722	0.058	
MARS				0.726(0.583,0.868) P=0.002
equation & coefficients: 1.81-(2.46*bf1)-(17.34*bf2)+(18.26*bf3) Where bf1=h(26.82-BMI), bf2=h(% wt loss at 1 month--0.78), bf3=h(% wt loss at 1 month--0.58)				

K kappa statistic, AUC area under the curve, ROC receiver operating characteristic curve, GLM generalised linear model (logistic regression), GAM generalised additive model, MARS multivariate adaptive regression splines, bf basis function, h hinge. *values from separate GAM model without percent weight loss at 1 month.

Figure 1. Decision Tree

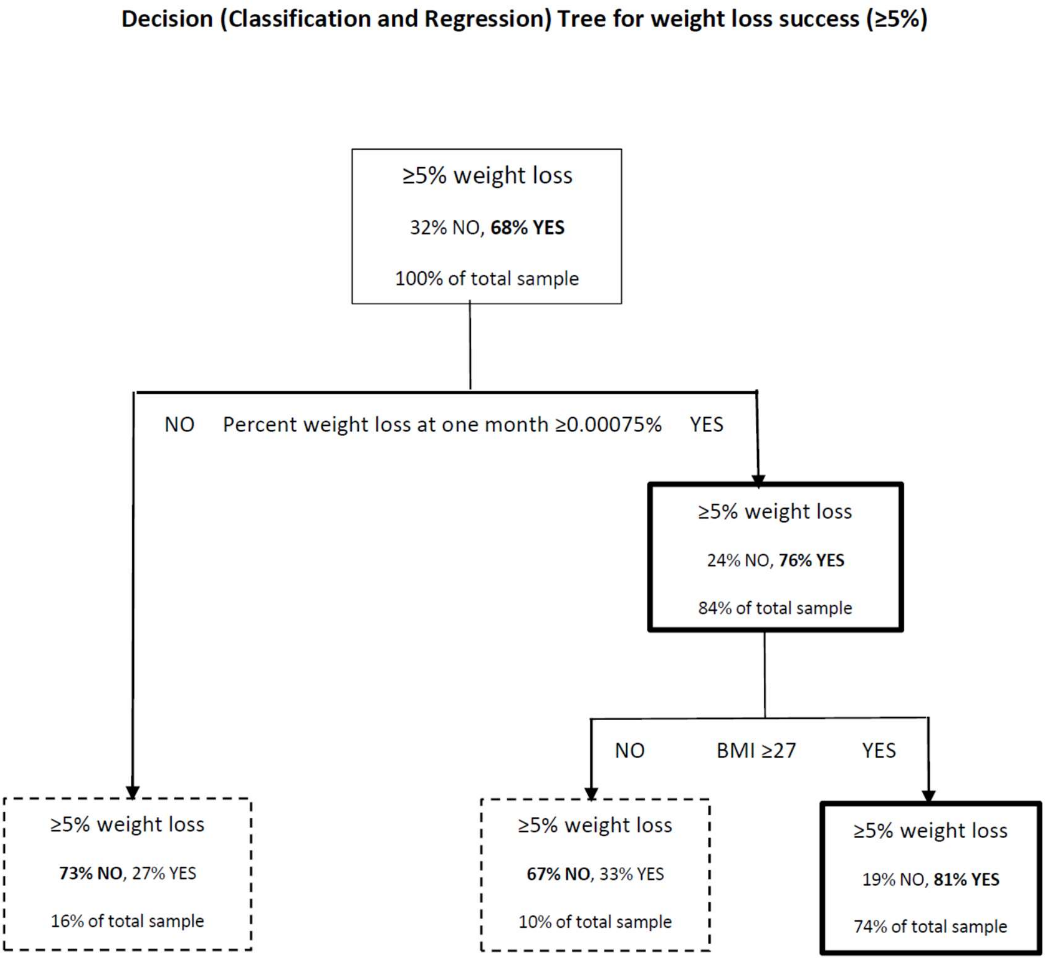
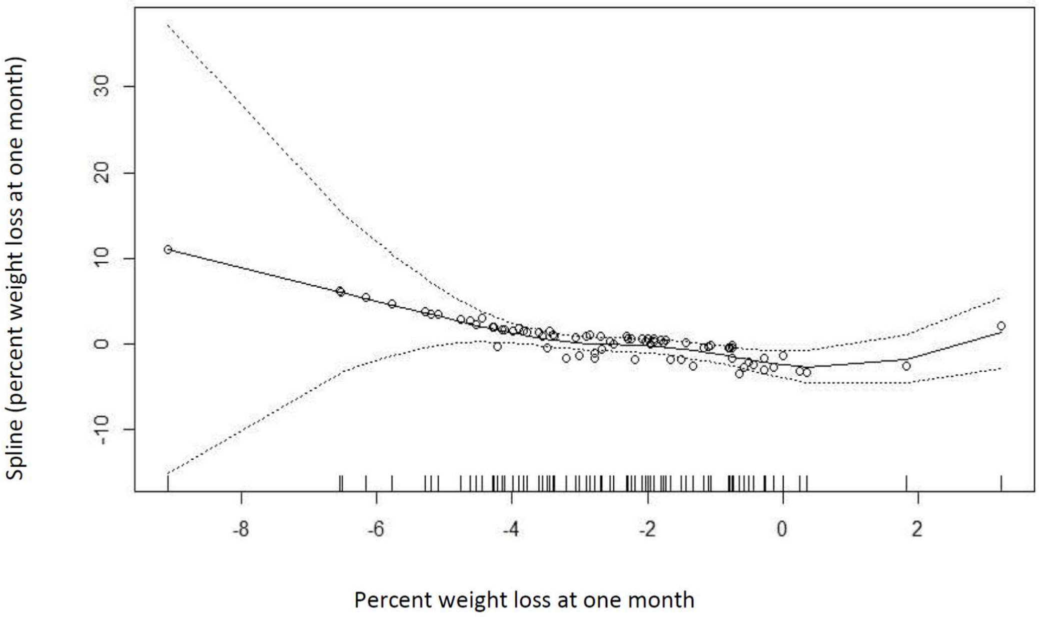


Figure 2. Non linear spline plot for percent weight loss at one month from the generalised additive model.


Spline (percent weight loss at one month)
Percent weight loss at one month