A Data-analytical System to Predict Therapy Success for Obese Children

Completed Research Paper

Nurten Öksüz
Saarland University
Campus A5 4, 2nd level
66123 Saarbrücken, Germany
nurten.oeksuez@iss.uni-saarland.de

Iaroslav Shcherbatyi
German Research Center for Artificial Intelligence
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
iaroslav.shcherbatyi@dfki.de

Tobias Kowatsch
University of St. Gallen
Dufourstrasse 40a
9000 St. Gallen, Switzerland
tobias.kowatsch@unisg.ch

Wolfgang Maass
Saarland University
Campus A5 4, 2nd level
66123 Saarbrücken, Germany
wolfgang.maass@iss.uni-saarland.de

Abstract

Childhood obesity is an increasingly pervasive problem. Traditional therapy programs are time- and cost-intensive. Furthermore, success of therapy is often not guaranteed. Typically, success of therapies is determined by comparison of body mass index (BMI) before and after a therapy. In this paper, we present a Data-analytical Information Systems (DAIS) that provides predictions of future BMI changes before conducting a therapy. The DAIS considers current parameters like age as well as heart rate during a standardized exercise. By predicting outcomes of a therapy, healthcare practitioners could personalize standard therapies and improve the outcome. We collected data from randomized clinical trial and trained Machine Learning models to estimate whether BMI will decrease after therapy with 85% accuracy. Accuracy of predictions is compared with domain experts' predictions. Further, we present empirical results of the domain experts' perception regarding the proposed DAIS. Our DAIS provides positive evidence as a tool for personalized medicine.

Keywords: Data-analytical Information System, Healthcare, Data Science, Predictive Modeling

Introduction

Professional obesity therapies rely on comparing data on fitness tests and in particular individual age- and gender-specific body-mass indices (BMI) before and after an intervention (Montesi et al. 2016; Reinehr 2003). BMI is used as a screening measure for overweight or obesity and gives an indication of whether the weight is higher than what is considered as a healthy weight for a given height (Montesi et al. 2016). Typical timeframes for obesity therapies are 3-6 months. General therapies are deployed on patient groups as an intervention and results are evaluated ex-post potentially leading to therapy adjustments. In contrast, patient-centric healthcare places individual patients at the center of therapies and analyses which treatments are optimal for each patient (Garvey et al. 2014; Garber et al. 2013; Jensen et al. 2014). This requires individual predictions on treatment success based on a broad understanding of patients in their social and economic environments. Combining existing methods with new and evolving technologies allows physicians to provide new and efficient patient-centric healthcare. Data-analytical Information Systems (DAIS) are a means for supporting physicians with predictive capabilities by leveraging data and Artificial Intelligence (AI) methods extensively (Shmueli and Koppius 2011). One class of methods in AI area is Machine Learning (ML). ML techniques are becoming more and more popular for applications with data streamed from devices and wearables such as heart rate sensors (Mannini and Sabatini 2010). The ML algorithms are able to provide the technical basis for accurate predictive modeling with continuous biological signals (Witten and Frank 2005). There are several scientific approaches considering data such as static parameters (e.g. age, gender, BMI) or lifestyle and environment factors to predicting future occurrence of childhood obesity and support physicians in decision-making (Zhang 2009; Dugan et al. 2015; Adnan et al. 2012a; Adnan et al. 2012b). However, there is no Information System (IS) or DAIS that uses medical static and dynamic physiological parameters, i.e. heart rate during standardized exercise, to predict whether the BMI will decrease after an obesity therapy, before conducting the therapy as such. With regard to our goal to provide such DAIS, the question arises whether this is feasible by only considering baseline data (data collected before intervention) including static (e.g. BMI, age) and dynamic parameters (heart rate during exercise). This motivates the following research question:

RQ1: Can data on static and dynamic healthcare parameters predict whether BMI will decrease after obesity treatment therapy?

RQ2: Do healthcare practitioners perceive these predictions as being useful?

There are several studies identifying heart rate during exercise as an indicator for characteristics about an individual (e.g. cardiovascular fitness), which in turn are indicators for therapy success (Singh et al. 2010; Bassett and Howley 2000; Warburton et al. 1999). We set up a study where we collected static parameters (e.g. BMI, age) and dynamic parameters[1], i.e. heart rate during a fittest of overweight and obese children before conducting therapy. We recorded the change in BMI after 6 months, which indicates whether the conducted therapy was successful or not. ML methods are used to build a predictive model, which estimates whether a patient's BMI will decrease after 6 months of therapy. In order to answer RQ1 we assess the statistical significance and quality of the results. To answer RQ2, we analyse the acceptance of our DAIS with the help of the Technology Acceptance Model (TAM) as well as situation-service-fit, user satisfaction and word-of-mouth research.

The paper is structured as follows. First, state-of-the-art is introduced, the conceptual model and data acquired from wearable sensors as they are processed by ML algorithms to predict whether BMI will decrease over time and the service design and implementation of the DAIS are described. Thereafter, a case study is outlined to compare ML with domain experts as well as to derive practical implications. Here, the experts are asked to predict whether BMI will decrease after therapy to compare their estimation accuracy with the one of the ML. We then report results with regard to the experts' perception of the proposed DAIS to assess the potential. We finally discuss the results and provide an outlook on future research.

State-of-the-Art

Child Obesity and Therapy Success

Child obesity is rather a complex problem that is as much a psychological as a physical one (Collins and Bentz 2009). It has been shown that several factors play a role in the obesity treatment success. The cardiorespiratory fitness of an individual is an essential factor considered to estimate the success of a therapy or decrease in BMI (Singh et al. 2010; Bassett and Howley 2000; Warburton et al. 1999). "Cardiorespiratory fitness is a health-related component of physical fitness defined as the ability of the circulatory, respiratory, and muscular systems to supply oxygen during sustained physical activity (Lee et al. 2009)". Dupuis et al. (2000) found out that among children and adolescents, cardiorespiratory fitness level was significantly negatively correlated with BMI. Consequently, obese children and adolescents are

[1] Collected data and the code used for the analysis can be found here: tinyurl.com/yarwr7hb

more likely to have lower cardiorespiratory fitness when compared to their normal-weight peers, mainly due the huge effort required to carry the larger amount of body fat (Dupuis et al. 2000). With regard to therapy success, Mota et al. (2008) have shown that children who have lower cardiorespiratory fitness and high BMI are likely to have an increase in BMI over time. Cardiorespiratory fitness is mostly expressed in maximal oxygen uptake (VO2max) measured by standardized fittest such as treadmill test. Due to the complexity of measuring the VO2max in practice without a well-equipped laboratory, more simple techniques are often used (Dimkpa 2009). Measuring heart rate during exercise, heart rate recovery (HRR) after exercise and the heart rate at rest represents one of the additional options for an assessment of cardiorespiratory fitness (Dimkpa 2009).

Heart Rate and its Relationship to Cardiorespiratory Fitness

The heart rate measuring during and after an exercise (HRR) can give information about an individual's (cardiorespiratory) fitness level (Bird et al. 1998). Physical activity or exertion elevates the heart rate for the duration of the physical activity and slows it down during the cool down period (Swain and Leutholtz 2007; Draper and Marshall 2014). The fitter an individual is, the lower the heart rate will be during training, the lower it will be during cool down and the faster it will return to the pre-exercise level (Bird et al. 1998). Repeating the test after a certain period of time can give information about the change of an individual's fitness (Bird et al. 1998). The HRR value depends on, amongst others, the cardiorespiratory fitness of an individual (Mazzeo and Marshall 1989). Furthermore, the results of a study conducted by Singh et al. (2008) have shown that children with higher BMI, especially those who are overweight, have slower 1-minute HRR after exercise and thus, a lower cardiorespiratory fitness than children with a lower BMI (Singh et al. 2008).

Existing Decision Support Systems for Treatment of Childhood Obesity

Clinical decision support tools offer potential for assisting health professionals in diagnosing overweight and obese children as well as improving obesity counselling (Ayash et al. 2013). Most clinical decision support system applications operate as components of Electronic health record (EHR) systems, which stores health information of patients. Rattay et al. (2009) demonstrated the potential effectiveness and efficiency of a medical alert system as well as decision support in health professionals' obesity management. For this, they modified their existing electronic medical record to embed clinical guidelines and best practice approaches that help providers work with the patients' families for making behaviour changes. Furthermore, EHR-based enhancements were developed to support clinical guidelines supported by the Expert Committee on Childhood Obesity (Barlow 2007) and to screen high-risk patients for liver disease and diabetes mellitus, and referrals to specialists, such as cardiologists if necessary. The Wellness Survey and the Educational hand-outs from the Hawaii Pediatric Weight Management Toolkit are instruments to assist physicians and dieticians to counsel children and their caregivers for healthier lifestyle behaviours (Brener et al. 2004; Chock and Kerr 2011). Beside EHR systems, which primarily serve as a management tool and help to improve diagnosis and prevent obesity, there are several scientific approaches applying Artificial Intelligence methods by using data mining and ML techniques to predict the future occurrence of early childhood obesity and facilitate physicians' decision-making (Zhang et al. 2009; Dugan et al. 2015; Adnan et al. 2012a; Adnan et al. 2012b). Zhang et al. (2009) analysed the Wirral child database with 16,653 instances, which was limited to static parameters (i.e. gender, height, weight, and BMI) prior to second birthday with the help of ML methods, more specifically Logistic Regression, Decision Trees, Association Rules, Neural Networks, Naïve Bayes, Bayesian Networks and Support Vector Machines. The highest accuracy was achieved with Naïve Bayes (accuracy = 91.9%). Dugan et al. used ML methods (i.e. Random Tree, Random Forest, J48, ID3, Naïve Bayes, and Bayes) to predict the future occurrence of childhood obesity by considering 176 parameters (e.g. gender, age, height, weight, former obesity occurrence) of 7519 children collected prior to the second birthday by CHICA, a clinical decision support system. The highest accuracy was achieved by ID3 with an accuracy of 85%. Adnan et al. (2012a) used Naive Bayesian Classifier to analyze several static parameters (e.g. gender, catch-up growth, adiposity rebound, and premature birth), lifestyle factors, and family/environment factors of 92 children aged 9-11 years for predicting future occurrence of childhood obesity. The used Naïve Bayesian Classifier achieved an accuracy of 71%. While most decision support system solutions focus on the management of childhood obesity and the prediction of future occurrence of childhood obesity, there is no DAIS which provide prediction of the therapy success of child obesity, meaning whether an intervention will lead to weight loss for a specific obese individual before conducting the therapy.

Conceptual Model

In order to create a model for the prediction of an individual's BMI change after 6 months of therapy, we relied on the existing theory in literature mentioned above. The initial situation of the conceptual model consists of static as well as physiological parameters (see Figure 1). The static parameters (profile data) considered in the model are height, weight, BMI, and age. BMI is an important indicator for the cardiorespiratory fitness and consequently predictive of the physical performance (e.g. number of laps run during 6-min-run test). Furthermore, age influences the heart rate values and thus, cardiorespiratory fitness and the physical performance of a subject too (Gilbert et al. 2014) and is therefore also considered for the conceptual model. Besides static parameters, we also take into account some physiological parameters namely heart rate during exercise as well as during cool down, as both give information about the cardiorespiratory fitness and thus, the performance of an individual

Cardiorespiratory fitness does not only have an impact on performance of an individual during exercise but also correlates with the BMI change over time. Thus, we assume, that the static (e.g. BMI, age) as well as dynamic parameters (heart rate during 6-min-run test and 3-min-cool down) are predictive of whether the BMI will decrease for an individual. ML methods serve as a basis for our analysis, since they are able to better identify non-trivial and complex patterns in continuous physiological signals and to detect previously unknown relationships in the data (Scime 2015).

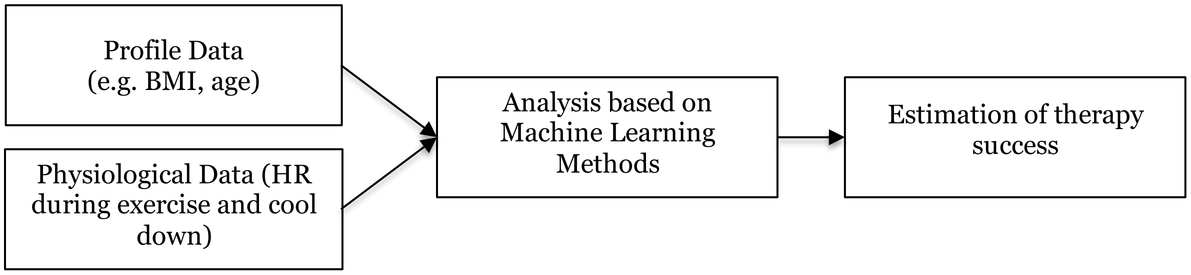

Profile Data
(e.g. BMI, age)
Physiological Data (HR during exercise and cool down)
Analysis based on Machine Learning Methods
Estimation of therapy success

Figure 1. Conceptual Model

Decision Support System Design & Implementation

In this section we describe design and implementation of our DAIS based on the conceptual model above. First, we provide a description of the DAIS. Then, we describe the necessary data collection to come up with underlying ML models. Finally, we provide details on how we obtain useful ML models based on the data collected, and conclude with a discussion of out - of - sample evaluation of the performance of these models.

System Description

The primary goal of our DAIS is to derive from dynamic physiological data (heart rate) and additional static medical parameters (e.g. BMI, age) whether the BMI of an obese child will decrease over time or not. Our DAIS is designed to collect sensory data using unobtrusive commodity hardware and additional static features from patients. The sensory data together with the static features are stored in a web - server where ML methods are used for predictions on whether BMI will decrease in the future. The analysis results by our DAIS are shown to health professionals via visualization with the help of a web app/graphical user interface (GUI). Based on the results, health professionals can make adjustments to the current therapy if necessary.

Data Collection

In order to be able to produce predictive models for our service, we conduct a data collection study at a Swiss children's hospital in St. Gallen[2]. 20 overweight and obese children aged between 11 and 16 years with high BMI values (25<BMI<37) participated in the Dordel-Koch fitness test (7 female and 13 male), a standardized test for children. Overweight is defined as a BMI at or above the 85th percentile and below the 95th percentile for children and teens of the same age and sex, while obesity is defined as a BMI at or above the 95th percentile (Reinehr 2003). The whole standardized fittest takes about 25 minutes on average and consists of multiple exercises. The final exercise of this test was a 6-min-run test, the basis for our ML analysis. The run test takes 6 minutes and requires children to run as many laps as possible (all out test). In our setting, one lap had a distance of 25 meters. Number of laps means the total count of laps a child was able to run during the 6-min-run test. Besides BMI, height, weight, age, gender and the number of laps run, the exercise heart rate measurements (about 25 minutes) as well as the heart rate measurements after the exercise (cool down period of 3 minutes) were measured. The participants were equipped with a Scosche Rhythm+ heart rate monitor and a Samsung Galaxy S6 smartphone, in which our data collection application was installed. The application collects heart rate data from the heart rate monitor when the *Exercise Button* or the *Cool down Button* is pressed. Then it sends the data to the server, where it is stored and processed. At the beginning of the exercise, the *Exercise Button* was pressed to measure the initial heart rate. After that, the children started the fittest while the heart rate was continuously measured. Right after the exercise, the *Cool down Button* was pressed to measure the heart rate of the children during the 3-min-cool down. Each child has been assigned to either obesity therapy A or obesity therapy B. Both therapies required patient's compliance and adherence (physical and dietary aspects). Subjects in one group were given challenges via smartphone, such as walking 10,000 steps on a particular day. After a therapy duration of 6 months, the fittest was repeated.

Data Preprocessing

Before we apply ML techniques, we apply a range of pre-processing steps to the data we collected. In particular, records of patient raw data are represented as vectors of fixed size (feature vector, independent variables) (Hastie et al. 2001). BMI change after therapy (dependent variable) is represented as a binary value. In order to convert raw heart rate data into feature vector, we first separate heart rate for the run test itself ("run test" heart rate), and the heart rate immediately after the run test ("cool down" heart rate). Thus we obtain two sequences of heart rate measurements. Then we subsample both sequences to m values, which are equidistantly located on the support of particular sequence (see Figure 2). Linear interpolation of original raw heart rate data is used to perform subsampling at equidistant points. Such representation of sequence of measurements conveys a general shape of heart rate curve. Concatenated vectors of heart rate values for "run test" and "cool down" during the fittest are used as feature representation of the raw heart rate data.

[2] The study has passed through the Institutional Review Board (The Cantonal Ethic Commission in Zurich). The approval is registered on clinicaltrials.gov with the ID number: NCT03270423.

You can find the registration here: https://clinicaltrials.gov/ct2/show/NCT03270423

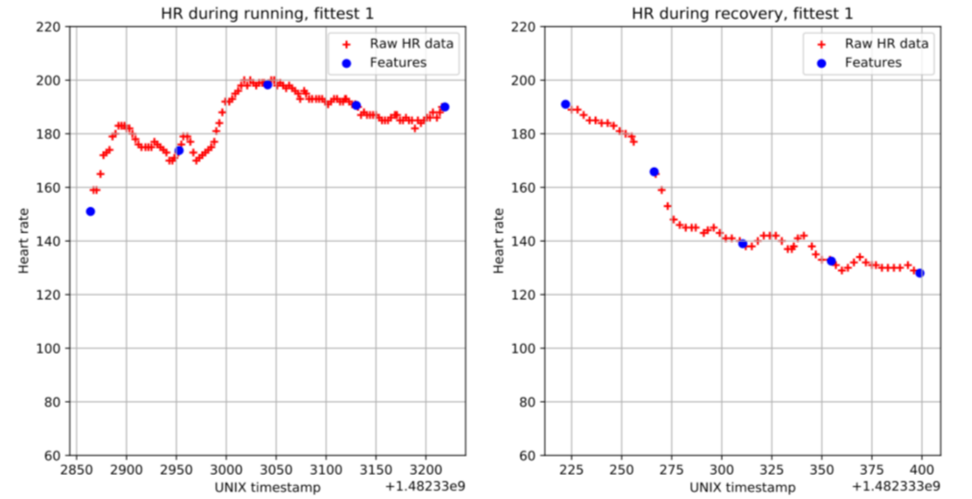


Figure 2. Feature Extraction From HR Measurements

In the picture, values of heart rate at blue points correspond to the values of the feature vector. More specifically, we obtain HR feature vectors as follows. We assume that the heart rate measurements curve is represented as a sequence of tuples of the form (t_i, h_i), where $i < n$, $n \in N$ is an index of one of n timestamps and heart rate value pairs, t denotes time and h denotes heart rate values. We assume that time of measurement always starts from zero, and the time of the last measurement is denoted as T. Let m > 1, m ∈ N be a desired number of the features in the feature vector. Let the function HR(t) denote a function that for some input time t outputs a linear interpolation of two closest data points along time axis of heart rate curve. Then we convert the curve to features using the following formula:

$$F_i=HR(T*i/(m-1)), i=0, 1, \ldots m-1,$$

where F_i is the i – th feature describing the HR curve. Such values are calculated for both the run test and cool down curves and concatenated with the rest of features (static parameters) to obtain a feature vector used for ML. Such approach allows to capture the general shape of the curve using a small set of real values. This is important in view of small size of our dataset, as increasing the number of features increases the likelihood of in – sample bias, referred to as "overfitting" in ML community (Hastie et al. 2001). As additional features for our predictive model we use subject's age, weight, height, BMI and a variable which indicates whether a patient is currently undergoing therapy A (labelled as 1) or therapy B (labelled as 0). Example of such feature vector with m=5 values that represent HR curves is depicted in the Table 1.

HR during run test					HR during recovery					Weight	Age	BMI	Height	Therapy
151	173	198	190	190	191	165	139	132	128	77.5	12	28.5	1.65	1

Table 1. Example Feature Vector Representation Of The Patient's Data.

As outputs for our model we define binary value, which takes values in the range of {-1, +1} and represents whether the BMI of an individual will decrease (value of -1) in 6 months or not (value of 1). Such value is sufficient to make decision on whether a therapy for a patient should be adjusted, if no weight decrease is forecasted. Furthermore, thresholding of BMI change reduces effect of outlier target values on training procedure of ML model (Rousseeuw and Leroy, 2005). We confirmed experimentally that we obtain better performance with classification problem than with regression problem.

We obtain target binary values from absolute change of BMI as follows. Successful decrease of BMI was defined as BMI change <= -0.4 BMI units (labelled as -1), whereas stable or increase of BMI was defined as BMI change > -0.4 BMI units (labelled as +1). The value of 0.4 was chosen as it on average means a change of more than 1 kg of body mass for a patient. By applying a threshold of 1 kg, we attempt to counteract noise in body weight measurements due to normal food intake before the fittest. According to Dixon et al. (2013), eating and drinking before testing (up to two hours) alter the body mass by approximately 1 kg (assumption: 750gr meal plus 250ml liquid). Some descriptive statistics of such binary value are given in Table 2.

+1 label, % of participants	**-1 label, % of participants**	**Median BMI increase**	**Median BMI decrease**
55%	45%	1.3	-1.63

Table 2. Statistics Of BMI Change For All Patients.

Predictive Analytics for Small Data Sets

An important consideration is that the dataset size we consider is relatively small (20 data points), which requires a careful ML approach in order to ensure that no in – sample bias or "overfitting" occurs (Hastie et al. 2001). We describe our approach that caters to this goal below. We use Python package "scikit-learn" for our implementation (Pedregosa et al. 2011).

There are a large number of ML models available in the literature (Hastie et al. 2001) that could be used in our ML pipeline. First, we concentrate on models with high empirical performance, as measured by PMLB benchmark (Olson et al. 2017). Second, we look for models from different model classes, such as parametric or nonparametric, linear or nonlinear, black box or white box (Hastie et al. 2001), in order to cover a variety of types of ML models. Our selection resulted in following models: Linear Support Vector Machines (Cortes and Vapnik 1995), K Nearest Neighbors (Hastie et al. 2001), Decision Trees (Quinlan 1986) and Gradient Boosting Model (GBRT) (Friedman 2002). We did not increase the number of model classes in our experiments in order to avoid overfitting due to hyperparameter and model class selection (Cawley et al. 2010) on our relatively small dataset.

We use nested cross validation for training and evaluation of our models. Nested cross – validation has been shown to be efficient for accurate estimation of true out of sample error in face of small dataset size (Varma and Simon 2006); Compared to simple training / testing partition of the data, it utilizes data more efficiently (Varma and Simon 2006). Such approach comprises of outer cross – validation loop, which is used to generate multiple training and testing partitions, and inner cross – validation loop, which is run separately on every training partitions and used to select hyperparameters of the models. For the nested cross validation, a number of partitions in both outer and inner loop needs to be specified. We experiment with different splits of the data in order to test how much of the variance is introduced into estimated error due to various numbers of validation folds (Krstajic et al. 2014). This allows to test whether performance of the model is too optimistic due to favorable setting of numbers of folds. In order to further scrutinize our results, we utilize a statistical test to estimate the probability of getting the test accuracy equal or greater than we obtain by accident, namely permutation test for classification (Ojala and Garriga 2010). In this test, a distribution of scores is obtained by repeatedly training a model on original dataset with labels permuted at random. Based on this distribution, statistical significance is estimated. We use a standard approach in ML in which we calculate the mean and standard deviation of the features on the training partition and subtract the found mean from any input provided to the model and divide it by standard deviation. This is a common procedure in supervised learning (Hastie et al. 2001), which often leads to improved performance of found models.

Results

We measure accuracy of the model as a percentage of samples where the model estimated correctly the ground truth label. Let the test dataset outputs be denoted as $y^{test} \in \{-1,1\}^n$, where $n \in N$ is the size of the test dataset. Let $y^{pred} \in \{-1,1\}^n$ denote estimations made by the model. Then we calculate accuracy (in %) as

$$\text{Accuracy} = 100\ n^{-1} \left(\sum_{i = \{1, 2, 3, \ldots , n\}} I(y_i^{test}\ , y_i^{pred})\right),$$

where function I is and indicator function, which returns 1.0 if it's two arguments are equal, and 0.0 otherwise.

The results of empirical evaluation of different predictive model classes according to methodology discussed above are given in the Table 3. For both inner and outer loops n, m = 3 of folds were used

Model	Out - of - sample accuracy estimate
Most frequent label as output	55%
Gaussian Kernel SVM	70%
K nearest neighbors	70%
Decision Trees	55%
GBRT	55%
Linear SVM	**85%**

Table 3. Empirical Evaluation Of Predictive Models.

The linear SVM classifier model performs best for the prediction task and achieves an accuracy of 85%, which is much higher than baseline accuracy of 55%. We use permutation test (Ojala and Garriga 2010), which is run for 2000 iterations to establish statistical significance of our results. As a result we establish the significance of our results with **p = 0.0075** (see Figure 3).

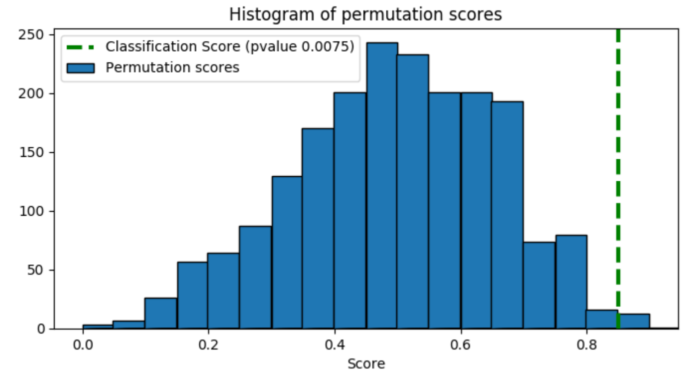


Figure 3. Permutation Scores of Linear SVM

In our experiments, we use m=5 values to which we downsample heart rate measurements (see Figure 2 and corresponding discussion). Values of m smaller than 5 lead to test accuracy of less or equal to 75%.

Furthermore, the test classification accuracy with m larger than 5 was consistently 85%. We kept m=5, as increasing the size of feature vector is more likely to lead to overfitting, known as "Curse of Dimensionality" effect (Hastie et al. 2001).

In order to access the significance of accuracy of the Linear SVM model, we utilize permutation test. Depicted in Figure 3 is the distribution of scores obtained by training Linear SVM on dataset with randomly permutated labels. The likelihood of obtaining accuracy like we do is less than 0.75%. We further analyze the accuracy variation for the Linear SVM w.r.t. different settings of numbers of folds used (see Table 4). The quality of our approach does not appear to depend on number of folds used, except for one case (3 inner folds, 5 outer folds), where accuracy is slightly worse. This shows that the performance that we achieve is unlikely a result of lucky configuration of the folds numbers.

We are using the stratified partitioning of the data in order to come up with results in the Table 4. Stratified partitioning ensures that statistics of the training and testing splits are as close as possible consistent with the original dataset. For some of the configurations of inner and outer number of folds, partitions necessarily consist of only single class instances, and thus are infeasible. Such configurations are marked with n/a.

	3 outer folds	**5 outer folds**	**10 outer folds**	**Leave one out**
3 inner folds	85%	80%	85%	85%
5 inner folds	85%	85%	85%	85%
10 inner folds	n/a	n/a	n/a	85%

Table 4. Effect Of Different Number Of Folds Selected For The Nested Cross – Validation Algorithm.

The results of the algorithm are consistent across different number of folds, except for one case, where accuracy is slightly lower. Finally, we train the Linear SVM using the inner procedure of nested cross – validation and use it as a core of our support system. In order to gain further insights into the model, we analyze the weights that we obtain from the model, which are shown in the Table 5.

We cross-checked validity of our results by using higher thresholds so that larger BMI decrease is considered for a label of -1. Even a BMI change of 0.8 led to significant results, but with smaller accuracy (75% accuracy). With high BMI change thresholds, the learning problem becomes more imbalanced (35% of "lost weight" labels, 65% of "gained weight"). As the distribution of labels becomes more skewed, it becomes harder to learn a meaningful classifier.

HR during run test					HR during recovery					Weight	Age	BMI	Height	Therapy
0.66	0.89	0.83	0.87	0.47	0.44	0.21	0.74	1.0	0.97	--0.29	-0.15	-0.22	-0.23	0.69

Table 5. Weights Of The Best Found Linear Model.

Higher score of the linear model indicates higher probability of a person gaining weight. First, we observe that no parameters were discarded by the linear model. This implies that all of the parameters contribute to the final output of the model. The highest weight values correspond to heart rate related parameters (heart rate during run test and heart rate during cool down). This means, that in our model, heart rate curve value features have the highest influence on the therapy success prediction.

Domain Experts' Prediction on Therapy Success

In the last section, we presented a data-analytical DAIS that provides predictions of whether BMI will decrease after therapy. For our DAIS, ML methods were applied to estimate whether BMI will decrease after 6 months of therapy with an accuracy of 85 % using Linear SVM. We only considered dynamic physiological (heart rate during 6-min-run test) and static medical parameters (e. g. BMI, age) for the estimation, as these parameters are indicators for cardiovascular fitness, which in turn is an indicator for the performance and also successful weight-loss. With the help of the following expert interview we want to compare the accuracy of our ML model with predictions from domain experts and derive practical implications. Furthermore, we are interested in the argumentation inherent in the critical decision-making of the domain experts for additional important parameters that might potentially increase our model accuracy in the future and thus, provide a more accurate DAIS.

Method

To receive predictions of the weight loss success by domain experts and thus, answers to our research question and also practical implications, we conducted an expert interview with mostly quantitative, but also qualitative parts in order to both, receive estimation for whether BMI will decrease after therapy based on domain expert knowledge and explain how argumentation is inherent in the experts' prediction. Instead of a randomly selected sample, an expert interview is based on a group of intentionally selected individuals who on the one hand have expert knowledge and on the other hand are of particular interest to a specific subject matter under study (Meuser and Nagel 1994). In general, expert interview studies involve a small number of participants (ibid.). Furthermore, there are few cardiologists having the expertise required for analyzing the data for the prediction of whether BMI will decrease after therapy.

The domain experts were interviewed for the expert estimation survey. Compared to the machine learning model, the experts were given more extensive information. First, the domain experts were given a short introduction to the fittest study of the children and were briefly informed about the applied therapy of the group A and group B. Second, they were handed over a report for each child with figures of the heart rate curve during the 6-min-run test as well as additional heart rate related parameters such as initial heart rate, maximum and average heart rate during the run test, 1-min-cool down (heart rate recovery after 1 minute) and 3-min-cool down (heart rate recovery after 3 minutes). Furthermore, the report contained information about static parameters such as BMI, weight status category, age, gender, whether the child belongs to therapy group A or therapy group B, and the number of laps the children run during the run test. The experts examined the 20 subjects one by one. The approach that both domain experts adopted in order to draw conclusions from the heart rate data is standard for cardiologists (Jouven, X. et al., 2005; Bird et al., 1998; Singh et al., 2008). In particular, both experts used similar approach by looking for initial heart rate, heart rate peaks, steepness of heart rate curve during run test and cool down as well as whether the heart rate came back down to the initial heart rate after 3-min-cool down.

Furthermore, it has to be noted that domain expert 2 also put a special focus on the heart rate during the run test. Here, the expert was looking for specific fluctuations that can give insights about cardiovascular fitness and motivation during the run test. Thus, interpretations and decisions regarding the prediction tasks might slightly differ due to the fact that both domain experts have prioritized slightly different heart rate-related features.

They estimated whether the BMI of a specific child decreased by at least 0.4 (yes, no) after 6 months. We also asked the experts for each estimation to indicate their confidence of prediction on a seven-point Likert scale ranging from very uncertain (1) to very certain (7). Furthermore, they should justify their estimation and indicate the parameters, which were most important for their estimation. At the end of the survey, the domain experts had the chance to indicate additional parameters they would find useful for the estimation of whether BMI will decrease over time.

Results

Two female cardiologists (Age of Expert 1 = 51 years, Age of Expert 2 = 49 years) took part in the expert interview. Both had on average 22 years of job experience (Std.Dev.= 1). We first assessed the inter-rater reliability between the two experts. For the binary classification task, Krippendorff's alpha (Krippendorf

2011) was calculated for 20 subjects, 2 raters and the nominal data level with R version 3.3.3 and the irr package version 0.84. Krippendorff's alpha resulted in -0.17 indicating disagreement between the two experts. While Expert 1 indicated a higher confidence of her predictions (Mean = 4.25, Std.Dev. = 0.97) than Expert 2 (Mean = 3.30, Std.Dev. = 0.98), the prediction accuracy of Expert 1 (60%) was higher than Expert 2 (40%) but both experts were clearly below the accuracy level of the Linear SVM with 85% as shown in Table 6. The sensitivity and specificity of the Linear SVM are also much higher than the one of the domain experts (specificity = 88.9%, sensitivity = 81.8%).

	TP	FP	TN	FN	Sens	Spec	Acc
Linear SVM	45%	10%	40%	5%	88.9%	81.8%	85%
Domain Expert 1	15%	30%	25%	30%	45.5%	33.3%	40%
Domain Expert 2	30%	25%	30%	15%	54.6%	66.7%	60%

Table 6. Comparison Between Linear SVM And The Two Domain Experts With TP=True Positives, FP=False Positives, TN=True Negatives, FN=False Negatives, Sen=Sensitivity, Spec=Specificity, Acc=Accuracy.

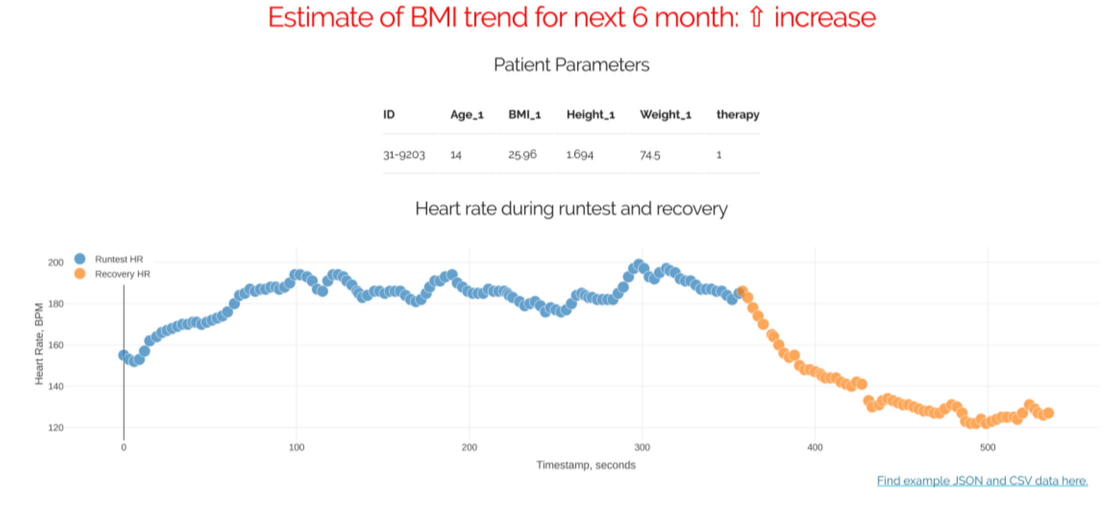
Empirical Adoption of Service among Experts

To assess the domain experts' perceptions of the decision support system with respect to its potential adoption in their everyday life, i.e. in their consultation hours, we designed a second survey and adopted constructs from user satisfaction, technology acceptance (Wixom & Todd 2005) and situation-service fit (Maass, Kowatsch, Janzen, & Filler 2012) and word-of-mouth research (Maxham et al 2002). The method and results of this evaluation are outlined in the following two subsections.

Method

The evaluation procedure consisted of two steps. First, the domain experts were shown a mock-up of the user interface of the DAIS as shown in Figure 4 and were then introduced in its use with regard to a particular situation. The experts had to play through a situation where an obese child that receives their treatment, comes to the consultation hour. Health professionals can use the IS to see static parameters such as age and BMI and heart rate related data during a 6-min-run test and 3-min-cool down of the obese child. Additionally, the system displays the estimation of whether BMI will decrease after 6 months of therapy duration. According to that, the health professional can give feedback and adjust the current treatment.

Second, each subject had to assess the IS with the help of survey instrument. For the development of the questionnaire, we used the constructs from (Wixom and Todd 2005) for the evaluation of the completeness, format, accuracy, quality, accessibility of and satisfaction with the information provided by the decision support system. Furthermore, we used from the same authors the constructs targeting system satisfaction, usefulness of the service, attitude towards the service and intention to use the service. We also used the situation-service fit and situation-behavior-fit items from situation-service-fit theory (Maass et al. 2012) and the intention to recommend construct from prior work (Maxham et al. 2002). Consistent with the anchors for the survey items of the work we used to design our survey instrument, seven-point Likert scales ranging from strongly disagree (unpleasant / very unfavorable for the attitude items) (-3) to neither (0) to strongly agree (pleasant / very favorable for the attitude items) (3) were employed. Domain experts were also asked whether they would recommend the decision support system to their colleagues on a seven-point Likert scale from strongly disagree (-3) to strongly agree (3). We then asked the domain experts to indicate aspects that must be addressed to improve the decision support system and to increase its potential adoption in the future. Two final survey items were included to ask the domain experts whether the questionnaire was understandable and not too long on seven-point Likert scales ranging from strongly disagree (-3) to neither (0) to strongly agree (3) to assess the face validity of the survey instrument.



Estimate of BMI trend for next 6 month: ⇧ increase

Patient Parameters

ID	Age_1	BMI_1	Height_1	Weight_1	therapy
31-9203	14	25.96	1.694	74.5	1

Heart rate during runtest and recovery

Find example JSON and CSV data here.

Figure 4. Mock-up Of The DAIS User Interface

Empirical Results

In addition to the two domain experts that were asked to make predictions on therapy success we invited three more domain experts for the assessment of DAIS. This results in two female and three male cardiologists (Age$_{\text{mean}}$ = 39.6; Std.Dev.= 9.8) with an average of 12.8 (Std.Dev.= 8.9) years of job experience. The mean values and standard deviation of constructs and questionnaire items outlined above are listed in Table 7. The instructions and survey items have been understood by the domain experts although they found the survey slightly too long. With respect to the user satisfaction, technology acceptance and situation service fit items, ratings are either neutral (0) or positive (>0) and there are no negative ratings. In particular, the information format and information accessibility constructs have been evaluated positive by the domain experts indicating an appropriate graphical layout of the information provided by DAIS. Moreover, the participating domain experts would, in general, recommend DAIS to their colleagues. They finally indicated that blood pressure before, during and after the run test, medication, waist-height-ratio, and VO2max need to be addressed to increase its utility and potential adoption of the decision support system in the future.

Construct	Source	# of items	Mean	Std. Dev.
Information completeness	Wixom and Todd (2005)	3 (Item 1, 2, 3)	0.40	0.83
Information format	Wixom and Todd (2005)	2 (Item 1, 3)	1.30	0.57
Information quality	Wixom and Todd (2005)	3 (Item 1, 2, 3)	0.27	0.72
Information accessibility	Wixom and Todd (2005)	2 (Item 1, 3)	1.40	0.89
Information satisfaction	Wixom and Todd (2005)	1 (Item 2)	0.20	1.30
System / service satisfaction	Wixom and Todd (2005)	2 (Item 1, 2)	0.20	1.10

Usefulness of the service	Wixom and Todd (2005)	3 (Item 1, 2, 3)	0.67	0.97
Attitude towards the service	Wixom and Todd (2005)	2 (Item 1, 3)	0.60	0.55
Intention to use the service	Wixom and Todd (2005)	2 (Item 1, 2)	0.00	0.94
Situation-service fit	Maass et al. (2012)	1 (Item 1)	0.60	0.55
Situation-behaviour fit	Maass et al. (2012)	1 (Item 2)	0.20	0.84
Recommendation of the service	Maxham et al (2002)	1 (Item 2)	0.80	0.84
Understandability of the survey	n/a	1	2.80	0.45
Survey was too long.	n/a	1	0.00	1.22

Table 7. Mean Values And Standard Deviations Of The Questionnaire Constructs For The Domain Experts (N=5). Note: The mean scores have been calculated first when a construct was measured with several questionnaire items; The Likert scales ranged from -3 (strongly disagree) to neither (0) to strongly agree (3) and (un)pleasant / very (un)favorable for the attitude items. The items in brackets indicate the corresponding item on the corresponding reference.

Discussion

First, the results of our analyses have shown that Linear SVM, which is the core model for our data-analytical DAIS, performs best with a significant accuracy of 85% (p-value = 0.007) and is much better than the baseline model (accuracy = 55%). The comparison of prediction accuracy between Linear SVM and the domain experts shows that the ML model is performing best compared to the two domain experts under given features (accuracy of domain expert 1 = 60%; accuracy of domain expert 2 = 40%). Moreover, considering additional parameters could lead to different results.

The low accuracy of the domain experts might be due to the fact that the decision process is complex and it is hard to detect and interpret all the specific and individual characteristics leading to weight-loss success. We interviewed experts in order to gain insights in how they make decisions. We found that all experts paid attention to features such as how high the initial heart rate is, how steep the increase of HR during run test is, how high the heart rate values get, and how close the HR during cooldown comes back to the value before run test. We set up an ML experiment where we used these values as features, instead of downsampled heart rate values. We found that test accuracy drops to less or equal to 65%, with any of the model classes we consider. This is close to the best accuracy achieved amongst the interviewed experts (60%). This suggests that such features may not be as informative for future BMI change as downsampled the heart rate values – at least with the ML model classes that we consider. It appears that the shape of the heart rate curve as such is important; This is supported by our experimental results, where a coarser representation of the curve with m<5 values used as features results in lower test accuracy of less or equal to 75%.

Experts suggested that information, which is necessary to make more accurate estimations, is missing, that is blood pressure before, during and after the exercise, VO2max, lactate, medication, and waist-to-height ratio. This information can also be used to increase our ML model accuracy in the future for a more accurate information support system tool. Consistent with literature, both domain experts agreed that BMI, age, initial heart rate before the fittest and the heart rate recovery play an important role for the estimation of whether BMI will decrease over time. However, domain expert 2 additionally considered heart rate during the 6-min-run test. According to domain expert 2, having ups and downs or a decrease in heart rate during the 6-min-run test means that “something is wrong”, i.e. “the child has lack of motivation", "the child didn't take the fittest seriously" or "the child's cardiorespiratory fitness is very low".

Theoretical and Practical Implications

The results of the predictive analytics indicate that we are able to estimate with high accuracy whether BMI will decrease after 6 months of study period by only considering baseline data, i.e. static medical parameters (BMI, height, weight, age) and dynamic physiological parameters (heart rate during 6-min-run test and 3-min cool down) before conducting a therapy. Table 5 shows that the higher the heart rate during the 6-min run test and 3-min cool down, the higher the likelihood that the BMI of the observed child will increase after 6 months, which is also supported by literature. As discussed above, higher heart rate during exercise and cool down are indicators for lower cardiorespiratory fitness, which in turn is an indicator for increased BMI over time. Additionally, ML algorithms might be able to further detect previously unknown or unclear relationships in the data that we have collected. Beside cardiorespiratory fitness, intrinsic motivation is another essential factor of any reliable model of human performance (Befort et al. 2008; Cerasoli et al. 2014). In the field of obesity treatment, the importance of intrinsic motivation has also been shown as a predictor of treatment success (Befort et al. 2008; Cerasoli et al. 2014). There are various studies addressing the relationship between intrinsic motivation and heart rate pattern under specific conditions such as physical exercise (Higginson 2016; Thøgersen-Ntoumani et al. 2015). Closer investigation of specific information leveraged by the ML model could lead to improved accuracy of the model via better feature engineering, and thus is a subject of future work.

The comparison of the prediction by our ML model and domain experts exemplarily shows that the decision making for predicting the therapy success is very complex and might be hard to accomplish even for experienced healthcare practitioners. We suggest a DAIS, which provides predictions of the therapy success and support healthcare practitioners in the decision making for suitable obesity therapies and thus, personalize therapies based on individual patient characteristics. This is also encouraged by the acceptance ratings of DAIS and the answers of the five domain experts as listed in Table 7. The experts particularly rated the information layout (i.e. its accessibility and format) positive and, with no negative ratings, agree that such data-analytical DAIS and its recommendations might be useful. Furthermore, the five experts tend to recommend DAIS to their colleagues. These results show that our predictive DAIS has the potential to be used as a tool for personalized medicine. It might help decreasing healthcare costs and also deriving time-efficient treatment without initiating unsuitable therapies.

Conclusion and Future Work

The origin of obesity represents a complex and constantly growing health problem, which is already widely spread amongst children and adolescents. Even though many scientists and healthcare practitioners developed various therapy programs, a well-defined solution that guarantees therapy success is still missing. An era of patient-centric healthcare is emerging, where individual patients are placed in the center of therapies and analysis are done to find out which treatments are optimal for each patient (Garvey et al. 2014; Garber et al. 2013; Jensen et al. 2014). We introduced a DAIS for health practitioners, which allows to make predictions whether BMI will decrease in the future, before conducting a therapy. The goal of the paper was to investigate if data on static and dynamic healthcare parameters can predict whether BMI will decrease after obesity treatment therapy, before conducting the therapy as such (see RQ1). Furthermore, we wanted to analyze whether healthcare practitioners perceive these predictions as being useful (see RQ2). The results of our analysis have shown that Linear SVM has the highest accuracy (85 %, p-value = 0.007) and works best for RQ1. The comparison between the prediction of Linear SVM and the domain expert estimation showed that the prediction accuracy provided by our DAIS is significantly higher than the one of the domain experts.

Furthermore, the domain experts are uncertain in making decisions about the prediction on whether BMI will decrease over time. Consequently, physicians might be motivated to use the DAIS as an additional clinical decision support in the treatment of obesity. This assumption is also encouraged by the results of the adoption survey as listed in Table 7, which provides the answer to RQ2. The idea is not to replace health professionals and their decisions but rather provide a decision supporting system, which allows a time- and cost-efficient treatment. ML is the basis of our information support system, since it provides a technical basis to better identify complex and non-trivial patterns in continuous biological signals (e.g.

heart rate). Nonetheless the complexity of the origin of obesity and the relationship between heart rate under physical activity and therapy success is too high and our understanding is still in its infancies for drawing final conclusions. Furthermore, there are several limitations, which need to be overcome in the future. First, we have a small sample size. Larger sample size could lead to a higher accuracy. Second, our analysis is only based on two obesity treatment therapies. Other therapy options could lead to different results. Third, the children considered in the study are of the same nationality. It would be interesting to compare results made with children of different nationalities. Moreover, in our study, we considered a therapy period of 6 months. In future studies, we plan to consider various other settings, such as variable therapy duration. Additionally, sensory readings over a prolonged period of time offer a huge potential for data analysis.

Furthermore, it should be mentioned that data-analytical DAIS used for predictions in the context of healthcare is critically discussed amongst health professionals and is still met with skepticism. In subsequent studies we want to further collaborate with domain experts to enhance our DAIS with their expert knowledge on the one hand and to better meet the needs of health professionals on the other hand. The future goal would be a DAIS which not only provides predictions whether BMI will decrease after therapy but also can give advices which therapy method might be best suited to an individual patient. Therefore, we want to conduct further studies with a larger sample size, where also normal weight children and adolescents will be considered to overcome problems due to sampling bias. Moreover, we will follow comments of the domain experts by considering additional parameters to increase our ML model accuracy and to improve our DAIS in terms of prediction quality. This also includes setting up the study environment to consider intrinsic motivation. As mentioned in the *Discussion*, an individual's intrinsic motivation to lose weight as well as self-motivation have been identified as predictors of successful weight loss in previous papers. Moreover, the relationship between intrinsic motivation and heart rate under physical exercise has been addressed in several studies. Further studies could attempt to leverage the relationship between intrinsic motivation and heart rate for the prediction of therapy success.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable feedback that helped us to improve the paper. Furthermore, we want to thank Prof. Dagmar l'Allemand (MD) and Björn Brogle for cooperation. We are also immensely grateful to Prof. Michael Böhm (MD) and his team from the Saarland University Medical Center as well as Ingrid Seipelt (MD) from Medical Care Center of the SHG-Kliniken Völklingen GmbH, for participating in the interviews. The work is part-funded by the Swiss National Science Foundation, project number 159289 (http://p3.snf.ch/Project-159289) and the German Research Foundation, project number: 273255463 (http://gepris.dfg.de/gepris/projekt/273255463).

References

Adnan, M., Husain, W., and Rashid, N. 2012a. "Parameter Identification and Selection for Childhood Obesity Prediction Using Data Mining," *2nd International Conference on Management and Artificial Intelligence* (35), pp. 75-80.

Adnan M., Hariz M., Husain, W., and Rashid, A. 2012b. "A hybrid approach using Naïve Bayes and Genetic Algorithm for childhood obesity prediction," *Computer & Information Science (ICCIS)* (1), pp. 281–285.

Ayash C. R., Simon, SR., Marshall R., et al. (2013). "Evaluating the impact of point-of-care decision support tools in improving diagnosis of obese children in primary care," *Obesity (Silver Spring)* (21:3), pp. 576–582.

Barlow, SE. 2007. "Expert Committee Expert committee recommendations regarding the prevention, assessment, and treatment of child and adolescent overweight and obesity: summary report," *Pediatrics* (120:Suppl 4), pp. 164–92.

Bassett, D. R., and Howley, E. T. 2000. "Limiting factors for maximum oxygen uptake and determinants of endurance performance," *Med. Sci. Sports Exerc.* (32:1), pp. 70–84.

Befort, CA., Nollen, N., Ellerbeck, EF., Sullivan DK., Thomas JL., and Ahluwalia JS. 2008. "Motivational interviewing fails to improve outcomes of a behavioral weight loss program for obese African American women: a pilot randomized trial," *J Behav Med.* (31:5), pp. 367–377.

Bird, SR., Smith, RA., James, K. 1998. *Exercise Benefits and Prescription*, Cheltenham, Nelson Thornes.

Brener, ND., Kann, L., Kinchen, SA., et al. 2004. "Methodology of the youth risk behavior surveillance system," *MMWR Recomm Rep.* (24:53), pp. 1–13.

Cawley, G.C. and Talbot, N.L., 2010. "On over-fitting in model selection and subsequent selection bias in performance evaluation." *Journal of Machine Learning Research*, (11:Jul), pp.2079-2107.

Cerasoli, CP., Nicklin, JM., and Ford, MT. 2014. "Intrinsic Motivation and Extrinsic Incentives Jointly Predict Performance: A 40-Year Meta-Analysis," *Psychological Bulletin*. (140:4), pp. 980-1008.

Chock, GY., and Kerr NA. 2011. "A Report on the development of the Hawai'i Pediatric Weight Management Toolkit," *Hawaii Med J.* (70:7 Suppl 1), pp. 49–51.

Collins, JC., and Bentz, JE. 2009. "Behavioural and psychological factors in obesity," *The Journal of Lancaster General Hospital* (4:4), pp. 124–127.

Cortes, C. and Vapnik, V., 1995. "Support-vector networks," *Machine learning* (20:3), pp. 273-297.

Dimkpa, U., (2009). "Post-Exercise Heart Rate Recovery: An Index of Cardiovascular Fitness," *Journal of Exercise Physiology* (12:1), pp. 10-22.

Dixon, C. B., Masteller, B., and Andreacci, J. L. 2013. "The effect of a meal on measures of impedance and percent body fat estimated using contact-electrode bioelectrical impedance technology," *European journal of clinical nutrition*, (*67*:9), p. 950-955.

Dupuis, JM., Vivant, JF., Daudet, G., Bouvet, A., Clement, M., Dazord, A., et al. 2000. "Personal sports training in the management of obese boys aged 12 to 16 years," *Arch Pediatr.* (7), pp. 1185-1193.

Dugan, TM., Mukhopadhyay, S., Carroll, A., and Downs, S. 2015. "Machine learning techniques for prediction of early childhood obesity," *Applied Clinical Informatics* (6:3), pp. 506-520.

Friedman, JH., 2002. "Stochastic gradient boosting." *Computational Statistics & Data Analysis* (38:4), pp. 367-378.

Garber, AJ., Abrahamson, MJ., Barzilay, JI., et al. 2013. "American Association of Clinical Endocrinologists' comprehensive diabetes management algorithm 2013 consensus statement," *Endocr Pract.* (19:Supplement 2), pp. 1–48.

Garvey, WT, Garber, AJ, Mechanick JI, et al. 2014. "American Association of Clinical Endocrinologists and American College of Endocrinology consensus conference on obesity: building an evidence base for comprehensive action," *Endocr Pract.* (20:9), pp. 956–976.

Gilbert, MJH., Zerulla, TC., and Tierney, KB. 2014. "Zebrafish (Danio rerio) as a model for the study of aging and exercise: physical ability and trainability decrease with age," *Experimental Gerontology* (50), pp. 106–113.

Hastie, T., Tibshirani, R. and Friedman, J. 2001. *Overview of supervised learning. In The elements of statistical learning,* New York: Springer, NY, pp. 9-41.

Higginson, K. 2016. "Distraction, Enjoyment, and Motivation During an Indoor Cycling Unit of High School Physical Education," PhD thesis. Brigham Young University.

Jensen, MD., Ryan, DH., Donato, KA., et al. 2014. "Guidelines (2013) for managing overweight and obesity in adults." *Obesity* (22:Suppl. 2), pp. 1-3.

Jouven, X., Empana, J. P., Schwartz, P. J., Desnos, M., Courbon, D., & Ducimetière, P. 2005. "Heart-rate profile during exercise as a predictor of sudden death," *New England Journal of Medicine*, (*352*:19), pp. 1951-1958.

Krippendorf, K. 2011. Computing Krippendor's Alpha-Reliability. Retrieved from hpttp://repository.upenn.edu/asc_papers/43.

Krstajic, D., Buturovic, LJ., Leahy, DE., and Thomas, S. 2014. "Cross-validation pitfalls when selecting and assessing regression and classification models," *Journal of Cheminformatics* (6:9), p. 10.

Lee, D., Artero, E. G., Sui, X., Blair, S. N. 2010. "Mortality Trends in the General Population: The Importance of Cardiorespiratory Fitness," *Journal of Psychopharmacology (Oxford, England)* (24:4_supplement), pp. 27–35.

Mannini, A., and Sabatini, A. M. 2010. "Machine learning methods for classifying human physical activity from on-body accelerometers," *Sensors* (10:2), pp. 1154–1175.

Maass, W., Kowatsch, T., Janzen, S., & Filler, A. 2012. "Applying Situation-Service Fit to Physical Environments Enhanced by Ubiquitous Information Systems," *20th European Conference on Information Systems (ECIS)* (221), pp. 1-12.

Maxham, JG. and Netemeyer, RG. 2002. "Modeling Customer Perceptions of Complaint Handling: The Effects of Perceived Justice on Complainant Attitudes and Intentions," *J Appl Physiol.* (78), pp. 239-252.

Mazzeo, RS., and Marhsall, P. 1989. " Influence of plasma catecholamine on the lactate threshold during graded exercise," *Journal of Retailing* (67:4), 1319–1322.

Meuser, M., and Nagel, U. 1994. *Expertenwissen und Experteninterview.* in *Expertenwissen. Die institutionelle Kompetenz zur Konstruktion von Wirklichkeit, R.* Hitzler, A. Honer, and C. Maeder (eds.) Opladen: Westdeutscher Verlag, pp. 180–192.

Montesi, L., El Ghoch, M., Brodosi, L., Marchesini, G., Calugi, S., Dalle Grave, R. 2016. “Long-term weight loss maintenance for obesity: a multidisciplinary approach,” *Diabetes Metab. Syndr. Obes. Targets Ther.* (26:9), pp. 37–46.

Mota, J., Ribeiro, JC., Carvalho, J., Santos, MP., and Martins, J. 2009. “Cardiorespiratory fitness status and body mass index change over time: a 2-year longitudinal study in elementary school children,” *Int J Pediatr Obes.* (4:4), pp. 338–342.

Ojala, M. and Garriga, GC. 2010. “Permutation tests for studying classifier performance,” *Journal of Machine Learning Research* (11:Jun), pp. 1833-1863.

Olson, R.S., La Cava, W., Orzechowski, P., Urbanowicz, R.J. and Moore, J.H., 2017. “PMLB: a large benchmark suite for machine learning evaluation and comparison,” *BioData mining*, (10:1), p. 36.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J. 2011. “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research* (12:Oct), pp. 2825-2830.

Quinlan, JR. (1986). “Induction of decision trees.” *Machine learning* (1:1), pp. 81-106.

Rattay, KT., Ramakrishman, M., Atkinson, A., Gilson, M., and Drayton, V. 2003. "Use of an electronic medical record system to support primary care recommendations to prevent, identify, and manage childhood obesity," *Pediatrics* (123:Suppl. 2), pp. 100-107.

Reinher, T., Brylak, K., Alexy, U., Kersting, M., and Andler, W. 2009., "Predictors to success in outpatient training in obese children and adolescents.," *Int J Obes Relat Metab Disord.* (27:9), pp. 1087-1092.

Rousseeuw, P.J. and Leroy, A.M., 2005. “Robust regression and outlier detection” (Vol. 589). John wiley & sons.

Scime, A. 2015. “Expert Knowledge in Data Mining,” in *Encyclopedia of Information Science and Technology*, M. Khosrow-Pour (eds.), Idea Group Reference, p. 1777.

Shmueli, G., and Koppius, O. 2011. "Predictive Analytics in Information Systems Research," *MIS Quarterly* (35:3), pp. 553-572.

Singh, T.P., Rhodes, J., and Gauvreau, K. 2008. “Determinants of Heart Rate Recovery Following Exercise in Children,” *Med Sci Sports Exerc.* (40:4), pp. 601-605.

Thøgersen-Ntoumani, C., Shepherd, S., Ntoumanis, N., Wagenmakers, A. and Shaw, C. 2015. “Intrinsic Motivation in Two Exercise Interventions: Associations With Fitness and Body Composition.” *Health Psychology.* (35:2), pp. 195-198.

Varma, S., and Simon, R. 2006. "Bias in error estimation when using cross-validation for model selection," *BMC bioinformatics* (7), p. 91.

Warburton, DE., Haykovsky MJ., Quinney, H. A., Humen, D. P. and Theo, KK. 1999. “Reliability and validity of measures of cardiac output during incremental to maximal aerobic exercise. Part I: conventional techniques,” *Sports Medicine* (27:1), pp. 23–41.

Witten, IH., and Frank, E. 2005. *Data Mining: Practical machine learning tools and techniques,* Burlington: Elsevier.

Wixom, BH., and Todd, PA. 2005. "A Theoretical Integration of User Satisfaction and Technology Acceptance," *Information Systems Research* (16:1), pp. 85-102.

Zhang, S., Tjortjis, C., Zeng, X., Qiao, H., Buchan I., and Keane, J. 2009. "Comparing data mining methods with logistic regression in childhood obesity prediction," *Information Systems Frontiers* (11:4), pp. 449–460.