Correlates of Physical Activity Behavior in Adults: A Data Mining Approach

Vahid Farrahi (Vahid.farrahi@oulu.fi)
Oulun Yliopisto https://orcid.org/0000-0001-8355-8488
Maisa Niemelä
University of Oulu
Mikko Kärmeniemi
University of Oulu
Soile Puhakka
University of Oulu
Maarit Kangas
University of Oulu
Raija Korpelainen
University of Oulu
Timo Jämsä
University of Oulu

Methodology

Keywords: Decision tree, CHAID, Multilevel model, Prediction, Classification

Posted Date: June 25th, 2020

DOI: https://doi.org/10.21203/rs.2.23726/v3

Version of Record: A version of this preprint was published on July 23rd, 2020. See the published version at https://doi.org/10.1186/s12966-020-00996-7.

Abstract

Purpose: A data mining approach was applied to establish a multilevel hierarchy predicting physical activity (PA) behavior, and to methodologically identify the correlates of PA behavior.

Methods: Cross-sectional data from the population-based Northern Finland Birth Cohort 1966 study, collected in the most recent follow-up at age 46, were used to create a hierarchy using the chi-square automatic interaction detection (CHAID) decision tree technique for predicting PA behavior. PA behavior is defined as *active* or *inactive* based on **machine-learned** activity profiles, which were previously created through a multidimensional (clustering) approach on continuous accelerometer-measured activity intensities in one week. The input variables (predictors) used for decision tree fitting consisted of individual, demographical, psychological, behavioral, environmental, and physical factors. Using generalized linear mixed models, we also analyzed how factors emerging from the model were associated with three PA metrics, including daily time (minutes per day) in sedentary (SED), light PA (LPA), and moderate-to-vigorous PA (MVPA), to assure the relative importance of methodologically identified factors.

Results: Of the 4,582 participants with valid accelerometer data at the latest follow-up, 2,701 and 1,881 had active and inactive profiles, respectively. We used a total of 168 factors as input variables to classify these two PA behaviors. Out of these 168 factors, the decision tree selected 36 factors of different domains from which 54 subgroups of participants were formed. The emerging factors from the model explained minutes per day in SED, LPA, and/or MVPA, including body fat percentage (SED: B=26.5, LPA: B=-16.1, and MVPA: B=-11.7), normalized heart rate recovery 60 seconds after exercise (SED: B=-16.1, LPA: B=9.9, and MVPA: B=9.6), average weekday total sitting time (SED: B=34.1, LPA: B=-25.3, and MVPA: B=-5.8), and extravagance score (SED: B=6.3 and LPA: B=-3.7).

Conclusions: Using data mining, we established a data-driven model composed of 36 different factors of relative importance from empirical data. This model may be used to identify subgroups for multilevel intervention allocation and design. Additionally, this study methodologically discovered an extensive set of factors that can be a basis for additional hypothesis testing in PA correlates research.

Introduction

The positive relationship between physical activity (PA) and health has been well established (1,2), yet many adults worldwide perform insufficient PA (3). Thus, understanding the factors associated with PA behavior is essential to develop and improve public health interventions (3–5). Many studies have investigated the association of various factors including personal, societal, and environmental factors with different PA behavior indices such as the daily amount of moderate-to-vigorous PA (MVPA) or sedentariness (5–7). Despite much progress in research into correlates, only a few studies have followed analytical approaches that account for both the existence of several levels of influence (5,7,8) and the complexity and multimodality of PA behavior (5,9,10). To further advance correlates research, there has

been calls for more research using both sophisticated statistical assessment that can capture the multilevel nature of correlates (4,5) and PA behavior definitions that better reflect everyday life rather than unidimensional metrics such as daily MVPA (1,5,9,10).

Using classical statistical modeling (such as regression analyses), studies have generally examined whether and how various factors are associated with different PA metrics (6,11). In classical statistics, these analyses could remain restricted to data analysts' decisions about how the association and interaction are hypothesized (knowledge-driven) mainly because the factors selected for inclusion in the analyses are primarily chosen subjectively according to their conceptual relevance and, in some cases, initial empirical associations (11,12). This may limit the recognition of new and innovative correlate categories, which are needed in this field for further progress (5,11). Ecological approaches that integrate ideas from several theories have been also used in correlates research, often to overcome classical statistical analysis limitations (5). They have been used to both conceptualize the factors and their interrelationships at all levels explaining PA behavior (such as the interconnections between individuals and their social and physical environments) (13) and guide variable selection for analyses (5,11). However, ecological approaches are also knowledge-driven (6) and, to some extent, rely on very well-established correlates (6,8), which might result in missing some factors and interrelationships associated with PA behavior.

We have now entered a data-intensive era, with an increasing popularity of data mining approaches (14). Such approaches originated from statistics but are known to capture hidden and novel insights buried in large amounts of data and generate data-driven hypotheses (14,15). These principles also regard the field of PA research, in which there is a need for more complex approaches to identify the next generation of PA behavior correlates, understand their relative importance, and capture the complex interrelations among the factors at different levels (5,6,8). Several studies have applied data mining approaches (16–19) mostly to establish data-driven correlate hierarchies (16,17) but using a limited number of factors and self-reported measurement of PA or sedentary behavior.

The present study applied a predictive data mining approach to classify individuals' PA behavior (defined as active or inactive) using an extensive list of individual, demographical, psychological, behavioral, environmental, and physical factors PA behavior, to better represent everyday life, was defined based on **machine-learned** activity profiles established preciously using a multidimensional (clustering) approach applied on continuous accelerometer-measured activity intensities in one week (20). This cross-sectional study sought to build a data-driven hierarchy of PA behavior correlates from empirical data and, as a secondary purpose, to methodologically identify PA behavior correlates from a wide list of factors.

Materials And Methods

Data for the present study were from the population-based Northern Finland Birth Cohort 1966 study (NFBC1966). NFBC1966 is a life-course study involving participants whose dates of birth were expected to be in 1966 in Finland's two northernmost provinces, Oulu and Lapland (n=12,058, 96.3% of all live

births in the study area). The present cross-sectional study included NFBC1966 cohort members who participated in the latest follow-up at age 46 and agreed to wear accelerometers for device-based physical activity measurements (21). A total of 10,321 NFBC1966 cohort members (85.6% of all cohort members) were alive in Finland in 2012 and were invited to the follow-up, of which 5,621 (46.6% of all cohort members and 54.4% of those who were invited) participated and wore accelerometers (Fig. 1). With respect to the measurement tools/techniques, the collected data can be categorized into four: self-reported measures, clinical measures, objective built and natural environmental measures, and objective physical activity measures.

Questionnaires and measurements

Questionnaires

A postal questionnaire was sent to all living cohort members with known addresses. The questionnaire included items on social background, frequency and type of habitual exercises, physical and psychological health and well-being, and work–life and socioeconomic situation. In addition, health-related behaviors were assessed by a separate questionnaire, the Quality Of Life Questionnaire (15D©), to rate health-related quality of life (22). Another additional separate survey was used to address opinions and experiences, covering questions from the Temperament and Character Inventory (TCI) questionnaire (23). The temperament and personality trait scores were then composed based on the responses to the items of the TCI questionnaire. More details on the self-reported measures can be found elsewhere (24).

Clinical examination and measurement of physical activity

Participants were also invited to attend a clinical examination. The clinical examinations included measurement of anthropometry, body composition, and cardiorespiratory fitness. Participants' height, weight, blood pressure and waist-hip ratio were measured and BMI (body mass index) calculated. Participants' body composition was measured with bio-impedance measurement (InBody720, InBody, Seoul, Korea). A static back muscle strength test (Biering-Sorensen trunk extension test) was performed to evaluate physical performance. A submaximal four-minute single-step test during which heart rate was continuously monitored was performed to assess cardiorespiratory fitness. Further details on the clinical examination protocol and measures are presented elsewhere (25,26).

Objective measurement of physical activity was initiated during clinical examination using a wrist-worn accelerometer (Polar Active, Polar Electro Oy, Kempele, Finland). Participants were instructed to wear the monitor on the wrist of their non-dominant hand continuously for 24 hours for 14 days. Polar Active has a uniaxial accelerometer that outputs estimated energy expenditure in metabolic equivalent (MET) values every 30 seconds. The validity of Polar Active under free-living conditions against the double-labeled water technique has been shown elsewhere (27).

Environmental measures

We obtained the residential coordinates of all participants whose residences were available at the time of the 46-year follow-up data collection (2012–2014) from the Finnish Population Register Centre. We used a geographic information system (ArcGIS 10.3) to calculate built, natural, and socioeconomic environment variables (Supplementary File 1, Table S1) that might describe the conduciveness of participants' residential environment to PA. We calculated all variables in the year the participant attended the 46-year data collection. We also determined quantitative environmental features using a one-kilometer-radius circular buffer around the residential locations, and the distances (as the crow flies) to amenities were measured using road network data.

Data related to community structure; land use; amenities such as retail, recreation, office, and community institutions; and socioeconomic factors were derived from the Finnish community structure database (28). Street network data, including the number of bus stops, intersection density, and length of cycle paths, were based on the Finnish national road and street database (Digiroad) (29). Data on indoor and outdoor sport facilities were obtained from the Finnish database of sport facilities (30). Natural environment features such as distances to the closest forests and parks and residential area greenness were assessed with the land cover data from the Finnish Environment Institute (31).

Data mining using a decision tree

We selected a decision tree technique to establish a data-driven model for classifying PA behavior. A decision tree model is created by partitioning the data on the basis of several independent input variables (or predictors) to form homogenous subgroups with respect to the outcome variable. A decision tree-produced hierarchy has a flow chart-like structure that enables identifying the relative importance of input variables in predicting the outcomes; the predictors in the higher layers of hierarchy are more important predictors (32). In clinical applications and several other areas in which interpreting the results is of vital importance, decision trees are one of the most widely used classification methods (12,14,32,33).

We used the Chi-squared Automatic Interaction Detection (CHAID) decision tree algorithm to create the model (34). CHAID has been repeatedly used in studies with clinical applications whose main purpose was to identify key factors related to the outcomes of interest (35,36). In this algorithm, homogenous groups may be formed by any possible combination of the known values of a categorical predictor, or by setting cut-off points at any values of a continuous predictor. The number of selected independent predictors for creating the model together with the number categories (for categorical and ordinal) and intervals (for continuous) for the selected independent predictors depends on results of the Chi-square analyses and whether the differences are significant or not. Since the correlates of PA behavior could be of mixed data types, CHAID is an appropriate candidate because it uses a nonparametric procedure with no assumptions of the underlying data and is designed to include continuous, ordinal, and categorical predictors (33).

Decision tree model construction and validation

Input variables (predictors) and physical activity behavior (outcome variable)

The questionnaire and clinical and environmental measures, except those with more than ~10% missing values, were used as input variables. Recent evidence suggests that any single unidimensional metric (including the most commonly used criterion that defines physical inactivity as the insufficient activity level to meet present recommendations (1)) might not be enough to define individuals' PA behavior (10,37–39). We therefore used participants' activity profiles, which we built in a previous study using a multidimensional approach and continuous accelerometer data to define the PA behaviors for the present study (20). A distinct aspect of this approach is that continuous accelerometer-measured activity intensities in one full week across the whole intensity continuum, including sedentary (SED), light PA (LPA), and MVPA were incorporated into a machine learning approach to create the activity profiles.

The details about how the activity profiles were established have been presented elsewhere (20). Briefly, X-means clustering algorithm was applied on accelerometer-based MET-level data of participants who had seven consecutive valid measurement days (N = 4,582), and four distinct activity profiles (clusters) were derived. A total of 1008 features/variables (10-minute averages of the original 30-sec MET data resulting in 144 MET values for each of the 7 valid measurement days) for each participant were fed into the clustering algorithm for creating the profiles (20). A valid measurement day was defined as at least 600 minutes of activity monitor wearing time per day during waking hours. Seven consecutive valid measurement days were used as a criterion to enable analyzing one full week including both weekdays and weekends. The activity profiles were named with respect to the temporal and intensity patterns of participants' daily activities in each cluster: Inactive (N = 1,881), Moderately active (N = 802), Evening active (N = 1,297), and Very active (N = 602). The results of our initial experiments revealed the decision trees induced for classifying the four activity clusters have unreasonable performance and generalizability, primarily because the outcome variable had both class imbalance (i.e., 41% Inactive, 18% Moderately active, 28% Evening active, and 13% Very active) and class overlap (i.e., those who were in the Moderately active, Evening active, and Very active had comparable activity profiles with different temporal patterns) problems (40). Previous research has shown that the effects of these two problems that associate with each other in limiting the performance and generalizability of classification trees is best minimized with near-balanced class distribution in the outcome variable (41). We therefore defined those in the Moderately active, Evening active, or Very active clusters as *active* (N = 2,701), and the remaining ones who were in the Inactive cluster as *inactive* (N = 1,881). We used the input variables in their original form to classify the two PA behavior categories: *active* and *inactive*.

Missing values and algorithm parameters

Missing values were included in the analysis as a separate category that was allowed to merge with other categories in the decision tree. The imputation of missing values of input variables was unnecessary (35). A previous study has shown that the a decision tree developed with the presence of missing values in their input variables has reasonable misclassification rates, especially when the missing values are not very high (e.g., 20%) (42).

Several parameters must be set prior to constructing a decision tree model. Of these parameters, pruning criteria are the most primary ones to limit the size of the tree and prevent overfitting (14). The pruning criteria were set such that groups smaller than 80 were not split any further (maximum number of participants in a parent node), and no group smaller than 40 was formed (maximum number of participants in a child node). The tree growth was limited to 10 layers, meaning that a maximum of 10 factors could be selected to form a group.

Model validation and visualization

We created and validated the model using 10-fold cross-validation. To evaluate the accuracy of the final decision tree model, we used the confusion matrix, which shows the proportion of participants with each outcome variable that was correctly and incorrectly classified. In the visualization of the final tree, the percentage of active and inactive participants in each subgroup, along with the response index (RI), was presented. The RI is the percentage of inactive participants in each subgroup relative to that of inactive participants in the total sample (i.e., 41.1%). Similar to an odds ratio, RI is an indicator of the direction and strength of the association (16).

Activity patterns in decision tree-formed subgroups of participants

Given that the outcome variable was formed with a multidimensional approach, we also calculated Z-scores of three PA metrics including average daily time (minutes per day [min/day]) spent in SED, LPA, and MVPA in each decision tree-formed subgroup of participants. A Z-score indicates how many standard deviations the mean of a measure in a subgroup is away from the corresponding mean in whole study population. As such, we could compare the variation of the three activity intensities across different subgroups with respect to the study population means. We calculated these three PA metrics from the same seven consecutive valid measurement days to establish the activity profiles (20) using previously validated cut-points (SED, 1–1.99 MET; LPA, 2–3.49 MET; and MVPA, ≥ 3.5 MET) by the accelerometer manufacturer (43).

Association analysis

The same above-mentioned PA metrics (SED, LPA, and MVPA) were also used for association analyses. We examined the association between factors emerging from the model and these PA metrics to determine the significance and relative importance of the methodologically identified factors. We used adjusted generalized linear mixed models, including urban–rural area as a random effect, to examine the associations between each independent variable (factor emerging in the decision tree) separately with min/day in SED, LPA, and MVPA. Age and gender were used as covariates in all models. We standardized the continuous independent variables to obtain a mean of zero and a standard deviation (SD) of 1 before including them in regression analyses. As such, we could interpret coefficients (B) from the models encompassing a continuous independent variable as a change in the outcome (e.g., min/day of LPA) for every 1 SD change in the independent variable and therefore compare them to each other across a similar outcome in terms of magnitude regardless of the unit. We included the categorical and ordinal

independent variables in the regression analyses in the form of dummy variables and set response categories at the lowest end as the reference category. A p-value of 0.05 was used to interpret significance. All analyses (including data mining) were performed with IBM SPSS Statistics for Windows, version 25.0 (IBM Corporation, Armonk, USA).

Results

Participants

A total of 4,582 participants (38% of all cohort members and 44.4% of those invited to the 46-year follow-up) had enough valid PA data to be included in the cluster analysis study (20) and, accordingly, sufficient information on the outcome value (active or inactive profile) for inclusion in the present study. The numbers of participants with an active and inactive profiles were 2,701 (58.9%) and 1,881 (41.1%), respectively. The characteristics of the study's participants for the whole sample, with respect to the two outcome variables, are shown in Table 1. These descriptive results are identical to those reported in cluster analysis study (20).

Input variables

We used a total of 168 factors as input variables after eliminating those with over ~10% missing values. Overall, the factors related to medication use and diseases had the highest number of missing values (~20%–50%) while the number of missing values in environmental and adiposity-related factors were lowest (~1%–5%). Of these 168 factors, 82 were continuous, 19 were categorical, and 67 were ordinal factors. All the 168 input variables are given in the Supplementary File 1, Tables S1–S3.

Decision tree model

The prediction results are presented in Table 2. The overall classification accuracy was 69.7%. The final decision tree is shown in Fig. 2. The decision tree algorithm selected a total of 36 different factors of different domains, by which 54 subgroups of participants were formed **(marked in Fig. 2 as S1-S54)**, 26 predicted as active and 28 as inactive. The most frequently appeared factor in the model, appearing three times, was 'average weekday total sitting time', followed by 'average weekday sitting time at the office or such places', 'body fat percentage', 'frequency of exercise through walking', 'urban-rural areas', and 'difficulty of a 5-kilometer run without breaks', which each appeared twice. Other variables appeared only once. The number of layers (or factors) for forming subgroups ranged from two to seven, even though the allowed maximum number of layers was 10.

Overall, participants with higher body fat percentage (>31%) were more likely to be inactive (RI range: 1.16–1.49) compared with those with lower body fat percentage (<28.3%). The largest subgroup of inactive participants (n=193, RI=1.55) included those with the highest body fat percentage who reported their physical activity frequency through gardening more than once a month, and were with a normalized heart rate recovery slope <55% per second. The largest active subgroup (n=335, RI=0.39) was composed

of participants with the lowest body fat percentage in the study population and with a normalized heart rate recovery 60 seconds after exercise >25 beats per minute. Participants who lived in city/rural centers and had a physically demanding occupation (i.e., process and transport workers, forestry workers and farmers, and other workers) had the least risk of being inactive (RI=0.11).

SED, LPA, and MVPA variations in the decision tree-formed subgroups of participants

The variations in the three activity intensities in the 54 decision tree-formed subgroups of participants are shown in Fig. 3. Most inactive and active subgroups had different accumulation patterns of SED, LPA, and MVPA. In general, although most active subgroups had lower SED level than the population mean, some subgroups had noticeably higher levels of MVPA (e.g., subgroups 3, 6, and 7), while others had noticeably higher levels of LPA (e.g., subgroups 20, 32, 33, and 52). Inactive subgroups had generally higher SED level and lower MVPA level than the population mean, while several subgroups had noticeably both lower LPA and MVPA levels (e.g., subgroups 41, 46, 49, and 51).

Association analysis

Tables 3 and 4 show the association between the continuous, categorical, and ordinal explanatory variables from the decision tree model and the three PA metrics in the total study population. All factors except fear of uncertainty and impulsiveness scores were associated with at least one PA metric. Most continuous factors (Table 3) in the relatively high layers of the decision tree model and larger subgroups significantly explained min/day in all the three PA metrics. For example, body fat percentage was positively associated with SED level (B=26.5) and inversely associated with LPA (B=-16.1) and MVPA (B=-11.7) levels. Higher normalized heart rate recovery 60 seconds after exercise was associated with lower SED (B=-16.1) and higher LPA (B=9.9) and MVPA (B=9.6). Categorical factors were also associated with min/day in SED, LPA, and/or MVPA (Table 4). For instance, those with physically strenuous occupations (workers, farmers, service, sales, and care staff compared with managers, advisers, office workers, etc.) spent less time in SED (B=-46.7) and more time in LPA (B=41.1) and MVPA (B=3.5). Those who reported a higher frequency of physical activity through gardening (2–3 times a month or higher compared with fewer than once month or not at all) had lower SED (B=-20.6) and higher LPA (B=14.4).

Overall, from the regression coefficients (B) in Tables 3 and 4 (indicative of changes in min/day of SB, LPA, and MVPA for every 1 SD change in the predictor and of changes from the reference response categories, respectively), the associations seemed generally stronger for those factors that emerged in the higher layer and larger subgroups. For instance, higher body fat percentage and lower normalized heart rate recovery slope were associated with lower and higher min/day in MVPA, respectively, but the former, which appeared in the higher level of the decision tree, was associated with MVPA to a greater extent (B=-11.7 vs. 9.5).

Discussion

This study applied the decision tree technique to establish a multilevel data-driven model that predicts adults' PA behavior, defined as active or inactive based on their machine-learned activity profiles, and to methodologically identify PA behavior correlates. From the 168 factors of different domains used as input variables to create the decision tree model, the final model selected 36 factors from which 54 different participant subgroups with different variations in SED, LPA, and MVPA were formed. The largest subgroup of inactive participants included those with the highest body fat percentage, who were frequently engaged in physically demanding activities through gardening, but who had rather slow heart rate recovery. The largest subgroup of active participants included those with the lowest body fat percentage in the study population with a relatively fast heart rate recovery. The factors that emerged from the decision tree model, such as body fat percentage, normalized heart rate recovery 60 seconds after exercise, urban–rural areas, average weekday total sitting time, and extravagance score, were associated with SED, LPA, and/or MVPA time. Thus, the present results may inform both multilevel intervention allocation and design.

Consistent with the results of studies focusing on understanding the causation of PA behaviors (5,8,13,44), the established model in the present study indicates that PA behavior is explained by a multilevel hierarchy composed of various factors in different domains. However, our results extend this finding by indicating that PA behavior predictors for different subgroups are different and come from various domains. In addition, our model was driven by empirical data consisting of a range of factors. Studies have generally conceptualized the influence of PA behaviors by theoretically combining common sense and well-established evidence, therefore primarily providing a broad view of PA behavior and its causation for general populations (5,8,44). While previous multilevel models have succeeded in hypothesizing the interaction among factors of different domains, their practical implications have remained limited (8) partially because of their theoretical nature. There were two studies that applied a data-driven approach to establish a decision tree–based model but with self-reported PA measure and a limited number of factors, and one of them used only demographical factors (17) while the other used only sociodemographic factors (16). Overall, the multilevel model presented here specifies the PA behavior correlates at different levels in each subgroup and may be utilized to target and tailor interventions.

Most emerged factors in the decision tree model have been recognized as factors associated with PA behavior in past works, such as education level, profession, overall health status, fitness status, and population density (5,6,11). However, there were also some factor in decision tree model were less established, such as those that were related to personality and temperament including extravagance, impulsiveness, and explorative excitability (6). Such factors (or factors similar to them) were assessed in a few studies but, mostly due to the limited or sometimes contradictory evidence, had not yet been identified as correlates nor been rejected. The other factors that can also be categorized as less established factors are body composition measures (i.e., lean body mass and skeletal muscle mass) and a few of the psychological and environmental factors (e.g., enjoyment of daily activities and number of road accidents) (6–8). A few measures related to heart rate recovery were also emerged in the decision tree model. Even though the association of PA with heart rate recovery measures have been well-studied

(45), they can be considered as novel factors associated with PA behavior that are identified in the present study because our results indicate the existence of another direction of relationship that has not been previously examined.

The less established and previously undiscovered factors found here may be candidates for the next generation of correlates (5). These factors have likely remained underreported (or unexamined) because of the subjective tendency in the existing literature toward examining only those factors for which evidence of significant associations (positive or negative) with different PA behavior indices has been well-established (11). It is important to consider that these factors were selected by the decision tree to create the final model from a wide list of input (independent) variables. This suggests that the less established and novel factors that emerged in the decision tree model might be relatively more important correlates and likely surrogates for the other previously less established or well-established factors that the decision tree excluded in creating the model, such as behavioral attributes (e.g., alcohol, smoking, etc.) or socioeconomic status (6). Nevertheless, one must infer the relative importance of the emergent factors with caution. The study's participants had a narrow age range (46–48 years), which might explain why some of the well-known PA behavior correlates, including age and gender, did not appear in the final model (5,6,11,46). This result agrees with the findings of a previous review, speculating that in studies including both men and women with sufficient age diversity, age was found to be inversely associated with PA participation, and significant differences in PA participation existed between men and women (higher in men) (11).

As far as we know, our study is the first to use machine-learned activity profiles to define the PA behavior of participants. Previous studies have generally examined the associations between different factors and unidimensional indices, typically including the daily amount of SED, LPA, and/or MVPA (16,17,47,48). However, recent evidence from time-use epidemiological studies and beyond suggests that these three activities are interrelated (10,37–39), and should all be considered when studying individuals' PA behavior (37–39). Although due to the methodological constraints (i.e., outcome variable imbalance and overlap) we merged all the active participants in one class to form a near-balance and non-overlapping outcome variable (40,41), it was apparent that the accumulation patterns of SED, LPA, and/or MVPA were varied in the 54 decision-tree formed subgroups and across different active and inactive subgroups. This is indicating that all the three activity intensities along with their interrelationships were considered in our definition of active and inactive individuals, which were based on machine-learned activity profiles (20). Hence, our multidimensional definition of PA behavior might limit the comparability of our results with those of other studies with unidimensional criteria for defining PA behaviors.

Body fat percentage, a direct measure of adiposity, was the most primary discriminator in the decision tree model. Even though it is typically assumed that PA impacts adiposity-related measures, this result is consistent with the findings of a previous systematic review suggesting a possible bidirectional relationship between adiposity and PA behavior (5). A number of other factors for which the other direction of relationship is generally assumed were also seen in the other layers of the final model including muscle strength and heart rate recovery measures. Of note is the prognostic value of most of

these factors for several chronic health conditions. For example, attenuated heart rate recovery is associated with an increased risk of diabetes (49), or can even indicate the presence of coronary artery disease (50). Chronic health conditions have been identified both as a barrier and as motivations towards PA in different populations (51). Even though the self-reported measures addressed the prevalence of diagnosed diseases (e.g., having diabetes, hypertension, etc.), these direct measures were eliminated from the list of input variables due to the high number of missing values. Besides, the study's participants did not consist of only healthy individuals. As a result, the factors with prognostic value of chronic diseases found in our model may be acting as partial surrogates for chronic health conditions/risks and their effects on different PA behaviors.

We also performed association analysis between all the emerged factors in the decision tree model and three PA metrics. Almost all the emerged factors in the decision tree model were significantly associated with SED, LPA, and/or MVPA. The results of association analyses were, at least for the well-established factors, in line with previous studies. For instance, a better health-related quality of life score was associated with lower levels of SED (52), and higher levels of LPA and MVPA (6). The results of association analyses also indicated the relative importance of the identified factors, supporting that our results can be used to highlight the factors associating with PA behavior in terms of priority.

The main strength of the present study is the inclusion of a wide list of factors rather than a few subjectively selected factors (5,11), which resulted in the discovery of the novel predictors. The use of objective measurement of daily PA is also a strength. Previous studies have typically used self-reported PA measures that are known to be imprecise and biased (53). Another strength is the discrimination of PA behaviors based on activity profiles built using the whole activity intensity spectrum over the course of one full week (10). However, the binary categorization of participants (active or inactive) might be a limitation. We used a binary outcome variable because the model's prediction accuracy degraded significantly when the number of PA behavior categories increased (for example, to Inactive, Moderately active and Evening active, and Very active), mostly because of the misclassification between the active categories (results not presented). This is not surprising because despite the different temporal pattern of activities in the active profiles, the overall activity levels were comparable (overlap problem) and the outcome variable was imbalanced. Although it was practically possible to reduce the dimension or select the relevant features in prior to decision tree induction (54), or use more complicated learning algorithms (such as ensemble methods) to achieve a better performance with higher number of categories in the outcome variable (33), these could have caused loss of key mechanistic information and obscured the interpretability of the final model (33,40,54), limiting the recognition of novel correlates categories that were identified here. Another limitation is the cross-sectional study design, which prevents any causal effects to be analyzed. Also, although more than 85% of the original cohort members were alive in Finland during the latest follow-up, less than 40% participated and provided valid accelerometer data—possibly those who were healthier and more active. This might induce selection bias and limit the generalizability of the results. Additionally, the study sample was homogenous in terms of age and ethnicity, and some of the emergent factors in the final model were related to cultural and health

behaviors. These might also limit the generalizability of the results, especially to more diverse populations with different cultural and health behaviors.

Conclusion

Using a data mining approach, we established a multilevel model that predicts PA behavior from empirical and large-scale data. The model consisted of 36 different factors of relative importance from different domains and may be used to target and tailor interventions. The factors emerging from the decision tree model such as body fat percentage, normalized heart rate recovery 60 seconds after exercise, urban-rural areas, average weekday total sitting time, and extravagance score were associated with SED, LPA, and/or MVPA time. The extensive set of factors that was methodologically discovered can be a basis for additional hypothesis testing in PA correlates research. Finally, data mining appeared to be a feasible approach and complex enough to identify different factors along with their interdependencies in explaining PA behavior.

List Of Abbreviations

BMI: Body mass index; CHAID: Chi-squared Automatic Interaction Detection; CI: Confidence interval; LPA: Light physical activity; MET: Metabolic equivalent; MVPA: Moderate-to-vigorous physical activity; NFBC1966: Northern Finland Birth Cohort 1966 study; PA: Physical activity; RI: Response index; SED: Sedentary; SD: Standard deviation; TCI: Temperament and Character Inventory;

Declarations

Ethics approval and consent to participate

The study was carried out in conformance with the declaration of Helsinki. It followed the legislation, decrees and ethical principles concerning medical research on humans in Finland. The Data Protection Ombudsman of Finland has reviewed the NFBC study, the ethical committee of Northern Ostrobothnia Hospital District has approved the study and the permission from the Finnish Ministry of Social Affairs and Health was obtained for the use of register data and patient records. All participants have been willing to participate in the study and have granted permission for the use of their information in unidentifiable format for the purposes of scientific research.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed in the present study are available from the NFBC Project Centre repository upon request, https://www.oulu.fi/nfbc/node/47960.

Competing interests

The authors declare that they have no competing interests.

Funding

NFBC1966 received financial support from University of Oulu Grant no. 24000692, Oulu University Hospital Grant no. 24301140, ERDF European Regional Development Fund Grant no. 539/2010 A31592. This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska–Curie grant agreement no. 713645, the Ministry of Education and Culture in Finland [grant numbers OKM/86/626/2014, OKM/43/626/2015, OKM/17/626/2016, OKM/54/626/2019], Infotech Oulu, Finland, and Northern Ostrobothnia Hospital District.

Authors' contributions

VF, MKa, RK and TJ designed the study. VF performed data mining, analyzed and interpreted the data and prepared the first version of the manuscript. MN contributed in data cleaning and measurement of physical activity-related variables from the activity monitors. MKä and SP contributed in measurement and cleaning of environmental variables. All authors contributed to revising and finishing the manuscript and read and approved the final manuscript.

Acknowledgments

We gratefully thank all cohort members and researchers who participated in the 46 yrs study. We also acknowledge the work of the NFBC project center.

References

1. Warburton DER, Bredin SSD. Health benefits of physical activity: a systematic review of current systematic reviews. Curr Opin Cardiol. 2017;32(5):541–56.
2. Lee I-M, Shiroma EJ, Lobelo F, Puska P, Blair SN, Katzmarzyk PT, et al. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. Lancet. 2012;380(9838):219–29.
3. Guthold R, Stevens GA, Riley LM, Bull FC. Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants. Lancet Glob Heal. 2018;6(10):e1077–86.
4. Kohl 3rd HW, Craig CL, Lambert EV, Inoue S, Alkandari JR, Leetongin G, et al. The pandemic of physical inactivity: global action for public health. Lancet. 2012;380(9838):294–305.
5. Bauman AE, Reis RS, Sallis JF, Wells JC, Loos RJF, Martin BW, et al. Correlates of physical activity: why are some people physically active and others not? Lancet. 2012;380(9838):258–71.
6. Choi J, Lee M, Lee J, Kang D, Choi J-Y. Correlates associated with participation in physical activity among adults: a systematic review of reviews and update. BMC Public Health. 2017;17(356).

7. O'donoghue G, Perchoux C, Mensah K, Lakerveld J, Van Der Ploeg H, Bernaards C, et al. A systematic review of correlates of sedentary behaviour in adults aged 18-65 years: a socio-ecological approach. BMC Public Health. 2016;16(163).
8. Chastin SFM, De Craemer M, Lien N, Bernaards C, Buck C, Oppert J-M, et al. The SOS-framework (Systems of Sedentary behaviours): an international transdisciplinary consensus framework for the study of determinants, research priorities and policy on sedentary behaviour across the life course: a DEDIPAC-study. Int J Behav Nutr Phys Act. 2016;13(83).
9. Pate RR, Berrigan D, Buchner DM, Carlson SA, Dunton G, Fulton JE, et al. Actions to improve physical activity surveillance in the United States. In: NAM Perspectives. Discussion Paper, National Academy of Medicine, Washington, DC; 2018.
10. Silva KS, Garcia LMT, Rabacow FM, de Rezende LFM, de Sá TH. Physical activity as part of daily living: Moving beyond quantitative recommendations. Prev Med. 2017;96:160–2.
11. Trost SG, Owen N, Bauman AE, Sallis JF, Brown W. Correlates of adults' participation in physical activity: review and update. Med Sci Sport Exerc. 2002;34(12):1996–2001.
12. Venkatasubramaniam A, Wolfson J, Mitchell N, Barnes T, JaKa M, French S. Decision trees in epidemiological research. Emerg Themes Epidemiol. 2017;14(11).
13. Sallis JF, Cervero RB, Ascher W, Henderson KA, Kraft MK, Kerr J. An ecological approach to creating active living communities. Annu Rev Public Heal. 2006;27:297–322.
14. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform. 2008;77(2):81–97.
15. Bellazzi R, Diomidous M, Sarkar IN, Takabayashi K, Ziegler A, McCray AT. Data analysis and data mining: current issues in biomedical informatics. Methods Inf Med. 2011;50(6):536–44.
16. Lakerveld J, Loyen A, Schotman N, Peeters CFW, Cardon G, van der Ploeg HP, et al. Sitting too much: a hierarchy of socio-demographic correlates. Prev Med. 2017;101:77–83.
17. Patterson F, Lozano A, Huang L, Perkett M, Beeson J, Hanlon A. Towards a demographic risk profile for sedentary behaviours in middle-aged British adults: a cross-sectional population study. BMJ Open. 2018;8(7):e019639.
18. Yoon S, Suero-Tejeda N, Bakken S. A data mining approach for examining predictors of physical activity among urban older adults. J Gerontol Nurs. 2015;41(17):14–20.
19. Buck C, Loyen A, Foraita R, Van Cauwenberg J, De Craemer M, Mac Donncha C, et al. Factors influencing sedentary behaviour: A system based analysis using Bayesian networks within DEDIPAC. PLoS One. 2019;14:e0211546.
20. Niemelä M, Kangas M, Farrahi V, Kiviniemi A, Leinonen A-M, Ahola R, et al. Intensity and temporal patterns of physical activity and cardiovascular disease risk in midlife. Prev Med (Baltim). 2019;124:33–41.
21. University of Oulu Web site [Internet]. NFBC 1966 data collection. [cited 2020 Feb 10]. Available from: https://www.oulu.fi/nfbc/node/19663

22. Sintonen H. The 15D instrument of health-related quality of life: properties and applications. Ann Med. 2001;33(5):328–36.
23. Cloninger CR, Przybeck TR, Svrakic DM, Wetzel RD. The Temperament and Character Inventory (TCI): A guide to its development and use. Center for Psychobiology of Personality, Washington University; 1994.
24. University of Oulu Web site [Internet]. 46-year follow-up study. [cited 2020 Feb 10]. Available from: https://www.oulu.fi/nfbc/node/26627
25. Kiviniemi AM, Perkiömäki N, Auvinen J, Niemelä M, Tammelin T, Puukka K, et al. Fitness, Fatness, Physical Activity, and Autonomic Function in Midlife. Med Sci Sport Exerc. 2017;49(12):2459–68.
26. University of Oulu Web site [Internet]. 46-year follow-up study, Clinical examination. Available from: https://www.oulu.fi/nfbc/node/30371
27. Kinnunen H, Häkkinen K, Schumann M, Karavirta L, Westerterp KR, Kyröläinen H. Training-induced changes in daily energy expenditure: Methodological evaluation using wrist-worn accelerometer, heart rate monitor, and doubly labeled water technique. PLoS One. 2019;14(7):e0219563.
28. Finnish Community Structure data base: Statistics Finland [Internet]. Grid Database. Available from: http://www.stat.fi/tup/ruututietokanta/index_en.html.
29. Street network data: Finnish Transport Agency [Internet]. Digiroad - National Road and Street Database. Available from: https://www.liikennevirasto.fi/web/en/open-data/digiroad
30. Sport facilities: University of Jyväskylä [Internet]. Finnish database of sport facilities. Available from: https://www.lipas.fi/etusivu
31. Finnish Environment Institute [Internet]. Corine land cover. Available from: https://www.syke.fi/openinformation
32. Kingsford C, Salzberg SL. What are decision trees? Nat Biotechnol. 2008;26:1011–3.
33. Loh W-Y. Fifty years of classification and regression trees. Int Stat Rev. 2014;82(3):329–48.
34. Kass G V. An exploratory technique for investigating large quantities of categorical data. J R Stat Soc Ser C, R Stat Soc. 1980;29(2):119–27.
35. Murphy EL, Comiskey CM. Using chi-Squared Automatic Interaction Detection (CHAID) modelling to identify groups of methadone treatment clients experiencing significantly poorer treatment outcomes. J Subst Abuse Treat. 2013;45(4):343–9.
36. Rodrı´guez AH, Avile´s-Jurado FX, Dı´az E, Schuetz P, Trefler SI, Solé-Violán J, et al. Procalcitonin (PCT) levels for ruling-out bacterial coinfection in ICU patients with influenza: A CHAID decision-tree analysis. J Infect. 2016;72(2):143–51.
37. Stamatakis E, Ekelund U, Ding D, Hamer M, Bauman AE, Lee I-M. Is the time right for quantitative public health guidelines on sitting? A narrative review of sedentary behaviour research paradigms and findings. Br J Sport Med. 2019;53(6):377–82.
38. Chastin SFM, Palarea-Albaladejo J, Dontje ML, Skelton DA. Combined effects of time spent in physical activity, sedentary behaviors and sleep on obesity and cardio-metabolic health markers: a

novel compositional data analysis approach. PLoS One. 2015;10(10):e0139984.

39. Rosenberger ME, Fulton JE, Buman MP, Troiano RP, Grandner MA, Buchner DM, et al. The 24-Hour Activity Cycle: A New Paradigm for Physical Activity. Med Sci Sports Exerc. 2019;51(3):454–64.
40. Ali A, Shamsuddin SM, Ralescu AL, others. Classification with class imbalance problem: a review. Int J Adv Soft Compu Appl. 2015;7(3):176–204.
41. Weiss GM, Provost F. Learning when training data are costly: The effect of class distribution on tree induction. J Artif Intell Res. 2003;19:315–54.
42. Zhang S, Qin Z, Ling CX, Sheng S. " Missing is useful": missing values in cost-sensitive decision trees. IEEE Trans Knowl Data Eng. 2005;17(12):1689–93.
43. Jauho A-M, Pyky R, Ahola R, Kangas M, Virtanen P, Korpelainen R, et al. Effect of wrist-worn activity monitor feedback on physical activity behavior: A randomized controlled trial in Finnish young men. Prev Med reports. 2015;2:628–34.
44. Garcia LMT, Roux AVD, Martins ACR, Yang Y, Florindo AA. Development of a dynamic framework to explain population patterns of leisure-time physical activity through agent-based modeling. Int J Behav Nutr Phys Act. 2017;14(111).
45. Carnethon MR, Jacobs JDR, Sidney S, Sternfeld B, Gidding SS, Shoushtari C, et al. A longitudinal study of physical activity and heart rate recovery: CARDIA, 1987-1993. Med Sci Sports Exerc. 2005;37(4):606–12.
46. Molanorouzi K, Khoo S, Morris T. Motives for adult participation in physical activity: type of activity, age, and gender. BMC Public Health. 2015;15(66).
47. Van Dyck D, Cardon G, Deforche B, Giles-Corti B, Sallis JF, Owen N, et al. Environmental and psychosocial correlates of accelerometer-assessed and self-reported physical activity in Belgian adults. Int J Behav Med. 2011;18(3):235–45.
48. Gonçalves PB, Hallal PC, Hino AAF, Reis RS. Individual and environmental correlates of objectively measured physical activity and sedentary time in adults from Curitiba, Brazil. Int J Public Health. 2017;62(7):831–40.
49. Qiu SH, Xue C, Sun ZL, Steinacker JM, Zügel M, Schumann U. Attenuated heart rate recovery predicts risk of incident diabetes: insights from a meta-analysis. Diabet Med. 2017;34(12):1676–83.
50. Akyüz A, Alpsoy Ş, Akkoyun DÇ, Değirmenci H, Güler N. Heart rate recovery may predict the presence of coronary artery disease. Anatol J Cardiol Kardiyol Derg. 2014;14(4):351–6.
51. Costello E, Kafchinski M, Vrazel J, Sullivan P. Motivators, barriers, and beliefs regarding physical activity in an older adult population. J Geriatr Phys Ther. 2011;34(3):138–47.
52. Rhodes RE, Mark RS, Temmel CP. Adult sedentary behavior: a systematic review. Am J Prev Med. 2012;42(3):e3–28.
53. Steene-Johannessen J, Anderssen SA, van der Ploeg HP, Hendriksen IJ, Donnelly AE, Brage S, et al. Are Self-report Measures Able to Define Individuals as Physically Active or Inactive? Med Sci Sports Exerc. 2016;48(2):235–44.

54. Lee CH, Yoon H-J. Medical big data: promise and challenges. Kidney Res Clin Pract. 2017;36(1):3–11.

Tables

Table 1. The characteristics of the study participants

	Inactive profile (n=1,881)	Active profile (n=2,701)	Total population (n=4,582)
Height, cm (SD)	168.4 (8.9)	172 (9.0)	170.5 (9.1)
Weight, kg (SD)	77.3 (17.1)	78.6 (16.2)	78.1 (16.6)
Body mass index, kg/m^2 (SD)	27.2 (5.2)	26.4 (4.5)	26.7 (4.8)
Body fat, % (SD)	31.2 (9.3)	27.7 (8.9)	28.9 (9.2)
Alcohol consumption, grams/day (SD)	9.8 (4.1)	10.4 (16.8)	10.1 (4.3)
Gender			
Men	648 (34)	1268 (47)	1916 (42)
Women	1233 (66)	1433 (53)	2666 (58)
Education			
No professional education	53 (3)	85 (3)	138 (3)
Vocational/college level education	1119 (60)	1765 (65)	2884 (67)
Polytechnic/university degree	573 (30)	670 (25)	1243 (29)
Employment status			
Employed	1499 (78)	2265 (84)	3764 (88)
Student	38 (2)	32 (1)	70 (2)
Unemployed	117 (6)	101 (4)	218 (5)
Other	95 (5)	88 (3)	183 (4)
Marital status			
Married/cohabiting	1421 (75)	2072 (77)	3493 (86)
Divorced	182 (10)	229 (8)	411 (10)
Unmarried	192 (11)	272 (10)	464 (11)
Widowed	11 (0.5)	6 (0.2)	17 (0.4)
Smoking			
Non-smoker	990 (55)	1413 (55)	2403 (55)
Current smoker	335 (19)	447 (18)	782 (18)
Former smoker	458 (26)	702 (27)	1160 (27)

SED, min/day (SD)	675.1 (78.9)	608.2 (91.7)	635.6 (92.7)
LPA, min/day (SD)	242.4 (59.3)	309.8 (71)	282.1 (74.3)
MVPA, min/day (SD)	48.3 (20.6)	83.8 (36.6)	69.2 (33.6)

Values are numbers (%) if not otherwise stated. SD = standard deviation, SED = sedentary, LPA = light physical activity, MVPA = moderate-to-vigorous physical activity.

Figures

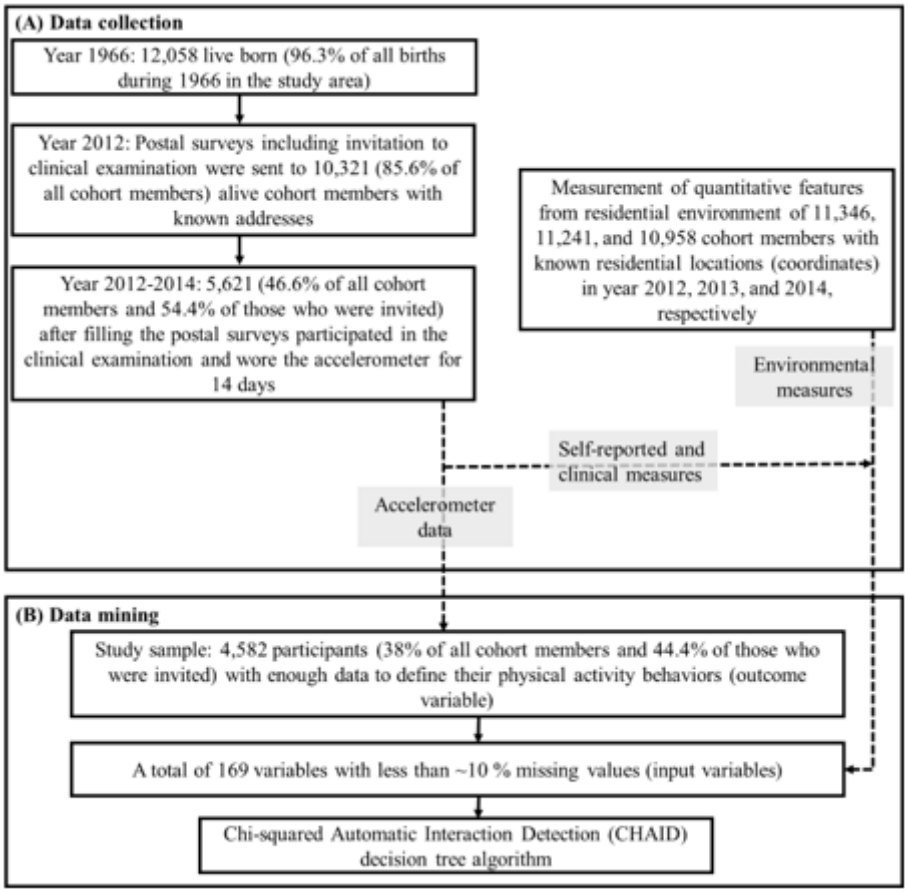

Figure 1

The collected data in the latest follow–up of Northern Finland Birth Cohort 1966 (A), and the selection of study population, input variables, and outcome variables for data mining in the present study (B).

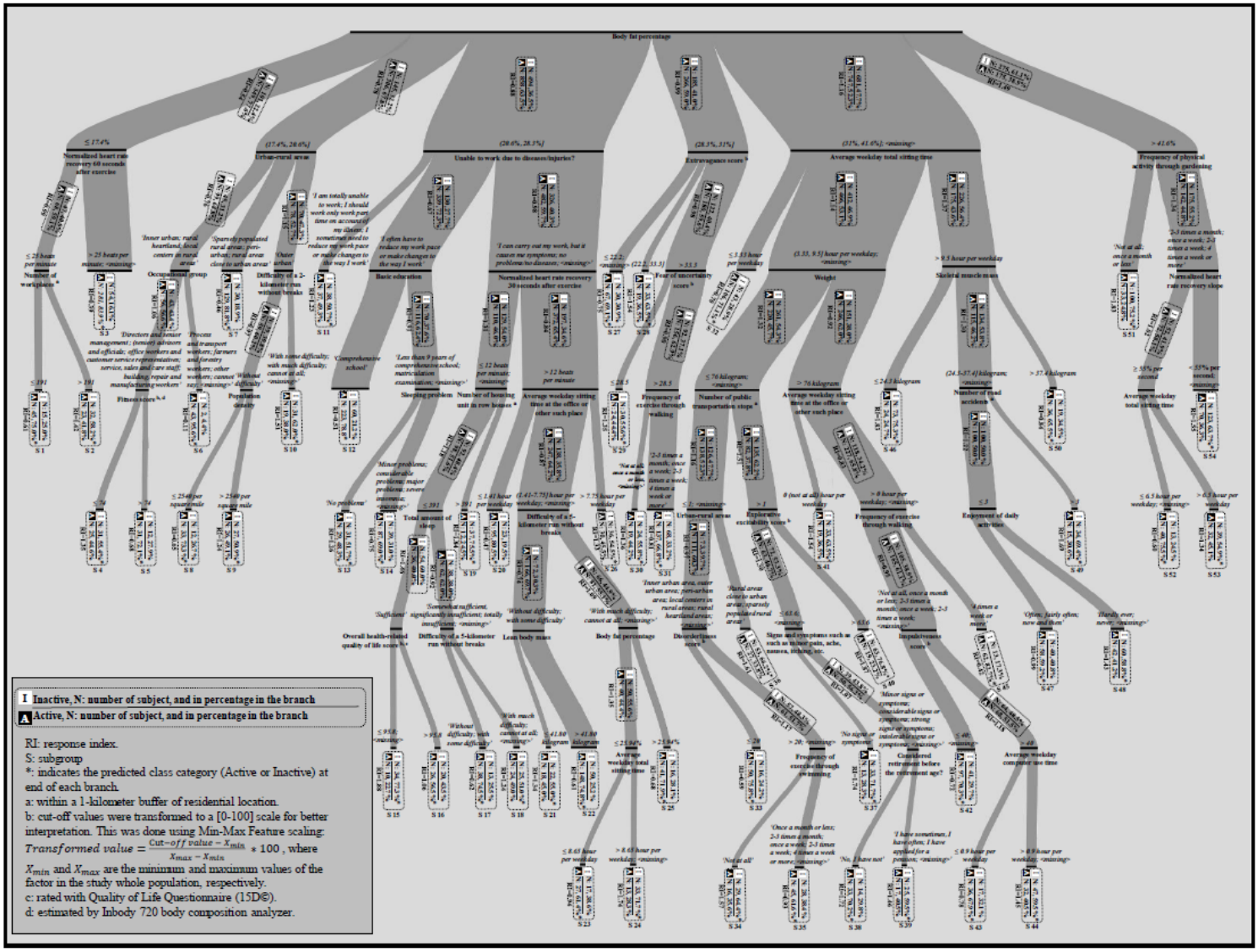

Figure 2

The Chi Squared Automatic Interaction Detection tree illustrating the hierarchy of the factors predicting Active and Inactive participants. The thickness of branches is based on the number of participants in the branch. Categories (for categorical and ordinal variables) and cut-off values (for continuous variables) are shown in italicized text, and the variables in normal text. In interval notations between brackets, inclusiveness and exclusiveness are shown with squared and round brackets, respectively.

Figure 3

Z-scores of sedentary (SED), light physical activity (LPA), and moderate-to-vigorous physical activity (MVPA) in the 54 decision tree-formed subgroups of participants. S = Subgroup.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FarrahiSupplementaryMaterial1.docx