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

CURRENT STATUS: UNDER REVIEW

International Journal of Behavioral Nutrition and Physical Activity BMC

Vahid Farrahi
Oulun Yliopisto

Vahid.farrahi@oulu.fi*Corresponding Author*
ORCiD: 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

DOI:
10.21203/rs.2.23726/v1

SUBJECT AREAS
Physical Medicine & Rehab *Nutrition & Dietetics*

KEYWORDS
Decision tree, CHAID, Multilevel model, Prediction, Classification

Abstract

Purpose

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

Methods

The 46-year follow-up data from the population-based Northern Finland Birth Cohort 1966 were used to create a hierarchy using Chi-square Automatic Interaction Detection (CHAID) decision tree technique for predicting PA behavior. The study's subjects were classified as physically active or physically inactive based on their activity profiles derived from objective measurement of PA. The variables were a wide list of potentially modifiable factors including self-reported, clinical, and environmental measures. We then analyzed the association of the factors emerging from the model with three PA metrics including sedentary (SED), light PA (LPA), and moderate-to-vigorous PA (MVPA) minutes per day.

Results

Model fitting was performed using a total of 168 factors as input variables to classify the PA behavior of 2,701 physically active and 1,881 physically inactive subjects. The decision tree selected a total of 36 factors of different domains by which 54 subgroups of subjects were formed. Factors emerging from the model were associated with the PA metrics, 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, a data-driven model was established from empirical data that can be potentially utilized to identify subgroups for multilevel intervention allocation. An extensive set of factors was methodologically discovered that can be a basis for additional hypothesis testing in PA correlates research.

Introduction

The positive relationship between physical activity (PA) and health is well established (1), yet the level of PA in many populations is low. A recent study with pooled data from more than 350 population-based surveys reported that more than one-fourth of adults (27.5% or 1.4 billion people) globally have insufficient PA (2). Physical inactivity has a negative impact on health and is associated with increased risk of developing several diseases (3). There is now a clear need to resolve the physical inactivity crisis and reverse its high prevalence (2–4).

Numerous studies have investigated the association between various factors of different domains and different indices of PA behavior (5,6). For instance, there is a consistent body of evidence suggesting that better health status or perceived fitness is associated with higher level of PA (5,6). Despite the existence of a large number of studies on correlates of PA, most PA promotion strategies have remained unsuccessful (4). This may be because previous studies have often neglected to account for the complexity and multimodality of PA behavior (4,6–8), and also the existence of several levels of influence in their analytical approaches (6,9).

Previous works have mostly used traditional statistical approaches, such as regression analyses, to examine the relationships between covariates and PA behavior metrics. This has led previous studies to remain restricted by data analysts' decisions about how association and interaction are hypothesized (knowledge-driven) (10), mainly because the selected factors for inclusion in regression analyses are chosen subjectively according to their conceptual relevancy and, in some cases, initial empirical associations (10,11). As a result, a limited number of correlates has been repeatedly reported in the existing literature (6,10). To complement regression-based studies, ecological approaches have been used to conceptualize the predictors and their interdependencies in explaining PA behavior such as the interrelations between individuals and their social and physical environments (12). Ecological approaches are knowledge-driven (5) and generally rely on well-established correlates (5,9), while there could potentially exist some missing correlates due to the aforementioned limitations of traditional statistical approaches (11). For further advancement in correlates research, there has been requests in the existing literature to increase the complexity of statistical assessment to multilevel and more sophisticated approaches to both capture and

understand different levels of influences on PA behavior (4,6,8).

We have now entered a data-intensive era, where the popularity of data mining approaches is increasing (13). Data mining approaches originated from statistics and are known to be more powerful techniques compared to traditional statistical methods (13,14). They have the ability to capture hidden and novel insights buried in large amounts of data and generate data-driven hypotheses (13,14). These principles are consistent with the field of PA research, in which there is a need for more complex approaches to identify correlates of PA behaviors, understand their relative importance, and capture the complex interrelations between the factors at different levels (5,6). The present study applies a predictive data mining approach on a variety of factors of different domains to build a data-driven hierarchy predicting PA behavior. Our primary purpose was twofold, (a) to establish a multilevel data-driven model explaining PA behavior and (b) to methodologically identify the associated factors (correlates) of PA behavior from a wide list of potentially modifiable factors.

Materials And Methods

Study population

This study was performed with data from the population-based Northern Finland Birth Cohort 1966 study (NFBC1966). NFBC1966 is a lifelong study involving subjects born between January 1 and December 31, 1966, in Finland's two northernmost provinces, Oulu and Lapland (n = 12,058 live births). The cohort members have been monitored on a regular basis prospectively since birth using different measurement techniques and tools including health care records, questionnaires, and clinical examinations (15).

For the present study, cross-sectional data collected in the most recent follow-up (at the age of 46 years), in which participants underwent a wide series of measurements, were utilized. With respect to the measurement tools/techniques, the measured factors can be categorized into four groups: self-reported measures, clinical measures, objective built and natural environmental measures, and objective physical activity measures (Fig. 1).

Questionnaires and measurements

Questionnaires

A postal questionnaire was sent to all living cohort members with known addresses in 2012–2014 (n

= 10,321). The questionnaire included items on social background, participation in light and brisk PA, 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 (16). Another additional separate survey was used to address opinions and experiences, covering questions from the Temperament and Character Inventory (TCI) questionnaire (17). 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 (18).

Clinical examination

Participants were invited to attend a clinical examination and a total of 5,861 subjects participated. 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 (19,20).

Objective measurement of physical activity

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. The accelerometers were blinded, not showing any feedback to the users. Polar Active has a uniaxial accelerometer that outputs estimated energy expenditure in metabolic equivalent (MET) values every 30 seconds (21).

Environmental measures

Geographic Information System (ArcGIS 10.3) was used to measure quantitative built and natural environment features related to community structure, road network, amenities and socioeconomic

factors that describe the conduciveness of participants' residential environment for PA. We obtained residential coordinates of all participants whose places of residence were available from the Finnish Population Register Centre at the time of the 46-year follow-up data collection. Quantitative environmental features were calculated using a one-kilometer circular buffer around residential location and distances to amenities were measured using road network data.

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 top-down approach and recursive partitioning are used to select and split input variables in such a way that the subgroups differ significantly with respect to a designated criterion. The final 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 (22). 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 (11,13,22).

We used the Chi-squared Automatic Interaction Detection (CHAID) decision tree algorithm to create the model (23). CHAID has been repeatedly used in studies with clinical applications whose main purpose was to identify key factors related to the outcomes of interest (24,25). 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 categorical, ordinal, and continuous predictors.

Decision tree model construction and validation

Outcome variable

The outcome variable for creating the decision tree model was a binary categorical variable. It was formed by categorizing the subjects into physically active or physically inactive. These two PA behaviors (categories) were based on activity profiles derived from a range of objectively measured activity intensities, measured over the course of one full week including sedentary (SED), light PA (LPA), and moderate-to-vigorous PA (MVPA). The activity profiles were established using a clustering approach in our previous study (26). Briefly, X-means clustering algorithm was applied on accelerometer-based MET-level data of subjects who had seven consecutive valid measurement days (N = 4,582), and four distinct activity profiles (clusters) were derived. 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 both weekdays and weekends. The activity profiles were named with respect to the temporal and intensity patterns of subjects' daily activities in each cluster: Inactive, Moderately active, Evening active, and Very active. For the purpose of the present study, we used these activity profiles to assign the subjects to either physically active or inactive category. The subjects in the Moderately active, Evening active, or Very active clusters were defined as physically active, and those in the Inactive cluster were defined as physically inactive.

Input variables

The input variables were the self-reported, clinical, and objective environmental measures, with exclusion of those that had more than ~ 10% missing values or were non-modifiable. Similar to a previous study on the correlate research priorities (9), the decision for inclusion of only potentially modifiable factors was made to enable intervention development in future studies. To select the factors, we asked domain experts in each domain to select the potentially modifiable factors.

The input variables were used in their original form to classify the study's subjects into physically active and physically inactive. To create a decision tree model using CHAID algorithm, no pre-assumption is needed to be made for the input variables (25). This means that there was no need to create a cut-off point for continuous input variables or those that have different cut-offs for men and

women (e.g., waist circumference). There was also no need to combine response categories of categorical or ordinal.

Dealing with missing values

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 (24). 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%), and therefore it has been concluded that in some scenarios where missing value imputation is not meaningful or feasible, there is no need for imputation (27).

Algorithm parameters

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 (13). The pruning criteria were set such that groups smaller than 80 were not split any further (maximum number of subjects in a parent node), and no group smaller than 40 was formed (maximum number of subjects 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.

Validation

To create and validate the model, 10-fold cross-validation approach was used. To evaluate the classification and misclassification rate of the decision tree model, the confusion matrix for the two outcomes was used to show the proportion of subjects with each of the outcome variable that were correctly and incorrectly classified.

Association analysis

We performed bivariate association analysis between each emerged factor in the decision tree and three objectively measured physical activity metrics with the whole study population. The physical activity metrics were average daily time (minutes per day) in SED, LPA, and MVPA over the course of the seven consecutive valid measurement days used for defining PA profiles in our previous study (26). Association analysis was done to assure the importance of identified data-driven factors in

explaining different levels of PA intensity. The PA metrics were calculated using previously validated cut-points (SED, 1–1.99 MET; LPA, 2–3.49 MET; and MVPA, ≥ 3.5 MET) by the accelerometer manufacturer (28).

Generalized linear mixed models were used to test the associations. Each emerged factor in the decision tree was included in a separate model with each outcome variable. This shows the association between each factor and PA metrics independent of the other variables. For all the models, age and gender were entered as covariates, and urban-rural area as a random effect (different areas may have different facilities and structures, or even polices that affect PA behavior). Continuous independent variables were standardized to have a mean of zero and standard deviation (SD) of 1, so the coefficients (B) with 95% confidence intervals (CI) of different models with similar outcome could be compared and interpreted as change in PA metric for every 1 SD change in the predictor. 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

Data

In total, 4,582 subjects had enough valid PA data to be included in the cluster analysis study (26), and accordingly, the information on the outcome value (a physically active or inactive profile) was available for the present study. The numbers of subjects with physically active and physically inactive profiles were 1,881 (42%) and 2,701 (58%), respectively. The characteristics of the study's subjects for the whole sample, with respect to the two outcome variables, are shown in Table 1. A total of 168 factors were used as input variables after eliminating non-modifiable factors and factors with over ~ 10% missing values. Of these, 82 were continuous, 19 were categorical, and 67 were ordinal factors. All the 168 input variables are given in the Supplementary File 1.

Table 1

The characteristics of the study subjects

	Physically active profile (n = 2701)	Physically inactive profile (n = 1881)	Total population (n = 4582)
Height, cm (SD)	172 (9.0)	168.4 (8.9)	170.5 (9.1)
Weight, kg (SD)	78.6 (16.2)	77.3 (17.1)	78.1 (16.6)
Body mass index, kg/m^2 (SD)	26.4 (4.5)	27.2 (5.2)	26.7 (4.8)
Gender			
Men	1268 (47)	648 (34)	1916 (42)
Women	1433 (53)	1233 (66)	2666 (58)
Education			
No professional education	85 (3)	53 (3)	138 (3)
Vocational/college level education	1765 (65)	1119 (60)	2884 (67)
Polytechnic/university degree	670 (25)	573 (30)	1243 (29)
Employment status			
Employed	2265 (84)	1499 (78)	3764 (88)
Student	32 (1)	38 (2)	70 (2)
Unemployed	101 (4)	117 (6)	218 (5)
Other	88 (3)	95 (5)	183 (4)
Marital status			
Married/cohabiting	2072 (77)	1421 (75)	3493 (86)
Divorced	229 (8)	182 (10)	411 (10)
Unmarried	272 (10)	192 (11)	464 (11)
Widowed	6 (0.2)	11 (0.5)	17 (0.4)

Values are numbers (%) if not otherwise stated, SD = standard deviation.

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 subjects were formed, 26 predicted as physically active and 28 as physically 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.

Table 2

Confusion matrix showing the performance of model with leave–one–subject–out cross validation

	Predicted outcome		Percent correct
Actual outcome	Physically active, n	Physically inactive, n	
Physically active, n	2014	687	74.6%
Physically inactive, n	705	1176	62.5%

Six main subgroups as divided by the first predictor were formed; subgroup I: body fat percentage ≤ 17.4%, subgroup II: body fat percentage of 17.4–20.6%, subgroup III: body fat percentage of 20.6–

28.3%, subgroup IV: body fat percentage of 28.3–31%, subgroup V: body fat percentage of 31–41.6%, and subgroup VI: body fat percentage > 41.6%. For subgroup I, there were two other predictors including one environmental (number of workplaces) and one fitness measure (heart rate recovery at 60 second after exercise). While heart rate recovery was not a predictor in subgroup II, the predictors included environmental measures (e.g., urban-rural areas), one socioeconomic measure (occupational group), and two physical capacity measures (e.g., difficulty of a 2-kilometer run without breaks). For subgroup III, other predictors were measures of demography (i.e., basic education), heart rate recovery (i.e., heart rate recovery 30 seconds after exercise), physical capacity (e.g., difficulty of a 5-kilometer run without breaks), body compositions (i.e., lean body mass), behavior (sleeping problem), and environment (e.g., number of housing unit in row houses).

In subgroup IV, the other predictors were two measures of temperament and personality (e.g., fear of uncertainty score) and one measure of behavior (frequency of exercise through walking). The environmental (e.g., urban-rural areas), temperament and personality (e.g., explorative excitability score), and psychology (e.g., enjoyment of daily activities) measures appeared together in subgroup V along with some other measures related to behavior (e.g., frequency of exercise through walking, average weekday total sitting time), and physical health and well-being (e.g., signs and symptoms such as pain, ache, nausea, itching, etc.). Finally, the predictors of subgroup VI, who had the highest body fat percentage, were one heart rate recovery measure (normalized heart rate recovery slope) and two behavioral measures (average weekday total sitting time, frequency of exercise through walking).

Association analysis

Table 3 shows the results of association performed with the entire study population between the factors emerged from the decision tree model and three PA metrics including SED, LPA and MVPA time. All the factors except fear of uncertainty and impulsiveness scores were associated with at least one PA metric. Most predictors that appeared in relatively high layers of the decision tree model and larger subgroups were significantly associated with all three PA metrics, including body fat percentage (SED: B = 26.5 (95% confidence interval: 23.5, 29.6), LPA: B = -16.1 (-18.5, -13.6), and

MVPA: B = -11.7 (-12.9, -10.6)), normalized heart rate recovery 60 seconds after exercise (SED: B = -16.1 (-18.1, -13.4), LPA: B = 9.9 (7.7, 12.1), and MVPA: B = 9.6 (8.6, 10.6)), normalized heart rate recovery slope (SED: B = -17.7 (-15, -20.4), LPA: B = 10.6 (12.8, 8.4), and MVPA: B = 9.5 (10.5, 8.4)), occupational group (SED: B = -46.7 (-52, -41.3), LPA: B = 41.1 (36.9, 45.4), and MVPA: B = 3.5 (1.4, 5.6)), and average weekday total sitting time (SED: B = 34.1 (31.5, 36.7), LPA: B = -25.3 (-27.4, -23.3), and MVPA: B = -5.8 (-6.8, -4.7)). The magnitude of associations (i.e., change in minutes per day of PA metrics for every 1 SD change in the predictor) was roughly related to the layer and size of subgroup in which the factor appeared; relatively higher in those that appeared in higher layers and larger subgroups.

Table 3
Associations with the whole study population (N = 4582) between the emerged factors in the decision tree model and time spent on sedenteriness (SED), light physical activity (LPA), and moderate-to-vigorous physical activity (MVPA).

	Missing values (n)	SED (min/day)		LPA (min/day)		MVPA (min/day)	
Factors (correlates) emerged in the decision tree model		B (95% CI)		B (95% CI)	p	B (95% CI)	p
Body fat percentage	75	26.5 (23.5, 29.6)	< 0.001**	-16.1 (-18.5, -13.6)	< 0.001**	-11.7 (-12.9, -10.6)	< 0.001**
Normalized heart rate recovery 60 seconds after exercise	421	-16.1 (-18.1, -13.4)	< 0.001**	9.9 (7.7, 12.1)	< 0.001**	9.6 (8.6, 10.6)	< 0.001**
Urban-rural areas (rural area)	10	-20.9 (-40.3, -1.4)	0.35	21.7 (5.3, 38.2)	0.009**	0.2 (-4.7, 5.1)	0.921
Unable to work due to diseases/injuries (no problems/no diseases; I can carry out my work, but it causes me symptoms)	209	6 (-1, 13.1)	0.093	-1.4 (-7, 4.1)	0.618	4.4 (1.8, 7.1)	0.001**
Extravagance score	336	6.3 (3.5, 9)	< 0.001**	-3.7 (-5.9, -1.50)	0.001**	-0.6 (-1.6, 0.5)	0.273
Average weekday total sitting time	204	34.1 (31.5, 36.7)	< 0.001**	-25.3 (-27.4, -23.3)	< 0.001**	-5.8 (-6.8, -4.7)	< 0.001**
Frequency of exercise through gardening (2–3 times a month or higher)	207	-20.6 (-27.3, -14)	< 0.001**	14.4 (9.1, 19.6)	< 0.001**	1.4 (-1, 3.9)	0.254

Number of workplaces	10	2.9 (-0.1, 6)	0.058	-3.2 (-5.6, -0.7)	0.011*	0.6 (-0.5, 1.7)	0.302
Occupational group (workers and service, sales and care staffs)	261	-46.7 (-52, -41.3)	< 0.001**	41.1 (36.9, 45.4)	< 0.001**	3.5 (1.4, 5.6)	0.001**
Difficulty of a 2-kilometer run without breaks (without difficulty; with some difficulty)	211	-22.9 (-28.7, -17.1)	< 0.001**	8.9 (4.3, 13.5)	< 0.001**	14.2 (12.1, 16.4)	< 0.001**
Basic education (less than 9 years of comprehensive school; comprehensive school)	206	-17.9 (-23.6, -12.2)	< 0.001**	19.9 (15.5, 24.4)	< 0.001**	-2.1 (-4.2, 0.1)	0.051
Normalized heart rate recovery 30 seconds after exercise	399	-16.9 (-19.6, -14.2)	< 0.001**	11 (8.8, 13.2)	< 0.001**	9.1 (8 to 10.1)	< 0.001**
Fear of uncertainty score	336	-1.8 (-4.6, 0.9)	0.198	0.7 (-1.6, 2.9)	0.553	-0.6 (-1.7, 0.4)	0.235
Weight	4	13.3 (10.3, 16.3)	< 0.001**	-8.4 (-10.7, -6)	< 0.001**	-3.4 (-4.5, -2.2)	< 0.001**
Skeletal muscle mass	75	-8.4 (-13.1, -3.8)	< 0.001**	3.5 (-0.1, 7.2)	0.060	9.6 (7.8, 11.3)	< 0.001**
Normalized heart rate recovery slope	425	-17.7 (-15, -20.4)	< 0.001**	10.6 (12.8, 8.4)	< 0.001**	9.5 (10.5, 8.4)	< 0.001**
Fitness score	75	-23.4 (-26.2, -20.7)	< 0.001**	15.1 (12.9, 17.3)	< 0.001**	10.7 (9.7, 11.8)	< 0.001**
Population density	12	7.5 (3.7, 11.4)	< 0.001**	-7 (-10.1, -3.1)	< 0.001**	0.8 (-0.5, 2.2)	0.227
Sleeping problems (no problems; minor problems)	186	-20.6 (-30.3, -10.9)	< 0.001**	14.6 (6.9, 22.2)	< 0.001**	5.3 (7.71, 8.9)	0.004**
Number of housing unit in row houses	12	3.2 (0.2, 6.2)	0.036*	-2.9 (-5.3, -0.5)	0.018*	-0.2 (-1.3, 0.9)	0.703
Average weekday sitting time at the office or other such place	382	27.9 (25.2, 30.6)	< 0.001**	-22.4 (-24.6, -20.3)	< 0.001**	-2.6 (-3.6, -1.5)	< 0.001**
Frequency of exercise through walking (2–3 times a month or more)	195	-16.9 (-24.7, -9.2)	< 0.001**	-3.89 (-2.2, 10)	.214	9.4 (6.5, 12.3)	< 0.001**
Number of public transportation stops	10	4.1 (1.8, 8.1)	0.002**	-3.6 (-6.1, -1)	0.006**	0.3 (-0.8, 1.5)	0.572
Number of	10	5.9 (2.5,	0.001**	-3.8 (-6.6,	0.006**	0.1 (-1.13,	0.876

road accidents		9.4)		-1.1)		1.3)	
Total amount of sleep (Sufficient, somewhat sufficient)	268	-10.1 (-19.9, -0.4)	0.041*	-0.6 (-8.3, 7.2)	0.88	2.5 (-1.1, 6.2)	0.176
Difficulty of a 5-kilometer run without breaks (without difficulty; with some difficulty)	201	-24.2 (-29.8, -18.7)	< 0.001**	8.3 (3.8, 12.7)	< 0.001**	16.1 (14, 18.1)	< 0.001**
Explorative excitability score	336	3.1 (0.4, 5.9)	0.026*	-2.2 (-4.4, 0.03)	0.054	0.6 (-0.4, 1.7)	0.248
Enjoyment of daily activities (often, fairly often)	205	-10.4 (-17.5, -3.3)	0.004**	6.8 (1.2, 12.4)	0.017*	3.8 (1.1, 6.4)	0.005**
Overall health-related quality of life score	334	-8.4 (-11.1, -5.7)	< 0.001**	5.81 (3.6, 7.9)	< 0.001**	4.1 (3.1, 5.1)	< 0.001**
Lean body mass	75	-8.8 (-13.4, -4.3)	< 0.001**	3.1 (0.3, 7.6)	0.033*	9.4 (7.8, 11.2)	< 0.001**
Disorderliness score	337	6.4 (3.6, 9.1)	< 0.001**	-4.8 (-6.1, -2.6)	< 0.001**	-0.7 (-1.8, 0.3)	0.188
Signs and symptoms such as pain, ache, nausea, itching, etc. (no signs or symptoms; minor signs or symptoms)	195	-18.7 (-29.3, -8)	0.001**	15 (6.5, 23.4)	0.001**	9.3 (5.3, 13.3)	< 0.001**
Impulsiveness score	336	2.2 (-0.5, 4.1)	0.112	-1.5 (-3.7, 0.7)	0.176	-0.4 (-1.4, 0.6)	0.436
Frequency of exercise through swimming (2–3 times a month or more)	212	-1.5 (-8.6, 5.5)	0.668	2.3 (-3.2, 8)	0.404	3.6 (0.9. 6.2)	0.008**
Considered retirement before the retirement age (no, I have not; I have sometimes)	213	3.5 (-5.1, 12.3)	0.424	2.3 (-4.6, 9.2)	0.510	3.4 (0.1, 6.7)	0.044*
Average weekday computer use time	456	14.7 (11.9, 17.5)	< 0.001**	-11.4 (-13.7, -9.2)	< 0.001**	-4.5 (-5.5, -3.4)	< 0.001**
The regression coefficients (B) with (95% confidence interval) from generalized linear mixed model controlling for sex, age, and with urban-rural area as a random effect (except for urban-rural area itself that no random effect was considered) are presented. $*p < 0.05$; $**p < 0.01$							
Reference categories in categorical variables: urban-rural areas: inner urban; outer urban; peri-urban; unable to work due to diseases/injuries: sometimes reduce in work pace or changes to work is needed; often reduce in work pace or changes to work is needed; part time work on account of illness is needed; totally unable to work; frequency of exercise through gardening: not at all; once a month or less; occupational group: directors and senior management; senior advisors and senior officials; advisors and officials; office workers and customer service representatives; cannot say; difficulty of a 2-kilometer run without breaks: with much difficulty; cannot at all; basic education: matriculation examination; sleeping problems: considerable problems; major problems; severe							

insomnia; frequency of exercise through walking: not at all; once a month or less; total amount of sleep: significantly insufficient; totally insufficient; difficulty of a 5-kilometer run without breaks: with much difficulty; cannot at all; enjoyment of daily activities: now and then; hardly ever; never; signs and symptoms: considerable signs or symptoms; strong signs or symptoms; intolerable signs or symptoms; frequency of exercise through swimming: not at all; once a month or less; considered retirement before the retirement age: I have often; I have applied for a pension.

Discussion

This study applied a decision tree technique to establish a multilevel data-driven model predicting PA behavior defined based on activity profiles and to methodologically identify the correlates of PA behavior. The decision tree fitting was performed using 168 factors of different domains as input variables to predict physically active and inactive individuals. The final model selected a total of 36 factors of different domains by which formed 54 different subgroups of subjects. 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 multilevel model and can be potentially informative for both multilevel intervention allocation and design, since it specifies the correlates of PA behavior at different level in each subgroup.

In agreement with the results of prior studies focused on understanding the causation of PA behaviors (6,9,12), the established model in the present study indicates that PA behavior is explained by a multilevel hierarchy composed of various factors of different domains. However, our results go beyond previous reports, showing that the predictors of PA behavior for different subgroups are different and from various domains. This is new and noteworthy because it can have potential implications for designing targeted, multilevel interventions including effective screening of subgroups followed by suggesting key modifiable factors for each subgroup. The data-driven nature of our model is also important, since prior studies have conceptualized the influences of PA behaviors mainly on the basis of theoretical combining of common sense and well-established evidence, and therefore primarily provided a broad view of PA behavior and its causation for general populations (5,9). Even though previous multilevel models have been successful in hypothesizing the interaction between factors of different domains, their implications in practice have remained limited (9), partially due to their theoretical nature. There are two studies that have applied a data-driven approach and established a decision tree-based model but with self-reported measurement of PA and

a limited number of factors, one of which using only demographical factors (29) and the other using only sociodemographic factors (30).

The emerged factors in the decision tree model included well-established, less established, and novel (unknown) correlates of PA behavior with regards to the previous studies focusing on both identifying and prioritizing the correlates of PA and sedentary behavior (5,6,9,10,31). The well-established correlates include factors that have been assessed in several studies and recognized as correlates. Most of the demographical, psychological, and environmental factors in the decision tree model have been recognized as factors associated with PA behavior in the past works (5,6,10). Some examples of these factors are education level, profession, overall health status, fitness status, and population density. The decision tree model also included some less established factors such as the factors related to personality and temperament such as extravagance, impulsiveness, and explorative excitability (5). Such factors (or factors similar to them) were assessed in a few or several studies but, mostly due to the limited or sometimes contradictory evidence, had not yet been identified as correlates nor been rejected. The body composition measures (i.e., lean body mass and skeletal muscle mass) and a few of the psychological and environmental factors such as enjoyment of daily activities and number of road accidents can also be categorized as less established factors (5,9,31). The decision tree model also included a few measures related to heart rate recovery. Even though the association of PA with heart rate recovery measures have been well-studied (32), 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 that are found here are potential candidates for the next generation of correlates. These factors were selected by the decision tree for creating 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 were not selected by the decision tree to create the model, such as behavioral attributes (e.g.,

alcohol, smoking, etc.) or socioeconomic status (5). The less established and novel factors that are found here have most probably remained underreported (or unexamined) due to the subjective tendency in the existing literature towards examining only those factors for which there has been a very well-established evidence of significant associations (positive or negative) with different indices of PA behavior (10). This has resulted in numerous correlates studies typically focusing on only psychosocial and environmental factors (6,10) and, accordingly, the emergence of calls for correlates research across different domains to discover the next generation of correlates (6).

Nevertheless, relative importance of the emerged factors should be inferred with caution. There might be some other possible reasons, at least in some cases, rather than higher relative importance of the factors in the final model compared to those that did not appear in the final model. First, the study's subjects had a narrow age range (46–48 years). This might explain why some of the very well-known correlates of PA behavior including age and gender did not appear in our final model (5,6,10,33). According to a previous systematic review (10), age was inversely associated with PA participation and when the age of the study's subjects was diverse enough, there were significant differences in PA participation between men and women (higher in men). Second, the outcome variables in our study were defined differently compared to prior works. To date, there is no agreed upon approach to differentiate between individuals with different PA behaviors. Previous works have used different, sometimes arbitrary, cut-offs (e.g., 150 minutes of MVPA per week, more than 7.5 hours of SED time, etc.) to define PA behavior (30,34). It has been argued that cut-offs are inappropriate for defining PA behavior because they do not reflect the whole range of activity intensities performed by individuals in everyday life (8). Instead, we defined PA behavior based on activity profiles derived from a range of activity intensities including SED, LPA, and MVPA, assessed objectively over the course of one full week (26). Therefore, the relative importance of the appeared factors in the final model should be interpreted with respect to our definition of PA behavior that is more complicated and composed of a range of activity across the intensity spectrum.

Body fat percentage, a direct measure of adiposity, was the most important factor explaining PA behavior 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 that has suggested the existence of a possible bidirectional relationship between adiposity and PA behavior on the basis of satisfactory evidence from longitudinal studies in which PA was predicted by adiposity (6). 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 (35), or can even indicate the presence of coronary artery disease (36). Another example is low muscle mass that is associated with increased risk of type 2 diabetes, metabolic syndrome, and coronary artery disease (37). Chronic health conditions have been identified both as a barrier and as motivations towards PA in different populations (38). 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 subjects 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 (39), and higher levels of LPA and MVPA (5). 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 potentially modifiable factors rather than a few subjectively selected factors (6,10), 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 (40). 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 (8).

This study is not without limitations. The study design was cross-sectional, and therefore the causality of identified factors remains unknown. The causality of the identified factors, especially the less established and novel ones, needs to be tested in future studies. Additionally, some of the emerged factors in the final model were related to cultural and health behaviors. This might limit the results of this study to only the current study population and might not be generalizable to different study populations with different cultural and health behaviors. Further studies are required to confirm the generalizability of the identified factors across different populations.

Conclusion

Using a data mining approach, a multilevel model was established from empirical and large-scale data. The model was composed of 36 different factors of relative importance from different domains explaining PA behavior. 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 multilevel model presented here specifies correlates of PA behavior at different level in each subgroup, and can be potentially utilized for multilevel intervention allocation and design. Additionally, 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.

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; 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 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. 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.
3. 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.
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. 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).
6. 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.
7. 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.

8. 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.
9. 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).
10. 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.
11. Venkatasubramaniam A, Wolfson J, Mitchell N, Barnes T, JaKa M, French S. Decision trees in epidemiological research. Emerg Themes Epidemiol. 2017;14(11).
12. 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.
13. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform. 2008;77(2):81–97.
14. 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.
15. University of Oulu Web site. NFBC 1966 data collection. Available from: https://www.oulu.fi/nfbc/node/19663. Accessed 1 Feb 2020.
16. Sintonen H. The 15D instrument of health-related quality of life: properties and applications. Ann Med. 2001;33(5):328–36.
17. 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.

18. University of Oulu Web site. 46-year follow-up study. Available from: https://www.oulu.fi/nfbc/node/26627. Accessed 1 Feb 2020.

19. 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.

20. University of Oulu Web site. 46-year follow-up study, Clinical examination. Available from: https://www.oulu.fi/nfbc/node/30371. Accessed 1 Feb 2020.

21. 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.

22. Kingsford C, Salzberg SL. What are decision trees? Nat Biotechnol. 2008;26:1011–3.

23. 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.

24. 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.

25. 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.

26. 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. 2019;124:33–41.

27. 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.

28. 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.

29. 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.

30. 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.

31. 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).

32. 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.

33. 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).

34. Trinh OTH, Nguyen ND, Dibley MJ, Phongsavan P, Bauman AE. The prevalence and correlates of physical inactivity among adults in Ho Chi Minh City. BMC Public Health. 2008;8(204).

35. 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.

36. 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.

37. Ko B-J, Chang Y, Jung H-S, Yun KE, Kim C-W, Park HS, et al. Relationship between low relative muscle mass and coronary artery calcification in healthy adults. Arterioscler Thromb Vasc Biol. 2016;36(5):1016–21.

38. 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.

39. Rhodes RE, Mark RS, Temmel CP. Adult sedentary behavior: a systematic review. Am J Prev Med. 2012;42(3):e3–28.

40. 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.

Figures

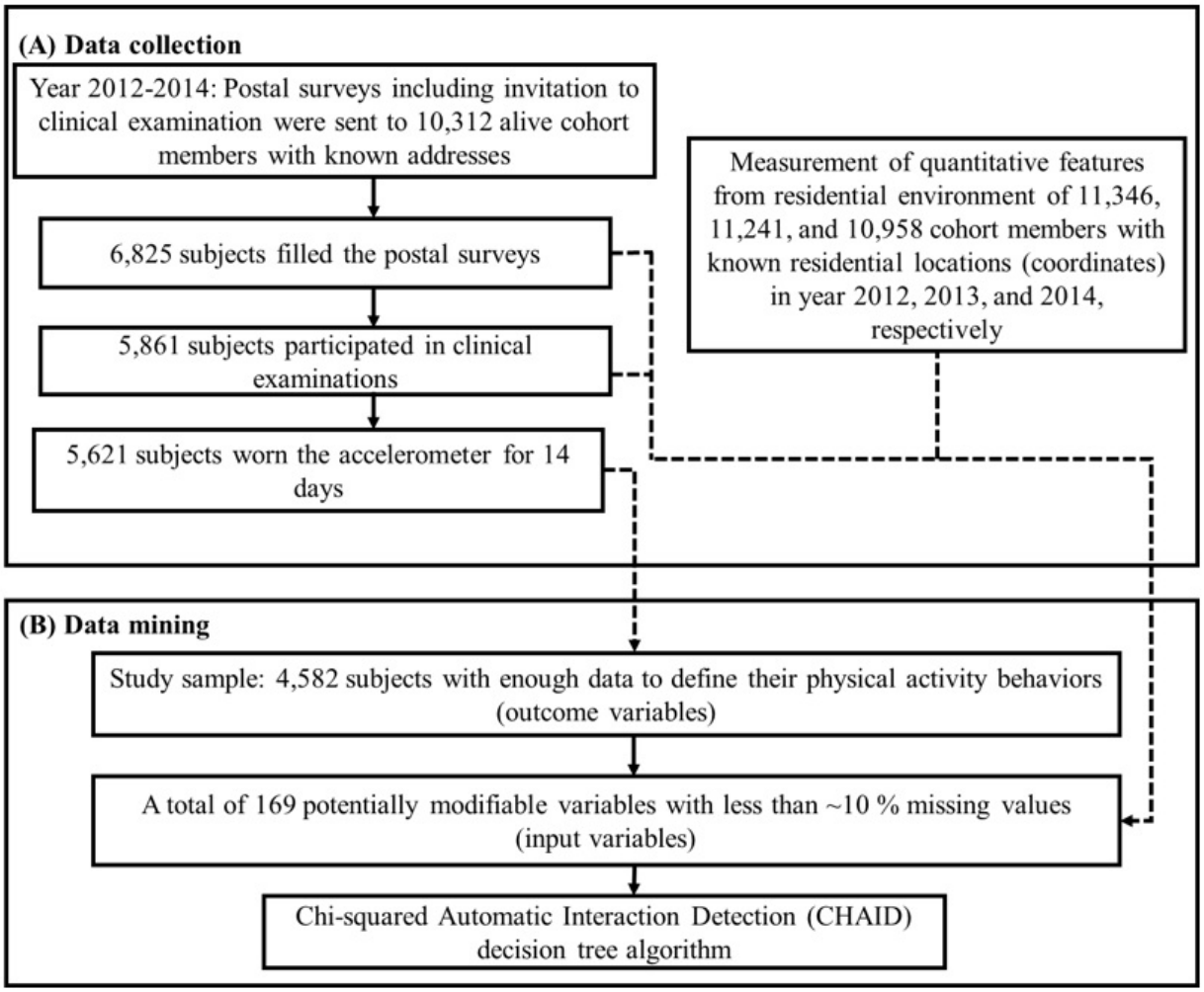

Figure 1

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).

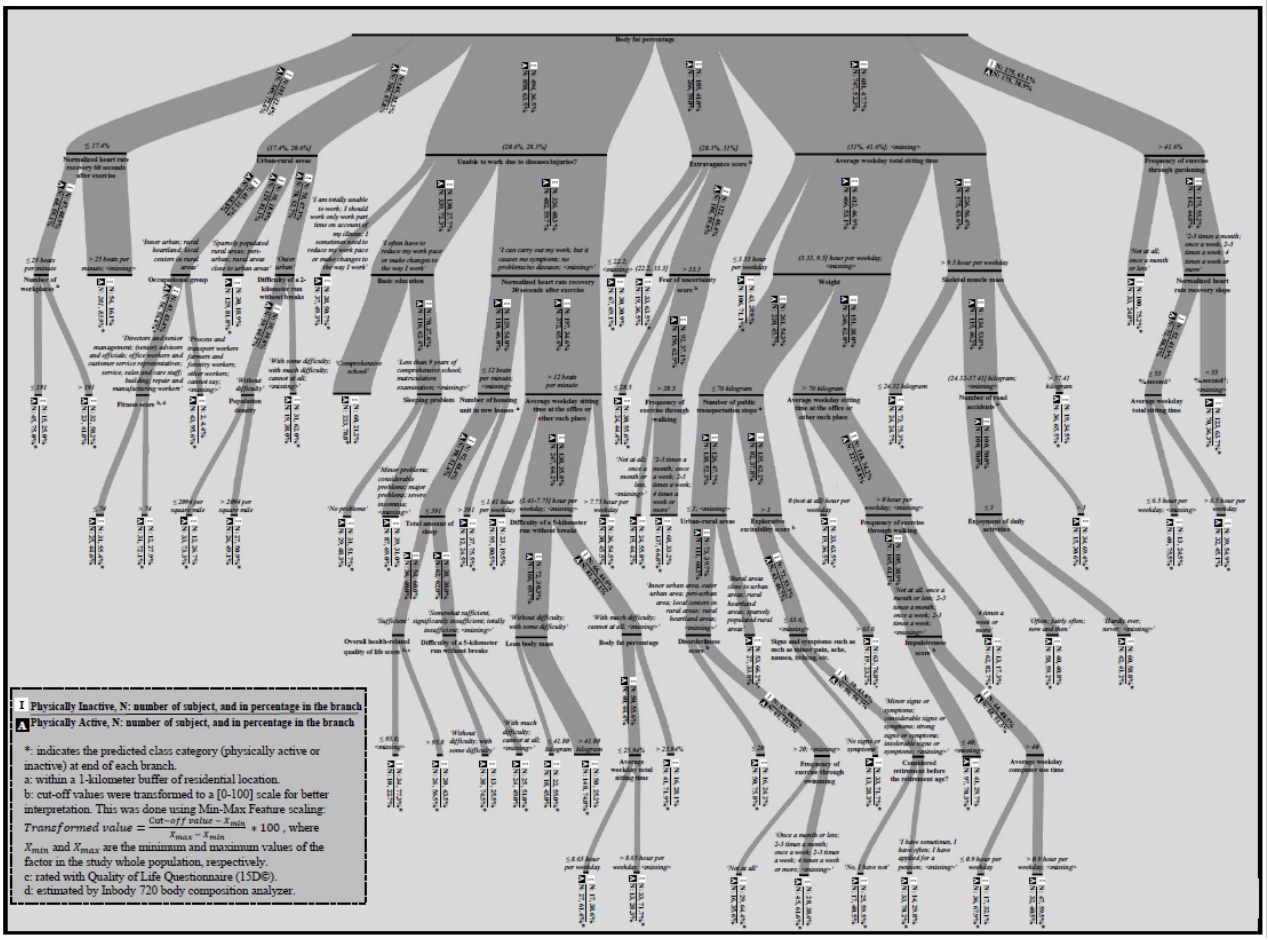

Figure 2

The Chi Squared Automatic Interaction Detection tree illustrating the hierarchy of the factors predicting physical activity and inactivity. The thickness of branches is based on the number of subjects in the branch. Categories (for categorical 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. (Original file provided in the Supplementary Files section).

Supplementary Files

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

Farrahi_Figure 02.pdf
Farrahi_Supplementary Material 01.docx