Prediction of multisite pain incidence in adolescence using a machine learning approach

Laura Joensuu\textsuperscript{a} Ph.D*, Ilkka Rautiainen\textsuperscript{a} MSc*, Arto Hautala\textsuperscript{a} Ph.D., Kirsti Siekkinen\textsuperscript{b} MSc,
Katariina Pirnes\textsuperscript{a} Ph.D, Tuija H Tammelin\textsuperscript{b} Ph.D.

\textsuperscript{a}Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland, \textsuperscript{b}Likes, Jamk University of Applied Sciences, Jyväskylä, Finland

*Authors contributed equally

Corresponding Author:
Ph.D. Postdoctoral Researcher Laura Joensuu,
Faculty of Sport and Health Sciences, University of Jyväskylä, Rautpohjankatu 8, 40700 Jyväskylä, Finland;
laura.p.joensuu@jyu.fi, @laurajoensuu, Orcid ID: 0000-0002-9544-6552

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
**Key points**

**Question:** What is the ability of machine learning approach to predict multisite pain incidence during adolescence?

**Findings:** With a random forest machine learning method, a broad variety of predictive physical, lifestyle and psychosocial factors were identified. Prediction ability reached AUC 0.65.

**Meaning:** The findings highlight that predictors of multisite pain incidence in adolescence are multifaceted, although the prediction ability of machine learning remained under clinical relevance (AUC <0.7). These findings support the adoption of holistic and multidisciplinary prevention approaches for multisite pain in adolescence in the future.
Abstract

Importance
Multisite pain is a major adverse health outcome in the adolescent population, affecting the daily lives of up to every third adolescent and their families.

Objective
To 1) predict multisite pain incidence in the whole body and in the musculoskeletal locations in adolescents, and 2) explore the sex-specific predictors of multisite pain incidence with a novel machine learning approach.

Design
A 2-year observational study (2013-2015). Three different baseline data sets were utilized to predict multisite pain incidence during the follow-up.

Setting

Participants
Apparently healthy adolescents.

Exposures
The first data set included 48 selected baseline variables relevant for adolescents’ health and wellbeing. Data included information on students self-reported, objectively measured, and device-based demographics, physical and psychosocial characteristics, and lifestyle factors. The second data set included nine physical fitness variables related to the Finnish national ‘Move!’ monitoring and surveillance system for health-related fitness. The third data set included all available baseline data (392 variables).

Main Outcome and Measures
Onset of multisite pain (=weekly pain during the past three months manifesting in at least three sites and not related to any known disease or injury) during the 2-year follow-up in the whole body or musculoskeletal locations. Musculoskeletal pain sites included the neck/shoulder, upper extremities, chest, upper back, low back, buttocks, and lower extremities. Whole body pain sites also the head and abdominal areas. A machine learning algorithm random forest was utilized.

Results
Among 410 participants (57% girls) aged on average 12.5-years (SD 1.2), 16 % of boys and 28 % of girls developed multisite pain in the whole body and 10 % and 15 % in the musculoskeletal area during follow-up.
The prediction ability of the machine learning approach with 48 predictive variables reached an AUC 0.65 at highest. With ML, a broad variety of predictors were identified, with up to 33 variables showing predictive power in girls and 13 in boys.

Conclusions and Relevance

Findings highlight that rather than any isolated variable, a variety of factors pose a risk for future multisite pain. More emphasis on holistic and multidisciplinary approaches is recommended to prevent multisite pain in adolescence.
Introduction

Pain is common in adolescents. Long-lasting pain in at least two bodily locations is reported by at least every tenth and up to every third adolescent in large cohort studies, with musculoskeletal locations most common sites for pain. This co-occurrence of pain is typical in the adolescent population, and hence, multisite pain is recommended to be considered in clinical practice over isolated pain sites. While the adolescent population is relatively free from many disabling health outcomes, experiences of pain are associated with a lesser ability to conduct daily activities. This association follows a dose-response pattern where more pain sites are associated with a higher degree of disability. The relevance of multisite pain in adolescence is stressed by pediatric pain researchers. Pain in adolescence is associated with a broad range of adverse outcomes, from limitations in school attendance and hobbies to reduced quality of life and depressive symptoms. Furthermore, pain experiences tend to track from childhood and adolescence into adulthood, with relevance to, e.g., future work disability.

Previous studies have found several cross-sectional correlates with pain. Various biological, psychosocial, and lifestyle factors, such as age, pubertal status, overweight or obesity, symptoms of anxiety and depression, chronic health problems, frequent change of residence, poor academic achievement, leisure screen time, fewer interactions with peers, unhealthy lifestyles (e.g., sedentary behavior, screen time, inadequate sleep and smoking), and excessive physical activity, especially in a technical, team, strength, or extreme sports increase the odds for overall or musculoskeletal pain. In addition, physical fitness is suggested to be associated with pain in adolescents. Especially the associations of flexibility and muscular fitness with musculoskeletal pain have been of great interest. Findings remain persistently inconclusive, although the majority of studies have focused on examining a specific pain
Despite the inconclusive evidence, proper functioning of the musculoskeletal system, i.e., adequate exertion of force, fatigue resistance, and range of motion in the body, is a rationale for many health-related large-scale monitoring and surveillance systems to implement fitness testing at the population level. Previous findings also indicate that girls report pain more often than boys, and the correlates of pain might be sex-specific, potentially due to differences in maturation, pain tolerance, or coping behaviors between sexes.

Less is known about the predictors of pain incidence. Predicting the future onset of multisite pain has proven to be challenging despite the broad range of potential explanatory factors. For example, Paananen et al. (2010) did not find statistically significant predictors for multisite musculoskeletal pain incidence in a 2-year follow-up study. Recently, machine learning (ML)-based pattern recognition approaches have emerged as promising alternatives to traditional statistical approaches in modeling complex phenomenon in various areas of society, including health care. Therefore, the aim of this study was to 1) predict multisite pain incidence in the musculoskeletal and whole body sites in adolescents and 2) explore the sex-specific determinants of multisite pain incidence utilizing a novel machine learning approach.

**Methods**

**Study population**

This study was part of a research entity related to the Finnish Schools' on the Move program focusing on physical activity and wellbeing among children and adolescents. A longitudinal observational study was conducted between January 2013 and June 2015. A total of 1778 students from nine Finnish schools were invited to participate. Out of these, 970 students...
(53% girls) provided signed written consent with their main caregiver and participated in the study. After excluding students with possible confounding factors at baseline (such as guardian-reported existing chronic illnesses or disorders, injuries, existing multisite pain, and more than 50% of missing data), the final sample consisted of 410 apparently healthy participants (57% girls) (Figure 1).

The study setting and measurements were approved by the Ethics Committee of the University of Jyväskylä, and all procedures were conducted in accordance with the principles outlined in the Declaration of Helsinki. Participants had the option to discontinue their involvement at any point during the research. All measurements were conducted by trained personnel.

**Outcome**

Pain incidence was considered the new onset of multisite pain at any time point during the 2-year follow-up. Pain symptoms were screened four times after baseline with a structured questionnaire at six-month intervals: “How often you have had symptoms in the last three months (in body parts A-I in the pictures below)? Mark the appropriate option. Headache (A), Neck and shoulder pain / ache (B), Upper extremities pain / ache (C), Chest pain / ache (D), Upper back pain / ache (E), Low back pain / ache (F), Stomach ache (G), Buttocks pain / ache (H), Lower extremities pain o/ ache (I)”.

The question was supported by an illustration of the described body parts. The answering options were: “Almost daily, More than once a week, About once a week, About once a month, Seldom or never.” Students also reported if the pain originated from trauma: ”Have you injured any of the above-mentioned and pictured pain areas during the previous 3 months (for example, fallen, stumbled, breached during sport, etc.)?” The answer options were “yes” or “no” and provided additional information.
related to the injured body area.

Multisite pain was defined as reported weekly pain (almost daily, more than once a week, or about once a week) in at least three sites during the past three months. Pains due to traumatic causes were excluded from the analysis. Multisite pain was reported separately for the whole body and musculoskeletal locations. The pain reported in at least three sites was selected to reflect the disabling form of multisite pain.\textsuperscript{3} Musculoskeletal pain sites included the neck/shoulder, upper extremities, chest, upper back, low back, buttocks, and lower extremities. Whole body pain sites included additionally head and abdominal areas.

**Predictive variables**

We aimed to determine the predictors of multisite pain incidence using three different data sets. First, we included 48 selected baseline variables relevant to adolescents’ physical activity, fitness, health, and wellbeing.\textsuperscript{22} The data included information on participants’ basic demographics, physical, psychosocial, and lifestyle characteristics and is presented in detail elsewhere.\textsuperscript{22} The results of these analyses are presented in the main text.

Secondly, we used baseline physical fitness measurements belonging to the Finnish national ‘Move!’ monitoring and surveillance system for health-related fitness\textsuperscript{23} (nine baseline variables described in detail elsewhere)\textsuperscript{24}. Annually, approximately 100,000 children and adolescents (approximately 96% of the relevant age groups) participate in ‘Move!’ creating a unique database for health-enhancing policies.\textsuperscript{18}

Thirdly, a data-driven approach was used with the whole available data (392 baseline variables) to explore potential novel predictors of multisite pain incidence. Data included
extensive information on students’ self-reported, objectively measured, and device-based
demographics, physical and psychosocial characteristics, and physical activity.

Analytical procedures

The random forest (RF) method was applied. All analyses were performed using MATLAB
R2022b with the Statistics and Machine Learning Toolbox and conducted separately for both
sexes. Initial preprocessing and creation of the outcome variable were made using the Python
programming language.

RF is an ML method where multiple de-correlated decision trees are grown to form a forest.
Afterward, this forest is employed as a voting ensemble, where each tree provides an answer
for the prediction task. The final prediction of the forest is the class that gets the most votes
from the individual trees.25,26

10-fold cross-validation (CV) was employed for model assessment. During CV, the data for
each prediction task was divided into 10 subsamples called folds. Nine of these folds, 90% of
the whole data set, were used as the training data to fit the RF model, while one fold, 10% of
the data, was used as the validation data. This procedure was repeated ten times in a rotating
manner, where eventually all the folds had been employed for training and validation. Thus,
all the presented results are based on ten separate data-driven prediction models.

For each of the 10 CV folds, the trained model was employed to predict the out-of-bag
(OOB) observations i.e., those observations which were not utilized during the training of
each tree, and the validation portion of the data. The main metric recorded was the area the
under receiver operating characteristic curve (AUC). T-tests were performed in MATLAB
for the OOB and validation data AUC results to determine if the means of the CV folds were
significantly (p<0.05) above the random level of 0.5. Further analyses regarding the predictive power of each variable were conducted only in those cases where AUC 95% Confidence Interval did not violate the 0.5 threshold.

RF requires choosing several hyperparameters i.e., options that define the model creation. F-measure for the OOB observations was used as a target during Bayesian optimization, where several hyperparameters of the RF model were chosen in an automated fashion. Please see Supplementary methods in Supplement 1 for further information on the target measure, the hyperparameters, and other details concerning the RF model.

The contribution of each variable to prediction was estimated using the OOB observations by a permutation importance measure. A baseline result for the model in each CV fold was the accuracy of the model with the original data. To estimate the contribution of each variable, the values of the variables were permuted randomly. The procedure was repeated for all the variables separately, and the accuracy of the model with permutations was recorded for each variable. The accuracy obtained using the permuted variable was then subtracted from the baseline accuracy. The final permutation importance estimate for each variable was the mean of accuracy change for the 10 CV folds. T-tests were employed also for the importance estimates. If the change was significantly (p<0.05) over zero, the variable was seen as having predictive power. Furthermore, if the mean change was near zero or negative, the variable did not have importance in the prediction. MATLAB’s predict function in the TreeBagger class was utilized to calculate the OOB predictions on the trained model. This function computed the weighted average of the class posterior probabilities over the trees.
In further sensitivity analyses the class imbalance was considered, meaning that there are considerably less observations in diffuse idiopathic pain class, is a challenge in all explored settings. This issue was approached during the modelling in two separate ways. Firstly, in the RF model by changing the default cost matrix of misclassification. Cost of misclassifying true pain class observations to no pain class (false negative classification) was increased to 2, while the other misclassification (false positive) was left to its default value 1.

Furthermore, as an alternate view, a synthetic minority oversampling technique for nominal and continuous data SMOTE-NC,\textsuperscript{28} was utilized to see if artificially balancing the training data by oversampling the pain class observations provided any performance improvements. Since SMOTE-NC, available in Themis library in R, expected that there are no missing values in data, missForest imputation for mixed-type data was utilized before oversampling. As a limitation, due to artificially manipulating the training data, the OOB observations could not be meaningfully utilized during this experiment and only non-manipulated validation data for each CV fold was used when estimating the performance measures.

When estimating the importance of individual variables, the associated risk for each variable was examined with simple ROC analysis while acknowledging how the variables were coded in the data. The analysis was done separately from the RF model for the whole age-adjusted data once without utilizing cross-validation. The analysis was performed only for continuous and ordinal variables. The identified risk variables are presented in permutation importance estimate figures with a red panel.

**Results**

At baseline, headache (22.5%, 30.4%) and neck and shoulder pain (13.5%, 18.5%) were the most prevalent pain symptoms among boys and girls, respectively (Table 1). Sixteen percent
of boys and 28.1% of girls experienced multisite pain incidents in the whole body area during
the 2-year follow-up. Multisite pain incidence in the musculoskeletal area was 9.6% and
15.3% in boys and girls, respectively (Table 2).

The ability of the machine learning approach to predict whole body multisite pain incidence
reached an AUC 0.54 (95% Confidence Interval 0.49 to 0.58) for boys and 0.65 (0.63 to
0.67) for girls (Table 2). The prediction ability for multisite musculoskeletal pain incidence
was AUC 0.65 (0.62 to 0.68) in boys and 0.51 (0.48 to 0.54) in girls.

The tasks where prediction ability reached above random level (AUC >0.5) were further
analyzed for variable importance. Altogether, 33 variables out of 48 baseline variables
showed predictive power for whole body multisite pain incidence among girls. All variables
are illustrated in Figure 2, and the top ten are described in detail here. Poorer perceived
health, higher perceived fitness, more frequent tiredness on schoolday mornings, having
overweight or obesity based on body mass index, more frequent participation in sports
competitions and matches, more frequent breakfast eating during the school week, a lower
grade point in physical education, a lower amount moderate-to-vigorous physical activity
during leisure time, higher school enjoyment, and higher pubertal status increased the
probability of multisite pain incidence in the whole body area in girls.

In boys, a total of 13 variables out of 48 showed predictive power for multisite
musculoskeletal pain incidence. The top ten predictors for pain incidence included higher
school strain, lower school enjoyment, a higher participation rate in sports competitions or
matches, lower amounts of continuous device-measured sedentary time, better muscular
fitness measured with the number of push-ups conducted within one minute, a lower body
mass index, more active participation in physical activity clubs in school, a later bedtime on schooldays, lower total sedentary time, and parents’ higher willingness to help with schoolwork (Figure 3).

Prediction ability with the Move! variables reached above the random level only in boys and in the whole body area (AUC 0.59 (0.56 to 0.62), eTable 1) and indicated that better muscular and cardiorespiratory fitness but poorer motor fitness predict higher multisite pain incidence (eFigure 1). With the full available data set ML, was able to predict multisite pain incidence only in girls (AUC 0.68 (0.66 to 0.70) and 0.58 (0.56 to 0.60) in whole body and musculoskeletal sites, respectively) (eTable 2). With the full data, along with physical, psychosocial, and lifestyle factors, individual pain sites at baseline rose as predictive factors of future multisite pain (eFigures 2 and 3). Balancing the data artificially with SMOTE-NC did improve prediction ability (up to AUC 0.72, with high a standard deviation as a result of the small size of each validation fold), but due to automatic risk threshold selection designed for earlier tasks, it created suboptimal sensitivity and specificity values (eTables 3-5).

**Discussion**

In this study, we aimed to investigate determinants of multisite pain incidence among the adolescent population with a novel ML approach. Multisite pain incidence in the study population was considerable, with up to 16% of boys and 28% of girls developing multisite weekly pain during the 2-year follow-up. The prediction ability of the ML approach with selected predictive variables reached an AUC 0.65 at its highest. With ML, a broad variety of variables predicting multisite pain incidence in adolescents were identified. Out of 48 selected variables, up to 33 variables showed predictive power in girls and 13 in boys. These findings highlight that rather than any isolated variable, a variety of factors may possess an
increased risk for multisite pain and indicate the paradoxical nature of some variables, especially in girls.

Multisite pain is a major adverse health outcome in the adolescent population, affecting the daily lives of more than every fourth adolescent and their families.\textsuperscript{1} Predicting the future onset of multisite pain, identifying individuals potentially experiencing disabling pain in the future, and recognizing the predictors of pain hold the potential to enhance the quality of life in this important demographic through better health education and policies. Pediatric experts have long stressed the importance of pain research and further understanding of pain epidemiology and underlying pathophysiology through innovative study designs.\textsuperscript{5}

Machine learning-based pattern recognition algorithms, a subgroup of artificial intelligence, have emerged as promising alternatives to traditional statistical methods in developing next-generation tools to enhance public health. In contrast to theory-based and often restricted traditional statistical models, the ML approach enables near unlimited learning capacity from the available data,\textsuperscript{25} providing the potential to develop more precise methods for screening and predicting adverse health outcomes. ML-based approaches are acknowledged to hold significant potential for reforming public health policies in the future.\textsuperscript{29}

Previous studies have shown that prediction of multisite pain incidence is demanding,\textsuperscript{13} and the isolated correlates have modest effect sizes.\textsuperscript{8} In this current study, the ML approach was able to predict pain incidence above the random level, however remaining under clinical relevance (AUC <0.7).\textsuperscript{30} Through the ML approach, we found various predictors for multisite pain incidence, reflecting the previously reported physical, lifestyle, and psychosocial correlates,\textsuperscript{1,2,12–15} and complementing these findings by illustrating the risky variables in a
holistic framework alongside the paradoxical nature of some variables. For example, with whole body multisite pain incidence among girls, indicators of both lower and higher psychosocial wellbeing (e.g. lower life enjoyment vs. higher school enjoyment), low and high physical activity (lower amount of moderate-to-vigorous physical activity during leisure time vs. more days with physical activity for at least 60 min per day), lower body mass index, and obesity or overweight classification were identified as risk factors. These findings illustrate that risk factors, especially for whole body multisite pain incidence in girls are complex, associations are not linear, and individuals with both healthy and unhealthy lifestyles, favorable or unfavorable psychosocial status might develop multisite pain in the future. In boys, findings indicated more consistently that poorer psychosocial wellbeing, higher physical activity, a leaner body, and better physical fitness predict multisite musculoskeletal pain incidence and support the acknowledgment of overall wellbeing and health-enhancing physical activity practices among boys to prevent musculoskeletal pain.

The strengths of this study were the novel application of ML in pain prediction, the longitudinal study design, and the extensiveness of predictors. The ML approach considerably extends pain research and provides potential avenues for screening and modeling complex phenomena in the future. The data was however limited by information (no data on current medication) and cases (e.g., <66 cases in the data set) with possibly influencing the generalizability of the findings. ML explores patterns in the data and does not explain underlying mechanisms or causality. The multicollinearity of the variables might affect the interpretation of variables with similar phenomenal origins. Self-reported data may suffer from recall bias, although the reliability of the utilized questionnaire has shown to be reasonable.21
In conclusion, these novel findings highlight the multifaceted predictors of multisite pain incidence in adolescents and support the adoption of holistic and multidisciplinary prevention approaches in the future.

**Statements**

This study was funded by Ella and Georg Ehrnrooth foundation. Data collection for this study was supported by the Juho Vainio Foundation (201410342) and the Finnish Ministry of Education and Culture (OKM/92/626/2013).

**Data sharing**

Data and utilized scripts are available upon reasonable request from IR (scripts) and THT (data).

**Declaration of interests**

We declare no competing interests.
References


Table 1. Selected subject demographics at baseline by sex.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>N=410</th>
<th>Boys</th>
<th>Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td></td>
<td>12.5 (1.2)</td>
<td>12.5 (1.2)</td>
</tr>
<tr>
<td>Height (cm)</td>
<td></td>
<td>156.7 (10.9)</td>
<td>155.1 (9.6)</td>
</tr>
<tr>
<td>Weight (kg)</td>
<td></td>
<td>46.1 (12.2)</td>
<td>45.4 (10.1)</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td></td>
<td>18.5 (3.2)</td>
<td>18.7 (3.0)</td>
</tr>
</tbody>
</table>

Pain at different body sites

Musculoskeletal pain sites

<table>
<thead>
<tr>
<th>Musculoskeletal pain sites</th>
<th>N=410</th>
<th>Boys</th>
<th>Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neck/Shoulder (%)</td>
<td></td>
<td>23 (13.6 %)</td>
<td>42 (18.5 %)</td>
</tr>
<tr>
<td>Upper extremities (%)</td>
<td></td>
<td>7 (4.1 %)</td>
<td>10 (4.4 %)</td>
</tr>
<tr>
<td>Chest (%)</td>
<td></td>
<td>1 (0.6 %)</td>
<td>2 (0.9 %)</td>
</tr>
<tr>
<td>Upper back (%)</td>
<td></td>
<td>5 (3.0 %)</td>
<td>5 (2.2 %)</td>
</tr>
<tr>
<td>Low back (%)</td>
<td></td>
<td>8 (4.7 %)</td>
<td>8 (3.5 %)</td>
</tr>
<tr>
<td>Buttocks (%)</td>
<td></td>
<td>2 (1.2 %)</td>
<td>6 (2.6 %)</td>
</tr>
<tr>
<td>Lower extremities (%)</td>
<td></td>
<td>9 (5.3 %)</td>
<td>13 (5.7 %)</td>
</tr>
</tbody>
</table>

Other pain sites

<table>
<thead>
<tr>
<th>Other pain sites</th>
<th>N=410</th>
<th>Boys</th>
<th>Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head (%)</td>
<td></td>
<td>38 (22.5 %)</td>
<td>69 (30.4 %)</td>
</tr>
<tr>
<td>Abdominal (%)</td>
<td></td>
<td>18 (10.7 %)</td>
<td>36 (15.9 %)</td>
</tr>
</tbody>
</table>

Physical activity and fitness

<table>
<thead>
<tr>
<th>Physical activity and fitness</th>
<th>N=410</th>
<th>Boys</th>
<th>Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moderate-to-vigorous physical activity (min/day)</td>
<td>58.3 (24.2)</td>
<td>47.4 (18.3)</td>
<td></td>
</tr>
<tr>
<td>Sedentary time (hours/day)</td>
<td></td>
<td>8.1 (1.3)</td>
<td>8.6 (1.1)</td>
</tr>
<tr>
<td>Physical fitness index (Move-index)</td>
<td>16.6 (4.1)</td>
<td>17.2 (3.6)</td>
<td></td>
</tr>
</tbody>
</table>

Values are means and standard deviations for continues variables and proportions of participants for others. BMI, Body mass index; Move-index, weighted sum of the Finnish national Move! monitoring system’s fitness items.
Table 2. Prediction ability of machine learning for multisite pain incidence among adolescents.

<table>
<thead>
<tr>
<th>Multisite pain incidence</th>
<th>Cases/N</th>
<th>AUC (95% CI)</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>All body sites</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boys</td>
<td>28/175</td>
<td>0.54 (0.49 to 0.58)</td>
<td>0.75 (0.64 to 0.86)</td>
<td>0.43 (0.30 to 0.55)</td>
</tr>
<tr>
<td>Girls</td>
<td>66/235</td>
<td>0.65 (0.63 to 0.67)</td>
<td>0.72 (0.67 to 0.78)</td>
<td>0.58 (0.53 to 0.62)</td>
</tr>
<tr>
<td>Musculoskeletal sites</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boys</td>
<td>17/178</td>
<td>0.65 (0.62 to 0.68)</td>
<td>0.72 (0.58 to 0.87)</td>
<td>0.65 (0.54 to 0.76)</td>
</tr>
<tr>
<td>Girls</td>
<td>36/235</td>
<td>0.51 (0.48 to 0.54)</td>
<td>0.81 (0.75 to 0.87)</td>
<td>0.37 (0.28 to 0.46)</td>
</tr>
</tbody>
</table>

AUC results are estimated from the out-of-bag observations using the 10-fold cross-validation.

**Figure legends**

Figure 1. Flow chart of the exclusion process

Figure 2. Permutation importance estimates for girls in the selected set for all sites (AUC 0.65). Red panel, risk factors; PA, physical activity; Counts, accelerometer total activity counts

Figure 3. Permutation importance estimates for boys in the selected set for musculoskeletal sites (AUC 0.65). Red panel, risk factors; PA, physical activity; Counts, accelerometer total activity counts
Original data

Whole body
N=970
Boys N=462
Girls N=508

Musculoskeletal locations

Exclusion criteria

1) Multisite pain reported at baseline.

<table>
<thead>
<tr>
<th>Whole body</th>
<th>Musculoskeletal locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excluded: 146 (67 boys, 79 girls)</td>
<td>Excluded: 71 (38 boys, 33 girls)</td>
</tr>
</tbody>
</table>

2) A known existing disorder at baseline.

<table>
<thead>
<tr>
<th>Whole body</th>
<th>Musculoskeletal locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excluded: 50 (30 boys, 20 girls)</td>
<td>Excluded: 58 (32 boys, 26 girls)</td>
</tr>
</tbody>
</table>

3) Self-reported injury at any location during the last three months at baseline.

<table>
<thead>
<tr>
<th>Whole body</th>
<th>Musculoskeletal locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excluded: 346 (181 boys, 165 girls)</td>
<td>Excluded: 373 (189 boys, 184 girls)</td>
</tr>
</tbody>
</table>

4) A student is completely excluded if none of the follow-ups have usable data i.e., data is either missing or there is a self-reported injury in all four follow-ups.

<table>
<thead>
<tr>
<th>Whole body</th>
<th>Musculoskeletal locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excluded: 10 (5 boys, 5 girls)</td>
<td>Excluded: 28 (2 boys, 26 girls)</td>
</tr>
</tbody>
</table>

Forming of the response variable for remaining students

N=410 (larger set)
Boys N=175, where multisite pain observations: 28
Girls N=235, where multisite pain observations: 66

Removing students where more than 50% of variables are missing (done separately for each three data configurations)

N=413 (larger set)
Boys N=178, where multisite pain observations: 17
Girls N=235, where multisite pain observations: 35
Figure 2. Permutation importance estimates for girls in the *selected set* for all sites (AUC 0.65). Red panel, risk factors; PA, physical activity; Counts, accelerometer total activity counts.
Figure 3. Permutation importance estimates for boys in the selected set for musculoskeletal sites (AUC 0.65). Red panel, risk factors; PA, physical activity; Counts, accelerometer total activity counts.