

1 Title page

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

4 Laura Joensuu^a Ph.D*, Ilkka Rautiainen^a MSc*, Arto Hautala^a Ph.D., Kirsti Siekkinen^b MSc,
5 Katariina Pirnes^a Ph.D, Tuija H Tammelin^b Ph.D.

6 *^aFaculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland, ^b Likes,*
7 *Jamk University of Applied Sciences, Jyväskylä, Finland*

8 **Authors contributed equally*

9 Corresponding Author:

10 Ph.D. Postdoctoral Researcher Laura Joensuu,
11 Faculty of Sport and Health Sciences, University of Jyväskylä, Rautpohjankatu 8, 40700 Jyväskylä, Finland;
12 laura.p.joensuu@jyu.fi, [@laurajoensuu](#), Orcid ID: 0000-0002-9544-6552

13 **Key points**

14

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

17

18 **Findings:** With a random forest machine learning method, a broad variety of predictive
19 physical, lifestyle and psychosocial factors were identified. Prediction ability reached AUC
20 0.65.

21

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

26 **Abstract**

27

28 **Importance**

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

31

32 **Objective**

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

36

37 **Design**

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

40

41 **Setting**

42 Population-based sample of Finnish adolescents.

43

44 **Participants**

45 Apparently healthy adolescents.

46

47 **Exposures**

48 The first data set included 48 selected baseline variables relevant for adolescents' health and
49 wellbeing. Data included information on students self-reported, objectively measured, and
50 device-based demographics, physical and psychosocial characteristics, and lifestyle factors.

51 The second data set included nine physical fitness variables related to the Finnish national
52 'Move!' monitoring and surveillance system for health-related fitness. The third data set
53 included all available baseline data (392 variables).

54

55 **Main Outcome and Measures**

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

62

63 **Results**

64 Among 410 participants (57% girls) aged on average 12.5-years (SD 1.2), 16 % of boys and
65 28 % of girls developed multisite pain in the whole body and 10 % and 15 % in the
66 musculoskeletal area during follow-up.

67

68 The prediction ability of the machine learning approach with 48 predictive variables reached
69 an AUC 0.65 at highest. With ML, a broad variety of predictors were identified, with up to 33
70 variables showing predictive power in girls and 13 in boys.

71

72 **Conclusions and Relevance**

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

76 **Introduction**

77 Pain is common in adolescents.¹ Long-lasting pain in at least two bodily locations is reported
78 by at least every tenth and up to every third adolescent in large cohort studies,² with
79 musculoskeletal locations most common sites for pain.³ This co-occurrence of pain is typical
80 in the adolescent population,⁴ and hence, multisite pain is recommended to be considered in
81 clinical practice over isolated pain sites.² While the adolescent population is relatively free
82 from many disabling health outcomes, experiences of pain are associated with a lesser ability
83 to conduct daily activities. This association follows a dose-response pattern where more pain
84 sites are associated with a higher degree of disability.³ The relevance of multisite pain in
85 adolescence is stressed by pediatric pain researchers.⁵ Pain in adolescence is associated with
86 a broad range of adverse outcomes, from limitations in school attendance and hobbies to
87 reduced quality of life and depressive symptoms.⁶⁻⁸ Furthermore, pain experiences tend to
88 track from childhood and adolescence into adulthood,^{9,10} with relevance to, e.g. future work
89 disability.¹¹

90

91 Previous studies have found several cross-sectional correlates with pain. Various biological,
92 psychosocial, and lifestyle factors, such as age, pubertal status, overweight or obesity,
93 symptoms of anxiety and depression, chronic health problems, frequent change of residence,
94 poor academic achievement, leisure screen time, fewer interactions with peers, unhealthy
95 lifestyles (e.g. sedentary behavior, screen time, inadequate sleep and smoking), and excessive
96 physical activity, especially in a technical, team, strength, or extreme sports increase the odds
97 for overall or musculoskeletal pain.^{1,2,12-15} In addition, physical fitness is suggested to be
98 associated with pain in adolescents. Especially the associations of flexibility and muscular
99 fitness with musculoskeletal pain have been of great interest. Findings remain persistently
100 inconclusive, although the majority of studies have focused on examining a specific pain

101 site.^{16,17} Despite the inconclusive evidence, proper functioning of the musculoskeletal system,
102 i.e adequate exertion of force, fatigue resistance, and range of motion in the body, is a
103 rationale for many health-related large-scale monitoring and surveillance systems to
104 implement fitness testing at the population level.^{16,18} Previous findings also indicate that girls
105 report pain more often than boys,^{2,3} and the correlates of pain might be sex-specific,
106 potentially due to differences in maturation, pain tolerance, or coping behaviors between
107 sexes.^{2,12}

108
109 Less is known about the predictors of pain incidence. Predicting the future onset of multisite
110 pain has proven to be challenging despite the broad range of potential explanatory factors.
111 For example, Paananen et al. (2010) did not find statistically significant predictors for
112 multisite musculoskeletal pain incidence in a 2-year follow-up study.¹³ Recently, machine
113 learning (ML)-based pattern recognition approaches have emerged as promising alternatives
114 to traditional statistical approaches in modeling complex phenomenon in various areas of
115 society, including health care.¹⁹ Therefore, the aim of this study was to 1) predict multisite
116 pain incidence in the musculoskeletal and whole body sites in adolescents and 2) explore the
117 sex-specific determinants of multisite pain incidence utilizing a novel machine learning
118 approach.

119 **Methods**

120 *Study population*

121 This study was part of a research entity related to the Finnish Schools' on the Move program
122 focusing on physical activity and wellbeing among children and adolescents.²⁰ A longitudinal
123 observational study was conducted between January 2013 and June 2015. A total of 1778
124 students from nine Finnish schools were invited to participate. Out of these, 970 students

125 (53% girls) provided signed written consent with their main caregiver and participated in the
126 study. After excluding students with possible confounding factors at baseline (such as
127 guardian-reported existing chronic illnesses or disorders, injuries, existing multisite pain, and
128 more than 50% of missing data), the final sample consisted of 410 apparently healthy
129 participants (57% girls) (Figure 1).

130

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

136 *Outcome*

137 Pain incidence was considered the new onset of multisite pain at any time point during the 2-
138 year follow-up. Pain symptoms were screened four times after baseline with a structured
139 questionnaire at six-month intervals: “How often you have had symptoms in the last three
140 months (in body parts A-I in the pictures below)? Mark the appropriate option. Headache
141 (A), Neck and shoulder pain / ache (B), Upper extremities pain / ache (C), Chest pain / ache
142 (D), Upper back pain / ache (E), Low back pain / ache (F), Stomach ache (G), Buttocks pain /
143 ache (H), Lower extremities pain o/ ache (I)”.²¹ The question was supported by an illustration
144 of the described body parts. The answering options were: “Almost daily, More than once a
145 week, About once a week, About once a month, Seldom or never.” Students also reported if
146 the pain originated from trauma: ”Have you injured any of the above-mentioned and pictured
147 pain areas during the previous 3 months (for example, fallen, stumbled, breached during
148 sport, etc.)?” The answer options were “yes” or “no” and provided additional information

149 related to the injured body area.

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

157 *Predictive variables*

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

163

164 Secondly, we used baseline physical fitness measurements belonging to the Finnish national
165 'Move!' monitoring and surveillance system for health-related fitness²³ (nine baseline
166 variables described in detail elsewhere)²⁴. Annually, approximately 100,000 children and
167 adolescents (approximately 96% of the relevant age groups) participate in 'Move!' creating a
168 unique database for health-enhancing policies.¹⁸

169

170 Thirdly, a data-driven approach was used with the whole available data (392 baseline
171 variables) to explore potential novel predictors of multisite pain incidence. Data included

172 extensive information on students' self-reported, objectively measured, and device-based
173 demographics, physical and psychosocial characteristics, and physical activity.

174 *Analytical procedures*

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

179

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

184

185 10-fold cross-validation (CV) was employed for model assessment. During CV, the data for
186 each prediction task was divided into 10 subsamples called folds. Nine of these folds, 90% of
187 the whole data set, were used as the training data to fit the RF model, while one fold, 10% of
188 the data, was used as the validation data. This procedure was repeated ten times in a rotating
189 manner, where eventually all the folds had been employed for training and validation. Thus,
190 all the presented results are based on ten separate data-driven prediction models.

191

192 For each of the 10 CV folds, the trained model was employed to predict the out-of-bag
193 (OOB) observations i.e., those observations which were not utilized during the training of
194 each tree, and the validation portion of the data. The main metric recorded was the area the
195 under receiver operating characteristic curve (AUC). T-tests were performed in MATLAB
196 for the OOB and validation data AUC results to determine if the means of the CV folds were

197 significantly ($p < 0.05$) above the random level of 0.5. Further analyses regarding the
198 predictive power of each variable were conducted only in those cases where AUC 95%
199 Confidence Interval did not violate the 0.5 threshold.

200

201 RF requires choosing several hyperparameters i.e., options that define the model creation. F-
202 measure for the OOB observations was used as a target during Bayesian optimization,²⁷
203 where several hyperparameters of the RF model were chosen in an automated fashion. Please
204 see Supplementary methods in Supplement 1 for further information on the target measure,
205 the hyperparameters, and other details concerning the RF model.

206

207 The contribution of each variable to prediction was estimated using the OOB observations by
208 a permutation importance measure. A baseline result for the model in each CV fold was the
209 accuracy of the model with the original data. To estimate the contribution of each variable,
210 the values of the variables were permuted randomly. The procedure was repeated for all the
211 variables separately, and the accuracy of the model with permutations was recorded for each
212 variable. The accuracy obtained using the permuted variable was then subtracted from the
213 baseline accuracy. The final permutation importance estimate for each variable was the mean
214 of accuracy change for the 10 CV folds. T-tests were employed also for the importance
215 estimates. If the change was significantly ($p < 0.05$) over zero, the variable was seen as having
216 predictive power. Furthermore, if the mean change was near zero or negative, the variable did
217 not have importance in the prediction. MATLAB's predict function in the TreeBagger class
218 was utilized to calculate the OOB predictions on the trained model. This function computed
219 the weighted average of the class posterior probabilities over the trees.

220

221 In further sensitivity analyses the class imbalance was considered, meaning that there are
222 considerably less observations in diffuse idiopathic pain class, is a challenge in all explored
223 settings. This issue was approached during the modelling in two separate ways. Firstly, in the
224 RF model by changing the default cost matrix of misclassification. Cost of misclassifying
225 true pain class observations to no pain class (false negative classification) was increased to 2,
226 while the other misclassification (false positive) was left to its default value 1.

227

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

236

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

243 **Results**

244 At baseline, headache (22.5%, 30.4%) and neck and shoulder pain (13.5%, 18.5%) were the
245 most prevalent pain symptoms among boys and girls, respectively (Table 1). Sixteen percent

246 of boys and 28.1% of girls experienced multisite pain incidents in the whole body area during
247 the 2-year follow-up. Multisite pain incidence in the musculoskeletal area was 9.6% and
248 15.3% in boys and girls, respectively (Table 2).

249

250 The ability of the machine learning approach to predict whole body multisite pain incidence
251 reached an AUC 0.54 (95% Confidence Interval 0.49 to 0.58) for boys and 0.65 (0.63 to
252 0.67) for girls (Table 2). The prediction ability for multisite musculoskeletal pain incidence
253 was AUC 0.65 (0.62 to 0.68) in boys and 0.51 (0.48 to 0.54) in girls.

254

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

265

266 In boys, a total of 13 variables out of 48 showed predictive power for multisite
267 musculoskeletal pain incidence. The top ten predictors for pain incidence included higher
268 school strain, lower school enjoyment, a higher participation rate in sports competitions or
269 matches, lower amounts of continuous device-measured sedentary time, better muscular
270 fitness measured with the number of push-ups conducted within one minute, a lower body

271 mass index, more active participation in physical activity clubs in school, a later bedtime on
272 schooldays, lower total sedentary time, and parents' higher willingness to help with
273 schoolwork (Figure 3).

274

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

286 **Discussion**

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

295 increased risk for multisite pain and indicate the paradoxical nature of some variables,
296 especially in girls.

297

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

305

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

313

314 Previous studies have shown that prediction of multisite pain incidence is demanding,¹³ and
315 the isolated correlates have modest effect sizes.⁸ In this current study, the ML approach was
316 able to predict pain incidence above the random level, however remaining under clinical
317 relevance (AUC <0.7).³⁰ Through the ML approach, we found various predictors for multisite
318 pain incidence, reflecting the previously reported physical, lifestyle, and psychosocial
319 correlates,^{1,2,12-15} and complementing these findings by illustrating the risky variables in a

320 holistic framework alongside the paradoxical nature of some variables. For example, with
321 whole body multisite pain incidence among girls, indicators of both lower and higher
322 psychosocial wellbeing (e.g. lower life enjoyment vs. higher school enjoyment), low and high
323 physical activity (lower amount of moderate-to-vigorous physical activity during leisure time
324 vs. more days with physical activity for at least 60 min per day), lower body mass index, and
325 obesity or overweight classification were identified as risk factors. These findings illustrate
326 that risk factors, especially for whole body multisite pain incidence in girls are complex,
327 associations are not linear, and individuals with both healthy and unhealthy lifestyles,
328 favorable or unfavorable psychosocial status might develop multisite pain in the future. In
329 boys, findings indicated more consistently that poorer psychosocial wellbeing, higher
330 physical activity, a leaner body, and better physical fitness predict multisite musculoskeletal
331 pain incidence and support the acknowledgment of overall wellbeing and health-enhancing
332 physical activity practices among boys to prevent musculoskeletal pain.

333

334 The strengths of this study were the novel application of ML in pain prediction, the
335 longitudinal study design, and the extensiveness of predictors. The ML approach
336 considerably extends pain research and provides potential avenues for screening and
337 modeling complex phenomena in the future. The data was however limited by information
338 (no data on current medication) and cases (e.g., <66 cases in the data set) with possibly
339 influencing the generalizability of the findings. ML explores patterns in the data and does not
340 explain underlying mechanisms or causality. The multicollinearity of the variables might
341 affect the interpretation of variables with similar phenomenal origins. Self-reported data may
342 suffer from recall bias, although the reliability of the utilized questionnaire has shown to be
343 reasonable.²¹

344

345 In conclusion, these novel findings highlight the multifaceted predictors of multisite pain
346 incidence in adolescents and support the adoption of holistic and multidisciplinary prevention
347 approaches in the future.

348 **Statements**

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

352 **Data sharing**

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

355 **Declaration of interests**

356 We declare no competing interests

357 **References**

- 358 1. King S, Chambers CT, Huguet A, et al. The epidemiology of chronic pain in children and
359 adolescents revisited: A systematic review. *Pain*. 2011;152(12):2729-2738.
360 doi:10.1016/j.pain.2011.07.016
- 361 2. Gobina I, Villberg J, Välimaa R, et al. Prevalence of self-reported chronic pain among
362 adolescents: Evidence from 42 countries and regions. *Eur J Pain*. 2019;23(2):316-326.
363 doi:10.1002/ejp.1306
- 364 3. Hoftun GB, Romundstad PR, Zwart JA, Rygg M. Chronic idiopathic pain in adolescence –
365 high prevalence and disability: The young HUNT study 2008. *Pain*. 2011;152(10):2259-
366 2266. doi:10.1016/j.pain.2011.05.007
- 367 4. Kujala UM, Taimela S, Viljanen T. Leisure physical activity and various pain symptoms
368 among adolescents. *Br J Sports Med*. 1999;33(5):325-328. doi:10.1136/bjsm.33.5.325
- 369 5. López-Solà M, Suñol M, Timmers I. Brain predictors of multisite pain onset in children.
370 *Pain*. 2022;163(4):e502-e503. doi:10.1097/j.pain.0000000000002430
- 371 6. Holden S, Rathleff MS, Roos EM, Jensen MB, Pourbordbari N, Graven-Nielsen T. Pain
372 patterns during adolescence can be grouped into four pain classes with distinct profiles:
373 A study on a population based cohort of 2953 adolescents. *Eur J Pain*. 2018;22(4):793-
374 799. doi:10.1002/ejp.1165
- 375 7. Gauntlett-Gilbert J, Eccleston C. Disability in adolescents with chronic pain: Patterns and
376 predictors across different domains of functioning. *Pain*. 2007;131(1):132-141.
377 doi:10.1016/j.pain.2006.12.021
- 378 8. Auvinen J, Eskola PJ, Ohtonen HR, et al. Long-term adolescent multi-site musculoskeletal
379 pain is associated with psychological distress and anxiety. *J Psychosom Res*.
380 2017;93:28-32. doi:10.1016/j.jpsychores.2016.12.006
- 381 9. Lucas R, Brandão M, Gorito V, Talih M. Refining the prediction of multisite pain in 13-
382 year-old boys and girls by using parent-reported pain experiences in the first decade of
383 life. *Eur J Pain*. 2022;26(3):695-708. doi:10.1002/ejp.1898
- 384 10. Kamaleri Y, Natvig B, Ihlebaek CM, Benth JS, Bruusgaard D. Change in the number of
385 musculoskeletal pain sites: A 14-year prospective study. *Pain*. 2009;141(1):25-30.
386 doi:10.1016/j.pain.2008.09.013
- 387 11. Kamaleri Y, Natvig B, Ihlebaek CM, Bruusgaard D. Does the number of
388 musculoskeletal pain sites predict work disability? A 14-year prospective study. *Eur J*
389 *Pain*. 2009;13(4):426-430. doi:10.1016/j.ejpain.2008.05.009
- 390 12. Hoftun GB, Romundstad PR, Rygg M. Factors Associated With Adolescent Chronic
391 Non-Specific Pain, Chronic Multisite Pain, and Chronic Pain With High Disability: The
392 Young–HUNT Study 2008. *J Pain*. 2012;13(9):874-883.
393 doi:10.1016/j.jpain.2012.06.001

- 394 13. Paananen MV, Taimela SP, Auvinen JP, et al. Risk factors for persistence of multiple
395 musculoskeletal pains in adolescence: A 2-year follow-up study. *Eur J Pain*.
396 2010;14(10):1026-1032. doi:10.1016/j.ejpain.2010.03.011
- 397 14. Guddal MH, Stensland SØ, Småstuen MC, Johnsen MB, Zwart JA, Storheim K.
398 Physical Activity Level and Sport Participation in Relation to Musculoskeletal Pain in a
399 Population-Based Study of Adolescents: The Young-HUNT Study. *Orthop J Sports*
400 *Med*. 2017;5(1):232596711668554. doi:10.1177/2325967116685543
- 401 15. Pirnes KP, Kallio J, Hakonen H, Hautala A, Häkkinen AH, Tammelin T. Physical
402 activity, screen time and the incidence of neck and shoulder pain in school-aged
403 children. *Sci Rep*. 2022;12(1):10635. doi:10.1038/s41598-022-14612-0
- 404 16. Plowman S, Meredith M. *Fitnessgram/Activitygram ReferenceGuide*. 4th ed. The
405 Cooper Institute; 2013.
- 406 17. Pirnes KP, Kallio JJ, Hakonen HJ, et al. Physical fitness characteristics and neck and
407 shoulder pain incidence in school-aged children—A 2-year follow-up. *Health Sci Rep*.
408 2022;5(6). doi:10.1002/hsr2.852
- 409 18. Joensuu L, Csányi T, Huhtiniemi M, et al. *How to Design and Establish a National*
410 *School-Based Physical Fitness Monitoring and Surveillance System for Children and*
411 *Adolescents: The Ten-Step Approach Recommended by the FitBack Network*. Open
412 Science Framework; 2023. doi:10.31219/osf.io/zsnju
- 413 19. Gevaert AB, Adams V, Bahls M, et al. Towards a personalised approach in exercise-
414 based cardiovascular rehabilitation: How can translational research help? A ‘call to
415 action’ from the Section on Secondary Prevention and Cardiac Rehabilitation of the
416 European Association of Preventive Cardiology. *Eur J Prev Cardiol*. 2020;27(13):1369-
417 1385. doi:10.1177/2047487319877716
- 418 20. Blom A, Tammelin T, Laine K, Tolonen H. Bright spots, physical activity investments
419 that work: the Finnish Schools on the Move programme. *Br J Sports Med*.
420 2018;52(13):820-822. doi:10.1136/bjsports-2017-097711
- 421 21. Pirnes KP, Kallio J, Siekkinen K, Hakonen H, Häkkinen A, Tammelin T. Test-retest
422 repeatability of questionnaire for pain symptoms for school children aged 10–15 years.
423 *Scand J Pain*. 2019;19(3):575-582. doi:10.1515/sjpain-2018-0338
- 424 22. Joensuu L, Rautiainen I, Äyrämö S, et al. Precision exercise medicine: predicting
425 unfavourable status and development in the 20-m shuttle run test performance in
426 adolescence with machine learning. *BMJ Open Sport Exerc Med*. 2021;7(2):e001053.
427 doi:10.1136/bmjsem-2021-001053
- 428 23. Move! <https://www.oph.fi/en/move>
- 429 24. Joensuu L, Syväoja H, Kallio J, Kulmala J, Kujala UM, Tammelin TH. Objectively
430 measured physical activity, body composition and physical fitness: Cross-sectional
431 associations in 9- to 15-year-old children. *Eur J Sport Sci*. 2018;18(6):882-892.
432 doi:10.1080/17461391.2018.1457081

- 433 25. Breiman L. Random Forests. *Mach Learn.* 2001;45(1):5-32.
434 doi:10.1023/A:1010933404324
- 435 26. Hastie T, Friedman J, Tibshirani R. *The Elements of Statistical Learning*. Springer New
436 York; 2001. doi:10.1007/978-0-387-21606-5
- 437 27. Snoek J, Larochelle H, Adams RP. Practical Bayesian Optimization of Machine
438 Learning Algorithms. In: Pereira F, Burges CJ, Bottou L, Weinberger KQ, eds.
439 *Advances in Neural Information Processing Systems*. Vol 25. Curran Associates, Inc.;
440 2012.
441 [https://proceedings.neurips.cc/paper_files/paper/2012/file/05311655a15b75fab86956663](https://proceedings.neurips.cc/paper_files/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf)
442 [e1819cd-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf)
- 443 28. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority
444 Over-sampling Technique. *J Artif Intell Res.* 2002;16:321-357. doi:10.1613/jair.953
- 445 29. *Horizon Europe Work Programme 2023-2024, 4. Health.*; 2022.
- 446 30. Hosmer D, Lemeshow S. Chapter 5. In: *Applied Logistic Regression*. 2nd ed. John Wiley
447 and Sons; 2000:160-164.
- 448

449 **Tables**

450 Table 1. Selected subject demographics at baseline by sex.

N=410	Boys	Girls
Characteristic		
N	175 (42.6 %)	235 (57.3 %)
Age (years)	12.5 (1.2)	12.5 (1.2)
Height (cm)	156.7 (10.9)	155.1 (9.6)
Weight (kg)	46.1 (12.2)	45.4 (10.1)
BMI (kg/m ²)	18.5 (3.2)	18.7 (3.0)
Pain at different body sites		
Musculoskeletal pain sites		
Neck/Shoulder (%)	23 (13.6 %)	42 (18.5 %)
Upper extremities (%)	7 (4.1 %)	10 (4.4 %)
Chest (%)	1 (0.6 %)	2 (0.9 %)
Upper back (%)	5 (3.0 %)	5 (2.2 %)
Low back (%)	8 (4.7 %)	8 (3.5 %)
Buttocks (%)	2 (1.2 %)	6 (2.6 %)
Lower extremities (%)	9 (5.3 %)	13 (5.7 %)
Other pain sites		
Head (%)	38 (22.5 %)	69 (30.4 %)
Abdominal (%)	18 (10.7 %)	36 (15.9 %)
Physical activity and fitness		
Moderate-to-vigorous physical activity (min/day)	58.3 (24.2)	47.4 (18.3)
Sedentary time (hours/day)	8.1 (1.3)	8.6 (1.1)
Physical fitness index (Move-index)	16.6 (4.1)	17.2 (3.6)

451 Values are means and standard deviations for continues variables and proportions of participants for others. BMI, Body mass index; Move-
 452 index, weighted sum of the Finnish national Move! monitoring system's fitness items.

453

454

455 Table 2. Prediction ability of machine learning for multisite pain incidence among
456 adolescents.

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

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

458

459 **Figure legends**

460 Figure 1. Flow chart of the exclusion process

461 Figure 2. Permutation importance estimates for girls in the *selected set* for *all sites* (AUC
462 0.65). Red panel, risk factors; PA, physical activity; Counts, accelerometer total activity
463 counts

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

Original data

Whole body

Musculoskeletal locations

N=970
Boys N=462
Girls N=508

Exclusion criteria

1) Multisite pain reported at baseline.

Whole body

Excluded: 146 (67 boys, 79 girls)

Musculoskeletal locations

Excluded: 71 (38 boys, 33 girls)

2) A known existing disorder at baseline.

Excluded: 50 (30 boys, 20 girls)

Excluded: 58 (32 boys, 26 girls)

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

Excluded: 346 (181 boys, 165 girls)

Excluded: 373 (189 boys, 184 girls)

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.

Excluded: 10 (5 boys, 5 girls)

Excluded: 28 (2 boys, 26 girls)

Forming of the response variable for remaining students

Forming of the response variable for remaining students

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

N=410 (larger set)

Boys N=175, where multisite pain observations: 28
Girls N=235, where multisite pain observations: 66

N=413 (larger set)

Boys N=178, where multisite pain observations: 17
Girls N=235, where multisite pain observations: 35

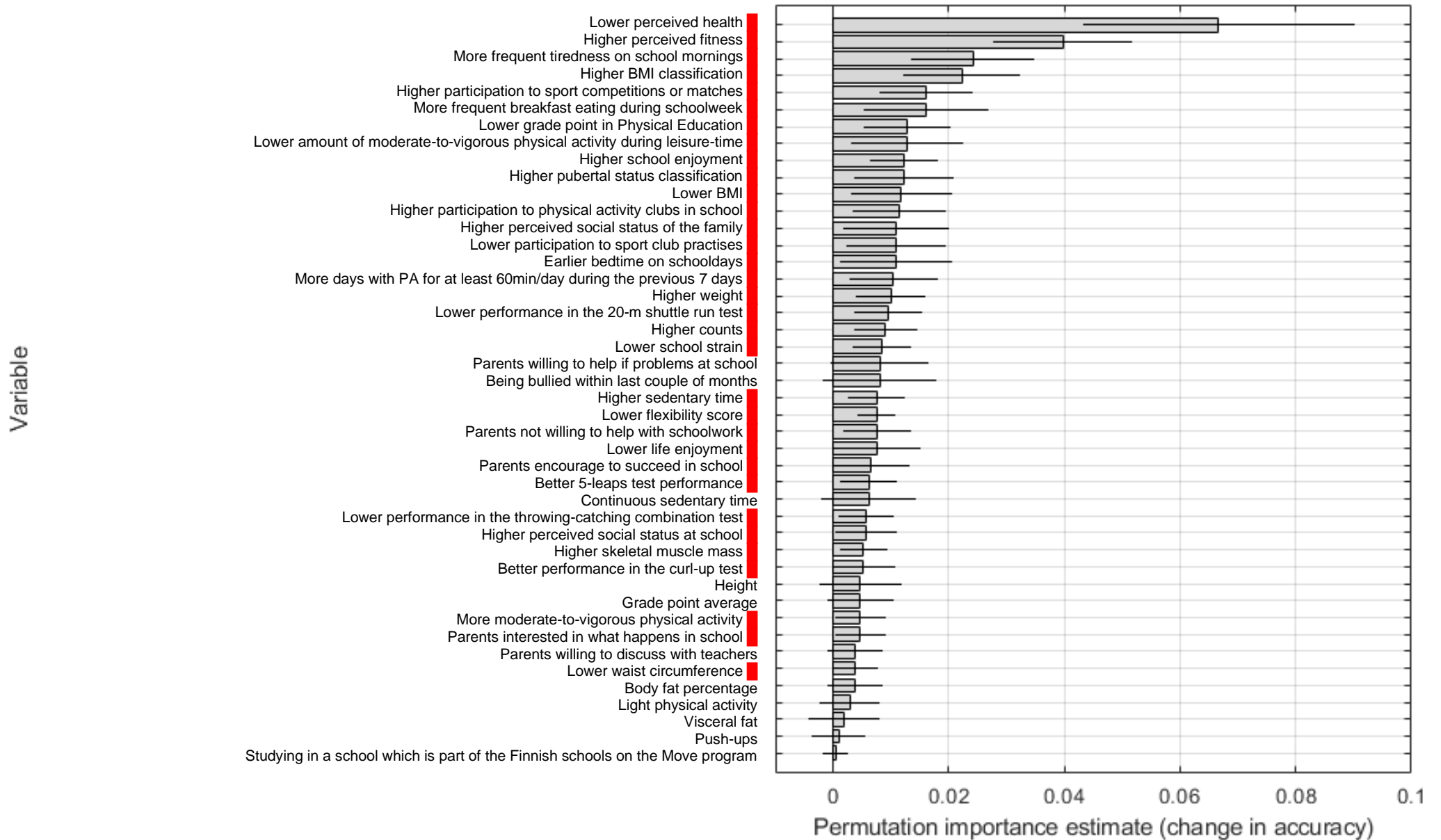


Figure 2. Permutation importance estimates for girls in the *selected set* for *all sites* (AUC 0.65). Red panel, risk factors; PA, physical activity; Counts, accelerometer total activity counts

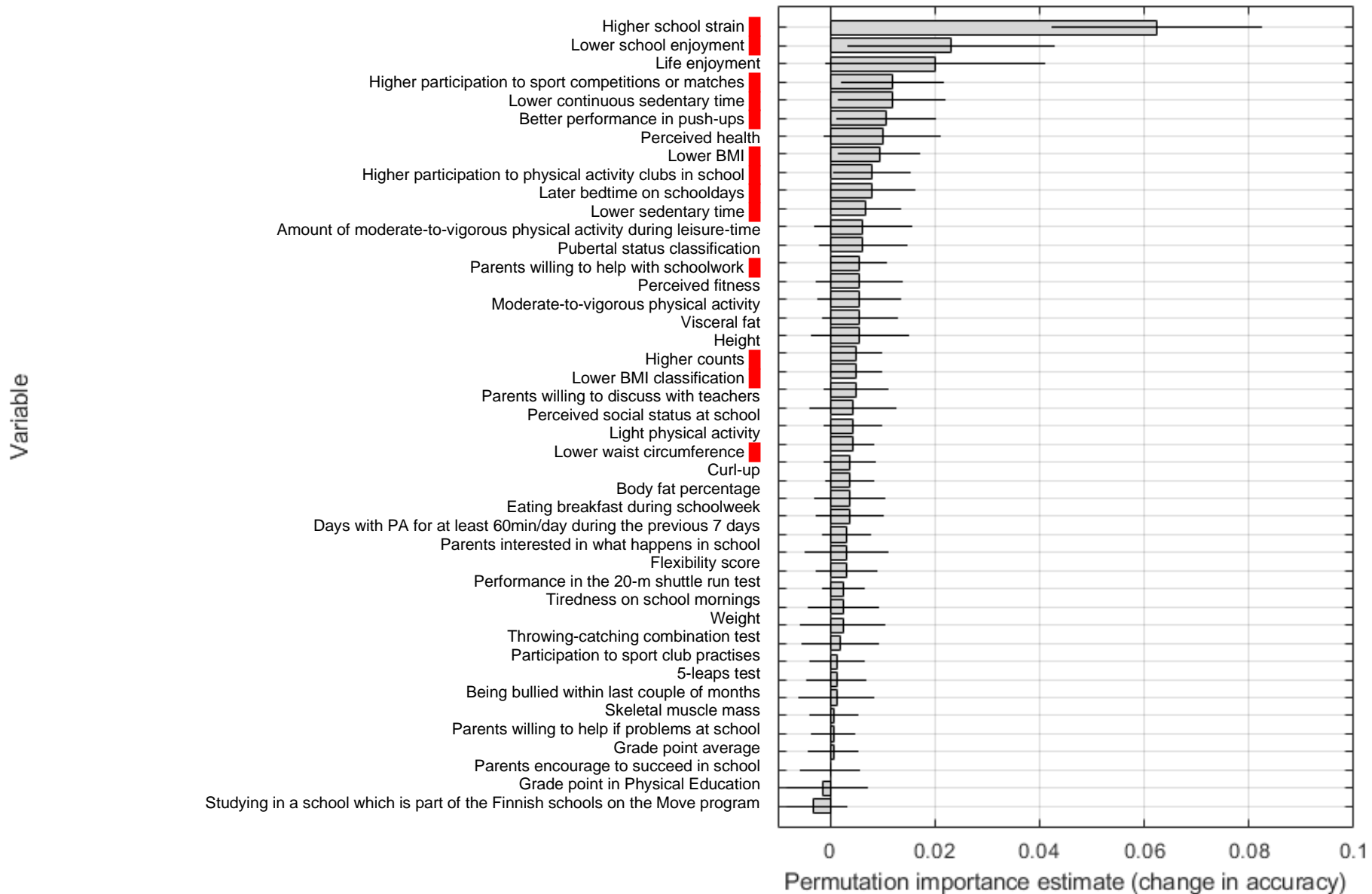


Figure 3. Permutation importance estimates for boys in the *selected set* for *musculoskeletal sites* (AUC 0.65). Red panel, risk factors; PA, physical activity; Counts, accelerometer total activity counts