

## The urban physical exposome and leisure-time physical activity in early midlife: a FinnTwin12 study

Zhiyang Wang (汪之洋)<sup>1</sup>, Sari Aaltonen<sup>1</sup>, Roos Teeuwen<sup>2</sup>, Vasileios Miliadis<sup>2</sup>, Carmen Peuters<sup>3,4,5</sup>, Bruno Raimbault<sup>3,4,5</sup>, Teemu Palviainen<sup>1</sup>, Erin Lumpe<sup>6</sup>, Danielle Dick<sup>7</sup>, Jessica E. Salvatore<sup>7</sup>, Maria Foraster<sup>8</sup>, Payam Dadvand<sup>3,4,5</sup>, Jordi Júlvez<sup>3,9</sup>, Achilleas Psyllidis<sup>2</sup>, Irene van Kamp<sup>10</sup>, Jaakko Kaprio<sup>1\*</sup>

<sup>1</sup> Institute for Molecular Medicine Finland, Helsinki Institute of Life Science, University of Helsinki, Helsinki, Finland

<sup>2</sup> Department of Sustainable Design Engineering, Delft University of Technology, Delft, the Netherlands

<sup>3</sup> ISGlobal, Barcelona Biomedical Research Park (PRBB), C/ Doctor Aiguader 88, Barcelona, Spain

<sup>4</sup> Universitat Pompeu Fabra (UPF), C/ Doctor Aiguader 80, Barcelona, Spain

<sup>5</sup> CIBER Epidemiología y Salud Pública (CIBERESP), Instituto de Salud Carlos III, c/ Monforte de Lemos 3-5, Madrid, Spain

<sup>6</sup> Department of Psychology, Rutgers University, New Brunswick, New Jersey, USA

<sup>7</sup> Department of Psychiatry, Rutgers University, Piscataway, New Jersey, USA

<sup>8</sup> PHAGEX Research Group, Blanquerna School of Health Science, Universitat Ramon Llull (URL), Barcelona, Spain

<sup>9</sup> Clinical and Epidemiological Neuroscience (NeuroÈpia), Institut d'Investigació Sanitària Pere Virgili (IISPV), Reus, Spain

<sup>10</sup> National Institute for Public Health and the Environment, the Netherlands

\* Corresponding author:

Jaakko Kaprio: [jaakko.kaprio@helsinki.fi](mailto:jaakko.kaprio@helsinki.fi); +358-503715419; address: Institute for Molecular Medicine, University of Helsinki, PL 20 (Tukholmankatu 8), FI-00014, Helsinki, Finland

1 **Abstract**

2 Leisure-time physical activity is beneficial for health and is associated with various urban characteristics.  
3 Using the exposome framework, the totality of the environment, this study investigated how urban physical  
4 environments were associated with leisure-time physical activity during early midlife. A total of 394  
5 participants (mean age: 37, range 34-40) were included from the FinnTwin12 cohort residing in five major  
6 Finnish cities in 2020. We comprehensively curated 145 urban physical exposures at residential addresses of  
7 participants and measured three leisure-time physical activity measures: (1) total leisure-time physical  
8 activity (total LTPA) and its sub-domains (2) leisure-time physical activity without commuting activity  
9 (LTPA) and (3) commuting activity. Using K-prototypes cluster analysis, we identified three urban clusters:  
10 "original city center," "new city center," and "suburban". Results from adjusted linear regression models  
11 showed that participants in the "suburban" cluster had lower levels of total LTPA (beta: -0.13, 95% CI: -0.23,  
12 -0.03) and LTPA (beta: -0.17, 95% CI: -0.28, -0.05), compared to those in the "original city center" cluster.  
13 The eXtreme Gradient Boosting models ranked exposures related to greenspaces, pocket parks, and road  
14 junctions as the top important factors influencing outcomes, and their relationships with outcomes were  
15 largely non-linear. More road junctions and more pocket parks correlated with higher total LTPA and LTPA.  
16 When the all-year normalized difference vegetation index within a 500 m buffer fell below 0.4, it correlated  
17 with higher levels of total LTPA, whereas above 0.4, it correlated with lower levels. To conclude, our  
18 findings revealed a positive correlation between urbanicity and physical activity in Finnish cities and  
19 decomposed this complexity into crucial determinants. Importance rankings and nonlinear patterns offer  
20 valuable insights for future policies and projects targeting physical inactivity.

21

22 **Keywords**

23 exercise, machine learning, urbanization

## 24 **1 Introduction**

25 Regular physical activity has been widely demonstrated to prevent multiple non-communicable diseases and  
26 reduce the risk of premature death<sup>1</sup>. The economic and health burden arising from physical inactivity is  
27 substantial and continually rising, costing public health care systems an estimated USD 47.6 billion globally  
28 every year<sup>2</sup>. Since previous studies show a strong contribution of environmental factors to physical activity<sup>3-5</sup>,  
29 interventions targeting the environment may be a good entry point to promote physical activity.

30 Urbanization stands as a transformative trend, with more than half the world's population currently residing  
31 in urban areas<sup>6</sup>. Many reviews have summarized the salient link between the urban environment and physical  
32 activity<sup>7-9</sup>. The exposome offers a theoretical framework with an umbrella perspective to depict the totality  
33 of the environment that people experience<sup>10</sup> and examines health effects from the real-world urban  
34 environment, of which the urban physical component plays an important role. The exposome studies have  
35 the potential to unveil more comprehensive non-genetic predictors through large-scale characterization of the  
36 environment. Gorman et al. have outlined the bidirectional effects between the exposome and physical  
37 activity but pointed out the uncertainty in mechanism and interactions<sup>11</sup>. The urban physical exposome is  
38 ubiquitous and multifaceted, which makes it a complex entity to study.

39 Every environmental factor contributes to this complex totality of exposures, and no factor is isolated. Urban  
40 regeneration projects are a good example, usually designed to improve public health by implementing  
41 structural and risk-minimizing solutions. They often yield collateral effects on other aspects, such as bringing  
42 economic, social, and cultural benefits, within the city's complex system<sup>12-14</sup>. For example, an urban riverside  
43 park regeneration project in Barcelona, Spain was estimated to attract over five thousand adult users daily to  
44 perform different types of physical activity<sup>15</sup>. Beyond the project's basic objectives, an open-air museum will  
45 be built there, transforming social and built environments. Nowadays, regeneration projects around the world  
46 are often multi-component and intersectoral. In another Barcelona regeneration program, aiming to improve  
47 living conditions in the most disadvantaged neighborhoods (involving, for example, social services, green  
48 spaces, and household support), researchers found that the neighborhood with a bigger project budget was  
49 associated with a higher frequency of physical activity among residents<sup>16</sup>. Previous studies relying on single  
50 exposures or limited sets were relatively inadequate to depict the broader urban environment and its health  
51 effects.

52 In this study, we aimed to comprehensively study the impact of the urban physical component of the  
53 exposome on the level of leisure-time physical activity during early midlife through two objectives: 1)  
54 clustering people with heterogeneous urban environments and comparing physical activity levels between  
55 clusters, as well as sex-specific effect and 2) ranking urban physical exposures by importance on leisure-time  
56 physical activity, examining non-linear relationships, and detecting pairwise interactions between exposures.

## 57 **2 Material and methods**

58 The flow chart of this study is presented in Figure 1.

## 59 **2.1 Participants**

60 Participants were from the FinnTwin12 cohort, which is a nationwide prospective cohort of all Finnish twins  
61 born between 1983 and 1987. Briefly, at baseline (1994-1999), 5522 12-year-old twins were invited to  
62 participate and 87% of them agreed to take part. There were four follow-ups: age 14, age 17, young  
63 adulthood (mean age 24), and early midlife (mean age 37), with retention rates of 92%, 75%, 66%, and 41%,  
64 respectively. A recent study has detailed the latest follow-up of the cohort<sup>17</sup>. In this study, we included  
65 individual twins who lived in five large cities of Finland, namely Helsinki, Tampere, Espoo, Oulu, or  
66 Jyväskylä, in 2020. A

## 67 **2.2 Measures**

### 68 2.2.1 Leisure-time physical activity

69 Our study focuses on early midlife leisure-time physical activity, which is performed at the person's  
70 discretion along with essential daily living activity or work-related tasks<sup>18</sup>. This type of physical activity is  
71 considered one of the most effective ways to increase overall physical activity levels<sup>19</sup>. It was measured  
72 through structured and validated questions on the frequency, mean duration, and mean intensity of  
73 participants' leisure-time physical activity sessions, as well as a question on their commuting activity<sup>20,21</sup>.  
74 Based on these structured questions, we quantified mean metabolic equivalent of task (MET) hours per day,  
75 which expresses the energy cost of physical activities in the form of the resting metabolic rate<sup>22</sup>. Its  
76 calculation formula was the following: physical activity frequency  $\times$  physical activity duration  $\times$  physical  
77 activity intensity<sup>23</sup>. The MET values for activity intensity were: 4 for intensity corresponding to walking, 6  
78 for intensity corresponding to vigorous walking to jogging, 10 for intensity corresponding to jogging, and 13  
79 for the intensity corresponding to running. All types of leisure time physical activities were considered when  
80 MET hours per day were calculated. We assumed that commuting activity was done on 5 days per week and  
81 on the intensity of walking. The questions are listed in Supplemental Note 1.

82 The primary measure, *total leisure-time physical activity (total LTPA)*, was the sum of leisure-time and  
83 commuting-related physical activities. These two sub-domains were secondary measures: 1) *leisure-time*  
84 *physical activity without commuting activity (LTPA)* and 2) *commuting activity*. Participants with over mean  
85 45 MET hours/day of total LTPA were identified as outliers and removed. This threshold corresponds to, for  
86 example, approximately 3.5 hours of fast running daily, which is likely unrealistic<sup>24</sup>. The distributions of all  
87 three measures are shown in Supplemental Figure 1, and due to the skewness, we log-transformed them.

### 88 2.2.2 Urban physical exposome

89 We assigned 145 indicators of urban physical exposures to the residential address of each study participant.  
90 Detailed description and summary statistics of these indicators are presented in Supplemental Table 1. The

91 urban physical exposome set comprehensively depicted the urban environment including aspects such as  
92 traffic, streets, land use, green (i.e. parks, forests, and fields) and blue (i.e. lakes and seas) spaces, and so on.  
93 The computing and enriching process was on the geocode level and derived from multiple open sources,  
94 which is described in Supplemental Note 2 and elsewhere<sup>25-28</sup>. Most urban physical exposures were  
95 measured or modelled in 2018 and 2023, and the percentage of area covered by trees was measured in 2015.  
96 We used the residential history provided by the Digital and Population Data Services Agency, Finland  
97 between birth and 2021 to merge the urban physical exposures by EUREF-FIN geocodes. Exposures  
98 available in 2018 or 2015 were merged with residential addresses of participants in 2018 or 2015, while  
99 exposures available in 2023 were merged with residential addresses in 2020.

### 100 2.2.3 Other measures

101 Five sociodemographic variables were identified a priori: sex (categorical, female vs. male), age (continuous,  
102 year), work (categorical, not working or other situation vs. currently work), education (categorical, post-  
103 secondary or lower vs. bachelor/equivalent or above), and marital status (categorical, married, steady  
104 relationship, or living together vs. no). The latter three were self-reported at the early midlife follow-up. Sex  
105 was based on the register information obtained when the cohort was established, while age was computed  
106 from the difference between the date of response and the date of birth. There were another three behavioral  
107 variables: illicit substance use (categorical, never vs. at least once), ever smoker (smoked over 100 cigarettes  
108 lifetime) (categorical, no vs. yes), and alcohol drinking (categorical, monthly or less or even never vs. 2-4  
109 times a month or more), inquired also at the early midlife follow-up. Adult leisure-time physical activity was  
110 associated to most of the sociodemographic and behavioral variables, as shown in previous research<sup>29-31</sup>.

111 To depict the social environment, four neighborhood social variables at the postal code level were derived  
112 from Statistics Finland in 2018: the proportion of resident living alone (single household), of residents with  
113 the lowest education level, of residents with the lowest income quartile, and of unemployed residents. A  
114 neighborhood deprivation score was generated from the latter three social variables<sup>32</sup>. We first standardized  
115 the three variables to z-scores, and their mean value is the deprivation score. Using a median split, we then  
116 categorized neighborhoods where participants lived in 2018 into two levels: low- and high-deprived. Thus,  
117 there were two neighborhood social variables: the proportion of resident living alone and deprivation level,  
118 which were merged via residential history in 2018 too.

## 119 2.3 Analysis

### 120 2.3.1 Data processing

121 After excluding those people who did not have information on leisure-time physical activity,  
122 sociodemographic, behavioral, and neighborhood-level social variables, 394 twin individuals resident in  
123 these urban areas were included in this study. Given that there were only 44 twin pairs with both cotwins  
124 satisfying the inclusion criteria, we did not consider zygosity as a covariate and did not perform any pairwise

125 twin analysis. The distribution of sociodemographic and behavioral variables among included and excluded  
126 participants are presented in Supplemental Table 2, respectively. There were significant differences between  
127 included and excluded participants in education, illicit substance use, and alcohol drinking.

### 128 2.3.2 Clustering analysis

129 The k-prototypes cluster analysis was employed to distinguish distinct patterns in the urban environment. It  
130 combines dissimilarity measures from both k-means and -modes algorithms for mixed types of exposures,  
131 and has shown to have a good performance<sup>33,34</sup>. Continuous exposures were standardized by standard  
132 deviation (SD). All 145 urban physical exposures were included in the clustering algorithms. The Silhouette  
133 method was used to pre-specify the number of clusters<sup>35</sup>. One-step imputation within the algorithm was  
134 applied for missing values<sup>36</sup>. Since k-prototypes cluster analysis is sensitive to outliers, the principal  
135 component analysis of mixed data was conducted before. Participants whose first or second principal  
136 components fell outside the range of five standard deviations were identified as outliers<sup>37</sup>, as a practical way,  
137 and excluded from the cluster analysis; three participants were excluded.

138 Next, hierarchical linear regression was performed for the relationship between the urban cluster and leisure-  
139 time physical activity measures with three adjustment plans for covariates: 1) sociodemographic variables, 2)  
140 sociodemographic and behavioral variables, and 3) sociodemographic, behavioral, and neighborhood social  
141 variables. The cluster effect of sampling based on families of twin pairs was controlled by the robust  
142 standard error. We also performed the sex-stratified analysis.

### 143 2.3.3 Machine learning analysis

144 Before exploring the complexity within the urban environment via a pluralistic analysis platform,  
145 generalized linear regression models with the robust standard error were repeatedly performed between each  
146 leisure-time physical activity measure (total LTPA, LTPA, and commuting activity) and each urban physical  
147 exposure (missing values were imputed). The *a priori* significant threshold of 0.01 was used to identify  
148 noteworthy candidates. Dimensional reduction increases the model stability of subsequent analysis.

149 Then, we performed the eXtreme Gradient Boosting (XGBoost) model to assess the importance of urban  
150 physical exposures on each physical activity measure, uncover interactions, and identify nonlinear  
151 relationships<sup>38</sup>. It is an optimized distributed gradient boosting library designed for efficient and scalable  
152 training of machine learning models, with gradient-boosted decision trees algorithm<sup>38</sup>. The hyperparameters  
153 were tuned through the 5-fold cross-validation grid search<sup>39</sup>. The participants were randomly split into  
154 training and testing subsets in a ratio of 3:1. The model performance was evaluated by root-mean-square  
155 error (RMSE). All urban physical exposures, sociodemographic, behavioral, and neighborhood social  
156 variables were included in the model. After hyperparameter tuning, the model was repeated two additional  
157 times with different seeds for result robustness.

158 To increase model transparency, the SHapley Additive exPlanations (SHAP) value was used to interpret and  
159 visualize the results from the XGboost model, which features the exposures' importance on the outcome  
160 based on the cooperative game theory<sup>40</sup>. Its direction suggests the direction of impact on prediction, leading  
161 the model to predict either a higher or lower value of outcomes. Its magnitude is a measure of how strong the  
162 effect is. We quantified pairwise interaction SHAP values between included variables and summed their  
163 absolute value of all participants, with a high value indicating a strong interaction and synergistic effect<sup>41</sup>.  
164 Additionally, Group-Lasso INTERaction-NET was performed for interaction to compare with the XGBoost's  
165 result<sup>42</sup>.

#### 166 2.3.4 Sensitivity analysis

167 Due to missing values in urban physical exposures, we additionally performed sensitivity analyses of K-  
168 prototype cluster analysis and repeated generalized linear regression models between each urban physical  
169 exposure and each leisure-time physical activity measure, after removing participants with missing values  
170 (n=13).

### 171 3 Results

#### 172 3.1 Description of participants

173 Of the 394 included participants (mean age: 37, SD: 1.5) (Table 1), more individuals were female (55%).  
174 Altogether, 87%, 79%, and 75% of participants were employed, had at least bachelor-level education, and  
175 were married or in a stable relationship, respectively. In their early midlife, more than half of the participants  
176 drank alcohol at least 2-4 times a month (58%), but fewer had smoked over 100 cigarettes (45%) or had used  
177 illicit substances such as marijuana at least once (48%). Before log-transformation, the means of total LTPA,  
178 LTPA, and commuting activity (unit: MET hours/day) were 5.4 (SD: 4.7), 4.3 (SD: 4.4), and 1.1 (SD: 1.0),  
179 respectively. After log-transformation, Spearman correlations between total LTPA and LTPA, between total  
180 LTPA and commuting activity, and between LTPA and commuting activity were 0.9, 0.3, and 0.1,  
181 respectively

#### 182 3.2 Results from clustering and hierarchical regression

183 The Silhouette method identified the optimal number of clusters to be three (largest Silhouette index, total  
184 within-cluster sum of squares: 42323.69). Using the map of Helsinki and Espoo and the spatial layer of  
185 centers and shopping areas in 2019 from the community structure monitoring system, Finnish Environment  
186 Institute<sup>43</sup>, we visually classified Cluster 1, 2, and 3 as the "Original city center", "New city center", and  
187 "Suburban" clusters, respectively, based on the participants' residence in 2018, as the urban cluster variable  
188 (Figure 2).

189 After fully adjusting for sociodemographic, behavioral, and neighborhood social variables, compared to  
190 participants who lived in the "original city center" cluster, participants who lived in the "suburban" cluster

191 were associated with significantly lower log-transformed scores of total LTPA (beta: -0.13, 95% CI: -0.23, -  
192 0.03) and LTPA (beta: -0.17, 95% CI: -0.28, -0.05) (Table 2). The effect sizes did not change substantially  
193 after adjustment of sociodemographic variables only and adjustment of both sociodemographic and  
194 behavioral variables. Regardless of adjustment plans, there was no significant association between the urban  
195 cluster and commuting activity (Table 2). There was no significant difference in any outcome between  
196 participants who lived in the “suburban” and “new city center” clusters. The powers of full-adjusted models  
197 of total LTPA, LTPA, and commuting activity were all 1.0.

198 After stratifying the analyses based on sex, we observed that, in males, the result pattern and effect sizes  
199 were like the overall results between the urban cluster and total LTPA, while the association between the  
200 urban cluster and LTPA became null after full adjustment (Supplemental Table 3). However, in females,  
201 after additionally adjusting for behavioral variables only or for both behavioral and neighborhood social  
202 variables, no significant association of the urban cluster with total LTPA and LTPA was seen (Supplemental  
203 Table 3).

### 204 **3.3 Results from XGBoost**

205 Based on the repeated generalized linear regression, there were 25 urban physical exposures significantly  
206 associated with total LTPA and 24 with LTPA (Supplemental Table 4). No urban physical exposure met the  
207 threshold p-value of 0.01 for association with commuting activity (Supplemental Table 4), so there was no  
208 XGBoost analysis for it.

209 In the XGBoost model of total LTPA including all urban physical exposures, sociodemographic, behavioral,  
210 and neighborhood social variables, the top three important urban physical exposures were the count of any  
211 type of road junctions within a 500 m buffer (ints\_500), the total area of all interconnected pocket parks  
212 within an 800 m walking distance (sumarea\_pocketparks\_800), and the 5-years moving average of  
213 Normalized Difference Vegetation Index (NDVI), an indicator of general greenness, within a 500 m buffer  
214 around the home during whole year (ndvi\_5yrs\_all\_500) (Figure 3A). In dependence plots, SHAP values  
215 positively correlated with both the count of any type of road junctions within a 500 m buffer (Figure 3B) and  
216 the total area of all interconnected pocket parks within an 800 m walking distance (Figure 3C). When the two  
217 urban physical exposures were within a certain range, SHAP values remained constant, which this type of  
218 non-linearity made these two exposures look like threshold variables. When the count of any type of road  
219 junctions within a 500 m buffer was in the range of 1-40, 40-50, and over 50, the SHAP values were  
220 approximately -0.003, 0, and 0.003, respectively. When the total area of all interconnected pocket parks  
221 within an 800 m walking distance was in the range of 0-0.005, 0.005-0.01, and over 0.01 km<sup>2</sup>, the SHAP  
222 values were approximately -0.002, 0.001, and 0.004, respectively. In Figure 3D, the 5-years moving average  
223 of NDVI within a 500 m buffer during whole year also showed a pattern of a binary threshold variable.  
224 When it was below 0.4, the SHAP value was around 0.0012. When it was over 0.4, the SHAP value was  
225 around -0.0012.



226 In the XGBoost model of LTPA (Figure 3E), the most important urban physical exposures were the count of  
227 pocket parks within an 800 m walking distance (count\_pocketparks\_800), the total area of all interconnected  
228 pocket parks within an 800 m walking distance remained the second most important, and the count of any  
229 type of road junctions within a 500 m buffer became the third most important. The SHAP value revealed a  
230 switch in predictions from lower (negative SHAP values) to higher (positive SHAP values) log-transformed  
231 LTPA when the number (count) of pocket parks within an 800 m walking distance was more than two  
232 (Figure 3F). Interestingly, each count led to two different but close SHAP values. The patterns of the total  
233 area of all interconnected pocket parks within an 800 m walking distance (Figure 3G) and the count of any  
234 type of road junctions within a 500 m buffer (Figure 3H) were similar to the model of total LTPA.

235 Supplemental Figure 2 displays pairwise SHAP interaction values in the XGBoost model of total LTPA, and  
236 there was no pairwise interaction between urban physical exposures. Similarly, the XGBoost model of LTPA  
237 (Supplemental Figure 3) indicates slightly some interactions but with very low values ( $<0.001$ ). Group-Lasso  
238 INTERaction-NET models also did not capture any strong pairwise interaction for the two physical activity  
239 measure analyses.

240 For the XGBoost model of total LTPA, the RMSE is 0.27 in the training subset and 0.29 in the testing subset.  
241 Comparing the reported test with two extra tests, the importance rank varied, but the count of any type of  
242 road junctions within a 500 m buffer was always the most or second most important (Supplemental Table 5).  
243 For the XGBoost model of LTPA, the RMSE is 0.32 in both training and testing subsets. The importance  
244 rank also varied between reported results and two extra tests, but the most important urban physical exposure  
245 was the count of pocket parks within an 800 m walking distance, being always in the top two (Supplemental  
246 Table 6).

### 247 **3.4 Sensitivity analysis for missing value in urban physical exposures**

248 After excluding 13 participants with missing values in some urban physical exposures, the Silhouette method  
249 identified two clusters. In the following fully adjusted linear regression models, no significant differences in  
250 any of the physical activity measures were found between the clusters. Repeated generalized linear  
251 regression analyses revealed 25 urban physical exposures significantly associated with total LTPA,  
252 consistent with the analysis using imputed data, and 26 exposures with LTPA, two more than the analysis  
253 with imputed data. Still, no urban physical exposure reached the 0.01 P-value threshold for association with  
254 commuting activity.

## 255 **4 Discussion**

256 We used clustering analysis and XGBoost to simultaneously and comprehensively study the effect of 145  
257 urban physical exposures on leisure-time physical activity in 394 Finnish adults in their early midlife. Three  
258 clusters were identified: “original city center”, “new city center”, and “suburban”. We found people living in  
259 suburban areas had a lower level of physical activity in leisure time compared to those living in the original

260 city center. There was no difference between “original city center” and “new city center” clusters. The  
261 effects appeared more clearly in males, while behavioral and neighborhood social factors may account for  
262 the associations in females. XGBoost models revealed a complex relationship between the urban physical  
263 exposome and leisure-time physical activities, in which important exposures showed non-linearity and  
264 looked like threshold variables. Increased road junctions and more and bigger pocket parks correlated with  
265 higher levels of leisure-time physical activity. However, higher amounts of vegetation greenness (indicated  
266 by NDVI) were associated with low leisure-time physical activity levels. We did not find any considerable  
267 interaction between urban physical exposures contributing to leisure-time physical activities.

268 Previous research has documented the relationship between the degree of urbanization and physical activity  
269 but with inconsistent findings regarding the direction of effects. A cross-sectional study in Shanghai, China  
270 with 327 respondents (mean age: 40) similarly reported higher leisure-time physical activity among  
271 downtown residents compared to suburban dwellers, but in contrast to our results, significant results were  
272 also found for transportation activities<sup>44</sup>. Another Canadian study showed that the physical activity level was  
273 higher in urban than in suburban among adolescents from schools in lower socio-economic areas<sup>45</sup>.  
274 Nevertheless, a systematic review suggested that children and teenagers who live in suburban areas were  
275 more physically active than in rural and urban areas<sup>46</sup>, and, similar to the Shanghai study above, a nationwide  
276 study in China showed that rising urbanization correlates with longer commuting times among adults (mean  
277 age: 45)<sup>47</sup>. Sex-specific effects have also been also observed. In the US, only male adolescents living in  
278 urban areas engaged in more moderate-to-vigorous physical activity (MVPA) than those living in suburban  
279 areas<sup>48</sup>. Additionally, distinct patterns between sexes in the significance and direction of associations  
280 between urbanity level in different aspects and physical activity measures were noted in Mexico<sup>49</sup>.  
281 Socioeconomic status (SES) might explain the sex difference, as the association weakened to null in females  
282 after adjusting for behavioral and neighborhood social variables. Previous population studies have observed  
283 some interaction effects between sex and SES on physical activity<sup>50-52</sup>. The inconsistency between literature  
284 and our findings may be due to different population characteristics, sports cultures, country contexts, urban  
285 planning, or urbanicity definitions. Instead of a pre-definition of (sub)urban areas by governmental  
286 guidelines, we used an unsupervised data-driven clustering method to determine heterogeneous urban  
287 environments within urban areas reflecting real-life exposure modes and accounting for correlation, additive,  
288 and mixture effects<sup>53</sup>.

289 XGBoost models ranked the elements of pocket parks, road junctions, and greenspaces as strongly associated  
290 with leisure-time physical activities among early midlife adults. A natural experimental study in low-income  
291 American neighborhoods found increased leisure-time exercise among middle-aged residents after pocket  
292 parks were constructed<sup>54</sup>. Users of pocket parks, defined as living within a 0.5 mile (~800 m) radius, had  
293 higher exercise levels than traditional park users<sup>54</sup>. Researchers further summarized that pocket parks were  
294 cost-effective for promoting physical activity in inner-city areas<sup>54</sup>. A study in Chongqing, China, utilizing  
295 interviews on conceptual understanding of park images, revealed that the environmental characteristics of

296 pocket parks contributed to a restorative effect involving entertainment activities and relief<sup>55</sup>. Noteworthy, a  
297 recent Chinese study using Light Gradient-Boosting Machine model found that recreational facilities were  
298 the most important factor for walking behavior in old adults but the number of parks was the least important  
299 among 11 factors, highlighting the specific effect driven by the content inside parks or recreation areas<sup>56</sup>. For  
300 road junctions, a Finnish study found the density of intersections, defined as the junction of a minimum of  
301 three roads, was positively associated with the number of physical activity bouts and the level of moderate to  
302 vigorous physical activity among older adults<sup>57</sup>. Zang et al. used random forest models to identify the  
303 intersection density, as well as streetscape greenery, as the most important physical exposure contributing to  
304 light physical activity among older adults<sup>58</sup>. More intersections usually indicate a greater degree of  
305 connectivity, which creates a more convenient environment for people to walk or bike to their destinations.  
306 However, the relationship between street connectivity, involving the number of intersections, and physical  
307 activity in all age groups of adults varied across different buffer areas in urban environments, suggesting the  
308 complexity of urban living environments<sup>59</sup>. Where the association of greenspace with physical activity is  
309 relatively inconsistent<sup>60</sup>, our findings show an association in which surrounding greenness is positively  
310 associated with LTPA up until a threshold of 0.4 NDVI, with higher NDVI relating to lower LTPA. High  
311 levels of green space might reflect suburban living to some extent, and other greenspace indicators, such as  
312 accessibility, were not prominent. The relationship between greenspace and physical activity could be  
313 moderated by the level of urbanization<sup>60</sup>. Other studies have similar findings on the threshold effect. For  
314 example, the positive association of physical activity with multiple green space uses indicators reached to  
315 peak when indicators were within a 600 m buffer<sup>61</sup>. Zang et al. also found that streetscape greenery had a  
316 positive effect on light physical activity when it ranged from 0.12 to 0.15 point, corresponding to a low level  
317 of visible greenery<sup>58</sup>. Besides, another Chinese study also identified the 0.4 NDVI, corresponding to areas  
318 with sparse to moderate vegetation, as the turning point for its association with self-rated health among the  
319 old population<sup>62</sup>, and self-rated health closely correlated with physical activity<sup>63</sup>. This annotation added to  
320 current evidence has critical guidance on urban planning.

321 We selected different methods to depict the contour of association between the urban physical exposome and  
322 leisure-time physical activity, translating abstract characteristics into practical understanding. On one hand,  
323 clustering analysis has the advantage of providing insight into real-world scenarios and holding a high  
324 scalability to uncover hidden patterns. On the other hand, tree-based machine learning can be applied as a  
325 pluralistic analysis platform to synthesize evidence between a range of urban physical exposures and  
326 physical activity<sup>58,64,65</sup>. Comparing to conventional analyses, the XGBoost model enhances our assessments  
327 with several advantages: 1) unraveling nonlinear relationships through visualization, 2) disentangling  
328 complex interactions among multiple exposures, and 3) offering robust computation for multi-inference  
329 approaches<sup>66</sup>. By deepening the understanding of distinct and complex characteristics of the urban physical  
330 exposome, supported by detailed exposure profiling, policymakers can develop precise and cost-effective  
331 interventions and strategies to address the challenge of low physical activity levels.

332 Besides its strength, this study is not without limitations. First, the sample size was relatively small compared  
333 to other exposome studies. Although the sample size for K-prototype clustering (over 10 times the number of  
334 clusters) and subsequent regression seems to be adequate (but not for the sex-stratified analysis),  
335 inconsistency in additional tests highlights the need for a larger sample. Additionally, due to the complexity  
336 of the large-dimensional exposome set, the modest sample size made capturing relatively small interactions  
337 more challenging. Therefore, we should be cautious when interpreting results. Second, only participants from  
338 the five largest cities in Finland were included, limiting the generalizability. Besides, we did not include any  
339 participants living in rural areas. Not only the physical environment, but lifestyles may also differ between  
340 urban and rural areas. Therefore, the interpretation should be narrowed down to specific types of cities. Third,  
341 urban physical exposures were based on residential addresses, which overlook dynamic human behaviors  
342 outside the home, leading to measurement errors. In addition, the used residential geocodes corresponded to  
343 participants' residences in 2017, 2018, or 2020, without accounting for how long they lived at those  
344 addresses. Measurement errors could skew our identification of key determinants, as exposures with larger  
345 errors might show weaker associations and be classified as less influential, even if they are actually more  
346 important than those identified as most influential. More granular and accurate estimations of exposure and  
347 behavior could facilitate the exploration in the dynamic interaction between the environment and human  
348 behavior<sup>12</sup>. Fourth, some exposures were available in 2023 but merged with the address in 2020, posing a  
349 temporality issue. The relatively slow urban renewal and construction in Finland reduced the concern<sup>67</sup>. Fifth,  
350 missing values in exposures may introduce bias. Excluding participants with missing values altered the  
351 optimal number of clusters, while the number of significant associations between exposures and outcomes  
352 remained similar to the number based on imputed data. Given that only about 3% of participants had missing  
353 values, the effect is likely modest, but caution is still warranted. Sixth, leisure-time physical activity was  
354 self-reported. The device-based measurement of leisure-time physical activity would have been more  
355 accurate. However, the validity of leisure-time physical activity questions used in Finnish twins has been  
356 demonstrated<sup>20,21</sup>.

## 357 **5 Conclusion**

358 This study employed two analytical approaches to explore the intricate impact of the urban physical  
359 exposome on leisure-time physical activity in early midlife in Finland. Clustering analysis revealed three  
360 heterogeneous patterns of urban environments. Living in suburban areas was associated with lower levels of  
361 leisure-time physical activity than in original city center areas. XGBoost models identified pocket parks,  
362 road junctions, and greenspaces as influential factors with non-linear relationships, which behaved like  
363 threshold variables. Given limitations in sample size, generalizability, and measurement granularity, we call  
364 for further studies in other settings to replicate our analyses. We still advocate presenting the evidence to  
365 stakeholders and policymakers to develop tailored interventions on some urban features to achieve higher  
366 cost-effectiveness by focusing on the most influential determinants and their optimal ranges in addressing the  
367 challenge of the physically inactive lifestyle in our rapidly urbanizing world.

## 368 **Acknowledgments**

369 This research was partly funded by the European Union's Horizon 2020 research and innovation program  
370 under grant agreement No 874724 (Equal-Life). Equal-Life is part of the European Human Exposome  
371 Network. Data collection in FinnTwin12 has been supported by the National Institute on Alcohol Abuse and  
372 Alcoholism (grants AA-12502, AA-00145, and AA-09203 to Richard J. Rose, and AA015416 to Danielle  
373 Dick and Jessica Salvatore) and the Academy of Finland (grants 100499, 205585, 118555, 141054, 264146,  
374 308248, 312073, 336823, and 352792 to Jaakko Kaprio). Jaakko Kaprio acknowledges support by the  
375 Academy of Finland (grants 265240, 263278). ISGlobal acknowledges support from the grant CEX2018-  
376 000806-S funded by MCIN/AEI/10.13039/501100011033, and support from the Generalitat de Catalunya  
377 through the CERCA Program.

## 378 **Author's contribution**

379 M.F., P.D., J.J., A.P., I.v.K., and J.K. conceived the exposome framework. Z.W. developed the research  
380 question and designed the analysis and other authors commented to refine it. S.A., D.D., J.S., and J.K. led the  
381 FinnTwin12 cohort. R.T., V.M., B.R., and M.F. enriched urban physical exposures and T.P. managed the  
382 FinnTwin12 data. Z.W. performed the analysis and wrote the original draft. All authors reviewed the draft  
383 and approved for the submission.

## 384 **Competing interests**

385 The authors declare that they have no competing interests.

## 386 **Ethical requirement**

387 The ethics committee of the Department of Public Health of the University of Helsinki (Helsinki, Finland)  
388 and the Institutional Review Board of Indiana University (Bloomington, Indiana, USA) approved the  
389 FinnTwin12 study protocol from the start of the cohort. The ethical approval of the ethics committee of the  
390 Helsinki University Central Hospital District (HUS) is the most recent and covers the most recent data  
391 collection (early midlife) (HUS/2226/2021, dated September 22, 2021). All participants and their  
392 parents/legal guardians gave informed written consent to participate in the study. The authors assert that all  
393 procedures contributing to this work comply with the ethical standards of the relevant national and  
394 institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in  
395 2008.

## 396 **Data and Code Availability**

397 The FinnTwin12 data are not publicly available due to the restrictions of informed consent. However, the  
398 FinnTwin12 data are available through the Institute for Molecular Medicine Finland (FIMM) Data Access  
399 Committee (DAC) ([fimm-dac@helsinki.fi](mailto:fimm-dac@helsinki.fi)) for authorized researchers who have IRB/ethics approval and an  
400 institutionally approved study plan. To ensure the protection of privacy and compliance with national data

401 protection legislation, a data use/transfer agreement is needed, the content and specific clauses of which will  
402 depend on the nature of the requested data. Requests will be addressed in a reasonable time frame (generally  
403 two to three weeks), and the primary mode of data access is by either personal visit or remote access to a  
404 secure server.

405 Code for major analyses is available at [https://github.com/doge73/city\\_urban\\_PA](https://github.com/doge73/city_urban_PA).

406

## Reference

1. Booth FW, Roberts CK, Laye MJ. Lack of Exercise Is a Major Cause of Chronic Diseases. In: *Comprehensive Physiology*. 2012.p.1143–211. Available from: <https://doi.org/10.1002/cphy.c110025>  
<https://doi.org/https://doi.org/10.1002/cphy.c110025>.
2. Santos AC, Willumsen J, Meheus F, Ilbawi A, Bull FC. The cost of inaction on physical inactivity to public health-care systems: a population-attributable fraction analysis. *Lancet Glob Heal* 2023;**11**:e32–9. [https://doi.org/10.1016/S2214-109X\(22\)00464-8](https://doi.org/10.1016/S2214-109X(22)00464-8).
3. Duncan GE, Goldberg J, Noonan C, Moudon AV, Hurvitz P, Buchwald D. Unique Environmental Effects on Physical Activity Participation: A Twin Study. *PLoS One* 2008;**3**:e2019.
4. Boomsma DI, Cherkas L, Cornes BK, Harris JR, Kaprio J, Kujala UM, et al. Variance Components Models for Physical Activity With Age as Modifier: A Comparative Twin Study in Seven Countries. *Twin Res Hum Genet* 2012/02/21. 2011;**14**:25–34. <https://doi.org/DOI:10.1375/twin.14.1.25>.
5. Carlin A, Perchoux C, Puggina A, Aleksovskaja K, Buck C, Burns C, et al. A life course examination of the physical environmental determinants of physical activity behaviour: A “Determinants of Diet and Physical Activity” (DEDIPAC) umbrella systematic literature review. *PLoS One* 2017;**12**:e0182083.
6. Programme UNHS. World Cities Report 2022 [Internet]. United Nations, 2022 [cited 2024 Jan 23]. (World Cities Report). Available from: <https://www.un-ilibrary.org/content/books/9789210028592>  
<https://doi.org/10.18356/9789210028592>.
7. Durand CP, Andalib M, Dunton GF, Wolch J, Pentz MA. A systematic review of built environment factors related to physical activity and obesity risk: implications for smart growth urban planning. *Obes Rev* 2011;**12**:e173–82. <https://doi.org/https://doi.org/10.1111/j.1467-789X.2010.00826.x>.
8. Ding D, Sallis JF, Kerr J, Lee S, Rosenberg DE. Neighborhood Environment and Physical Activity Among Youth: A Review. *Am J Prev Med* 2011;**41**:442–55.  
<https://doi.org/https://doi.org/10.1016/j.amepre.2011.06.036>.
9. Kärmeniemi M, Lankila T, Ikäheimo T, Koivumaa-Honkanen H, Korpelainen R. The Built Environment as a Determinant of Physical Activity: A Systematic Review of Longitudinal Studies and Natural Experiments. *Ann Behav Med* 2018;**52**:239–51. <https://doi.org/10.1093/abm/kax043>.
10. Wild CP. The exposome: from concept to utility. *Int J Epidemiol* 2012;**41**:24–32.  
<https://doi.org/10.1093/ije/dyr236>.
11. Gorman S, Larcombe AN, Christian HE. Exposomes and metabolic health through a physical activity lens: a narrative review. *J Endocrinol* 2021;**249**:R25–41. <https://doi.org/10.1530/JOE-20-0487>.

12. Sonnenschein T, Scheider S, de Wit GA, Tonne CC, Vermeulen R. Agent-based modeling of urban exposome interventions: prospects, model architectures, and methodological challenges. *Exposome* 2022;**2**:osac009. <https://doi.org/10.1093/exposome/osac009>.
13. Wang H, Liu N, Chen J, Guo S. The Relationship Between Urban Renewal and the Built Environment: A Systematic Review and Bibliometric Analysis. *J Plan Lit* 2021;**37**:293–308. <https://doi.org/10.1177/08854122211058909>.
14. Chen Y, Liu G, Zhuang T. Evaluating the Comprehensive Benefit of Urban Renewal Projects on the Area Scale: An Integrated Method. Vol. 20, *International Journal of Environmental Research and Public Health*. 2023. <https://doi.org/10.3390/ijerph20010606>.
15. Vert C, Nieuwenhuijsen M, Gascon M, Grellier J, Fleming LE, White MP, et al. Health Benefits of Physical Activity Related to an Urban Riverside Regeneration. Vol. 16, *International Journal of Environmental Research and Public Health*. 2019. <https://doi.org/10.3390/ijerph16030462>.
16. Bartoll-Roca X, López MJ, Pérez K, Artazcoz L, Borrell C. Short-term health effects of an urban regeneration programme in deprived neighbourhoods of Barcelona. *PLoS One* 2024;**19**:e0300470.
17. Cooke M, Lumpe E, Stephenson M, Urjansson M, Aliev F, Palviainen T, et al. Alcohol use in early midlife: Findings from the age 37 follow-up assessment of the FinnTwin12 cohort. *OSF Prepr* 2024. <https://doi.org/10.31219/OSF.IO/A2N34>.
18. Caspersen CJ, Powell KE, Christenson GM. Physical Activity, Exercise, and Physical Fitness: Definitions and Distinctions for Health-Related Research. *Public Heal Reports* 1985;**100**:126–31.
19. Borodulin K, Laatikainen T, Juolevi A, Jousilahti P. Thirty-year trends of physical activity in relation to age, calendar time and birth cohort in Finnish adults. *Eur J Public Health* 2008;**18**:339–44. <https://doi.org/10.1093/eurpub/ckm092>.
20. Waller K, Kaprio J, Kujala UM. Associations between long-term physical activity, waist circumference and weight gain: a 30-year longitudinal twin study. *Int J Obes* 2008;**32**:353–61. <https://doi.org/10.1038/sj.ijo.0803692>.
21. Leskinen T, Waller K, Mutikainen S, Aaltonen S, Ronkainen PHA, Alén M, et al. Effects of 32-Year Leisure Time Physical Activity Discordance in Twin Pairs on Health (TWINACTIVE Study): Aims, Design and Results for Physical Fitness. *Twin Res Hum Genet* 2009;**12**:108–17. <https://doi.org/DOI:10.1375/twin.12.1.108>.
22. Jetté M, Sidney K, Blümchen G. Metabolic equivalents (METS) in exercise testing, exercise prescription, and evaluation of functional capacity. *Clin Cardiol* 1990;**13**:555–65. <https://doi.org/https://doi.org/10.1002/clc.4960130809>.



23. Kujala UM, Kaprio J, Sarna S, Koskenvuo M. Relationship of Leisure-Time Physical Activity and MortalityThe Finnish Twin Cohort. *JAMA* 1998;**279**:440–4. <https://doi.org/10.1001/jama.279.6.440>.
24. AINSWORTH BE, HASKELL WL, HERRMANN SD, MECKES N, BASSETT DRJR, TUDOR-LOCKE C, et al. 2011 Compendium of Physical Activities: A Second Update of Codes and MET Values. *Med Sci Sport Exerc* 2011;**43**.
25. Teeuwen R, Psyllidis A, Bozzon A. Measuring children’s and adolescents’ accessibility to greenspaces from different locations and commuting settings. *Comput Environ Urban Syst* 2023;**100**:101912. <https://doi.org/https://doi.org/10.1016/j.compenvurbsys.2022.101912>.
26. Miliias V, Psyllidis A. Assessing the influence of point-of-interest features on the classification of place categories. *Comput Environ Urban Syst* 2021;**86**:101597. <https://doi.org/https://doi.org/10.1016/j.compenvurbsys.2021.101597>.
27. van Kamp I, Persson Waye K, Kanninen K, Gulliver J, Bozzon A, Psyllidis A, et al. Early environmental quality and life-course mental health effects: The Equal-Life project. *Environ Epidemiol* 2022;**6**.
28. Wang Z, Zellers S, Whipp AM, Heinonen-Guzejev M, Foraster M, Júlvez J, et al. The effect of environment on depressive symptoms in late adolescence and early adulthood: an exposome-wide association study and twin modeling. *Nat Ment Heal* 2023. <https://doi.org/10.1038/s44220-023-00124-x>.
29. Thompson TP, Horrell J, Taylor AH, Wanner A, Husk K, Wei Y, et al. Physical activity and the prevention, reduction, and treatment of alcohol and other drug use across the lifespan (The PHASE review): A systematic review. *Ment Health Phys Act* 2020;**19**:100360. <https://doi.org/https://doi.org/10.1016/j.mhpa.2020.100360>.
30. Poortinga W. Associations of physical activity with smoking and alcohol consumption: A sport or occupation effect? *Prev Med (Baltim)* 2007;**45**:66–70. <https://doi.org/https://doi.org/10.1016/j.yjmed.2007.04.013>.
31. Abu-Omar K, Messing S, Sarshar M, Gelius P, Ferschl S, Finger J, et al. Sociodemographic correlates of physical activity and sport among adults in Germany: 1997–2018. *Ger J Exerc Sport Res* 2021;**51**:170–82. <https://doi.org/10.1007/s12662-021-00714-w>.
32. Kivimäki M, Batty GD, Pentti J, Shipley MJ, Sipilä PN, Nyberg ST, et al. Association between socioeconomic status and the development of mental and physical health conditions in adulthood: a multi-cohort study. *Lancet Public Heal* 2020;**5**:e140–9. [https://doi.org/10.1016/S2468-2667\(19\)30248-8](https://doi.org/10.1016/S2468-2667(19)30248-8).
33. Preud’homme G, Duarte K, Dalleau K, Lacomblez C, Bresso E, Smail-Tabbone M, et al. Head-to-

- head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark. *Sci Rep* 2021;**11**:4202. <https://doi.org/10.1038/s41598-021-83340-8>.
34. Huang Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Min Knowl Discov* 1998;**2**:283–304. <https://doi.org/10.1023/A:1009769707641>.
  35. Al-Zoubi MB, Rawi M al. An Efficient Approach for Computing Silhouette Coefficients. *J Comput Sci* 2008;**4**:252–5. <https://doi.org/10.3844/jcssp.2008.252.255>.
  36. Aschenbruck R, Szepannek G, Wilhelm AFX. Imputation Strategies for Clustering Mixed-Type Data with Missing Values. *J Classif* 2023;**40**:2–24. <https://doi.org/10.1007/s00357-022-09422-y>.
  37. Jolliffe IT. Outlier Detection, Influential Observations, Stability, Sensitivity, and Robust Estimation of Principal Components BT - Principal Component Analysis. In New York, NY: Springer New York, 2002.p.232–68. Available from: [https://doi.org/10.1007/0-387-22440-8\\_10](https://doi.org/10.1007/0-387-22440-8_10) [https://doi.org/10.1007/0-387-22440-8\\_10](https://doi.org/10.1007/0-387-22440-8_10).
  38. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min* 2016. <https://doi.org/10.1145/2939672>.
  39. Yu T, Zhu H. Hyper-parameter optimization: A review of algorithms and applications. *arXiv Prepr arXiv200305689* 2020.
  40. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. 2017. <https://doi.org/10.48550/arxiv.1705.07874>.
  41. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020;**2**:56–67. <https://doi.org/10.1038/s42256-019-0138-9>.
  42. Lim M, Hastie T. Learning Interactions via Hierarchical Group-Lasso Regularization. *J Comput Graph Stat* 2015;**24**:627–54. <https://doi.org/10.1080/10618600.2014.938812>.
  43. Viinikka A, Tiitu M, Heikinheimo V, Halonen JI, Nyberg E, Vierikko K. Associations of neighborhood-level socioeconomic status, accessibility, and quality of green spaces in Finnish urban regions. *Appl Geogr* 2023;**157**:102973. <https://doi.org/https://doi.org/10.1016/j.apgeog.2023.102973>.
  44. Zhou R, Li Y, Umezaki M, Ding Y, Jiang H, Comber A, et al. Association between Physical Activity and Neighborhood Environment among Middle-Aged Adults in Shanghai. Rissel C, editor. *J Environ Public Health* 2013;**2013**:239595. <https://doi.org/10.1155/2013/239595>.
  45. Shearer C, Blanchard C, Kirk S, Lyons R, Dummer T, Pitter R, et al. Physical Activity and Nutrition Among Youth in Rural, Suburban and Urban Neighbourhood Types. *Can J Public Heal* 2012;**103**:S55–60. <https://doi.org/10.1007/BF03403836>.

46. Sandercock G, Angus C, Barton J. Physical activity levels of children living in different built environments. *Prev Med (Baltim)* 2010;**50**:193–8. <https://doi.org/https://doi.org/10.1016/j.ypmed.2010.01.005>.
47. Zhu Z, Li Z, Liu Y, Chen H, Zeng J. The impact of urban characteristics and residents' income on commuting in China. *Transp Res Part D Transp Environ* 2017;**57**:474–83. <https://doi.org/https://doi.org/10.1016/j.trd.2017.09.015>.
48. Moore JB, Beets MW, Morris SF, Kolbe MB. Comparison of Objectively Measured Physical Activity Levels of Rural, Suburban, and Urban Youth. *Am J Prev Med* 2014;**46**:289–92. <https://doi.org/https://doi.org/10.1016/j.amepre.2013.11.001>.
49. Hermosillo-Gallardo ME, Jago R, Sebire SJ. Association between urbanicity and physical activity in Mexican adolescents: The use of a composite urbanicity measure. *PLoS One* 2018;**13**:e0204739.
50. Wardle J, Waller J, Jarvis MJ. Sex Differences in the Association of Socioeconomic Status With Obesity. *Am J Public Health* 2002;**92**:1299–304. <https://doi.org/10.2105/AJPH.92.8.1299>.
51. Brodersen NH, Steptoe A, Boniface DR, Wardle J. Trends in physical activity and sedentary behaviour in adolescence: ethnic and socioeconomic differences. *Br J Sports Med* 2007;**41**:140 LP – 144. <https://doi.org/10.1136/bjism.2006.031138>.
52. Van Dyck D, Cerin E, De Bourdeaudhuij I, Salvo D, Christiansen LB, Macfarlane D, et al. Moderating effects of age, gender and education on the associations of perceived neighborhood environment attributes with accelerometer-based physical activity: The IPEN adult study. *Health Place* 2015;**36**:65–73. <https://doi.org/https://doi.org/10.1016/j.healthplace.2015.09.007>.
53. Guillien A, Cadiou S, Slama R, Siroux V. The Exposome Approach to Decipher the Role of Multiple Environmental and Lifestyle Determinants in Asthma. *Int J Environ Res Public Health* 2021;**18**. <https://doi.org/10.3390/ijerph18031138>.
54. Cohen DA, Marsh T, Williamson S, Han B, Derose KP, Golinelli D, et al. The Potential for Pocket Parks to Increase Physical Activity. *Am J Heal Promot* 2014;**28**:S19–26. <https://doi.org/10.4278/ajhp.130430-QUAN-213>.
55. Peng H, Li X, Yang T, Tan S. Research on the Relationship between the Environmental Characteristics of Pocket Parks and Young People's Perception of the Restorative Effects—A Case Study Based on Chongqing City, China. *Sustainability* 2023;**15**. <https://doi.org/10.3390/su15053943>.
56. Yang L, Yang H, Cui J, Zhao Y, Gao F. Non-linear and synergistic effects of built environment factors on older adults' walking behavior: An analysis integrating LightGBM and SHAP. *Trans Urban Data, Sci Technol* 2024:27541231241249864. <https://doi.org/10.1177/27541231241249866>.

57. Keskinen KE, Gao Y, Rantakokko M, Rantanen T, Portegijs E. Associations of Environmental Features With Outdoor Physical Activity on Weekdays and Weekend Days: A Cross-Sectional Study Among Older People. *Front Public Heal* 2020;**8**.
58. Zang P, Qiu H, Xian F, Yang L, Qiu Y, Guo H. Nonlinear Effects of the Built Environment on Light Physical Activity among Older Adults: The Case of Lanzhou, China. *Int J Environ Res Public Health* 2022;**19**. <https://doi.org/10.3390/ijerph19148848>.
59. McGinn AP, Evenson KR, Herring AH, Huston SL, Rodriguez DA. Exploring Associations between Physical Activity and Perceived and Objective Measures of the Built Environment. *J Urban Heal* 2007;**84**:162–84. <https://doi.org/10.1007/s11524-006-9136-4>.
60. Browning MHEM, Rigolon A, McAnirlin O, Yoon H (Violet). Where greenspace matters most: A systematic review of urbanicity, greenspace, and physical health. *Landsc Urban Plan* 2022;**217**:104233. <https://doi.org/https://doi.org/10.1016/j.landurbplan.2021.104233>.
61. Cardinali M, Beenackers MA, van Timmeren A, Pottgiesser U. The relation between proximity to and characteristics of green spaces to physical activity and health: A multi-dimensional sensitivity analysis in four European cities. *Environ Res* 2024;**241**:117605. <https://doi.org/https://doi.org/10.1016/j.envres.2023.117605>.
62. Huang B, Yao Z, Pearce JR, Feng Z, James Browne A, Pan Z, et al. Non-linear association between residential greenness and general health among old adults in China. *Landsc Urban Plan* 2022;**223**:104406. <https://doi.org/https://doi.org/10.1016/j.landurbplan.2022.104406>.
63. Guan M. Associations of fruit & vegetable intake and physical activity with poor self-rated health among Chinese older adults. *BMC Geriatr* 2022;**22**:10. <https://doi.org/10.1186/s12877-021-02709-6>.
64. Ping WX, Yan LZ, Meng Z, Yong LH, Ping WX, Yan LZ, et al. Machine-learning-assisted Investigation into the Relationship between the Built Environment, Behavior, and Physical Health of the Elderly in China. *Biomed Environ Sci* 2023, Vol 36, Issue 10, Pages 987-990 2023;**36**:987–90. <https://doi.org/10.3967/BES2023.125>.
65. Lee K, Wang J, Heo J. How the physical inactivity is affected by social-, economic- and physical-environmental factors: an exploratory study using the machine learning approach. *Int J Digit Earth* 2023;**16**:2503–21. <https://doi.org/10.1080/17538947.2023.2230944>.
66. Ohanyan H, Portengen L, Huss A, Traini E, Beulens JWJ, Hoek G, et al. Machine learning approaches to characterize the obesogenic urban exposome. *Environ Int* 2022;**158**:107015. <https://doi.org/https://doi.org/10.1016/j.envint.2021.107015>.
67. FIEC - Statistical Report 2023 [Internet]. [cited 2023 Nov 22]. Available from: <https://fiec-statistical-report.eu/>



Figure 1 Study flow

Figure 2: Twin participants' residence in the Helsinki and Espoo area in 2018 colored by cluster.

Note: The gray layer shows centers and shopping areas in 2019.

Figure 3: Results of XGBoost models for total leisure-time physical activity (total LTPA) and leisure-time physical activity without commuting activity (LTPA)

Note: The SHAP bar plots show the influence of each variable: total LTPA (a) and LTPA (e). The SHAP dependence plots show how a single individual influences the XGboost prediction on total LTPA (b, c, d) and LTPA (f, g, h). ints\_500 is the count of any type of road junctions within a 500 m buffer; sumarea\_pocketparks\_800 is the total area of all interconnected pocket parks within an 800 m walking distance; ndvi\_5yrs\_all\_500 is the 5-years moving average of Normalized Difference Vegetation Index within a 500 m buffer during whole year; count\_pocketparks\_800 is the count of pocket parks within an 800 m walking distance. Abbreviation: leisure-time physical activity (LTPA); SHapley Additive exPlanations (SHAP)

**Table 1: Characteristics of sociodemographic, behavior, and neighborhood social variables  
(participants n=394)**

<b>Characteristics</b>	<b>N. (%) / Mean (SD)</b>
<b>Sex</b>	
Male	179 (45.4)
Female	215 (54.6)
<b>Work</b>	
Not working or other situation	51 (12.9)
Currently work	343 (87.1)
<b>Education</b>	
Post-secondary or lower	84 (21.3)
Bachelor/equivalent or above	310 (78.7)
<b>Marital status</b>	
No	97 (24.6)
Married, steady relationship, or living together	297 (75.4)
<b>Age (years)</b>	37.1 (1.5)
<b>Illicit substance use</b>	
No	204 (51.8)
Yes	190 (48.2)
<b>Ever smoker (smoked over 100 cigarettes lifetime)</b>	
No	216 (54.8)
Yes	178 (45.2)
<b>Alcohol</b>	
Monthly or less, or even never	164 (41.6)
2-4 times a month or more	230 (58.4)
<b>Deprivation level</b>	
Low	225 (57.1)
High	169 (42.9)
<b>The proportion of single households in the neighborhood</b>	50.0 (10.8)

**Table 2: Results of the linear regression between the urban cluster and physical activity measures**

Outcome (log-transformed)	Characteristics	Beta (95% CI)		
		Model 1 <sup>a</sup>	Model 2 <sup>b</sup>	Model 3 <sup>c</sup>
	<b>Urban cluster</b>			
<b>Total LTPA</b>	1 (original city center)	Ref.	Ref.	Ref.
	2 (new city center)	-0.06 (-0.13, 0.01)	-0.06 (-0.13, 0.01)	-0.05 (-0.13, 0.03)
	3 (suburban)	-0.14 (-0.22, -0.07)*	-0.14 (-0.22, -0.06)*	-0.13 (-0.23, -0.03)*
	<b>Urban cluster</b>			
<b>LTPA</b>	1 (original city center)	Ref.	Ref.	Ref.
	2 (new city center)	-0.07 (-0.15, 0.01)	-0.06 (-0.15, 0.02)	-0.06 (-0.16, 0.03)
	3 (suburban)	-0.17 (-0.26, -0.08)*	-0.17 (-0.26, -0.07)*	-0.17 (-0.28, -0.05)*
	<b>Urban cluster</b>			
<b>Commuting activity</b>	1 (original city center)	Ref.	Ref.	Ref.
	2 (new city center)	-0.01 (-0.06, 0.03)	-0.01 (-0.05, 0.04)	0.00 (-0.04, 0.05)
	3 (suburban)	-0.03 (-0.08, 0.01)	-0.03 (-0.08, 0.02)	-0.01 (-0.07, 0.05)

\* P<0.05

<sup>a</sup> Adjusted for age, sex, education, work, and marital status

<sup>b</sup> Based on model 1, additionally adjusted for smoking, alcohol drinking, and illicit substance use

<sup>c</sup> Based on model 2, additionally adjusted for neighborhood deprivation level and the proportion of single households in the neighborhood



## Data preparation

- 58 exposures available in 2023
- 84 exposures available in 2018
- 3 exposures available in 2015

2120 individual twins responded to the early midlife follow-up

423 lived in Helsinki, Tampere, Espoo, Oulu, or Jyväskylä, in 2020

394 included in the analysis

- 13 removed for missing residential information in 2018
- 15 removed due to missing physical activity and other measures
- 1 removed as an outlier in total LTPA



### Objective 1

PCAmix to assess outliers of the urban exposome (3 excluded)

K-prototype clustering

Hierarchical linear regression between urban cluster and physical activity

Sex-stratification sensitivity analysis



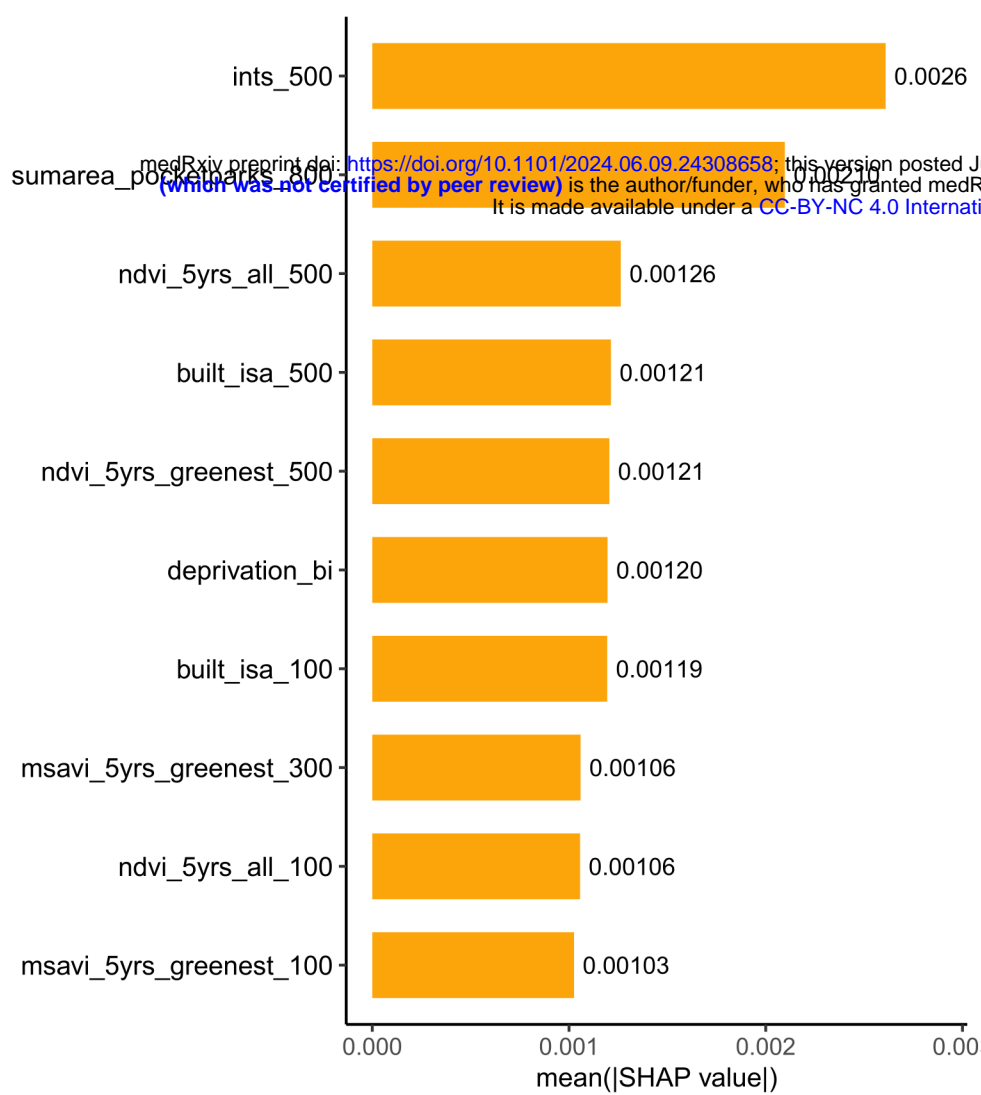
### Objective 2

Repeated generalized linear regression to select strong candidates ( $P < 0.01$ )

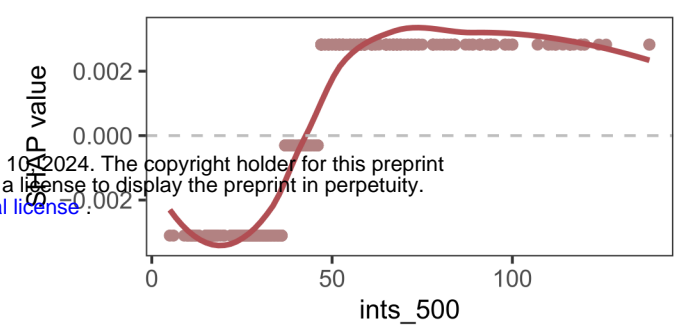
XGBoost to assess importance, linearity, and interaction

Group-Lasso INTERaction-NET for additional interaction assessment

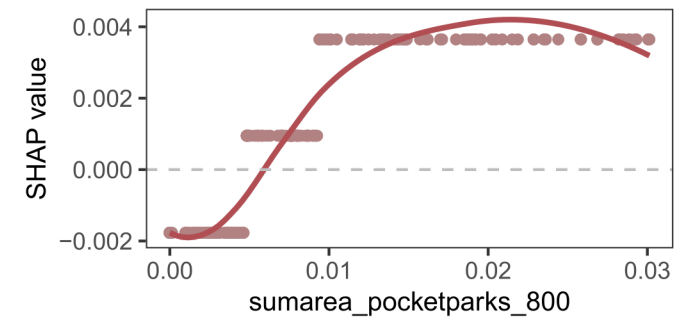
A. outcome: Total LTPA



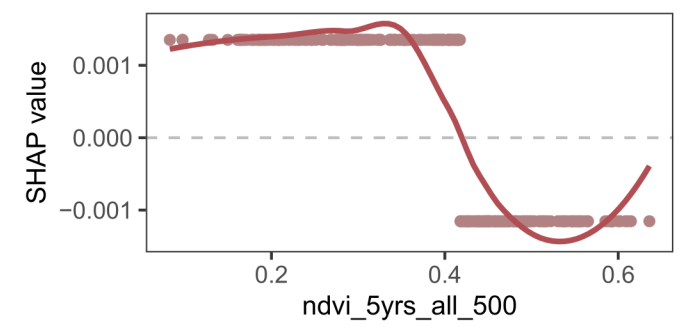
B. outcome: Total LTPA



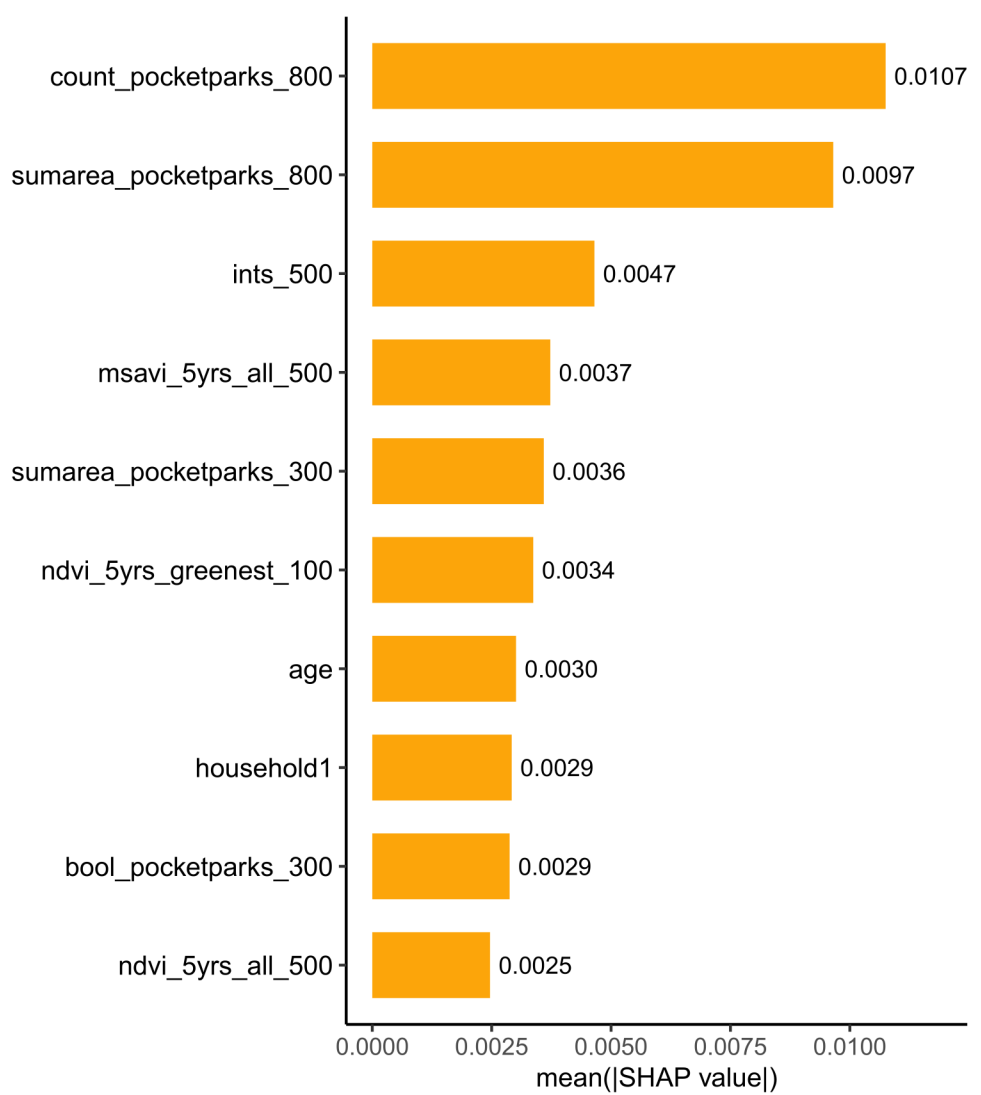
C. outcome: Total LTPA



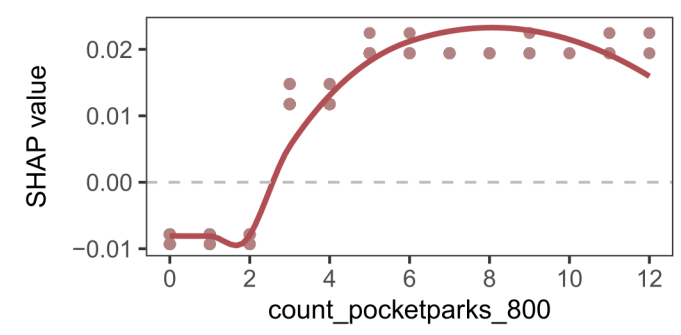
D. outcome: Total LTPA



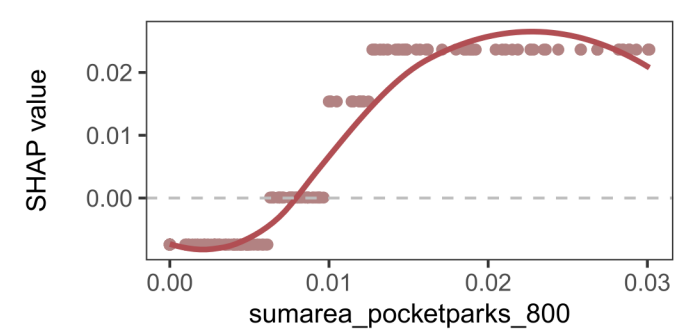
E. outcome: LTPA



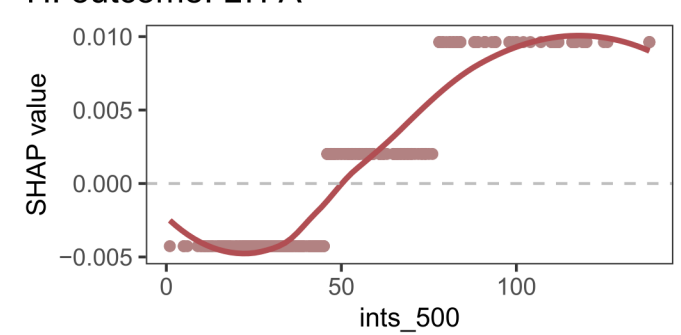
F. outcome: LTPA



G. outcome: LTPA



H. outcome: LTPA



medRxiv preprint doi: <https://doi.org/10.1101/2024.06.09.24308658>; this version posted June 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-NC 4.0 International license](https://creativecommons.org/licenses/by-nc/4.0/).

