1 **Title:** Built Environment Features Obtained from Google Street View Are Associated with Coronary
2 Artery Disease Prevalence: A Deep-Learning Framework
3
4 **Authors:** Zhuo Chen, PhD; Yassin Khalifa, PhD; Jean-Eudes Dazard, PhD; Issam Motairek, MD; Sanjay
5 Rajagopalan, MD*; Sadeer Al-Kindi, MD*
6
7 *Contributed Equally
8
9 **Affiliations**
10 Harrington Heart and Vascular Institute, University Hospitals, and School of Medicine, Case Western
11 Reserve University, Cleveland, OH
12
13 **Corresponding Authors**
14 Sadeer Al-Kindi, MD, FACC
15 Assistant Professor of Medicine
16 University Hospitals Harrington Heart and Vascular Institute
17 Case Western Reserve University School of Medicine
18 11100 Euclid Ave, Cleveland, OH 44106
19 Email: sadeer.al-kindi@uhhospitals.org
20
21 OR
22
23 Sanjay Rajagopalan, MD, FACC
24 Herman K Hellerstein Professor of Cardiovascular Research
25 Director, Cardiovascular Research Institute
26 University Hospitals Harrington Heart and Vascular Institute
27 Case Western Reserve University School of Medicine
28 11100 Euclid Ave, Cleveland, OH 44106
29 Email: sanjay.rajagopalan@uhhospitals.org
30
31 **Key words:** Cardiovascular risk, neighborhood, phenotypes, socio-environmental, machine learning
32
33 **Word count:**
34
35 **Disclosures:** None of the authors have conflicts of interest relevant to the contents of
36 this manuscript.
37

40 **Abstract**

41 Background: Built environment plays an important role in development of cardiovascular disease. Tools

42 to evaluate the built environment using machine vision and informatic approaches has been limited. We

43 sought to investigate the association between machine vision-based built environment and prevalence of

44 cardiometabolic disease in urban cities.

45 Methods: This cross-sectional study used features extracted from Google Street view (GSV) images to

46 measure the built environment and link them with prevalence of cardiometabolic disease. Convolutional

47 neural networks, light gradient boosting machines and activation maps were utilized to predict health

48 outcomes and identify feature associations with coronary heart disease (CHD). The study obtained 0.53

49 million GSV images covering 789 census tracts in 7 cities (Cleveland, OH; Fremont, CA; Kansas City, MO;

50 Detroit, MI; Bellevue, WA; Brownsville, TX; and Denver, CO). Analyses were conducted from February

51 2022 to December 2022. We used census tract-level data from the Centers for Disease Control and

52 Prevention's PLACES dataset. Main outcomes included census tract-level estimated prevalence of CHD

53 based on GSV built environment features.

54 Results: Built environment features extracted from GSV using deep learning predicted 63% of the census

55 tract variation in CHD prevalence. The ExtraTrees Regressor achieved the best result among all models

56 with the lowest average mean absolute error of 1.11% and Root mean square of error of 1.58. The

57 addition of GSV features outperformed and improved a model that only included census-tract level age,

58 sex, race, income and education. Activation maps from the features revealed a set of neighborhood

59 features represented by buildings and roads associated with CHD prevalence.

60 Conclusions: In this cross-sectional study, a significant portion of CHD prevalence were explained by GSV-

61 based built environment factors analyzed using deep learning, independent of census tract demographics.

62 Machine vision enabled assessment of the built environment could help play a significant role in designing

63 and improving heart-heathy cities.

64 **Introduction**
65 Coronary heart disease (CHD) accounts for over 50% of mortality from heart disease in the United States,
66 responsible for nearly 400,000 deaths in 2020[1]. Despite advances in prevention and treatment over the
67 past decade in the United States[2] CHD remains the leading cause of death in the United States since 1950
68 with increasing evidence for non-conventional risk factors playing a large than anticipated role than
69 previously suspected[1,3].
70
71 Socioenvironmental factors are amongst the leading non-traditional risk factors increasingly implicated in
72 CHD development[4–6]. These factors include social determinants such as race, income, education, and
73 culture as well as the factors in the built environment and factors in the ambient environment such as
74 noise, and air pollution all of which have been to exert significant effects on CHD. [5–8]
75
76 Large-scale integrated assessment of the environment at the neighborhood can facilitate rapid and
77 complete assessment of its impact on CHD. Such data is however scarce, partly because of the costly and
78 time-consuming nature of neighborhood audits, and inconsistent measurements and standards for data
79 collection. Machine vision approaches such as Google Street View (GSV) has become an increasingly
80 popular approach for virtual neighborhood audits since its launch in 2007. GSV image coverage has been
81 consistently expanding in recent years achieving almost full coverage in the United States[9]. Previous
82 studies have shown GSV results are comparable to field assessments and have been used to assess the
83 built environment features such as greenspace[10,11], buildings[12], and roads[13].
84
85 GSV images further become a favored data source for large-scale studies due to the open availability of
86 such data, arguably the largest compendium of machine vision enabled assessment of large tracts of the
87 earth, and the standardized approaches used. Deep learning approaches such as convolutional neural
88 networks (CNN) have been widely used in many studies and applications, given their excellent
89 performance in tasks such as image classification, object detection, and image segmentation[14]. The use of
90 such approaches to rapidly assess and extract built environment features from GSV images using deep
91 learning can help facilitate integrated assessment and capture other aspects that may not be otherwise
92 included. The goal of this study is to use GSV images to assess built environment and use them to estimate
93 CHD prevalence at the census tract level.
94
95 **Methods**
96 **Data source for coronary heart disease**
97 The prevalence of census-tract coronary heart disease (CHD) was obtained from the CDC PLACES, a project
98 that provided chronic disease risk factors, health outcomes, and clinical preventive services. This project,
99 is a collaboration between the Centers for Disease Control and Prevention (CDC), the Robert Wood
100 Johnson Foundation, and the CDC Foundation, measures CHD prevalence using data from Behavioral Risk
101 Factor Surveillance System (BRFSS), where people aged ≥18 are surveyed to report whether or not they
102 have been told by a doctor, nurse, or other health professional that they had angina or coronary heart
103 disease. We collected the CHD prevalence data for 789 census tracts in 7 cities: Bellevue, WA; Brownsville,
104 TX; Cleveland, OH; Denver, CO; Detroit, MI; Fremont, CA; and Kansas City, KS.
105
106 **Google street view data**
107 Environment information was derived from approximately 0.53 million GSV images for the 7 cities (143K
108 for Detroit, 59K for Kansas City, 70K for Cleveland, 65K for Brownsville, 38K for Fremont, 35K for Bellevue,
109 and 120K for Denver). The GSV images were downloaded via Google Street View Static Application
110 Programming Interface (API) from 2020-2021. GSV API provides users with street-level panoramic
111 imagery which captures the visual domain of pedestrians in thousands of cities worldwide. The GSV

112   images of each census tract were downloaded in a grid pattern in the corresponding tract with an interval
113   of 100m. At each location where GSV images were retrieved, four images were gathered from different
114   directions (i.e., the cardinal directions: N, E, S, and W.), which composes a panoramic view of the
115   surroundings at that location. When latitude and longitude coordinates are provided, the API searches
116   within a 50-meter radius for a photograph closest to this location. The API would not return any images
117   if no available images could be found.

118

119   To process these images and gain environment information from them, a pre-trained deep convolutional
120   neural network (DCNN) Place365 CNN [15] was used as the feature extractor to obtain the deep features of
121   the image. Here, the deep features are the outputs of the deep layers in the hierarchy of the network.
122   Compared with the shallow features in the shallow layers, these deep features represent the semantic
123   information of the GSV images. Details of how the extraction was performed can be found in e Figure 1 in
124   the Supplement. We used Place365 CNN as the feature extractor because the images trained on Place365
125   CNN are more similar to that of GSV. Place365 CNN was trained on the subset of Places Database, which
126   contains more than 10 million images consisting of 400+ unique scene categories such as towers, soccer
127   fields, streets, swimming pools, and train station platforms. Compared with the ImageNet database, the
128   diversity of environmental features found in the Places Database was believed to be representative of
129   what is contained in GSV images. Through feature extraction, we obtained 4096 features representing the
130   average built environment information for each census tract.

131

132   **Statistical Analysis**
133   **Features Visualization using Grad-CAM**
134   Elastic net regression models[16] were used to estimate the census tract-level CHD prevalence by using the
135   DCNN-extracted features from GSV images. There are 4096 features and elastic net can handle this high
136   dimensional data by applying L1 and L2 regularization. Ten-fold cross-validation was repeated 3 times to
137   find the best parameters of the elastic net. Elastic net can select important features by simultaneously
138   performing feature selection and feature shrinkage, so we used it to select top features according to the
139   coefficients of each feature. The top features can be evaluated by examining the magnitudes and signs of
140   their coefficients in the elastic net mode, thus understanding how each feature is associated with CHD
141   prevalence. The top features were then visualized as the saliency map in the original GSV images using
142   Grad-CAM technique[17], which provides certain explanations of what environmental features the CNN
143   thinks to be associated with neighborhood CHD prevalence.

144

145   **Machine Learning Models with CNN-extracted Features**
146   A variety of machine-learning predictive models were used and compared to explore the association
147   between the CNN-extracted features of GSV images and the tract-level CHD prevalence. The models for
148   this analysis included ExtraTrees regressor (ET), AdaBoost regressor (AB), Random Forest Regressor (RF),
149   Gradient Boosting Regressor (GB), Extreme Gradient Boosting Regressor (XGB), and Light Gradient
150   Boosted Machine Regressor (LGBM). All models were estimated using a 10-fold cross-validation technique
151   for a more robust result. For 10-fold cross-validation, the dataset is split into 10 equal-sized subsets, and
152   the model is trained on 9 subsets and tested on the remaining 1 subset. This process is repeated 10 times
153   until all 10 subsets were used once as the testing set. R-squared values were reported as the measure of
154   association between the CNN-extracted features of GSV images and the tract-level CHD prevalence. The
155   performance of each model was also evaluated using the mean absolute error (MAE) and root mean
156   squared error (RMSE).

157

158   **Multilevel Modeling with Demographics and Socio-Economic Factors**

159 We analyzed the effects of common demographic and socio-economic factors (DSE) as well as CNN-
160 extracted features of GSV images (GSV) associated with the CHD. We built multilevel regression models
161 to account for the effect of these factors including city, age, sex, race, income and education. A
162 multivariate Sparse Partial Least Squares (SPLS) regression[18] was applied first to the CNN-extracted
163 features to reduce the dimensionality issue and the effect of noise. The selected SPLS components and
164 the demographic and socio-economic factors were then used to fit a Linear Mixed-Effect regression model.
165 Three models were compared in this analysis: 1) a model containing both DSE factors and SPLS
166 components (Combined Model); 2) a model with DSE factors alone (DSE Model); 3) a model with the SPLS
167 components alone (GSV Model). Model performance was assessed using goodness of fits measures such
168 as Likelihood Ratios Tests (as it applied), AIC and BIC criteria. In addition, all models were compared by
169 using R squared values obtained from a Light Gradient Boosted Machine (LGBM) and Random Forest (RF)
170 predictive model, which represent the amount of variance explained by each set of independent variables.
171 The comparison was done using a 10-fold cross validation scheme.

172

173 **Results**
174 **Regression results with CNN Features**
175 The 4096 CNN-extracted features from GSV images were able to explain more than 63% of the variance
176 (R^2 = 0.634) on the tract-level CHD prevalence in 7 cities (Figure 1). The ET achieved the best result among
177 all models with the lowest average MAE of 1.11 and RMSE of 1.58. The actual estimate from CDC's CHD
178 prevalence and the model-predicted CHD prevalence were mapped for all census tracts in 7 cities (Figure
179 2). There was a good agreement between the actual estimates and predicted CHD prevalence across all
180 census tracts in 7 cities. We found a small number of extreme values that were underestimated by the
181 models in certain census tracts of Detroit and Cleveland. The CHD prevalence of these underestimated
182 census tracts was often more than 12%. When examining the CNN-extracted features using t-SNE, we
183 noticed clustering of census tracts with similar values of CHD prevalence (eFigure2 in the supplement)

184

185

186 **Visualization of Top CNN Features**
187 Grad-CAM was utilized to visualize top CNN-extracted features identified from the elastic net model. The
188 saliency maps generated by the Grad-CAM, suggested that feature #1555, which seemed to highlight
189 deteriorated buildings (suggesting neighborhood blight), had a positive association with CHD prevalence
190 (Figure 3a). Another feature (feature #484) that was positively associated with CHD was found to be
191 highlighting road cracks as shown in Figure 3b. In contrast, feature #204 in Figure 3c had a negative
192 association with CHD prevalence, and its heatmap highlighted trees along the road. Feature #1732,
193 seeming to focus on well built houses, also had a negative association with CHD prevalence (Figure 3d).

194

195

196 **Comparison of CNN Features with Demographics and Socio-Economic Factors**
197 With SPLS, an optimal model was obtained with h = 7 SPLS components (η = 0.6), yielding a model with
198 816 CNN-extracted features that explain $R^2_{XY} > 66.7\%$ variance of CHD prevalence in the census tracts.
199 All three models were compared with Table 1 shows model comparisons for all three models Likelihood
200 Ratios Tests (LRT, also see eMethods in the Supplement). eTable 1 in the Supplement shows the
201 corresponding regression estimates and ANOVA results. Table 2 and Supplemental eTable 2 show the
202 amount of total explained variance of the GSV models and demographic and socioeconomic (DSE)
203 variables (Table 2: LGBM, eTable 2: RF). After adjusting for each individual variable, we found that the
204 combined model (GSV + DSE) demonstrated a better Goodness of Fit, with statistically significant higher
205 log-likelihood and lower AIC/BIC when compared to GSV or DSE model alone (Table 1). Also, we found
206 that nearly all the SPLS components are statistically significant (eTable 1 in the Supplement). Although the

207   DSE model has lower AIC and BIC values, with a significant LRT, when compared to the GSV model alone
208   (Table 1), the GSV features alone explain more variance of CHD prevalence than the DSE variables (Table
209   2). Altogether, this indicates that GSV features and traditional demographics and socio-economics
210   variables are both significantly associated and predictive of CHD prevalence.
211
212   **Discussion**
213   While many epidemiological studies have examined associations between cardiovascular disease and
214   individual built environmental features (e.g. greenspace, urban architecture, street connectivity, food
215   availability), our approach focused on machine vision derived physical environment, relying on
216   convolutional neural networks (CNN) and its related techniques to extract features.
217
218   Our results showed a good association (R^2 = 0.634) between the CNN-extracted features from GSV and
219   CHD at the census tract level in 7 cities. This indicated that the CNN-extracted features could capture
220   neighborhood features impacting cardiovascular health. The predicted CHD prevalence using CNN-
221   extracted features tended to underestimate in certain areas compared to observed CHD prevalence
222   especially in Detroit and Cleveland. This may suggest that certain CHD-related factors may either not be
223   embedded in these environments at these locations or that perhaps features not captured by street view
224   images, such as demographic factors, ambient factors and other demographic and traditional variables
225   may play a much larger role in these environments.
226
227   Our approach took the advantage of the knowledge that fully connected layers in the CNN contain
228   condensed information of the input imagery that can be extracted and utilized for a variety of purposes.
229   We utilized a pre-trained deep convolutional neural network (DCNN) Place365 CNN[15], so that the deep
230   features from the CNN may be more representative of the built environment. One advantage of this
231   approach is that predefined relevant features in the built environment is not required. The 4096-
232   dimensional deep features embeds all essential information of the built environment in the imagery so
233   that we could retain relevant features as much as reasonably possible. Conversely, the disadvantage of
234   using deep features from a pre-trained CNN is that it becomes difficult to identify corresponding physical
235   features that impact CHD at the neighborhood level. To alleviate this issue and provide certain
236   interpretations of the deep features, we utilized Grad-CAM techniques to visualize the CHD-related
237   features with a saliency map.
238
239   Grad-CAM highlighted several potential built environment features that are either associated with higher
240   or lower CHD at the neighborhood level. Deteriorated houses and roads are a feature of urban blight
241   associated with higher CHD. This feature may in turn embody other features in the neighborhood that
242   drive cardiovascular risk, including lack of space for physical activity[7,19], limited access to nutritionally-
243   balanced food[20], lack of access to health care[21]. Street greenery on the other hand was highlighted as
244   associated with lower CHD prevalence. This agrees with previous studies that showed a robust association
245   between green space and decreased cardiovascular risks[22,23].
246
247   The results of multilevel modeling using demographics and socio-economic factors, indicate that
248   demographics and socio-economic variables, were still better predictors of CHD prevalence, than GSV
249   features. One explanation is obviously the fact that physical environmental feature even if they represent
250   a "meta" framework for other mediators, may not be sufficient to convey the risk conveyed by other
251   factors which may be sparsely represented. Another reason may be that GSV features may engender
252   increased model complexity, by virtue of including 4096 features (Table 1). However, GSV features alone
253   could still explain preponderant proportion of variance in CHD prevalence (Table 2 and eTable1 in the
254   Supplement). Therefore, by incorporating GSV features into regular DSE variables, one could help improve

255 the prediction of CHD prevalence at the neighborhood level. Our results further suggest that GSV features
256 indeed may be helpful in highlighting specific built environment information related to CHD prevalence at
257 the neighborhood level as illustrated by Grad-CAM methods, which provided a way of identifying built
258 environment information.
259
260 There are multiple limitations of this study that should be noted. Firstly, the GSV images used in the study
261 are only available along major streets and roads, and there are some populations who do not live in such
262 neighborhood. However, given the fact that most population live around the urban neighborhood where
263 GSV are abundant, we believe this would not significantly affect the results for majority of census tracts.
264 Further, although Place365 database contains 400+ unique scene categories, it may not include all
265 features that can be found in the built environment. Small objects such as trash, other environmental
266 pollutants and physical domains that may translate into better urban quality of life, may be difficult for
267 computer vision techniques like CNN to detect in a GSV image [24]. Additionally, the census tracts with CHD
268 prevalence data are from 7 representative U.S cities of CDC PLACES dataset, and may not generalize to all
269 census tracts in the U.S., especially rural areas[25]. Future work is needed to examine the disparities of
270 urban and rural areas and its cardiovascular-related built environment features.
271
272 **Conclusion**
273 Built environment impacts cardiovascular health outcome. In this study, we used Google Street View (GSV)
274 and a scene-pretrained convolutional neural network (CNN) to assess the built environment. We found
275 CNN-extracted features explain significant portion of coronary heart disease (CHD) prevalence at the
276 census tract level. Compared to traditional demographic and socio-economic factors, GSV provides unique
277 information that may relate to CHD such as buildings, greenspace and roads as suggested by the activation
278 maps from Grad-CAM technique. The outcomes of our study provides proof of concept for machine-vision
279 enabled identification of urban network features associated with risk that in principle, may enable rapid
280 identification and targeting interventions in at-risk neighborhoods to reduce cardiovascular burden.
281
282 **References**

283 1. Tsao CW, Aday AW, Almarzooq ZI, et al. Heart Disease and Stroke Statistics—2022 Update: A Report
284   From the American Heart Association. *Circulation*. 2022;145(8):e153-e639.
285   doi:10.1161/CIR.0000000000001052

286 2. *Health, United States, Annual Perspective, 2020-2021*. National Center for Health Statistics (U.S.);
287   2022. doi:10.15620/cdc:122044

288 3. Heron M, Anderson RN. Changes in the Leading Cause of Death: Recent Patterns in Heart Disease and
289   Cancer Mortality. *NCHS Data Brief*. 2016;(254):1-8.

290 4. Havranek EP, Mujahid MS, Barr DA, et al. Social Determinants of Risk and Outcomes for
291   Cardiovascular Disease. *Circulation*. 2015;132(9):873-898. doi:10.1161/CIR.0000000000000228

292 5. Al-Kindi SG, Brook RD, Biswal S, Rajagopalan S. Environmental determinants of cardiovascular
293   disease: lessons learned from air pollution. *Nature Reviews Cardiology*. 2020;17(10):656-672.
294   doi:10.1038/s41569-020-0371-2

295 6. Bhatnagar A. Environmental Determinants of Cardiovascular Disease. *Circulation Research*.
296   2017;121(2):162-180. doi:10.1161/CIRCRESAHA.117.306458

297   7. Sallis JF, Floyd MF, Rodríguez DA, Saelens BE. Role of built environments in physical activity, obesity,
298   and cardiovascular disease. *Circulation*. 2012;125(5):729-737.
299   doi:10.1161/CIRCULATIONAHA.110.969022

300   8. Rajagopalan S, Landrigan PJ. Pollution and the Heart. *N Engl J Med*. 2021;385(20):1881-1892.
301   doi:10.1056/NEJMra2030281

302   9. Google. How Street View works and where we will collect images next. Google Maps Street View.
303   Accessed February 2, 2023. https://www.google.com/streetview/how-it-works/

304   10.   Seiferling I, Naik N, Ratti C, Proulx R. Green streets – Quantifying and mapping urban trees with
305   street-level imagery and computer vision. *Landscape and Urban Planning*. 2017;165:93-101.
306   doi:10.1016/j.landurbplan.2017.05.010

307   11.   Lu Y. Using Google Street View to investigate the association between street greenery and
308   physical activity. *Landscape and Urban Planning*. 2019;191:103435.
309   doi:10.1016/J.LANDURBPLAN.2018.08.029

310   12.   Kang J, Körner M, Wang Y, Taubenböck H, Zhu XX. Building instance classification using street
311   view images. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2018;145:44-59.
312   doi:10.1016/j.isprsjprs.2018.02.006

313   13.   Nagata S, Nakaya T, Hanibuchi T, Amagasa S, Kikuchi H, Inoue S. Objective scoring of streetscape
314   walkability related to leisure walking: Statistical modeling approach with semantic segmentation of
315   Google Street View images. *Health & Place*. 2020;66:102428. doi:10.1016/j.healthplace.2020.102428

316   14.   LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444.
317   doi:10.1038/nature14539

318   15.   Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A. Places: A 10 Million Image Database for Scene
319   Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2018;40(6):1452-1464.
320   doi:10.1109/TPAMI.2017.2723009

321   16.   Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal
322   Statistical Society Series B: Statistical Methodology*. 2005;67(2):301-320. doi:10.1111/j.1467-
323   9868.2005.00503.x

324   17.   Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations
325   from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE international
326   conference on computer vision*. Published online 2017:618-626. doi:10.1007/s11263-019-01228-7

327   18.   Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction
328   and variable selection. *J R Stat Soc Series B Stat Methodol*. 2010;72(1):3-25. doi:10.1111/j.1467-
329   9868.2009.00723.x

330   19.   Chandrabose M, Rachele JN, Gunn L, et al. Built environment and cardio-metabolic health:
331   systematic review and meta-analysis of longitudinal studies. *Obesity Reviews*. 2019;20(1):41-54.
332   doi:10.1111/obr.12759

333 20. Gondi KT, Larson J, Sifuentes A, et al. Health of the Food Environment Is Associated With Heart
334   Failure Mortality in the United States. *Circulation: Heart Failure*. 2022;15(12):e009651.
335   doi:10.1161/CIRCHEARTFAILURE.122.009651

336 21. White-Williams C, Rossi LP, Bittner VA, et al. Addressing Social Determinants of Health in the
337   Care of Patients With Heart Failure: A Scientific Statement From the American Heart Association.
338   *Circulation*. 2020;141(22):e841-e863. doi:10.1161/CIR.0000000000000767

339 22. Pereira G, Foster S, Martin K, et al. The association between neighborhood greenness and
340   cardiovascular disease: an observational study. *BMC Public Health*. 2012;12:466. doi:10.1186/1471-
341   2458-12-466

342 23. Mitchell R, Popham F. Effect of exposure to natural environment on health inequalities: an
343   observational population study. *Lancet*. 2008;372(9650):1655-1660. doi:10.1016/S0140-
344   6736(08)61689-X

345 24. Ross CE, Mirowsky J. Disorder and Decay: The Concept and Measurement of Perceived
346   Neighborhood Disorder. *Urban Affairs Review*. 1999;34(3):412-432.
347   doi:10.1177/107808749903400304

348 25. Loccoh EC, Joynt MKE, Wang Y, Kazi DS, Yeh RW, Wadhera RK. Rural-Urban Disparities in
349   Outcomes of Myocardial Infarction, Heart Failure, and Stroke in the United States. *Journal of the
350   American College of Cardiology*. 2022;79(3):267-279. doi:10.1016/j.jacc.2021.10.045
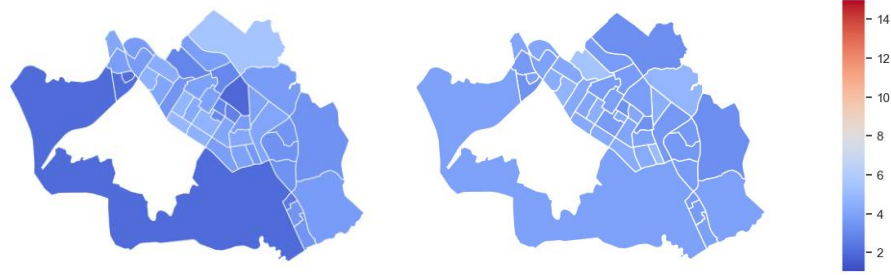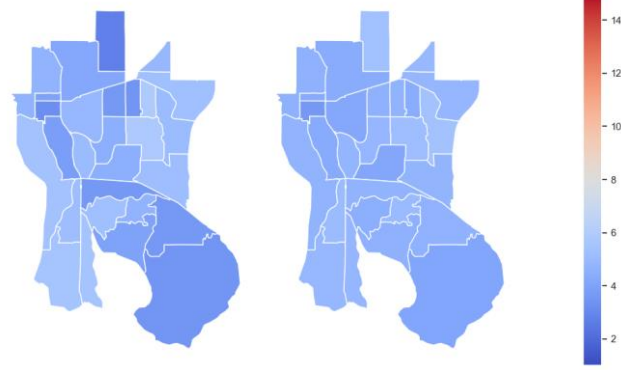
351
352

353 **Figures**
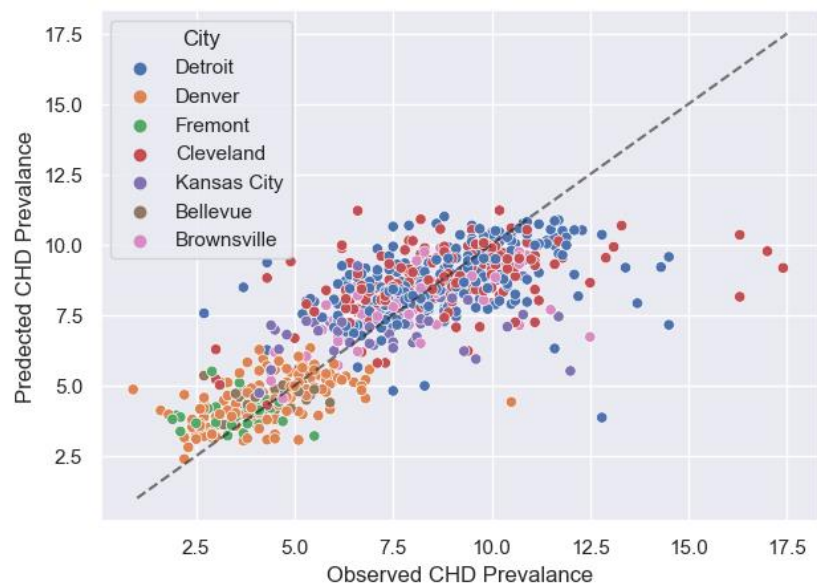
354
355 Figure 1. Scatterplot of the actual estimated (observed) and predicted CHD prevalence (in percentage) in
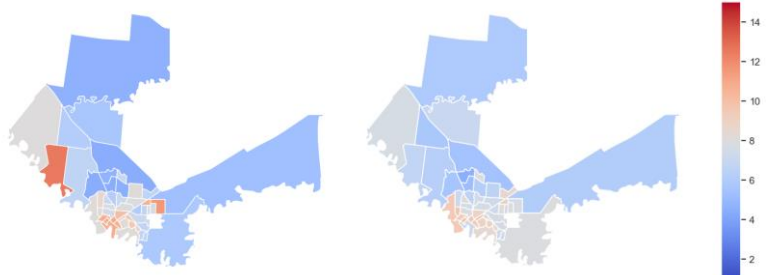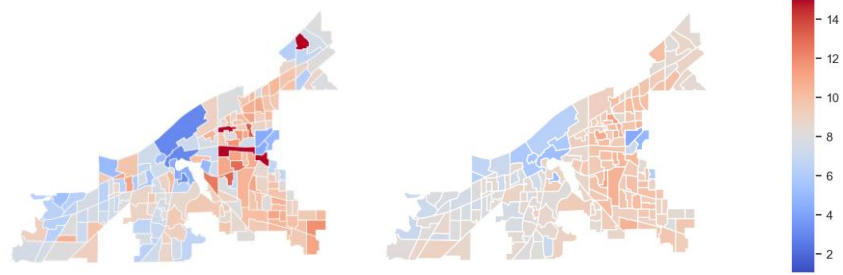356 seven cities. The Black dotted line represents the $y = x$ line.
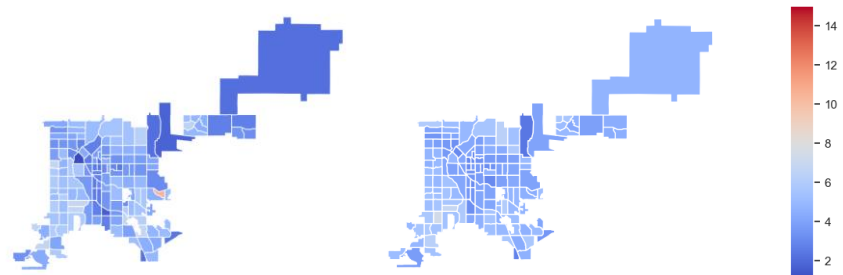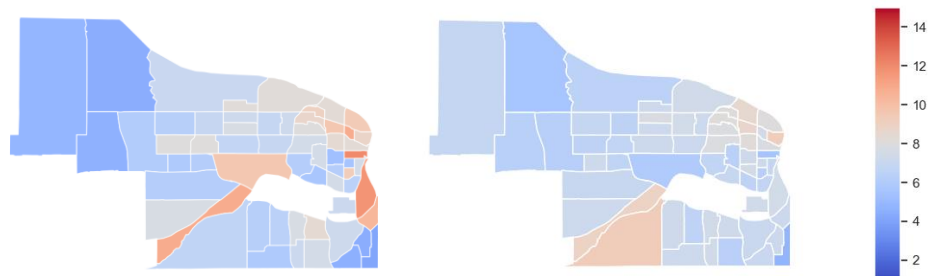357

Fremont, CA

Bellevue, WA

Brownsville, TX

Cleveland, OH

Denver, CO
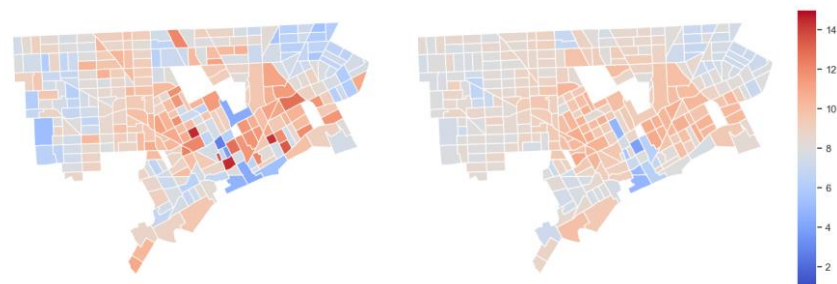
Detroit, MI

Kansas City, KS

Figure 2. Maps of the actual estimates of cardiovascular heart disease (CHD) prevalence (left) and predicted CHD prevalence (right, in percentage). The predicted CHD prevalence is obtained by averaging the results from 100 random trials based on k-fold cross-validation (with k = 10).
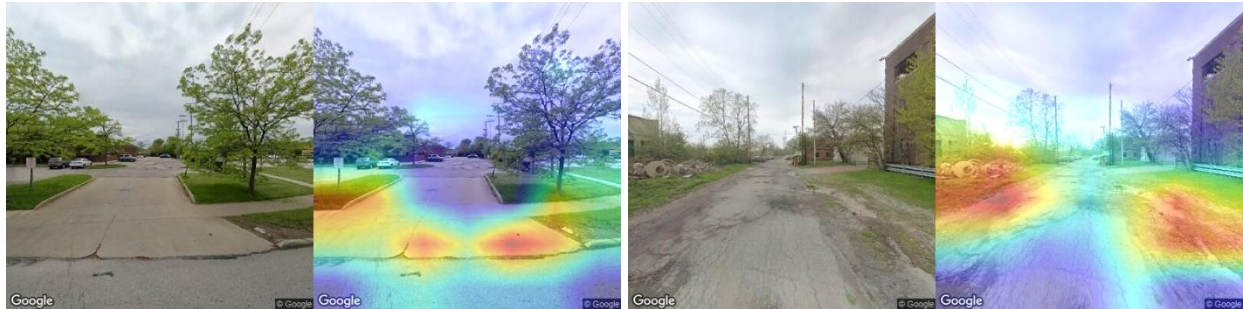
GSV associated with higher CHD

(a) Feature #1555

(b) Feature #484

GSV associated with lower CHD

(c) Feature #204

(d) Feature #1732

Figure 3. Feature interpretations using Grad-CAMs. Images (a) and (b) show the two pairs of GSV images (left) and their activation maps (right) for the features associated with higher CHD prevalence. Images (c) and (d) show the two pairs of GSV images (left) and their activation maps (right) for the features associated with lower CHD prevalence.

Tables

Table 1. AIC, BIC Criteria and Likelihood Ratios Tests (LRT) of LMM models by Model Combination for CHD Prevalence. Models: GSV = the reduced LMM model with only the selected SPLS components CHD: h = 7 obtained from the full CNN features; DSE = the reduced LMM model with only the Demographics and Socio-Economic variables; Combined = the full LMM model with both sets of independent variables from GSV and DSE.

CHD

LMM Model	AIC	BIC	Log. Lik.	Test	LRT	p-value
Combined	757.0	835.0	-368.0			
DSE	930.0	967.0	-457.0	Combined vs. DSE	178.0	< 1.00E-4
Combined	757.0	835.0	-368.0			
GSV	984.0	1029.0	-482.0	Combined vs. GSV	227.0	< 1.00E-4
DSE	930.0	967.0	-457.0			
GSV	984.0	1029.0	-482.0	DSE vs. GSV	49.0	< 1.00E-4

401   Table 2. Total Explained Variance of the LGBM Prediction Model by Model Combination for CHD
402   Prevalence.

CHD	R^2		
LGBM Model	Median	Mean	Std. Error
DSE	0.360	0.149	0.553
GSV	0.591	0.597	0.057

403

404   **SUPPLEMENTARY MATERIAL**

405

406   **Supplementary Information Text**

407   **Multilevel Modeling of Demographics and Socio-Economic Effects**

408   Population Demographics and Socio-Economic status variables are well-known factors associated with

409   the prevalence of disease such as those of interest in this study. To assess the contribution of these

410   effects besides the CNN-extracted features to the crude prevalence of CHD a multilevel-level regression

411   model was built to simultaneously account for the effect of City, as well as variables Age (Median), Sex

412   (Female %), Race (White %), Income (Median $), and Education (< High School %) in addition to the 4096

413   CNN-extracted features. Further, to reduce the dimensionality of the problem and to reduce the effect

414   of noise on the error rates of our inferences, we followed a two-step modeling strategy. In the first step,

415   a multivariate Sparse Partial Least Squares (SPLS) regression[27] was applied to the CNN-extracted

416   features to consider a reduced set of selected SPLS components (latent variables/factors), to which

417   dimension reduction was further applied by imposing sparsity on their loadings (CNN-extracted

418   features). This was done by using shrinkage estimates of regression coefficients (loadings of component)

419   with combined L1- and L2-penalized estimation. The two tuning parameters of the multivariate Sparse

420   PLS regression model are: (i) the number $h \in \left[1, \ min\ (p, (\frac{f-1}{f})n)\right]$ of components that enter the Linear

421   Mixed Effects Model (where f is the number of folds and p is the dimensionality and n is the sample

422   size), and (ii) the sparsity parameter that controls the amount of shrinkage by a combination of the L1-

423   and L2-penalties. Both were tuned simultaneously by 10-fold Cross-Validation. In the second step, a

424   Mixed-Effect regression model was fitted with the first few selected SPLS components augmented with

425   the other Demographics and Socio-Economic variables, all treated here as fixed-effects, and where City

426   was treated as a random effect. Departure from normality of the univariate dependent variable (CHD

427   Prevalence) was tested by EDA analysis and Shapiro-Wilks test. A Box-Cox transformation of the

428   response was applied to minimize departure from normality. Because the dependent variables are

429   continuous, modeling was done by fitting a Linear Mixed Effects Model (LMM). The modeling is entirely

430   supervised because multivariate SPLS regression seeks latent components that not only capture the

431   most variance in the X-space (multivariate independent variables) but also the most covariance with the

432   response Y (univariate dependent variable of disease prevalence). Finally, AIC and BIC criteria as well as

433   Likelihood Ratios Tests (as it applied) were used for model comparison between the full model and the

434   reduced models with the Demographics and Socio-Economic variables alone or the selected SPLS

435   components alone. Goodness of fit results of all models were also compared by the R^2 amount of

436   explained variance achieved by each set of independent variables in the cross validated Random Forest

437   (RF) and Light Gradient Boosted Machine (LGBM) predictive models as described above.

438   **Comparison of CNN Features with Recognized Factors**

439   An optimal model could be obtained after model parameter tuning by cross-validation with a maximum

440   of h = 7 SPLS components and η = 0.6, yielding a final model comprising s = 2031 selected loadings or

441   CNN-extracted features. This resulted in a SPLS model with a cumulative explained variance of CHD:

442    $R_{XY}^2 > 66.7\%$. Additional selection of the CNN-extracted features was done by excluding loadings with

443   0-containing 95% bootstrapped confidence intervals, resulting in a final model of CHD: s = 816 loadings.

444   Model specification only included main random and fixed effects and no random interaction effect was

445   found significant. Also, a structure for the covariance matrix was deemed not necessary (not statistically

446   significantly different).

447   We further refer to the Google Satellite View (GSV) model, as the reduced LMM model where only the

448   selected SPLS components (CHD: h = 7) obtained from the full CNN features enter into the model.

449   Likewise, we refer to the Demographics and Socio-Economic (DSE) model, as the reduced LMM model

450   where only the Demographics and Socio-Economic variables enter into the model. We compared the

451   regression estimates and Goodness of Fit (GOF) measures between these reduced models and the

452   Combined LMM model, where both sets of independent variables enter simultaneously. For all three

453   models, Table 1 shows model comparisons by GOF and Likelihood Ratios Tests (LRT), and Supplemental

454   eTable 1 shows the corresponding regression estimates and ANOVA results. Likewise, Table 2 shows the

455   amount of total explained variance by predictive modeling (Table 2: LGBM, Supplemental eTable2: RF).
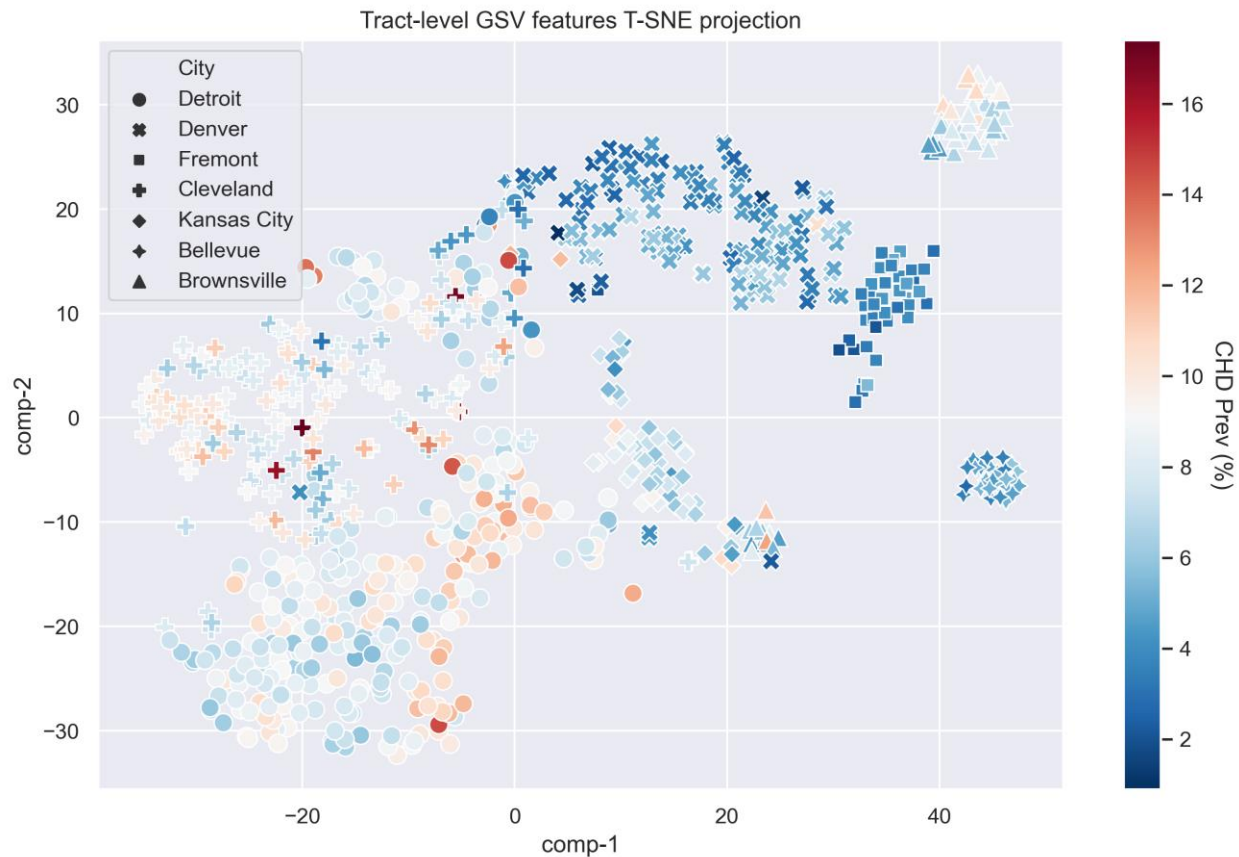
456

457   **Supplementary Figures**
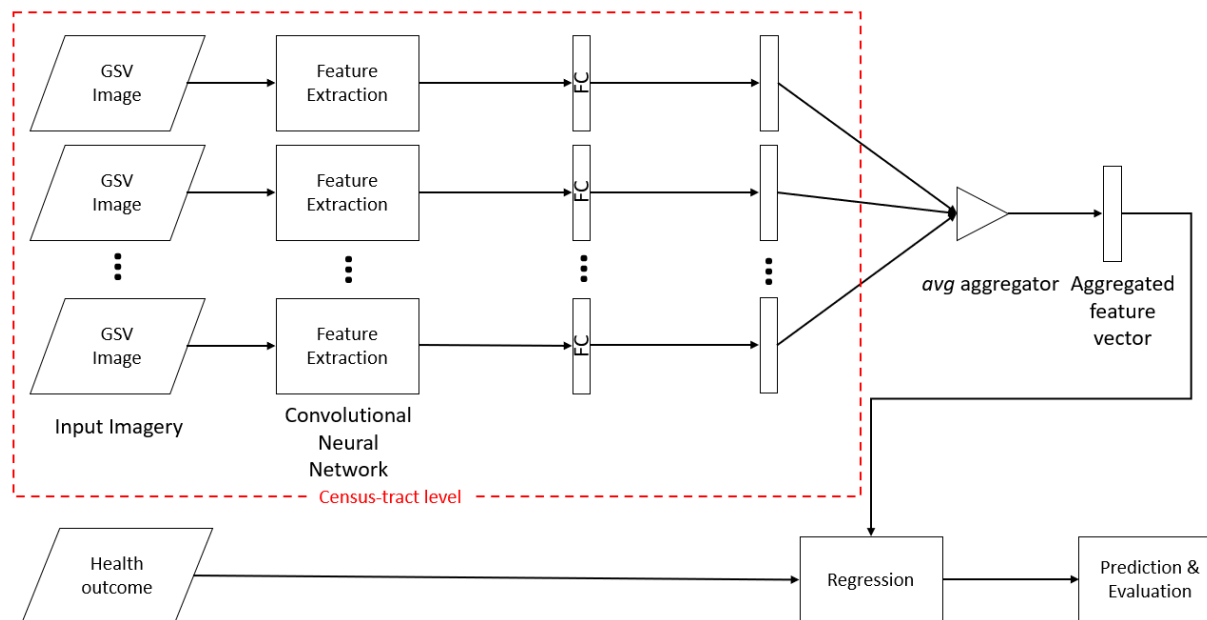
458

459   eFigure 1. Workflow of the GSV feature extract and regression method. A Places365 pre-trained
460   convolutional neural network – ResNet-50 was used as a feature extractor to obtain the deep features
461   from GSV images. The aggregated feature vector was used in regression to predict CHD prevalence.

462

463   eFigure2. T-SNE projection of the 4096 features from GSV data for the tracts of 7 cities. The projected
464   points are colored by CHD prevalence (%) of the census tract.

465

466 **Supplementary Tables**

467 eTable 1. Regression Estimates and ANOVA Results by Model Combination for CHD Prevalence.

CHD	Effects	Coeff.	Std. Error	t-value	p-value	F-value	p-value
Combined	sex	< 0.1	< 0.1	2.870	0.004	3428	< 0.0001
	Age	< 0.1	< 0.1	14.92	< 0.001	150	< 0.0001
	Race	< 0.1	< 0.1	-3.450	0.001	94.000	< 0.0001
	Income	< 0.1	< 0.1	-10.36	< 0.001	515	< 0.0001
	Education	1.2	0.18	6.790	< 0.001	132	< 0.0001
	Comp.1	44.2	19.56	2.260	0.024	52.000	< 0.0001
	Comp.2	-23.2	11.02	-2.11	0.035	6.000	< 0.0181
	Comp.3	36.5	9.65	3.780	< 0.001	59.000	< 0.0001
	Comp.4	32	9.95	3.220	0.001	7.000	0.0109
	Comp.5	-41.2	10.75	-3.83	< 0.001	1.000	0.4663
	Comp.6	35.7	10.25	3.480	0.001	10	0.0014
	Comp.7	49.5	10.67	4.640	< 0.001	21	< 0.0001
DSE	Sex	0.011	0.034	3.22	0.001	18	< 0.0001
	Age	0.042	0.025	17.06	< 0.001	142	< 0.0001
	Race	-0.006	0.001	-6.37	< 0.001	85	< 0.0001
	Income	0	0	-10.42	< 0.001	297	< 0.0001
	Education	1.903	0.181	10.49	< 0.001	110	< 0.0001
GSV	Comp.1	61.8	26.19	2.36	0.019	115	< 0.0001
	Comp.2	-25.4	13.3	-1.91	0.057	54	< 0.0001
	Comp.3	58.7	11.64	5.04	< 0.001	137	< 0.0001
	Comp.4	73.9	12.96	5.7	< 0.001	23	< 0.0001
	Comp.5	-51.9	14.09	-3.69	< 0.001	5	0.0248
	Comp.6	36.2	12.74	2.86	0.004	5	0.0258
	Comp.7	82.2	13.51	6.08	< 0.001	37	< 0.0001

468

469 eTable 2. Total Explained Variance of the RF Prediction Model by Model Combination for CHD
470 Prevalence.

471

CHD	R^2		
RF Model	Median	Mean	Std. Error
DSE	0.328	0.160	0.559
GSV	0.613	0.615	0.045

472