Greenspace and cardiovascular health outcomes in UK Biobank participants: assessing the environmental and physiological pathways

Charlotte Roscoe

Department of Epidemiology and Biostatistics

School of Public Health

Faculty of Medicine

Imperial College London

Thesis submitted for fulfilment of the requirements for

Doctor of Philosophy (PhD)

January 2020

Abstract

Urban greenspace is hypothesised to protect cardiovascular health via multiple mechanistic pathways. These include the *environmental* pathway, via which greenspace attenuates harmful pollution levels (e.g., air and noise pollution); and the *physiological* pathway, via which accessible greenspace is hypothesised to increase physical activity. My PhD thesis explored the impact of greenspace on *environmental* and *physiological* pathway-specific exposures.

I developed pathway-specific exposure models and applied them to participants of a large, adult cohort – UK Biobank (*n* ~500,000). I estimated residential air pollution exposure via a land-use regression model, with surrounding vegetation cover as a predictor variable to assess the *environmental* pathway. To assess the *physiological* pathway, I developed a novel 'green walkability' index within a street/path network buffer around participants' residential addresses. To enhance specificity of greenspace exposure, I assessed greenspace cover surrounding addresses using functional attribute data (e.g., public parks versus private gardens). I conducted survival analyses to examine the associations of greenspace cover surrounding residential addresses with cardiovascular and non-injury mortality, adjusting for relevant individual- and area-level confounders.

Integration of vegetation cover into a 'green walkability' index did not strengthen effect estimates for physical activity participation in UK Biobank participants when compared to a standard index. I examined the interrelationships of greenspace, air pollution, traffic noise and walkability, and showed that ignoring built environment components related to walkability might result in biased greenspace and physical activity effect

estimates. In epidemiological analyses, I showed a protective association across quintiles of surrounding greenspace (100 m circular distance buffer) and non-injury mortality, though not cardiovascular mortality.

My thesis points to important policy implications of exposure interrelationships. Greenspace exposure might protect against premature mortality in older adults, though indirect mechanisms (e.g., air flow and walkability of street networks) should be considered, alongside greenspace, to ameliorate specific exposures on the *environmental* and *physiological* pathways.

Contents

Acknowledgements

Thank you Daniela Fecht, John Gulliver and Susan Hodgson for, among other qualities, your patience and compassion during times of difficulty. Thank you also to Paolo Vineis for inviting me to join your team during my PhD.

Imperial Girls: Annalisa Sheehan, Aina Roco Barceló, Sophie Hamilton, Honor Bixby and Weiyi Wang – thank you for everything. I am forever grateful for your support and kindness. Annie, in particular, you were my rock during thesis write up, and an incredibly wonderful boat-mate!

Causality Journal Club taught me some epidemiology, and a touch of philosophy, thank you in particular to Tom Bond and Jessica Laine. Thank you to Samuel Cai for your input on the epidemiological chapter of my thesis. Thank you also to Zhenchun Yang for all the chats in Room 171. Special thanks to Charlotte Sheridan – it was a pleasure to brainstorm walkability with you!

For the talent and energy that you invested in GREEN SPACE, thank you to my wonderful collaborators and friends, Enya Lachman Curl and Robbie Thompson. Thank you also to Rupert Cole, Eduardo Pombo Seleiro and Liz Hayes for supporting GREEN SPACE. Thank you to Robbie Parks for pushing me in my final year, and helping me to believe in myself. Thank you for all the music.

Finally, thank you to my family, who always allow me to find my own path.

List of tables

List of figures

List of equations

List of abbreviations

ADMS	Atmospheric dispersion modelling system
CERC	Cambridge Environmental Research Consultants
CRTN	Calculation of Road Traffic Noise method
CEH	Centre for Ecology and Hydrology, UK
CVD	Cardiovascular disease
GAM	Generalized additive model
GIS	Geographic information system
IHD	Ischaemic heart disease
LAEI	London Atmospheric Emissions Inventory
LUR	Land use regression
ONS	Office for National Statistics
OS	Ordnance Survey
PM	Particulate matter
REML	Restricted maximum likelihood
RMSE	Root Mean Square Error
SASHU	Small Area Health and Statistics Unit
SDG	Sustainable development goal
SES	Socioeconomic status

UFORE	Urban Forest Effects (UFORE) model
WHO	World Health Organisation

Copyright statement

Declaration of originality

I hereby declare that the work in this thesis is my own original research and that I have appropriately cited any work that is not my own.

1 Introduction

In the United Kingdom, over 80% of the population lived in urban areas in 2018 (UN DESA, 2019). While high-density urban living offers social and economic opportunity, city inhabitants are exposed to a mixture of environmental features and pollutants that impact health. For example, cardiovascular health is adversely impacted by exposure to air pollution and traffic noise (Cai et al., 2018), which are the leading environmental risk factors for health in Europe (Hanninen et al., 2014), whereas exposure to some urban features, such as walkable neighbourhoods (Nieuwenhuijsen, 2018) and urban greenspace (James et al., 2015), have been shown to protect cardiovascular health.

High density urban populations provide an opportunity for government to improve the health of a large proportion of the national populace through urban planning and interventions. When health is adequately prioritised in decision-making, evidence suggests that small improvements in urban environmental exposures – to which many individuals in the population are exposed – can equate to large health benefits (Mueller et al., 2017b), and accompanying health cost savings (Pimpin et al., 2018).

Natural environments in cities, including urban greenspace, have been associated with health and wellbeing benefits, and an accumulating body of evidence suggests urban green spaces might buffer some of the harmful health effects associated with urban living (James et al., 2015, Markevych et al., 2017, van den Berg et al., 2015). Greenspace is defined as land predominantly covered with vegetation and, in the urban context, includes public parks, private gardens, playing fields, religious grounds, allotments, sports grounds and street trees. In a review of evidence, Nieuwenhuijsen et al. (2017b) outlined several mechanistic pathways linking urban greenspace and better health – broadly, these pertain to physical activity enhancement, psychological wellbeing and social enhancement, microbial exposure diversification, and reduction in exposure to environmental pollutants, such as air pollution and traffic noise (**Figure 1**).

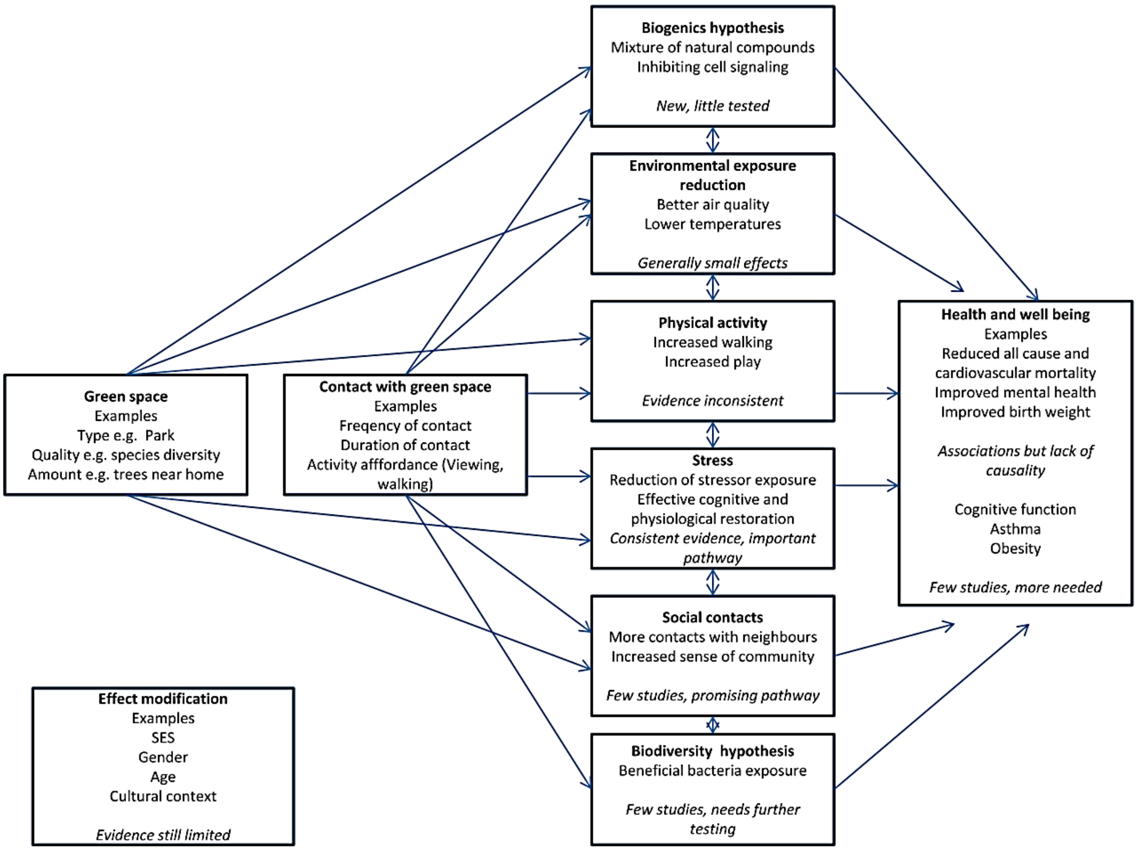

Biogenics hypothesis
Mixture of natural compounds
Inhibiting cell signaling
New, little tested
Environmental exposure reduction
Better air quality
Lower temperatures
Generally small effects
Physical activity
Increased walking
Increased play
Evidence inconsistent
Stress
Reduction of stressor exposure
Effective cognitive and physiological restoration
Consistent evidence, important pathway
Social contacts
More contacts with neighbours
Increased sense of community
Few studies, promising pathway
Biodiversity hypothesis
Beneficial bacteria exposure
Few studies, needs further testing
Green space
Examples
Type e.g. Park
Quality e.g. species diversity
Amount e.g. trees near home
Contact with green space
Examples
Freqency of contact
Duration of contact
Activity afffordance (Viewing, walking)
Health and well being
Examples
Reduced all cause and cardiovascular mortality
Improved mental health
Improved birth weight
Associations but lack of causality
Cognitive function
Asthma
Obesity
Few studies, more needed
Effect modification
Examples
SES
Gender
Age
Cultural context
Evidence still limited

Figure 1. Conceptual diagram by Nieuwenhuijsen et al. (2017b) linking greenspace via mechanistic pathways to health outcomes; status of evidence in italics.

The term *physiological* pathway is used in the context of *physical activity* in this thesis. While exposures on the *environmental* pathway (e.g., air pollution and traffic noise) are ‘passive’, the physiological pathway involves behavioural aspects (i.e. choice to participate in physical activity). Both, exposure to environmental pollutants and physical activity can result in downstream physiological impacts within the body. For example, exposure to some air pollutants (e.g., nitrogen dioxide) can result in oxidative stress; a harmful physiological response involving cell damage by free radicals (Kelly, 2003). In contrast, exposure to regular physical activity increases endogenous antioxidant activity, which rebalances antioxidant–prooxidant levels, and potentially provides protection from CVD (Elosua et al., 2003). Evidence suggests physiological responses to stress are improved through regular physical activity. Physiological biomarkers of stress, such as cortisol, have also been associated with greenspace exposure (Ward Thompson et al., 2012). Beneficial changes in behavioural physiological responses (e.g., response to stress) associated with regular physical activity, which in the context of this study are intertwined with non-focal greenspace pathway exposures, such as psychological mechanisms linking greenspace and health, are embedded within the *physiological* pathway. That is, physical activity, and improved (behavioural) physiological response to stress associated with physical activity, are encompassed within the *physiological* pathway.

Greenspace has long been considered beneficial for urban populations. *Rus in urbe* – meaning to create the illusion of countryside in the city – was originally coined by the Romans, and naturalistic landscapes are deeply embedded in the history of modern day urbanism (Gehl, 2010). In 19th Century Britain, municipal parks were made publically accessible to combat poor standards of urban living. Around this time, William Pitt the Elder coined the term ‘green lungs’ of the city – a metaphor that pervaded urban planning, both nationally (e.g., Victoria Park, London) and globally (e.g., Central Park, New York City), and was founded on limited mechanistic understanding of underlying beneficial processes. In alignment with *green lungs* rhetoric, patch size of greenspace has been shown to positively associated with the respiratory and cardiovascular health of city inhabitants via air pollution reduction (Shen and Lung, 2016).

Patrick Geddes proposed an interconnected network of small green spaces, acting as the ‘green lungs’ of Edinburgh’s cramped Old Town in the late 19th Century, and, throughout the 20th Century, greenspace provisioning expanded beyond the design and preservation of large parks to the creation of more numerous, smaller patches of vegetation distributed throughout community neighbourhoods. This shift encouraged both deliberate greenspace exposure (e.g., park visitation) and incidental exposure (e.g., regular exposure from *passing through* greenspace in the vicinity of residential addresses).

Though greenspace provisioning was often driven by aesthetic goals, and benefits were expected from deliberate greenspace visitation, contemporary evidence using detailed data on street trees is supportive of the benefits of incidental exposure. For example, by linking the residential addresses of respondents of the London Travel Demand Survey and data on trees surrounding respondents’ addresses, a study by Sarkar et al. (2015b) found a positive association of street tree density and odds of physical activity (walking). In this example, respondents’ exposure was presumably incidental; they did not walk *to* the trees surrounding their address, but walked *through* the trees to access other destinations in their local neighbourhood.

Data used in epidemiological analyses of greenspace are typically linked to the place of residence of study participants via spatial identifiers such as addresses or neighbourhood. Such spatial linkage requires geospatial methods. Geographic information systems (GIS) have become a widely utilised tool for the assessment of spatial variables in environmental epidemiology because of their capacity to capture, process and query the spatial components of environment and health data at different scales (Labib et al., 2019, Markevych et al., 2017). Using GIS, exposure to greenspace is typically assessed within a specific distance around an individual's place of residence (e.g., within a distance buffer), or in the corresponding administrative boundary, depending on the available health data for epidemiological analysis.

A seminal ecological study conducted on the English population showed that greenspace cover at the small-area level (administrative boundaries) was associated with reduced premature mortality, particularly for circulatory disease mortality (Lachowycz and Jones, 2014a, Mitchell and Popham, 2008a). Findings were corroborated by large scale cohort studies that used residential address exposure to greenspace (within circular distance buffers) in Canada (Villeneuve et al., 2012a) and the USA (James et al., 2016b). Non-fatal outcomes have also shown associations with greenspace exposure, including incidence of cardiovascular disease (CVD) (e.g., Donovan et al., 2015) and CVD risk factors – such as obesity (e.g., Sarkar, 2017), physical activity (e.g., Mytton et al., 2012), and metabolic syndrome (e.g., de Keijzer et al., 2019).

Scale is a critical factor in the investigation of greenspace and health. Studies conducted in England (Bixby et al., 2015) and the USA (Richardson et al., 2012) found no association of premature all-cause and circulatory disease mortality with greenness at the city scale. These findings demonstrate that the health effects of urban greenspace observed at the neighbourhood scale (e.g., Mitchell and Popham, 2008b), are not transferred to the city scale. This is unsurprising given the hypothesised pathways, such as the environmental and physiological (physical activity) pathway, which link greenspace and health Nieuwenhuijsen et al. (2017b). To illustrate this point, one can consider individuals who live in the East of a city, who are unlikely to benefit (on a regular basis) from a large park in the West of the city; reduction in traffic related air pollution and noise, provision of space for physical activity, and most psychological and social benefits offered by the park in the West do not reach those residing in the East. Though the reach of greenspace effect on health might not decrease linearly with distance, and might depend on individual mobility (high socioeconomic status groups tend to be most mobile), the distribution and accessibility of greenspace is an important consideration for improving health.

To promote greenspace accessibility, UK planning guidelines such as the Accessible Natural Greenspace Standard (ANGSt) were developed by Natural England in the early 1990s (English Nature Research Reports, 2008). ANGSt guidelines were based on research into the distances individuals were willing to travel to access the natural environment. The standard was reviewed in 2008 and supplemented with further guidance in 2010 (Natural England, 2010). The criteria recommended that all individuals should have: a) an accessible natural greenspace of at least 2 hectares in size, 300 metres (m) or less from their place of residence; b) at least one accessible 20 hectare site within two kilometres (km) of their place of residence; c) one accessible 100 hectare site within five km of their place of residence; d) one accessible 500 hectare site within ten km of their place of residence; and e) a minimum of one hectare of statutory Local Nature Reserves per thousand population.

The ANGSt criteria incorporated the notion of cumulative opportunities, whereby access to greenspace with specific qualities or attributes is assessed at multiple distances, with willingness to travel a certain distance to a particular greenspace type dependent upon its designation or size. This cumulative approach to assessing access to different types of greenspace has a strong advantage over simpler guidelines. For example, a guideline of 1 hectare of greenspace within 300 m of all residences has been suggested as a European standard (Annerstedt van den Bosch et al., 2016). Such simplistic guidelines, however, could inadvertently encourage developers to achieve standards by creating high-rise blocks around homogenous parks (e.g., 1 hectare lawn), with little regard for variation in types and sizes of greenspace at layered proximities (Ekkel and de Vries, 2017).

Improving access to greenspace in cities was included in the United Nations (UN) Sustainable Development Goal (SDGs) 11.7 (United Nations General Assembly, 2015). In response, city leaders have rolled out large-scale greenspace interventions to ameliorate urban environments and health. For example, *National Park City* is a flagship urban greenspace scheme by the Mayor of London, Sadiq Khan, with the broad goal to, “make life in London better”. The Mayor has also pledged to plant two million trees across the city while in office (Greater London Authority, 2019). The effectiveness of large-scale urban greening schemes for health will depend on the research evidence, and translation to policy, on which they are founded.

In 2016, the World Health Organisation released a review of evidence on greenspace and health (WHO Regional Office for Europe, 2016), which summarised various mechanistic pathways leading to health effects, and authors advocated “the implementation and evaluation of targeted, evidence-based green space interventions for the health promotion of urban residents”. This call for evidence, and the recent surge in public health studies involving greenspace is, in part, driven by contemporary data availability. A range of data sources are currently available to represent greenspace exposure, including land cover data, tree canopy data and greenness indices derived from satellite images. These satellite-based indices, including a commonly used objective measure of greenness, the Normalized Difference Vegetation Index (NDVI), have allowed for unprecedented advances in greenspace exposure assessment. The NDVI is calculated using satellite imagery, whereby living vegetation is distinguished from other land cover via a combination of absorption of visible light (0.4 to 0.7 μm) by chlorophyll in photosynthesis, and reflectance of near-infrared light (0.7 to 1.1 μm) by leaf cell structures.

$$NDVI = \frac{NIR - VIS}{NIR + VIS}$$

Equation 1. Normalized Difference Vegetation Index (NDVI) is calculated by dividing near-infra red light (NIR) minus visible light (VIS) by the sum of near-infra red light and visible light.

The simple NDVI calculation (**Equation 1**) produces values ranging from −1 to 1, with water bodies corresponding to values near −1, bare ground and snow below 0.1, grassy areas ranging from 0.2–0.3, and high-density vegetation (e.g., forest cover) corresponding to values over 0.6 (**Figure 2**).

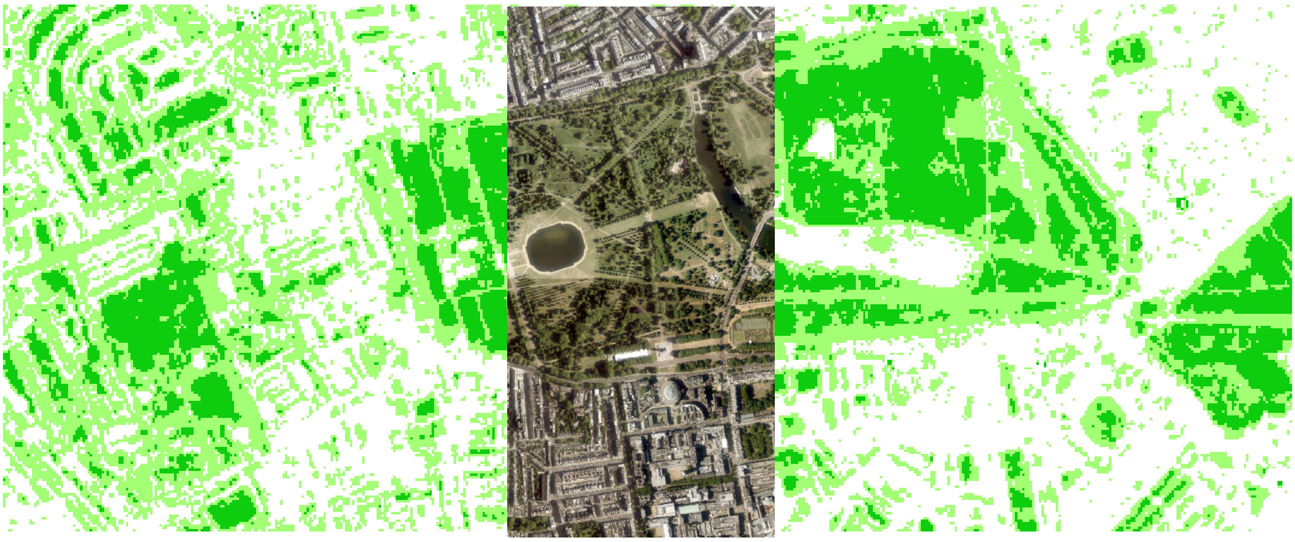


Figure 2. Hyde Park, London, and surrounding area captured via high resolution (25 cm x 25 cm) aerial photography (central panel; JPG geospatial data, updated 29/10/2018; openly accessible via EDINA Aerial Digimap Service: https://digimap.edina.ac.uk) overlaid on Normalised Difference Vegetation Index (NDVI) data (10 m x 10 m resolution) captured via Sentinel 2 satellite images (sensed 26/02/2019; openly accessible via https://scihub.copernicus.eu). In the left and right panel, water, concrete and bare ground are shown in white, sparse vegetation (0.2—0.6 NDVI values) are shown in lighter green and dense vegetation (values over 0.6 NDVI) are shown in darker green.

Indices of greenness have major advantages, including ubiquitous data coverage over large areas at an adequately fine scale for epidemiological analysis. However, simplifying *greenspace* to *greenness* does not permit investigation of the functional role or specific attributes of greenspace driving health associations, and furthermore might act as a barrier to translating research into practice due to limited specificity (Rugel et al., 2017b). For example, urban planners might increase NDVI by adding more street trees, or more sports grounds, or more public gardens into cities – but how should they choose the appropriate greenspace type for health? To answer this question requires greater specificity than NDVI in greenspace exposure assessment. That is, in light of the mechanistic pathways outlined above, some greenspace attributes and types have the potential to be particularly relevant to different pathways (e.g., sports grounds for the *physiological* pathway). Crude indices of greenness such as the NDVI might be limited for furthering understanding of greenspace and health associations.

1.1 Rationale

Interventions to improve urban environmental quality in countries like the United Kingdom are primarily driven by motivations to protect and promote health and wellbeing. Estimating health effects of urban environmental exposures is therefore a crucial input to public health research and urban planning. While evidence of harmful associations of environmental exposures, such as air pollution, and health are well understood (Brook et al., 2010a, Cohen et al., 2017, Hoek et al., 2013), there is a less consistent and comprehensive evidence base available on salutogenic features of the urban environment, such as greenspace (Nieuwenhuijsen et al., 2017b).

There is accumulating evidence of a protective effect of greenspace on cardiovascular health outcomes (Bowler et al., 2010, Fong et al., 2018, Gascon et al., 2016, James et al., 2015), though typically studies rely on non-specific measures of greenness (e.g., Normalised Difference Vegetation Index), which can limit inference on the contribution of specific pathways to the total greenspace-health effect. The *environmental* pathway, for example, which is hypothesised to impact cardiovascular health via regulation of the local environment (e.g., by reducing exposure to air pollution and noise), could be better understood by using categorized vegetation data (e.g., tree versus groundcover), as oppose to non-specific vegetation cover. That is, the height and density of vegetation can alter (i.e. increase or decrease) pollution concentrations depending on neighbouring built features and meteorology (Abhijith et al., 2017), therefore height-categorised data might provide further insight. Whereas the *physiological* pathway, via which greenspace is hypothesised to facilitate physical activity, might be better understood by using: a) data that includes functional categorisation of greenspace (e.g., public parks and sports grounds versus private residential gardens) and b) assessment of high resolution vegetation data in concert with known drivers of physical activity in the urban context (e.g., walkability metrics).

The rationale for focusing on these two mechanistic greenspace pathways in this PhD thesis, as oppose to others, are two-fold: firstly, available height-categorised greenspace data for London was particularly suited to the study of these pathways; secondly, UK Biobank has rich data on physical activity, including self-report and accelerometry data. Therefore, although the psychosocial pathway is likely to strongly mediate greenspace and health associations (Banay et al., 2019, James et al., 2016a, Rugel et al., 2019), geographical and UK Biobank data available were not optimal for study of this pathway. Exploration of the *environmental* and *physiological* pathways offered the opportunity to train in exposure assessment methods while adding novel insight to two hypothesised greenspace and health pathways. The interplay of greenspace with built environment features (e.g., buildings, neighbourhood design) and health could also be demonstrated in exposure assessment via focusing on these pathways. Further, as previously noted, small improvements in physical activity and air quality in densely populated cities can offer vast improvements in public health.

Furthermore, urban environmental exposures, including greenspace, air pollution, traffic noise and walkability are often assessed in isolation, despite impacting health as a complex mixture of interrelated urban exposures. Interrelation of exposures can be explored via statistical analysis of exposures (e.g., correlation), and can also be explored at the exposure assessment stage via integration of exposures (e.g., integration of vegetation into air pollution and walkability models). By studying greenspace integration in established exposure assessment fields (i.e. air pollution and walkability), one might estimate greenspace contribution along specific hypothesised mechanistic pathways, namely, the *environmental* and *physiological* pathways.

Identifying pathway-specific contributions is crucial to inform effective planning and health policies. Research in this area is also motivated by global urbanisation trends, which will exacerbate effects of urban environmental planning on health as urban populations increase. The New Urban Agenda, adopted at Habitat III 2016 (United Nations, 2017), committed members to “promoting the creation and maintenance of well-connected and well distributed networks of open, multipurpose, safe, inclusive, accessible, green and quality public spaces” and “to improving […] physical and mental health, and household and ambient air quality, to reducing noise and promoting attractive and liveable cities”, with a strong emphasis on urban environmental equality and social cohesion.

The rationale for conducting this PhD research therefore was to deepen understanding of specific mechanistic pathways linking greenspace and cardiovascular health by using detailed greenspace data across multiple exposure assessment methods associated specifically with the *environmental* and *physiological* pathways.

1.2 Aims and objectives

I aim to estimate the effect of *environmental* and *physiological* pathway-specific exposures on cardiovascular outcomes in UK Biobank using pathway-relevant exposures, including air pollution and walkable neighbourhood exposures.

In order to achieve this, specific objectives relating to the assessment of *environmental* and *physiological* pathway-specific exposures at UK Biobank participant residential addresses, include:

1. To assign high-resolution, categorised greenspace exposure metrics in multiple distance buffers at residential address-level;
2. To assess traffic-related air pollution and noise levels at residential address using high-resolution models;
3. To assess neighbourhood walkability at residential address using high resolution data within a street-network buffer;
4. To examine the interrelationship of the greenspace, walkability, air pollution and traffic noise exposures assigned to UK Biobank participants at residential address;
5. To demonstrate the use of address-level exposures assigned in this PhD project in epidemiological analyses of non-injury and cardiovascular mortality using UK Biobank record linkage.

1.3 Structure of the thesis

In **Chapter 2**, I critically review relevant, and pathway-specific, literature to provide the reader with an up-to-date overview of current greenspace and cardiovascular health research, as well as pathway-specific exposure assessment techniques. I then summarise, in **Chapter 3**, the UK Biobank cohort and the process of acquiring participant addresses for environmental exposure assessment from UK Biobank. In **Chapter 4**, I describe exposure assessment of greenspace cover and neighbourhood deprivation. Alongside exposure assessment methods, I summarise corresponding high-resolution greenspace data available for this project. In **Chapter 5**, I describe exposure assessment methods used to assess the environmental pathway relevant exposures (i.e. air pollution and traffic noise). I also summarise exposure assessment results and discuss air pollution model variables in detail. The physiological pathway exposure assessment, namely walkability analysis using a transport network buffer, is described in **Chapter 6**, and I provide validation of walkability scores using physical activity data from UK Biobank. In **Chapter 7**, I present findings from the UK Biobank address-level exposure assessment, including assessment of greenspace, air pollution, traffic noise and walkability, with a focus on the interrelation of these exposures and their relationship with neighbourhood-level deprivation. In **Chapter 8**, I present the findings of the association of greenspace and premature mortality in UK Biobank participants residing in England. I conclude, in **Chapter 9**, with an overview of my findings, and the implications of my findings from a public health perspective. I also suggest some potential future research directions.

2 Background: Greenspace and cardiovascular health

Cardiovascular disease (CVD) is a collective term for diseases affecting the heart and circulatory system. In the UK, CVD is responsible for 28% of deaths each year (BHF, 2019). CVD incidence and mortality and behavioural risk factors (e.g., physical inactivity) have been linked to multiple environmental exposures. Ischaemic heart disease (IHD) and stroke, for example, have an estimated 35% and 42% environmental contribution, respectively (Prüss-Üstün et al., 2016). Evidence is growing on the potential biological mechanisms via which environmental exposures impact the heart and circulatory system. Long-term exposure to air pollution and traffic noise, for example, might induce development of carotid atherosclerosis, which is a common precursor to IHD and ischemic stroke incidence (Brauer, 2015, Kaufman et al., 2016).

Political interest in the role of the natural environment in achieving better health outcomes has been revived following an increase in studies investigating associations of greenspace exposure and health, especially in urban contexts (Parliamentary Office of Science and Technology, 2016). Meta-analyses and reviews synthesising these associations have found that natural environments, including urban greenspace, are largely salutogenic, despite some disservices, including allergy (Dadvand and Nieuwenhuijsen, 2018, Fong et al., 2018, Gascon et al., 2015, Gascon et al., 2016, James et al., 2015, Labib et al., 2019, Markevych et al., 2017, Twohig-Bennett and Jones, 2018). More specifically, studies have shown that urban greenness (e.g., NDVI) surrounding residential addresses is associated with reduced risks of CVD incidence and CVD mortality (James et al., 2016a, Mitchell and Popham, 2008b, Seo et al., 2019, Villeneuve et al., 2012a) and CVD risk factors, such as diabetes (Astell-Burt et al., 2014, Dalton et al., 2016, Klompmaker et al., 2019b) and obesity (Lachowycz and Jones, 2011, Sarkar, 2017).

Using data on a cohort of retired civil servants in the UK (Whitehall II), de Keijzer et al. (2019) showed an interquartile range increase in greenness surrounding participants' residential addresses (NDVI in a 500 m distance buffer) was associated with a 13% (95% confidence interval (CI): 1%, 23%) lower risk of metabolic syndrome, which is an important CVD risk factor. Protective associations with higher levels of vegetation cover were found for all components of metabolic syndrome, including lower risk of high levels of fasting glucose, high triglyceride levels, low high-density lipoprotein cholesterol, and hypertension.

A study of the English population below retirement age showed greenspace cover was associated with lower all-cause and circulatory disease mortality at the small-area level (Mitchell and Popham, 2008b). Findings were suggestive of a more pronounced effect of greenspace on health in more deprived sub-populations. For example, the Incidence Rate Ratio (IRR) for cardiovascular mortality for the most income-deprived sub-population (quartile) versus the least deprived was 2·19 (95% CI = 2·04, 2·34) in the least green areas, whereas it was 1·54 (95% CI = 1·38, 1·73) in the most green (Mitchell and Popham, 2008b). A study in the US found that greenspace cover attenuated the effect of air pollution ($PM_{2.5}$) on cardiovascular mortality, though only in deprived areas (Yitshak-Sade et al., 2019). The interrelationship of greenspace alongside other built environment exposures (e.g., air pollution), and deprivation requires careful consideration.

In this PhD thesis I focus on the *environmental* and *physiological* pathways linking greenspace and health. Greenspace is hypothesized to protect health, including cardiovascular health, via multiple mechanistic pathways, which include: the *environmental* pathway, via which vegetation in greenspace is thought to reduce exposure to harmful pollutants (e.g. air pollution and noise); and the *physiological* pathway, via which accessible greenspace is hypothesized to facilitate physical activity (e.g., walking, jogging and cycling), with accompanying health benefits (Nieuwenhuijsen et al., 2017b).

To date, evidence for the *environmental* pathway is suggestive of small protective effect sizes (Nieuwenhuijsen et al., 2017b). However, major public health improvements could be provided by minor improvements in important environmental determinants for health (e.g., air pollution and traffic noise), particularly in associated, high prevalence diseases, such as IHD and stroke (BHF, 2019).

Physical activity is also an important determinant of cardiovascular health. though evidence of associations with greenspace via the *physiological* pathway is inconsistent (Nieuwenhuijsen et al., 2017b). Physical activity was suggested to be *the* most important exposure for improving urban population health in a recent urban Health Impact Assessment of a western European city (Mueller et al., 2017a). The protective effect of physical activity against chronic conditions, including cardiovascular disease, has led to research into strategic interventions aimed at encouraging physical activity in populations via ‘active neighbourhood’ urban planning (Nieuwenhuijsen, 2016). Again, small improvements in physical activity levels across the population would lead to substantial cardiovascular health benefits (UK Chief Medical Officers, 2019).

In the following sections of this review, I critically review specific literature on the *environmental* and *physiological* pathways in detail, and argue that our understanding of greenspace and health associations can be deepened via the use of existing, well-developed environmental exposure assessment methods related to specific pathways.

2.1 Environmental (air pollution and noise) pathway

Exposure to ambient air pollution increases morbidity and mortality, and is the global leading environmental determinant of health (Cohen et al., 2017). Associations of long-term outdoor air pollution and a range of cardiovascular outcomes have been identified through epidemiological investigation (Newby et al., 2015). Susceptible individuals, such as older people and those with pre-existing conditions such as diabetes, high blood pressure and cardiovascular disease, are at higher risk than others of myocardial infarction, arrhythmias and stroke when exposed to elevated levels of air pollution (Brook et al., 2010a). For example, an increased risk of stroke in older people with long-term exposures to fine particulate matter (aerodynamic diameter ≤2.5 μm; $PM_{2.5}$) was reported by a harmonized exposure assessment study of 11 European cohorts (ESCAPE; Beelen et al., 2014).

Cesaroni et al. (2014) reported positive associations of annual mean concentrations of course particulate matter (aerodynamic diameter ≤10 μm; PM_{10}) modelled in ESCAPE and coronary events. The authors also highlighted that risks were increased at levels below the European Limit Value (ELV) of 40 μg/m^3 (hazard ratio (HR) 1.12 (95% CI 1.00, 1.27), for a 10 μg/m^3 increase in PM_{10}). In 2018, 7 out of 43 reporting zones in the UK were compliant with the ELV for NO_2 of 40 μg/m^3, and London had some of the worst pollution levels in the country (DEFRA, 2019). Based on the rate of downward trends in the capital between 2010 and 2016, London's roads are predicted to require between 2 and 193 years (average of 21 years) to attain the ELV for NO2 (Font et al., 2019).

2.1.1 *Relationship of air pollution and vegetation*

Ambient air pollution is a mixture of gases and airborne particulate matter (PM). By convention, particulate matter is categorised into three major groups, irrespective of their source: particulate matter 10 (PM_{10}) is the mass of all particles with an aerodynamic diameter of <10 µm; particulate matter 2.5 ($PM_{2.5}$) is the mass of all particles with an aerodynamic diameter of <2.5 µm; ultrafine particles are 100nm or less. Pollutants that are directly emitted from a source, as opposed to formed via secondary atmospheric processes (e.g., ozone), are particularly problematic in UK cities. These primary pollutants include nitrogen dioxide gas (NO_2) and black carbon, which are directly emitted from combustion engines (e.g., diesel powered vehicles).

Emissions from road traffic sources are the main driver of air quality in UK urban environments; other sources of ambient pollution include industry, aviation and biomass burning. From point of source, air pollution concentrations are impacted by meteorological conditions, as well as the physical properties of the near source environment (Wania et al., 2012). These physical properties – such as greenspace and buildings – within the vicinity of emission sources affect the deposition and dispersion of pollutants and therefore the spatial variation of pollutant concentrations.

Janhäll (2015) reviewed the impact of vegetation on the deposition and dispersion of pollutants in the urban context. *Deposition* of air pollution is the transfer of airborne particles or gas molecules to the surface of vegetation (and all other built environment surfaces) when they come into physical contact or close proximity; the absorption of gaseous pollutants through leaf stomata is considered a form of deposition. *Dispersion* of air pollution is the movement of pollutants in space, and is dependent on ventilation and airflow.

The distance from source, height and porosity of vegetation, and its interaction with built features, can impact pollution concentrations in the built environment (**Figure 3**) (Abhijith et al., 2017). For example, in open road configurations, tall and dense vegetation near to roads can act as a barrier to divert pollutants upwards, away from pavement-level and neighbouring buildings (directed dispersion) (Tong et al., 2016). Conversely, pavement-level pollution concentrations can be elevated in built environments where dispersion of pollutants is limited, such as in a built-up street canyon with perpendicular wind direction, where concentrations can be augmented by poorly positioned vegetation, such as overhanging trees that trap pollutants by limiting ventilation (Vos et al., 2013).

The relationship of built and green infrastructure in altering pollution concentrations at the micro-scale has been explored in dispersion models (e.g. Wania et al., 2012). Recommendations based on dispersion modelling in street canyon scenarios include the planting of low-lying vegetation and hedges, and the planting of well-spaced or isolated trees to avoid entrapment of pollutants (Vos et al., 2013, Wania et al., 2012). Deposition of pollutants on vegetation can be maximised through porous designs that neither redirect airflow (as solid barriers), nor reduce ventilation, though do allow for pollutant deposition. A study by Pugh et al. (2012) suggested that in street canyons where air circulation is carefully considered and deposition velocities to the surface of vegetation are maximized, green walls could offer reductions of 40% and 60% for nitrogen dioxide (NO_2) and course particulate matter (PM_{10}), respectively.

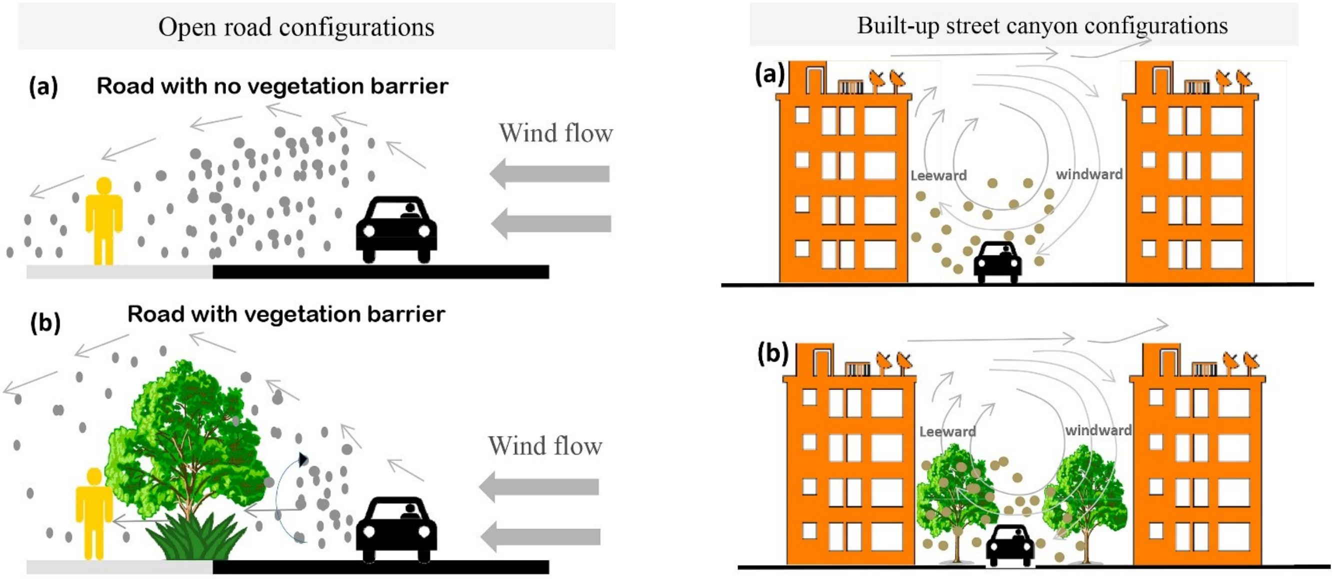

Open road configurations
(a) Road with no vegetation barrier
Wind flow
(b) Road with vegetation barrier
Wind flow
Built-up street canyon configurations
(a)
Leeward
windward
(b)
Leeward
windward

Figure 3. Open road and built-up street canyon configurations with: (a) no vegetation vs. (b) tall (tree cover) roadside vegetation. Placement of vegetation between pavement and traffic can redirect pollution (directed dispersion) away from pedestrians (open road configuration; lower left panel) or entrap pollutants by reducing airflow (street canyon configuration; lower right). Adapted from Abhijith et al. 2017.

The spatial and temporal scale of analysis is of critical importance in accurately determining the relationship of vegetation and air pollution. The effect of trees on air pollution concentrations is potentially bi-directional (deposition versus entrapment) depending on built environment context and scale. At the street scale, particularly in canyon configurations, trees might weaken overall positive associations of greenspace and air pollution due to neutral or negative associations when poorly placed (Wang et al., 2018). Whereas at the national scale tree cover is associated with net-reductions in air pollution (Nowak et al., 2008). Furthermore, inter-species variation in deposition capabilities and inter-seasonal differences in foliage retention add complexity to estimating deposition of pollutants throughout the year, that is, deposition capacity of deciduous trees is reduced compared to coniferous evergreens in winter months.

2.1.2 *Modelling of vegetation and air pollution*

Using a personal monitoring method of exposure assessment in a small population of pregnant women in Barcelona (n = 54), Dadvand et al. (2012) highlighted the association of higher average greenness (NDVI) in a buffer of 100 m around each maternal residential address and lower exposure to particulate matter (aerodynamic diameter ≤2.5 µm; $PM_{2.5}$) and nitric oxides (NOx). Though personal measurements are possible for a small-scale study, personal air pollution monitoring is not currently feasible for a large number of people. An alternative, model based approach is required to investigate the effects of greenspace on ambient air quality at the residential addresses in large cohorts. Below I outline the key literature on greenspace and air pollution modelling approaches.

2.1.2.1 *Deposition models*

Overall, trees improve air quality in urban environments through the removal of pollutants (Nowak, 2006; Rao et al., 2014). To estimate air quality improvement, the United States Department of Agriculture Forest Service developed a user-friendly toolkit to estimate air pollution ($PM_{2.5}$) deposition attributable to trees – i-tree – https://www.itreetools.org/. Using the Urban Forest Effects (UFORE) model (Nowak et al., 2008), i-tree can estimate, amongst other ecosystem services, the reduction in air pollution attributable to forests via multiple inputs (e.g., number of trees, species composition, tree sizes). By scaling this model to trees and forests across the conterminous United States, Nowak et al. (2014) estimated that 17.4 million tonnes of air pollution was removed in 2010 (range: 9.0–23.2 million tonnes). This pollution removal equated to an average air quality improvement of less than one percent and most of the pollution removal occurred in rural areas. However, significant health impacts from small air quality improvements were estimated in urban areas, including the avoidance of more than 850 incidences of mortality.

The Centre for Ecology and Hydrology (CEH) conducted a UK-wide analysis (Jones et al., 2017b) for the UK Office for National Statistics (ONS) on pollutant deposition (including NO_2 and $PM_{2.5}$) across vegetated land, including woodland, farmland, grassland, moorland and wetlands, at 1 km x 1 km grid resolution. As with the UFORE model, the highest absolute levels of air pollution deposition on vegetation were in rural areas. CEH also estimated savings in healthcare spending attributable to removed pollution. This valuation analysis demonstrated the contribution of pollutant deposition on vegetation in cities, which, although lower in absolute terms compared to rural areas, resulted in major health cost reductions due to the combination of high pollution concentrations and high population density. Built up urban areas were also found to benefit from deposition on vegetation in neighbouring rural areas via a beneficial 'spillover effect'. Deposition models are particularly useful for quantification and valuation of greenspace, though are low-resolution and sub-optimal for epidemiological assessment of individual (address-level) cohort data.

2.1.2.2 *Land-use regression models*

Air pollution varies spatially within cities. Interaction of pollutants with the physical environment (e.g., with buildings and greenspace) can result in spatial variations in pollutant concentrations. Land use regression models use geographical variables (e.g., roads, land cover, topography) as a proxy for source emissions and sinks to estimate spatial variability in concentrations of air pollution. LUR models can be used to estimate air pollution concentrations to which individuals are exposed (e.g., at their residential address) without costly monitoring equipment. LURs are, therefore, a cost effective and robust approach for assessing spatial variation in pollutants in an urban context (Gulliver et al., 2011).

LUR studies typically use long-term pollution concentration averages (e.g., annual), measured at multiple air pollution monitoring sites across the study area, and predict measured concentrations based on traffic and the spatial distribution of different land uses surrounding monitoring sites (Beelen et al., 2013). Potential traffic related variables and land-use predictor variables (e.g. greenspace, ports, water, high- and low-density residential areas) are regressed against monitored concentrations. Variables are retained in regression models based on their coefficients and their contribution to the ability of the model to predict monitored air pollution concentrations. The model is tested for robustness via validation techniques and then applied to locations where measured concentrations are not available – e.g., an individual's place of residence (Gulliver and de Hoogh, 2015). The spatial resolution of predictor variables is a key factor in determining the prediction capabilities of LUR models. In order to develop accurate models of intra-urban levels of pollution, models that include site characteristics (e.g. street configuration, building dimensions) have been suggested (Eeftens, 2013).

The ESCAPE project was set up to address a deficit in investigations (compared to the US) of the long-term health impacts of air pollution in European populations (Beelen et al., 2014, Eeftens et al., 2012). The ESCAPE project used LURs to estimate air pollution exposure. Vegetation cover was not selected as a predictor in the ESCAPE model for London, though vegetation data from land use datasets have been selected as predictor variables in other intra-urban air pollution models of NO_2 (Rao et al., 2014, Tang et al., 2013). Typically, vegetation cover is derived from land use information such as Ordnance Survey MasterMap™ in the UK. Satellite image derived indices of greenness (NDVI) have also been used as a predictor variable, for example, NDVI was selected for a national (South Korea) model for NO_2 (Kim and Song, 2017). To date, high resolution vegetation data differentiated by height has not been used in air pollution prediction models, despite the potential for taller vegetation (e.g., trees) to entrap pollutants and augment concentrations in some built environment scenarios (Abhijith et al., 2017).

2.1.2.3 *Dispersion models*

Another type of model used to predict intra-urban scale air pollution concentrations are dispersion models. Dispersion models estimate emissions from specific sources (e.g., traffic, industry, domestic) and use meteorological data, atmospheric chemistry, and attributes of the built environment to model how pollutants disperse in the environment. Air pollution dispersion models that do not include vegetation data in estimations of pollutant concentrations nonetheless show a pollutant gradient across large areas of greenspace. This gradient across greenspace is not derived from estimations of deposition of pollutants on vegetation. Instead, in these models, it is a ‘by-product’ (non-causal relationship) of the low level of emission within large, green and open spaces, combined with unobstructed airflow and ventilation.

The non-causal association of vegetation cover and air pollution can be highlighted with an example. Researchers at King’s College London (KCL, 2019) estimated annual average NO_2 concentrations from the London Atmospheric

Emissions Inventory (LAEI) 2016 via pollutant dispersion (kernel) estimations from multiple sources. To calculate high resolution air pollution concentration estimates (20 m × 20 m grid surface), model inputs included emissions from road-traffic, industrial, commercial, domestic and miscellaneous sources from multiple datasets (e.g., Environment Agency's Pollution Inventory and Local Authority records), as well as hourly meteorological data and background levels of pollutants. The dispersion from different pollution sources was modelled using Atmospheric Dispersion Modelling System (ADMS) – a comprehensive software for air quality modelling produced by Cambridge Environmental Research Consultants (www.cerc.co.uk).

Figure 4 shows the decline in annual average NO_2 concentrations away from congested roads that border a large, open greenspace in central London, UK – Hyde Park. Distance decay from sources (e.g., diesel vehicle emissions), combined with unobstructed airflow in open greenspace, is responsible for lower pollution concentrations in central areas of the park compared to the road-side fringe.

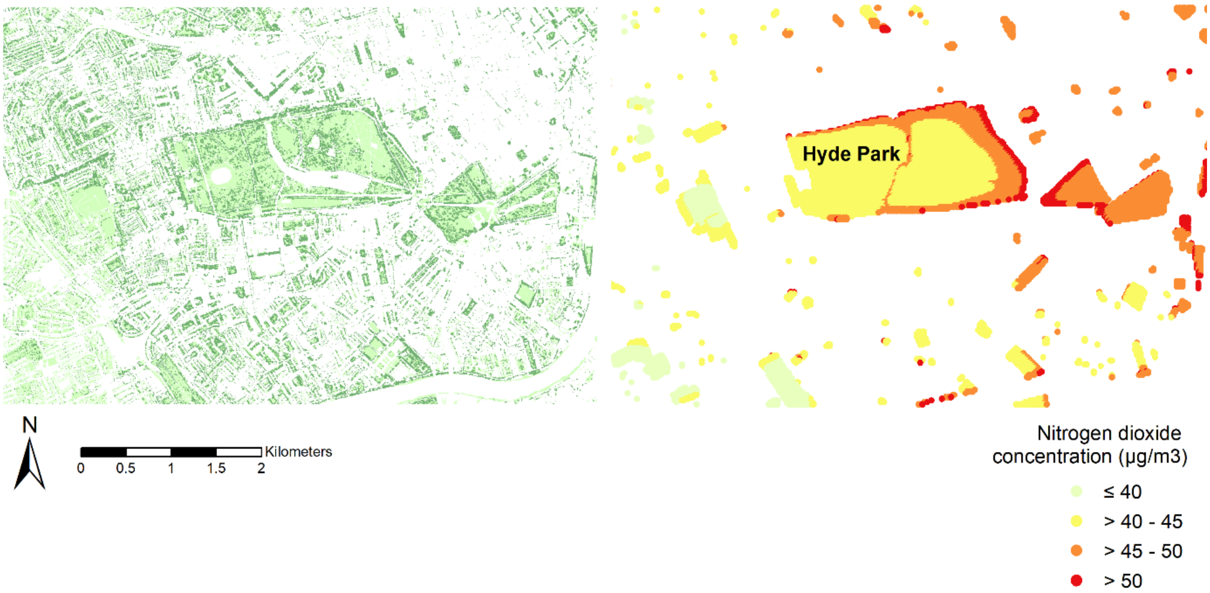


Figure 4. Vegetation cover (GeoInformation Group data) in Hyde Park, London, and the surrounding area (left panel) shown with modelled nitrogen dioxide concentrations (London Atmospheric Emissions Inventory, 2016, 20 m x 20 m grid square centroids) within vegetation cover (e.g., Hyde Park) boundaries (Ordnance Survey Open Greenspace data; right panel).

ADMS does not include vegetation data in calculations, and is modelled as if no trees or groundcover are present; to improve estimates of traffic related air pollution concentrations in street canyons from ADMS-Urban (the version of ADMS used to produce intra-urban estimates of pollutant concentrations), Tang et al. (2013) created supplementary variables representing the near-source built environment. Tang et al. (2013) combined built environment (land use) variables with output from the ADMS dispersion model in a regression model, hereafter this type of model is referred to as a Dispersion-LUR model.

2.1.2.4 *Dispersion-LUR model*

Tang et al. (2013) developed a dispersion-LUR model to better represent pollution trapping (i.e. reduced ventilation) caused by near-source buildings around major roads. In their model, inclusion of greenspace data improved estimation of pollutant concentrations; pollutant trapping from building topography increased concentrations and vegetation cover decreased them.

2.1.3 *Epidemiological findings on greenspace and air pollution*

A study by de Keijzer et al. (2017) concurrently assessed the contribution of air pollution and greenspace to mortality rates in Spain (nationwide, small area study), and showed a decrease in life expectancy of almost one year for a 5 μg/m^3 increase in particulate matter (PM_{10}). Higher level of residential surrounding greenspace was positively associated with life expectancy only in the lower socio-economic status (SES) group, and the authors identified the need to clarify the influence of vegetation levels on air pollution.

Currently, it is not common practice to adjust air pollution and cardiovascular disease risk estimates for confounding by greenspace cover (Crouse et al., 2019). In contrast, adjustment for confounding by air pollution is commonplace in greenspace epidemiological analyses (Crouse et al., 2017, de Keijzer et al., 2017, Villeneuve et al., 2012a). For example, using data on ~1.3 million adults across 30 Canadian cities, Crouse et al. (2017) calculated the HR of cardiovascular mortality in relation to greenness (NDVI 250 m), adjusted for personal and contextual covariates, and adjusted for NO_2, which reduced the HR by approximately 5%. Additionally, in a cohort study relating urban greenness (NDVI 500 m) with mortality in Ontario, Canada, after adjustment for personal and contextual covariates, adjustment for NO_2 slightly attenuated rate ratios (RR) for cardiovascular disease mortality (RR 0.94 (95% CIs = 0.92, 0.96) vs. 0.95 (95% CIs = 0.93, 0.97)) (Villeneuve et al., 2012a).

Some greenspace analyses have assessed mediation of the greenspace-health effect by air pollution. For example, using the Swiss National Cohort, (Vienneau et al., 2017) reported a HR for greenness (NDVI) per interquartile range within 500 m of residential address of 0.95 (95% CIs = 0.94, 0.96) for CVD mortality, and estimated air pollution (PM_{10}) to explain a small proportion of the total effect (3.1% (95% CI = 0.6%, 8.5%)). In a study conducted in the Netherlands on cardiometabolic disease – a cardiovascular disease risk factor – Klompmaker et al. (2019b) found that NO_2 partially mediated associations of greenness 300 m and 1000 m with diabetes; the proportion mediated was 20% (95% CI 8%, 33%) and 34% (95% CIs = 15%, 53%), respectively.

Another study on the Dutch national cohort showed that greenness (NDVI 300 m) associations with cardiometabolic disease, which were adjusted for physical activity to exclude the effect of the *physiological* pathway in risk estimates, were partially explained (mediated) by air pollution (e.g., NO_2 20% mediation, 95% CI = 8%, 33%) (Klompmaker et al., 2019b). The authors suggested that other mechanistic pathways (e.g., stress reduction) accounted for the remaining effect (~80%).

As described above, studies of greenspace and cardiovascular health outcomes have controlled for air pollution, or conducted mediation analysis, though air pollution exposure estimates are often produced from models that do not include greenspace as a predictor variable. Models are inevitably associated with the predictor variables that they contain. Therefore, correlation of modelled air pollution and greenspace will vary depending on the specific predictors in the air pollution model. To give a tangible example, if vehicle flow and population density are predictors in an air pollution model, the association of greenspace with modelled air pollution estimates would be a product of *low* vehicle flow or *lack* of domestic pollution sources in greenspace, and therefore does not necessarily represent a causal link between greenspace and air pollution. This is an important consideration in epidemiological analysis of greenspace and air pollution, particularly for mediation analysis, whereby the association of the focal exposure (e.g., greenspace) and the mediating variable (e.g., air pollution) can impact findings.

2.1.4 *Implication of correlation of air pollution and traffic noise*

Alongside air pollution, traffic-related noise is one of the most important environmental risk factors for cardiovascular disease (Münzel et al., 2014). It is estimated that at least one million disability-adjusted life years are lost every year due to traffic noise in the western part of Europe (World Health Organization, 2011). In London, one of Europe's largest cities, daytime levels of traffic noise are estimated to exceed 55 dB for ~1.6 million people (Gulliver et al., 2015) – a threshold above which the World Health Organisation (WHO) defines as harmful to human health (World Health Organization, 1999). A small-area level analysis in London showed that long-term (annual average) traffic noise levels >60 vs. <55 decibels (dB) were associated with increased risk of all-cause mortality in adults (relative risk (RR) = 1.04; 95% CIs = 1.00, 1.07), and risk of hospital admission for stroke in both adults ≥25 years (RR = 1.05; 95% CIs = 1.02, 1.09) and older adults ≥75 years (RR = 1.09; 95% CIs = 1.04, 1.14) (Halonen et al., 2015).

In a meta-analysis of 14 studies, Babisch (2014) reported an association of road traffic noise and coronary heart disease (RR (pooled estimate) 1.08, 95% CIs = 1.04, 1.13, per 10 dB(A) increase in weighted day-night noise level (L_{den})). A review of traffic-noise and hypertension – a traditional CVD risk factor – corroborated these findings (van Kempen et al., 2002), as did a review of aircraft noise and hypertension (Babisch and Kamp, 2009). Associations of traffic noise and blood biochemical markers (e.g., high-density lipoprotein cholesterol) in a harmonised multi-cohort study were also suggestive of a biologically-plausible link between long-term noise exposure and cardio-metabolic disease risk (Cai et al., 2017).

The mechanisms via which traffic noise is hypothesised to adversely interact with cardiovascular health differs from those of air pollution. Noise is hypothesised to affect health via direct (e.g., sleep disturbance) and indirect (e.g., annoyance) mechanistic pathways (Babisch, 2014). Long term exposure to harmful levels of noise is thought to generate an adaptive physiological response to the repeated release of stress hormones (e.g., cortisol), which might result in adverse changes to blood pressure, glucose and lipid levels, and contribute to cardiovascular disease (Daiber et al., 2019, Münzel et al., 2014, Recio et al., 2016). In contrast, air pollution is hypothesised to affect the circulatory system via particle translocation into cardiovascular tissues, which triggers inflammation, and can result in autonomic nervous system imbalance (Brook et al., 2010a, Franklin et al., 2015). While the size and state of air pollutants determines the mechanistic pathway via which pollutants impacts health, physical properties of sound and an individual's susceptibility to noise annoyance (psychological pathway) determines how noise affects health (Basner et al., 2014, Münzel et al., 2017). Traffic related air pollution (e.g. NO_2 and $PM_{2.5}$) and traffic noise share a common source (i.e. motor vehicles) and might therefore be correlated. Local built environment characteristics can also affect correlation of traffic noise and air pollution (Foraster et al., 2011). Measurement campaigns have highlighted strong-to-moderate correlation of traffic noise and NO_2 from traffic sources, which are impacted by traffic flow and road layout (Davies et al., 2009).

Using modelled estimates, the effect of scale and area characteristics on traffic noise and air pollution correlations have been investigated in London (Fecht et al., 2016). Traffic noise and air pollution were modelled, respectively, via the UK Calculation of Road Traffic Noise (CRTN) method and the KCLurban dispersion model for pollutants, including NO_2, and the total and traffic-only component of $PM_{2.5}$ and PM_{10}. Fecht et al. (2016) emphasised the need to select a relevant geographic unit of analysis in epidemiological investigations and to assess health associations of correlated environmental exposures with caution. Correlation of traffic noise and traffic-related air pollution near roads within London was found to be moderate, indicating that independent effects of air pollution and noise can be reliably determined when input data is adequately detailed.

Epidemiological evidence of independent effects of noise and air pollution has started to emerge (Gan et al., 2012). Some studies showed independent effects (adjusted single exposure models), and/or additive effects of air pollution and noise exposure (multi-exposure models), though typically did not show evidence of interaction (i.e. multiplicative effects) of air pollution and traffic noise on cardiovascular outcomes and risk factors (Floud et al., 2013, Klompmaker et al., 2019b, Tétreault et al., 2013), except in a single study of air pollution, noise and stroke incidence (Sørensen et al., 2014).

Advances in exposure assessment using high-resolution data have allowed for assignment of individual traffic noise and air pollution exposures to large number of addresses. Analyses across regionally diverse areas is strengthening the evidence base on traffic noise and air pollution effects on cardiovascular outcomes, such as incident CVD(Cai et al., 2018). Individual-level exposure assessment for cohorts requires address location accuracy (Brugge et al., 2013), and careful consideration of the scale of inputs in exposure assessment models (Fecht et al., 2016). For example, adjustment of an air pollution-health model for confounding by traffic noise, in which data inputs into the air pollution model are detailed (high resolution) and data inputs into the noise model are course (low resolution), might result in biased effect estimates due to incomplete adjustment. It is critical that environmental exposures, whether used as confounders or otherwise in health models, are matched in terms of detailed inputs to reduce bias.

2.2 Physiological (physical activity) pathway

2.2.1 *Physical activity and cardiovascular health*

The UK Chief Medical Officers' guidelines recommend that older adults participate in a minimum of 150 minutes of moderate intensity physical activity or 75 minutes of vigorous intensity physical activity per week (UK Chief Medical Officers, 2019). Compared to previous guidelines (Davies et al., 2011), the newer (2019) guidelines put a stronger emphasis on regular, light intensity physical activity for older adults, and removed recommendation of a 10 minute minimum session duration that was previously advised (UK Chief Medical Officers, 2019). Recent evidence on the benefits of cumulative, low to moderate intensity physical activity on likelihood of developing cardiovascular disease risk factors (e.g., diabetes) and premature mortality in older adults prompted these changes (Chastin et al., 2019, Füzéki et al., 2017, Hupin et al., 2015, Jefferis et al., 2019, LaMonte et al., 2017). The current guidelines are also based on earlier evidence that showed associations of physical activity with lower likelihood of developing cardiovascular diseases, such as coronary heart disease and stroke (e.g., Haskell et al., 2007).

Physical activity can be achieved in multiple contexts including the commute (active travel), during working hours (occupational), and during leisure time. In a meta-analysis of 21 prospective cohort studies, the RR of overall CVD in women with high levels of physical activity during leisure time was 0.73 (95% CIs = 0.68, 0.78), compared to the reference group with low leisure time physical activity (Li and Siegrist, 2012). A similar effect was observed among men (RR 0.76, 95% CIs = 0.70, 0.82). In the UK, active travel physical activity (walking) compared to private transport use was associated with lower likelihood of cardiovascular risk factors, including hypertension (adjusted OR = 0.83, 95% CIs = 0.71, 0.97) (Laverty et al., 2013).

In UK Biobank, Celis-Morales et al. (2017) found an association of self-reported active commuting and a lower risk of CVD incidence, for cycling commute (HR 0.54, 95% CIs = 0.33, 0.88) and walking commute (HR 0.73, 95% CIs = 0.54, 0.99). For CVD mortality, protective associations were also found for cycling commute (HR 0.48, 95% CIs = 0.25, 0.92), and walking commute (HR 0.64, 95% CIs = 0.45, 0.91). Further, a UK Biobank longitudinal analysis showed an association of body mass index (BMI) – a cardiovascular risk factor – and transition to active commuting from vehicular transport (BMI reduction estimate: -0.30 kg/m^2; 95% CIs = -0.13 kg/m^2, -0.47 kg/m^2) (Flint et al., 2016). This study used repeat assessment data that was available for 20,346 UK Biobank participants (Stockport UK Biobank assessment centre only). Moreover, a graded obesity response was observed in UK Biobank across seven categories of commuting behaviours, ranging from active to inactive travel (Flint and Cummins, 2016), adding further evidence in support of physical activity interventions that promote active behaviours (Sallis, 2016).

2.2.2 *Relationship of greenspace with physical activity*

A review of neighbourhood greenness and health associations reported “fairly strong” evidence of a positive association of greenness and physical activity, and greenness and cardiovascular disease, and suggested that future research identify effect modifiers and mediators of these associations (James et al., 2015). The *physiological* pathway linking greenspace and health, which is explored in this PhD thesis, is based on the assumption that access to greenspace increases physical activity in the local population via the provision of amenable spaces to carry out exercise (for e.g., walking, jogging and cycling), with accompanying health benefits (Nieuwenhuijsen et al., 2017a). Accessibility of greenspace from an individual’s place of residence is considered a key driver to the *physiological* pathway, with better access expected to increase physical activity (Rigolon, 2016). The definition of *accessible* is context-specific, and often subjective, which has resulted in inconsistencies in the literature (Giles-Corti et al., 2019).

In England, using the Generalised Land Use Database (GLUD), a large scale ecological study assessed greenspace exposure at the administrative middle super-output area level (MSOA; average population of ~6000), which was deemed representative of an accessible environment for physical activity, and reported higher odds of achieving the national recommended amount of physical activity (OR 1.27, 95% CIs = 1.13, 1.44) for people living in the greenest quintile (nationally) compared to those living in the least green quintile (Mytton et al., 2012). However, in this example, associations were found for gardening and do-it-yourself, and occupational physical activity, rather than expected greenspace-based activities (e.g. walking, running), suggesting that recreational use of greenspace was not a driver of the association. In another England-based study, the percentage of greenspace at the MSOA level (and within 5 km and 10 km of the MSOA) was not associated with self-reported walking (Lachowycz and Jones, 2014b).

Some studies on the relationship of greenness (NDVI) and physical activity have reported a limited, and sometimes inverse, association when assessing quantity of greenspace surrounding the residential address and physical activities. For example, Maas et al. (2008) found no association of greenness within a 800m circular distance buffer of an individual's place of residence and whether they met Dutch public health recommendations for physical activity, sports and walking for commuting purposes. Once again, gardening was associated with greenness, as was cycling for commuting, although these activities did not explain the greenness-health relationship.

In a study conducted using the Dutch National Health Survey (*n* = 387,195 adults), greenness surrounding the residential address (NDVI 300 m) in the highest quintile compared to the lowest quintile was associated with increased odds for outdoor physical activity (OR 1.14, 95% CIs = 1.10, 1.17) (Klompmaker et al., 2018). However, the importance of scale, particularly in this example, should be highlighted – it has been suggested that the wider neighbourhood greenspace environment (e.g., up to 1600 m network distance) influences physical activity levels (Giles-Corti et al., 2019), though associations in this study were generally stronger for smaller buffers (strongest in 300 m NDVI circular buffer). The authors also noted that associations were not found for greenspace exposure derived from Dutch land use data (TOP10NL), and speculate this is because NDVI takes private gardens into account, whereas TOP10NL does not. The authors did not clarify if private garden cover was potentially indicative of gardening physical activity, or potentially acting as a proxy variable in the model for socioeconomic status.

Other studies on the association of greenness (NDVI) exposure and incidence of health outcomes in the US have used physical activity levels in mediation analysis. For example, James et al. (2016a) found total PA in an all-female cohort to explain the smallest proportion of the association of greenness (NDVI 250 m and 1250 m) and non-accidental mortality (2.1% mediation, 95% CIs = 0.2%, 19.3%; and 1.1% mediation, 95% CIs = 0.1%, 15.8%, respectively), compared to other assessed mediators (i.e. air pollution, social engagement, mental health). In another all-female cohort study, Villeneuve et al. (2018) used the US National Land Cover Database to identify natural environments (e.g., forest, shrubland, herbaceous land covers) and assessed residential surrounding exposure in a 500 m circular distance buffer. Associations with obesity were estimated to be partially mediated by physical activity (32% mediation, 95% CIs not reported).

The above studies on greenspace and physical activity share several limitations that may hinder identification of associations: 1) greenspace exposure was assessed via NDVI or land use data, both of which are unspecific quantifiers of greenspace coverage (e.g., qualitative features related to greenspace function are not represented); and 2) multiple well-evidenced environmental predictors of light to moderate intensity physical activity (e.g., walking) are ignored in geospatial and statistical analysis. More specifically regarding the last point, in studies of greenspace and physical activity, it is commonplace to adjust associations of greenness and physical activity for urbanicity, which might partially account for population density, however several other known drivers of physical activity in the urban context, are often ignored. For example, junction density and destination density are known predictors of walking for transport in cities (Frank et al., 2017, Giles-Corti et al., 2019, James et al., 2017).

Regarding unspecific quantifiers of greenspace (point 1 above), the type and quality of greenspace – e.g. amenities and safety – have been proposed to contribute to its perceived usability for physical activity and health associations (Giles-Corti et al., 2019, Jorgensen et al., 2013). Coombes et al. (2010) used transport network analysis in a UK city (Bristol) to assess access to greenspace categorised by type, and association with physical activity levels. Formal park access (though not other greenspace types) was associated with lower likelihood of obese or overweight (a cardiovascular risk factor), and a higher likelihood of achieving national physical activity recommendations (Coombes et al., 2010). Formal parks are potentially distinct from other greenspace due to the path networks running through them, which can be used for multiple physical activities (e.g. walking, cycling). Investigation of specific greenspace function (e.g., sports grounds versus public parks versus private gardens) with associations of physical activity and health outcomes in a large sample has not been conducted to date, though is critical for informing built environment interventions targeted at improving physical activity. NDVI, and other unspecific quantifiers of greenness, cannot determine *physiological* pathway specific contributions and might hinder research translation, if findings are not triangulated with other exposure assessment methods (Rugel et al., 2017a).

2.2.3 *Accessibility via road/path network analysis*

The definition of accessibility in the Dictionary of Human Geography is, "the ease with which goods and services in one location can be accessed by people living in another location". Importantly, accessibility is a measure of opportunity rather than usage, which is an important consideration for greenspace epidemiological investigation, especially along the physiological pathway (Tamosiunas et al., 2014). While all accessibility measures have an origin and destination(s), they vary based on how destinations are incorporated into the calculation (Kwan and Weber, 2003). In greenspace accessibility literature, accessibility measures can be divided into two main groups: 1) service area measures, and 2) proximity-based measures (Higgs et al., 2012). Service area measures use either simplistic geographic containers, such as administrative boundaries, to measure density of destinations, or can use more sophisticated transport network distance buffers to measure density of destinations within a set distance or time along a transport network (e.g., cumulative opportunities to access greenspace within 10 minutes walking distance of residential address). Proximity based measures use either simplistic measures, such as distance to the nearest destination (e.g., nearest park entry point), or add weights for destination features (e.g., size of park), and balance both weight and distance in calculations.

Transport network distance buffers are a form of service area accessibility measure. They are created by tracing the road (or path) networks a pre-determined distance (e.g., 1000 m) from the residential address, and buffering the line network (e.g., by 50 m). Such network buffers capture a more accurate representation of the area that can be traversed, for example, when walking compared to circular distance buffers. Network buffers are, therefore, thought to better capture the spatial attributes of the neighbourhood which may influence physical activity (Frank et al., 2017, Giles-Corti et al., 2019).

A transport network study in London showed street connectivity and street tree density, surrounding the residential addresses of respondents to the London Travel Demand Survey (*n* = 15,354), were associated with higher odds of walking (Sarkar et al., 2015b). In this case, the connectivity of the walkable street network was potentially driving associations of greenspace (street trees) and PA (authors did not mutually adjust these exposures). Findings might also point to the impact of incidental greenspace exposure, as oppose to deliberate exposure, on physical activity. In the literature, greenspace has been considered as the final destination endpoint, as by Coombes et al. (2010), or as an environmental feature that individuals are exposed to on journeys to other destination endpoints (e.g., shops, transport links), as by (Sarkar et al., 2015b).

For environmental epidemiological analysis, the representation of greenspace as a single destination (e.g., access point) as oppose to land cover across a buffer is potentially reductive. That is, simplistic proximity analysis (nearest destination) is potentially inadequate for representing greenspace exposure. For example, distance to the nearest park entrance was not associated with being overweight or outdoor physical activity in a study conducted in the Netherlands (Klompmaker et al., 2018). In the UK, an analysis that explored proximity of greenspace (with adjustment for greenspace size and quality) found no evidence of an association with leisure time PA (Hillsdon et al., 2006).

Greenspace exposure is potentially accumulated when travelling through-, as well as when travelling to- greenspace, hence some association with surrounding greenness might be expected. Surrounding greenness, particularly surrounding traversable routes, has the potential to impact individual decisions on transport mode (e.g., preference for walking for transport versus driving a motor vehicle) and should be further explored.

2.2.4 *A case for walkability*

Walkability assessment is a form of service area based accessibility measurement, and service areas for walkability assessment can be generated based on distance along road/path network from study participants' residential addresses (transport network analysis). Forsyth (2015) provided a 'minimal definition' of walkability, in which she stated, "traversability and closeness with some basic level of safety—these are the core requirements for walking". This definition broadly summarises epidemiological considerations of walkability. Essentially, in epidemiological research, walkability is an estimate of the number of people and places to visit (e.g., shops, transport stops, amenities), and the road networks that connect them, typically, in a pre-determined vicinity of an individual's address (Adams et al., 2014b). Walkability can then be assessed statistically to derive its impact on a health outcome; for example, cardio-metabolic disease (Braun et al., 2016).

The International Physical Activity and the Environment Network (IPEN) study of adults measured variation in built environment features relevant to walkability using geographic information systems (GIS) across 12 countries, including the UK (Adams et al., 2014b). Walkability components (i.e., residential density, street connectivity, mixture of land uses) were derived around each participant’s residential address using 500 m and 1000 m street network buffers. Higher walkability in the street network buffers generated by the IPEN study showed an association with lower cardio-metabolic risk (Coffee et al., 2013).The study also showed the potential deficiency of using administrative district boundaries in analyses of walkability, as results were not replicable at the administrative boundary level.

Traditionally, land use mixture is used to represent accessibility and heterogeneity of features within walking distance (Adams et al., 2014a, Adams et al., 2014b, Frank et al., 2010). Land use mixture is a common component of walkability indices, though the specific role of vegetation cover surrounding walkable networks has not been assessed in detail. High greenspace cover (in a circular buffer) has been shown to be associated with low walkability (James et al., 2017), though the impact of greenspace cover surrounding the walkable network, taking account of known predictors of walking for transport (i.e. density of population, street intersection density and business density) warrants further investigation.

2.3 Interrelation of greenspace, long term air pollution and walkability

To improve urban environments for health, it is crucial to understand the interrelation of urban environment exposure (e.g., air pollution, walkability, greenspace) and their separate and combined impacts on cardiovascular health. Urban interventions should be optimised for the improvement of health via across multiple exposures. Exposures, therefore, must be analysed concurrently in epidemiological assessment to provide evidence for net-positive interventions. For example, air pollution reduction and walkability can be inversely associated, and should be considered jointly to avoid worsening of one exposure with improvement of the other (Hankey et al., 2012).

Analysis of the interrelation of multiple built environment exposures surrounding residential addresses in the USA found negative associations of greenspace with walkability and air pollution (James et al., 2017). Following convention, walkability exposure in this analysis was constructed by equally weighting spatial predictors to generate a single (z) score, these predictors were: density of population, street intersection density and registered business density. It is unsurprising therefore that high surrounding greenness (NDVI) levels, typically accompanied by low population, few road connections, and low business density, were associated with low walkability.

Klompmaker et al. (2019b) showed the impact of the interrelation of air pollution, road traffic noise and greenness (NDVI) on odds of diabetes – a cardiovascular disease risk factor. The authors ran multiple models with: a) mutual adjustment for confounding exposures in single exposure models; and b) mediation and/or interaction terms in multi-exposure models. They concluded that single exposure studies might overestimate the association of diabetes attributed to exposure. Further, the combined impact of exposure to a combination of surrounding greenness (NDVI) and air pollution might be underestimated by the associations from single-exposure models. Assessment of multiple environmental exposures to evaluate individual and combined associations, along with exposure interaction, is becoming increasingly important as evidence from single exposure studies accumulates.

3 UK Biobank cohort

In this chapter, I describe the UK Biobank cohort study which was used in this study. I focus on the health and lifestyle information used here as well as the process of assigning environmental exposures to UK Biobank participants' residential addresses.

3.1 Background

Established with funding from the Wellcome Trust and Medical Research Council, the UK Biobank is a population-based observational cohort study designed to provide sufficient power to assess a wide range of chronic disease outcomes (Collins, 2012). These outcomes include CVD and CVD risk factors, such as diabetes. Detailed phenotyping of participants at baseline was anticipated to facilitate research into the aetiology and mechanisms driving associations of exposure and health outcomes (Sudlow et al., 2015).

The UK Biobank cohort contains over 500,000 volunteer participants, aged between 40-69 years at recruitment in 2006–2010. During baseline assessments, participants completed lifestyle questionnaires, gave biological samples (i.e. blood, saliva and urine), were assessed physically, and gave consent to health follow-up via medical record linkage (Allen et al., 2014). The automated and ongoing record linkage, which is provided via National Health Service (NHS) and Office for National Statistics (ONS) reports, is used to link participants' UK Biobank records with hospital-based healthcare events, health outcomes and mortality.

Middle- and older-aged British residents were recruited based on NHS primary care registrations. Following enrolment, participants attended an initial baseline assessment visit at one of 22 centres located throughout England, Scotland and Wales, three of which were located in London. Study protocols have been described in detail (Elliott et al., 2008). In brief, UK Biobank delivered questionnaires via touch-screen devices to assess lifestyle factors that included dietary habits, physical activity behaviour via the short-form Recent Physical Activity Questionnaire (Besson et al., 2009), psychological wellbeing, cognitive functioning, social factors, personal and familial self-reported medical histories, and sociodemographic information. Responses were clarified and expanded upon with a trained member of staff (e.g., regarding self-reported illnesses and medication) after the touch-screen session was completed. Research nurses recorded physical and anthropometric measures, including blood pressure, height, and weight, using standardised equipment, and drew blood samples. Participants provided their own urine and saliva samples. Processing of samples followed standardised protocols mandated by UK Biobank to minimise random and systematic error, and produce adequately harmonised data across the 22 assessment centres. UK Biobank catalogued the collected data in an online database – the Data Showcase (ukbiobank.ac.uk/data-showcase) – with detailed protocols of data collection and handling for each variable. The UK Biobank study received ethical approval from the UK National Research Ethics Committee North West.

3.2 Geographic location of study participants

UK Biobank collected information on residential addresses at baseline assessment to maintain contact with participants. In addition to providing a contact database, address information also allows for assessment of associations of health outcomes and neighbourhood environmental factors, which have an inherent spatial dimension. The residential addresses of the full cohort (~500,000 participants) were previously geocoded by the UK Small Area Health and Statistics Unit (SASHU), Imperial College as part of the EU-funded BioSHaRE-EU project (Doiron et al., 2013). Geocoding was performed using Quick Address Software, which resulted in ~97% geocoded addresses. This procedure results in an X,Y-coordinate representing the geometric centroid of the building corresponding to the address, based on Ordnance Survey AddressBase.

3.3 Application process to obtain residential address points

The preliminary application to UK Biobank to get approval for this project was initiated by my former supervisor Dr Susan Hodgson before I started work on my PhD. I drafted the main UK Biobank application, requesting specific variables and justifying our request for residential address geocodes (Appendix). Due to the need to assign new environmental exposures to UK Biobank participants, the application process had to be divided into two stages to ensure anonymity of study participants. In Stage 1, I requested a unique identifier and the geocode of the residential address for each study participant from UK Biobank for exposure assignment. This is the information I used for exposure assessment and which I handed back to UK Biobank. UK Biobank then integrated the exposure data with the UK Biobank data base, stripping off all geographic identifiers. In Stage 2, I requested these exposure data, together with participant health and lifestyle variables.

3.4 Participant health, characteristics and lifestyle variables

3.4.1 *Cardiovascular disease outcomes in UK Biobank*

Though the focus of this PhD project is on exposure assessment (for CVD epidemiology), I requested cardiovascular disease outcome data from UK Biobank with the intention of performing preliminary epidemiological analysis. Participant health, characteristic and lifestyle variables were received in autumn of 2019 (after UK Biobank had integrated the residential exposure assessment variables). My preliminary epidemiological analysis was on non-injury and cardiovascular mortality (Chapter 6). I hereafter describe the mortality data and preparation for survival analysis. I also requested CVD hospital episode data, though, in the preliminary epidemiological analysis shown in this thesis, CVD incidence summary data were used only to exclude prevalent CVD at baseline (see below), not to conduct CVD incidence analysis. This was due to an error on my part in the original data request (hospitalisation plus date were required to be requested as two separate columns from the showcase, I only requested hospitalisation).

UK Biobank receives notice of death of participants via the ONS Death Register. Primary cause of death is classified by a medical doctor according to the International Classification of Diseases, Tenth Revision (ICD-10). At the date of this PhD thesis submission, the UK Biobank mortality censor date for participants residing in England and Wales was 31st January 2018, and in Scotland was 30 November 2016. I used ICD-10 groupings as per the Global Burden of Disease Study (James et al., 2018, WHO, 2018), that is, I grouped non-injury deaths by excluding ICD-10 categories coded, 'V'—'Z', and I grouped circulatory disease deaths as those coded, 'I00'—'I99'.

3.4.2 *Prevalent disease in UK Biobank*

Through routine linkage with NHS Digital Hospital Episode Statistics, UK Biobank participants can be identified who were diagnosed with either myocardial infarction (ST segment elevation or non-ST segment elevation) or stroke (ischaemic stroke, intracerebral haemorrhage or subarachnoid haemorrhage). Information on any other prevalent cardiovascular diseases was also collected by a practice nurse at the assessment centre. Summary information on CVD hospitalisation (data fields 42000 and 42006 in the UK Biobank Data Showcase) was used, prior to conducting survival analysis, to exclude participants who had prevalent CVD before the date of their baseline assessment.

Given that questionnaire items in UK Biobank related to self-reported disability were strong predictors of 5-year mortality (Ganna and Ingelsson, 2015), I used the self-reported item *unable to work due to sickness or disability* to exclude participants from survival analysis. Further, diabetes is a known risk factor for cardiovascular disease (Tsao and Vasan, 2015), I therefore excluded all participants who self-reported diabetes, though I retained those who reported only gestational diabetes

A large proportion of participants self-reported high blood pressure (hypertension) diagnosis. Added to self-reported hypertension status, two blood pressure measurements were taken at baseline, 2 minutes apart (after a seated rest) using a cuff and an Omron HEM-7015IT digital monitor. I used high blood pressure as a confounder in this analysis, as oppose to excluding participants with high blood pressure diagnosis prior to baseline, which would have resulted in many exclusions. To maximize accuracy, I used measured as opposed to self-reported hypertension, and used 140 mmHg systolic over 90 mmHg diastolic blood pressure as the cut off for high blood pressure (Pazoki et al., 2018). I calculated mean systolic (SBP) and diastolic blood pressure (DBP) from 2 automated or 2 manual readings. For individuals with 1 manual and 1 automated blood pressure reading, I used the mean of the 2 values. For individuals with a single BP measurement (1 manual or 1 automated BP reading), I used the single measurement. I replicated this approach from another UK Biobank study (Pazoki et al., 2018).

3.4.3 *Participants' characteristic variables*

Participant age at baseline (in years) is available in UK Biobank (data field 21003). I calculated a more precise baseline age by using month (data field 52) and year of birth (data field 34), combined with median day of month (15th) to estimate age at baseline assessment (+/-16 days maximum difference). I calculated person-days of follow-up from participant baseline assessment date until either death or end of follow-up (censor date: 31st January 2018), whichever came first. Gender of participant (Female, Male) was available. Ethnicity pf participants was predominantly White, with 27,932/~500,000 participants categorising themselves as non-White; I binned responses to create a binary ethnicity variable (non-White, White).

3.4.4 *Participant lifestyle variables*

I requested lifestyle variables for cardiovascular epidemiological analysis based on previous greenspace and circulatory disease mortality literature (Crouse et al., 2017, de Keijzer et al., 2017, James et al., 2016a, Vienneau et al., 2017, Villeneuve et al., 2012a) and known cardiovascular disease risk factors (Tsao and Vasan, 2015).

Variables requested included: 'total (annual) household income level after tax' (categories: 'less than £18,000', '£18,000 to £30,999', '£31,000 to £51,999', '£52,000 to £100,000', 'greater than £100,000'), pack years smoking (continuous variable), alcohol intake (grams/week). Smoking pack years is a unit for measuring the amount a person has smoked over the life course. It is calculated by multiplying the numbers of packs of cigarettes smoked per day by the number of years the person has smoked. For example, 1 pack-year is equal to smoking 20 cigarettes (1 pack) per day for 1 year. This variable was available pre-calculated from UK Biobank, though required cleaning via the variable ‘smoking status’ (categories: ‘current’, ‘previous’, ‘never’), whereby participants with no response (NA) for ‘smoking status’ received NA for ‘pack years smoking’, and those who responded ‘never’ smoker received ‘0’ for pack years.

For alcohol intake at recruitment, I calculated a ‘grams of alcohol per week’ variable based on self-reported total alcohol (beers, wines, spirits) consumed per week. Participants were asked via touchscreen questionnaire about the number of drinks of each alcoholic beverage consumed weekly. For beer consumption, participants were asked, “how many pints of beer or cider would you drink in an average week?”; for both red wine and white wine (including champagne), participants were asked, “how many glasses would you drink in an average week (typically there are six glasses per bottle)?”; for spirits, participants were asked, “how many measures of spirits or liqueurs would you drink in an average week (there are 25 standard measures in a normal sized bottle)?”. In the UK, a unit of alcohol is used as a measure to quantify alcohol consumption and one unit is equal to eight grams of alcohol, as defined by the House of Commons Science and Technology committee in 2012. As with a previous UK Biobank publication (Cai et al., 2018), one pint of beer or cider is equal to 2.5 units of alcohol, one medium-sized glass of wine or champagne is equal to 2.3 units of alcohol and 1 measure of spirits or liqueurs is equal to 1 unit alcohol. As such, actual alcohol content of one serving size for each beverage was calculated by multiplying the units of alcohol by eight, which is 20 grams of alcohol per pint of beer or cider, 18.4 grams of alcohol per one medium sized glass of wine or champagne, and 8 grams of alcohol per one measure of spirits or liqueurs. To derive the ‘grams of alcohol per week’ variable, the number of drinks per week for each beverage was multiplied by the above mentioned grams of alcohol content per drink for each beverage, and these numbers were summed up to give an average consumption of total alcohol per week for each participant. For all non-alcohol drinkers, a value of zero was assigned.

3.5 UK Biobank Greater London subset

The Greater London-based UK Biobank subset used in this PhD project for some exposure assessments (e.g., air pollution and walkability modelling) comprises ~13% of the total cohort. It is defined as all participants with addresses falling within the administrative Greater London Authority boundary. The rationale for focusing on the UK Biobank London subset in exposure assessments was two-part; firstly, high-resolution vegetation data categorised by height (tree versus groundcover) was available only for Greater London (see Chapter 4.1) and, secondly, Greater London was found to be a computationally feasible option for modelling multiple detailed exposures. That is, London has high sample density in a relatively small area (compared to UK-wide analysis), with sufficient exposure contrast (see Chapter 7), and sufficient heterogeneity in demographics. London is unique with respect other UK cities, and findings might not necessarily be transferable to all parts of the UK, though this was deemed an acceptable limitation.

Participants in the UK Biobank London subset were enrolled at one of three UK Biobank assessment centres – Bart's, Croydon or Hounslow. The subset, hereafter referred to as UK Biobank London.

3.6 Awareness of bias in UK Biobank

Due to the untargeted and systematic canvassing of eligible individuals residing within of 25 miles (40 kilometres) of an assessment centre, UK Biobank achieved a relatively low response rate of 5.5% (Fry et al., 2017). In those who enrolled in the cohort, a 'healthy-volunteer' effect – whereby volunteer participants are typically healthier than non-volunteers (non-responders) in the general population, has been reported (Delgado-Rodríguez and Llorca, 2004). In the context of this PhD study, this form of non-response bias does not preclude valid assessment of epidemiological associations if there is sufficient variability of environmental exposures within the cohort (Rothman et al., 2013). Though, small effect sizes should be interpreted with caution due to the potential for unmeasured confounding (associated with enrolment) in self-selected cohorts (Ebrahim and Davey Smith, 2013, Swanson, 2012).

A limitation of the UK Biobank study design, in the context of environmental epidemiological investigations, is the weak representation of some groups (e.g., non-White individuals) in the cohort, which has implications for effect modification analyses. This limitation is not due to representativeness; a population representative cohort would not balance potential effect modifiers in the sample. Instead, it is due to weak representation of some subgroups in the sample (e.g., non-White, low SES), which stems from UK Biobank's objective to predominantly explore genetic, as oppose to social, determinants of health (Fry et al., 2017). Nonetheless, UK Biobank collected extensive phenotypic information (including many known social determinants of health) for all participants, which should provide adequate information for confounder adjustment in epidemiological analyses. The recruitment of older age groups in UK Biobank should not lead to bias in this PhD project as the focal outcome is cardiovascular mortality.

4 Greenspace and deprivation exposure assessment

In this chapter, I introduce exposure assessment. I provide an overview of the data sets that I used for exposure assessment in UK Biobank, which are described alongside the corresponding exposure assessment methods and findings. I introduce the two greenspace datasets used in exposure assessment, which were: 1) Ordnance Survey MasterMap™ Greenspace (categorised by function) greenspace data; and 2) The GeoInformation Group – Greater London vegetation data. I also give an overview of exposure assessments that were conducted in preparation for confounder adjustment in epidemiological statistical analysis, namely, assessment of the Index of Multiple Deprivation (neighbourhood deprivation).

4.1 Greenspace exposure assessment

4.1.1 *Ordnance Survey MasterMap™ Greenspace*

Greenspace is land predominantly covered with vegetation and, in the urban context, might comprise of public parks, private gardens, playing fields, allotments, etc. In light of the mechanistic pathways that I outlined in Chapter 1 and 2, some greenspace types, or combinations of types, might be more protective for health than others. Crude measures of greenness (e.g., NDVI) have advantages, such as relatively complete data coverage across large geographical areas, though do not permit investigation of the functional role or qualities of greenspace that are most important for health.

To my knowledge, no study has assigned categorised greenspace data that allows for investigation of the effect of specific greenspace types (e.g., public parks versus private gardens) on health in a large cohort that has data available on important individual and area-level confounders. Such qualitative greenspace information (function) might be particularly important in assessing amenable greenspace for physical activity (i.e. the physiological pathway). Therefore, in this exposure assessment, I assessed categorised greenspace land cover (Ordnance Survey MasterMap™ Greenspace) in multiple circular buffers at addresses in UK Biobank. This is the only UK Biobank-wide exposure assessment in this PhD project. That is, addresses of some participants from all UK Biobank assessment centres were covered by the OS MasterMap™ Greenspace data, including in Scotland and Wales. This assessment was conducted in preparation of assessing specific categories of greenspace (e.g., public parks and sports grounds) and cardiovascular outcomes, and testing for mediation of associations by physical activity (see *future research*, Chapter 9).

4.1.1.1 Data

Ordnance Survey (OS) produced a greenspace layer that gives a comprehensive overview of greenspace in urban areas of the UK (**Figure 5**). The dataset comprises of topographical areas, which give the boundaries for greenspace as released in OS MasterMap™ Topography Layer (1:1250 scale), with additional greenspace attributes to describe their function. It includes both publicly accessible and private greenspace, sports facilities and natural environment features.

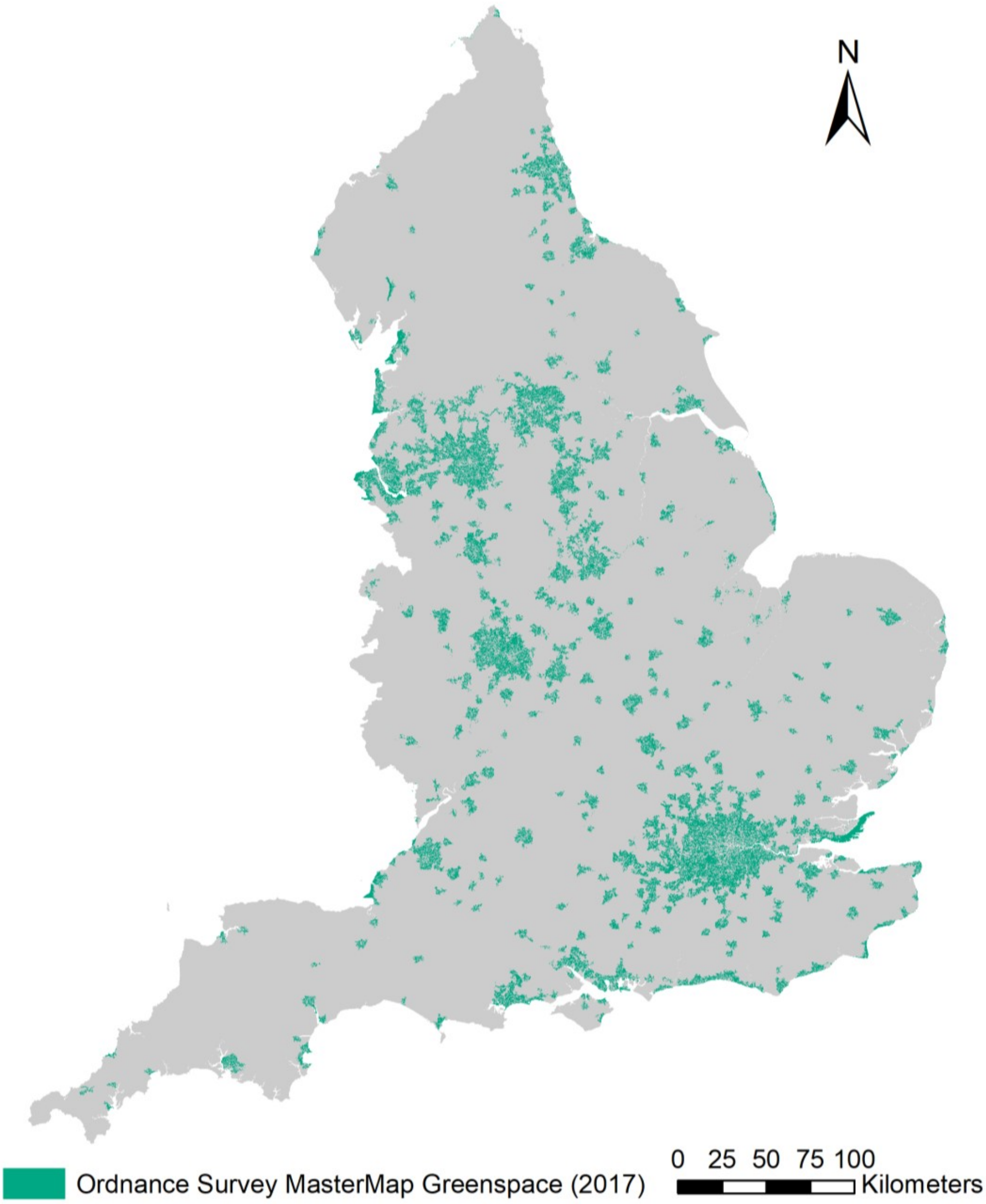


Figure 5. Distribution of urban areas in England with Ordnance Survey MasterMap™ Greenspace data used in this study.

OS described data coverage as the following in the OS MasterMap™ Greenspace product guide: “For England and Wales, urban areas are included where they are greater than 6km². For Scotland, urban areas are defined as those with a population in excess of 500 people. This is based on data provided by the National

Records of Scotland. In Scotland a buffer of 500 m has been added by OS to the urban extents (with over 500 people) to define the OS MasterMap Greenspace product coverage. Where a greenspace site crosses the boundary of a buffered urban area, all features within the site are included in OS MasterMap™ Greenspace, even where these are outside the buffered urban area. This applies up to a limit of 1,500 m from the urban boundary." More information on the data can be found online: https://www.ordnancesurvey.co.uk/docs/product-guides/osmm-greenspace-product-guide.pdf.

Only UK Biobank addresses within the data extent were used in exposure assessment. To select these addresses, I used Office for National Statistics (ONS) Built-up areas (v2) shapefile as a data extent boundary for England and Wales (https://data.gov.uk/dataset/15e3be7f-66ed-416c-b0f2-241e87668642/built-up-areas-december-2011-boundaries-v2), and the National Records of Scotland Settlements (reflective of mid-2016 populations) in Scotland, plus a 500 m buffer (as specified in the technical document linked above).

Table 1. Ordnance Survey (OS) MasterMap™ Greenspace primary function classifications (18 categories) with descriptions as per OS technical specification document.

Function	Description
Private Garden	Areas of land normally enclosed and associated with private residences and reserved for private use.
Golf Course	A large area of land that is specially prepared for playing golf.
Tennis Court	A specially prepared area intended for playing tennis.
Amenity - Transport	Landscaped areas providing visual amenity or separating different buildings or land uses for environmental, visual or safety reasons when related to a transport function, such as a road, or within a transport hub.
Cemetery	Areas of land associated with burial areas or crematoriums.
Natural	Land use areas with no other function but with Form attribute of woodland, open semi-natural, open water, beach or foreshore.
Land Use Changing	Areas of land that are currently under development or awaiting redevelopment.

Play Space	Areas providing safe and accessible opportunities for children's play, usually linked to housing areas or parks and containing purpose-built equipment. Not captured if within schools or paid-for tourist attractions.
Playing Field	Large, flat areas of grass or specially designed surfaces, generally with marked pitches, used primarily for outdoor sports, i.e. football, rugby, cricket.
Bowling Green	A specially prepared area intended for playing bowls.
Camping or Caravan Park	An organised area of ground designated for tents or caravans, intended for temporary occupation by holidaymakers.
Allotments or Community Growing Spaces	Areas of land for growing fruit, vegetables, and other plants, either in individual allotments or as a community activity. Produce is for the grower's own consumption and not primarily for commercial activity.
Amenity - Residential or Business	Landscaped areas providing visual amenity or separating different buildings or land uses for environmental, visual or safety reasons. Where the area is better described by another category this will be used in preference (e.g. playing field, public park, play space).
Institutional Grounds	Areas of land normally enclosed and associated with institutions. Grounds may be reserved for private use or have restricted access. Includes Universities, Hospitals, Nursing homes, Emergency Services, Prisons, Military Sites, Government and Community Buildings providing public services, Libraries, Museums, Zoos and Theatres.
Public Park or Garden	Areas of land normally enclosed, designed, constructed, managed and maintained as a public park or garden. These normally have a defined perimeter and free public access, and generally sit within or adjacent to urban areas. Access is granted for a wide range of uses and not usually restricted to paths or tracks within the area. May include areas with managed facilities such as benches and flowerbeds, and more natural areas.
School Grounds	Areas of land normally enclosed and associated with a school and primarily reserved for their use.
Other Sports Facility	Land used for other sports not specifically described by other categories. Includes facilities for sports spectating (e.g. stadiums) as well as participation.
Religious Grounds	Areas of land associated with churches and other places of worship.

4.1.1.2 Geospatial analysis of large datasets using PostGIS

I used PostGIS (v. 2.3.3) – a spatial extender for the database software, Postgres (v. 9.6) – to implement geospatial queries in this project (unless otherwise stated). PostGIS functions are coded using structured query language (SQL), making geospatial analyses reproducible. Further, when a geographical file (e.g., shapefile) is added to a Postgres database, it is stored as a table with a single column containing geometry data, which can be indexed. Indexing of the spatial component of the data vastly increases processing efficiency compared to alternative geospatial software (e.g., ArcGIS). My motivation for using PostGIS was the efficiency offered when working with large geographical datasets.

4.1.1.3 Data access

Edina – a data and digital services centre at the University of Edinburgh – created a bespoke data extract of this data for this project as a bulk download with permission granted from OS (free of charge for academic research), via the Digimap platform – https://digimap.edina.ac.uk/. The data release used in this project was October 2017.

For this exposure assessment, Edina provided data as multiple File Geodatabase. To prepare for exposure analysis, I used the translator library GDAL (https://gdal.org/index.html), OSGeo4W function *ogr2ogr*, to bulk upload the File Geodatabases via Command Prompt. Once uploaded to a Postgres database, I wrote a function to loop over all geographic tables (OS tiled, slightly overlapping chunks of geographic data), and insert them into a single table based on a tablename wildcard (e.g., %_greenspace%). Once all tables were inserted into a master table, I dissolved spatially overlapping areas (at the edges of OS tiled chunks) based on the TOID attribute – a unique identifier harmonised across all OS MasterMap™ products.

4.1.1.4 Exposure assessment

In PostGIS, I assigned OS MasterMap™ Greenspace cover in multiple circular distance buffers (100 m, 300 m, 500 m, 1000 m and 1500 m) surrounding each participant’s residential address geocode (i.e. X, Y-coordinate). The 100 m and 300 m buffer were selected to represent the near-home environment, with the 100 m buffer expected to capture private gardens surrounding the residence, and also corresponding to the buffer size selected in the air pollution model, and the 300m distance corresponding to a threshold distance proposed as a European standard for public greenspace access (Annerstedt van den Bosch et al., 2016), the larger (500 m, 1000 m and 1500 m) 'neighbourhood' buffer sizes were selected for comparability with greenness data (NDVI) that has previously been integrated into UK Biobank (Sarkar et al., 2015a). I provided greenspace exposure assessment to UK Biobank as percentage cover of the total buffer area.

I calculated greenspace cover categorised by greenspace function (**Table 1**) for all eligible addresses (England, Scotland and Wales). The eligibility of an address for exposure assessment was dependent on the OS MasterMap™ Greenspace Layer extent. As mentioned before, the OS mm greenspace data "covers all major urban areas in Great Britain". All addresses outside of the OS MasterMap™ Greenspace Layer extent were incrementally excluded from assessment. For instance, to ensure that the greenspace data fully covered the circular distance buffer around each participant's address, only addresses located inwards of the built-up area boundary at the specified buffer distance for each analysis were included in each assessment. For example, geocoded addresses located at least -100m from the built-up area boundary were used for the 100m analysis, geocoded address located at least -300m from the built-up area boundary were used for the 300m analysis, etc. This approach maximised the number of geocoded addresses eligible for analysis for each assessment. The total number of addresses excluded due to lack of coverage is therefore greater at larger buffer sizes.

Due to differences in the OS MasterMap™ Greenspace coverage definition of urban areas in England/Wales versus Scotland, decreases in address data coverage started after the 500 m circular distance marker in Scotland, whereas they were incremental in England and Wales. **Table 2** gives a summary of the number of addresses that fell within the OS MasterMap™ data extent per buffer size, and were therefore eligible for exposure assessment.

Table 2. UK Biobank participant addresses within the Ordnance Survey MasterMap™ Greenspace data extent for circular distance buffer sizes which have been assigned greenspace function.

Distance buffer size (m)	**Geocoded UK Biobank participant addresses (n)**		
	England and Wales	**Scotland**	**Total**
100	311,295	33351	344646
300	221,562	33351	254913
500	174,322	33351	207673
1000	116,508	17822	134330
1500	91,204	11130	102334

4.1.2 *The GeoInformation Group – Greater London Vegetation*

4.1.2.1 Data

On behalf of the Greater London Authority Urban Greening Unit, the GeoInformation Group (https://www.geomni.co.uk/) produced the Greater London Vegetation Layer used in this project. The GeoInformation Group released the layer in 2015. Data were derived from a combination of aerial imagery, Light Detection and Ranging (LiDAR) data and manually digitised tree data. LiDAR (http://www.lidar-uk.com) is an aerial mapping system, which uses lasers to establish the distance between an aeroplane and land. Remote sensors on aircraft detect reflected light to create accurate three-dimensional images of the Earth’s surface. In the data, vegetation cover greater or equal to 2.5m in height is hereafter referred to as tree cover, and vegetation below 2.5m in height is referred to as ground cover. The vegetation data is highly detailed, capturing raster tree canopy and ground cover data at a resolution of 2.5m x 2.5m (see **Figure 6**), however, as with NDVI, this data does not discriminate between public and private land, or greenspace function.

The Greater London Vegetation Layer covers the Greater London Authority – Greater London boundary, as shown in **Figure 7**. The high resolution data, and differentiation of tree and ground cover, offered the potential for novel insight into the relationship of greenspace exposure and health outcomes, such as if trees, or groundcover, or total vegetation is a better predictor of air pollution in an air pollution model. This is one of the justifications for focusing on the Greater London UK Biobank subset for the majority of exposure assessments conducted in this PhD project.

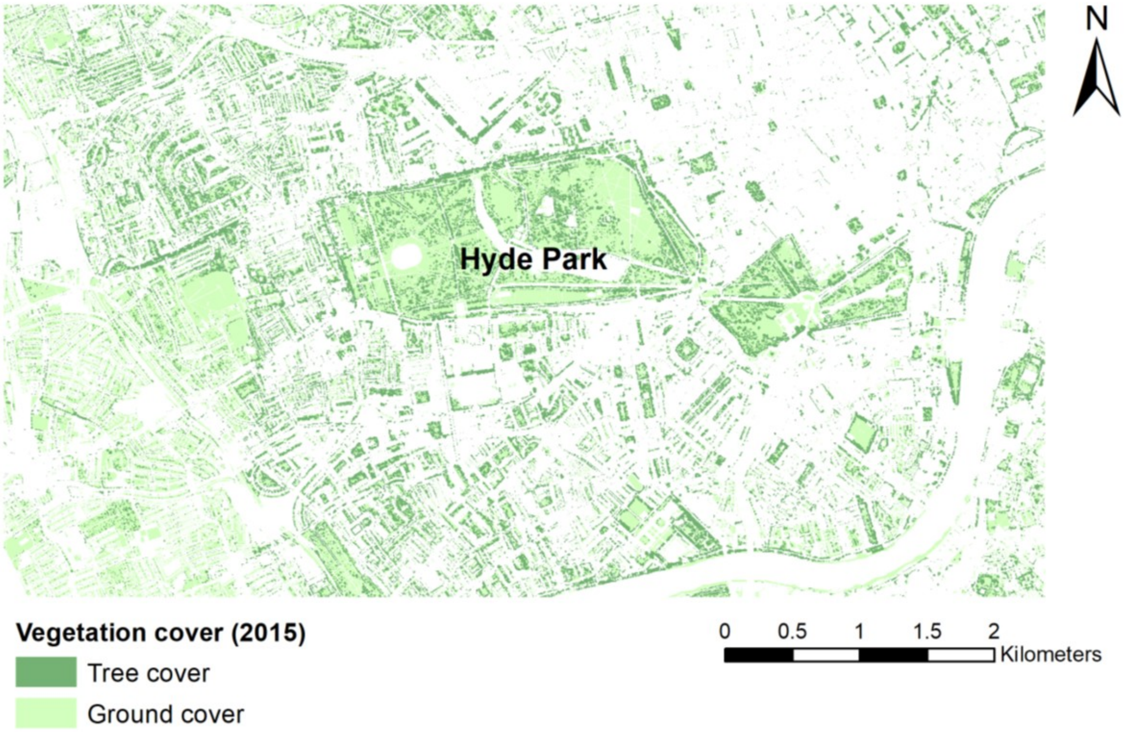


Figure 6. An example of the GeoInformation Group 2015 vegetation cover data (2.5 m x 2.5 m resolution) in and around Hyde Park, London. Tree cover (dark green) and ground cover (light green) are shown. White areas represent built-up zones with no vegetation cover, as well as water cover – the River Thames is visible on the bottom right.

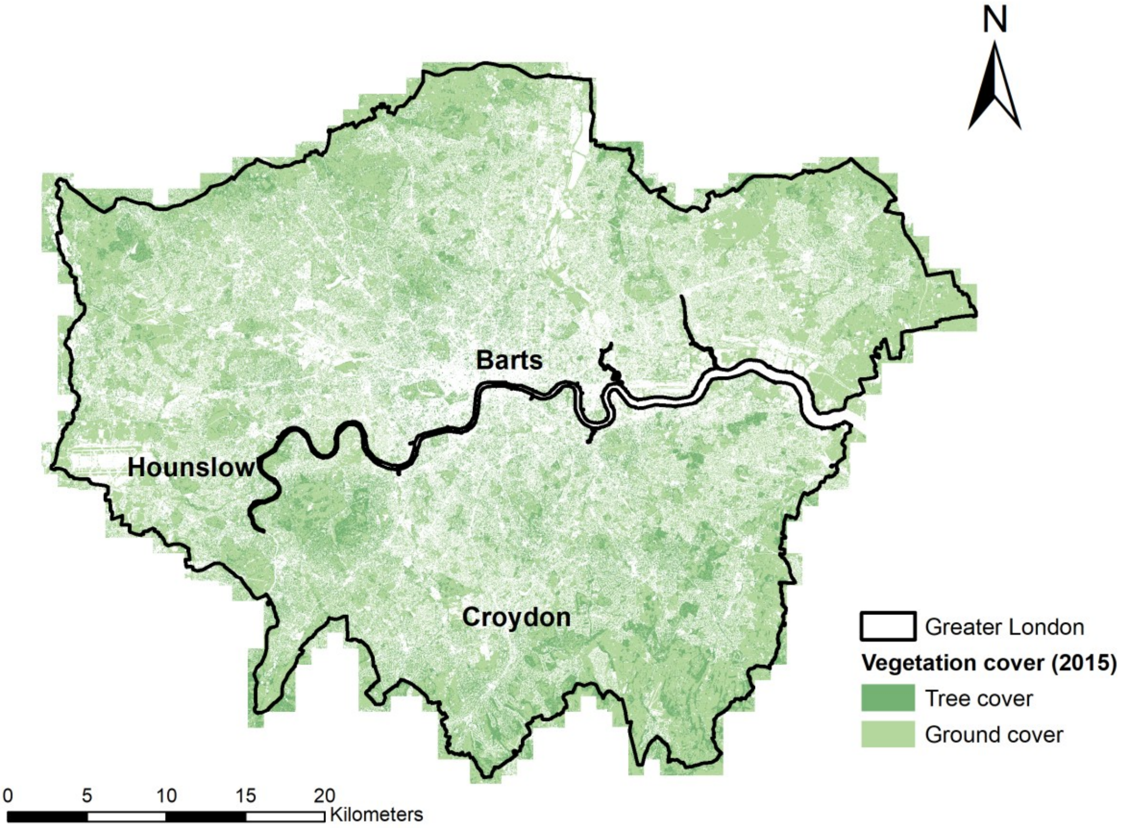


Figure 7. Greater London study area boundary with Thames River (black line) overlaid on The GeoInformation Group vegetation cover layer (2.5 m x 2.5 m resolution). White areas represent built-up zones with no vegetation cover. The locations of the three London-based UK Biobank assessment centres (Barts, Croydon and Hounslow) are shown.

4.1.2.2 Exposure assessment

I assessed vegetation cover data, categorised as tree canopy or ground cover, for each participant's geocoded residential address in multiple circular distance buffers (100 m, 300 m, 500 m, 1000 m and 1500 m). I assigned percentage cover of vegetation categorised as tree canopy and ground cover to all eligible addresses.

To do this, vegetation data, stored in a shapefile format, were uploaded to PostGIS via a graphical user interface uploader, *PostGIS 2.0 DBF and Shapefile Loader Exporter* (EPSG for British National Grid: 27700), with the UK Biobank address geocodes (points) shapefile. I created a table of circular distance buffers around address geocodes in PostGIS, and intersected distance buffers with the vegetation data, saving the area of the intersected vegetation to a new table. This was scripted in a single function that passed output from the previous step to the next step (e.g., stored as a temporary table of intersected vegetation data used to calculate percentage area cover).

The eligibility of an address for exposure assessment was dependent on the data extent, which covers the Greater London Authority (**Figure 7**). To ensure that the vegetation data fully covered the circular distance buffers around each participant's address, only addresses located inwards by the size of the buffer (e.g., 1000m inwards of the data extent for 1000m analysis) of the Greater London Vegetation Layer boundary were included in this exposure assessment. This was achieved by clipping the data extent as shown in **Figure 8**.

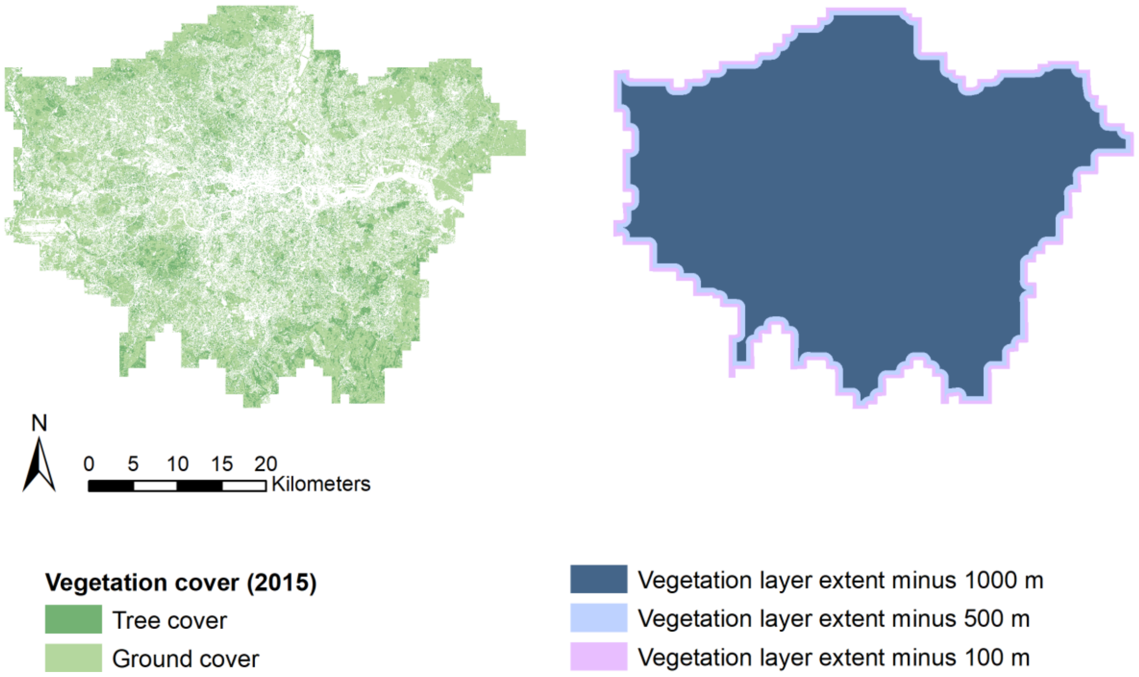

N
0 5 10 15 20
Kilometers
Vegetation cover (2015)
Tree cover
Ground cover
Vegetation layer extent minus 1000 m
Vegetation layer extent minus 500 m
Vegetation layer extent minus 100 m

Figure 8. The GeoInformation Group vegetation cover layer (left) was clipped inwards of the layer extent at 100 m, 500 m, 1000 m (right) to maximise the number of addresses (and corresponding distance buffers) completely covered by the vegetation cover data for each analysis.

4.2 Neighbourhood deprivation level exposure assessment

The Index of Multiple Deprivation (IMD) is a composite, area-level measure of deprivation released by the government at Lower layer Super Output Area (LSOA) level. The IMD comprises six domains representing different dimensions of deprivation: income, employment, education, health, crime, housing, and living environment. Each domain consists of indicators that vary year-to-year (Department for Communities and Local Government, 2016). IMD data are produced by the Ministry of Housing, Community and Local Government, and are used by local and national government to rank areas by deprivation. IMD data are freely-available online (UK GOV, 2019)..

LSOAs are geographic areas designed for the dissemination of census data (e.g., 2011 census), and are specifically designed to improve the reporting of small area statistics in England and Wales. They contain on average 1500 people (minimum 1000) and represent neighbourhoods in the UK. LSOA geographical boundaries, represented as polygons in a shapefile, can be downloaded online from the UK Office of National Statistics portal (ONS).

Studies conducted in the UK have found neighbourhood deprivation level to modify relationships of greenspace and health (Mitchell and Popham, 2008b). I therefore used the 2015 IMD ranks to assign socioeconomic deprivation deciles to corresponding 2011 LSOAs in ArcGIS. I linked them using the unique LSOA identifier code shared by the IMD and LSOA administrative boundary datasets. I collapsed the deciles to quintiles for analysis with greenspace and other environmental variables. To visualise this data, I mapped IMD at LSOA level for Greater London as shown in **Figure 9**.

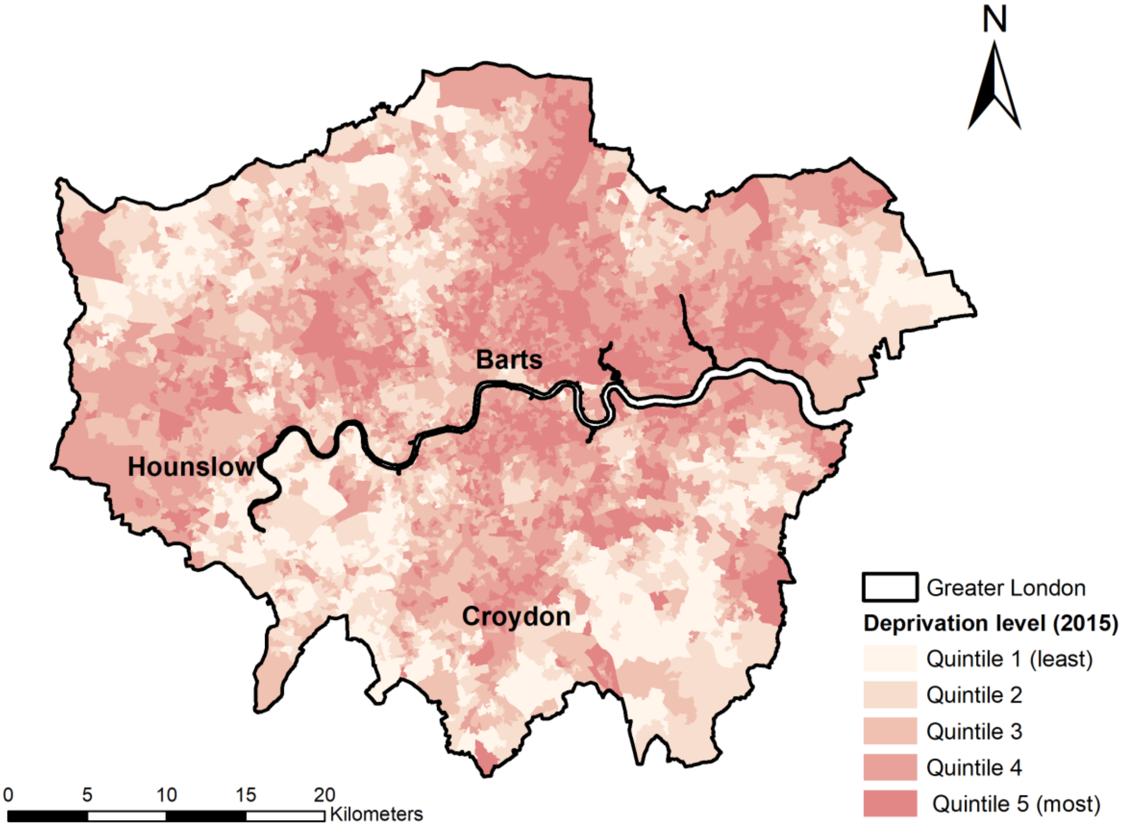


Figure 9. Greater London study area boundary with Thames River (black line) overlaid on lower super output area (LSOA) geometries (2011). LSOA level deprivation (2015) quintiles are mapped to LSOA geometries. The more deprived an LSOA, the darker the shade of red fill in the LSOA geometry. Quintiles are based on national deprivation levels as opposed to an intra-London comparison. The locations of the three London-based UK Biobank assessment centres (Barts, Croydon and Hounslow) are shown.

4.3 Chapter summary

The results of the greenspace and deprivation exposure assessments are in the form of address-level exposures. Data was available both across all UK-Biobank assessment centres (OS MasterMap Greenspace and IMD) and London-wide (The GeoInformation Group – Greater London vegetation layer), and exposures are currently integrated into UK Biobank for all approved researchers to use (see UK Biobank Data Showcase metadata in Appendix). I made greenspace exposure data available at multiple buffer sizes (100 m, 300 m, 500 m, 1000 m, 1500 m) to allow for scale to be adjusted to fit pathway-specific greenspace research questions.

5 Environmental pathway

In order to assess the contribution of the *environmental* pathway to associations of greenspace and cardiovascular health, I focused in this project on air pollution exposure assessment. I estimated outdoor air pollution concentrations through modelling of a commonly assessed air pollutant in the UK – nitrogen dioxide (NO_2). The aim of this approach was to assign modelled air pollution exposure estimates that are partially predicted by vegetation cover to the residential addresses of UK Biobank participants residing within Greater London, where high-resolution vegetation data are available. I used a commercial air pollution dispersion modelling software that is commonly used in research, consultancy and by local authorities – ADMS-Urban – combined with a wide range of potential spatial predictors that have been applied in other air pollution modelling studies, with a primary focus on estimating the modelled reduction in nitrogen dioxide attributable to near-residence vegetation at UK Biobank addresses across Greater London, using attribution methods developed by (Eeftens et al., 2013). My approach in this section was informed by findings from previous NO_2 modelling efforts in London, particularly by an Imperial College London PhD project that included dispersion model output, building topography and greenspace cover in a dispersion-LUR model to estimate annual average NO_2 concentration in street canyons in London (Tang et al., 2013).

Using high resolution vegetation data differentiated by height (e.g., tree versus groundcover), and other high resolution built-environment variables, I built upon the findings of Tang et al. (2013) and developed a high resolution dispersion-LUR model that can be used to estimate annual average concentrations of NO_2 for address locations.

I produced traffic noise estimates based on high resolution data inputs to adequately exclude the confounding effect of traffic noise in epidemiological analysis, which I deemed necessary for estimation of the association of greenspace (high resolution data) and/or air pollution (high resolution data model) with cardiovascular health outcomes. At the end of this chapter, I provide correlations of outputs from low resolution (ESCAPE and CNOSSOS low resolution) and high resolution models (models that I developed and adapted in this PhD project) of air pollution and noise assessed at UK Biobank London, along with my thoughts on how spatial resolution mismatch in modelled exposure data might bias effect estimates in health studies.

5.1 Development of a dispersion-LUR model

I chose to focus efforts on modelling NO_2 to demonstrate this approach because: a) it has been associated with adverse cardiovascular health effects in previous studies (e.g. Atkinson et al., 2013, Brook et al., 2010b); b) it is more spatially variable at the intra-urban scale and therefore is more effectively predicted using spatial variables than particulate matter (i.e. models perform better than those for particulate matter); and c) NO_2 is one of the most widely monitored pollutants in London. As an example of this last point, in the year 2010 there were 128 monitoring stations for NO_2 compared to 34 sites for PM2.5 in the London Air Quality Network (https://www.londonair.org.uk/london/reports/2010%20LAQN%20Summary%20Report.pdf). To match UK Biobank baseline (2006-2010) and for comparability with NO_2 estimates from ESCAPE LUR models (Beelen et al., 2013, Eeftens et al., 2012), which were assigned to all UK Biobank residential addresses in 2010, the year 2010 was designated as the study period for exposure assessment.

5.1.1 *Monitoring network*

The number of measurement sites has previously been demonstrated to impact the accuracy of air pollution exposure estimates, with a large number of sites (>80 if feasible) recommended in a study of exposure assessment bias (Basagaña et al., 2012). The London Air Quality Network (LAQN), which is managed and operated by the Environmental Research Group at King's College London, provides measurements of monitored concentrations, which are used for both air quality management and epidemiological investigation. Monitored concentrations are publicly available online (London Air, 2019). In the year of exposure assessment for this project (2010), there were 128 monitoring stations measuring NO_2 concentrations across Greater London and the surrounding area. However, to avoid producing models based on unrepresentative pollutant concentration averages, it was necessary to assess the completeness of data collection at monitoring sites over the study period. The inclusion criteria for monitoring sites in this project was the following:

- Available hourly measurements for a minimum of 50% of hours in 2010; and
- Available measurements for a minimum of 50% of days in each month of the study period (12 months).

The inclusion criteria resulted in the retention of 56 monitoring stations. This is a lower number of monitors than that recommended by Basagaña et al., (2012), though I considered the impact of unrepresentative pollutant averages (due to missing data) to be more detrimental to model accuracy than a suboptimal number of monitors. Information on retained monitoring sites included monitor name and site code, British National Grid X, Y-coordinates, annual average of hourly monitored NO_2 concentration and site type (roadside, kerbside, urban background, suburban, industrial; see **Table 3**). Monitoring sites were plotted as a point layer in ArcGIS.

Table 3. Classification of air pollution monitoring sites (LAQN, 2016).

Site type	**Description**
Kerbside	Sites with sample inlets within 1m of the kerb of a busy road
Roadside	Sites with sample inlets between 1m and 5m of the kerbside
Urban Background	Urban locations away from major sources and broadly representative of town/city-wide background concentrations (e.g. urban residential areas)
Suburban	Sites of residential areas on the outskirts of a town or city
Industrial	Sites where industrial emissions make a significant contribution to the level of pollution

Industrial sites are typically unrepresentative for modelling residential air pollution and were removed from the dataset (n = 2). Kerbside sites were also removed from the dataset due to their proximity to heavily traffic (within 1 m), which was deemed unrepresentative of residential address locations (n = 4). All roadside, urban background and suburban sites that met inclusion criteria were retained in the dataset to be used for modelling annual average NO_2 concentrations (n = 50).

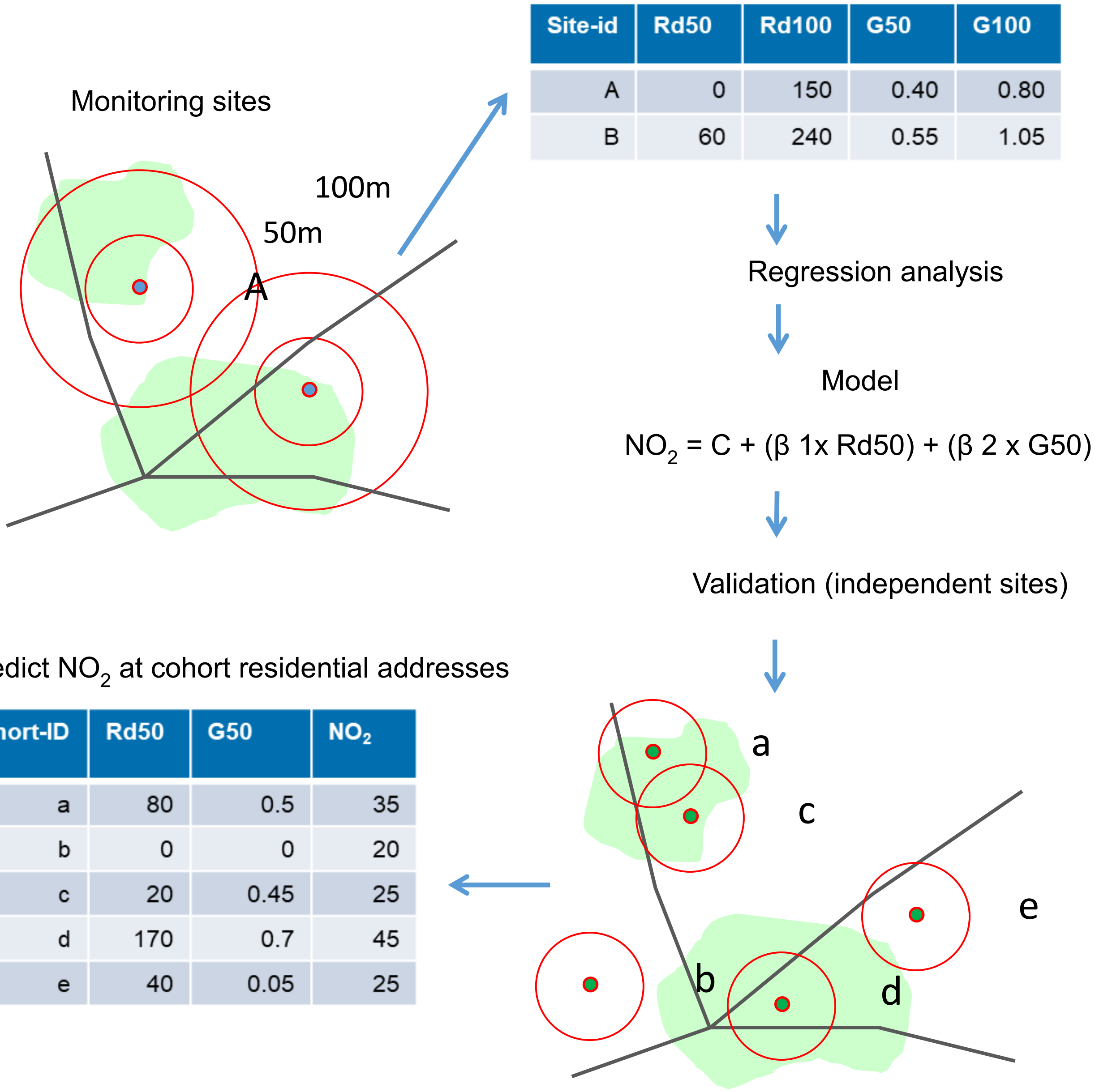
5.1.2 *GIS predictor variable generation*

For the dispersion model output, I modelled NO_2 concentrations related to road traffic (on major roads) using ADMS-Urban (variable name: ADMS-Roads), a proprietary dispersion model produced by Cambridge Environmental Research Consultants (CERC; www.cerc.co.uk). ADMS-Urban is an advanced air pollution dispersion model that relies on emission rates that factor speed of vehicles into calculations. It also includes the effects of meteorology to represent upwind/downwind differences in concentrations, and uses a Gaussian distribution horizontally and a non-Gaussian distribution vertically, offering many advantages over more simplistic traffic related LUR variables. Model inputs for ADMS-Urban were hourly meteorology (wind speed/direction, cloud cover, temperature), emission rates (g/km/s) calculated using EMIT emissions inventory (CERC) from traffic flows/speeds for each road link attributed to road geography (London Atmospheric Emissions Inventory (LAEI) - freely available). Project supervisor, Professor John Gulliver, provided the emission rates data as 51 tiled areas coving Greater London, which had been tiled in preparation for use with ADMS-Urban in another project.

For all other predictor variable generation, I used ArcGIS v10.3, a proprietary GIS by Esri, to map spatial datasets and obtain LUR predictor variables. I generated predictor variables for each monitoring site and linked to annual average concentrations of NO_2. To generate predictor variables, I plotted concentric circles of multiple distances around each monitoring site (radii in meters: 25, 50, 100, 200, 300, 400, 500, 1000, 2000 and 5000), and used the intersect function to clip all available vector/vectorised environmental datasets (described below) to produce predictor variables that characterised the monitoring site at different scales. Buffer sizes were similar to those used in ESCAPE (Beelen et al., 2014, Eeftens et al., 2012).

Figure 10 adapted from (Gulliver and de Hoogh, 2015), shows the basic principles of using geographic data for LUR. Geographical data (e.g., vegetation cover) was extracted at different scales using concentric circular distance buffers (red) at monitoring stations (A, B), and regressed on monitored nitrogen dioxide concentration levels measured at these monitoring stations. Environmental variables that best explain the variance of measured NO_2 concentrations across monitoring sites were retained in the model, and the model was validated (e.g., via cross validation methods).

The beta coefficients of the geographical variables retained in the validated LUR model are used to estimate the concentration of nitrogen dioxide levels at residential address (a, b, c, d, e) where no monitored NO_2 data is available.



Monitoring sites

Site-id	Rd50	Rd100	G50	G100
A	0	150	0.40	0.80
B	60	240	0.55	1.05

Regression analysis

Model

$$NO_2 = C + (\beta\ 1x\ Rd50) + (\beta\ 2\ x\ G50)$$

Validation (independent sites)

Predict NO_2 at cohort residential addresses

Cohort-ID	Rd50	G50	NO_2
a	80	0.5	35
b	0	0	20
c	20	0.45	25
d	170	0.7	45
e	40	0.05	25

Figure 10. Diagram of Land Use Regression (LUR) steps used in this project. Adapted from Gulliver and de Hoogh, 2015.

Available data for this project included output from the ADMS dispersion model, traffic related variables, population and household density, building footprints attributed with height, land-use and high-resolution vegetation data (**Table 4**).

I calculated the traffic related variables used in this project from length of roads, length of major roads (>10,000 vehicles), traffic flow and heavy traffic flow in multiple circular distance buffer sizes (range 25 m – 1000 m) around monitoring sites. I derived the variables from Department for Transport and London Atmospheric Emissions Inventory (2008) data, for which road geometry is based on the Ordnance Survey Integrated Transport Network. To derive distance related predictor variables (e.g. inverse distance), proximities were calculated using the 'Near' function in ArcGIS. The 'Near' function measures Euclidean distance between the target and multiple source objects and reports the source object with the shortest distance to the target, 'Spatial join' is used to append data from the nearest source and target (e.g., major roads to monitoring sites). The in-program 'field calculator' was used for all transformations and calculations.

I derived population and household density from Office of National Statistics census data for 2011 (most recent). This data represents background levels of emissions and was calculated in multiple circular distance buffers around each monitoring station (range: 100 m – 5000 m). I expected this variable to contribute to elevated emissions from housing sources.

I derived data on building volume (area x height of buildings) from Ordnance Survey MasterMap™ topography layer building footprints, with Ordnance Survey Building Height Attribute. Data are high-resolution and accurately represent the dimensions of buildings (+/-0.5 m horizontal and vertical). I calculated building variables in multiple circular distance buffers to provide potential predictor variables that represent potential pollutant-trapping environments (e.g., street canyons). Furthermore, following Tang et al. (2013), major roads (>10,000 vehicles) were buffered (20 m) and the building area falling within this road buffer was then intersected with a standard circular buffer around monitoring sites (e.g., 20 m road buffer of buildings, re-buffered by 50 m circular buffer at monitoring sites) to generate a building topography predictor variable that is only present around traffic-related sources.

The GeoInformation Group provided detailed (2.5 m x 2.5 m pixel) vegetation data for Greater London, which had attribution for vegetation height categorisation of tree cover (≥2.5 m) and groundcover (<2.5 m). **Figure 11** shows the intersection of the vegetation dataset with multiple circular distance buffers around an example monitoring station, Marylebone road), with trees shown in dark green and groundcover shown in light green. Using the in-program *Field Calculator*, I calculated tree cover, ground cover and total area cover as potential predictor variables for the LUR model.

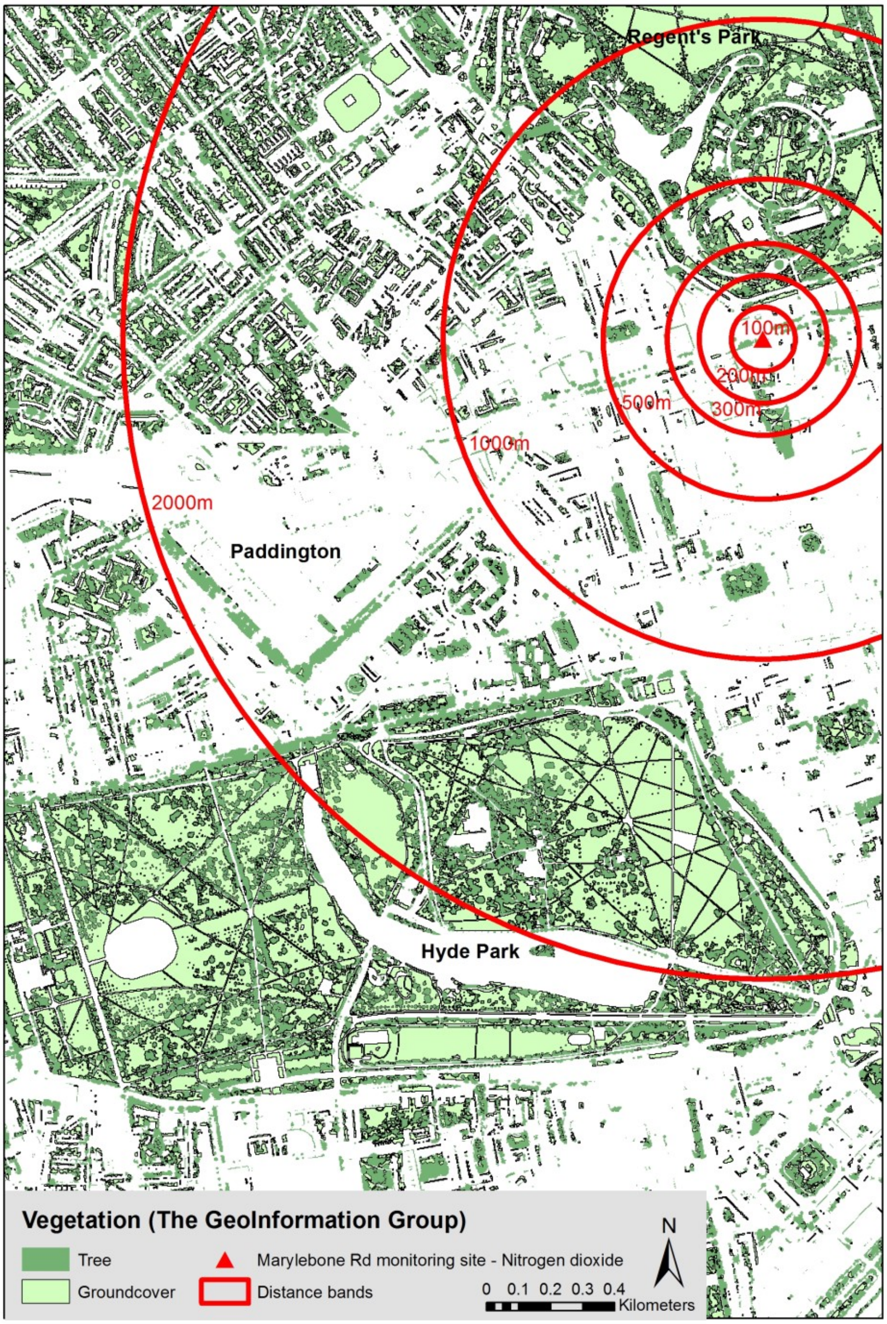


Figure 11. An example London Air Quality Network monitoring station for nitrogen dioxide (Marylebone Road) with concentric buffers of multiple distance radii (red lines) with high-resolution vegetation data (2.5 m x 2.5 m pixel resolution) provided by The GeoInformation Group.

Table 4. Potential predictor variables with unit, buffer sizes and a priori estimate of direction of effect.

Predictor variable	Unit	Circular buffer sizes (radii in m)	Direction of effect
Vegetation cover (The GeoInformation Group)			
Trees	m^2	25, 50, 100, 200, 300, 400, 500, 1000, 1500	+/-
Ground cover vegetation	m^2		-
Total vegetation	m^2		-
Census (Office of National Statistics)			
Number of households	N	100, 200, 300, 400, 500, 1000, 2000, 5000	+
Number of inhabitants (population)	N		+
Traffic (Department for Transport)			
Road length	m^2	25, 50, 100, 300, 500, 1000	+
Road length of major roads	m^2		+
Traffic intensity on nearest road	veh*day^{-1}	N/A	+
Traffic intensity on nearest major road	veh*day^{-1}		+
Heavy-duty traffic intensity on nearest road	veh*day^{-1}		+
Heavy-duty traffic intensity on nearest major road	veh*day^{-1}		+
Traffic intensity of all roads	veh*day^{-1}*m	25, 50, 100, 300, 500, 1000	+
Traffic intensity of major roads	veh*day^{-1}*m		+
Heavy-duty traffic intensity of all roads	veh*day^{-1}*m		+
Heavy-duty traffic intensity of major roads	veh*day^{-1}*m		+
Inverse distance to nearest road	m^{-1}	N/A	+
Inverse distance to nearest road, squared	m^{-2}		+
Inverse distance to nearest major road	m^{-1}		+
Inverse distance to nearest major road, squared	m^{-2}		+
Traffic intensity, inverse distance to nearest road	veh*day^{-1}*m^{-1}		+
Traffic intensity, inverse distance to nearest road squared	veh*day^{-1}*m^{-2}		+
Traffic intensity, inverse distance to	veh*day^{-1}		+

nearest major road	*m^{-1}		
Traffic intensity, inverse distance to nearest major road, squared	veh*day^{-1} *m^{-2}		+
Heavy-duty traffic intensity. inverse of distance to nearest road	veh*day^{-1} *m^{-1}		+
Heavy-duty traffic intensity. inverse of distance to nearest road, squared	veh*day^{-1} *m^{-2}		+
Building dimensions (Ordnance survey MasterMap™ Topography)			
Building volume (area*height)	m^3	25, 50, 100, 300, 500, 1000	
Building volume (area*height), restricted to major road buffer of 20m	m^3	25, 50, 100	+
ADMS-Urban (variable: ADMS-Roads)			
Estimate of NO_2 from dispersion model by Cambridge Environmental Research Consultants (CERC)	µg	N/A	+

5.1.3 *Air pollution model development method*

As a preliminary indication of association with monitored annual average NO_2 concentrations, all potential predictor variables were regressed individually. Those with the highest R^2 were selected (the dispersion model variable and other traffic-related variables) and combined iteratively with variables representing the near-source environment (e.g., buildings, vegetation) and wider background sources (e.g., household and population density). Additional variables were retained based on their contribution to the model based on coefficients to optimize overall performance. I followed the conditions for variable retention that were reported in a paper describing development of ESCAPE LUR models for NO_x and NO_2 (Beelen et al., 2013). That is, variables were only retained if they satisfied the following conditions: 1) the increment of adjusted R^2 was greater than 1%; 2) the coefficient conformed to the pre-determined direction of effect (see **Table 4**); and 3) the *p* value was no greater than 0.05. Variable variance inflation factors (VIF) were checked to avoid collinearity in the model (VIF ≤3 acceptable). All statistical analyses were carried out in R Studio (v.0.99.489).

5.1.4 *Air pollution model evaluation method*

A commonly used evaluation method for LUR models is leave-one-out cross validation (LOOCV), whereby an independent prediction for each site is derived from the air pollution model. That is, the model variables are fixed, though the coefficients are allowed to change each iteration. In LOOVC, predictions are pooled for all sites and performance statistics are produced. However, it has been demonstrated that LOOCV tends to over-estimate the predictive accuracy of the model (Babyak, 2004). I used an alternative method of evaluation in this project, whereby n-20% of sites is used to predict concentrations of the remaining 20% of sites (i.e. grouped cross-validation; GCV). A stratified random sampling approach was used to assign sites to five groups (n = 20); site types were evenly distributed across groups. Performance of models was assessed by combining the predictions of each site from all groups (obtained via cross-validation of the five groups), and predicted concentrations were compared to monitored concentrations and model performance was summarised in terms of R^2 and root mean squared error (RMSE).

As conventional R^2 is measured around the best-fit regression line (i.e. the square of the correlation, which does not account for bias), Basagaña et al. (2012) suggested a method of transforming mean squared error (MSE) into a R^2-like formula to act as a more stringent measure of model performance (i.e. around the 1:1 line), defined as:

$$\text{MSE R}^2 = 1 - \frac{\text{MSE}}{\left(\frac{1}{\text{N}}\sum_{i=1}^{N}(y_i - \bar{y_t})^2\right)}$$

Equation 2. MSE R^2 as per Basagaña et al. (2012).

Where *yi* is the monitored concentration at each site and *yt* is the averaged monitored concentration. MSE-R^2 is thus one minus the mean squared error divided by the variance of the measurements. I used this more stringent evaluation measure to assess models.

5.1.5 *Dispersion-LUR model developed in this project*

Table 5 presents the final Dispersion-LUR model used in this project to estimate NO$_2$, and includes incremental values of adjusted R^2, standard error of the estimate (SEE), regression coefficients (β), p-values, variance inflation factors (VIF).

Table 5. Dispersion-LUR model variables and model descriptives. Note: Adjusted R^2 and standard error estimates are incremental values.

Variable description	Adj. R^2	SEE	β	p-value	VIF
Constant			41.2019334	<0.001	
Modelled dispersion variable for roads in 2010	0.67	9.96	1.4144094	<0.001	1.474
Inverse distance to nearest major road, squared	0.80	7.85	725.174287	<0.001	1.369
Groundcover in a 100m circular distance buffer	0.83	7.09	-0.0006535	0.001	1.210

As expected, the first variable retained in all annual average NO_2 models produced in this project was associated with roads and traffic flow. The final dispersion-LUR model that I developed in this project included the ADMS-Urban dispersion model output as the predictor variable, which represented vehicle flow emissions, and yielded an adjusted R^2 of 0.83. For comparison with this model, I created multiple alternative LUR models, without the ADMS-Urban dispersion variable (see alternative traffic related variables in **Table 4**), for which the model with the highest adjusted R^2 value was lower (R^2 = 0.78) than the dispersion-LUR model (R^2 = 0.83). I deemed the best LUR model suboptimal when compared to the dispersion-LUR, not only as it had a lower adjusted R^2, but it also predicted less well at Urban background and Suburban sites (where the majority of UK biobank addresses are located) compared to the dispersion-LUR model (i.e. it was more optimised for roadside sites than background sites due to underlying model variables). Further, SEE was lowest for the dispersion-LUR compared to alternative LUR models.

In the final model (dispersion-LUR), the dispersion modelling (ADMS-Urban) output variable for road sources explained a large portion of the variance in NO_2 concentrations (67% variance). Further variance was explained by the next two variables retained in the model – the *Inverse distance to nearest major road, squared* variable (extra 13% variance), and the *Groundcover in a 100m circular distance buffer* variable (extra 3% variance). The order that the variables are selected into the model impacts the extra contribution. There was low collinearity of variables (VIF < 3).

A high maximum Cook's Distance value of 0.90 was reported from the final model due to one Central London roadside site – City of London, Walbrook Wharf, 117.43µg / m^3 – which had the highest concentrations of annual average NO_2 of all roadside sites. I retained this site in the dataset as I deemed the particularly high influence an artefact of removing kerbside sites from the monitoring site dataset (i.e. having few high concentration sites remaining in the dataset). Most UK Biobank address locations were not expected to be exposed to concentrations as high as this monitoring site, however some Central London roadside addresses might near such high concentrations, and I retained the influential variable to help with prediction in such cases. The next most influential site was urban background (Hillingdon – Harlington – 34.26) and had an acceptable Cook's Distance value (0.12), signifying lower influence.

Coefficients multiplied by the 90th percentile minus 10th percentile were calculated (as in Eeftens et al., 2013) to give their 'typical' contributions to predicted values. The contribution of *Groundcover in a 100 m circular distance buffer* was larger at urban background and suburban sites than at roadside sites, whereas *Inverse distance to nearest major road, squared* contributed primarily in roadside locations (**Table 6**). In effect, the *Inverse distance to major road, squared* variable substantially contributes to estimates only when an address is located at close proximity to a major road.

Table 6. Measured annual average nitrogen dioxide concentrations at all, roadside, and urban background + suburban sites, and coefficients for dispersion-LUR model variables multiplied by the 90th percentile minus 10th percentile (β*(P90-P10) as an estimate of typical contribution to modelled annual average nitrogen dioxide concentrations for all sites, roadside and urban background + suburban sites.

	Modelled dispersion variable for roads in 2010 (ADMS-Urban) contribution (µg/m^3)	**Inverse distance to nearest major road, squared contribution (µg/m^3)**	**Groundcover in a 100m circular distance buffer contribution (µg/m^3)**
All sites (n = 50)	19.68	14.17	-10.00
Roadside (n = 25)	31.89	18.59	-8.08
Urban background + Suburban (n = 25)	9.08	0.81	-11.62

Predicted NO_2 concentrations associated with each variable by site can be found in the stacked bar plot presented in **Figure 12**. Note again, the *Inverse distance to nearest major road, squared* variable typically contributes to concentration estimates in roadside locations, and *Groundcover in a 100 m circular buffer* can act as an important predictor of concentrations at urban background and suburban sites. The stacked bar plot shows the additive nature of the variables to the NO_2 concentration estimates and demonstrates the strength of the model in accounting for intra-urban variability. For example, at multiple sites, major roads are located far from the monitor and the contribution of the *Inverse distance to major road, squared* variable to the total nitrogen dioxide concentration at that monitoring site is zero.

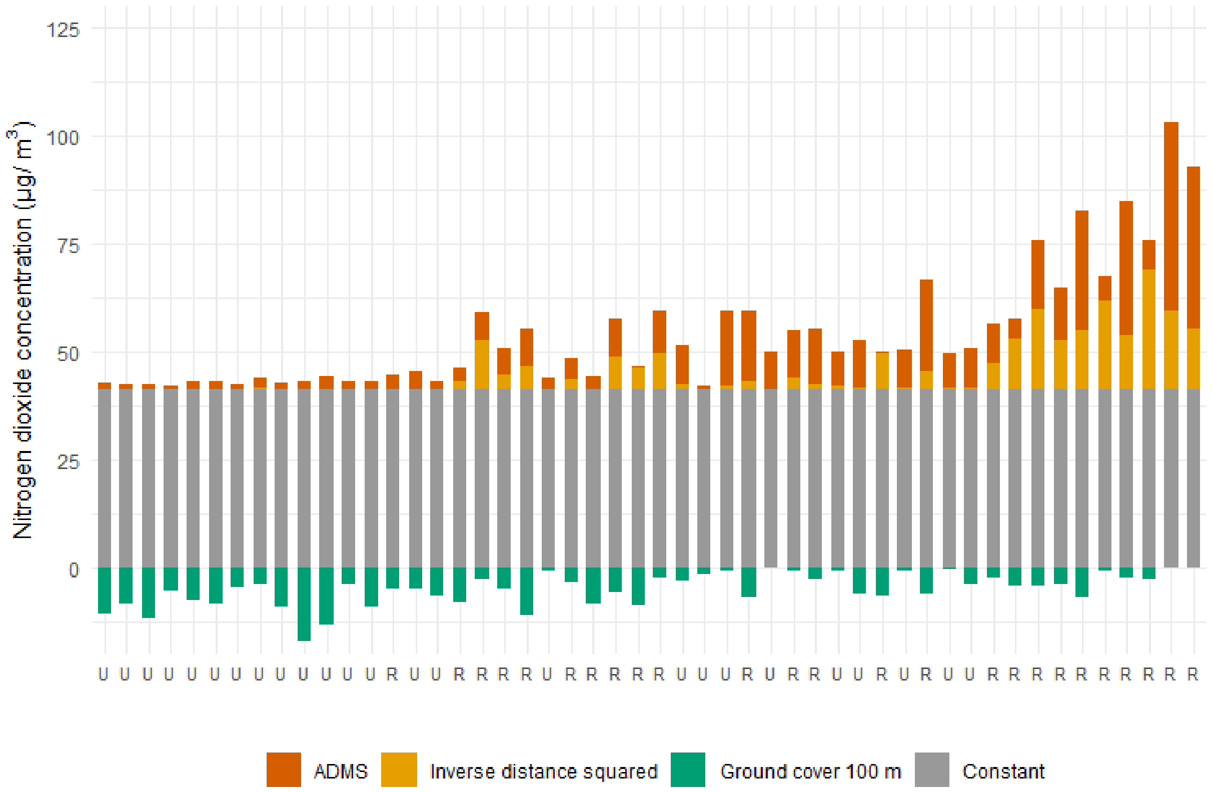


Figure 12. The proportional contribution of model variables at each site to total concentrations, for the Dispersion-LUR (with ADMS-Urban) model ordered (left-to-right) by measured annual average nitrogen dioxide concentration at each site. Site type is displayed on the x-axis (R = roadside, U = urban background + suburban sites). Note: the negative values for the variable ‘groundcover within 100m’ should be subtracted from the positive values (constant and other variables) to produce the correct estimated concentrations of NO_2 at each site.

5.1.6 *Dispersion-LUR model validation*

Table 7 shows results of the grouped (leave-20%-out) cross-validation (GCV) of the dispersion-LUR model. The R^2 and MSE R^2 values obtained from GCV were 0.79 and 0.80, respectively, indicating a relatively small reduction in overall performance. That is, compared to the model, the GVC explained 4% lower variation as per the R^2. The root mean squared error (RMSE) was adequately low (7.89), indicating a good absolute measure of fit between the measured concentrations and the predicted values from the model. The beta coefficient (β) of 0.83 (95% CI = 0.71, 0.95) indicates adequately accurate prediction.

Table 7. Summary statistics of results of grouped (leave-20%-out) cross-validation (GCV).

Model	R^2	MSE-R^2	RMSE	β	95% CI (lower–upper)
Dispersion-LUR (with ADMS-Urban)	0.79	0.80	7.89	0.83	0.71–0.95

Figure 13 shows observed (monitored) concentrations (X-axis) plotted against predicted concentrations from GCV for the model. The regression line (grey line) versus the 1:1 line (dotted line) shows very good fit. There is slightly greater variance between observed and estimated concentrations at the extremes of the distribution; bias (under-estimation) at very high concentrations (roadside locations) and bias (slight over-estimation) at low concentrations (urban background and suburban locations), though overall the model fit is good.

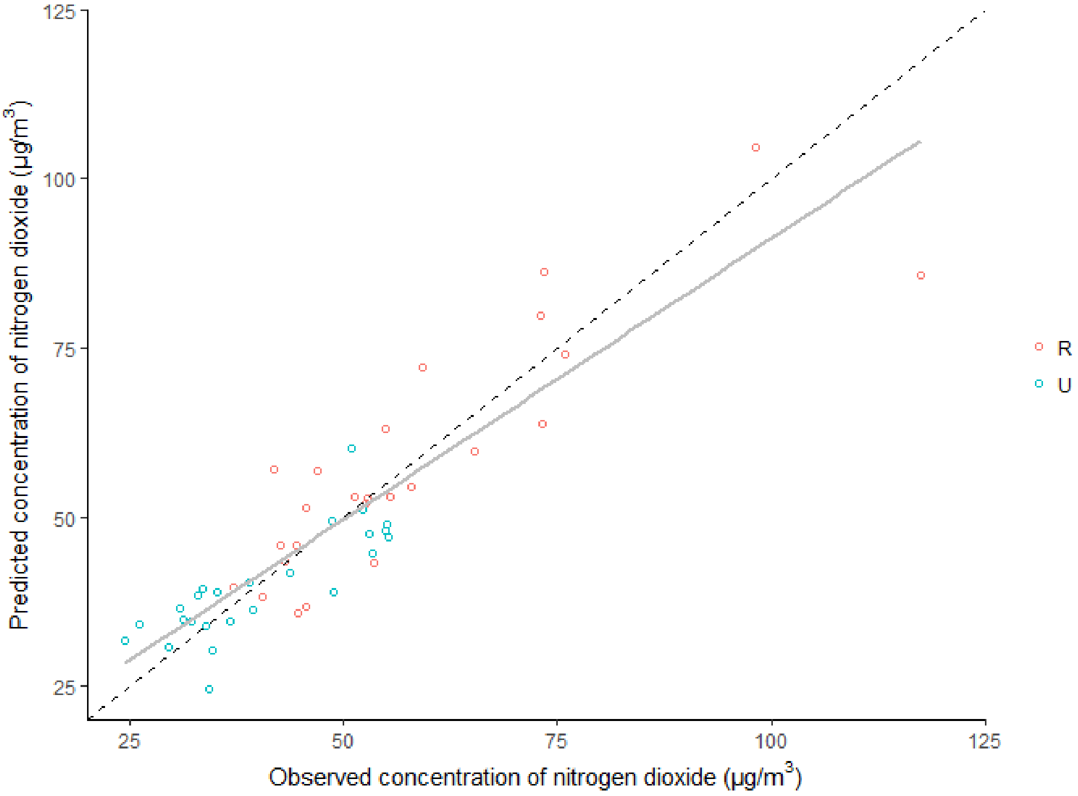


Figure 13. Predicted annual average nitrogen dioxide concentrations from grouped cross-validation (y-axis) plotted against monitored annual average nitrogen dioxide concentrations (x-axis) at 50 monitoring sites for the dispersion-LUR (with ADMS-Urban) model developed in this PhD project. Roadside sites (R) are represented by red circles and urban and suburban background sites (U) by blue circles. For comparison with the regression slope of the model (grey line), the dotted line indicates a 1:1 relationship of observed vs. predicted concentrations.

5.2 Application of model to UK Biobank London addresses

5.2.1.1 Standardised receptor placement

To standardise placement of address receptor points, I linked each address point with the nearest building polygon from OS MasterMap™. In most cases, this was the building that contained the geocoded address point. Following this, I generated address receptors (points) following a method developed by Gulliver et al. (2015). In brief, I modified and ran a PostGIS script that placed receptor points 1 m from the façade of the building associated with each geocode, on the side of the building closest to a main road (Figure 14). This was the assumed entry point of the building. The primary reason for creating standardized receptors on the exterior of the building was for the application of the traffic noise model (see section 5.4), and the use of these receptors in the air pollution modelling was primarily for continuity and comparability across exposure variables.

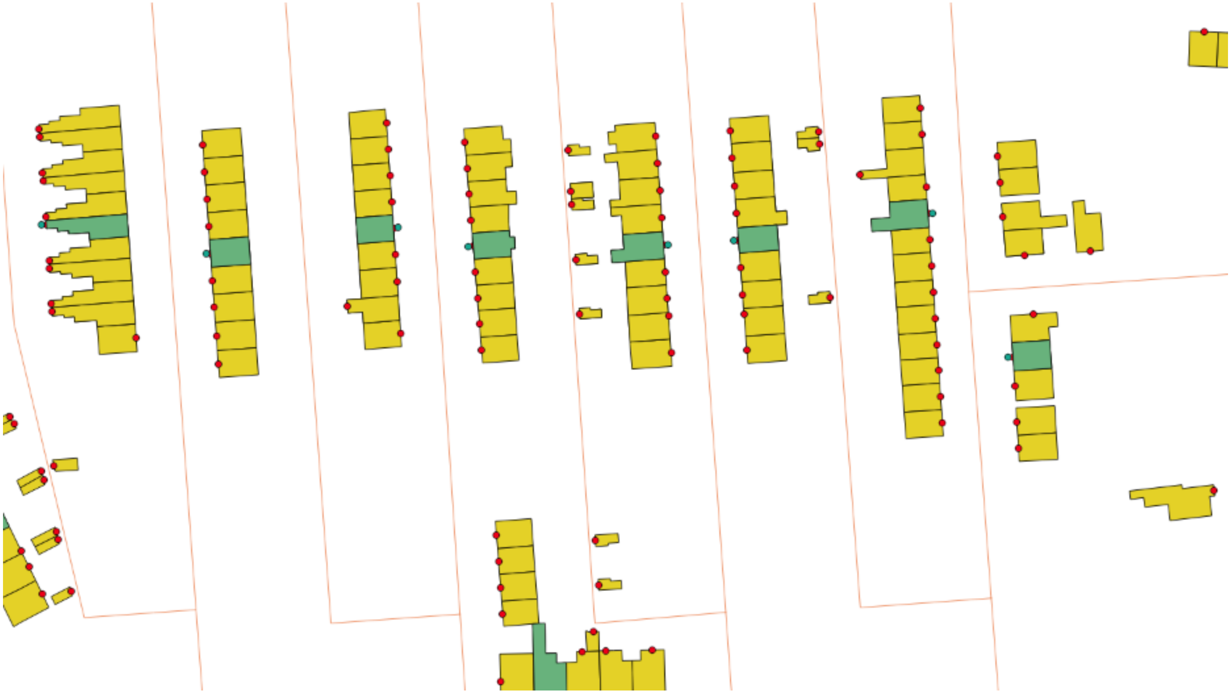


Figure 14. Diagram of building polygons (yellow and green), with green polygons representing addresses of cohort participants (note: example addresses not from UK Biobank). Receptors (red points) were produced for all buildings that were entered into the receptor generation script, only the receptors associated with the buildings of interest (green) were moved 1 m from the façade of the building and retained for air pollution and noise modelling.

5.2.1.2 Scaling the dispersion (ADMS-Urban) modelling

To scale the dispersion (ADMS-Urban) modelling for estimation at all address receptors in Greater London (n = 58,801), I used the same model inputs as those used to model ADMS at monitoring sites, including the 51 tiles of emission rates along major roads (points located every 10 m on road geometry with vehicle emissions rate attribution). However, for the monitoring site dataset (n = 50), NO_2 contributions from all 51 tiles were modelled for each monitoring site and summed. Often the contribution from a tile located far (e.g. > 300 m) from a monitoring site was zero. To run 58,801 address receptors over 51 model runs was estimated to far exceed the remaining time in which I had to complete the PhD project. To improve runtime, I assigned addresses to the emissions tile in which they fell, and to all neighbouring tiles within 300 m of the address using ArcGIS. This ensured that addresses located towards the edge of tiles received modelled NO_2 contributions from neighbouring tiles, whilst reducing run time (maximum 4 runs per receptor versus 51 times, previously). Given that this description is potentially easier for the reader to interpret graphically, I provide **Figure 15** that shows emissions 'tiles' and overlap. From output tables, average annual contributions from all tiles (maximum 4) per unique receptor point were summed to produce final dispersion model estimates to be used in the LUR.

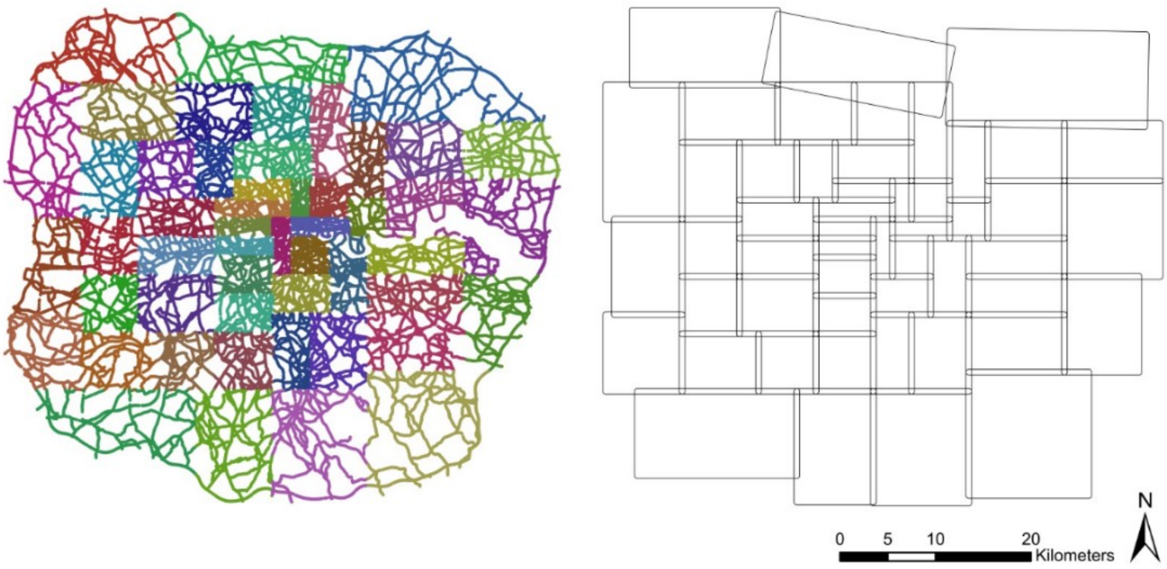


Figure 15. Data on emissions rates assigned to points at 10 m intervals along major roads in London, split into 51 tiles (shown in multiple colours) in preparation for running in ADMS-Urban dispersion modelling software (left). Bounding boxes created around the 51 tiles with a buffer of 300 m to capture UK Biobank receptors (addresses) falling within each tile and within neighbouring tiles (right). UK Biobank receptors were modelled only for tiles in which they intersected to increase model run efficiency.

5.2.1.3 Ground cover within 100 m circular distance buffer

Ground cover (2.5 m x 2.5 m) within 100 m circular distance buffer was calculated for each address receptor using an SQL script in PostGIS, as described in

Section 4.1. The use of PostGIS, improved efficiency when working with detailed data surrounding over 58,000 addresses.

5.2.1.4 Inverse distance to nearest major road, squared variable

This variable was calculated in ArcGIS using the *Near* function as per the model development methodology. In a few instances (<100 occurrences) the distance calculated by the *Near* function was unrealistically close to the road (e.g., 1 m from road centre line, or in the road). Given that the contribution of the *Inverse distance to nearest major road, squared* variable was based on monitoring stations that were not located on the road (<3 m from the road centre line), this artefact of address receptor generation would have been problematic for prediction of NO_2 concentrations. To account for this, I set a minimum distance to road centre line of 3 m (i.e. no address receptors were permitted to be <3m from the road).

5.3 Findings and discussion

I modelled NO_2 estimates for the year 2010 for addresses in Greater London using a dispersion-LUR model. Predictor variables on land cover classes and distance to source (e.g. road) were derived using a GIS. In addition, traffic-related air pollution concentration estimates were modelled using an air pollution dispersion model (ADMS-Urban). I regressed the output from the ADMS-Urban (variable: ADMS Roads), with distance to source and high-resolution vegetation (ground cover) data, against air pollution measurements (i.e. dependent variable) to derive a Dispersion-LUR model. I used the variables and their coefficients from the dispersion-LUR modelling stage to estimate NO_2 concentration estimates (μg/m^3) for Greater London UK Biobank addresses. Addresses outside of Greater London were not modelled due to the extent of the underlying vegetation data. The average (mean) concentration of annual average NO_2 concentration across all Greater London UK Biobank addresses was 36.63 (±5.26) µg/m^3 in NO_2 concentrations, which suggests residences are largely located in background, as oppose to roadside, locations. From the modelled estimates at UK Biobank addresses, the average contribution of *groundcover within a 100 m circular distance buffer* was a reduction of 6.88 (±2.55) µg/m^3. As expected, *Inverse distance to nearest major road, squared*, contributed little to estimations (i.e. only in addresses situated close to major roads, of which there were few), and therefore on average this variable added 0.88 (±2.78) to modelled NO_2 concentrations.

ADMS-Urban is a proprietary air pollution dispersion modelling tool developed to incorporate emissions from individual sources (e.g., roads, industrial point sources and area sources). It provides local-scale air pollution estimates within cities. I used ADMS modelling software to estimate NO_2 concentrations at address-level receptor points. I used the NO_2 concentration output from this model as a predictor variable in the LUR model. I included the *Inverse distance to nearest major road, squared* variable to increase the gradient in modelled concentrations around roads, potentially compensating for the air pollution dispersion (ADMS-Urban) model having shallower air pollution gradients around roads than expected due to high levels of ventilation conditions in the model (modelled on 'flat world'). I made the contribution of '*inverse distance squared to the nearest major road'* to the estimated NO_2 concentration (µg/m^3) at each address (receptor point) available in the exposure assessment data integrated into UK Biobank. I also made the (subtractive) contribution of the variable to the total annual average estimated NO_2 concentration (µg/m^3) at each address (receptor point) available in UK Biobank.

5.3.1 *Further consideration of model variables*

5.3.1.1 Ground cover within 100m circular distance buffer

Vegetation (e.g., ground cover) is not included in the air pollution dispersion model ADMS-Urban and is therefore not directly represented in the model estimates. The contribution of *ground cover in a 100 m circular distance buffer* in the dispersion-LUR model developed in this project, which moderately improved prediction of NO_2 concentrations at monitoring stations, theoretically, might be explained by a) the deposition of pollutants on nearby vegetation; b) air flow and dispersion of pollutants over open space (particularly low height ground cover); c) a lack of NO_2 sources in the space occupied by vegetation (e.g., traffic or household sources); or, most likely, d) a combination of the above. Ground cover in all buffer sizes was a stronger predictor of NO_2 concentrations than tree cover and total cover (tree cover and ground cover combined), potentially indicating that air flow and dispersion of pollutants over low-lying vegetation could explain some of the predictive contribution of the variable. *Groundcover within a 100m circular buffer* improved model prediction in the expected direction, and, although the groundcover variable explained comparatively less of the variance (adjusted R^2) than the other variables in the model, it was substantially more important in absolute terms (typical contribution in $\mu g/m^3$) in the estimation of NO_2 concentrations in urban background and suburban sites, which were expected to represent the majority of UK Biobank addresses, than *Inverse distance to nearest major road, squared* variable (**Table 6**).

In comparing the typical (expected) proportional contribution of groundcover at urban background and suburban sites, and roadside sites, the difference between the 10th and 90th percentile of groundcover contribution was equivalent to a difference in NO_2 concentration of -11.62 μg/m^3 and -8.08 μg/m^3, respectively. Urban background and suburban sites typically had ~28% groundcover cover within a 100 m buffer, whereas Roadside sites typically had ~22%, which explains the typical (estimated) difference (-3.54 μg/m^3) in modelled contribution between site types. Typical (expected) contributions of groundcover from the dispersion-LUR model are feasible in relation to average observed (measured) concentrations at roadside and urban background site monitors (58.32 μg/m^3 and 40.46 μg/m^3, respectively). These had an average difference of 17.86 μg/m^3. The highest monitored annual average value in dataset was 117.43 μg/m^3 and the lowest was 24.34 μg/m^3, a range of 93.09 μg/m^3. In light of differences between sites in London, the typical (expected) contribution of vegetation cover is in the expected order of magnitude. *Inverse distance to nearest major road, squared variable*

ADMS-Urban (without buildings) is modelled on a 'flat world'. Due to this, steep pollutant concentration gradients around major roads are not accounted for in the model. The *Inverse distance to nearest major road, squared* variable is a proxy for NO2 gradients perpendicular to roads in situations where buildings cause pollution trapping, and contributes to predictions primarily at roadside locations where pollutant concentrations are high. The inclusion of this variable in the model can be considered as a secondary road proximity-related variable that allows for greater flexibility (steeper gradients) in the Gaussian distribution of ADMS-Urban. In effect, the improvement in estimation of NO_2 concentrations at roadside sites is improved by superposing a second distribution, which allows for deviation from the original normal distribution. The result of this is a reduction in the underestimation of the model at roadside sites or, in application to the cohort, reduced underestimation of concentrations at addresses located at busy roadside locations.

In the model development stage, variables based on the work of Robert Tang's PhD thesis (Tang et al., 2013), that represented street canyon topology were created. The expected strength of the *Building volume, restricted to major road buffer of 20 m* variable that Tang developed was that limited ventilation is a strong contributor to higher concentrations only when a source is present, thus the model improves estimates through better representation of pollutant trapping (Tang et al., 2013). Indeed, in recreating this variable and assessing its predictive contribution in the dispersion-LUR on monitored concentrations, this was true. Buildings and groundcover variables, when present within a site buffer, refined estimates that were based principally on traffic flow emissions (ADMS-Urban) by adding more detailed information about the near-site physical environment, which alter pollution dispersion. However, the application of the *Building volume, restricted to major road buffer of 20 m* variable in the context of a cohort was problematic. If included, the variable would introduce an arbitrary cut off (e.g., 20 m distance from a major road) for receiving the contribution from the building volume variable, addresses over 20 m from the major road would receive zero. This could not be overcome by using the *Building volume* (on all addresses) variable, as this offered little predictive contribution to the dispersion-LUR model, as where there were low emissions, the building height topography was unimportant. I deemed the 20 m cut off likely to produce misclassification for addresses located around major road (e.g., steep decline in concentrations after 20 m distance), and opted for the predictor, *Inverse distance to nearest major road, squared* variable, which I expected to be less precise for capturing canyon scenarios accurately, though I considered to introduce adequate flexibility to the dispersion model (ADMS) output distribution to better predict higher concentrations, without the introduction of an arbitrary cut off surrounding major roads. That is, the *Building volume* variable would not be applicable to most residential locations, whereas *Inverse distance to* *nearest major road, squared* applies everywhere even if the contribution is very small away from major roadside locations.

Comparisons of the typical (expected) proportional contribution of each variable in the model via the method of Eeftens et al. (2013) – calculated by multiplying the variable regression coefficient by the 90th minus 10th percentile – identified the very low contribution of *Inverse distance to nearest major road, squared* variable at urban background and suburban sites (located away from major roads). Given that the majority of UK Biobank participants are expected to reside in non-roadside locations, a stronger influence of *groundcover 100 m* was expected in the cohort address-level exposure assessment, and *Inverse distance to nearest major road, squared* was expected to contribute primarily at addresses located at (major) roadside locations (i.e. improving estimation for a small minority of addresses at the upper tail end of the exposure distribution).

NO_2 deposition on vegetation in the UK was previously estimated to be negligible (Jones et al., 2017a), the improvement in prediction provided by *groundcover within 100 m circular distance buffer* variable is likely to linked to improve dispersion of pollutants over lower lying groundcover surrounding the receptor. The inclusion of groundcover, as oppose to total vegetation cover or tree cover supports this proposition, as low lying vegetation avoids entrapment of pollutants in built up urban contexts. It should be noted that the model was developed from measured NO_2 concentration data from 50 monitoring sites in roadside and background locations. At these sites, high percentage cover of groundcover in the 100 m surrounding a monitoring site, indicates that other landcover (e.g., tall buildings) are not dominating the area. The variable therefore provides information on groundcover versus ‘other’ cover within 100 m distance to the model; both pieces of information are likely informative for prediction. If the area surrounding the monitor is covered in low lying vegetation, it is important to recognise that it is well ventilated compared to a monitor that is tightly sandwiched between a road and building. I suggest that dispersion and air flow, as oppose to deposition of NO_2 on vegetation surfaces, could be responsible mechanisms. In the dispersion-LUR model, however, *groundcover in a 100 m circular distance buffer* simply acts as a ‘sink’ in the pollutant surface produced by ADMS.

5.4 Traffic noise model

The common framework for noise assessment methods (CNOSSOS-EU) was developed to harmonise assessment of noise levels from major sources (road and rail traffic, aircraft and industry) across Europe (Kephalopoulos et al., 2014a). Morley et al. (2015) coded (in Structured Query Language) the CNOSSOS model as an algorithm. The algorithm calculates traffic noise levels by creating ray paths (straight lines) from each address receptor point to all road source points (points located at 20 m intervals along the road network) within a 500 m circular distance buffer. Propagation of modelled sound along each ray path (source to address receptor) is adjusted for sound absorption and/or diffraction based on traffic flows and speeds, diurnal traffic profiles, source geometry, ray length (distance), land cover data, topography, building heights, and prevailing wind direction. Sound levels from all ray paths surrounding each address receptor are summed logarithmically to produce a single modelled noise exposure value per address receptor (hourly average L_{Aeq}), which is then A-weighted to represent the relative loudness of sound as received by the human ear. **Figure 16** by Morley et al. (2015) graphically summarises the CNOSSOS model algorithm.

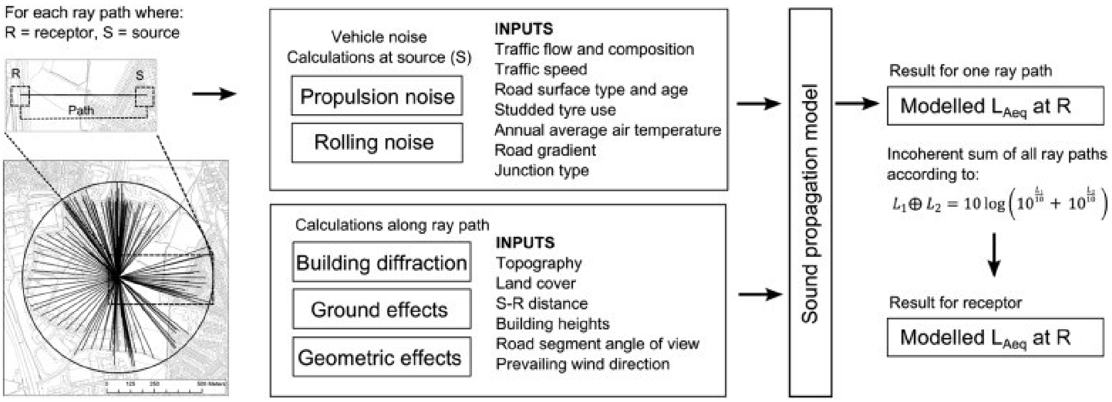


Figure 16. Schematic representation of the CNOSSOS-EU road noise model data processing flow by Morley et al. (2015).

The CNOSSOS model has previously been used for residential address-level noise exposure assessment in UK Biobank as part of the Biobank Standardisation and Harmonisation for Research Excellence in the European Union (BioSHaRE-EU) project, which required a harmonised noise exposure measure across several EU cohorts. BioSHaRE-EU used input data that was available for all countries in the study (i.e. across the EU). In the UK, where detailed input data were available, Morley et al. (2015) showed the accuracy of traffic noise estimation was improved via the use of high resolution data inputs in the CNOSSOS modelling framework versus lower resolution inputs (i.e. inputs used in BioSHaRE-EU). Morley et al. conducted validation of the algorithm, including a measurement campaign in Norwich, UK.

Added to this, Morley and Gulliver (2016b) developed a method to enhance traffic noise model input data (traffic flow data). Typically, traffic flow data are not available for minor roads, which often service residential areas, and are entered into noise models as a constant number of vehicles per day. To estimate variations in traffic flow on minor roads and better predict noise levels at residential addresses serviced by minor roads, Morley and Gulliver (2016b) developed a routing algorithm to rank roads by importance based on simulated journeys through the road network in 2013.

Validation from measurements of traffic counts and noise by Morley and Gulliver (2016b) showed that estimation of minor road flow, which was derived from simulations, improved traffic noise prediction capability when compared to models that did not estimate minor road variability (Spearman's rho. increased from 0.46 to 0.72).

5.4.1 *Implementation of CNOSSOS algorithm in this project*

Using detailed data for London, I implemented a CNOSSOS algorithm, which was scripted by Dr David Morley for use with high resolution (HR) inputs (CNOSSOS-HR), to assess annual average traffic noise exposure for UK Biobank residential addresses in Greater London (*N* = 58,881). I incorporated the 'minor roads' dataset (see above) produced by Morley and Gulliver (2016a) as the traffic flow component of the noise model. I implemented traffic noise modelling and receptor generation using PostGIS (v. 2.3.3) – a spatial extender for Postgres (v. 9.6), which I coded in Structured Query Language (SQL). I generated building receptors (i.e. residential addresses points) used for traffic noise modelling following a method developed by Morley and Gulliver (Gulliver et al., 2015). Receptors were set 1 m from the façade of the building associated with each participant's geocoded residential address, on the side of the building closest to a main road. This was the assumed entry point of the building and standardised receptor placement across addresses (Gulliver et al., 2015). Morley wrote the original receptor placement script that I used in this project. I updated Morley's script for use with a recent release of the OS Integrated Transport Network (2018). This was required due to changes in the road hierarchy classification (*descriptive term* attribute) compared to older versions. I optimised Morley's script for use with cohort addresses; the original script created a receptor for all buildings (OS MasterMap™ Topography building polygons) across the city, as oppose to specific polygons linked to specific geocoded residential addresses (cohort setting). Noise modelling used the same receptors as described in air pollution modelling methodology.

Input datasets for the noise model required manipulation in preparation for use with the CNOSSOS-HR algorithm script. Due to the large size of the datasets, I conducted the data manipulation in Postgres. When a shapefile was added to a Postgres database, it was stored as a table with a single column containing geometry data. The table can be indexed using the geometry column, which vastly increased processing efficiency compared to alternative geospatial software (e.g., ArcGIS), hence my motivation for using PostGIS. Preparation of the UK ‘minor roads’ dataset in PostGIS included: 1) partitioning daily (24 hour average) minor road flow estimates from the routing algorithm into hourly flow (diurnal profile of 24 timepoints; see **Figure 17**) using proportions corresponding to weekday traffic distribution on all roads by time of day (Department for Transport, 2013); and 2) creating a function that replaced the road hierarchy descriptive term (text) in the original table with a vehicle speed limit (numeric) corresponding to road hierarchy and vehicle classes (e.g., cars, light goods vehicles, heavy goods vehicles, 2 wheeled vehicles).

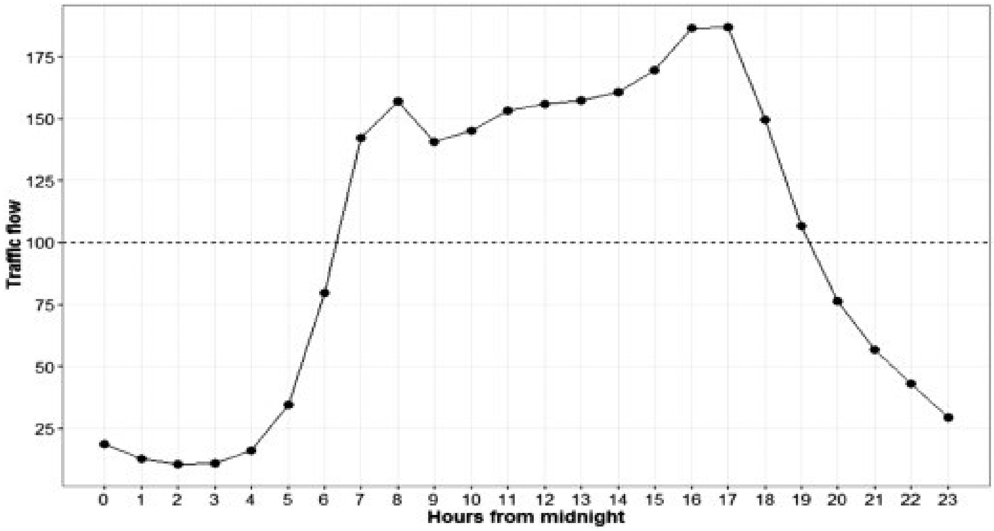


Figure 17. Example provided by Morley et al. (2015) of UK annual average hourly (Monday to Friday) traffic flow profile based on automated traffic counters. Traffic flow is represented as an index where 100 is the hourly average.

Additionally, I joined (by attribute) OS MasterMap™ Topography (land cover) with OS MasterMap™ Topography – Building Height Attribute (building height) on the building polygon TOID, which is an OS MasterMap™ unique identifier harmonised across OS MasterMap™ datasets. I set all other (non-building) polygons to zero height. OS datasets were provided for research purposes, free of charge, by Edina via Digimap (https://digimap.edina.ac.uk/). As required by the CNOSSOS algorithm, I added geographical (point) tables to the database (via the open-access Graphical User Interface *PostGIS 2.0 Shapefile Loader and Exporter*) with annual average temperature and annual average wind direction proportion (NE, SE, SW, NW) at the nearest Meteorological Office (MET) station (Heathrow Airport).

To calculate noise levels at each receptor, I ran (a modified version for high resolution data inputs) of the CNOSSOS algorithm script in PostGIS. For each ray path, sound level at the source point was derived from the 'minor roads' traffic flow data and empirical relationships defined by CNOSSOS-EU. Specific factors along each ray path were used to correct the level of sound received at the receptor from each traffic noise source point according to CNOSSOS-EU sound propagation guidelines.

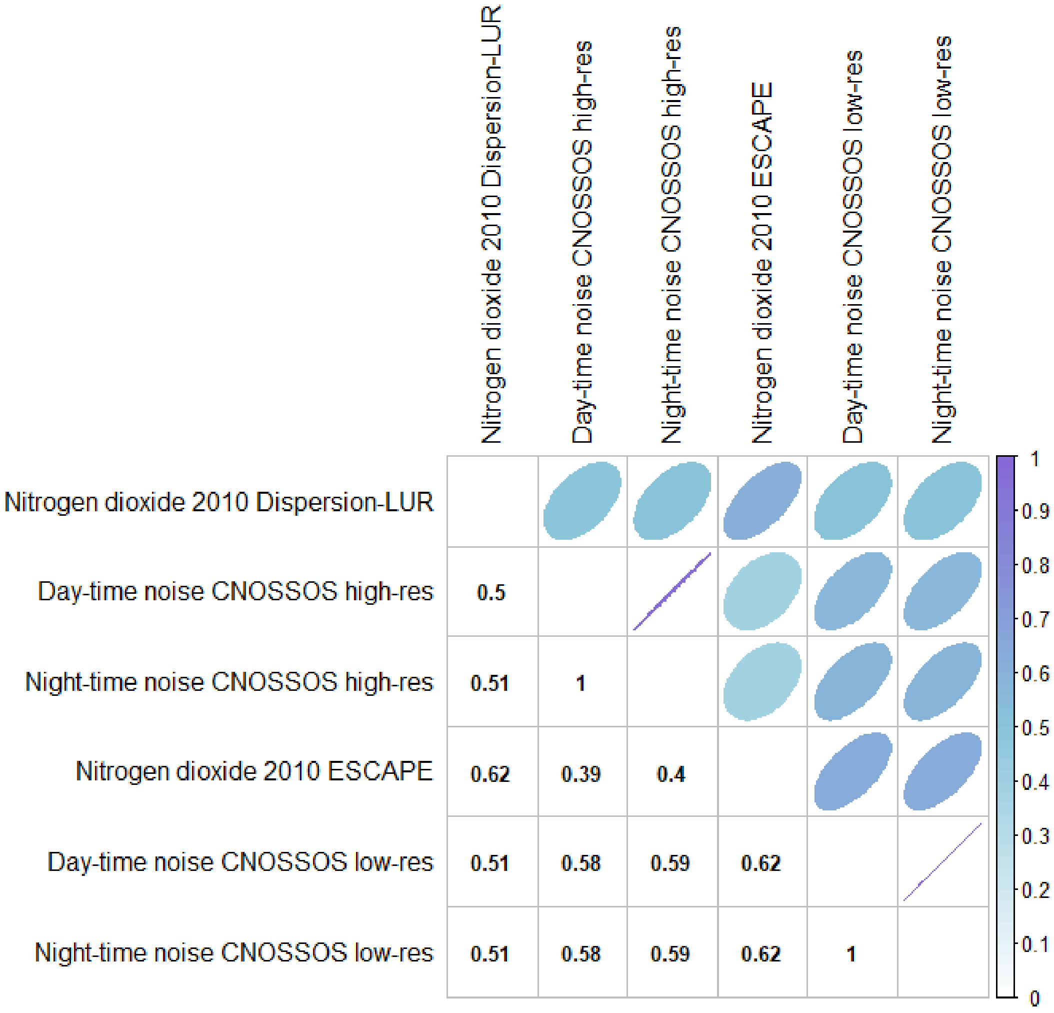
To attenuate propagation due to land cover absorbance, the algorithm divided each ray path into segments according to the OS MasterMap™ Topography category that it traversed (e.g. building, manmade surface, natural surface), and each segment was classed as sound absorbent (e.g. natural surface) or not (e.g. manmade surface) for adjustments to be applied (maximum adjustment = -3dB). Sound barriers (e.g., buildings) along ray paths were ascertained via height data for London (derived from LiDAR Digital Surface Model and Digital Terrain Model), and the relative heights of the source, receptor and barriers along the ray path were used for sound propagation correction. Further corrections for geometrical divergence and meteorological conditions were applied and a final sound level estimate for each ray path was produced.

The algorithm produced hourly A-weighted noise values (in decibels; dB(A)) that were averaged for time periods 00:00–23:00, 07:00–23:00 and 23:00–06:00 – denoted as $L_{Aeq,24\ h}$, $L_{Aeq,16\ h}$ and L_{night}, respectively. These averages are typical noise exposures used in epidemiological investigations (e.g., Vienneau et al., 2015). The algorithm also calculated another 24 hour averaged noise metric used in epidemiological studies (e.g., Clark et al., 2017), L_{den}. L_{den} is weighted with penalties of 5 dB for noise in the evening (19:00–23:00) and 10 dB for noise at night (23:00–07:00) to capture potential noise annoyance.

5.5 Findings and discussion

Though Spearman's rank correlation amongst exposures assigned in this project for annual average NO_2 concentrations and traffic noise were correlated with those previously assigned to UK Biobank addresses in Greater London, correlation was reduced from r = 0.62 between previous estimates to r = 0.51 between the high resolution model estimates assigned in this project. All models are likely to have some misclassification, though the use of high resolution inputs aimed to reduce misclassification, and aid with future identification of pathway-specific epidemiological effects.

Correlation of traffic-related air pollution (e.g., NO_2) and traffic noise exposures can limit epidemiological investigation of separate and combined effects of air pollution and traffic noise exposures due to problems caused by covariance in statistical models. However, a potentially important problem for environmental epidemiological interpretation using modelled environmental exposures is shown in **Figure 18**. The correlation plot shows low correlation of the high-resolution CNOSSOS noise model estimates that I assigned to UK Biobank, and the ESCAPE air pollution model estimates of annual average nitrogen dioxide concentrations that were assigned by other researchers (r = 0.39), which might be a product of using mismatched models (or mismatched resolution input data) and could result in biased epidemiological effect estimates. In this case, the ESCAPE model was developed using data available across EU study countries in the ESCAPE project, and data inputs were not optimised specifically for the UK or London. To provide meaningful findings in multi-exposure assessments of the built environment, it is crucial to consider the specificity of models (and model input data resolution) used to produce confounders (e.g., traffic noise in this PhD project), as well as focal exposure variables (e.g., air pollution).



	Nitrogen dioxide 2010 Dispersion-LUR	Day-time noise CNOSSOS high-res	Night-time noise CNOSSOS high-res	Nitrogen dioxide 2010 ESCAPE	Day-time noise CNOSSOS low-res	Night-time noise CNOSSOS low-res
Nitrogen dioxide 2010 Dispersion-LUR						
Day-time noise CNOSSOS high-res	0.5					
Night-time noise CNOSSOS high-res	0.51	1				
Nitrogen dioxide 2010 ESCAPE	0.62	0.39	0.4			
Day-time noise CNOSSOS low-res	0.51	0.58	0.59	0.62		
Night-time noise CNOSSOS low-res	0.51	0.58	0.59	0.62	1	

Figure 18. Correlation plot showing exposures assigned to UK Biobank London addresses in this project (nitrogen dioxide from dispersion-LUR model and high-resolution CNOSSOS noise model) and exposures previously assigned to the same addresses (nitrogen dioxide from ESCAPE LUR model and low-resolution CNOSSOS noise model).

5.6 Chapter summary

Development of a dispersion-LUR model and the contribution of the variables at roadside and background sites is discussed in depth in this chapter. In particular, the greenness variable included in the dispersion-LUR model – *Groundcover within a 100 m circular distance buffer* – is considered. Modelling environmental pathway-specific exposures (air pollution and traffic noise) with high resolution data was proposed to reduce exposure misclassification of participant addresses, which were mostly located in residential locations serviced by minor roads. Exposure assessment across Greater London UK Biobank addresses is currently integrated into the UK Biobank Data Showcase. I anticipate that updated traffic noise exposures, in concert with updated air pollution exposures, will enhance detection of environmental associations with adverse health outcomes. Finally, the risk of introducing systematic effect estimate bias in epidemiological analyses by using modelled environmental exposures based on different spatial resolution and quality inputs is also discussed at the end of the chapter.

6 Physiological pathway

The *physiological* pathway is posited to increase physical activity levels via the provision of amenable greenspace to carry out exercise (e.g., walking, jogging and cycling), with accompanying health benefits. To investigate this pathway, I developed a novel greenspace exposure assessment approach grounded in the established research framework of 'walkability'.

As oppose to disregarding known urban environment correlates of physical activity and walking, I aimed to assess the added contribution of vegetation cover surrounding a walkable transport network to walkability exposure after accounting for conventional walkability metrics. In this chapter I draw on some of the details of the principal studies that informed my methodological approach, and justify my choice of built environment walkability components in the context of this study. I provide a technical description of how I built road/path based network distance buffers for all UK Biobank addresses in Greater London (n = 58,234) using pgRouting, an extension of the PostGIS/Postgres geospatial database. I also provide an overview of how I improved the accuracy of network buffer creation beyond that of pgRouting integrated functions. Finally, I provide associations of self-reported physical activity levels of UK Biobank London participants and walkability of their residential addresses, as measured by the conventional walkability score (without vegetation integration), and with vegetation integration – i.e. via the 'green walkability' index.

6.1 Walkability index

In the context of the updated UK chief medical officer's physical activity recommendations for older adults, all physical activity – including low- and moderate-intensity physical activity, such as walking, for any duration of time – is currently considered to contribute to health in older adults. Substantial health improvements from creating more walkable environments would be expected, if the positive association of neighbourhood walkability and walking is proven to be causal. Based on self-reported physical activity data from the Whitehall II study – an occupational cohort of mid- to older-aged civil servants in the UK – Stockton et al. (2016a) developed walkability indices specifically for adults based in Greater London. Of the walkability indices developed, which varied in complexity, the 'basic' index integrated three built environment components: residential density, street connectivity, and land-use mix (entropy). The three-component walkability index showed a dose-response association with time spent walking per week.

Three core walkability components – population density, street connectivity and land-use mix (entropy) – have conventionally been included in walkability indices in varied geographical regions, such as the USA (Frank et al., 2007) and Australia (Giles-Corti et al., 2019), where most walkability research has been conducted (Grasser et al., 2017). Less research on walkability has been conducted in Europe, though the walkability indices produced by Stockton et al. (2016a) for Greater London were the foundation upon which I built the walkability methodology for this project. Broadly, I used the same (three) walkability components as Stockton et al. (2016a) for the walkability index developed in this study. However, due to data availability and my decision to assess walkability at individual address level (within road/path based network distance buffers), as oppose to administrative area-level, which is more computationally expensive, I used an alternative dataset to Stockton et al. (2016a) for one of the three components of the walkability index model. That is, I opted to use destination density rather than land-use entropy to quantify walking destinations; destination density (within a 3 component walkability index) has been shown to predict physical activity in mid to older-age adults (females) (Orstad et al., 2018). Computationally, such an approach reduced expense as spatial analysis (e.g., intersection) with point (destination) data is less demanding than with polygon (land cover) data. The index therefore includes the variables: population density, street connectivity (road/path junction density) and destination density.

6.1.1 *Input data*

6.1.1.1 Population density

To estimate population density, I used ONS estimates of number of residents per postcode based on 2011 census data. Census collection in the UK occurs every 10 years; I used 2011 headcount estimates as the census which most closely represents the UK Biobank population at baseline (2006-2010). Charlotte Sheridan (colleague) joined ONS 2011 population headcount estimates to the corresponding 2011 postcode centroid (point geometries with British National Grid X, Y-coordinates), which I used in this analysis. In England, postcodes represent on average 19 households, and they are the smallest geographic unit available from ONS.

6.1.1.2 Street Connectivity

I used OS Integrated Transport Network (ITN) with OS Urban Paths Theme Network extension (UPN), to represent roads with vehicle access, as well as footpaths, cycle paths and other pedestrianised throughways (**Figure 19**). Edina provided these data sets, free of charge, via Digimap (https://digimap.edina.ac.uk/). The road and path link geometries (lines) in this dataset fall within polygons that represent these features in the Ordnance Survey MasterMap™ data set, and road/path lines are typically positioned along the polygon centreline. OS ITN and UPN links are connected by ‘connecting links’, which have no real world geometry, but act as logical connectors between road and path geometries in the two data sets. OS ITN and UPN are detailed (scale 1:1250) and attributed using a standardised approach, and are accurately positioned relative to other OS products. For these reasons, I opted to use this data as oppose to open source data (e.g., Open Street Map).

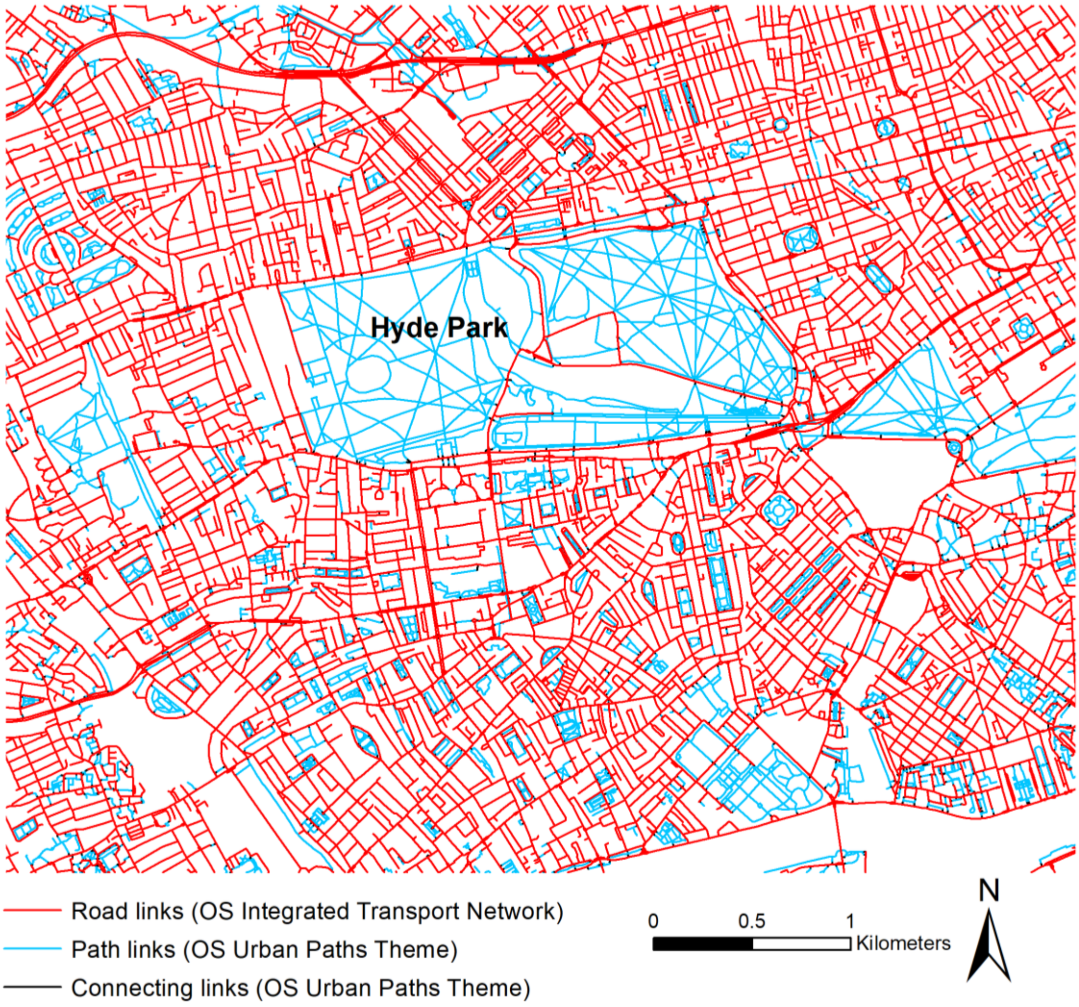

Hyde Park
Road links (OS Integrated Transport Network)
Path links (OS Urban Paths Theme)
Connecting links (OS Urban Paths Theme)
0
0.5
1
Kilometers
N

Figure 19. An example of Integrated Transport Network data (road geometry; red) and Urban Paths Network data (path geometry; blue), with connecting links (black), in and around Hyde Park, London.

6.1.1.3 Destination density

The Ordnance Survey Points of Interest (POI) dataset contains around 4 million unique geographical features across Great Britain, including businesses, services, transport and public infrastructure. I used this dataset to estimate destination density in the walkability model. Features (e.g., businesses) represented in the data set have national grid coordinates (point geometry; coordinate precision <1 m) and functional categorisation. I compiled a list of POI functional categories that I deemed relevant to the assessment of neighbourhood walkability. I used category codes to select and retain relevant POIs. Categories retained included restaurants, shops, markets, banks, sports facilities, hair and beauty services, schools, health centres, post offices, libraries, places of worship, public transport stations and bus stops. Edina provided national OS POI data for this project via the Digimap portal (https://digimap.edina.ac.uk/).

6.1.2 *Road/path network buffers*

Network buffers capture a more accurate representation of the area that can be traversed – e.g., when walking – than a circular buffer and are thought to better capture the spatial attributes of the neighbourhood that may influence physical activity (Giles-Corti et al., 2019, Oliver et al., 2007).

Figure 20 shows distance bands representing walking time in minutes based on the assumption that individuals walk along provided infrastructure. Note, in the lower panel, the greenspace in the North East cannot be accessed within a 5 minute walk despite being equidistant (as-the-crow-flies) to the northernmost point of the 4 minute walk band.

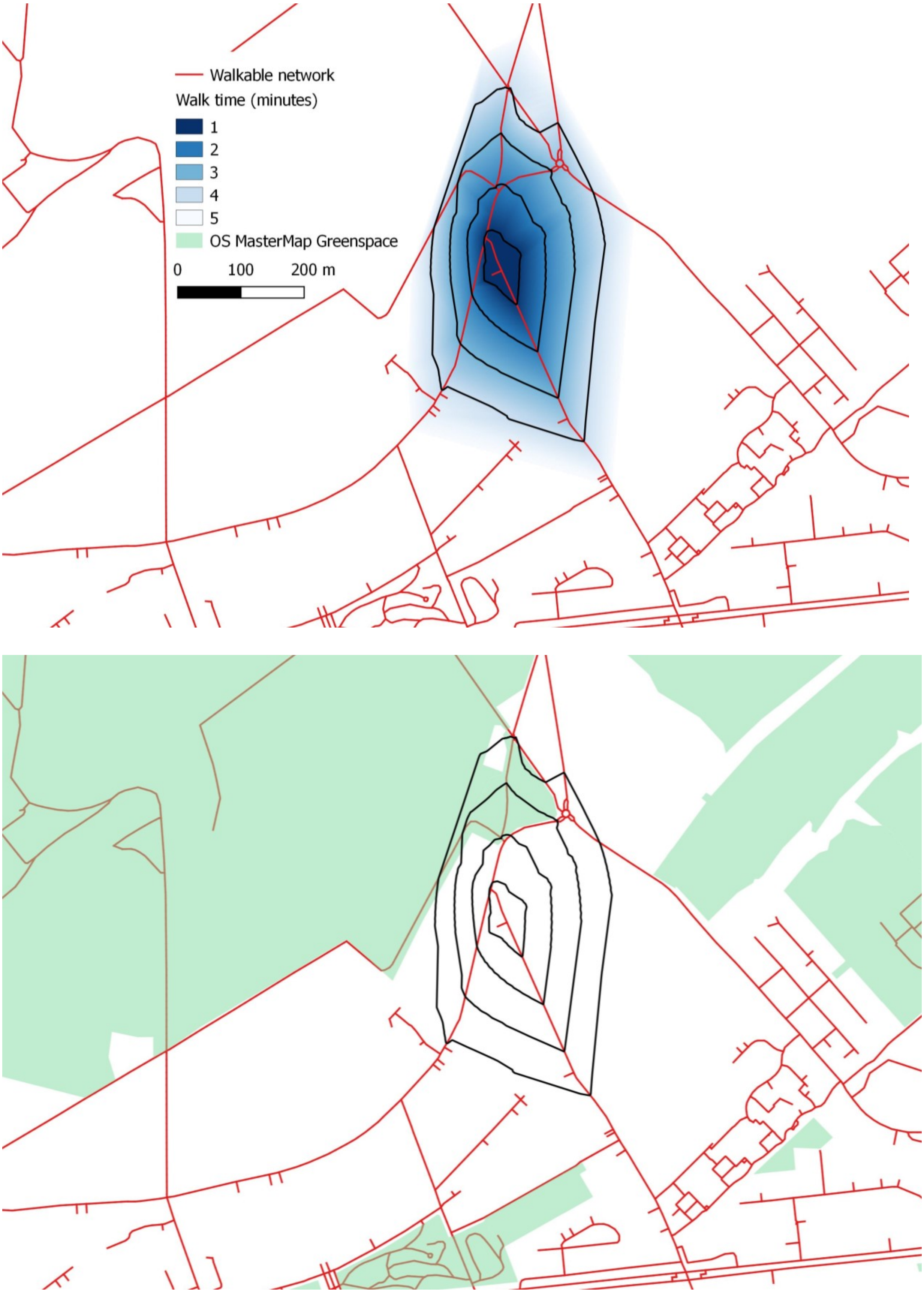


Figure 20. Distance bands representing walking time in minutes based on the assumption that individuals walk at 3 mph along provided infrastructure. The lower panel shows greenspace that can and cannot be accessed within a 5 minute walk from the start location (a dummy residential address).

After reviewing buffer sizes used in the literature, I selected a road and path network distance of 1000 m to represent a ~10-15-minute walk to reflect the near walkable neighbourhood. My justification for this selection included: a) a 1000 m street network distance had previously been used in a UK Biobank environmental exposure assessment (Sarkar et al., 2015a); b) a recent UK Biobank study showed associations in the expected direction of effect for physical activity establishment density (e.g., sports centres) and fast-food environments within a 1000 m street network and adiposity (Mason et al., 2018); and c) a large walkability study across 12 countries (including the UK) – the International Physical Activity and the Environment Network (IPEN) study –used a 1000 m (and 500 m) street network buffer (Adams et al., 2014b), and an ancillary study found protective associations of IPEN walkability 1000 m with cardiometabolic risk (Coffee et al., 2013).

6.1.2.1 Integration of Ordnance Survey Integrated Transport Network and Urban Paths Network

I created a walkable road and path network by routing OS ITN with the OS UPN extension, to capture roads and pedestrianised routes. Stockton et al. (2016a) previously applied this integration approach in area-level walkability analyses in London. To replicate their integration approach as closely as possible, I converted roads (OS ITN) and paths (OS UPN) data to ESRI File Geodatabase format using ESRI ProductivitySuite 3.5 *OS Data Convertor tool*, an extension of ArcGIS10.4.1. Following online advice (Ordnance Survey, 2013), I added road and path networks into a single File Geodatabase, using the same table name prefix (ITN_). I used the OSMM Data Preparation Tool to create a network dataset (stored in a new File Geodatabase) from the File Geodatabase containing the road and path data sets.

When building the network, I defined several network parameters of note: I penalised Motorway road links by setting them at a lower hierarchy than all other road links (it is illegal to walk along Motorways in the United Kingdom, hence the penalty) and I altered the speed of all road/path links to three miles per hour as an estimate of average walking speed (Ordnance Survey, 2013). The single, integrated transport network that was built enabled routing along all road/path links that are accessible to pedestrians. I validated the road/path network in ArcMap (v. 10.4.1) using the *Route Finding Tool* (ESRI Network Analyst extension, v. 10.4.1) by routing the shortest path between random points (i.e. routing from A to B) and ensuring that the calculated shorted path used a mixture of road and path links (i.e. routed across the two datasets).

6.1.2.2 Routing analysis using ArcGIS Network Analyst Service Area tool

In ArcMap, I used the Network Analyst *Service Area* tool to calculate network distance buffers for a subset of UK Biobank address locations. Though these were successfully routed, the process was computationally expensive. The Service Area options to create either 'generalised' or 'detailed' service area polygons was suboptimal for walkability and greenspace exposure assessment; 'generalised' network polygons (A in **Figure 21**) covered areas that were not accessible from the road/path network (e.g., were behind buildings) and 'detailed' polygons (B in **Figure 21**) varied in width around the road and path lines in the network. The lines generated via the Service Area creation could be buffered in a second step, however the generation of the lines in the network added a further time costs in the Service Area generation. I deemed the efficiency of the Network Analyst Service Area approach suboptimal for generation of road/path network buffers for UK Biobank Greater London addresses (n ~58,000), which echoed sentiments of (Stockton et al., 2016b), who conducted Network Analysis on a subset of the total Whitehall sample, and deemed network analysis on the cohort too computationally expensive to further pursue. In summary, ArcGIS network analysis was inefficient for large data and did not produce a standardised width of polygon surrounding the road/path network via the *Service Area* tool.

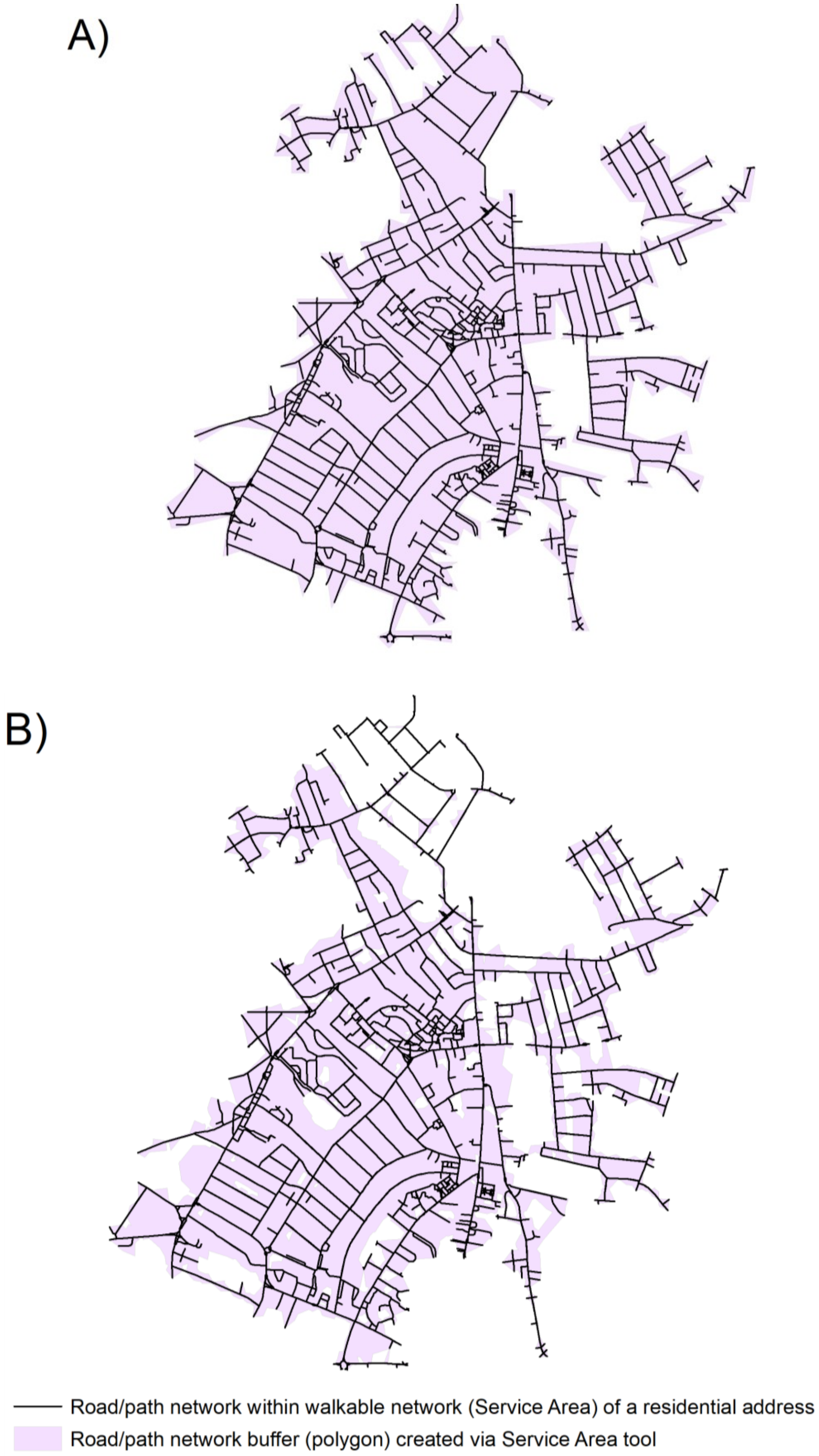

A)
B)
Road/path network within walkable network (Service Area) of a residential address
Road/path network buffer (polygon) created via Service Area tool

Figure 21. Roads and paths (lines) within a walkable network of an example residential address shown with ArcGIS Network Analyst Service Area tool polygons generated using options: a) 'generalised' network polygon and b) 'detailed' network polygon.

6.1.2.3 Network analysis using pgRouting and PostgreSQL

To address slow run time when routing networks via ArcMap Network Analyst, I opted to run the routing assessment in pgRouting, an extension of the open source geospatial database management system PostgreSQL/PostGIS. My motivation for transferring to the PostgreSQL platform was primarily driven by spatial analysis efficiency for large datasets. Reproducibility of analysis – via code – was another advantage. Using pgRouting, I created road and cycle/footpath network distance buffers by tracing a road/path network a given distance from UK Biobank residential address locations (1000 m road/path length), and adding a buffer (50 m width) around the traced road/path network to create a network polygon buffer. In agreement with other studies on walkability and network attributes (Frank et al., 2017, Oliver et al., 2007), I selected a width of 50m (25m either side of the traced network line) surrounding the network, which was sufficient to capture walkable network attributes (e.g., vegetation cover). In the following sections I provide a technical description of how this was achieved.

6.1.2.4 Creating a routable network graph of roads and paths in pgRouting

In order to route the ITN and UPN in pgRouting, I added the geometry (shapefile of line features) from the previous (ArcGIS) network integration step via the uploader, *PostGIS 2.0 DBF and Shapefile Loader Exporter* (EPSG for British National Grid: 27700). This created a single table containing the integrated network, which is comparable to an ArcGIS Attribute Table, with a single *geometry* column. I adapted code from a combination of online, open source resources that provided guidance for preparing ITN data for routed network analysis in pgRouting. These resources included: 1) Building ITN for pgRouting by Ross McDonald (https://mixedbredie.github.io/pgrouting-workshop/) and pgRouting by Anita Graser (https://anitagraser.com/?s=pgrouting). I also used *PgRouting: A Practical Guide* (Obe et al., 2017). In brief, the steps I used to build the network included: adding indices to link sources and targets, adding a distance based cost (and reverse cost) to each link, setting a single average speed for all links (4.83 kmph or 3 mph), building the network with the inbuilt function *pgr_createTopology*, and analysing for errors with *pgr_analyzegraph*. Checks were made by using the function *pgr_dijkstra* to route from A to B (random nodes in the network) and visualizing the results. If the network was incorrectly routed due to unexpected gaps between links, routes would (visibly) deviate from the shortest route. OS geometries are highly accurate so unexpected gaps were not present in this dataset.

6.1.2.5 Routing using pgRouting

I used the function *pgr_withPointsDD* to capture all vertices (nodes) in the line network within a set distance of each residential address (≤ 1000m). The output of this function reported all unique identifiers of the road and path links (edges) leading to all vertices within 1000 m. I linked the geometry of the edge links to the output table via the unique identifier. This step produced the black line network shown in **Figure 22**.

Routing via the *pgr_withPointsDD* function begins at the nearest point on the route network to the residential address (i.e. fractional start edges are permitted) by creating a temporary node table on the routed network. This is an improvement on previous functions whereby only the nearest network vertex could be used as a start point (https://docs.pgrouting.org/2.2/en/src/withPoints/doc/pgr_withPointsDD.html).

6.1.2.6 Improving accuracy of transport network length with fractional edges

I further improved the accuracy of the transport network by extending it to exactly 1000 m in length by adding fractional edges/lines to the terminal vertex of the routed network. This addition was necessary as the in-built function reports vertices under 1000 m distance along the network, and in some instances these can be significantly less than 1000 m distance. For example, if a vertex falls at 750 m along the transport network, and the next vertex along the network is at 1300 m, the inbuilt *pgr_withPointsDD* algorithm will report the 750 m vertex only (1300 m > 1000 m). To modify the output, I used the unique identifier of the terminal vertex (e.g., the node located at 750 m along network), which was reported in the output of *pgr_withPointsDD*. This identifier was associated with road links/edges that it touched, that is, the vertices/node was either the start or the target node of all edges that they touched in the routed ITN/UPN master table. I adapted the *pgr_withPointsDD* function so that it reported all start nodes and all target nodes associated with the terminal node. In effect, this allowed the extra road link/edge (e.g., the 550 m length edge running from 750 m to 1300 m) to be identified and joined to output (see extra terminal links from source (red) and target (blue) nodes in Figure 22).

The result was a road/path link network, which in some places was significantly over 1000 m (e.g. 1300 m). Therefore I used the aggregated cost column in the *pgr_withPointsDD* output, which reported the cost (in meters) to the terminal vertex (e.g., 750 m), to calculate the required length along the additional road/path link required (e.g., 1000 m – 750 m = 250 m). I transformed the length to a fraction of the corresponding additional road/path link edge (e.g., 250/550 = 0.45), and used the function ST_LineSubstring to split the link at precisely 1000 m. This was applied across all terminal nodes in a two-step process (first joining and splitting edges from the source nodes, then the target nodes) before merging and dissolving the output.

To achieve precise 1000 m network distance an extra step was required for terminal target nodes. As each road/path link had a routed directionality running from source node to target node (arbitrarily assigned in the ITN UPN network routing stage), the fraction (e.g., 0.45) to split required reversing from target nodes, and therefore I inversed target node fractions (e.g., 1 – 0.45 = 0.55) before using the function *ST_LineSubstring* on the corresponding additional road/path link. Some road/path links had a fraction of 1 or over, indicating that both the source and the target node were within 1000 m network distance of the address location, I retained these links to create the network buffers (shown in grey in Figure 22).

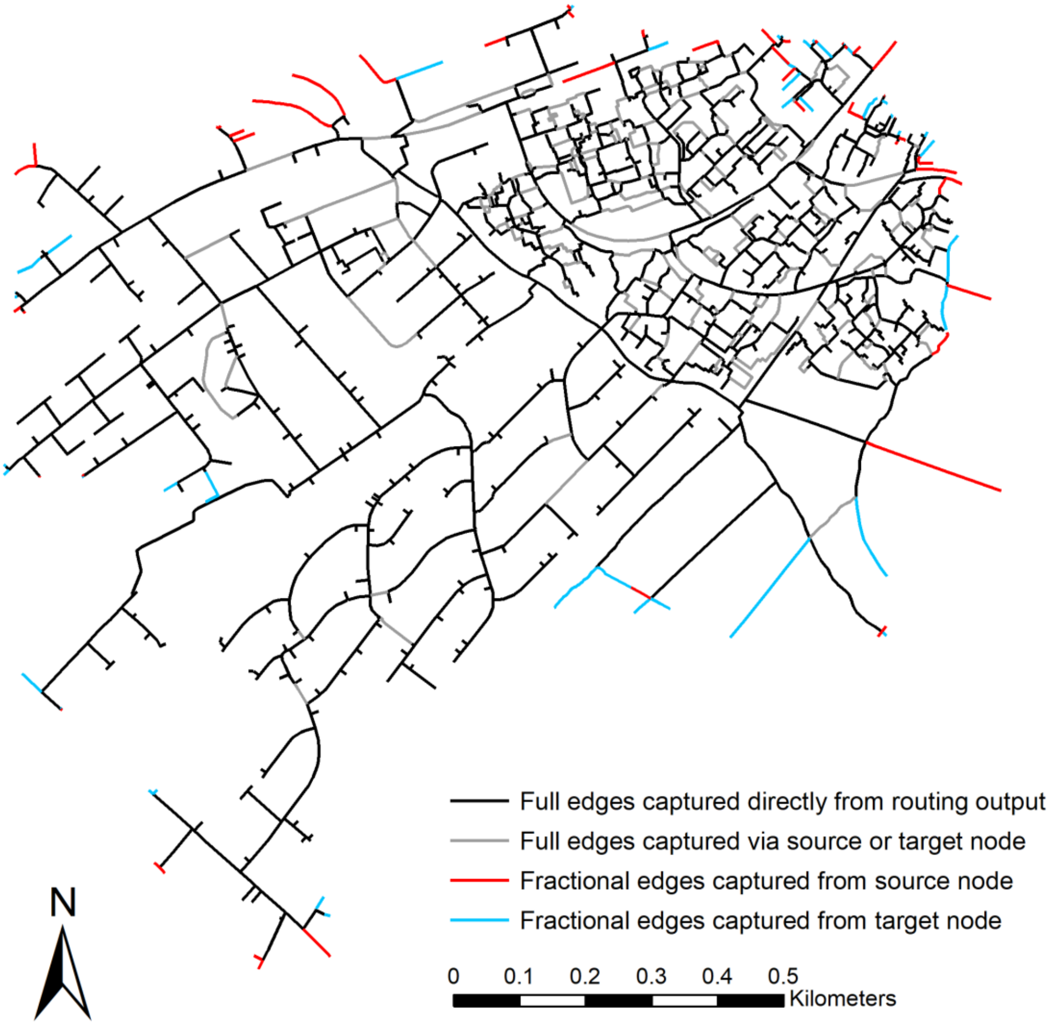

Full edges captured directly from routing output
Full edges captured via source or target node
Fractional edges captured from source node
Fractional edges captured from target node
0 0.1 0.2 0.3 0.4 0.5
Kilometers
N

Figure 22. An integrated road/path transport network showing full edges captured directly from output from function *pgr_withPointsDD* in pgRouting, plus full edges and fractional edges associated with output from the modified version of *pgr_withPointsDD* that I developed in this project. The *pgr_withPointsDD* modification provides a precise 1000 m road/path network from each UK Biobank residential address.

6.1.2.7 *Buffering the line based network*

I found the inbuilt function to create service areas in pgRouting – *alphashape* – efficient, yet unspecific in buffering the transport network as, similar to in ArcGIS, the width of the buffer was variable surrounding the network (see **Figure 20** for a visualisation of *alphashape* output). I therefore scripted SQL to efficiently create walkable network buffers of specific width (25 m each side of the line) for standardised walkability assessment across addresses. **Figure 23** shows a network with a 50 m (sausage) buffer traced around all walkable routes within a ~15 minute walk from the address point.

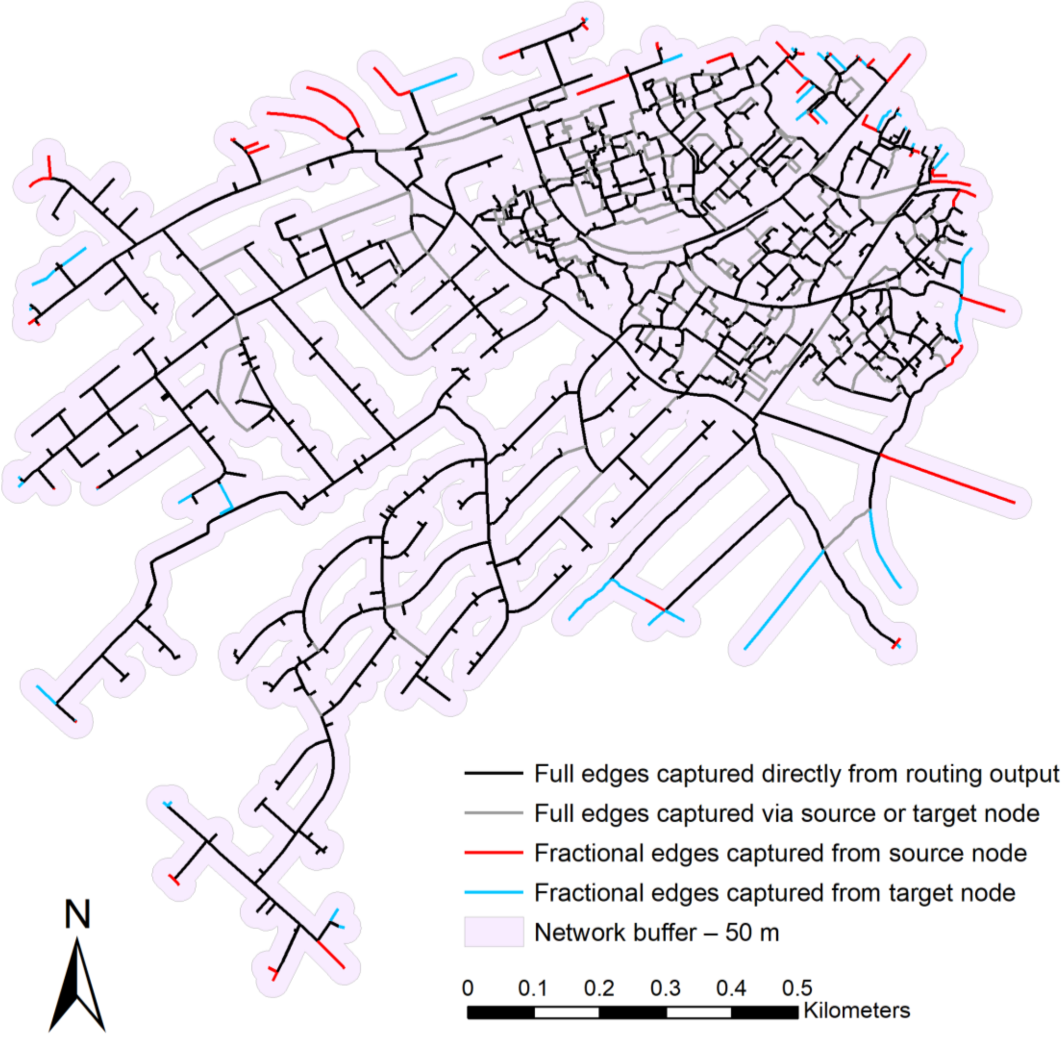

Full edges captured directly from routing output
Full edges captured via source or target node
Fractional edges captured from source node
Fractional edges captured from target node
Network buffer – 50 m
N
0
0.1
0.2
0.3
0.4
0.5
Kilometers

Figure 23. An accurate 1000 m length integrated road/path transport network with a 50 m buffer from a dummy address point in London.

6.2 Walkability exposure assessment

Physical activity is likely to be influenced by other environmental features aside from vegetation, I therefore created a walkability score for each participant. The walkability z-score combined three commonly used metrics to assess walkability: population density (derived from summed postcode headcounts (postcode points) located within the line based network buffer); three-way intersection density (derived from the vertices/nodes of the underlying road/path network); destination (business) counts (derived from OS Points of Interest in the line based network buffer). **Figure 24** shows POI overlaid on an example 1000m network buffer.

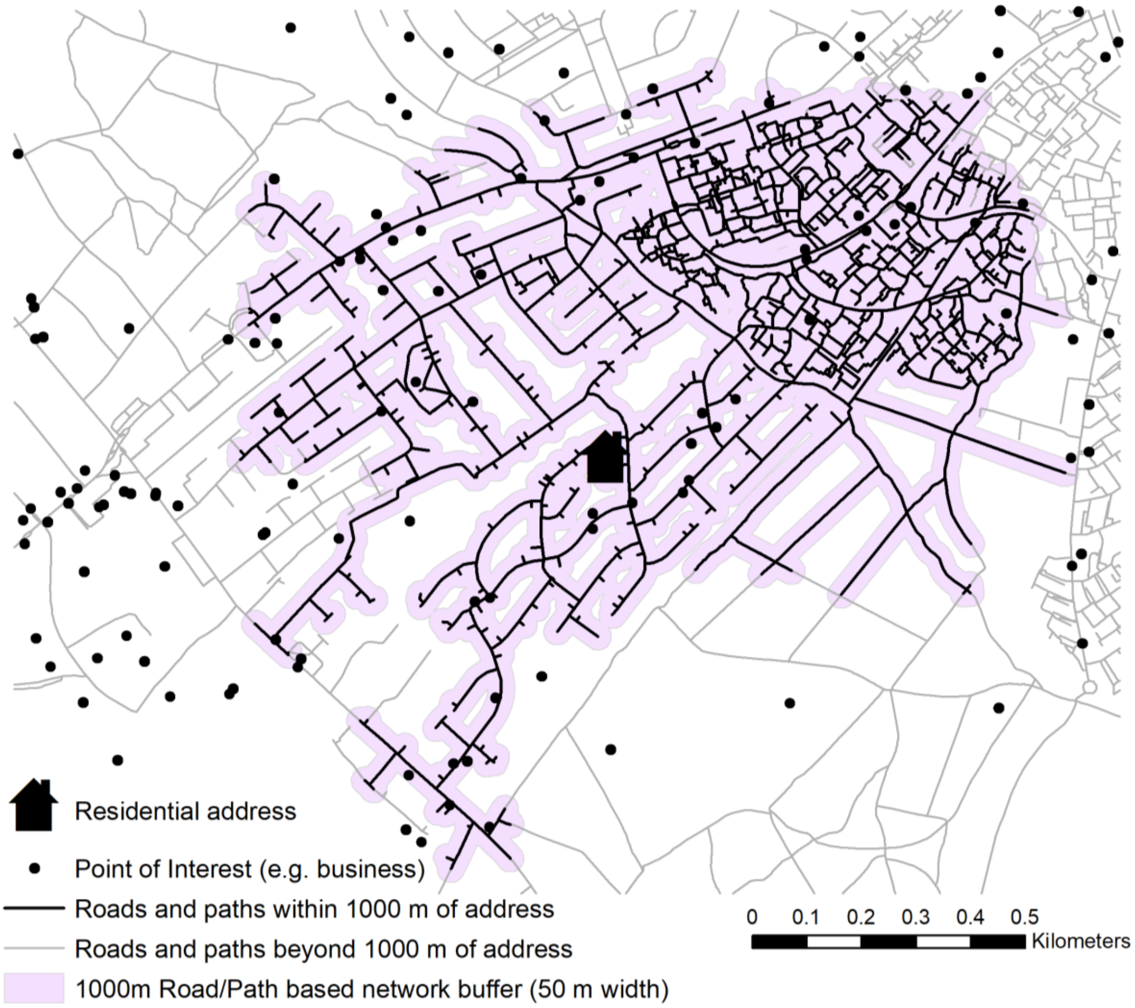

Residential address
Point of Interest (e.g. business)
Roads and paths within 1000 m of address
Roads and paths beyond 1000 m of address
1000m Road/Path based network buffer (50 m width)
0 0.1 0.2 0.3 0.4 0.5
Kilometers

Figure 24. Road/path network within 1000 m of (dummy) cohort participant residential address (black lines), with 50 m width buffer surrounding the network (purple), showing Ordnance Survey Points of Interest data such as businesses, services and facilities (black points) within and beyond the extent of the walkable network buffer.

I used the postcode headcounts for population density estimation. I used all junction types – e.g., road-to-road, road-to-footpath and footpath-to-footpath – to represent walkable network connectivity. To capture true network junctions, as oppose to line segment breaks, only junctions that connected a minimum of three line segments were considered a true junction. To derive this, I used a freely-available geoprocessing tool from Esri by L. Beale, *Line and Junction Connectivity* (https://www.arcgis.com/home/item.html?id=3fa41b1f8b764879be8f21b4e7ffbabd). As previously discussed, points of Interest (POI) counts were included to represent potential destinations – as a proxy for land-use mix – and included retail, facilities and services that an individual may walk to. This included newsagents, bus stops, underground stations, sports facilities, restaurants, banks, libraries, etc. Intersections of datasets corresponding to walkability metrics with transport network buffers was efficient, as: a) data were geographically indexed and queried in PostGIS; and b) the three underlying datasets were point data, as oppose to polygon data (e.g. land cover), which simplified the intersect process in PostGIS. I inserted the counts of intersected points from the three datasets into in a single table and exported in csv format. I used R statistical software to calculate z-score from the three walkability metrics (population density, street connectivity, and destination density within the walkable network buffer). I divided the counts (population, junctions, and destinations) by the network buffer area, to provide a density estimate. I summed (equal weighting) the three z-scores from the metrics to calculate a single walkability z-score. Z-scoring was necessary to standardize across data outputs that would otherwise be incomparable across the three datasets. Using PostGIS, I intersected the vegetation cover data, categorised as tree canopy (≥2.5 m height) or ground cover (<2.5 m height), which is described in Section 4.1, with the road/path network distance buffer for each participant's address. This was provided for integration to UK Biobank as percentage cover of the total network buffer area. To ensure that the vegetation data fully covered the road/path network distance buffers around each participant's address geocode, I included only addresses located inwards (-1000 m) of the Greater London Vegetation Layer boundary in walkability exposure assessment. Two example buffers, with intersected vegetation, are shown in **Figure 25**.

To assess the difference between a ‘basic’ walkability score (three component: population, connectivity and businesses) and a ‘green’ walkability score in association with PA in UK Biobank, I converted the percentage cover vegetation in the walkable network buffer surrounding each UK Biobank address to a z-score, and summed this z-score with the three-component z-score to create 4-component ‘green’ walkability z-score.

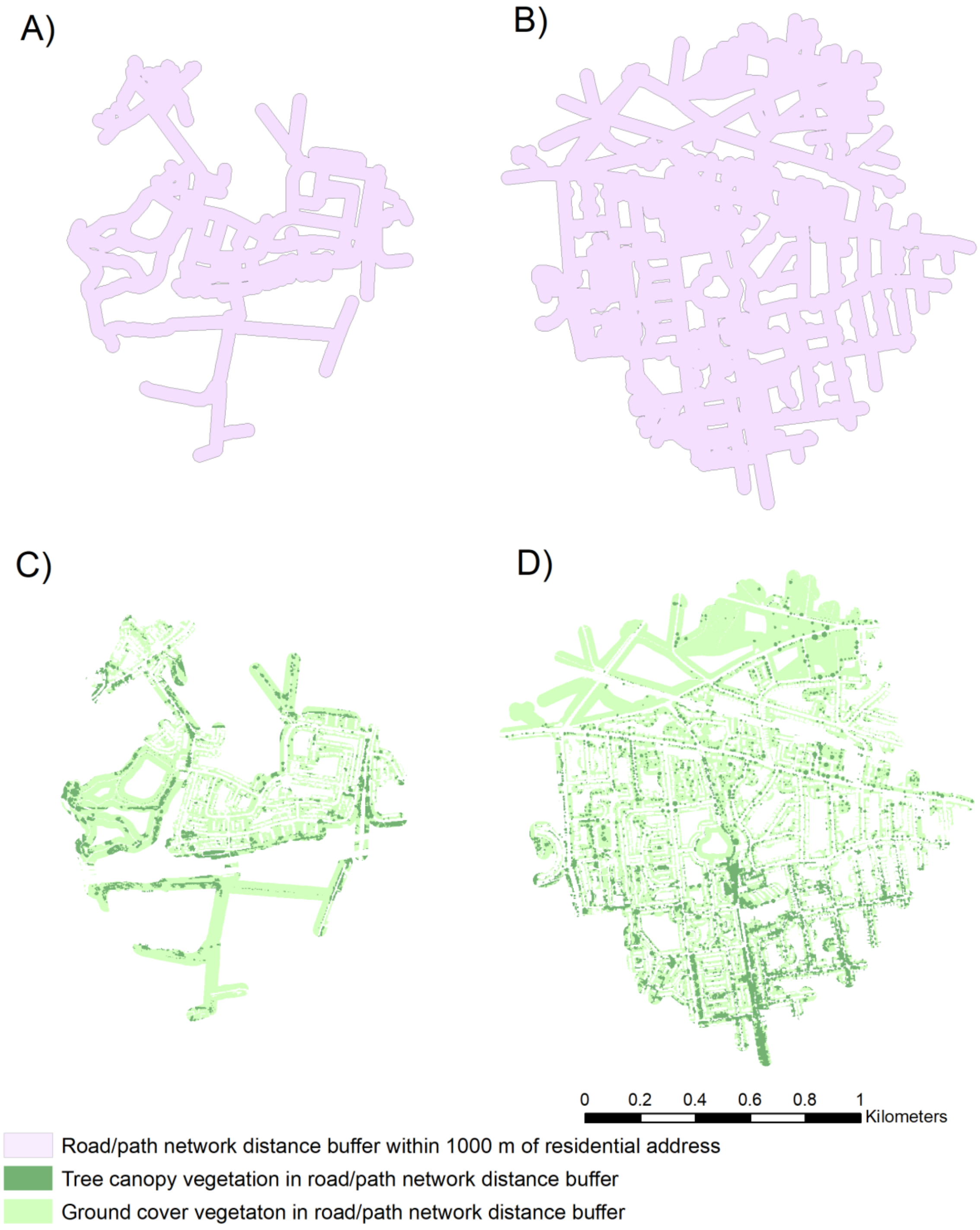

A)
B)
C)
D)
0
0.2
0.4
0.6
0.8
1
Kilometers
Road/path network distance buffer within 1000 m of residential address
Tree canopy vegetation in road/path network distance buffer
Ground cover vegetaton in road/path network distance buffer

Figure 25. A sparse (A and C) and a dense (B and D) road/path network distance buffer within 1000 m of two (example) residential addresses in Greater London. Vegetation cover produced by the GeoInformation Group (2.5 m x 2.5 m resolution), categorized as ground cover (<2.5 m height) and tree canopy (≥2.5 m height), is shown within the two road/path network distance buffers (C and D).

6.3 Walkability and physical activity in UK Biobank London

The two-stage process of assigning exposures to UK Biobank address coordinates before obtaining UK Biobank health and covariate data with integrated environmental exposures for statistical analysis did not allow for the testing of the predictive capacity of walkability indexes before exposures were integrated into UK Biobank. Instead, I validated the input variables (population density, street connectivity, destination density and vegetation cover) used in the assigned walkability index in UK Biobank London. I explored the strength of these exposures on physical activity and transport modal choice in UK Biobank using statistical models, which I adjusted for potential confounders. I provide the methods, findings and discussion of the walkability assessment validation below.

6.3.1 *Statistical methods*

To assess if the walkability score (independent variable) impacted modal choice for commuting purposes (categorical dependent variable: transport type for commuting to workplace), I fit multinomial log-linear models, with car/motor vehicle commute as the reference category. I exponentiated the coefficients from the model to obtain odds (risk ratio) of modal choice (compared to reference category odds of 1) for a one unit increase in the walkability score, and calculated 95% confidence intervals. I adjusted the model for: age at baseline assessment, sex, average household income before tax, neighbourhood deprivation level and distance between home and work place. I repeated this methodology with transport mode for non-commute purposes (dependant outcome), though I removed the variable *distance between home and work place,* as this was unnecessary for confounder adjustment in a non-commute context. To improve interpretability of findings, I conducted analyses using quintiles of walkability (independent variable), with quintile 1 (least walkable 20% of UK Biobank addresses) used as the reference category.

I used multinomial log-linear models to assess associations of UK Biobank participant weekly average physical activity levels recorded by the short form International Physical Activity Questionnaire (IPAQ) at baseline, This outcome was categorised by UK Biobank as low, medium and high weekly physical activity following IPAQ protocol and categories are available to all researchers via the showcase (https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/ipaq_analysis.pdf). In brief, UK Biobank developed an algorithm to calculate categories of IPAQ physical activity, which included total volume and the number of days/sessions of physical activity, in order to capture regular participation in physical activity across all domains (leisure, domestic, work-related and transport physical activity). To do this, the algorithm calculated metabolic equivalent of task (MET)-minutes based on each participant's IPAQ responses, whereby MET values were based on the metabolic intensity of an activity compared to sitting quietly (e.g., metabolising 3.5 ml of oxygen per kilogram body weight per minute). Following IPAQ convention, this was represented as a ratio for sitting versus walking (3.3 METs), moderate physical activity (4.0 METs) and vigorous physical activity (8.0 METs). The algorithm multiplied MET-minutes by the number of days an activity was conducted per week. The high physical activity category corresponded to either: a) vigorous-intensity physical activity on at least 3 days, achieving a minimum total physical activity of at least 1500 MET-minutes/week; or b) 7 or more days of any combination of walking, moderate-intensity or vigorous-intensity physical activity, achieving a minimum total physical activity of at least 3000 MET-minutes/week. The moderate physical activity category corresponded to either a) 3 or more days of vigorous-intensity physical activity of at least 20 minutes per day; or

b) 5 or more days of moderate-intensity physical activity and/or walking of at least 30 minutes per day; or c) 5 or more days of any combination of walking, moderate-intensity or vigorous intensity physical activity, achieving a minimum total physical activity of at least 600 MET-minutes/week. The low physical activity category corresponded to any level of physical activity below the moderate category threshold. UK Biobank metadata stated that category thresholds were based on pedometer studies (Tudor-Locke and Bassett, 2004).

I fitted a generalized linear model (GLM) in R using base package, *stats*. I specified the GLM family binomial (link = logit) to conduct logistic regression, and examine if likelihood of achieving UK recommended physical activity levels (binary variable: Yes, No) was associated with surrounding walkability. I repeated this with the 4-componet walkability score with vegetation included (hereafter referred to as the *Green walkability score*). UK Biobank provided the binary outcome variable based on responses to the IPAQ short form questionnaire. In alignment with models described above, I adjusted the model for: age at baseline assessment, sex, average household income before tax and neighbourhood deprivation level. I also used quintiles of walkability exposure to improve interpretability of results.

6.3.2 *Findings and discussion*

Higher walkability score surrounding the residential address (1000 m network buffer) was associated with higher odds of walking, cycling and taking public transport compared to driving a motor vehicle to work in UK Biobank London (see **Table 8**). Moreover, for a 1 unit increase in walkability score, the odds of walking and cycling to work were higher compared to car/motor vehicle commute in the green walkability score, compared to the walkability score without vegetation. For example, walking for transport (non-commute) was 30% more likely than driving a car for a 1 unit increase in the walkability score (OD 1.30, 95% CI 1.29-1.32), and was 34% more likely than driving a car for a 1 unit increase in the green walkability score (OD 1.34, 95% CI 1.32-1.36). In general, the alternative classification of walkability, which included vegetation, showed a stronger association with walking as a transport modal choice. For non-commute transport, adjusted odds were increased by a similar order of magnitude as for commute purposes by walkability, and by green walkability.

In UK Biobank, 38% of addresses were reclassified into a different quintile when vegetation was added to the walkability score. When assessing walkability and green walkability by quintiles, a positive trend across quintiles of increasing cycling walking and taking public transport versus taking the car/motor vehicle for commute and non-commute transport was shown in more walkable quintiles (compared to the least walkable quintile) (see **Table 9** and **Table 10**). For walking transport, the quintile regression showed the difference between the least walkable and most walkable quintile was larger for the standard walkability score (e.g., non-commute transport walking OR 6.37, (95% CI 5.85, 6.94)) than for the green walkability score (e.g., non-commute transport walking OD 4.46 (95% CI 4.12, 4.84)).

Table 8. Odds ratios and 95% confidence intervals (CIs) for self-reported commute transportation mode and non-commute transportation mode in UK Biobank London associated with surrounding walkability (i.e. population density, street connectivity and destination density) and Green walkability (i.e. population density, street connectivity, destination density, and vegetation in the network buffer) scores. Multinomial log-linear models were adjusted for: age at baseline assessment, sex, average household income before tax, neighbourhood deprivation level and distance between home and work place (commute only). Driving a car/motor vehicle was used as the reference category.

	Walkability (adjusted model) **odds ratios (95% CI)**	**Green Walkability** (adjusted model) **odds ratios (95% CI)**
Commute mode (ref = car/motor vehicle)		
Cycle	1.30 (1.26, 1.33)	1.37 (1.32, 1.42)
Public transport	1.23 (1.21, 1.25)	1.24 (1.21, 1.26)
Walk	1.31 (1.29, 1.34)	1.33 (1.31, 1.36)
Non-commute transport (ref = car/motor vehicle)		
Cycle	1.29 (1.25, 1.32)	1.36 (1.32, 1.41)
Public transport	1.22 (1.21, 1.24)	1.24 (1.22, 1.27)
Walk	1.30 (1.29, 1.32)	1.34 (1.32, 1.36)

Table 9. Odds (risk ratio) and 95% confidence intervals (CIs) of active commute transport mode and active non-commute transport mode compared to using car/motor vehicle for commute and non-commute transport shown by quintiles of walkability (Q1 being the lowest 20% of walkability scores; Q5 the highest 20%), for walkability score. Models were adjusted for: age, sex, average household income before tax, neighbourhood deprivation level and distance between home and work place (commute only).

	Walkability quintiles (adjusted model) **odds ratios (95% CI)**				
	Q1	**Q2**	**Q3**	**Q4**	**Q5**
Commute mode (ref = car/motor vehicle)					
Cycle	1	1.58 (1.25, 1.99)	1.92 (1.52, 2.42)	2.93 (2.33, 3.69)	6.69 (5.29, 8.46)
Public transport	1	1.35 (1.22, 1.48)	1.79 (1.63, 1.97)	2.74 (2.48, 3.02)	4.03 (3.61, 4.50)
Walk	1	1.28 (1.14, 1.44)	1.70 (1.52, 1.91)	2.52 (2.24, 2.83)	5.06 (4.46, 5.74)
Non-commute transport (ref = car/motor vehicle)					
Cycle	1	1.43 (1.13, 1.80)	2.09 (1.68, 2.62)	2.66 (2.13, 3.33)	6.54 (5.21, 8.20)
Public transport	1	1.51 (1.37, 1.67)	2.00 (1.81, 2.20)	2.67 (2.42, 2.95)	4.09 (3.68, 4.55)
Walk	1	1.66 (1.53, 1.79)	2.47 (2.29, 2.67)	3.39 (3.13, 3.67)	6.37 (5.85, 6.94)

Table 10. Odds (risk ratio) and 95% confidence intervals (CIs) of active commute transport mode and active non-commute transport mode compared to using car/motor vehicle for commute and non-commute transport shown by quintiles of walkability (Q1 being the lowest 20% of walkability scores; Q5 the highest 20%), for green walkability score. Models were adjusted for: age, sex, average household income before tax, neighbourhood deprivation level and distance between home and work place (commute only).

	Green walkability quintiles (adjusted model) odds ratios (95% CI)				
	Q1	**Q2**	**Q3**	**Q4**	**Q5**
Commute mode (ref = car/motor vehicle)					
Cycle	1	1.63 (1.28, 2.08)	2.14 (1.69, 2.71)	2.99 (2.37, 3.76)	5.77 (4.57, 7.27)
Public transport	1	1.25 (1.14, 1.38)	1.65 (1.50, 1.81)	2.09 (1.90, 2.29)	2.92 (2.64, 3.24)
Walk	1	1.25 (1.11, 1.40)	1.65 (1.47, 1.85)	2.20 (1.97, 2.47)	3.64 (3.23, 4.11)
Non-commute transport (ref = car/motor vehicle)					
Cycle	1	2.13 (1.67, 2.72)	2.42 (1.90, 3.09)	3.30 (2.60, 4.18)	6.77 (5.34, 8.59)
Public transport	1	1.47 (1.33, 1.62)	1.99 (1.81, 2.19)	2.33 (2.12, 2.56)	3.28 (2.96, 3.63)
Walk	1	1.54 (1.43, 1.66)	2.18 (2.02, 2.35)	2.91 (2.70, 3.13)	4.46 (4.12, 4.84)

Walkability and green walkability scores were associated in the expected direction of effect with IPAQ category (as categorised by UK Biobank), that is, the higher the walkability score for an address, irrespective of walkability scale components, the higher the likelihood of greater levels of physical activity (medium or high IPAQ categories). The medium category showed a slightly higher odds ratio than the high category, when compared to the low category (see **Table 11**). Odds ratios across quintiles of walkability and green walkability can be found in

Table 12 and **Table 13**, respectively.

Table 11. Odds ratios and 95% confidence intervals (CIs) for physical activity category derived from IPAQ responses in UK Biobank London participants associated with surrounding walkability and green walkability scores at residential addresses. Multinomial log-linear models were adjusted for: age at baseline assessment, sex, average household income before tax, neighbourhood deprivation level. Low activity level was used as the reference category.

	Walkability quintiles (adjusted model) odds ratios (95% CI)				
	Q1	**Q2**	**Q3**	**Q4**	**Q5**
IPAQ physical activity category (ref = Low)					
Moderate	1	1.13 (1.03, 1.23)	1.20 (1.10, 1.31)	1.28 (1.17, 1.40)	1.52 (1.37, 1.68)
High	1	1.15 (1.05, 1.26)	1.20 (1.10, 1.32)	1.24 (1.13, 1.36)	1.49 (1.34, 1.65)

Table 12. Odds ratios and 95% confidence intervals (CIs) for physical activity category derived from IPAQ responses in UK Biobank London participants associated with surrounding walkability score at residential addresses, shown by quintiles of walkability (Q1 being the lowest 20% of walkability scores; Q5 the highest 20%). Models was adjusted for: age at baseline assessment, sex, average household income before tax, neighbourhood deprivation level. Low activity level was used as the reference category.

IPAQ category	**Walkability (adjusted model) odds (95% CI)**	**Green walkability (adjusted model) odds (95% CI)**
Low (reference)	1	1
Medium	1.06 (1.04, 1.08)	1.07 (1.05, 1.09)
High	1.05 (1.03, 1.07)	1.06 (1.04, 1.08)

Table 13. Odds ratios and 95% confidence intervals (CIs) for physical activity category derived from IPAQ responses in UK Biobank London participants associated with surrounding green walkability score at residential addresses, shown by quintiles of walkability (Q1 being the lowest 20% of walkability scores; Q5 the highest 20%). Model was adjusted for: age at baseline assessment, sex, average household income before tax, neighbourhood deprivation level. Low activity level was used as the reference category.

	Green walkability quintiles (adjusted model) odds ratios (95% CI)				
	Q1	**Q2**	**Q3**	**Q4**	**Q5**
IPAQ physical activity category (ref = Low)					

Moderate	1	1.12 (1.02, 1.22)	1.13 (1.03, 1.23)	1.18 (1.08, 1.29)	1.42 (1.28, 1.57)
High	1	1.10 (1.00, 1.20)	1.16 (1.06, 1.26)	1.22 (1.11, 1.33)	1.37 (1.24, 1.51)

Likelihood of achieving recommended physical activity guidelines via a combination of walking, moderate and vigorous activity was significantly associated with walkability and green walkability score, however achieving physical activity recommendations via moderate and vigorous activity (excluding walking activity) was not associated with surrounding walkability, and green walkability (**Table 14**).

Table 14. Odds ratios and 95% confidence intervals (CIs) for achieving UK weekly physical activity guidelines via moderate and vigorous physical activity, or via walking, moderate and vigorous physical activity in UK Biobank London participants associated with surrounding walkability and green walkability scores at residential address. Binomial models were adjusted for: age at baseline assessment, sex, average household income before tax and neighbourhood deprivation level. Not achieving recommendations and quintile 1 level of walkability were used as reference categories.

	Walkability (adjusted model) odds (95% CI)	**Green walkability (adjusted model) odds (95% CI)**
Achieved UK physical activity recommendations…		
…via moderate and vigorous activity combined	1 (1.00, 1.01)	1.01 (1.00, 1.02)
…via walking, moderate and vigorous activity combined	1.05 (1.04,1.07)	1.06 (1.05, 1.08)

Table 15. Odds ratios and 95% confidence intervals (CIs) for achieving UK weekly physical activity guidelines via walking, moderate and vigorous physical activity in UK Biobank London participants associated with surrounding walkability score at residential address, shown by quintiles of walkability (Q1 being the lowest 20% of walkability scores; Q5 the highest 20%). Binomial model was adjusted for: age at baseline assessment, sex, average household income before tax and neighbourhood deprivation level. Not achieving recommendations and quintile 1 level of walkability were used as reference categories.

	Walkability quintiles (adjusted model) odds ratios (95% CI)				
	Q1	**Q2**	**Q3**	**Q4**	**Q5**
Achieved UK physical activity recommendations via moderate and vigorous activity					
Achieved*	1	1.03 (0.97, 1.09)	1.03 (0.97, 1.09)	1.02 (0.96, 1.08)	1.07 (1.00, 1.14)
Achieved UK physical activity recommendations via walking, moderate, vigorous activity					
Achieved*	1	1.09 (1.01, 1.18)	1.17 (1.08, 1.26)	1.25 (1.15, 1.35)	1.46 (1.34, 1.60)

*Reference category: did not achieve UK physical activity recommendations.

Table 16. Odds ratios and 95% confidence intervals (CIs) for achieving UK weekly physical activity guidelines via walking, moderate and vigorous physical activity in UK Biobank London participants associated with surrounding green walkability scores at residential address, shown by quintiles of walkability (Q1 being the lowest 20% of walkability scores; Q5 the highest 20%). Binomial model was adjusted for: age at baseline assessment, sex, average household income before tax and neighbourhood deprivation level. Not achieving recommendations and quintile 1 level of walkability were used as reference categories.

	Green walkability quintiles (adjusted model) odds ratios (95% CI)				
	Q1	**Q2**	**Q3**	**Q4**	**Q5**
Achieved UK physical activity recommendations via moderate and vigorous activity					
	1	1.00 (0.94, 1.08)	1.03 (0.97, 1.09)	1.03 (0.98, 1.10)	1.06 (1.00, 1.13)
Achieved UK physical activity recommendations (walking, moderate and vigorous activity)					
Achieved*	1	1.06 (0.98, 1.15)	1.09 (1.01, 1.18)	1.17 (1.08, 1.26)	1.36 (1.25, 1.49)

*Reference category: did not achieve UK physical activity recommendations.

Walkability showed a positive association with healthy modal choice for commute transport and non-commute transport after adjustment for confounding (e.g., age, sex, household income after tax, neighbourhood deprivation level, and distance to workplace from residence). For the commute, walking and cycling were both 37% more likely than taking the car per unit increase in walkability score, and for non-commute transport they were 32% and 33% higher likelihood, respectively. Studies have shown the benefits of healthy modal choices (walking and cycling) on cardiovascular health outcomes. Adding vegetation cover in the network buffer to the walkability score increased the odds of cycling and walking during the commute (7% and 1%, respectively) and for non-commute related transport (8% and 3%).

UK Biobank London participants living at residences in more walkable neighbourhoods (1000 m walking distance) had higher likelihood of medium or high than low physical activity category classification from their IPAQ responses. Further, when dichotomised into those who do achieve and do not achieve UK physical activity recommendations, participants residing at addresses in more walkable neighbourhoods (1000 m network) had a higher likelihood of achieving recommendations than those in less walkable neighbourhoods. However, achieving UK guidelines was only associated with network walkability when walking (plus moderate and vigorous intensity) physical activity were summed in assessments. In other words, when only moderate and vigorous physical activity were accounted for (without summing walking MET), no relationship with neighbourhood walkability was shown. In light of environmental exposures that are accounted for in walkability assessment (e.g., destination density), findings were suggestive that walking was partially driving associations of neighbourhood walkability and physical activity, which is logical in that moderate and vigorous intensity physical activity might be driven by specific destinations, for example, sport centres/facilities.

Use of the green walkability in analysis attenuated the difference in odds between the least and most walkable quintile compared to the standard walkability score for all physical activity outcomes. That is, the addition of vegetation cover into the walkability score reclassified walkability of UK Biobank participants' addresses (38% change of quintile with inclusion). This resulted in attenuation of the difference in odds between most and least walkable areas for: a) cycling, public transport and walking for commute and non-commute transport (versus car/motor vehicle use); b) moderate and high physical activity (versus low physical activity categorisation); and 3) achieving UK physical activity guidelines via walking, moderate and vigorous physical activity (versus not achieving guidelines). Given that defining propensity to walk related to features of the residential neighbourhood is the objective of creating a walkability score. Adding vegetation cover to the score might have enhanced differentiation of addresses in relation to walking behaviours. This was not shown between most and least walkable quintiles, however the trend across quintiles (more, walkable, more reported walking) remained stable. Potentially, adding vegetation cover reduced differentiation of walking behaviour in the most and least walkable quintiles of the sample as suburban (green) neighbourhoods were reclassified as more walkable, and urban (less green) areas were reclassified as less walkable, which might not reflect modal choices or walking behaviour in the most and least walkable areas.

While I integrated all score components with equal weighting, other studies have ranked walkability score components using principal component analysis to assign contribution weights (Creatore et al., 2016). Optimising the walkability score by re-weighting the vegetation component of the score is possible in UK Biobank in the future as I added the variable *vegetation cover within the 1000 m network* in addition to the walkability and green walkability scores. For example, destination density potentially far outweighs the importance of the greenness of the walkable network in individual modal choice decision-making. Further work on the direct and combined effect of individual walkability components might clarify the magnitude of effect (if any) of vegetation on modal choice and walking physical activity, though my preliminary analyses suggest, irrespective of the inclusion of vegetation in the index, a positive trend across walkability quintiles with walking.

Also, analyses were cross sectional based on UK Biobank baseline, limiting causal interpretation. A study conducted on UK Biobank Stockport (Greater Manchester assessment centre), where repeat assessment data is available, showed modal shift in commute to walking and cycling to be associated with lower BMI and protective cardiovascular effects (Celis-Morales et al., 2017, Flint et al., 2016). Further work on the Stockport sub-cohort would offer the opportunity to assess potential environmental drivers of modal shift.

Associations presented were based on self-reported physical activity levels, which could be further validated in future work by using accelerometry data available in UK Biobank. However, the sample of UK Biobank London participants who had walkability assigned as part of this project and participated in the 7-day accelerometer study (with valid wear time) was less than a fifth (n = 13,023) of those with available self-reported physical activity data. Any assessment would therefore require careful consideration of the loss of power if the self-report and accelerometry data were compared.

6.4 Chapter summary

I developed this approach with the aim of integrating greenspace exposure assessment and walkability assessment to better understand the separate and combined effects of vegetation cover and other built environment components (walkability metrics) on physical activity levels in UK Biobank participants. In this chapter, I provided a method of integrating greenspace into a walkability assessment model. I showed that walkability surrounding UK Biobank London residential addresses was associated with self-reported physical activity levels, commute and non-commute transport mode choice, and achieving UK physical activity recommendations (the latter, only when walking activity was included in the total). Integrating total vegetation cover surrounding the walkable network into the walkability score attenuated associations with all physical activity measures. In conclusion, this chapter demonstrates that, as oppose to assessing greenspace as a unique, isolated construct in relation to physical activity, known built environment correlates of physical activity should not be ignored in assessments of the *physiological pathway*.

7 Interrelation of exposures in UK Biobank London

In this chapter, I explore the interrelation of vegetation cover surrounding UK Biobank London participants' addresses and other built environment exposures (air pollution, traffic noise and walkability). I provide descriptive summaries of the built environment exposures at UK Biobank London addresses that were assessed in this PhD thesis. Further, some greenspace epidemiological studies have shown effect modification by area level deprivation (Mitchell and Popham, 2008b, Yitshak-Sade et al., 2019), and others have stressed the importance of adjusting for covariates linked to personal and neighbourhood deprivation (Markevych et al., 2017, Villeneuve et al., 2012b), I graphically summarise exposure variables stratified by: a) neighbourhood deprivation quintiles; and b) household income. I provide findings of correlations of all exposure variables. I also report other statistical methods that I used to assess linear and non-linear relationships of the exposure variables assessed at UK Biobank London addresses, and provide findings and implications for future epidemiological analysis of linearity of relationships.

7.1 Graphical and statistical methods used to explore interrelation

I tabularised descriptive summaries of the distributions (e.g., mean, standard deviation, etc.) of environmental exposure levels at UK Biobank participants' residential addresses in Greater London. I also show the spread of the distribution using box plots (showing median, two hinges corresponding to the first and third quartiles (the 25th and 75th percentiles) and two whiskers that extend from hinges to the largest/smallest value no further than 1.5 * inter-quartile range (IQR; i.e. distance between the first and third quartiles). Box plots were stratified by a) quintiles of area level deprivation; and b) self-reported average total household income groups using the ggplot2 package (Wickham, 2016) in R (v 3.3.1; R Development Core Team).

Quintiles of deprivation corresponded to national, as oppose to within sample, quintiles of deprivation. Income groups corresponded to a questionnaire item from UK Biobank baseline assessment (i.e. is self-reported). Additionally, I produced 'notched' box plots to visually assess overlap of quintiles of deprivation and of household income (not shown graphically). The notches corresponded to 1.58 * IQR / square root(*n*), which gives a roughly 95% confidence interval for comparing overlap between groups as per the method of McGill et al. (1978). I also produced urban versus rural plots (not shown), though these were uninformative as ~99% of UK Biobank in England is categorised at *Urban (less sparse)*.

To further explore interrelations, I used correlation plots to assess and visualise correlations between exposures. Due to the skew of the built environment exposure data, I used Spearman's rank correlation coefficient estimations to assess the monotonic relationship between exposures (rank based assessment), as oppose to Pearson's correlation (true-value based assessment). I used R package corrplot (Wei and Simko, 2017) to visualise correlations.

Several studies have shown non-linear relationships of environmental exposure variables, I therefore fitted pairwise exposure models with a natural cubic spline (3 degrees of freedom (df)) using vegetation as the independent variable and walkability, air pollution (NO_2) and traffic noise as independent variables. I used cross-validation criterion (R^2 and RMSE) to assess if a non-linear (3 df) relationship of exposure variables was a better predictor of the data than a linear one (i.e. a regression with 1 df). I used a generalised additive model (GAM), employing the Restricted Maximum Likelihood (REML) method, to further explore non-linear relationships between variables that were better explained by natural cubic spline (3 df) fit than a linear fit (1 df). While different methods have their advantages for extrapolating smooth terms, REML is most likely to give a reliable, stable results, and penalizes overfitting more strongly than other methods, such as generalized cross validation (Wood, 2011). I presented the smooth term for the GAM graphically. I also visualised the relationship of the data, with line of best fit corresponding to the model that best explained the relationship (i.e. linear or natural cubic spline smooth). I conducted analysis in R (v. v 3.3.1) using package, mgcv (Wood, 2017) and visualised results using ggplot2 (Wickham, 2016).

7.2 Findings and discussion

7.2.1 *Description of UK Biobank London*

After exclusions of participant address located outside of the modelling domain for any exposure, there remained 58,587 addresses that were assigned all built environment exposures (greenspace, air pollution, traffic noise and walkability). The mean and distribution of the OS MasterMap™ Greenspace and Greater London Authority (GLA) vegetation cover total was similar, indicating agreement in the two datasets (see **Table 17**). GLA ground cover and tree cover averages remained in a similar ratio (~35% cover to ~15% cover) across all circular distance buffer sizes. In the road/path network buffer created for walkability assessment (1000 m length network, with 50 m buffer), ratio of groundcover, but not tree cover, was on average lower than in the circular buffer analyses (ratio: ~30% cover to ~15% cover, respectively). Modelled estimates of NO_2 concentration at UK Biobank addresses was 36.60 μg/m^{3}on average, and generally below the European Limit Value (40 μg/m^3) but reached a maximum of 113.47 μg/m^3. On average, modelled day-time and night-time noise levels exceeded thresholds above which the World Health Organisation (WHO) defines as harmful to human health (e.g., 55 dB for day-time noise). Exposure distributions were skewed for NO_2 and traffic noise exposure estimates.

Table 17. Description of built environment exposure variables assigned to UK Biobank London addresses (*n* = 58,587). OS = Ordnance Survey; GLA = Greater London Authority vegetation data; Lday, Leve, Lnight and LAeq,16h and are A-weighted averages of hourly sound levels for time periods 07:00–18:00, 19:00–22:00, 23:00–06:00, and 07:00–23:00, respectively. Lden is equivalent to LAeq over 24 hours, with added penalty weights of 5dB for noise in the evening (19:00–23:00), and 10dB for noise at night (23:00–07:00).

Exposure	Buffer size	Variable	Mean	Standard deviation	Range	Skew
Greenspace (% cover)	100 m	OS MasterMap™ Greenspace	50.17	15.61	97.32	-0.35
		GLA total	48.43	15.33	93.17	-0.38
		GLA ground cover	33.60	12.43	86.16	-0.24
		GLA tree cover	14.87	8.78	63.06	0.95
	500 m	OS MasterMap™ Greenspace	52.55	14.46	91.87	-0.32
		GLA total	50.12	14.42	93.78	-0.25
		GLA ground cover	34.80	11.92	81.84	-0.23
		GLA tree cover	15.36	6.97	61.98	0.99
	1000 m	OS MasterMap™ Greenspace	53.73	13.41	86.30	-0.45
		GLA total	51.44	13.84	90.23	-0.32
		GLA ground cover	35.89	11.32	74.21	-0.43
		GLA tree cover	15.59	6.27	47.38	0.88
	1500 m	OS MasterMap™ Greenspace	54.38	12.81	79.45	-0.55
		GLA total	52.17	13.48	81.16	-0.38
		GLA ground cover	36.53	10.91	69.05	-0.59
		GLA tree cover	15.68	5.77	40.16	0.74
Walkability	1000 m network	Walkability z-score	0.12	2.44	19.14	1.03
		Walkability z-score with GLA total	-2.90	2.45	19.14	1.03
		Walkability z-score with GLA ground	2.46	1.81	18.76	0.03
		Walkability z-score with GLA tree	-1.58	2.37	19.20	1.02
		GLA total (% cover)	39.40	12.96	67.14	0.35
		GLA ground cover (% cover)	29.26	8.54	59.31	-0.40
		GLA tree cover (% cover)	14.79	6.11	51.16	1.03
		Road/Path network length (m)	37197	12659	85647	0.35

		Proportion pedestrianised length (%)	26.06	9.60	55.26	1.14
Air pollution (µg/m^3)		Annual average NO_2 2010	36.60	5.26	89.60	2.96
Traffic noise (dB(A))		L_{day} 2013	55.46	4.53	38.21	1.88
		L_{eve} 2013	51.86	4.53	38.21	1.91
		L_{night} 2013	47.47	4.51	38.20	1.99
		$L_{Aeq,16h}$ 2013	54.80	4.53	38.22	1.88
		L_{den} 2013	54.27	4.52	38.21	1.97

7.2.2 *Interrelation of built environment exposure variables*

Area-level deprivation and household income might impact the relationship of greenspace and health outcomes. The box plots by IMD quintiles (neighbourhood deprivation) and household income level in **Figure 26** and **Figure 27**, respectively, show the distribution of each exposure variable across strata. Addresses in the highest quintile of IMD (most deprived) were exposed to lower greenness (500 m buffer shown - the relationship was stable across all buffer sizes, not shown) and higher walkability score, and were exposed to higher concentrations of NO_2 and traffic noise. When stratified by total household income, trends were similar but weaker.

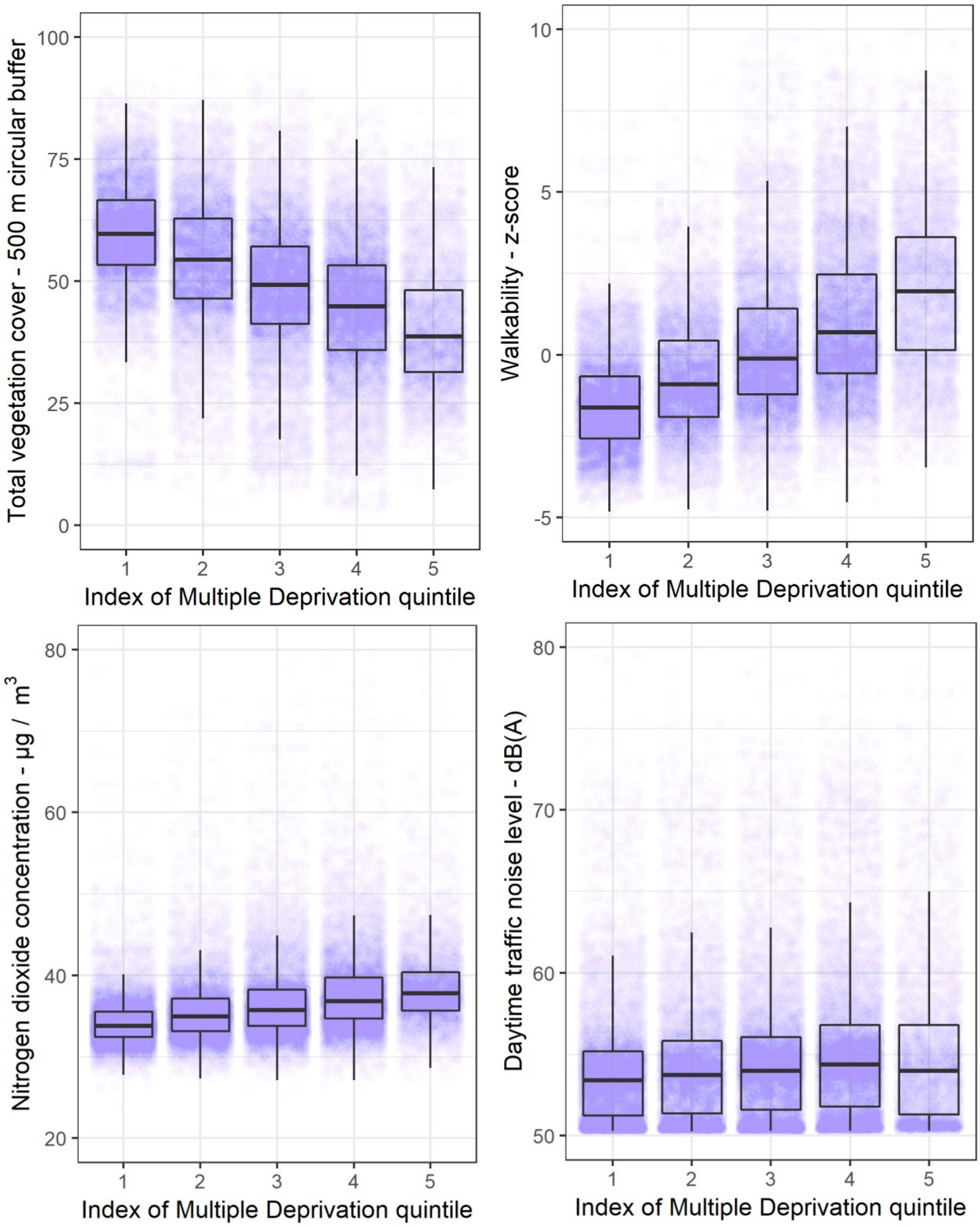


Figure 26. Box plots showing the distribution of built environment exposures at UK Biobank London addresses. Clockwise from top left: vegetation cover in a 500 m circular distance buffer, walkability z-score, annual average nitrogen dioxide concentration (µg/m^3), annual average traffic noise (dB(A)), by quintiles of area-level deprivation (Index of Multiple Deprivation; IMD) – quintile 1 is the least deprived, and 5 is the most deprived. The violet colour is made up of small points (1 per address) to show the range of exposure across Greater London UK Biobank addresses (*n* = 58,587).

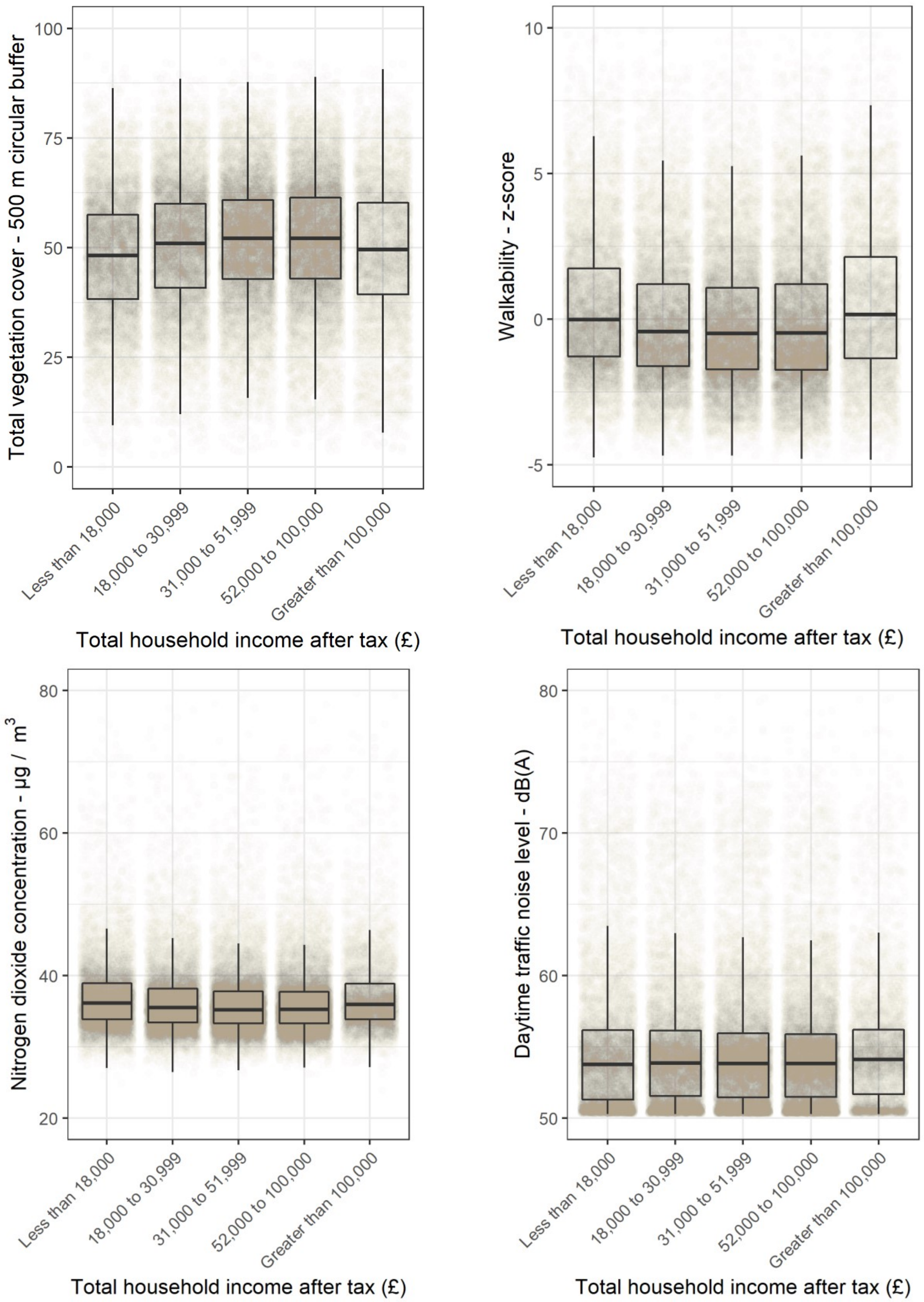
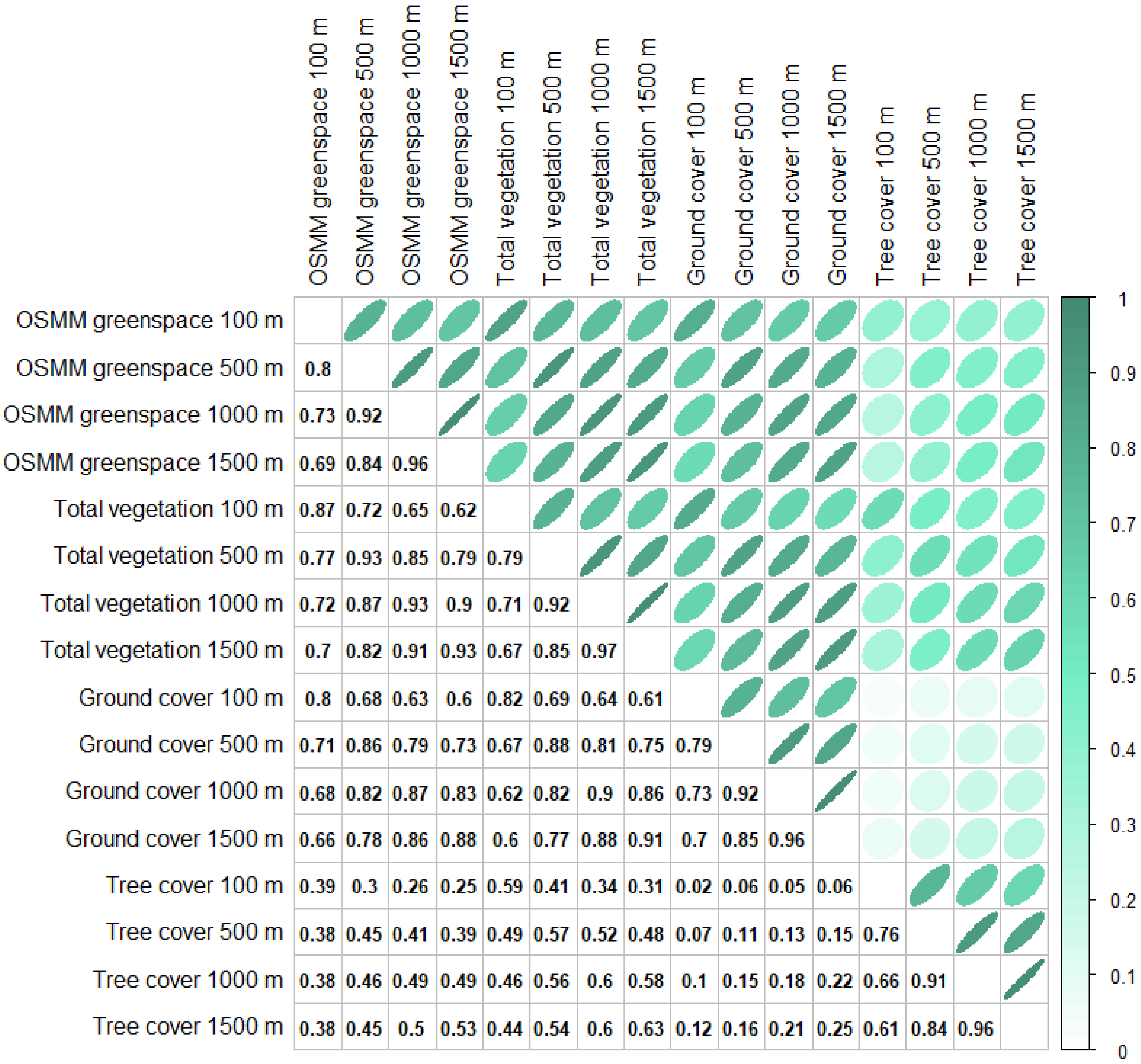


Figure 27. Box plots showing the distribution of built environment exposures assessed at UK Biobank London addresses. Clockwise from top left: vegetation cover in a 500 m circular distance buffer, walkability z-score, annual average nitrogen dioxide concentration (µg / m^3), annual average traffic noise (dB(A)), by (self-reported) categories of annual average total household income after tax (£). The brown colour is made up of small points (1 per address) to show the range of exposure across Greater London UK Biobank addresses (n = 58,587).

Strong positive correlation of the OS MasterMap™ Greenspace cover and the GLA total vegetation cover is shown in **Figure 28**. Correlations were stronger across the two datasets within the same buffer size, than between different buffer sizes within the same dataset, signalling agreement between the two datasets. In the GLA data, ground cover makes up most of the total vegetation cover, hence showed stronger correlation with total cover than tree cover, irrespective of buffer size. Tree cover showed weak correlation with ground cover, potentially due to the either/or categorisation within the dataset (a pixel is either tree canopy cover or ground cover). Further, the abundance of one type of vegetation (e.g., street trees) does not necessitate the abundance of the other (e.g. open ground cover), and vice-versa, which would weaken overall positive associations of ground cover and tree cover.

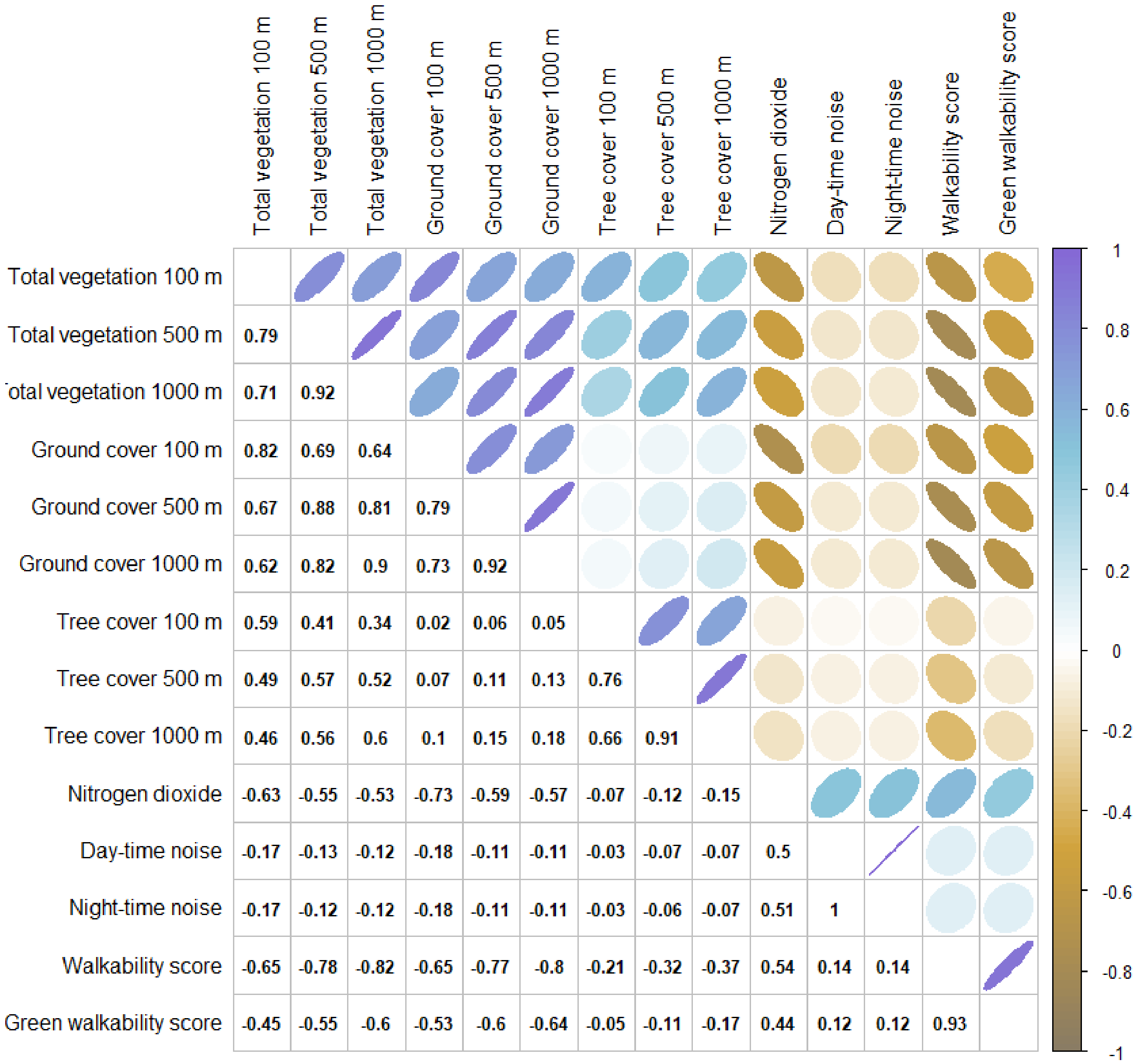


	OSMM greenspace 100 m	OSMM greenspace 500 m	OSMM greenspace 1000 m	OSMM greenspace 1500 m	Total vegetation 100 m	Total vegetation 500 m	Total vegetation 1000 m	Total vegetation 1500 m	Ground cover 100 m	Ground cover 500 m	Ground cover 1000 m	Ground cover 1500 m	Tree cover 100 m	Tree cover 500 m	Tree cover 1000 m	Tree cover 1500 m
OSMM greenspace 100 m																
OSMM greenspace 500 m	0.8															
OSMM greenspace 1000 m	0.73	0.92														
OSMM greenspace 1500 m	0.69	0.84	0.96													
Total vegetation 100 m	0.87	0.72	0.65	0.62												
Total vegetation 500 m	0.77	0.93	0.85	0.79	0.79											
Total vegetation 1000 m	0.72	0.87	0.93	0.9	0.71	0.92										
Total vegetation 1500 m	0.7	0.82	0.91	0.93	0.67	0.85	0.97									
Ground cover 100 m	0.8	0.68	0.63	0.6	0.82	0.69	0.64	0.61								
Ground cover 500 m	0.71	0.86	0.79	0.73	0.67	0.88	0.81	0.75	0.79							
Ground cover 1000 m	0.68	0.82	0.87	0.83	0.62	0.82	0.9	0.86	0.73	0.92						
Ground cover 1500 m	0.66	0.78	0.86	0.88	0.6	0.77	0.88	0.91	0.7	0.85	0.96					
Tree cover 100 m	0.39	0.3	0.26	0.25	0.59	0.41	0.34	0.31	0.02	0.06	0.05	0.06				
Tree cover 500 m	0.38	0.45	0.41	0.39	0.49	0.57	0.52	0.48	0.07	0.11	0.13	0.15	0.76			
Tree cover 1000 m	0.38	0.46	0.49	0.49	0.46	0.56	0.6	0.58	0.1	0.15	0.18	0.22	0.66	0.91		
Tree cover 1500 m	0.38	0.45	0.5	0.53	0.44	0.54	0.6	0.63	0.12	0.16	0.21	0.25	0.61	0.84	0.96	

Figure 28. Correlation plot of vegetation surrounding UK Biobank addresses in Greater London (n = 58,801) in circular distance buffers (100 m, 500 m and 1000 m) from two data sources – Ordnance Survey MasterMap™ (OSMM) Greenspace and the GLA GeoInformation Group vegetation. The latter dataset is further categorised as tree canopy cover (≥2.5 m height) and groundcover cover (<2.5 m height). Darker colours show stronger correlations, and the shape of the ellipse shows strength of the association, with narrower ellipses indicating a stronger correlation, which is also shown numerically (lower portion of the plot).

Vegetation cover in a circular distance buffer showed strong negative correlation with air pollution, and walkability. Including vegetation in the air pollution and walkability model at the exposure assessment stage resulted in the anticipated amplification (air pollution) and attenuation (walkability) of negative correlation with vegetation cover. That is, the strongest negative correlation observed between air pollution and vegetation was with ground cover 100m (-0.73). This strong correlation was expected as this was the variable included in the air pollution model. Further, compared to the standard 3-component walkability score, negative correlations of vegetation with green walkability were attenuated. For example, green walkability and vegetation 500 m showed attenuation of the negative correlation (-0.55), compared to walkability and vegetation 500 m (-0.78). This was due to the inclusion of vegetation within the network buffer in the green walkability score.

Compared to air pollution, traffic noise exposure was more weakly correlated with vegetation cover at all buffer sizes. Day-time and night-time noise were produced by the same model and so show complete correlation. NO_2 and traffic noise showed moderate positive correlation, which was of similar order of magnitude as the correlation of noise and walkability, and NO_2 and walkability (~0.50) (see **Figure 29**).



	Total vegetation 100 m	Total vegetation 500 m	Total vegetation 1000 m	Ground cover 100 m	Ground cover 500 m	Ground cover 1000 m	Tree cover 100 m	Tree cover 500 m	Tree cover 1000 m	Nitrogen dioxide	Day-time noise	Night-time noise	Walkability score	Green walkability score
Total vegetation 100 m														
Total vegetation 500 m	0.79													
'otal vegetation 1000 m	0.71	0.92												
Ground cover 100 m	0.82	0.69	0.64											
Ground cover 500 m	0.67	0.88	0.81	0.79										
Ground cover 1000 m	0.62	0.82	0.9	0.73	0.92									
Tree cover 100 m	0.59	0.41	0.34	0.02	0.06	0.05								
Tree cover 500 m	0.49	0.57	0.52	0.07	0.11	0.13	0.76							
Tree cover 1000 m	0.46	0.56	0.6	0.1	0.15	0.18	0.66	0.91						
Nitrogen dioxide	-0.63	-0.55	-0.53	-0.73	-0.59	-0.57	-0.07	-0.12	-0.15					
Day-time noise	-0.17	-0.13	-0.12	-0.18	-0.11	-0.11	-0.03	-0.07	-0.07	0.5				
Night-time noise	-0.17	-0.12	-0.12	-0.18	-0.11	-0.11	-0.03	-0.06	-0.07	0.51	1			
Walkability score	-0.65	-0.78	-0.82	-0.65	-0.77	-0.8	-0.21	-0.32	-0.37	0.54	0.14	0.14		
Green walkability score	-0.45	-0.55	-0.6	-0.53	-0.6	-0.64	-0.05	-0.11	-0.17	0.44	0.12	0.12	0.93	

Figure 29. Correlation plot of vegetation cover (The GeoInformation Group data) surrounding UK Biobank addresses in Greater London (n = 58,801) in circular distance buffers (100 m, 500 m and 1000 m), and annual average nitrogen dioxide concentrations, day-time and night-time noise levels, walkability score in a 1000 m network distance buffer, and 'green' walkability (supplemented with vegetation data) in a 1000 m network buffer. Blue colours show positive associations, and golden colours show negative associations, with darker colours showing stronger correlations. The shape of the ellipse shows the strength of the association, with narrower ellipses indicating a stronger correlation. Correlations are also shown numerically (lower portion of the plot).

7.2.3 *Assessing deviations from linearity*

I assessed if built environment exposure variables (greenspace, air pollution, traffic noise, walkability) at UK Biobank addresses had non-linear relationships. Findings supported a non-linear association of vegetation (The GeoInformation Group data; 500 m) and walkability, which are shown in **Table 18**. In terms of R^2 and RMSE, the model fitted with a natural cubic spline (3 df) predicted better than the model fitted with a linear relationship (1 df) for vegetation (500 m) and walkability; this was not the case for vegetation and NO_2, or for vegetation and noise, which were best explained by the linear model. I verified this finding with all vegetation buffer sizes (not shown). There was a weaker relationship between vegetation and traffic noise, so noise was not well predicted by the linear or the cubic spline model. The interrelationship of vegetation with other built environment variables and residuals are presented graphically. The best fit line (with 95% confidence intervals) is fitted linearly for total vegetation (The GeoInformation Group data; 100 m) and annual average NO_2 concentration (**Figure 30)** and traffic noise (**Figure 31)**. The linear fit in **Figure 32** can also be compared with the marginally better fit (natural cubic spline, 3 df) for total vegetation (The GeoInformation Group data; 500 m) and walkability shown in **Figure 33**.

To further explore the non-linear relationship of total vegetation and walkability, I used a generalised additive model (GAM) with smoothing via REML method. To show the smooth term I represented this graphically. However, irrespective of potential overfitting (see **Figure 34**), the GAM model offered no advantage over the model fitted with a natural cubic spline (3 df) in terms of R2 or RMSE (**Table 18**) for vegetation 500 m and walkability.

Table 18. R^2 and RMSE of models of vegetation cover and other built exposures (walkability, air pollution and traffic noise) fit with linear (1 df), natural cubic spline (3 df) and generalised additive models (GAM; Restricted maximum likelihood smooth term). Higher R^2 and lower RMSE indicate a better model fit.

Exposure (fit)	**R^2**	**RMSE**
Walkability (1 df)	0.61	1.51
Walkability (3 df)	0.62	1.49
Walkability (GAM, REML)	0.62	1.49
Air pollution (1 df)	0.39	4.10
Air pollution (3 df)	0.34	4.26
Traffic noise (1 df)	0.03	4.44
Traffic noise (3 df)	0.03	4.44

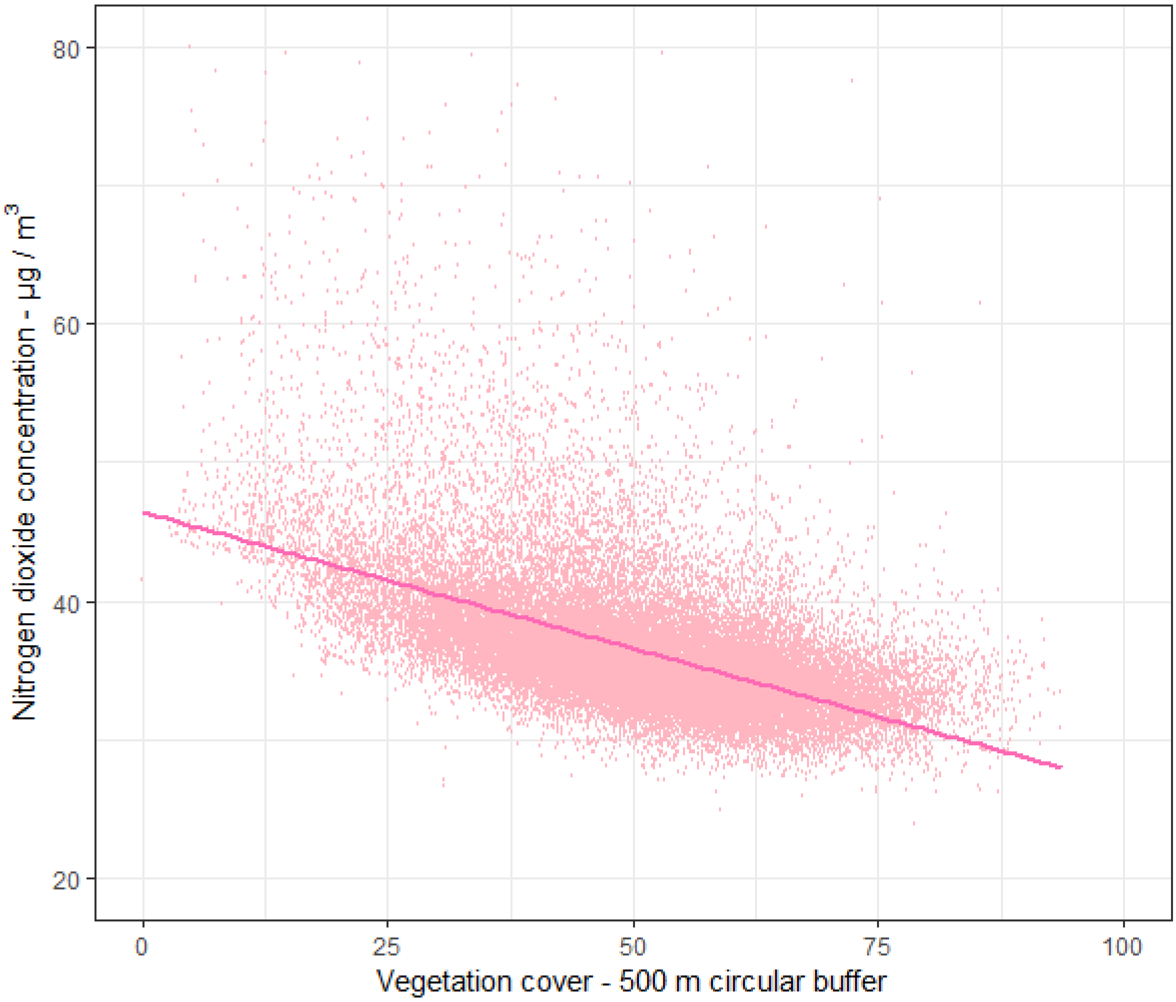


Figure 30. Relationship of GLA total vegetation cover (The GeoInformation Group data; 100 m circular distance buffer) and annual average nitrogen dioxide concentration (2010) at residential addresses in UK Biobank London. Residuals shown in light pink, line of best fit (1 df) shown in hot pink, with 95% confidence intervals in grey.

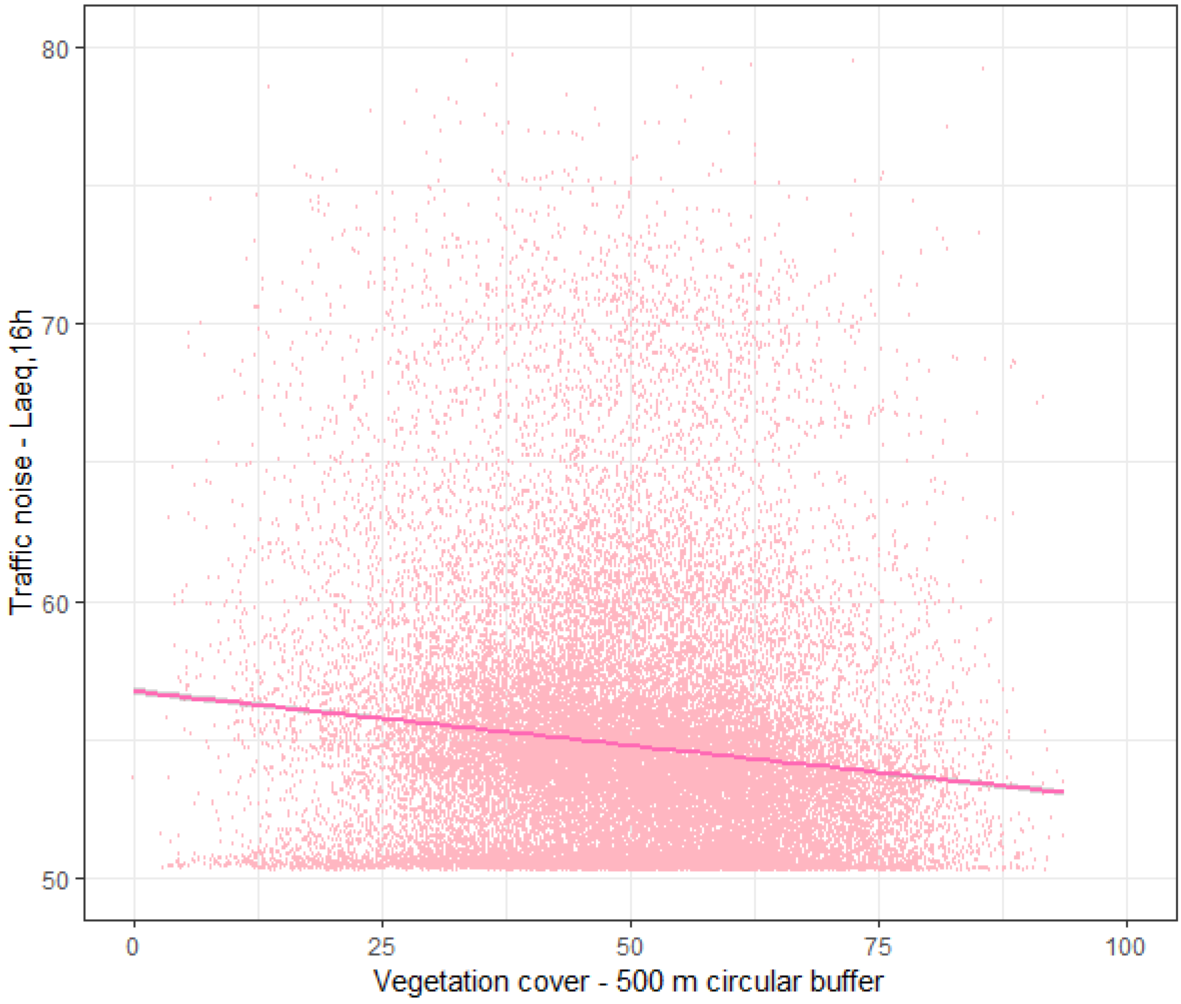


Figure 31. Relationship of GLA total vegetation cover (The GeoInformation Group data; 100 m circular distance buffer) and annual average traffic noise for 07:00 to 23:00 ($L_{Aeq,\ 16h}$ (2013) in dB(A)) at residential addresses in UK Biobank London. Residuals shown in light pink, line of best fit (1 df) shown in hot pink, with 95% confidence intervals in grey.

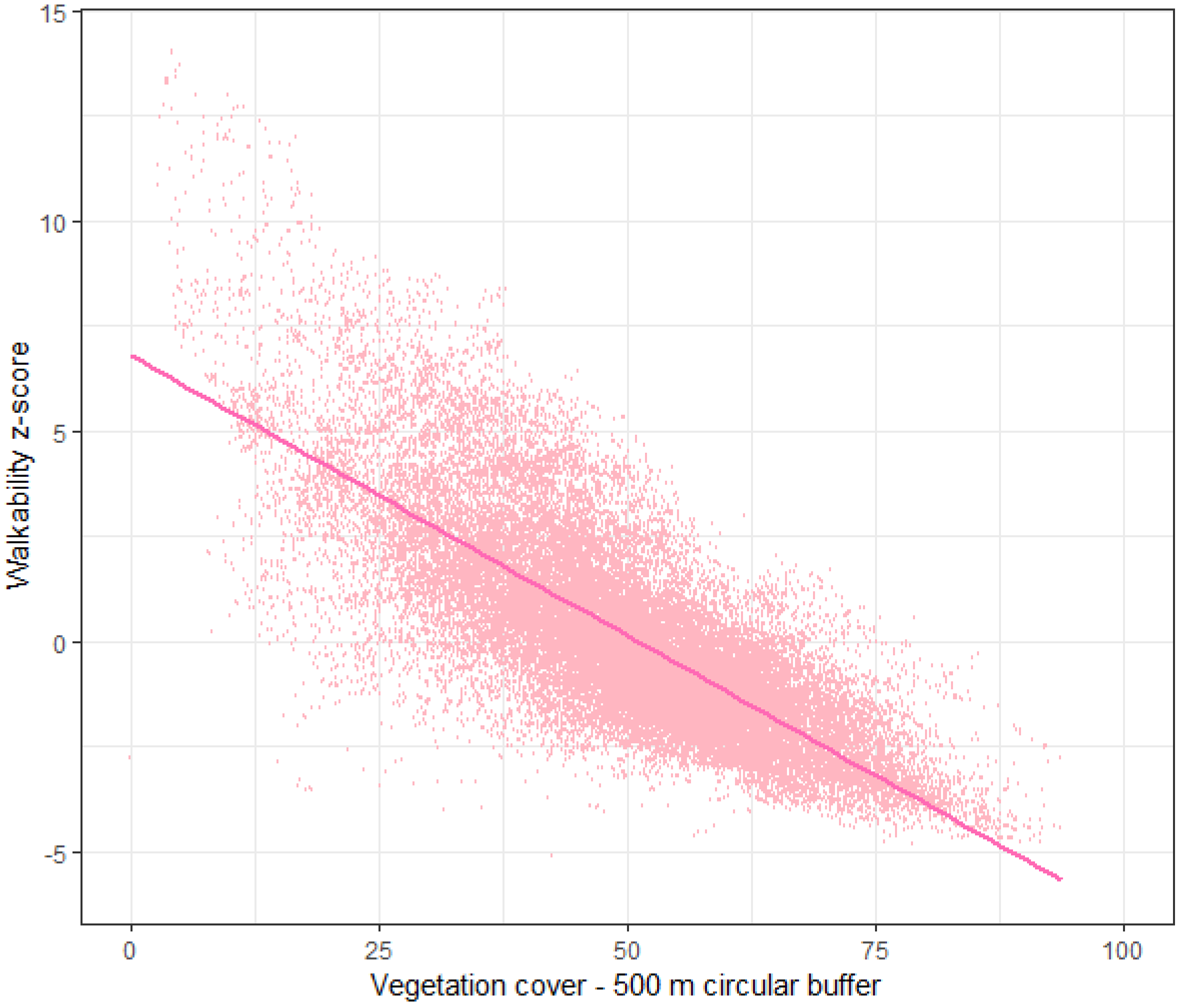


Figure 32. Relationship of GLA total vegetation cover (The GeoInformation Group data; 500 m circular distance buffer) and walkability z-score at residential addresses in UK Biobank London. Residuals shown in light pink, line of best fit (1 df) shown in hot pink, with 95% confidence intervals in grey.

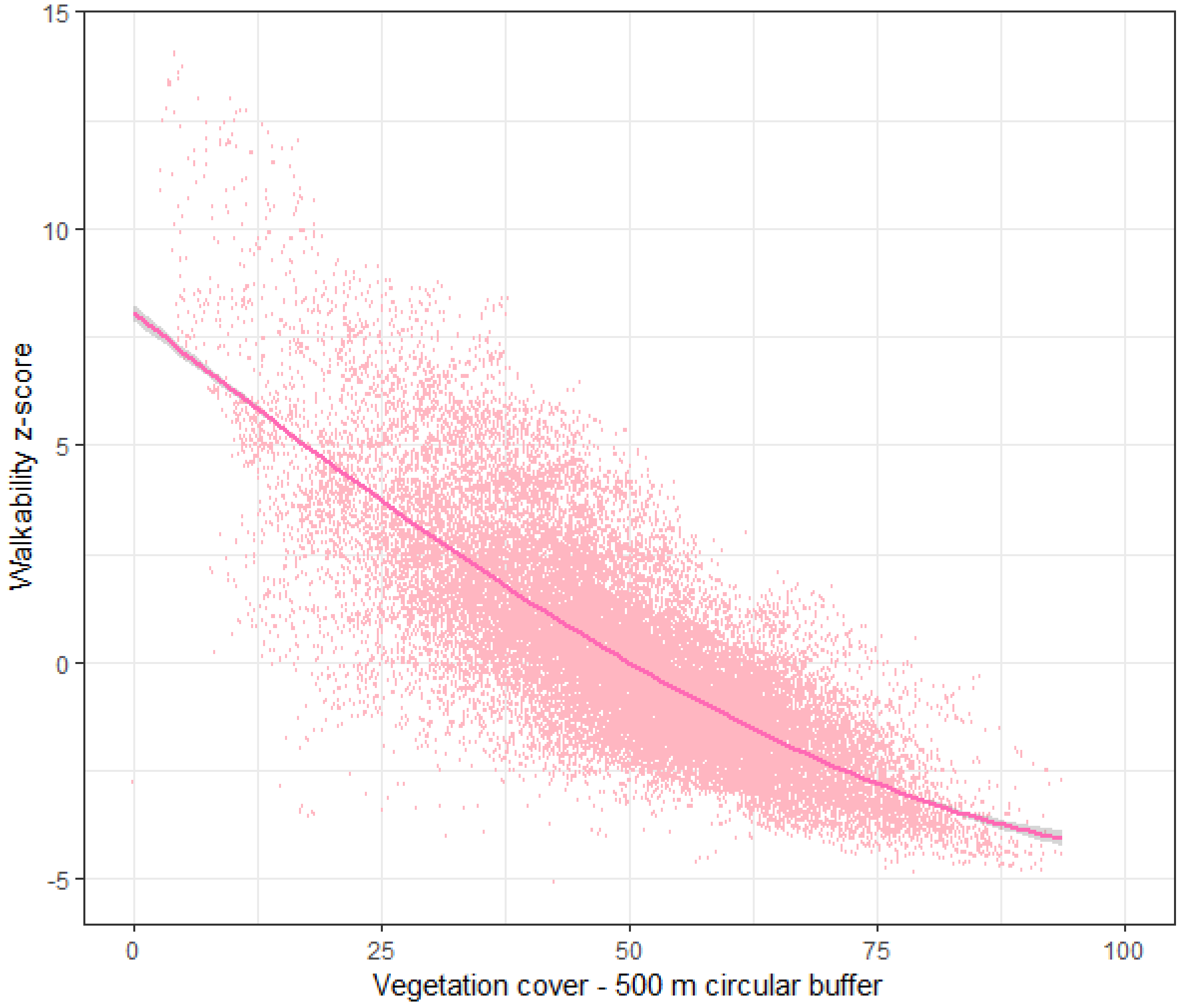


Figure 33. Relationship of GLA total vegetation cover (The GeoInformation Group data; 500 m circular distance buffer) and walkability z-score at residential addresses in UK Biobank London. Residuals shown in light pink, fitted with natural cubic spline (3 df) shown in hot pink, and 95% confidence intervals shown in grey.

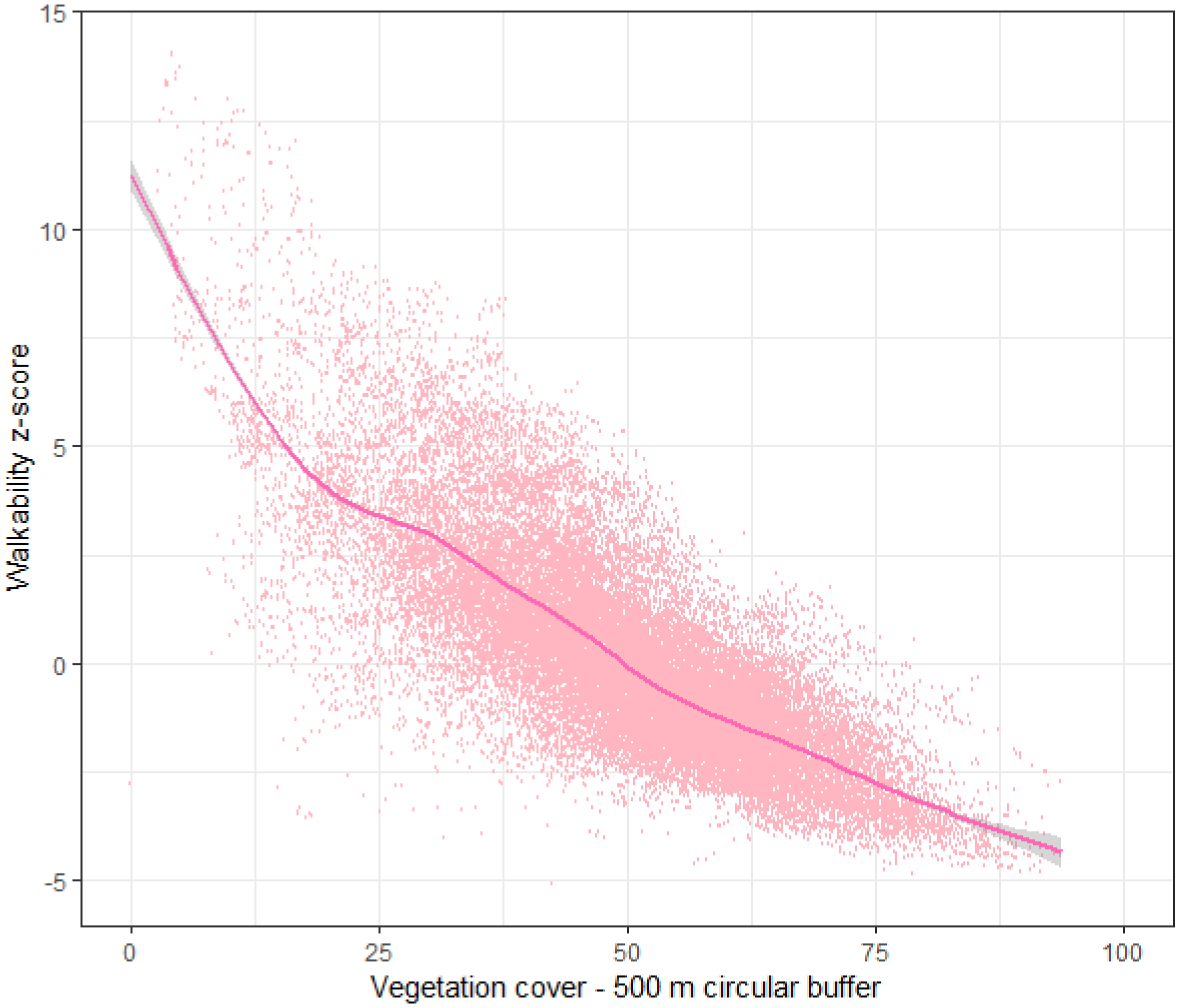


Figure 34. Relationship of GLA total vegetation cover (The GeoInformation Group data; 500 m circular distance buffer)and walkability z-score at residential addresses in UK Biobank London. Residuals shown in light pink, fitted with smooth produced by a general additive model (GAM, REML method) shown in hot pink, and 95% confidence intervals in grey.

7.2.3.1 Implications of interrelationship of built environment exposures

UK Biobank London addresses in the most deprived neighbourhood quintile compared to the least deprived, had higher annual average NO_2 levels, higher traffic noise, higher walkability score, and lower levels of surrounding greenness. Incomplete adjustment for neighbourhood deprivation level could therefore lead to biased effect estimates in epidemiological analyses. Further, the shallow, U-shape distribution of the exposure variables across total household income categories potentially corresponds to the most and least wealthy households being located in central London neighbourhoods (i.e. both the lowest and the highest income categories have lower green, higher walkability, higher NO_2 and higher traffic noise levels compared to moderate income categories). Trends across categories of household income, however, were less distinct than those across neighbourhood deprivation quintiles, which suggests that household income level is not interchangeable with neighbourhood deprivation level in adjusting for SES, and is not as important a confounder as neighbourhood deprivations level for future epidemiological analyses. Given the clear trend of vegetation cover, walkability and air pollution across deprivation quintiles, the quality of the deprivation index used for confounder adjustment is an important consideration for environmental epidemiological research.

Exposures to surrounding greenness (100 m, 500 m and 1000 m), walkability z-scores, annual average NO_2, and annual average traffic noise were, overall, moderately correlated, suggesting that confounding of exposures is possible. In instances of strong correlation (e.g., walkability and greenness 1000 m), multicollinearity of exposures should be reviewed when using multiple exposure variables in epidemiological analyses. Though, analyses using moderately correlated variables across UK Biobank London addresses would be expected to be able to reliably determine independent effects.

The moderate negative correlations of vegetation cover with air pollution and traffic noise that I found were similar to those reported in other studies (Hystad et al., 2014, Klompmaker et al., 2019a, Tétreault et al., 2013). As detailed in the literature review, the association of vegetation cover and air pollution is potentially due to causal and non-causal relationships between these exposures. That is, direct removal of air pollutants from the air by uptake via leaf stomata or deposition to leaf surfaces offers a causal explanation, whereas a non-causal explanation of correlation could be the lack of air pollution sources in greenspace, and the enhanced dispersion of pollutants across open areas compared to built-up areas (Janhäll, 2015). According to studies on deposition of pollutants on vegetation, the effect is likely small, and strongest for PM_{10} as oppose to NO_2 (Nowak et al., 2014). A combination of deposition and dispersion is likely to explain the correlation.

In this PhD thesis, air pollution concentrations (NO_2) and road traffic noise exposures were predicted by models, which are partly based on traffic flow and land-use variables. Due to the use of similar traffic and land-use predictor variables, correlations between modelled exposures might be higher than between true exposures. Also, in the LUR model ground cover 100 m was included as a predictor, potentially leading to overestimation of the relationship between greenness (all buffer sizes) and annual average NO_2. However, the correlation coefficients I found between modelled air pollution and greenness 500 m exposure were similar to those reported in another study that used NDVI and ESCAPE air pollution estimates, which do not include vegetation cover as an air pollution predictor in the model (Klompmaker et al., 2019a). Further, modelled air pollution and noise correlations were similar to a study that evaluated correlations between measured (short-term) air pollution and noise exposures in the urban context (Davies et al., 2009). Moreover, correlation between modelled air pollution and traffic noise exposure across all London postcodes, which were estimated using alternative prediction models, showed correlations of the same order of magnitude (Fecht et al., 2016).

Vegetation and walkability showed a strong negative correlation. Correlation of greenness (100 m and 500 m) and walkability in UK Biobank London (r = -0.65 and -

0.78, respectively) was slightly stronger than that of a study in the US (James et al., 2017). The US study assessed walkability using similar input variables (e.g., population, road junction and business density) and assessed greenness using NDVI 250 m (r = -0.6). Although, the slightly stronger correlation in UK Biobank London could be due to differences in the underlying input data (e.g., OS Points of Interest, OS Urban Paths network data compared to less detailed US data sources). James et al. (2017) reported a non-linear association of walkability and greenness across the US addresses, similar to the non-linear association in UK Biobank London shown here. The negative, non-linear relationship of walkability and greenness warrants further investigation, particularly in light of the inconsistent evidence supporting the physiological pathway (Nieuwenhuijsen et al., 2017a). Effect modification of surrounding greenness and physical activity by walkability, for example, should be explored in future work.

Though there is evidence for associations between air pollution, walkability, greenness, and cardiovascular disease incidence, studies have tended to assess these associations linearly or using categorical variables. These approaches either ignore potential deviations from linearity in the dose–response curve or lack efficiency. Some studies have shown non-linear associations of built environment exposures (James et al., 2016a, James et al., 2017, Klompmaker et al., 2019a). By exploring linearity, I aimed to highlight the potential for non-linear confounding, which could result in biased observed associations with epidemiological outcomes (e.g., cardiovascular disease incidence) if not adequately adjusted.

7.3 Chapter summary

In this chapter, I graphically showed environmental exposure variables stratified by quintiles of neighbourhood deprivation and annual average household income level. Results were indicative of a strong linear association of neighbourhood deprivation and greenspace (negative association), and walkability (positive association) in UK Biobank London. Annual average household income after tax showed a weaker, shallow U-shape association. The interrelation of surrounding greenness cover and air pollution, traffic noise and walkability were explored iteratively. Greenspace and air pollution showed a (linear) negative association. The greenspace and walkability association was also negative, though was non-linear, and was best described using a natural cubic spline.

8 Epidemiological analysis

In this chapter I conduct epidemiological analyses using the OS MasterMap™ Greenspace exposure assigned at UK Biobank residential addresses (see Chapter 4.1). I provide the results of survival analysis using UK Biobank linked mortality data. Specifically, I explore the association of all categories of OS MasterMap™ Greenspace and cardiovascular disease mortality using UK Biobank addresses in England. I also explore OS MasterMap™ Greenspace and non-injury mortality. I adjust associations for important participant characteristics, lifestyle variables and area-level confounders, and discuss the potential of residual confounding by deprivation in analyses of greenspace and mortality based on findings.

8.1 Exposure and address study inclusion

In this analysis, I used OS MasterMap™ Greenspace cover in circular distance buffers, as described in Chapter 4.1. In this analysis, I focused on the 500 m circular distance buffers, with sensitivity analyses at 100 m and 1000 m to assess across a range of residential-neighbourhood scales. I chose the 500 m buffer as the focal buffer as other cohort studies have shown greenness (NDVI) 500 m to be associated with cardiovascular and non-injury mortality at this scale (Crouse et al., 2017, Villeneuve et al., 2012b).

I excluded addresses in Scotland and Wales from analysis (with or without OS MasterMap™ Greenspace data coverage) as Scotland and Wales' deprivation index differs from that of England, and is not directly comparable. Further, in UK Biobank, Scotland's death register update cycle is also not aligned with England and Wales' cycle (Scotland censoring date: 30 November 2016). **Figure 35** shows the distribution of urban areas in England with Ordnance Survey MasterMap™ Greenspace data in this study.

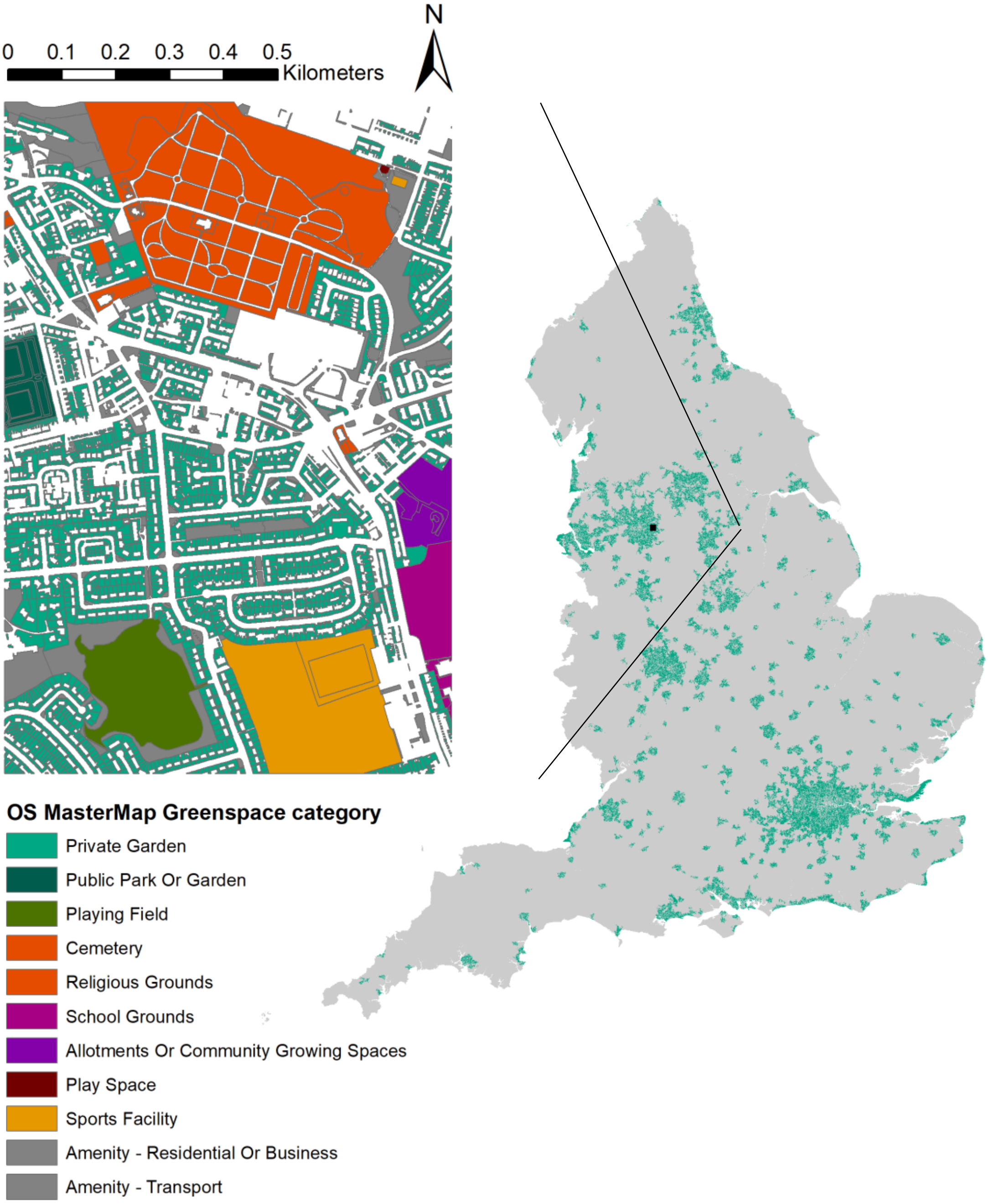


Figure 35. Distribution of urban areas in England with Ordnance Survey MasterMap™ Greenspace data in this study (right), with inset map of an area in Greater Manchester showing some (11 of 18) primary function categories in the data (e.g., Allotments, Public Parks, and Private Gardens).

8.2 Outcome

I assessed death occurring from the date of baseline assessment to the current (England) UK Biobank censor date (31st January 2018). I conducted survival analysis on non-injury deaths (excluding ICD-10 categories coded 'V'—'Z'). I also conducted cause-specific analysis of all circulatory disease deaths (I00—I99). I used ICD-10 groupings as per the Global Burden of Disease Study (James et al., 2018, WHO, 2018) following advice and coding input from Dr Robbie M. Parks.

8.3 Methods

8.3.1 *Prevalent cardiovascular disease study exclusion*

I excluded participants who were diagnosed with myocardial infarction (ST segment elevation or non-ST segment elevation; data field 42000) or stroke (ischaemic stroke, intracerebral haemorrhage or subarachnoid haemorrhage; data field 42006) prior to baseline assessment. The exclusion of prevalent cardiovascular disease, with validation via hospital records, allowed for exclusion of prevalent CVD irrespective of self-report via the touchscreen questionnaire.

Given that questionnaire items in UK Biobank related to self-reported disability were strong predictors of 5-year mortality (Ganna and Ingelsson, 2015), I used the item *unable to work due to sickness or disability* to exclude participants from the analysis. A large proportion (n = 120,163) of participants reported high blood pressure (hypertension) diagnosis, and, as opposed to excluding these participants, I adjusted for hypertension in the final model to avoid loss of power.

8.3.2 *Covariates*

I selected confounding factors for this analysis a-priori based on previous greenspace and circulatory disease mortality literature (Crouse et al., 2017, de Keijzer et al., 2017, James et al., 2016a, Vienneau et al., 2017, Villeneuve et al., 2012a) and known cardiovascular disease risk factors (Tsao and Vasan, 2015). Age of participant was used as the underlying timescale so was not adjusted for as a covariate. I used high blood pressure as a confounder in this analysis, as oppose to excluding participants with high blood pressure diagnosis prior to baseline, which resulted in many exclusions (see previous section). To maximize accuracy, I used measured as opposed to self-reported hypertension in final analyses, and used 140 mmHg systolic over 90 mmHg diastolic blood pressure as the cut off for high blood pressure. Measured blood pressure was available from baseline assessment: participants gave two blood pressure measurements that were taken after 2 minutes rest (seated) using a cuff and an Omron HEM-7015IT digital monitor. I calculated mean systolic (SBP) and diastolic blood pressure (DBP) from 2 automated or 2 manual readings. For individuals with 1 manual and 1 automated blood pressure reading, I used the mean of the 2 values. For individuals with a single BP measurement (1 manual or 1 automated BP reading), I used the single measurement. This method was replicated from another UK Biobank study (Pazoki et al., 2018).

UK Biobank collected detailed data on participant characteristics and lifestyle variables at baseline. I adjusted models incrementally, firstly, for sex in **Model 1**. In **Model 2**, I adjusted for multiple personal contextual variables, which, as outlined in Section 3.4., are typical in cardiovascular epidemiology and greenspace epidemiological analyses. My rationale for showing the results of **Model 2** versus **Model 3** in my findings was to show the change in effect estimates expected due to the strong (inverse) relationship of greenspace cover and IMD area-level deprivation (see Chapter 7). I supplemented the main analysis (**Model 3**) by adding air pollution and traffic noise exposure into the model (**Model 4**) to assess if effect estimates were substantially altered via adjustment, therefore showing effect change (reduction in effect expected) from adjusting for associated environmental exposures after the effect of greenspace, personal contextual and area-level deprivation have been accounted for in the model. A large effect reduction would indicate that associations of greenspace with air pollution and traffic noise either could not be disentangled (due to correlation), or that air pollution and noise were mediators of the greenspace-mortality association, providing impetus to conduct mediation analysis in future work. Covariates included in models were as follows:

- **Model 1** adjusted for sex;
- **Model 2** adjusted for personal contextual covariates – sex, ethnicity (White versus non-White), total household income level after tax (two highest groups of income binned into single group to meet proportional hazards assumptions), smoking status (‘never’ ‘previous’, ‘current’), pack years smoking, alcohol intake (grams/week), high blood pressure, and diabetes (excluding gestational diabetes);
- **Model 3** adjusted for personal contextual covariates (as in **Model 2**) + Index of Multiple Deprivation (IMD) neighbourhood deprivation level (2010);
- **Model 4** adjusted for personal contextual and area-level deprivation covariates (as in **Model 3**) + Annual average NO_2 concentration + annual average $L_{Aeq,\ 16hr}$ traffic noise exposure level

Covariates that I defined as ‘on the causal pathway’ were not adjusted for in analyses. For example, UK Biobank studies have shown body mass index (BMI) to be associated with both greenspace, cardiometabolic and cardiovascular outcomes (Celis-Morales et al., 2017, Lyall et al., 2017, Sarkar, 2017), however, physical activity levels impact BMI and adjusting for BMI could result in a dampening of the true overall effect of the physiological pathway linking greenspace and cardiovascular outcomes. As oppose to adjustment for such factors, I deemed mediation appropriate for future analysis.

Air pollution and traffic noise exposure have also been associated with both greenspace and cardiovascular outcomes (Brook et al., 2010b, Cai et al., 2018, de Keijzer et al., Klompmaker et al., 2019a), though, in this case, I opted to exclude the effect of air pollution and traffic noise in **Model 4** to observe changes in effect estimates when the environmental pathway is excluded.

8.3.3 *Missing confounder study exclusion*

I did not carry out imputation. I excluded participants from mortality analyses who had missing data (including ‘Prefer not to answer’ and ‘Don’t know’ responses) at baseline assessment. **Figure 36** shows participant exclusions due to address location, missing covariates and prevalent CVD at baseline.

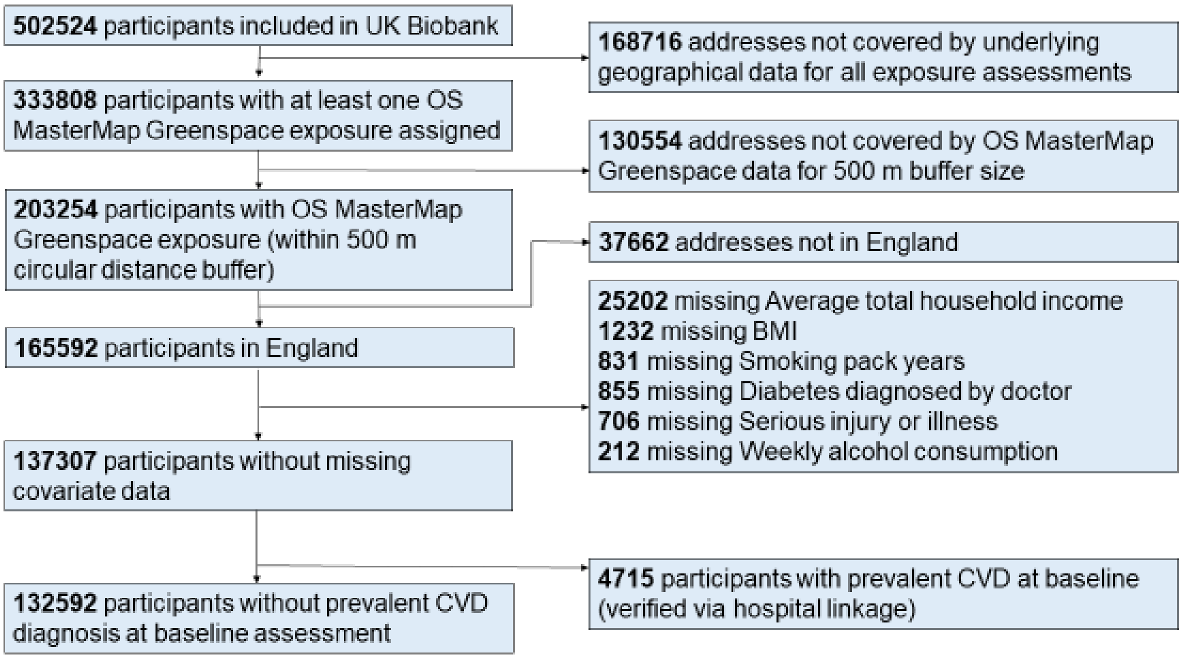

502524 participants included in UK Biobank
168716 addresses not covered by underlying geographical data for all exposure assessments
333808 participants with at least one OS MasterMap Greenspace exposure assigned
130554 addresses not covered by OS MasterMap Greenspace data for 500 m buffer size
203254 participants with OS MasterMap Greenspace exposure (within 500 m circular distance buffer)
37662 addresses not in England
165592 participants in England
25202 missing Average total household income
1232 missing BMI
831 missing Smoking pack years
855 missing Diabetes diagnosed by doctor
706 missing Serious injury or illness
212 missing Weekly alcohol consumption
137307 participants without missing covariate data
4715 participants with prevalent CVD at baseline (verified via hospital linkage)
132592 participants without prevalent CVD diagnosis at baseline assessment

Figure 36. Flow chart of UK Biobank participant study exclusions for the OS MasterMap™ Greenspace 500 m analysis.

8.3.4 *Statistical analysis*

I fit Cox proportional hazards models, with participant age as the underlying timescale, to calculate hazard ratios (HRs) and 95% confidence intervals (CIs) for associations of greenspace cover (500 m) at residential address and each mortality outcome group (all-cause, non-injury and circulatory disease mortality). Participant's age, as oppose to time-in-study, as the underlying timescale allows for comparison of individuals of the same age (Thiébaut and Bénichou, 2004). As the most appropriate choice of scale at which to assess greenspace for mortality assessment is unclear (James et al., 2016a, Mitchell et al., 2011), I conducted sensitivity analyses for alternative buffer sizes (100 m and 1000 m). I also conducted the 100 m analysis on the 1000 m sample, that is, I carried out further sensitivity analysis to assess if geographical scale or statistical power (higher *n*) was driving associations.

8.4 Findings and discussion

8.4.1 *Description of UK Biobank sample*

Individual-level data from 132,592 UK Biobank participants, were available for this analysis after exclusions for covariate missingness, prevalent CVD at baseline and lack of data coverage at residential address (**Table 19**). The mean age of the pooled population was 56.4 years; 56% were women. Non-injury and cardiovascular deaths were lower (*n*) in the most-green quintile (Q5) than in the least-green quintile (Q1).

Table 19. UK Biobank participants' characteristics by quintiles of Ordnance Survey MasterMap™ Greenspace cover within a 500 m buffer of their residential address (after exclusions for prevalent CVD; *n* = 132,592). Addresses in Quintile 1 (Q1) had the least surround greenspace cover in a 500 m circular distance buffer, and quintile 5 (Q5) addresses had the most.

	Ordnance Survey MasterMap™ Greenspace cover (500 m)				
Characteristic	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**
Cohort participants (n)	26525	26522	26539	26497	26509
Deaths (n)					
All-causes	1883	1726	1805	1809	1588
Non-injury	1198	1058	1095	1113	997
Cardiovascular	233	224	206	206	189
OS MasterMap™ Greenspace (all types 500 m) (percentage cover mean ±SD)	37.48 (±7.76)	51.57 (±2.60)	59.02 (±1.76)	64.56 (±1.52)	72.02 (±4.00)
Age (years) (mean ±SD)	55.18 (±8.32)	55.71 (±8.28)	56.34 (±8.26)	56.85 (±8.16)	57.41 (±8.00)
Sex (Female %)	54	56	55	56	56
Ethnicity (White %)	82	86	90	92	92
Smoking status (%)					
Never	48.73	52.33	54.23	55.18	57.31
Previous	34.58	34.36	33.92	34.06	33.56
Current	16.69	13.31	11.84	10.76	9.14
IPAQ physical activity group					
Low	16.72	18.04	18.93	18.96	19.08

Moderate	42.74	42.09	41.33	40.76	41.36
High	40.55	39.87	39.74	40.28	39.56
Smoking pack years (mean ±SD)	8.31 (±16.32)	7.37 (±15.03)	7.05 (±14.72)	6.77 (±14.18)	6.08 (±13.51)
Weekly alcohol intake in grams (mean ±SD)	130.60 (±174.25)	125.18 (±165.87)	122.83 (±160.19)	119.56 (±153.63)	119.81 (±148.87)
Body mass index	27.20 (±5.16)	27.33 (±4.98)	27.47 (±4.90)	27.46 (±4.83)	27.22 (±4.69)
High blood pressure (Yes %)	33.62	36.89	39.48	40.81	40.14
Diabetes diagnosed by doctor (Yes %)	6.29	5.81	5.49	5.05	4.82
Serious injury or other illness (Yes %)	20.31	19.37	18.90	19.68	18.91
Average annual total household income before tax (%)					
Less than 18,000	28.9	24.43	24.53	23.39	19.38
18,000 to 30,999	22.5	23.69	25.2	25.99	23.81
31,000 to 51,999	20.79	25.2	25.69	26.7	26.8
52,000 to 100,000	17.82	20.21	19.54	19.61	23.22
Greater than 100,000	9.98	6.48	5.04	4.32	6.79
Index of multiple deprivation (%)					
Low neighbourhood deprivation	10.23	23.69	30.61	38.90	48.90
Medium neighbourhood deprivation	37.93	44.51	44.41	38.17	33.07
High neighbourhood deprivation	51.84	31.8	24.98	22.93	18.03
Annual average nitrogen dioxide concentration (µg / m^3)	39.95 (±6.30)	36.13 (±3.88)	35.03 (±3.55)	34.19 (±3.21)	33.69 (±3.19)
Annual average noise level ($L_{Aeq,16h}$)	55.44 (±4.93)	54.78 (±4.32)	54.67 (±4.48)	54.22 (±4.15)	54.05 (±4.05)

8.4.2 *Non-injury mortality findings*

Overall, the main model (Model 3), indicated no statistically significant associations of all categories of OS MasterMap™ Greenspace cover 500 m and non-injury mortality after adjustment for personal contextual covariates and area-level deprivation (**Table 20**). Further adjusting for NO_2 + traffic noise (Model 4) slightly reduced odds of premature mortality compared to model 3, though results remained insignificant.

Table 20. Hazard ratios (95% confidence intervals (CIs)) of non-injury mortality by quintiles of OSMM Greenspace (500 m), with incremental adjustment for covariates.

Ordnance Survey MasterMap™ Greenspace 500 m (*n* = 132,592)	**Model 1: Adjusted for sex HR (95% CI)**	**Model 2: Adjusted for sex + personal contextual covariates HR (95% CI)**	**Model 3: Adjusted for sex + personal contextual covariates + area-level deprivation HR (95% CI)**	**Model 4: Adjusted for sex + personal contextual covariates + area-level deprivation + NO_2 + traffic noise HR (95% CI)**
Quintile 1	Reference	Reference	Reference	Reference
Quintile 2	0.85 (0.78, 0.93)	0.90 (0.82, 1)	0.94 (0.85, 1.04)	0.92 (0.83, 1.02)
Quintile 3	0.83 (0.77, 0.90)	0.89 (0.81, 0.99)	0.94 (0.85, 1.04)	0.91 (0.82, 1.02)
Quintile 4	0.82 (0.75, 0.89)	0.91 (0.83, 1)	0.97 (0.88, 1.08)	0.94 (0.84, 1.05)
Quintile 5	0.72 (0.66, 0.79)	0.86 (0.78, 0.95)	0.92 (0.83, 1.02)	0.89 (0.79, 1)

8.4.2.1 Sensitivity analysis

Due to the underlying data coverage, a larger sample was available at the smallest buffer size assessed (100 m). **Table 21** shows the results of survival analysis conducted on the 100 m buffer size with the full available sample after exclusions (*n* =

277,236) and **Table 22** shows the same analysis conducted on the 1000 m buffer size with the full available sample after exclusions (*n* = 108,574). The main model (Model 3) indicated statistically significant associations of all categories of OS MasterMap™ Greenspace cover 100 m and non-injury mortality after adjustment for personal contextual covariates and area-level deprivation. For example, in the most green quintile of addresses (Quintile 5) compared to the least green (Quintile 1), a protective association of greenspace cover and non-injury mortality was observed (HR = 0.91 (95% CI = 0.84, 0.99)). For all categories of OS MasterMap™ Greenspace cover 1000 m and non-injury mortality, the main model (Model 3), indicated no statistically significant associations of after adjustment for personal contextual covariates and area-level deprivation. When the 100 m circular distance was assessed in the smaller sample size used in the 500 m buffer size analysis (n = 132,592) (**Table 23**), no statistically significant associations remained after adjustment for personal contextual covariates and area-level deprivation (Model 3), indicating that different number of deaths (power) at each buffer size might be influencing findings.

Table 21. Hazard ratios (95% confidence intervals (CIs)) of non-injury mortality by quintiles of OSMM Greenspace (100 m), with incremental adjustment for covariates.

Ordnance Survey MasterMap™ Greenspace 100 m exposure (*n* = 277,236)	**Model 1: Adjusted for sex HR (95% CI)**	**Model 2: Adjusted for sex + personal contextual covariates HR (95% CI)**	**Model 3: Adjusted for sex + personal contextual covariates + area-level deprivation HR (95% CI)**	**Model 4: Adjusted for sex + personal contextual covariates + area-level deprivation + NO_2 + traffic noise HR (95% CI)**
Quintile 1	Reference	Reference	Reference	Reference
Quintile 2	0.87 (0.82, 0.93)	0.93 (0.86, 0.97)	0.95 (0.89, 1.03)	0.95 (0.88, 1.02)
Quintile 3	0.80 (0.75, 0.85)	0.89 (0.83, 0.96)	0.93 (0.87, 1)	0.93 (0.86, 1)

Quintile 4	0.77 (0.73, 0.82)	0.89 (0.82, 0.96)	0.93 (0.86, 1)	0.93 (0.85, 1)
Quintile 5	0.70 (0.66, 0.74)	0.87 (0.80, 0.94)	0.91 (0.84, 0.99)	0.90 (0.83, 0.99)

Table 22. Hazard ratios (95% confidence intervals (CIs)) of non-injury mortality by quintiles of OSMM Greenspace (1000 m), with incremental adjustment for covariates.

Ordnance Survey MasterMap™ Greenspace 1000 m (*n* = 108,574)	**Model 1: Adjusted for sex HR (95% CI)**	**Model 2: Adjusted for sex + personal contextual covariates HR (95% CI)**	**Model 3: Adjusted for sex + personal contextual covariates + area-level deprivation HR (95% CI)**	**Model 4: Adjusted for sex + personal contextual covariates + area-level deprivation + NO_2 + traffic noise HR (95% CI)**
Quintile 1	Reference	Reference	Reference	Reference
Quintile 2	0.96 (0.87, 1.06)	0.94 (0.83, 1.06)	0.97 (0.86, 1.10)	0.98 (0.87, 1.12)
Quintile 3	0.86 (0.78, 0.95)	0.89 (0.79, 1)	0.94 (0.83, 1.06)	0.96 (0.84, 1.09)
Quintile 4	0.82 (0.74, 0.91)	0.87 (0.77, 0.99)	0.93 (0.82, 1.06)	0.96 (0.83, 1.10)
Quintile 5	0.77 (0.70, 0.86)	0.90 (0.80, 1.02)	0.97 (0.85, 1.10)	1 (0.87, 1.17)

Table 23. Hazard ratios (95% confidence intervals (CIs)) of non-injury mortality by quintiles of OSMM Greenspace (100 m), with incremental adjustment for covariates, using the sample available for the 500 m analysis (n = 132,592).

Ordnance Survey MasterMap™ Greenspace 100 m exposure (*n* = 132,592)	**Model 1: Adjusted for sex HR (95% CI)**	**Model 2: Adjusted for sex + personal contextual covariates HR (95% CI)**	**Model 3: Adjusted for sex + personal contextual covariates + area-level deprivation HR (95% CI)**	**Model 4: Adjusted for sex + personal contextual covariates + area-level deprivation + NO_2 + traffic noise HR (95% CI)**
Quintile 1	Reference	Reference	Reference	Reference
Quintile 2	0.85 (0.79, 0.92)	0.90 (0.82, 0.98)	0.92 (0.84, 1.01)	0.92 (0.84, 1.01)
Quintile 3	0.83 (0.77, 0.90)	0.89 (0.81, 0.98)	0.93 (0.84, 1.02)	0.92 (0.83, 1.02)
Quintile 4	0.81 (0.74, 0.88)	0.91 (0.83, 1)	0.96 (0.87, 1.06)	0.95 (0.86, 1.06)
Quintile 5	0.71 (0.65, 0.77)	0.87 (0.79, 0.96)	0.92, 0.83, 1.02)	0.91 (0.82, 1.02)

8.4.3 *Cardiovascular mortality findings*

Overall, the main model (Model 3), indicated no statistically significant associations of all categories of OS MasterMap™ Greenspace cover 500 m and cardiovascular mortality after adjustment for personal contextual covariates and area-level deprivation (**Table 24**).

Table 24. Hazard ratios (95% confidence intervals (CIs)) of cardiovascular mortality by quintiles of OSMM Greenspace (500 m), with incremental adjustment for covariates.

Ordnance Survey MasterMap™ Greenspace 500 m (*n* = 132,592)	**Model 1: Adjusted for sex HR (95% CI)**	**Model 2: Adjusted for sex + personal contextual covariates HR (95% CI)**	**Model 3: Adjusted for sex + personal contextual covariates + area-level deprivation HR (95% CI)**	**Model 4: Adjusted for sex + personal contextual covariates + area-level deprivation + NO_2 + traffic noise HR (95% CI)**
Quintile 1	Reference	Reference	Reference	Reference
Quintile 2	0.94 (0.78, 1.13)	0.96 (0.76, 1.20)	1 (0.80, 1.25)	1 (0.79, 1.26)
Quintile 3	0.81 (0.67, 0.98)	0.88 (0.70, 1.11)	0.95 (0.75, 1.20)	0.94 (0.73, 1.20)
Quintile 4	0.78 (0.65, 0.94)	0.88 (0.70, 1.11)	0.97 (0.77, 1.23)	0.96 (0.74, 1.24)
Quintile 5	0.71 (0.58, 0.86)	0.84 (0.66, 1.07)	0.95 (0.74, 1.21)	0.94 (0.71, 1.25)

8.4.3.1 Sensitivity analysis

Due to the underlying data coverage, a larger sample was available at the smallest buffer size assessed (100 m). **Table 25** shows the results of survival analysis conducted on the 100 m buffer size with the full available sample after exclusions (*n* = 277,236) and

Table 26 shows the same analysis conducted on the 1000 m buffer size with the full available sample after exclusions (*n* = 108,574). The main model (Model 3) indicated statistically significant associations of OS MasterMap™ Greenspace cover 100 m and cardiovascular mortality after adjustment for personal contextual covariates and area-level deprivation when comparing the most and least green addresses (i.e. quintile 5 versus quintile 1). For example, in the most green quintile of addresses (Quintile 5) compared to the least green (Quintile 1), a protective association of greenspace cover and cardiovascular mortality was observed (HR = 0.80 (95% CI = 0.68, 0.97)). However, results should be interpreted with caution due to the smaller number of cardiovascular deaths than non-injury deaths in the sample; confidence intervals were large and no clear dose response across quintiles was observed. For all categories of OS MasterMap™ Greenspace cover 1000 m and cardiovascular mortality, the main model (Model 3), indicated no statistically significant associations of after adjustment for personal contextual covariates and area-level deprivation.

Table 25. Hazard ratios (95% confidence intervals (CIs)) of cardiovascular mortality by quintiles of OSMM Greenspace (100 m), with incremental adjustment for covariates.

Ordnance Survey MasterMap™ Greenspace 100 m exposure (*n* = 277,236)	**Model 1: Adjusted for sex HR (95% CI)**	**Model 2: Adjusted for sex + personal contextual covariates HR (95% CI)**	**Model 3: Adjusted for sex + personal contextual covariates + area-level deprivation HR (95% CI)**	**Model 4: Adjusted for sex + personal contextual covariates + area-level deprivation + NO_2 + traffic noise HR (95% CI)**
Quintile 1	Reference	Reference	Reference	Reference
Quintile 2	0.83 (0.73, 0.95)	0.88 (0.75, 1.03)	0.92 (0.78, 1.08)	0.92 (0.78, 1.09)
Quintile 3	0.74 (0.65, 0.85)	0.79 (0.67, 0.94)	0.84 (0.71, 1)	0.85 (0.71, 1.02)
Quintile 4	0.72 (0.62, 0.82)	0.86 (0.73, 1.01)	0.92 (0.77, 1.09)	0.93 (0.77, 1.11)
Quintile 5	0.57 (0.49, 0.66)	0.75 (0.63, 0.90)	0.80 (0.68, 0.97)	0.82 (0.67, 1)

Table 26. Hazard ratios (95% confidence intervals (CIs)) of cardiovascular mortality by quintiles of OSMM Greenspace (1000 m), with incremental adjustment for covariates.

Ordnance Survey MasterMap™ Greenspace 1000 m (*n* = 108,574)	**Adjusted for sex HR (95% CI)**	**Adjusted for sex + personal contextual covariates HR (95% CI)**	**Model 3: Adjusted for sex + personal contextual covariates + area-level deprivation HR (95% CI)**	**Model 4: Adjusted for sex + personal contextual covariates + area-level deprivation + NO_2 + traffic noise HR (95% CI)**
Quintile 1	Reference	Reference	Reference	Reference
Quintile 2	1.04 (0.82, 1.30)	1.08 (0.81, 1.44)	1.14 (0.86, 1.52)	1.17 (0.87, 1.57)
Quintile 3	1 (0.80, 1.26)	1.06 (0.78, 1.41)	1.15 (0.86, 1.54)	1.20 (0.88, 1.62)
Quintile 4	0.95 (0.76, 1.20)	1.07 (0.81, 1.43)	1.20 (0.90, 1.61)	1.26 (0.92, 1.73)
Quintile 5	0.77 (0.60, 0.98)	0.97 (0.72, 1.31)	1.09 (0.81, 1.49)	1.16 (0.82, 1.65)

In my preliminary epidemiological analysis of the UK Biobank cohort (England), I found a protective association of total (all categories) OS MasterMap Greenspace cover surrounding UK Biobank England residential addresses and non-injury mortality after adjusting for sex (age was used as the underlying timescale and therefore was not adjusted). Associations by quintiles of exposure were attenuated after adjustment for personal contextual covariates, and statistical significance (p <0.05) remained across all quintiles after adjustment for IMD neighbourhood-level deprivation only in the 100 m buffer size analysis. Association did not remain significant after adjustment for IMD across quintiles in the 500 m and 1000 m. My findings were consistent with those of James et al. (2016a), who found that smaller greenspace buffer sizes (e.g., 250 m) were more consistently associated with non-injury mortality than larger buffer sizes (e.g., 1250 m). However, in my analysis, larger confidence intervals in larger buffer sizes in my analysis might be due to loss of power with increasing buffer size, as oppose to a greenspace scale effect on increased risk of non-injury mortality. This is potentially driven by differences in sample sizes for different buffer sizes due to the underlying geographical data extent used for exposure assessment (see Chapter 4.1). Importantly, when reassessed using the 500 m buffer analysis sample (*n* =132, 592), the 100 m buffer analysis was no longer statistically significant after adjustment for confounders, suggesting difference in effect across buffer sizes should be interpreted with caution. Alternatively, residences in the outlying areas of the built up area, which were excluded when using the 500m buffer sizes, were driving the effect in the 100 m buffer analysis.

Results in the 100 m buffer are consistent with previous cohort studies of greenspace and non-injury mortality, which have shown a protective association, after adjustment for personal contextual and area-level confounding factors (James et al., 2016a). A meta-analysis of studies of associations of greenness and mortality (Rojas-Rueda et al., 2019) derived a pooled HR of all-cause mortality of 0.96 (95% CI 0.94, 0.97) for each increment of 0·1 NDVI in a residential buffer zone of 500 m or less. Of 9 studies included in the analysis, 7 showed a protective association. Further 3 of the studies included air pollution as a covariate in models, and 1 included noise, though similarly to the findings presented in this preliminary analysis, there was no appreciable effect on hazard ratios after adjusting for air pollution and noise. James et al. (2016a) found that air pollution explained a small proportion (4%) of the greenspace and non-injury mortality effect via mediation analysis.

Though air pollution and traffic noise did not substantially reduce greenspace-mortality association effect estimates. The lack of a substantial effect estimate change should not be interpreted as a lack of association of air pollution, traffic noise and mortality; the order that the variables are added to the model and the variance in death rate linked to greenspace (adjusted for other variables in the model) led to a small effect estimate change, indicating that reduction in air pollution and noise exposure were not primary drivers of the protective greenspace effect. If air pollution and traffic noise, however, had been added into the model as focal exposures it is likely, based on the literature, that harmful associations with cardiovascular health would be found.

A review by Gascon et al. (2016) concluded that there was good evidence for reduction of the risk of mortality from cardiovascular disease in areas with higher residential greenness. In my analysis, fewer cases of cardiovascular disease mortality compared to non-injury mortality resulted in lower confidence of the protective association of greenspace and mortality across all buffer sizes, and results were non-significant following adjustment for confounding factors. In the 100 m buffer size analysis, some evidence of an association with cardiovascular mortality was found though results were inconsistent across quintiles, and results should be interpreted with caution.

A national cohort study in Switzerland found lower mortality rates to be associated with greenness (NDVI) and greenspace (land use) data, with similar effect estimates found across the two greenspace exposure metrics (Vienneau et al., 2017). In my preliminary analysis on non-injury and cardiovascular mortality, I used all categories of greenspace (18 in total) for comparability with other analyses that have used non-specific greenness (NDVI) exposure (Crouse et al., 2019, Crouse et al., 2017, de Keijzer et al., 2017, James et al., 2016a, Villeneuve et al., 2012c, Zijlema et al., 2019). However, one of the strengths of the OS MasterMap Greenspace exposure is that specific functions (e.g., private gardens, public parks, sport facilities) have been attributed to the data by OS. This will allow for investigation of the effect of specific types of greenspace on mortality, and is expected to be a strength of the full analysis.

Another strength of this UK Biobank study lies in the potential for longitudinal analysis, with rich data available on individual level confounders. An ecological study conducted on the English population showed that greenspace cover at the small-area level (administrative boundaries) was associated with reduced premature mortality, and cardiovascular disease mortality (Mitchell and Popham, 2008a), however the study could not adjust for important individual-level confounders, such as smoking. UK Biobank has rich covariate data available, as shown in this analysis, and offers the potential to reassess earlier effect estimates with adequate confounder adjustment.

There were several limitations to this preliminary analysis. The exposure used was released in 2017 and baseline for UK Biobank assessment was 2006-2010, the use of this dataset relies on the assumption that greenspace cover has not increased substantially surrounding residential addresses (which might result in biased estimation of the protective effect of greenspace cover on mortality). Greenspace, relative to other environmental exposures, is expected to be relatively stable over time, though it remains possible that effect estimates are inaccurate due to this mismatch in data. Sensitivity analyses from 2017 onwards will be viable in assessments conducted in the future, though the current censoring date does not currently allow for such analyses to be conducted. Neighbourhood self-selection cannot be ruled out in this analysis. If participants in better health selected to move to neighborhoods with higher levels of greenness, this confounding by neighborhood preference could explain the relationship between greenness and non-injury mortality. Furthermore, people living in greener areas may be more active, less obese, and generally lead more healthy lifestyles, physical activity and adiposity were not adjusted for in this analysis, however, variables relating to these traits are available in UK Biobank and can be assessed in subsequent analyses as mediators and/or confounders, depending on if the analysis is focused on the physiological pathway or excludes the physiological pathway. I also have not yet accounted for time at residence in sensitivity analyses, and have not yet had the opportunity to conduct longitudinal analysis over repeat assessments (currently available for Stockport assessment centre only (~20,000 participants) in UK Biobank).

Air pollution and traffic noise estimates used in adjustment of models in this analysis used lower resolution data than those produced in this PhD project. Though the air pollution and noise exposure estimates integrated nationally for air pollution and noise are validated and adequate for adjustment, sensitivity analyses for UK Biobank London using the high resolution air pollution and noise exposures (developed in this PhD project) offer an opportunity to observe effect estimate changes using updated confounders that were expected to reduce misclassification.

Epidemiological analyses were less detailed than originally planned. This limitation was imposed by UK Biobank data access requests. In autumn 2019, verification and linkage into UK Biobank, and release of data with usable cardiovascular health outcomes (mortality) by UK Biobank was completed. CVD incidence data, which is of interest for this project, in the form of HES data, has recently been provided as ICD-10 code, with the corresponding ‘date’ column, meaning that CVD incidence can be assessed in future work. This ‘date’ column was not received in time to report CVD incidence in the thesis. Fortunately, however, UK Biobank had verified a self-reported item on prevalent disease at baseline using HES data, which allowed for the accurate exclusion of prevalent CVD (e.g., myocardial infarction, and including stroke) at baseline for survival analysis within the timeframe of my PhD.

The relationship between greenness and neighbourhood deprivation is likely to strongly confound analyses of greenspace and health, however, I adjusted for individual- and area-level measures of SES. The potential for residual confounding by neighbourhood SES and personal wealth remains problematic in greenspace analyses, as the composite IMD score and (self-reported) annual household income might not capture all aspects of deprivation related to health. The potential to use the OS MasterMap Greenspace exposure to assess the association of private gardens in a small buffer size surrounding participants' residential addresses (versus public greenspace) will be an important development for assessing residual confounding by SES. That is, if the association is driven primarily by private garden cover, this will raise questions as to whether greenspace effects are driven by the proposed exposure pathways or if greenspace in small buffers is a proxy for individual assets (affording the luxury of a garden in an urban area).

9 Conclusion

9.1 Overall summary

Findings from this PhD thesis, involving multiple exposure assessments of UK Biobank addresses, contribute to the current evidence on greenspace and address some important research gaps regarding the interrelationship of greenspace, air pollution, traffic noise and walkability. Exposures assessed in this PhD thesis are now integrated into the UK Biobank Data Showcase and are available to all registered researchers to use (see meta-data in Appendix).

From the *environmental pathway* exposure assessment, I found that high-resolution groundcover in a 100 m circular distance buffer had a typical (expected) proportional contribution of -10µg/m^3 in modelled concentrations of annual average NO_2 concentrations across London receptors. When combined with an air pollution dispersion model (ADMS-Urban) output variable, and a LUR variable that introduced a steeper concentration gradient adjacent to major roads, the *groundcover in a 100 m circular distance buffer* variable introduced flexibility in the model ('sinks' of NO_2 concentration). Combined, the final dispersion-LUR model (3 variables) resulted in good prediction of annual average NO_2 prediction at both background and roadside locations in Greater London according to validation. To my knowledge, this is the first study to explore the environmental pathway-specific relationship of greenspace and air pollution at the exposure assessment stage, as oppose to via confounding adjustment and/or effect mediation/modification in environmental epidemiological analyses. Further, through updating traffic noise estimates using high-resolution data inputs that were available for Greater London, I expect more accurate estimations (i.e. lower misclassification) of traffic noise, which should minimise confounding by traffic noise in epidemiological assessments of health (e.g., cardiovascular outcomes) with both air pollution and greenspace.

In my approach to assessing the *physiological* pathway, I make a case for acknowledging important known environmental correlates of physical activity in assessments of greenspace and physical activity. I demonstrated this by focusing on walking physical activity, in part due to the older demographic of the UK Biobank cohort who might be less likely to conduct regular moderate/vigorous exercise compared to a younger cohort, and in part based on walkability indices being particularly well documented and validated. Due to the two-stage process of, firstly, using UK Biobank addresses (X,Y-coordinates) for exposure assessment, before returning exposures and, secondly, receiving participant health and lifestyle variables, it was not possible to validate the walkability index before it was integrated into the UK Biobank Data Showcase. I had to rely, therefore, on literature to select conventional environmental correlates of walking in urban populations (i.e. walkability index components: population density, street connectivity and destination density). Once walkability scores were integrated into the UK Biobank database, I validated the walkability indices produced in this project using physical activity variables provided by UK Biobank, which have been used by other UK Biobank research projects (e.g., Cassidy et al., 2016), and have shown associations with cardiovascular disease prevalence (as reported at baseline). Both the walkability score with vegetation, and the walkability score without vegetation, integrated into the score was associated in the expected direction of effect with: a) physical activity levels; b) commute and non-commute transport modal choice; and c) achieving UK physical activity recommendations. However, the integration of vegetation into the walkability score attenuated associations compared to the conventional (without vegetation) walkability score. The attenuation indicates that adding greenness surrounding traversable routes does not increase associations with physical activity and mode choice in UK Biobank London participants. Similarly to the findings from the *environmental pathway*, the attenuation shown might not be driven by causal effects of vegetation (i.e. vegetation does not cause individuals to walk less), but due to low density of destinations, population and road/path junctions in green spaces. In light of these findings, some suggestions for further analyses to assess the relationship of walkability scores and vegetation cover in UK Biobank are provided in the future work section. It should be noted that during the physiological pathway exposure assessment, I developed reproducible methods (coded in SQL) for creating accurate transport network buffers over the relatively large UK Biobank London sample (*n* = 58,234), which has the potential to be scaled to national level analyses.

The ability to assess the interrelation of IMD neighbourhood deprivation, greenspace and other built environment variables (air pollution, traffic noise, and walkability) is one of the strengths of using UK Biobank in this PhD project; confounding by deprivation is an important consideration for cross-sectional greenspace research. Some studies have reported non-linear associations of built environmental variables, including greenspace and walkability (e.g., James et al., 2017), my findings were suggestive of a similar relationship. That is, the association of greenspace and walkability was negative and non-linear, and was best described using a natural cubic spline. This finding has important implications for epidemiological analysis. For example, non-linear confounding effects should be considered in future analyses.

In preliminary epidemiological analysis of non-injury and cardiovascular mortality, I found a protective association of total (all categories) OS MasterMap™ Greenspace cover surrounding UK Biobank England residential addresses and non-injury mortality after adjusting associations for sex, personal contextual covariates, and IMD neighbourhood-level deprivation only in the 100 m buffer size analysis. However, larger buffer sizes that were assessed (500 m and 1000 m) cannot be ruled out as analyses might have lacked power to detect a significant association in fully adjusted models. In the 100 m buffer size analysis, some evidence of an association with cardiovascular mortality was found, though results were inconsistent across quintiles, and results should be interpreted with caution.

9.2 Future research and policy implications

More research is needed on greenspace and cardiovascular health, with qualitative attribution of data, and assessment of mixtures of built environment exposures critical to furthering our understanding of causal mechanisms driving protective effects. My preliminary analyses on non-injury mortality and cardiovascular mortality in Chapter 6 were cross-sectional in design, and longitudinal studies investigating changes of premature mortality in relation to greenspace are required to confirm these findings.

In terms of qualitative attribution of data, greater depth of understanding can be achieved by assessing associations of specific types (functions) of greenspace and cardiovascular mortality in UK Biobank, For example, the effect of private gardens versus public parks and sports grounds might reveal pathway-specific effects related to specific types of greenspace via mediation analysis, which in turn would inform urban policy geared towards specific targets (e.g., increasing physical activity). Further, the quantification of 'private garden effect' might redirect future research through the lens of environmental equality should findings indicate that residual confounding for SES is in reality driving protective associations of greenspace and mortality. Given the growing evidence on greenspace and cardiovascular outcomes, including cardiovascular mortality (Gascon et al., 2016), it will be important to investigate *greenspace,* as oppose to *greenness* (NDVI), in future research, which will permit investigation of the functional role or specific attributes of greenspace that are associated with outcomes. Furthermore, moving away from NDVI might facilitate translation of greenspace research into practice, by improving specificity of findings (Rugel et al., 2017b). Studies on cumulative effects of greenspace exposure across the life-course will also be important in future research.

Analyses presented here on mortality should be supplemented with more cause specific outcomes (e.g., cerebrovascular mortality) in due course. Further, CVD incidence and CVD effect biomarkers – both of which are available in UK Biobank – should be used to assess if greenspace associations observed in mortality studies are consistent across earlier stages of cardiovascular health decline. UK Biobank recently (autumn 2019) released blood biochemical assays (effect biomarkers) associated with cardiovascular health, such as high-density lipoprotein (HDL)-cholesterol, low-density lipoprotein (LDL)-cholesterol and c-reactive protein, as well as inflammation markers, such as insulin-like growth factor 1 (IGF-1). The association of air pollution and noise with CVD might operate via systemic inflammation (Cai et al., 2017, Recio et al., 2016), and there is some limited evidence on biological mechanisms (e.g., inflammation) linking greenspace and health (Chaparro et al., 2018, Rook, 2013). Epidemiological analyses using detailed environmental exposures and CVD effect biomarker and inflammation markers have the potential to improve our understanding of underlying biological mechanisms.

Besides CVD outcomes, emerging studies in the last 2 to 3 years showed that other health outcomes such as cognitive performance, depression, obesity and diabetes are also related to greenspace, air pollution and noise exposures, which warrant further investigation. Added to this, studies in other populations (e.g. those in developing countries with more extreme levels of urban exposures) should be prioritised. The New Urban Agenda, adopted at Habitat III 2016 (United Nations, 2017), had a strong emphasis on urban environmental equality and social cohesion, and committed members to "promoting the creation and maintenance of well-connected and well distributed networks of open, multipurpose, safe, inclusive, accessible, green and quality public spaces" and "to improving […] physical and mental health, and household and ambient air quality, to reducing noise and promoting attractive and liveable cities". Health impact assessments have the potential to highlight synergies among existing policies and increase return on urban environmental interventions. Further, studies on the possible independent and joint effects of greenspace, walkability, air pollution and traffic noise exposure are needed to better clarify the current knowledge, and elucidate environment co-benefits of greenspace interventions for health in a range of contexts.

The identification of salutogenic and harmful effects of concomitant urban environmental exposures should be prioritised in future environmental epidemiological research. Exploration of 'urban exposome' (Andrianou and Makris, 2018) effects on health could be achieved via novel methods for analysing mixtures. For example, methods are in development for the assessment of correlated environmental exposure mixtures, such as Bayesian kernel machine regression (BKMR) (Bobb et al., 2018, Bobb et al., 2014), though further statistical advancements are required to assess mixtures in a survival analysis setting. However, the field is rapidly progressing, and a recent analysis of a mixture of exposures (metals) and cardiovascular incidence used a probit extension of BKMR (fitted iteratively across time points), and demonstrated a viable method for the assessment of correlated urban exposure mixtures and time-to-event outcomes (Domingo-Relloso et al., 2019).

The correlation of air pollution and traffic noise (Fecht et al., 2016), and the correlation of air pollution and walkability (Hankey et al., 2012), have implications for conducting analyses of cardiovascular epidemiology (i.e. biased effect estimates might be produced without due consideration of confounding), and have implications for urban interventions, which might inadvertently worsen one exposure while improving another. More attention should be focused on studying the interrelation of walkability and greenspace exposure in urban areas; greater levels of both exposures are hypothesised to enhance physical activity, with accompanying cardiovascular health benefits. However, walkability and greenspace showed a strong negative correlation at UK Biobank London addresses. Another study in the United States also showed an inverse association of walkability and greenness (NDVI) (James et al., 2017). It will be crucial to assess the association (if any) of greenspace and physical activity levels, and accompanying health benefits, along the physiological pathway, whilst taking into account known correlates of physical activity (e.g., for walkability: destination density, street connectivity and population density).

Due to the potential effect of weights of walkability score components (Shashank and Schuurman, 2019), which, in the walkability score that I developed in this PhD project were equally weighted, reducing the weight of greenness cover compared to destination density might optimize prediction of physical activity. Alternative, using exposure variables that I integrated into UK Biobank, a greenspace effect modification approach could be borrowed from the greenspace-air pollution literature (Crouse et al., 2019), whereby quintiles of greenspace cover in the network buffer could be used as an effect modifier of walkability and cardiovascular outcomes.

Further, tree cover can be relatively high in areas with high walkability (e.g., high density areas with many street trees); Sarkar et al. (2015b) found a positive association of street tree density and walking in Greater London, as well as an association of walkability metrics and walking, though did not mutually adjust models. An effect modification approach by total vegetation cover, ground cover and tree cover might disentangle effects of greenness, walkability and physical activity, and better inform strategies for increasing physical activity and greenspace exposure in urban populations, particularly in dense, space-limited urban cores.

Using findings form environmental pathway exposures assessment (dispersion-LUR modelling), the typical 'expected' reduction in annual average nitrogen dioxide concentrations estimated from the model developed in this project could be used in a counterfactual framework for all London addresses (e.g., an assessment of all postcodes), which would provide policy relevant findings. This could be used to assess, for example, expected health costs (health impact assessment) if all individuals were exposed to pollution reduced to the lowest quintile (bottom 20% of cover) of vegetation cover versus highest quintile (top 20%) across the all postcodes. No assumptions would be made regarding an increase in source emissions related to land use change, though the approach would provide an estimate that would build on findings from the vegetation and air pollution deposition model (courser 1 km x 1 km scale) described in the literature review (Jones et al., 2017b), by highlighting potentially indirect pathways linking vegetation and air quality. Layering of evidence on the contribution of vegetation to reducing air pollutants via measurement campaigns, and single street, city and national level exposure assessment modelling will contribute to understanding of the potential (and the limits) of using vegetation as a pollution mitigation tool for health protection.

The configuration of greenspace is important to air quality and cardiovascular health. For example, Shen and Lung (2016) explored the effect of the spatial arrangement of greenspace on cardiovascular mortality in Taiwan. Potential greenspace-health mediation pathways were assessed, including air pollution (e.g., NO_2, PM_{10}, and $PM_{2.5}$) and temperature. Results showed that fragmentation of greenspace and largest patch percentage (i.e. percentage cover attributable to the largest patch within a circular buffer) were associated with cardiovascular mortality, and were mediated by air pollution reduction, as oppose to temperature reduction. Moreover, fragmentation was associated with an increase in secondary air pollutants (e.g., ozone), demonstrating the deficiency of current greening policies that primarily focus on the ratio of built and greenspace coverage.

In 2016, the World Health Organisation released a review of evidence on greenspace and health (WHO Regional Office for Europe, 2016), which summarised various mechanistic pathways leading to health effects, and authors advocated "the implementation and evaluation of targeted, evidence-based green space interventions for the health promotion of urban residents". In London, the current Mayor, Sadiq Khan, has also pledged to plant two million trees across the city while in office (Greater London Authority, 2019). If geographical and qualitative information can be acquired, this large tree planting intervention offers a unique opportunity for evaluation. However, the effectiveness of greening schemes for health will depend on the research evidence, and translation to policy, on which they are founded, combined with the provisioning of resources for maintenance of green infrastructure into the future.

Improving access to greenspace in cities was included in the UN SDGs (SDG 11.7) (United Nations General Assembly, 2015). However, an important question to ask is: *who is benefiting from greenspace access?* That is, unintended risks and impacts associated with greenspace interventions in cities are emerging, such as green gentrification and displacement of vulnerable communities due to greening initiatives (Anguelovski et al., 2019). To avoid unintended consequences of translation of greenspace research into practice, it is critical for greenspace researchers to further explore the strong association of greenspace and neighbourhood deprivation that is pervasive across the literature.

In summary, from the literature review and modelling work conducted in this PhD, I suggest that interventions specifically targeting air pollution reduction should focus resources on reducing sources (e.g., diesel vehicles in the case of Greater London), rather than overstating the capacity of air pollution mitigation by vegetation. Retrofitting heavily polluted street canyons with greenspace, for example, can have a variable impact on concentrations, and can worsen pollution levels (Abhijith et al., 2017). Trees, for example, on the busy Marylebone Road, London, have been shown to augment air pollution concentrations overall by reducing air flow (Jeanjean et al., 2017). Whereas, vehicle emissions reduction strategies have been shown to be effective in reducing concentrations (Font et al., 2019). In planning of new urban areas, consideration of prevailing wind direction and ventilation or roads should be optimised to avoid pollutant trapping, by vegetation, or by other built environment features (e.g., buildings).

Importantly, *greenspace* is a catch-all term as highlighted by the OS MasterMap Greenspace exposure assessment; greenspace can encompass a variety of functional spaces, and the opportunity to visit local greenspace relies on quality of design, maintenance, safety and provision of activities (e.g., via community allotment groups, park exercise groups, etc.), which in turn relies on adequate government investment in environment and community groups. Greenspace psychosocial pathway benefits (not assessed in this PHD thesis) are suggested to be important for health (Houlden et al., 2019, Rugel et al., 2019), landscape architects and designers are well placed to integrate pathway-specific interventions, though should ensure that an intervention does not worsen exposures on non-focal pathways.

Together, the findings presented in this thesis highlight complex interrelationships of built environment exposures. I do not anticipate that the attenuation of odds of physical activity when vegetation cover is added to the walkability score versus when it is not, represents a causal relationship (i.e. individuals do not choose to actively avoid physical activity due to dislike of the vegetation cover surrounding the transport network), instead the attenuation might be due to the dampening of the importance of destination density in the score, which is an artefact of my assessment methodology. The inverse relationship of greenspace and walkability exposure highlighted in this thesis is an important take-away message. Policies that tackle walkability (density and adequate heterogeneity of environment to make walking a feasible transport option), while lowering the need for private vehicle use, would in turn provide many km^2 of open space, which would be freed up by releasing parking space and vehicle related infrastructure space. This newly freed space could be greened, and provide health benefits. Some major European cities such as Barcelona, Spain (‘Superblocks’ plan), and Paris, France (Mayor Anne Hidalgo’s 15-minute [walkable] city plan), are introducing holistic plans that ameliorate multiple environmental exposures by reducing car-reliance in urban cores, alongside greening strategies. In the UK, efforts are being made to tackle air pollution (e.g., via the ULEZ in London) and to promote greenspace (e.g. via Mayor Sadiq Khan’s National Park City plan for London), though researchers have recently criticised design of housing estates and neighbourhood developments across the country (Place

Alliance, 2020), in part due to car-centric planning. Walkability and people-centred planning will be important, alongside greenspace provisioning, to ameliorate specific exposures on the *environmental* and *physiological* pathways.

10 References

Abhijith, K. V., Kumar, P., Gallagher, J., Mcnabola, A., Baldauf, R., Pilla, F., Broderick, B., Di Sabatino, S. & Pulvirenti, B. 2017. Air pollution abatement performances of green infrastructure in open road and built-up street canyon environments – A review. *Atmospheric Environment,* 162**,** 71-86.

Adams, M. A., Frank, L. D., Schipperijn, J., Smith, G., Chapman, J., Christiansen, L. B., Coffee, N., Salvo, D., Du Toit, L., Dygryn, J., Hino, A. A., Lai, P. C., Mavoa, S., Pinzon, J. D., Van De Weghe, N., Cerin, E., Davey, R., Macfarlane, D., Owen, N. & Sallis, J. F. 2014a. International variation in neighborhood walkability, transit, and recreation environments using geographic information systems: the IPEN adult study. *Int J Health Geogr,* 13**,** 43.

Adams, M. A., Frank, L. D., Schipperijn, J., Smith, G., Chapman, J., Christiansen, L. B., Coffee, N., Salvo, D., Du Toit, L., Dygrýn, J., Hino, A. a. F., Lai, P.-C., Mavoa, S., Pinzón, J. D., Van De Weghe, N., Cerin, E., Davey, R., Macfarlane, D., Owen, N. & Sallis, J. F. 2014b. International variation in neighborhood walkability, transit, and recreation environments using geographic information systems: the IPEN adult study. *International Journal of Health Geographics,* 13**,** 43.

Allen, N. E., Sudlow, C., Peakman, T. & Collins, R. 2014. UK Biobank Data: Come and Get It. *Science Translational Medicine,* 6**,** 224ed4.

Andrianou, X. D. & Makris, K. C. 2018. The framework of urban exposome: Application of the exposome concept in urban health studies. *Science of The Total Environment,* 636**,** 963-967.

Anguelovski, I., Connolly, J. J. T., Pearsall, H., Shokry, G., Checker, M., Maantay, J., Gould, K., Lewis, T., Maroko, A. & Roberts, J. T. 2019. Opinion: Why green “climate gentrification” threatens poor and vulnerable populations. *Proceedings of the National Academy of Sciences,* 116**,** 26139-26143.

Annerstedt Van Den Bosch, M., Mudu, P., Uscila, V., Barrdahl, M., Kulinkina, A., Staatsen, B., Swart, W., Kruize, H., Zurlyte, I. & Egorov, A. I. 2016. Development of an urban green space indicator and the public health rationale. *Scandinavian Journal of Public Health,* 44**,** 159-167.

Astell-Burt, T., Feng, X. & Kolt, G. S. 2014. Is neighbourhood green space associated with a lower risk of Type 2 Diabetes Mellitus? Evidence from 267,072 Australians. *Diabetes Care.,* 37.

Atkinson, R. W., Carey, I. M., Kent, A. J., Van Staa, T. P., Anderson, H. R. & Cook, D. G. 2013. Long-term exposure to outdoor air pollution and incidence of cardiovascular diseases. *Epidemiology,* 24**,** 44-53.

Babisch, W. & Kamp, I. 2009. Exposure-response relationship of the association between aircraft noise and the risk of hypertension. *Noise Health,* 11**,** 161-8.

Babisch, W. 2014. Updated exposure-response relationship between road traffic noise and coronary heart diseases: a meta-analysis. *Noise Health,* 16**,** 1-9.

Babyak, M. A. 2004. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med,* 66**,** 411-21.

Basagaña, X., Rivera, M., Aguilera, I., Agis, D., Bouso, L., Elosua, R., Foraster, M., De Nazelle, A., Nieuwenhuijsen, M., Vila, J. & Künzli, N. 2012. Effect of the number of measurement sites on land use regression models in estimating local air pollution. *Atmospheric Environment,* 54**,** 634-642.

Basner, M., Babisch, W., Davis, A., Brink, M., Clark, C., Janssen, S. & Stansfeld, S. 2014. Auditory and non-auditory effects of noise on health. *Lancet,* 383**,** 1325-32.

Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., Tsai, M.-Y., Künzli, N., Schikowski, T., Marcon, A., Eriksen, K. T., Raaschou-Nielsen, O., Stephanou, E., Patelarou, E., Lanki, T., Yli-Tuomi, T., Declercq, C., Falq, G., Stempfelet, M., Birk, M., Cyrys, J., Von Klot, S., Nádor, G., Varró, M. J., Dėdelė, A., Gražulevičienė, R., Mölter, A., Lindley, S., Madsen, C., Cesaroni, G., Ranzi, A., Badaloni, C., Hoffmann, B., Nonnemacher, M., Krämer, U., Kuhlbusch, T., Cirach, M., De Nazelle, A., Nieuwenhuijsen, M., Bellander, T., Korek, M., Olsson, D., Strömgren, M., Dons, E., Jerrett, M., Fischer, P., Wang, M., Brunekreef, B. & De Hoogh, K. 2013. Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe – The ESCAPE project. *Atmospheric Environment,* 72**,** 10-23.

Beelen, R., Stafoggia, M., Raaschou-Nielsen, O., Andersen, Z. J., Xun, W. W., Katsouyanni, K., Dimakopoulou, K., Brunekreef, B., Weinmayr, G., Hoffmann, B., Wolf, K., Samoli, E., Houthuijs, D., Nieuwenhuijsen, M., Oudin, A., Forsberg, B., Olsson, D., Salomaa, V., Lanki, T., Yli-Tuomi, T., Oftedal, B., Aamodt, G., Nafstad, P., De Faire, U., Pedersen, N. L., Ostenson, C. G., Fratiglioni, L., Penell, J., Korek, M., Pyko, A., Eriksen, K. T., Tjonneland, A., Becker, T., Eeftens, M., Bots, M., Meliefste, K., Wang, M., Bueno-De-Mesquita, B., Sugiri, D., Kramer, U., Heinrich, J., De Hoogh, K., Key, T., Peters, A., Cyrys, J., Concin, H., Nagel, G., Ineichen, A., Schaffner, E., Probst-Hensch, N., Dratva, J., Ducret-Stich, R., Vilier, A., Clavel-Chapelon, F., Stempfelet, M., Grioni, S., Krogh, V., Tsai, M. Y., Marcon, A., Ricceri, F., Sacerdote, C., Galassi, C., Migliore, E., Ranzi, A., Cesaroni, G., Badaloni, C., Forastiere, F., Tamayo, I., Amiano, P., Dorronsoro, M., Katsoulis, M., Trichopoulou, A., Vineis, P. & Hoek, G. 2014. Long-term exposure to air pollution and cardiovascular mortality: an analysis of 22 European cohorts. *Epidemiology,* 25**,** 368-78.

Besson, H., Brage, S., Jakes, R. W., Ekelund, U. & Wareham, N. J. 2009. Estimating physical activity energy expenditure, sedentary time, and physical activity intensity by self-report in adults. *The American Journal of Clinical Nutrition,* 91**,** 106-114.

Bhf 2019. UK Factsheet, August 2019. UK: British Heart Foundation.

Bixby, H., Hodgson, S., Fortunato, L., Hansell, A. & Fecht, D. 2015. Associations between Green Space and Health in English Cities: An Ecological, Cross-Sectional Study. *PLoS ONE,* 10**,** e0119495.

Bobb, J. F., Valeri, L., Claus Henn, B., Christiani, D. C., Wright, R. O., Mazumdar, M., Godleski, J. J. & Coull, B. A. 2014. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics,* 16**,** 493-508.

Bobb, J. F., Claus Henn, B., Valeri, L. & Coull, B. A. 2018. Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. *Environmental Health,* 17**,** 67.

Bowler, D. E., Buyung-Ali, L. M., Knight, T. M. & Pullin, A. S. 2010. A systematic review of evidence for the added benefits to health of exposure to natural environments. *BMC Public Health,* 10**,** 1-10.

Brauer, M. 2015. Air pollution, stroke, and anxiety. *BMJ : British Medical Journal,* 350**,** h1510.

Braun, L. M., Rodríguez, D. A., Evenson, K. R., Hirsch, J. A., Moore, K. A. & Diez Roux, A. V. 2016. Walkability and cardiometabolic risk factors: Cross-sectional and longitudinal associations from the Multi-Ethnic Study of Atherosclerosis. *Health & Place,* 39**,** 9-17.

Brook, R. D., Rajagopalan, S., Pope, C. A., Brook, J. R., Bhatnagar, A. & Diez-Roux, A. V. 2010a. Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association. *Circulation,* 121.

Brook, R. D., Rajagopalan, S., Pope, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R. V., Mittleman, M. A., Peters, A., Siscovick, D., Smith, S. C., Whitsel, L. & Kaufman, J. D. 2010b. Particulate Matter Air Pollution and Cardiovascular Disease. *An Update to the Scientific Statement From the American Heart Association,* 121**,** 2331-2378.

Brugge, D., Lane, K., Padró-Martínez, L. T., Stewart, A., Hoesterey, K., Weiss, D., Wang, D. D., Levy, J. I., Patton, A. P., Zamore, W. & Mwamburi, M. 2013. Highway proximity associated with cardiovascular disease risk: the influence of individual-level confounders and exposure misclassification. *Environmental Health,* 12**,** 84.

Cai, Y., Hansell, A. L., Blangiardo, M., Burton, P. R., De Hoogh, K., Doiron, D., Fortier, I., Gulliver, J., Hveem, K., Mbatchou, S., Morley, D. W., Stolk, R. P., Zijlema, W. L., Elliott, P. & Hodgson, S. 2017. Long-term exposure to road traffic noise, ambient air pollution, and cardiovascular risk factors in the HUNT and lifelines cohorts. *European Heart Journal,* 38**,** 2290-2296.

Cai, Y., Hodgson, S., Blangiardo, M., Gulliver, J., Morley, D., Fecht, D., Vienneau, D., De Hoogh, K., Key, T., Hveem, K., Elliott, P. & Hansell, A. L. 2018a. Road traffic noise, air pollution and incident cardiovascular disease: A joint analysis of the HUNT, EPIC-Oxford and UK Biobank cohorts. *Environ Int,* 114**,** 191-201.

Cai, Y., Hodgson, S., Blangiardo, M., Gulliver, J., Morley, D., Fecht, D., Vienneau, D., De Hoogh, K., Key, T., Hveem, K., Elliott, P. & Hansell, A. L. 2018b. Road traffic noise, air pollution and incident cardiovascular disease: A joint analysis of the HUNT, EPIC-Oxford and UK Biobank cohorts. *Environment International,* 114**,** 191-201.

Cassidy, S., Chau, J. Y., Catt, M., Bauman, A. & Trenell, M. I. 2016. Cross-sectional study of diet, physical activity, television viewing and sleep duration in 233 110 adults from the UK Biobank; the behavioural phenotype of cardiovascular disease and type 2 diabetes. *BMJ Open,* 6**,** e010038.

Celis-Morales, C. A., Lyall, D. M., Welsh, P., Anderson, J., Steell, L., Guo, Y., Maldonado, R., Mackay, D. F., Pell, J. P., Sattar, N. & Gill, J. M. R. 2017. Association between active commuting and incident cardiovascular disease, cancer, and mortality: prospective cohort study. *BMJ,* 357.

Cesaroni, G., Forastiere, F., Stafoggia, M., Andersen, Z. J., Badaloni, C., Beelen, R., Caracciolo, B., De Faire, U., Erbel, R., Eriksen, K. T., Fratiglioni, L., Galassi, C., Hampel, R., Heier, M., Hennig, F., Hilding, A., Hoffmann, B., Houthuijs, D., Jöckel, K.-H., Korek, M., Lanki, T., Leander, K., Magnusson, P. K. E., Migliore, E., Ostenson, C.-G., Overvad, K., Pedersen, N. L., J, J. P., Penell, J., Pershagen, G., Pyko, A., Raaschou-Nielsen, O., Ranzi, A., Ricceri, F., Sacerdote, C., Salomaa, V., Swart, W., Turunen, A. W., Vineis, P., Weinmayr, G., Wolf, K., De Hoogh, K., Hoek, G., Brunekreef, B. & Peters, A. 2014. Long term exposure to ambient air pollution and incidence of acute coronary events: prospective cohort study and meta-analysis in 11 European cohorts from the ESCAPE Project. *BMJ : British Medical Journal,* 348.

Chaparro, M. P., Benzeval, M., Richardson, E. & Mitchell, R. 2018. Neighborhood deprivation and biomarkers of health in Britain: the mediating role of the physical environment. *BMC Public Health,* 18**,** 801.

Chastin, S. F. M., De Craemer, M., De Cocker, K., Powell, L., Van Cauwenberg, J., Dall, P., Hamer, M. & Stamatakis, E. 2019. How does light-intensity physical activity associate with adult cardiometabolic health and mortality? Systematic review with meta-analysis of experimental and observational studies. *British Journal of Sports Medicine,* 53**,** 370-376.

Clark, C., Sbihi, H., Tamburic, L., Brauer, M., Frank, L. D. & Davies, H. W. 2017. Association of Long-Term Exposure to Transportation Noise and Traffic-Related Air Pollution with the Incidence of Diabetes: A Prospective Cohort Study. *Environmental Health Perspectives,* 125**,** 087025.

Coffee, N. T., Howard, N., Paquet, C., Hugo, G. & Daniel, M. 2013. Is walkability associated with a lower cardiometabolic risk? *Health & Place,* 21**,** 163-169.

Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L., Pope, C. A., Shin, H., Straif, K., Shaddick, G., Thomas, M., Van Dingenen, R., Van Donkelaar, A., Vos, T., Murray, C. J. L. & Forouzanfar, M. H. 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *The Lancet,* 389**,** 1907-1918.

Collins, R. 2012. What makes UK Biobank special? *The Lancet,* 379**,** 1173-1174.

Coombes, E., Jones, A. P. & Hillsdon, M. 2010. The relationship of physical activity and overweight to objectively measured green space accessibility and use. *Soc Sci Med,* 70.

Creatore, M. I., Glazier, R. H., Moineddin, R., Fazli, G. S., Johns, A., Gozdyra, P., Matheson, F. I., Kaufman-Shriqui, V., Rosella, L. C., Manuel, D. G. & Booth, G. L. 2016. Association of Neighborhood Walkability With Change in Overweight, Obesity, and Diabetes. *JAMA,* 315**,** 2211-2220.

Crouse, D. L., Pinault, L., Balram, A., Hystad, P., Peters, P. A., Chen, H., Van Donkelaar, A., Martin, R. V., Ménard, R., Robichaud, A. & Villeneuve, P. J. 2017. Urban greenness and mortality in Canada's largest cities: a national cohort study. *The Lancet Planetary Health,* 1**,** e289-e297.

Crouse, D. L., Pinault, L., Balram, A., Brauer, M., Burnett, R. T., Martin, R. V., Van Donkelaar, A., Villeneuve, P. J. & Weichenthal, S. 2019. Complex relationships between greenness, air pollution, and mortality in a population-based Canadian cohort. *Environment International,* 128**,** 292-300.

Dadvand, P., De Nazelle, A., Figueras, F., Basagaña, X., Su, J., Amoly, E., Jerrett, M., Vrijheid, M., Sunyer, J. & Nieuwenhuijsen, M. J. 2012a. Green space, health inequality and pregnancy. *Environment International,* 40**,** 110-115.

Dadvand, P., De Nazelle, A., Triguero-Mas, M., Schembari, A., Cirach, M. & Amoly, E. 2012b. Surrounding greenness and exposure to air pollution during pregnancy: an analysis of personal monitoring data. *Environ Health Perspect,* 120.

Dadvand, P., De Nazelle, A., Triguero-Mas, M., Schembari, A., Cirach, M., Amoly, E., Figueras, F., Basagana, X., Ostro, B. & Nieuwenhuijsen, M. 2012c. Surrounding greenness and exposure to air pollution during pregnancy: an analysis of personal monitoring data. *Environ Health Perspect,* 120**,** 1286-90.

Dadvand, P. & Nieuwenhuijsen, M. 2018. Greenspace and Health (Chapter 20). *In:* NIEUWENHUIJSEN, M. & KHREIS, H. (eds.) *Integrating Human Health into Urban and Transport Planning: A Framework.* Switzerland: Springer International Publishing.

Daiber, A., Kröller-Schön, S., Frenis, K., Oelze, M., Kalinovic, S., Vujacic-Mirski, K., Kuntic, M., Bayo Jimenez, M. T., Helmstädter, J., Steven, S., Korac, B. & Münzel, T. 2019. Environmental noise induces the release of stress hormones and inflammatory signaling molecules leading to oxidative stress and vascular dysfunction—Signatures of the internal exposome. *BioFactors,* 45**,** 495-506.

Dalton, A. M., Jones, A. P., Sharp, S. J., Cooper, A. J. M., Griffin, S. & Wareham, N. J. 2016. Residential neighbourhood greenspace is associated with reduced risk of incident diabetes in older people: a prospective cohort study. *BMC Public Health,* 16**,** 1171.

Davies, H. W., Vlaanderen, J. J., Henderson, S. B. & Brauer, M. 2009. Correlation between co-exposures to noise and air pollution from traffic sources. *Occupational and Environmental Medicine,* 66**,** 347.

Davies, S., Burns, H., Jewell, T. & Mcbride, M. 2011. Start active, stay active: a report on physical activity from the four home countries. *Chief Medical Officers,* 16306**,** 1-62.

De Keijzer, C., Agis, D., Ambrós, A., Arévalo, G., Baldasano, J. M., Bande, S., Barrera-Gómez, J., Benach, J., Cirach, M., Dadvand, P., Ghigo, S., Martinez-Solanas, È., Nieuwenhuijsen, M., Cadum, E. & Basagaña, X. The association of air pollution and greenness with mortality and life expectancy in Spain: A small-area study. *Environment International*.

De Keijzer, C., Agis, D., Ambrós, A., Arévalo, G., Baldasano, J. M., Bande, S., Barrera-Gómez, J., Benach, J., Cirach, M., Dadvand, P., Ghigo, S., Martinez-Solanas, È., Nieuwenhuijsen, M., Cadum, E. & Basagaña, X. 2017. The association of air pollution and greenness with mortality and life expectancy in Spain: A small-area study. *Environment International,* 99**,** 170-176.

De Keijzer, C., Basagaña, X., Tonne, C., Valentín, A., Alonso, J., Antó, J. M., Nieuwenhuijsen, M. J., Kivimäki, M., Singh-Manoux, A., Sunyer, J. & Dadvand, P. 2019. Long-term exposure to greenspace and metabolic syndrome: A Whitehall II study. *Environmental Pollution,* 255**,** 113231.

Defra 2019. Air Pollution in the UK 2018; Compliance Assessment Summary. UK: Department for the Environment, Fisheries and Rural Affairs.

Delgado-Rodríguez, M. & Llorca, J. 2004. Bias. *Journal of Epidemiology and Community Health,* 58**,** 635-641.

Department for Communities and Local Government 2016. The English Indices of Deprivation 2015 – Frequently Asked Questions (FAQs).

Department for Transport. 2013. *Table TRA0307. Traffic Distribution by Time of Day on All Roads in Great Britain, 2009* [Online]. Available: https://www.gov.uk/government/statistical-data-sets/tra03-motor-vehicle-flow [Accessed].

Doiron, D., Burton, P., Marcon, Y., Gaye, A., Wolffenbuttel, B. H. R., Perola, M., Stolk, R. P., Foco, L., Minelli, C., Waldenberger, M., Holle, R., Kvaløy, K., Hillege, H. L., Tassé, A.-M., Ferretti, V. & Fortier, I. 2013. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. *Emerging Themes in Epidemiology,* 10**,** 12.

Domingo-Relloso, A., Grau-Perez, M., Briongos-Figuero, L., Gomez-Ariza, J. L., Garcia-Barrera, T., Dueñas-Laita, A., Bobb, J. F., Chaves, F. J., Kioumourtzoglou, M.-A., Navas-Acien, A., Redon-Mas, J., Martin-Escudero, J. C. &

Tellez-Plaza, M. 2019. The association of urine metals and metal mixtures with cardiovascular incidence in an adult population from Spain: the Hortega Follow-Up Study. *International Journal of Epidemiology,* 48**,** 1839-1849.

Donovan, G. H., Michael, Y. L., Gatziolis, D., Prestemon, J. P. & Whitsel, E. A. 2015. Is tree loss associated with cardiovascular-disease risk in the Women's Health Initiative? A natural experiment. *Health & Place,* 36**,** 1-7.

Ebrahim, S. & Davey Smith, G. 2013. Commentary: Should we always deliberately be non-representative? *International Journal of Epidemiology,* 42**,** 1022-1026.

Eeftens, M., Tsai, M.-Y., Ampe, C., Anwander, B., Beelen, R., Bellander, T., Cesaroni, G., Cirach, M., Cyrys, J., De Hoogh, K., De Nazelle, A., De Vocht, F., Declercq, C., Dėdelė, A., Eriksen, K., Galassi, C., Gražulevičienė, R., Grivas, G., Heinrich, J., Hoffmann, B., Iakovides, M., Ineichen, A., Katsouyanni, K., Korek, M., Krämer, U., Kuhlbusch, T., Lanki, T., Madsen, C., Meliefste, K., Mölter, A., Mosler, G., Nieuwenhuijsen, M., Oldenwening, M., Pennanen, A., Probst-Hensch, N., Quass, U., Raaschou-Nielsen, O., Ranzi, A., Stephanou, E., Sugiri, D., Udvardy, O., Vasкövi, É., Weinmayr, G., Brunekreef, B. & Hoek, G. 2012. Spatial variation of PM2.5, PM10, PM2.5 absorbance and PMcoarse concentrations between and within 20 European study areas and the relationship with NO2 – Results of the ESCAPE project. *Atmospheric Environment,* 62**,** 303-317.

Eeftens, M., Beekhuizen, J., Beelen, R., Wang, M., Vermeulen, R., Brunekreef, B., Huss, A. & Hoek, G. 2013. Quantifying urban street configuration for improvements in air pollution models. *Atmospheric Environment,* 72**,** 1-9.

Ekkel, E. D. & De Vries, S. 2017. Nearby green space and human health: Evaluating accessibility metrics. *Landscape and Urban Planning,* 157**,** 214-220.

Elliott, P., Biobank, O. B. O. U., Peakman, T. C. & Biobank, O. B. O. U. 2008. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *International Journal of Epidemiology,* 37**,** 234-244.

English Nature Research Reports 2008. Accessible Natural Green Space Standards in Towns and Cities: A Review and Toolkit for their Implementation. 2 ed. Peterborough.

Fecht, D., Hansell, A. L., Morley, D., Dajnak, D., Vienneau, D., Beevers, S., Toledano, M. B., Kelly, F. J., Anderson, H. R. & Gulliver, J. 2016. Spatial and temporal associations of road traffic noise and air pollution in London: Implications for epidemiological studies. *Environment International,* 88**,** 235-242.

Flint, E. & Cummins, S. 2016. Active commuting and obesity in mid-life: cross-sectional, observational evidence from UK Biobank. *The Lancet Diabetes & Endocrinology,* 4**,** 420-435.

Flint, E., Webb, E. & Cummins, S. 2016. Change in commute mode and body-mass index: prospective, longitudinal evidence from UK Biobank. *The Lancet Public Health,* 1**,** e46-e55.

Floud, S., Blangiardo, M., Clark, C., De Hoogh, K., Babisch, W., Houthuijs, D., Swart, W., Pershagen, G., Katsouyanni, K., Velonakis, M., Vigna-Taglianti, F., Cadum, E. & Hansell, A. L. 2013. Exposure to aircraft and road traffic noise and associations with heart disease and stroke in six European countries: a cross-sectional study. *Environmental Health,* 12**,** 89.

Fong, K. C., Hart, J. E. & James, P. 2018. A Review of Epidemiologic Studies on Greenness and Health: Updated Literature Through 2017. *Current Environmental Health Reports*.

Font, A., Guiseppin, L., Blangiardo, M., Ghersi, V. & Fuller, G. W. 2019. A tale of two cities: is air pollution improving in Paris and London? *Environmental Pollution,* 249**,** 1-12.

Foraster, M., Deltell, A., Basagaña, X., Medina-Ramón, M., Aguilera, I. & Bouso, L. 2011. Local determinants of road traffic noise levels versus determinants of air pollution levels in a Mediterranean city. *Environ Res,* 111.

Forsyth, A. 2015. What is a walkable place? The walkability debate in urban design. *URBAN DESIGN International,* 20**,** 274-292.

Frank, L. D., Saelens, B. E., Powell, K. E. & Chapman, J. E. 2007. Stepping towards causation: Do built environments or neighborhood and travel preferences explain physical activity, driving, and obesity? *Soc Sci Med,* 65.

Frank, L. D., Sallis, J. F., Saelens, B. E., Leary, L., Cain, K., Conway, T. L. & Hess, P. M. 2010. The development of a walkability index: application to the Neighborhood Quality of Life Study. *Br J Sports Med,* 44**,** 924-33.

Frank, L. D., Fox, E. H., Ulmer, J. M., Chapman, J. E., Kershaw, S. E., Sallis, J. F., Conway, T. L., Cerin, E., Cain, K. L., Adams, M. A., Smith, G. R., Hinckson, E., Mavoa, S., Christiansen, L. B., Hino, A. a. F., Lopes, A. a. S. & Schipperijn, J. 2017. International comparison of observation-specific spatial buffers: maximizing the ability to estimate physical activity. *International Journal of Health Geographics,* 16**,** 4.

Franklin, B. A., Brook, R. & Arden Pope, C. 2015. Air Pollution and Cardiovascular Disease. *Current Problems in Cardiology,* 40**,** 207-238.

Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R. & Allen, N. E. 2017. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology,* 186**,** 1026-1034.

Füzéki, E., Engeroff, T. & Banzer, W. 2017. Health Benefits of Light-Intensity Physical Activity: A Systematic Review of Accelerometer Data of the National Health and Nutrition Examination Survey (NHANES). *Sports Medicine,* 47**,** 1769-1793.

Gan, W. Q., Davies, H. W., Koehoorn, M. & Brauer, M. 2012. Association of Long-term Exposure to Community Noise and Traffic-related Air Pollution With Coronary Heart Disease Mortality. *American Journal of Epidemiology,* 175**,** 898-906.

Ganna, A. & Ingelsson, E. 2015. 5 year mortality predictors in 498103 UK Biobank participants: a prospective population-based study. *The Lancet,* 386**,** 533-540.

Gascon, M., Triguero-Mas, M., Martínez, D., Dadvand, P., Forns, J. & Plasència, A. 2015. Mental health benefits of long-term exposure to residential green and blue spaces: a systematic review. *Int J Environ Res Public Health,* 12.

Gascon, M., Triguero-Mas, M., Martínez, D., Dadvand, P., Rojas-Rueda, D., Plasència, A. & Nieuwenhuijsen, M. J. 2016. Residential green spaces and mortality: A systematic review. *Environment International,* 86**,** 60-67.

Gehl, J. 2010. *Cities for people,* Washington DC, Island press.

Giles-Corti, B., Gunn, L., Hooper, P., Boulange, C., Diomedi, B. Z., Pettit, C. & Foster, S. 2019. Built Environment and Physical Activity. *In:* NIEUWENHUIJSEN, M. & KHREIS, H. (eds.) *Integrating Human Health into Urban and Transport Planning: A Framework.* Cham: Springer International Publishing.

Grasser, G., Van Dyck, D., Titze, S. & Stronegger, W. J. 2017. A European perspective on GIS-based walkability and active modes of transport. *European Journal of Public Health,* 27**,** 145-151.

Greater London Authority 2019. Mayor hosts summit as London becoming world’s first National Park City. London: Mayor of London - London Assembly.

Gulliver, J. & De Hoogh, K. 2015. Environmental exposure assessment: modelling air pollution concentrations. *Oxford Textbook of Global Public Health* Oxford, UK: Oxford University Press.

Gulliver, J., Morley, D., Vienneau, D., Fabbri, F., Bell, M., Goodman, P., Beevers, S., Dajnak, D., J Kelly, F. & Fecht, D. 2015. Development of an open-source road traffic noise model for exposure assessment. *Environmental Modelling & Software,* 74**,** 183-193.

Halonen, J. I., Hansell, A. L., Gulliver, J., Morley, D., Blangiardo, M., Fecht, D., Toledano, M. B., Beevers, S. D., Anderson, H. R., Kelly, F. J. & Tonne, C. 2015. Road traffic noise is associated with increased cardiovascular morbidity and mortality and all-cause mortality in London. *European Heart Journal,* 36**,** 2653-2661.

Hankey, S., Marshall, J. D. & Brauer, M. 2012. Health Impacts of the Built Environment: Within-Urban Variability in Physical Inactivity, Air Pollution, and Ischemic Heart Disease Mortality. *Environmental Health Perspectives,* 120**,** 247-253.

Haskell, W. L., Lee, I.-M., Pate, R. R., Powell, K. E., Blair, S. N., Franklin, B. A., Macera, C. A., Heath, G. W., Thompson, P. D. & Bauman, A. 2007. Physical Activity and Public Health. Updated Recommendation for Adults From the American College of Sports Medicine and the American Heart Association. *Circulation*.

Higgs, G., Fry, R. & Langford, M. 2012. Investigating the Implications of Using Alternative GIS-Based Techniques to Measure Accessibility to Green Space. *Environment and Planning B: Planning and Design,* 39**,** 326-343.

Hillsdon, M., Panter, J., Foster, C. & Jones, A. 2006. The relationship between access and quality of urban green space with population physical activity. *Public Health,* 120**,** 1127-1132.

Hoek, G., Krishnan, R. M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B. & Kaufman, J. D. 2013. Long-term air pollution exposure and cardio- respiratory mortality: a review. *Environmental Health,* 12**,** 43.

Hupin, D., Roche, F., Gremeaux, V., Chatard, J.-C., Oriol, M., Gaspoz, J.-M., Barthélémy, J.-C. & Edouard, P. 2015. Even a low-dose of moderate-to-vigorous physical activity reduces mortality by 22% in adults aged ⩾60 years: a systematic review and meta-analysis. *British Journal of Sports Medicine,* 49**,** 1262-1267.

Hystad, P., Davies, H. W., Frank, L., Van Loon, J., Gehring, U. & Tamburic, L. 2014. Residential greenness and birth outcomes: evaluating the influence of spatially correlated built-environment factors. *Environ Health Perspect,* 122.

James, P., Banay, R. F., Hart, J. E. & Laden, F. 2015. A Review of the Health Benefits of Greenness. *Current Epidemiology Reports,* 2**,** 131-142.

James, P., Hart Jaime, E., Banay Rachel, F. & Laden, F. 2016a. Exposure to Greenness and Mortality in a Nationwide Prospective Cohort Study of Women. *Environmental Health Perspectives,* 124**,** 1344-1352.

James, P., Hart, J. E., Banay, R. F. & Laden, F. 2016b. Exposure to Greenness and Mortality in a Nationwide Prospective Cohort Study of Women. *Environmental Health Perspectives,* 124**,** 1344-1352.

James, P., Kioumourtzoglou, M.-A., Hart, J. E., Banay, R. F., Kloog, I. & Laden, F. 2017. Interrelationships Between Walkability, Air Pollution, Greenness, and Body Mass Index. *Epidemiology,* 28**,** 780-788.

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., Abdollahpour, I., Abdulkader, R. S., Abebe, Z., Abera, S. F., Abil, O. Z., Abraha, H. N., Abu-Raddad, L. J., Abu-Rmeileh, N. M. E., Accrombessi, M. M. K., Acharya, D., Acharya, P., Ackerman, I. N., Adamu, A. A., Adebayo, O. M., Adekanmbi, V., Adetokunboh, O. O., Adib, M. G., Adsuar, J. C., Afanvi, K. A., Afarideh, M., Afshin, A., Agarwal, G., Agesa, K. M., Aggarwal, R., Aghayan, S. A., Agrawal, S., Ahmadi, A., Ahmadi, M., Ahmadieh, H., Ahmed, M. B., Aichour, A. N., Aichour, I., Aichour, M. T. E., Akinyemiju, T., Akseer, N., Al-Aly, Z., Al-Eyadhy, A., Al-Mekhlafi, H. M., Al-Raddadi, R. M., Alahdab, F., Alam, K., Alam, T., Alashi, A., Alavian, S. M., Alene, K. A., Alijanzadeh, M., Alizadeh-Navaei, R., Aljunid, S. M., Alkerwi, A. A., Alla, F., Allebeck, P., Alouani, M. M. L., Altirkawi, K., Alvis-Guzman, N., Amare, A. T., Aminde, L. N., Ammar, W., Amoako, Y. A., Anber, N. H., Andrei, C. L., Androudi, S., Animut, M. D., Anjomshoa, M., Ansha, M. G., Antonio, C. a. T., Anwari, P., Arabloo, J., Arauz, A., Aremu, O., Ariani, F., Armoon, B., Ärnlöv, J., Arora, A., Artaman, A., Aryal, K. K., Asayesh, H., Asghar, R. J., Ataro, Z., Atre, S. R., Ausloos, M., Avila-Burgos, L., Avokpaho, E. F. G. A., Awasthi, A., Ayala Quintanilla, B. P., Ayer, R., Azzopardi, P. S., Babazadeh, A., Badali, H., Badawi, A., Bali, A. G., et al. 2018. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990&#x2013;2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet,* 392**,** 1789-1858.

Janhäll, S. 2015. Review on urban vegetation and particle air pollution – Deposition and dispersion. *Atmospheric Environment,* 105**,** 130-137.

Jefferis, B. J., Parsons, T. J., Sartini, C., Ash, S., Lennon, L. T., Papacosta, O., Morris, R. W., Wannamethee, S. G., Lee, I.-M. & Whincup, P. H. 2019. Objectively measured physical activity, sedentary behaviour and all-cause mortality in older men: does volume of activity matter more than pattern of accumulation? *British Journal of Sports Medicine,* 53**,** 1013-1020.

Jones, L., Vieno, M., Morton, D., Hall, J., Carnell, E., Nemitz, E., Beck, R., Reis, S., Pritchard, N. & Hayes, F. 2017. Developing estimates for the valuation of air pollution removal in ecosystem accounts. Final report for Office of National Statistics by Centre for Ecology and Hydrology.

Jorgensen, L. J., Ellis, G. D. & Ruddell, E. 2013. Fear Perceptions in Public Parks: Interactions of Environmental Concealment, the Presence of People Recreating, and Gender. *Environment and Behavior,* 45**,** 803-820.

Kaufman, J. D., Adar, S. D., Barr, R. G., Budoff, M., Burke, G. L., Curl, C. L., Daviglus, M. L., Roux, A. V. D., Gassett, A. J., Jacobs, D. R., Kronmal, R., Larson, T. V., Navas-Acien, A., Olives, C., Sampson, P. D., Sheppard, L., Siscovick, D. S., Stein, J. H., Szpiro, A. A. & Watson, K. E. 2016. Association between air pollution and coronary artery calcification within six metropolitan areas in the USA (the Multi-Ethnic Study of Atherosclerosis and Air Pollution): a longitudinal cohort study. *The Lancet,* 388**,** 696-704.

Kcl. 2019. *TRAFFIC (project)* [Online]. London: King's College London. Available: https://www.kcl.ac.uk/lsm/research/divisions/aes/research/erg/research-projects/traffic/index [Accessed 20th December 2019].

Kephalopoulos, S., Paviotti, M., Anfosso-Ledee, F., Van Maercke, D., Shilton, S. & Jones, N. 2014. Advances in the development of common noise assessment methods in Europe: The CNOSSOS-EU framework for strategic environmental noise mapping. *Sci Total Environ,* 482-483**,** 400-10.

Kim, S.-Y. & Song, I. 2017. National-scale exposure prediction for long-term concentrations of particulate matter and nitrogen dioxide in South Korea. *Environmental Pollution,* 226**,** 21-29.

Klompmaker, J. O., Hoek, G., Bloemsma, L. D., Gehring, U., Strak, M., Wijga, A. H., Van Den Brink, C., Brunekreef, B., Lebret, E. & Janssen, N. a. H. 2018. Green space definition affects associations of green space with overweight and physical activity. *Environmental Research,* 160**,** 531-540.

Klompmaker, J. O., Janssen, N. a. H., Bloemsma, L. D., Gehring, U., Wijga, A. H., Van Den Brink, C., Lebret, E., Brunekreef, B. & Hoek, G. 2019a. Associations of Combined Exposures to Surrounding Green, Air Pollution, and Road Traffic Noise with Cardiometabolic Diseases. *Environmental Health Perspectives,* 127**,** 087003.

Klompmaker, J. O., Janssen, N. a. H., Bloemsma, L. D., Gehring, U., Wijga Alet, H., Van Den Brink, C., Lebret, E., Brunekreef, B. & Hoek, G. 2019b. Associations of Combined Exposures to Surrounding Green, Air Pollution, and Road Traffic Noise with Cardiometabolic Diseases. *Environmental Health Perspectives,* 127**,** 087003.

Kwan, M.-P. & Weber, J. 2003. Individual Accessibility Revisited: Implications for Geographical Analysis in the Twenty-first Century. *Geographical Analysis,* 35**,** 341-353.

Labib, S. M., Lindley, S. & Huck, J. J. 2019. Spatial dimensions of the influence of urban green-blue spaces on human health: A systematic review. *Environmental Research***,** 108869.

Lachowycz, K. & Jones, A. P. 2011. Greenspace and obesity: a systematic review of the evidence. *Obes Rev.,* 12.

Lachowycz, K. & Jones, A. P. 2014a. Does walking explain associations between access to greenspace and lower mortality? *Social Science & Medicine,* 107**,** 9-17.

Lachowycz, K. & Jones, A. P. 2014b. Does walking explain associations between access to greenspace and lower mortality? *Social Science & Medicine (1982),* 107**,** 9-17.

Lamonte, M. J., Lewis, C. E., Buchner, D. M., Evenson, K. R., Rillamas‐Sun, E., Di, C., Lee, I. M., Bellettiere, J., Stefanick, M. L., Eaton, C. B., Howard, B. V., Bird, C., Lacroix, A. Z., Rossouw, J., Ludlam, S., Burwen, D., Mcgowan, J., Ford, L., Geller, N., Anderson, G., Prentice, R., Kooperberg, C., Manson, J. E., Jackson, R., Thomson, C. A., Wactawski‐Wende, J., Limacher, M., Wallace, R., Kuller, L. & Shumaker, S. 2017. Both Light Intensity and Moderate&#x2010;to&#x2010;Vigorous Physical Activity Measured by Accelerometry Are Favorably Associated With Cardiometabolic Risk Factors in Older Women: The Objective Physical Activity and Cardiovascular Health (OPACH) Study. *Journal of the American Heart Association,* 6**,** e007064.

Laqn 2016. Monitoring sites classification [Online]. Available at: http://www.londonair.org.uk/london/asp/publicdetails.asp?region=0 [Last access: 21/05/2017].

Laverty, A. A., Mindell, J. S., Webb, E. A. & Millett, C. 2013. Active Travel to Work and Cardiovascular Risk Factors in the United Kingdom. *American Journal of Preventive Medicine,* 45**,** 282-288.

Li, J. & Siegrist, J. 2012. Physical Activity and Risk of Cardiovascular Disease—A Meta-Analysis of Prospective Cohort Studies. *International Journal of Environmental Research and Public Health,* 9**,** 391.

Lyall, D. M., Celis-Morales, C., Ward, J., Iliodromiti, S., Anderson, J. J., Gill, J. M. R., Smith, D. J., Ntuk, U. E., Mackay, D. F., Holmes, M. V., Sattar, N. & Pell, J. P. 2017. Association of Body Mass Index With Cardiometabolic Disease in the UK Biobank: A Mendelian Randomization Study. *JAMA Cardiology,* 2**,** 882-889.

Maas, J., Verheij, R. A., Spreeuwenberg, P. & Groenewegen, P. P. 2008. Physical activity as a possible mechanism behind the relationship between green space and health: a multilevel analysis. *BMC Publ Health,* 8.

Markevych, I., Schoierer, J., Hartig, T., Chudnovsky, A., Hystad, P., Dzhambov, A. M., De Vries, S., Triguero-Mas, M., Brauer, M., Nieuwenhuijsen, M. J., Lupp, G., Richardson, E. A., Astell-Burt, T., Dimitrova, D., Feng, X., Sadeh, M., Standl, M., Heinrich, J. & Fuertes, E. 2017. Exploring pathways linking greenspace to health: Theoretical and methodological guidance. *Environmental Research,* 158**,** 301-317.

Mason, K. E., Pearce, N. & Cummins, S. 2018. Associations between fast food and physical activity environments and adiposity in mid-life: cross-sectional, observational evidence from UK Biobank. *The Lancet Public Health,* 3**,** e24-e33.

Mcgill, R., Tukey, J. W. & Larsen, W. A. 1978. Variations of box plots. *The American Statistician,* 32**,** 12-16.

Mitchell, R. & Popham, F. 2008a. Effect of exposure to natural environment on health inequalities: an observational population study. *The Lancet,* 372.

Mitchell, R. & Popham, F. 2008b. Effect of exposure to natural environment on health inequalities: an observational population study. *The Lancet,* 372**,** 1655-1660.

Mitchell, R., Astell-Burt, T. & Richardson, E. A. 2011. A comparison of green space indicators for epidemiological research. *Journal of Epidemiology and Community Health,* 65**,** 853-858.

Morley, D. W., De Hoogh, K., Fecht, D., Fabbri, F., Bell, M., Goodman, P. S., Elliott, P., Hodgson, S., Hansell, A. L. & Gulliver, J. 2015. International scale implementation of the CNOSSOS-EU road traffic noise prediction model for epidemiological studies. *Environmental Pollution,* 206**,** 332-341.

Morley, D. W. & Gulliver, J. 2016a. Methods to improve traffic flow and noise exposure estimation on minor roads. *Environ Pollut,* 216**,** 746-54.

Morley, D. W. & Gulliver, J. 2016b. Methods to improve traffic flow and noise exposure estimation on minor roads. *Environmental Pollution,* 216**,** 746-754.

Mueller, N., Rojas-Rueda, D., Basagana, X., Cirach, M., Cole-Hunter, T., Dadvand, P., Donaire-Gonzalez, D., Foraster, M., Gascon, M., Martinez, D., Tonne, C., Triguero-Mas, M., Valentin, A. & Nieuwenhuijsen, M. 2017a. Urban and Transport Planning Related Exposures and Mortality: A Health Impact Assessment for Cities. *Environ Health Perspect,* 125**,** 89-96.

Mueller, N., Rojas-Rueda, D., Basagaña, X., Cirach, M., Cole-Hunter, T., Dadvand, P., Donaire-Gonzalez, D., Foraster, M., Gascon, M., Martinez, D., Tonne, C., Triguero-Mas, M., Valentín, A. & Nieuwenhuijsen, M. 2017b. Urban and Transport Planning Related Exposures and Mortality: A Health Impact Assessment for Cities. *Environmental Health Perspectives,* 125**,** 89-96.

Münzel, T., Gori, T., Babisch, W. & Basner, M. 2014. Cardiovascular effects of environmental noise exposure. *European Heart Journal,* 35**,** 829-836.

Münzel, T., Sørensen, M., Gori, T., Schmidt, F. P., Rao, X., Brook, J., Chen, L. C., Brook, R. D. & Rajagopalan, S. 2017. Environmental stressors and cardio-metabolic disease: part I–epidemiologic evidence supporting a role for noise and air pollution and effects of mitigation strategies. *European Heart Journal,* 38**,** 550-556.

Mytton, O. T., Townsend, N., Rutter, H. & Foster, C. 2012. Green space and physical activity: An observational study using Health Survey for England data. *Health & Place,* 18**,** 1034-1041.

Natural England 2010. Nature Nearby. Accessible Natural Greenspace Guidance. Peterborough.

Newby, D. E., Mannucci, P. M., Tell, G. S., Baccarelli, A. A., Brook, R. D., Donaldson, K., Forastiere, F., Franchini, M., Franco, O. H., Graham, I., Hoek, G., Hoffmann, B., Hoylaerts, M. F., Kunzli, N., Mills, N., Pekkanen, J., Peters, A., Piepoli, M. F., Rajagopalan, S. & Storey, R. F. 2015. Expert position paper on air pollution and cardiovascular disease. *Eur Heart J,* 36**,** 83-93b.

Nieuwenhuijsen, M. J. 2016. Urban and transport planning, environmental exposures and health-new concepts, methods and tools to improve health in cities. *Environmental Health,* 15**,** S38.

Nieuwenhuijsen, M. J., Khreis, H., Triguero-Mas, M., Gascon, M. & Dadvand, P. 2017a. Fifty Shades of Green. *Epidemiology,* 28**,** 63-71.

Nieuwenhuijsen, M. J., Khreis, H., Triguero-Mas, M., Gascon, M. & Dadvand, P. 2017b. Fifty Shades of Green: Pathway to Healthy Urban Living. *Epidemiology,* 28.

Nieuwenhuijsen, M. J. 2018. Influence of urban and transport planning and the city environment on cardiovascular disease. *Nature Reviews Cardiology,* 15**,** 432-438.

Nowak, D. J., Crane, D. E., Stevens, J. C., Hoehn, R. E., Walton, J. T. & Bond, J. 2008. A ground-based method of assessing urban forest structure and ecosystem services. *Aboriculture & Urban Forestry. 34 (6): 347-358.,* 34.

Nowak, D. J., Hirabayashi, S., Bodine, A. & Greenfield, E. 2014. Tree and forest effects on air quality and human health in the United States. *Environmental Pollution,* 193**,** 119-129.

Obe, R. O., Hsu, L. S. & Sherman, G. E. 2017. *pgRouting: A Practical Guide,* Chugiak AK, Locate Press.

Oliver, L. N., Schuurman, N. & Hall, A. W. 2007. Comparing circular and network buffers to examine the influence of land use on walking for leisure and errands. *International journal of health geographics,* 6**,** 41-41.

Ordnance Survey. 2013. *RE: Urban paths ITN layer difficulties, 11.10.2013* [Online]. Available: https://www.ordnancesurvey.co.uk/forums/media/get/userfiles/de/f1d8a62865819d19d4ef8df0c683b7.docx [Accessed 1st February 2018].

Orstad, S. L., Mcdonough, M. H., James, P., Klenosky, D. B., Laden, F., Mattson, M. & Troped, P. J. 2018. Neighborhood walkability and physical activity among older women: Tests of mediation by environmental perceptions and moderation by depressive symptoms. *Preventive Medicine,* 116**,** 60-67.

Parliamentary Office of Science and Technology 2016. Green Space and Health. London: Houses of Parliament.

Pazoki, R., Dehghan, A., Evangelou, E., Warren, H., Gao, H., Caulfield, M., Elliott, P. & Tzoulaki, I. 2018. Genetic Predisposition to High Blood Pressure and Lifestyle Factors. *Circulation,* 137**,** 653-661.

Pimpin, L., Retat, L., Fecht, D., De Preux, L., Sassi, F., Gulliver, J., Belloni, A., Ferguson, B., Corbould, E., Jaccard, A. & Webber, L. 2018. Estimating the costs of air pollution to the National Health Service and social care: An assessment and forecast up to 2035. *PLOS Medicine,* 15**,** e1002602.

Prüss-Üstün, A., Wolf, J., Corvalán, C., Bos, R. & Neira, M. 2016. *Preventing disease through healthy environments: A global assessment of the burden of disease from environmental risks,* Geneva, WHO.

Rao, M., George, L. A., Rosenstiel, T. N., Shandas, V. & Dinno, A. 2014. Assessing the relationship among urban trees, nitrogen dioxide, and respiratory health. *Environmental Pollution,* 194**,** 96-104.

Recio, A., Linares, C., Banegas, J. R. & Díaz, J. 2016. Road traffic noise effects on cardiovascular, respiratory, and metabolic health: An integrative model of biological mechanisms. *Environmental Research,* 146**,** 359-370.

Richardson, E. A., Mitchell, R., Hartig, T., De Vries, S., Astell-Burt, T. & Frumkin, H. 2012. Green cities and health: a question of scale? *Journal of Epidemiology and Community Health,* 66**,** 160.

Rigolon, A. 2016. A complex landscape of inequity in access to urban parks: A literature review. *Landscape and Urban Planning,* 153**,** 160-169.

Rojas-Rueda, D., Nieuwenhuijsen, M. J., Gascon, M., Perez-Leon, D. & Mudu, P. 2019. Green spaces and mortality: a systematic review and meta-analysis of cohort studies. *The Lancet Planetary Health,* 3**,** e469-e477.

Rook, G. A. 2013. Regulation of the immune system by biodiversity from the natural environment: An ecosystem service essential to health. *Proceedings of the National Academy of Sciences,* 110**,** 18360-18367.

Rothman, K. J., Gallacher, J. E. & Hatch, E. E. 2013. Why representativeness should be avoided. *International Journal of Epidemiology,* 42**,** 1012-1014.

Rugel, E. J., Henderson, S. B., Carpiano, R. M. & Brauer, M. 2017a. Beyond the Normalized Difference Vegetation Index (NDVI): Developing a Natural Space Index for population-level health research. *Environmental Research,* 159**,** 474-483.

Rugel, E. J., Henderson, S. B., Carpiano, R. M. & Brauer, M. 2017b. Beyond the Normalized Difference Vegetation Index (NDVI): Developing a Natural Space Index for population-level health research. *Environmental Research,* 159**,** 474-483.

Sallis, J. F. 2016. New evidence for the role of transportation in health. *The Lancet Public Health,* 1**,** e38-e39.

Sarkar, C., Webster, C. & Gallacher, J. 2015a. UK Biobank Urban Morphometric Platform (UKBUMP) – a nationwide resource for evidence-based healthy city planning and public health interventions. *Annals of GIS,* 21**,** 135-148.

Sarkar, C., Webster, C., Pryor, M., Tang, D., Melbourne, S., Zhang, X. & Jianzheng, L. 2015b. Exploring associations between urban green, street design and walking: Results from the Greater London boroughs. *Landscape and Urban Planning,* 143**,** 112-125.

Sarkar, C. 2017. Residential greenness and adiposity: Findings from the UK Biobank. *Environment International,* 106**,** 1-10.

Seo, S., Choi, S., Kim, K., Kim, S. M. & Park, S. M. 2019. Association between urban green space and the risk of cardiovascular disease: A longitudinal study in seven Korean metropolitan areas. *Environment International,* 125**,** 51-57.

Shashank, A. & Schuurman, N. 2019. Unpacking walkability indices and their inherent assumptions. *Health & Place,* 55**,** 145-154.

Shen, Y.-S. & Lung, S.-C. C. 2016. Can green structure reduce the mortality of cardiovascular diseases? *Science of The Total Environment,* 566-567**,** 1159-1167.

Sørensen, M., Lühdorf, P., Ketzel, M., Andersen, Z. J., Tjønneland, A., Overvad, K. & Raaschou-Nielsen, O. 2014. Combined effects of road traffic noise and ambient air pollution in relation to risk for stroke? *Environmental Research,* 133**,** 49-55.

Stockton, J. C., Duke-Williams, O., Stamatakis, E., Mindell, J. S., Brunner, E. J. & Shelton, N. J. 2016a. Development of a novel walkability index for London, United Kingdom: cross-sectional application to the Whitehall II Study. *BMC Public Health,* 16**,** 416.

Stockton, J. C., Duke-Williams, O., Stamatakis, E., Mindell, J. S., Brunner, E. J. & Shelton, N. J. 2016b. Development of a novel walkability index for London, United Kingdom: cross-sectional application to the Whitehall II Study. *BMC Public Health,* 16**,** 416.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T. & Collins, R. 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine,* 12**,** e1001779.

Swanson, J. M. 2012. The UK Biobank and selection bias. *The Lancet,* 380**,** 110.

Tamosiunas, A., Grazuleviciene, R., Luksiene, D., Dedele, A., Reklaitiene, R. & Baceviciene, M. 2014. Accessibility and use of urban green spaces, and cardiovascular health: findings from a Kaunas cohort study. *Environ Health,* 13.

Tang, R., Blangiardo, M. & Gulliver, J. 2013. Using Building Heights and Street Configuration to Enhance Intraurban PM10, NOX, and NO2 Land Use Regression Models. *Environmental Science & Technology,* 47**,** 11643-11650.

Tétreault, L.-F., Perron, S. & Smargiassi, A. 2013. Cardiovascular health, traffic-related air pollution and noise: are associations mutually confounded? A systematic review. *International Journal of Public Health,* 58**,** 649-666.

Thiébaut, A. C. M. & Bénichou, J. 2004. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Statistics in Medicine,* 23**,** 3803-3820.

Tong, Z., Baldauf, R. W., Isakov, V., Deshmukh, P. & Max Zhang, K. 2016. Roadside vegetation barrier designs to mitigate near-road air pollution impacts. *Science of The Total Environment,* 541**,** 920-927.

Tsao, C. W. & Vasan, R. S. 2015. Cohort Profile: The Framingham Heart Study (FHS): overview of milestones in cardiovascular epidemiology. *International Journal of Epidemiology,* 44**,** 1800-1813.

Tudor-Locke, C. & Bassett, D. R. 2004. How Many Steps/Day Are Enough? *Sports Medicine,* 34**,** 1-8.

Twohig-Bennett, C. & Jones, A. 2018. The health benefits of the great outdoors: A systematic review and meta-analysis of greenspace exposure and health outcomes. *Environmental Research,* 166**,** 628-637.

Uk Chief Medical Officers 2019. UK Chief Medical Officers' Physical Activity Guidelines UK: Department of Health and Social Care.

Un Desa 2019. World Urbanization Prospects: The 2018 Revision. New York: Population Division of the United Nations Department of Economic and Social Affairs.

United Nations 2017. New Urban Agenda. New York, NY, USA: United Nations Habitat III.

United Nations General Assembly 2015. Transforming our world: the 2030 Agenda for Sustainable Development.

Van Den Berg, M., Wendel-Vos, W., Van Poppel, M., Kemper, H., Van Mechelen, W. & Maas, J. 2015. Health benefits of green spaces in the living environment: A systematic review of epidemiological studies. *Urban Forestry & Urban Greening,* 14**,** 806-816.

Van Kempen, E. E. M. M., Kruize, H., Boshuizen, H. C., Ameling, C. B., Staatsen, B. a. M. & De Hollander, A. E. M. 2002. The association between noise exposure and blood pressure and ischemic heart disease: a meta-analysis. *Environmental Health Perspectives,* 110**,** 307-317.

Vienneau, D., Schindler, C., Perez, L., Probst-Hensch, N. & Röösli, M. 2015. The relationship between transportation noise exposure and ischemic heart disease: A meta-analysis. *Environmental Research,* 138**,** 372-380.

Vienneau, D., De Hoogh, K., Faeh, D., Kaufmann, M., Wunderli, J. M. & Röösli, M. 2017. More than clean air and tranquillity: Residential green is independently associated with decreasing mortality. *Environment International,* 108**,** 176-184.

Villeneuve, P. J., Jerrett, M., G. Su, J., Burnett, R. T., Chen, H., Wheeler, A. J. & Goldberg, M. S. 2012a. A cohort study relating urban green space with mortality in Ontario, Canada. *Environmental Research,* 115**,** 51-58.

Villeneuve, P. J., Jerrett, M., Su, J. G., Burnett, R., Chen, H. & Wheeler, A. J. 2012b. A cohort study relating urban green space with mortality in Ontario, Canada. *Environ Res,* 115.

Villeneuve, P. J., Jerrett, M., Su, J. G., Burnett, R. T., Chen, H., Wheeler, A. J. & Goldberg, M. S. 2012c. A cohort study relating urban green space with mortality in Ontario, Canada. *Environ Res,* 115.

Villeneuve, P. J., Jerrett, M., Su, J. G., Weichenthal, S. & Sandler, D. P. 2018. Association of residential greenness with obesity and physical activity in a US cohort of women. *Environmental Research,* 160**,** 372-384.

Vos, P. E. J., Maiheu, B., Vankerkom, J. & Janssen, S. 2013. Improving local air quality in cities: To tree or not to tree? *Environmental Pollution,* 183**,** 113-122.

Wang, C., Li, Q. & Wang, Z.-H. 2018. Quantifying the impact of urban trees on passive pollutant dispersion using a coupled large-eddy simulation–Lagrangian stochastic model. *Building and Environment,* 145**,** 33-49.

Wania, A., Bruse, M., Blond, N. & Weber, C. 2012. Analysing the influence of different street vegetation on traffic-induced particle dispersion using microscale simulations. *Journal of Environmental Management,* 94**,** 91-101.

Wei, T. & Simko, V. 2017. *R package "corrplot": Visualization of a Correlation Matrix* [Online]. Available: https://github.com/taiyun/corrplot [Accessed].

Who 2018. WHO methods and data sources for global burden of disease estimates 2000-2016. Geneva: Department of Information, Evidence and Research, World Health Organisation.

Who Regional Office for Europe 2016. Urban green spaces and health. A review of evidence. . Copenhagen: WHO Regional Office for Europe.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.

Wood, S. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)***,** 3–36.

Wood, S. 2017. *Generalized Additive Models: An Introduction with R*, Chapman and Hall/CRC.

World Health Organization. 1999. *Guidelines for community noise.* [Online]. Geneva: World Health Organization. Available: http://apps.who.int/iris/handle/10665/66217 [Accessed 10 September 2018].

World Health Organization 2011. Burden of disease from environmental noise: Quantification of healthy life years lost in Europe.

Yitshak-Sade, M., James, P., Kloog, I., Hart, J. E., Schwartz, J. D., Laden, F., Lane, K. J., Fabian, M. P., Fong, K. C. & Zanobetti, A. 2019. Neighborhood Greenness Attenuates the Adverse Effect of PM2.5 on Cardiovascular Mortality in Neighborhoods of Lower Socioeconomic Status. *International Journal of Environmental Research and Public Health***,** 814.

Zijlema, W. L., Stasinska, A., Blake, D., Dirgawati, M., Flicker, L., Yeap, B. B., Golledge, J., Hankey, G. J., Nieuwenhuijsen, M. & Heyworth, J. 2019. The longitudinal association between natural outdoor environments and mortality in 9218 older men from Perth, Western Australia. *Environment International,* 125**,** 430-436.

Appendix

UK Biobank Application 4236
Documentation of exposure variables linked to geocoded participants' residential addresses

Charlotte Roscoe
MRC Centre for Environment and Health
Imperial College London
c.roscoe16@imperial.ac.uk

Daniela Fecht
MRC Centre for Environment and Health
Imperial College London
d.fecht@imperial.ac.uk

John Gulliver
Centre for Environmental Health and Sustainability
University of Leicester
jg435@leicester.ac.uk

The exposure variables provided are the result of a Medical Research Council funded PhD project. Before using any of the exposure estimates in any study we advise that the planned analyses are discussed with the contacts above. The authors of the exposure work should be acknowledged in any publication or included as co-authors.

All derived variables are subject to the citation requirements of underlying data. Several bulk downloads of Ordnance Survey data were provided for this project by Edina via Digimap (https://digimap.edina.ac.uk/). Re-use of derived variables in any study should cite Edina and the underlying data accordingly as detailed below.

Data received from UK Biobank
Geocoded UK Biobank participants' residential addresses (XY coordinates) were supplied for exposure assignment. As per the latest (16/10/2018) participant withdrawal update, we removed 73 participants from the data.

Data returned to UK Biobank
Data file: UKB_ app4236.csv (Number of records: 339,078)

This file contains internal UK Biobank IDs and environmental exposure variables assessed at participants' residential addresses. The returned records have at least one of the following environmental exposures assigned to addresses: 1) percentage cover of greenspace in circular distance buffers, categorised by greenspace function; 2) percentage cover of greenspace in circular distance buffers, categorised by vegetation height; 3) percentage cover of greenspace in a road/path network buffer, categorised by vegetation height; 4) attributes of the road/path network, including walkability; 5) modelled outdoor nitrogen dioxide concentrations; and 6) modelled outdoor traffic noise levels. Methods and models used to derive these variables are detailed below.

Environmental exposure could not been assigned to some participants due to their residential address falling outside the modelling domain or restricted geographical coverage of input data. These cases have been assigned as missing data (-999).

Percentage cover of greenspace in circular distance buffers, categorised by greenspace function

Input data description
Ordnance Survey (OS) MasterMap Greenspace Layer gives a comprehensive overview of greenspace in urban areas of the UK. The dataset comprises of topographical areas, as released in OS MasterMap Topography Layer, with additional greenspace attributes to describe their function. It includes both publicly accessible and private greenspace, sports facilities and natural environment features.

Data update cycle: October 2017

Accuracy of topography for urban data capture: 1.0 m (1:1250 scale)

Coverage
Coverage of the OS Mastermap Greenspace data is defined as the following in the OS Mastermap Greenspace product guide: "For England and Wales, urban areas are included where they are greater than 6km^2. For Scotland, urban areas are defined as those with a population in excess of 500 people. This is based on data provided by the National Records of Scotland. In Scotland a buffer of 500m has been added to the urban extents to define the product coverage. Where a site crosses the boundary of an area, all features within the site are included in MasterMap Greenspace, even where these are outside the urban area. This applies up to a limit of 1,500m from the urban boundary." More information can be found online: https://www.ordnancesurvey.co.uk/docs/product-guides/osmm-greenspace-product-guide.pdf

Exposure assessment
OS Mastermap Greenspace cover was assigned in multiple circular distance buffers (100m, 300m, 500m, 1000m and 1500m) around each participant's residential address geocode (i.e. X/Y coordinate). The 100m and 300m buffer were selected to represent the near-home environment, the other larger 'neighbourhood' buffer sizes were selected for comparability with greenness data (Normalized Difference Vegetation Index; NDVI) that has previously been integrated into UK Biobank (Sarkar et al., 2015a). Data are provided as percentage cover of the total buffer area.

Greenspace cover categorised by greenspace primary function (see Table 1) were assigned to all eligible addresses. The eligibility of an address for exposure assessment was dependent on the OS Mastermap Greenspace Layer extent. This data "covers all major urban areas in Great Britain" (see *Coverage* section). All addresses outside of the OS Mastermap Greenspace Layer extent were excluded from assessment (-999).

To ensure that the greenspace data fully covered the circular distance buffer around each participant's address, only addresses located inwards of the built-up area boundary at the specified buffer distance for each analysis were included in each assessment. For example, geocoded addresses located at least -100m from the built-up area boundary were used for the 100m analysis, geocoded address located at least -300m from the built-up area boundary were used for the 300m analysis, etc. This approach maximised the number of geocoded addresses eligible for analysis for each assessment.

In Table 1, 'Field' denotes column headers in the exposure data, which follow the naming convention, *code_buffersize*. For example, the percentage cover of 'Allotments Or Community Growing Spaces' within a 500m circular distance buffer of an address can be found in field, 'allot_500'. Depending on research focus, these greenspace variables from different categories within the same buffer can be grouped together as required, or totalled to provide total greenspace cover.

Table 1. Ordnance Survey (OS) Mastermap Greenspace primary function classifications (18 categories) with descriptions. Full descriptions of the data set can be found in the technical specification:https://www.ordnancesurvey.co.uk/docs/technical-specifications/os-mastermap-greenspace-layer-technical-specification.pdf

Field	**Greenspace function**	**Description**
privg	Private Garden	Areas of land normally enclosed and associated with private residences and reserved for private use.
sport	Other Sports Facility	Land used for other sports not specifically described by other categories. Includes facilities for sports spectating (e.g. stadiums) as well as participation.
golf	Golf Course	A large area of land that is specially prepared for playing golf.
tenni	Tennis Court	A specially prepared area intended for playing tennis.
amtra	Amenity - Transport	Landscaped areas providing visual amenity or separating different buildings or land uses for environmental, visual or safety reasons when related to a transport function, such as a road, or within a transport hub.
cem	Cemetery	Areas of land associated with burial areas or crematoriums.
nat	Natural	Land use areas with no other function but with Form attribute of woodland, open semi-natural, open water, beach or foreshore.
luc	Land Use Changing	Areas of land that are currently under development or awaiting redevelopment.
plays	Play Space	Areas providing safe and accessible opportunities for children's play, usually linked to housing areas or parks and containing purpose-built equipment. Not captured if within schools or paid-for tourist attractions.
playf	Playing Field	Large, flat areas of grass or specially designed surfaces, generally with marked pitches, used primarily for outdoor sports, i.e. football, rugby, cricket.
bowl	Bowling Green	A specially prepared area intended for playing bowls.
camp	Camping or Caravan Park	An organised area of ground designated for tents or caravans, intended for temporary occupation by holidaymakers.
allot	Allotments or Community Growing Spaces	Areas of land for growing fruit, vegetables, and other plants, either in individual allotments or as a community activity. Produce is for the grower's own

		consumption and not primarily for commercial activity.
amres	Amenity - Residential or Business	Landscaped areas providing visual amenity or separating different buildings or land uses for environmental, visual or safety reasons. Where the area is better described by another category this will be used in preference (e.g. playing field, public park, play space).
inst	Institutional Grounds	Areas of land normally enclosed and associated with institutions. Grounds may be reserved for private use or have restricted access. Includes Universities, Hospitals, Nursing homes, Emergency Services, Prisons, Military Sites, Government and Community Buildings providing public services, Libraries, Museums, Zoos and Theatres.
pubpg	Public Park or Garden	Areas of land normally enclosed, designed, constructed, managed and maintained as a public park or garden. These normally have a defined perimeter and free public access, and generally sit within or adjacent to urban areas. Access is granted for a wide range of uses and not usually restricted to paths or tracks within the area. May include areas with managed facilities such as benches and flowerbeds, and more natural areas.
sch	School Grounds	Areas of land normally enclosed and associated with a school and primarily reserved for their use.
relig	Religious Grounds	Areas of land associated with churches and other places of worship.

Indices of Multiple Deprivation

Participants' residential address locations (XY coordinates) were assigned to the Lower-layer Super Output Area (LSOA; England and Wales) or Data Zones (Scotland) in which they were contained (using PostGIS). This allowed Index of Multiple Deprivation (IMD) deciles (standardised across LSOAs/DataZones for each country) to be assigned – via LSOA or Data Zone code attribution – to all participants that had a least one greenspace variable assigned (from Section 1). As they are calculated via a country-specific methodology, IMD for each country is not directly comparable. For more explanation see the documentation for the country-specific indexes:

England 2015 IMD: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015
Scotland 2016 IMD: https://www2.gov.scot/Topics/Statistics/SIMD
Wales 2014 IMD: https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation/WIMD-2014

Percentage cover of greenspace in circular distance buffers, categorised by vegetation height

Input data description

The Greater London Vegetation Layer was produced by the GeoInformation Group (https://www.geomni.co.uk/) on behalf of the Greater London Authority – Urban Greening Unit. The data was derived from aerial imagery, Light Detection and Ranging (LiDAR) data and manually digitised tree data from UKMap. LiDAR is an aerial mapping system, which uses lasers to establish the distance between an aeroplane and land. Remote sensors detect reflected light to create accurate three-dimensional images of the Earth's surface. The vegetation data is highly detailed, capturing raster tree canopy and ground cover data at a resolution of 2.5m, however, does not discriminate between public and private land. In the data, tree canopy (tree_) is vegetation cover over or equal to 2.5m in height, ground cover (ground_) is below 2.5m in height.

Data release year: 2015

Resolution: 2.5m x 2.5m

Coverage

To ensure that the vegetation data fully covered the circular distance buffers around each participant's address geocode, only addresses located 1500m inwards of the Greater London Vegetation Layer boundary were included in exposure assessment.

Exposure assessment

Vegetation cover data, categorised as tree canopy or ground cover, was assessed for each participant's geocoded residential address in multiple circular distance buffers (50m, 100m, 500m, 1000m and 1500m). Percentage cover of vegetation categorised as tree canopy (tree_) and ground cover (ground_) was assigned to all eligible addresses.

The eligibility of an address for exposure assessment was dependent on the data extent. The data covers the Greater London Authority – Greater London boundary (available here: https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london). All addresses outside of the data extent were excluded from assessment (-999).

Table 2 descries the added fields. The column headers in the added data follow the naming convention, *covertype_buffersize* i.e. tree canopy within a 500m circular distance buffer are added in column, 'tree_500'.

Table 2. Field descriptions for Greater London Vegetation within a circular buffer in UKB_ app4236.csv

Field	Description
tree	Percentage cover of vegetation ≥2.5 m height in circular buffer (%)
ground	Percentage cover of vegetation <2.5 m height in circular buffer (%)

Percentage cover of greenspace in a road/path network buffer, categorised by vegetation height

Road and cycle/footpath network distance buffers were created by tracing the road/path network a given distance from the residential address location, and adding a small buffer (e.g., 50 m) around the traced line to create a polygon. These buffers capture a more accurate representation of the area that can be traversed – e.g., when walking – than a circular buffer and are thought to better capture the spatial attributes of the neighbourhood that may influence physical activity (Giles-Corti et al., 2019, Oliver et al., 2007).

To address this, a road and path network distance of 1000m was selected to represent a 15-minute walk, which adequately represents the near walkable neighbourhood. This distance has previously been used in UK Biobank network buffer assessment (Sarkar et al., 2015a). The walkable road and path network was created by routing Ordnance Survey (OS) Integrated Transport Network with OS Urban Paths Theme extension, to capture roads with vehicle access, as well as footpaths, cycle paths and other pedestrianised routes. This integration approach has previously been applied in area-level walkability analyses in London (Stockton et al., 2016a).

Road/path networks were produced using pgRouting, an extension of the open source database Postgres, via the function pgr_withPointsDD (https://docs.pgrouting.org/2.2/en/src/withPoints/doc/pgr_withPointsDD.html#pgr-withpointsdd). This function captures all vertices in the line network within a set distance of the residential address (≤1000m), and the edges leading to them. Routing via this function begins at the nearest point on the route network to the residential address (fractional start edges are permitted). Accuracy of networks was improved by extending the network to exactly 1000m in length via addition of fractional edges/lines to the terminal vertex of the routed network.

In agreement with other studies on walkability and network attributes (Frank et al., 2017, Oliver et al., 2007), a buffer width of 50m (25m either side of the traced network line) was deemed sufficient to capture walkable network attributes (e.g., vegetation).

Coverage

To ensure that the vegetation data fully covered the line-based network, distance buffers around each participant's address geocode (only addresses located inwards (-1000m) of the Greater London Vegetation Layer boundary) were included in exposure assessment.

Exposure assessment

The vegetation cover data, categorised as tree canopy (LNB_tree) or ground cover (LNB_ground), was intersected with the road/path network distance buffer for each participant and is provided as percentage cover (see Table 3). More information on the vegetation data is provided in Section 2. The total area of the buffer was also added (LNB_area).

Table 3. Field descriptions for Greater London Vegetation within the road/path network buffer in UKB_ app4236.csv

Field	Description
LNB_area	Area of line network buffer (m^2)
LNB_tree	Percentage cover of vegetation ≥2.5m height in line network buffer (%)

LNB_ground	Percentage cover of vegetation <2.5m height in line network buffer (%)

Attributes of the road/path network buffer (including walkability)

Attributes for the road/path network buffer include i) total route length within 1000m for residential address, ii) pedestrianised route length within 1000m, and iii) walkability score. The walkability score (z-score) combines three commonly used metrics to assess walkability: population density, three-way intersection count, and destination count. The walkability z-scores provide an intra-sample comparison of walkability for UK Biobank participants residing in Greater London, as oppose to a cohort wide or national comparison.

We used Office for National Statistics (ONS) data on 2011 postcode locations and headcount totals to calculate the population density of each network buffer. The population in the network buffer was totalled and divided by the network buffer area to provide a population density estimate.

The number of junctions within each road/path network buffer was used to represent walkable network connectivity. Junctions were derived from the intersection points of the underlying line-based network (see Section 3). All junction types – e.g., road-to-road, road-to-footpath and footpath-to-footpath – were used. To capture true network junctions, as opposed to line segment breaks, only junctions that connected a minimum of three line segments were included.

Potential destinations that a participant may walk to within their corresponding network buffer were summed. Destinations considered include retail, facilities and services, for example, newsagents, bus stops, underground stations, sports facilities, restaurants, banks, libraries, etc. Please contact us for a detailed list of included destination points. The destination points were derived from Ordnance Survey Points of Interest (POI). More information on the underlying Ordnance Survey data can be found here: https://www.ordnancesurvey.co.uk/business-and-government/products/points-of-interest.html

The length (meters) of roads and pedestrianised routes in each participant's 1000m line-based network is also provided in the data (see Table 4).

Table 4. Field descriptions of road/path network attributes in UKB_ app4236.csv

Field	Description
LN_length	Total length (m) in 1000m line network of road links, path links and connecting links derived from Ordnance Survey Integrated Transport Network and Urban Paths Theme extension
LN_lenped	Length (m) in 1000m road/path network of 'Footpath', 'Bridleway', 'Canal Path' and 'Pedestrianised Street' links

	derived from Ordnance Survey Integrated Transport Network and Urban Paths Theme extension
Walk_z	Walkabilty score generated by summing population density z-score, 3-way intersection (junction connectivity) z-score, and destination count z-score

Air pollution

Annual average nitrogen dioxide concentrations, 2010

Nitrogen dioxide (NO_2) estimates for the year 2010 were modelled for addresses in Greater London using a Land Use Regression (LUR) model. Predictor variables on land cover classes and distance to source (e.g. road) were derived using a geographic information system (GIS). In addition, traffic-related air pollution concentration estimates were modelled using an air pollution dispersion model (ADMS-Urban). The output from the ADMS-Urban along with selected land cover variables and distance to source were regressed against air pollution measurements (i.e. dependent variable) to derive a LUR model (see Table 5). Hence, the added air pollution concentrations are estimated using a 'Dispersion LUR' model and are available in column, 'no2_dlur10'.

Standardised receptor placement

To standardise the placement of address receptor points, each address point was associated with its nearest building polygon from Ordnance Survey (OS) MasterMap™. Following this, address receptors (points) were generated following a method described elsewhere (Gulliver et al., 2015). In brief, the receptor points were set 1m from the façade of the building associated with each postcode, on the side of the building closest to a main road. This was the assumed entry point of the building.

LUR predictor variables

ADMS-Urban

ADMS-Urban is a proprietary air pollution dispersion modelling tool developed to incorporate emissions from individual sources (e.g., roads, industrial point sources and area sources). It provides local-scale air pollution estimates within cities. The ADMS modelling software was used to estimate NO_2 concentrations at address-level receptor points (see Section 5.2). The NO_2 concentration output from this model was used as a predictor variable in the LUR model. The contribution of the variable 'ADMS-Urban' to the estimated NO_2 concentration (μg/m^3) at each receptor is available in column, 'ADMS'. More information on ADMS can be found here: https://www.cerc.co.uk/environmental-software/ADMS-Urban-model.html

Ground cover vegetation within 100m circular buffer

The LUR model includes high-resolution (2.5m x 2.5m) vegetation data as an air pollution predictor variable. The vegetation data – 'ground cover' – is described in more detail in Section 2. The (subtractive) contribution of 'ground cover within a 100m circular buffer' to the estimated nitrogen dioxide concentration (μg/m^3) at each address (receptor point) is available in column, 'gc100'.

Inverse distance squared to nearest major road

Inverse distance to roads (m^{-1}) squared is based on the road network geometry of the 2008 London Atmospheric Emissions Inventory. More information can be found here: https://www.kcl.ac.uk/lsm/research/divisions/aes/research/erg/modelling/emissions-inventory and data are available here: https://data.london.gov.uk/dataset/laei-2008. The 'inverse distance squared' variable was included to increase the gradient in modelled concentrations around roads, compensating for the air pollution dispersion (ADMS-Urban) model being run without buildings (not feasible for the whole of London), which resulted in shallower air pollution gradients around roads than expected due to greater levels of ventilation conditions in the model. The contribution of 'inverse distance squared to the nearest major road' to the estimated nitrogen dioxide concentration (µg/m^3) at each address (receptor point) is available in column, 'distinvmaj'.

Table 5. Field description for NO_2 variables in UKB_app4236.csv

FIELD	DESCRIPTION
no2_dlur10	Nitrogen dioxide; Dispersion-LUR estimate for annual average 2010 (µg/m^3) based on constant value, plus 'ADMS', 'GC100' and 'distinvmaj' contribution (detailed below).
ADMS	Contribution of dispersion air pollution model output (ADMS-Urban) to LUR NO_2 estimate (µg/m^3)
GC100	Contribution of ground cover (vegetation) within a 100m circular buffer variable to LUR NO_2 estimate (µg/m^3)
distinvmaj	Contribution of variable 'inverse distance squared to the nearest major road'* variable to NO_2 estimate (µg/m^3)

* Inverse distance to the nearest major road based upon 2008 London Atmospheric Emissions Inventory road network

Coverage

NO_2 concentration estimates (µg/m^3) are available for Greater London UK Biobank addresses due to the extent of the underlying vegetation data; 58828 addresses were within the modelling domain. Other addresses have missing data (-999).

Road traffic noise

Annual average traffic noise estimates, 2013

A-weighted noise (dB) estimates for the year 2013 were modelled using a high-resolution version of the CNOSSOS-EU noise model (Morley et al., 2015). Common NOise aSSessment methOdS (CNOSSOS), mandatory for modelling in relation to the European Noise Directive 2002/49/EC, was developed to standardise practices in noise estimation across Europe (Kephalopoulos et al., 2014b). The CNOSSOS algorithm was implemented using PostGIS (v. 2.3.3) – a spatial extender for Postgres (v. 9.6). Road traffic noise was modelled at the same address receptor points for air pollution modelling (see Section 5.2). The high-resolution input data used in the model are described in more detail below.

Standardised receptor placement

To standardise the placement of address receptor points, each address point was associated with its nearest building polygon from Ordnance Survey (OS) MasterMap™. Following this, address receptors (points) were generated following a method described elsewhere (Gulliver et al., 2015). In brief, the receptor points were set 1 m from the façade of the building associated with each postcode, on the side of the building closest to a main road. This was the assumed entry point of the building.

Input data

Estimated traffic flow on minor roads

Traffic flow data on minor roads is not captured by the UK Department for Transport. This data limitation impacts the accuracy of traffic noise estimations at residential locations serviced by minor roads. To better predict noise levels at these locations, a routing algorithm was developed to rank roads by importance based on simulated journeys through the road network and included in statistical model to estimate traffic flows (Morley and Gulliver, 2016b) – 2013 traffic flow estimates were adapted (converted to hourly flow over an average diurnal profile) and used as CNOSSOS model input.

High resolution land cover data

OS MasterMap™ Topography data (1:1250 scale) was used to calculate land cover between source (traffic) and receptor (address) points. Land cover was classified based on sound absorbance. For example, natural surfaces were classed as more sound absorbent than manmade surfaces. Classified land cover was used in the model to attenuate propagation of sound due to land cover absorbance. The maximum adjustment in the model is -3dB. More information on the underlying data can be found here: https://www.ordnancesurvey.co.uk/business-and-government/products/topography-layer.html

Building height data

Geometry of buildings with heights was included to model noise diffraction along lines of noise propagation between road sources and receptors (address). Building heights were derived from LiDAR Digital Surface Model and Digital Terrain Model, and were overlaid with OS MasterMap™ Topography building polygons.

Table 6. Field descriptions for 2013 noise estimates in UK_Biobank_app4236.csv

Field	Description
Lday_13	LDay (day equivalent level): Average sound level pressure LAeq over the 12-hour period 07:00 to 19:00 (dB)
Leve_13	LEve (evening equivalent level): Average sound level pressure LAeq between the hours of 19:00 to 23:00 (dB)
Lnight_13	LNight (night equivalent level): Average sound level pressure LAeq overnight 23:00 to 07:00 (dB)

Laeq16_13	LAeq,16hr (A-weighted equivalent sound level): Average sound level pressure LAeq between the hours of 07:00 to 23:00 (dB)
Lden_13	LDen: (day-evening-night equivalent level): A weighted Leq noise level measured over the 24 hour period with a 10 dB penality added to the levels between 23:00 and 07:00 (dB)