Preprint Cambridge Centre for Computational Chemical Engineering ISSN 1473 – 4273

Cold homes, unsafe communities, and limited healthy opportunities: Evidence on neighbourhood exposures and non-communicable diseases via The World Avatar

Jiying Chen[1], Jethro Akroyd[1,2], Sebastian Mosbach[1,2],
Jingfeng Zhou[1,2], Feroz Farazi[1,2], Xiaochi Zhou[1],
Søren Brage[2,3], Nicholas J. Wareham[2,3], Markus Kraft[1,2,4]

released: December 27, 2025

[1] Department of Chemical Engineering
and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom

[2] CARES
Cambridge Centre for Advanced
Research and Education in Singapore
1 Create Way
CREATE Tower, #05-05
Singapore, 138602

[3] MRC Epidemiology Unit
University of Cambridge
Cambridge
United Kingdom

[4] Department of Chemical Engineering
Massachusetts Institute of Technology (MIT)
77 Massachusetts Avenue, Room E17-504
Cambridge, MA 02139
United States of America

Preprint No. 342

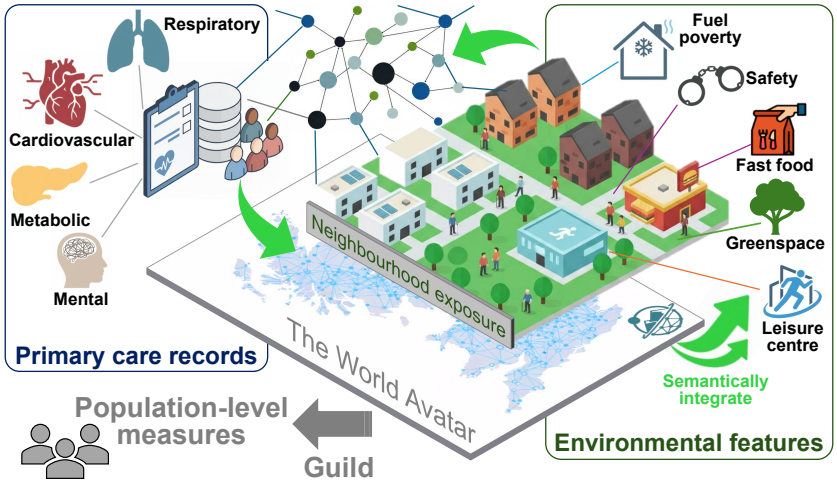
Keywords: non-communicable diseases, neighbourhood exposures, semantic representation, digital twin, evidence-based intervention

Edited by
Computational Modelling Group
Department of Chemical Engineering and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom

E-Mail: mk306@cam.ac.uk
World Wide Web: https://como.ceb.cam.ac.uk/

Abstract

Urban environments shape non-communicable disease (NCD) patterns, yet the transformation of public data into granular, policy-actionable evidence remains challenging as cross-sector data silos and heterogeneous standards impede at-scale harmonisation and holistic analysis. We built a UK-wide, small-area environmental health evidence base by semantically integrating national datasets including housing, neighbourhood amenities, community safety, and accessibility within *The World Avatar* (TWA), an interoperable knowledge architecture enabling machine-readable, consistent profiling of adjusted associations between determinants, covariates, and primary-care disease prevalence. Results show that fuel poverty emerged as a leading correlate of respiratory morbidity: scenarios consistent with improved home heating efficiency suggest that reducing fuel poverty could be associated with averting over 13% of the population burden of chronic obstructive pulmonary disease. Spatial profiling identified critical proximity thresholds: leisure-centre access was protective within an approximately 4-minute drive, while health risks linked to fast-food outlets declined beyond roughly 6 minutes. Mental-health burden showed strong, monotonic associations with antisocial behaviour, domestic abuse, and drug-related crime. Green space supported metabolic health but exhibited a U-shaped association with severe mental illness, with the lowest prevalence at intermediate vegetation density (NDVI ≈0.5–0.6). These findings show how semantic integration can generate actionable, scalable environmental intelligence for digital health systems and place-based prevention.



Highlights

- Semantic integration turns open data into actionable population-health evidence.
- Leisure centres within 4 min and fast food beyond 6 min signal healthier places.
- Fuel poverty links to COPD; warmer homes could cut the burden by over 13%.
- Rising crime and fuel poverty track rising mental illness; mid-level greenness offers protection.

Contents

1 Introduction

The escalating global burden of non-communicable diseases (NCDs) represents a primary public health challenge of the 21st century [39]. This burden is dominated by a complex syndemic of metabolic, cardiovascular, respiratory and cancer conditions [29, 70, 90]. Evidence indicates that the burden of NCDs is shaped not only by individual biology but also by the complex interplay of context and place, operating through patterns of everyday living and exposure to the surrounding environment [59, 100]. Understanding how environmental determinants are associated with the prevalence of NCDs at the population level remains a core aim of public health research.

Large cohort studies have provided powerful information on exposure-specific risks, causal inference, and aetiological understanding, with depth in genetic, biomarker, and individual-level longitudinal data. Although such studies offer deep longitudinal evidence, their substantial resource requirements, long follow-up period and limited geographic coverage can constrain the ability to capture variability in environmental features and social contexts. This challenge is particularly salient for public health surveillance, which needs to characterise risk distributions across the entire population [53]. The digital transformation of public health services and government administration creates a transformative opportunity to bridge this gap. Public open data, including primary care records, geospatial surveys and demographic statistics, provide a complementary resource for both estimating disease burden and characterising environmental exposures [74, 107].

Routinely collected data have the potential to provide a high-resolution, nationwide contextual layer for public health studies. In parallel, they can facilitate the development of digital cohorts and deepen exposure assessment, providing infrastructure for population scale exposome epidemiology [13, 103]. Yet the rich environmental, social, and administrative data needed to approximate a more complete exposome often reside in heterogeneous and disconnected silos [14, 107]. They are typically maintained by separate public bodies and organisations, stored in heterogeneous formats ranging from static tabular data to satellite imagery, referenced to misaligned spatial units [43], and often lack interaction standards. The resulting impediment to interoperability caused by this fragmentation is more than a technical inconvenience. It can introduce avoidable bias hinder the integration of socioeconomic and health information and obscure mechanisms that drive health inequalities and algorithmic unfairness [47, 80].

Despite the breadth of public administrative and geospatial data now available, three practical challenges limit their use for decision-making about environmental determinants of health. First, evidence is often produced in domain silos, making it difficult to understand how multiple, correlated determinants (housing energy deprivation, crime, greenspace, and the retail and activity environment) jointly relate to health outcomes and inequalities. Second, planners require *actionable* spatial signals, for example, proximity thresholds for facilities and hazards, yet these are rarely reported consistently at the national scale. Third, while cohort studies provide deep individual-level inference, routine public health surveillance needs tools that can use the data that are *already* produced nationally (and regularly) by government and public bodies, even if those data are cross-sectional snapshots rather than full longitudinal histories.

Semantic technologies are increasingly used in public health to make heterogeneous data interoperable, which enables multi-scale analytics at population scale and supports surveillance that spans many data holders [89]. The FAIR Guiding Principles[105] define findability, accessibility, interoperability, and reusability as shared goals for data and workflows, and they have become a foundation for linking and reusing public health information in practice. Complementary standards and associated terminology services help systems not only exchange records but also interpret them consistently, while health knowledge graphs represent entities and relations so that clinical, environmental, and administrative sources can be queried together with provenance and context. One of the semantic approaches is The World Avatar (TWA) [49], an open-source, community-driven initiative that aims to create an all-encompassing knowledge model of the real world using machine-readable models, with autonomous agents to keep the digital representation updated, extensible and computable. It has been used to link data and models across domains applied to the chemistry space [4], buildings [73] and urban environments to national-scale representations [20, 91] for sustainable planning and infrastructure management. We apply this idea of a digitally interconnected world to nationwide public health by integrating a population-level digital infrastructure that semantically aligns publicly available environmental and residential data with routinely collected health and administrative records into TWA.

The **purpose of this paper** is to advance fine-scale, policy-ready environmental health evidence by 1) building a national, interoperable evidence layer that links publicly available environmental determinants to routine primary-care health outcomes; 2) estimating adjusted associations while accounting for key covariates including age, urban-rural classification, access to healthcare, and socioeconomic deprivation; and 3) extracting policy-relevant benchmarks that can inform targeted intervention. We implement this by semantically integrating heterogeneous government data sources and harmonising determinants and covariates to a census-based small-area geography, enabling consistent analysis from local to national levels. This cross-sector interoperability supports precision public health and geographically informed planning by allowing systematic examination of spatial patterns of disease and their potential environmental and socioeconomic correlates.

2 Background

Chronic disease burden reflects cumulative exposures in the residential environment. Policy frameworks increasingly treat neighbourhood conditions as modifiable determinants of non-communicable disease and health inequalities. The WHO European Healthy Cities framework highlights healthier urban environments, equity, and cross-sector action [106]. In the UK, national planning policy frames the planning system as a lever to support healthy, safe, and inclusive communities [98]. These agendas create a clear evidence need: decision-makers require small-area, policy-ready measures of modifiable neighbourhood determinants linked to routine health outcomes.

To situate this study, we mapped recent cross-sectional population studies on urban environmental characteristics and the prevalence of metabolic, cardiovascular, respiratory, and mental health conditions. We searched Web of Science Core Collection for studies

published in the last five years and generated a keyword co-occurrence network (Figure 1).

The network points to recurring environmental domains relevant to non-communicable diseases. Keywords cluster around 1) the local food environment and obesity-related outcomes; 2) greenspace and mental health and wellbeing; and 3) air pollution, traffic, and noise in relation to cardiovascular and respiratory outcomes. Crime and area deprivation connect across clusters. Housing and energy terms bridge socioeconomic disadvantage and cardiometabolic risk. Direct small-area measures of dwelling quality and indoor temperatures are often sparse, whereas fuel poverty is routinely available nationwide and reflects affordability and thermal adequacy, making it a practical proxy for housing and energy deprivation in this analysis [20]. Guided by data availability and policy relevance, our knowledge-based analytical framework covers three dimensions: energy affordability, community safety, and everyday exposure settings and amenities. We operationalise these dimensions using five small-area measures: fuel poverty prevalence, police-recorded crime incidents, satellite-derived greenness as a proxy for neighbourhood vegetation, access to leisure centres, and fast-food outlet accessibility.

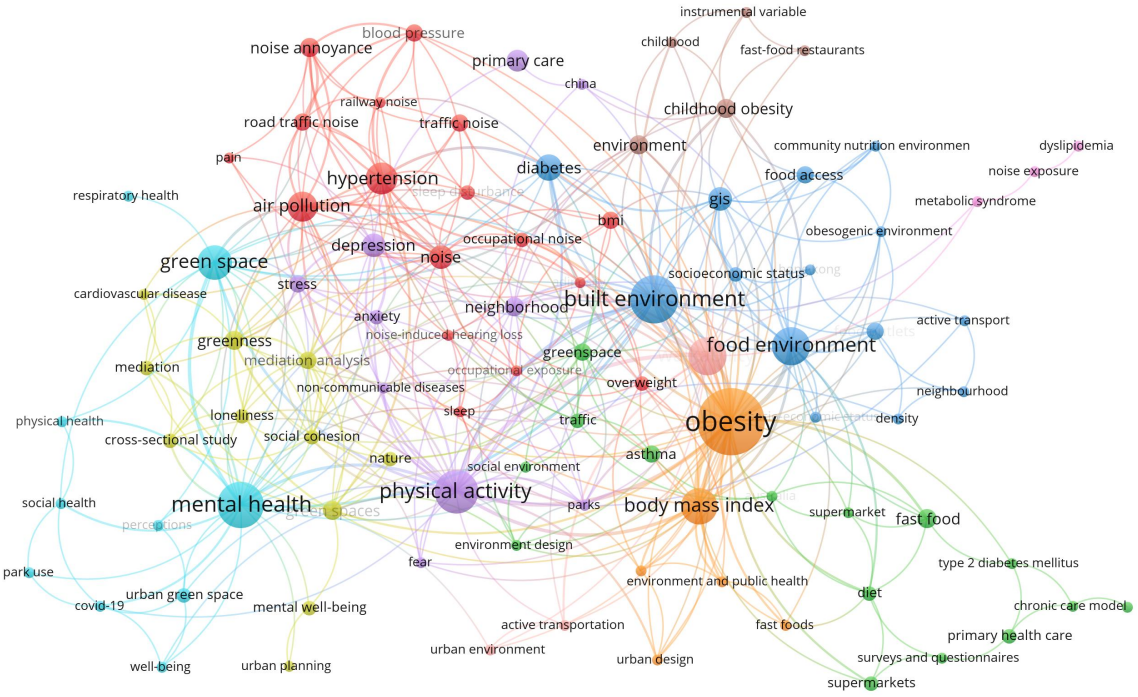


Figure 1: *Keyword co-occurrence network of literature keywords in population-based cross-sectional studies on urban environmental characteristics and the prevalence of metabolic, cardiovascular, respiratory and mental health conditions retrieved from Web of Science in the last five years. Node size reflects keyword frequency and link thickness represents co-occurrence strength.*

2.1 Fuel poverty and health

Fuel poverty is commonly described as the inability to keep the home adequately warm due to unaffordable energy and poor building energy efficiency, with cold homes contributing to a wider burden of ill health and inequalities [15, 97]. Quantitative studies link fuel poverty and cold homes to respiratory and cardiometabolic outcomes, with evidence spanning area-level analyses, cohorts, and intervention studies. A national LSOA ecological analysis linked mean EPC rating and building retrofit indicators to adult Hospital Episode Statistics (HES) emergency admissions. In adjusted negative binomial models, a one-point higher mean EPC rating was counter-intuitively associated with slightly higher admission rates (RR$\approx$1.005 for asthma and CVD; RR$\approx$1.010 for COPD). However, associations for specific measures were mixed. For instance boiler upgrades exhibited a protective effect for COPD (RR 0.98–1.00) [83]. As a downstream endpoint, emergency admissions capture severe exacerbations and service use and are not equivalent to routine prevalence burden. In a survey of social housing tenants in Cornwall (southwest England), self-reported fuel poverty and colder bedroom conditions were associated with poorer mental health (SF-12 MCS), while none of the fuel poverty measures showed a significant association with SF-12 physical component score (PCS) [97]. Panel analyses using Understanding Society reported more pronounced associations with reduced life satisfaction and raised inflammatory biomarkers when fuel poverty was defined using composite indicators that incorporate heating inadequacy, rather than expenditure-based indicators alone [24]. A natural experiment around Winter Fuel Payment eligibility (24,651 observations) found no overall effect on care needs, quality of life, or cold-related housing problems, but reported benefits for some subgroups and fewer cold-related housing problems in more modern homes [19]. Intervention evidence is mixed: a before–after study installing central heating in homes of children with asthma reported improvements in symptoms and school absence [87], whereas a controlled before–after study of domestic energy efficiency investments in low-income areas found limited evidence of changes in self-reported health despite improvements in wellbeing and related outcomes [32]. A meta-analysis across 36 studies (33,313 participants) reported a small overall health benefit from energy efficiency measures (sample-weighted $d^{+} = 0.08$) [52]. In summary, interpretation requires recognising heterogeneity in energy affordability measures and study scales. Local studies are confined to selected populations and condition-specific outcomes. National analyses often rely on service-based endpoints that may not align with routine morbidity burden.

2.2 Crime statistics and health

A growing body of work has examined how neighbourhood safety, often proxied by crime statistics, relates to health outcomes. Longitudinal linkage analyses of the Scottish Longitudinal Study combine police-recorded crime at the data-zone level with census health measures and National Health Service (NHS) prescribing records, and indicate that increases in local crime are associated with higher odds of self-reported mental illness and new prescriptions for antidepressants and antipsychotics, with stronger effects in younger adults and socioeconomically disadvantaged groups [6, 7]. Complementary evidence from South London, using electronic mental healthcare records, shows that res-

idents with mental disorders living in higher-crime neighbourhoods have higher odds of physical victimisation after adjustment for individual and area-level factors, consistent with close coupling between community safety and mental ill-health [12].

Cohort evidence in adolescence further suggests that the crime-related neighbourhood context operates alongside individual victimisation. In a UK adolescent twin cohort, neighbourhood social adversity indices incorporating crime statistics and personal violent-crime victimisation were each associated with psychotic experiences, and joint exposure was linked with higher odds [60]. A related analysis in the same cohort indicates that adolescents' perceptions of neighbourhood disorder were associated with psychotic experiences even after accounting for official neighbourhood crime rates [61], highlighting that recorded crime and lived experience can capture partially distinct dimensions of neighbourhood safety. At the national scale, descriptive analyses link higher area crime rates with lower life expectancy, underscoring the public health salience of community safety [93]. A recent systematic review and meta-analysis similarly concludes that neighbourhood crime is consistently associated with depression and psychological distress beyond the effects of socioeconomic deprivation, although evidence on physical disease outcomes remains sparse [5].

Taken together, existing studies provide convergent signals across linked administrative records, electronic health records, and cohort designs, yet they often operationalise crime exposure and neighbourhood context in study-specific ways (e.g., recorded crime, composite adversity, or perceived disorder) and focus on selected populations or geographies. As a result, evidence synthesis is constrained by differences in how crime exposure is defined and linked to health endpoints, and by limited consistency in spatial and temporal resolution across data sources.

2.3 Leisure centre accessibility and health

Formal leisure facilities within the built environment are recognised as critical determinants of physical activity [1, 76]. In the UK Biobank, the presence of six or more formal facilities within 1 km of the home was associated with lower body mass index (-0.57 kg/m^2, 95% CI -0.74 to -0.39), waist circumference (-1.22 cm, 95% CI -1.64 to -0.80), and body fat percentage (-0.81%, 95% CI -1.03 to -0.59) compared with having no facilities [55]. These associations were stronger among women and higher-income groups but genetic predisposition to obesity did not modify this effect [55, 56]. Similarly, linking Scottish health survey data to a national facility registry showed that each additional facility within a 20-minute walk was associated with a 0.015 unit decrease in body mass index (BMI) ($P = 0.02$), although self-reported physical activity did not mediate this result [28]. Comparisons of home and workplace neighbourhoods suggest that private facilities near the home may have a greater link to exercise frequency than those near the workplace [51].

Evidence from natural experiments strengthens causal inference. In a deprived English authority, removing financial barriers to swimming pools and gyms led to significant increases in participation and highlighted the role of cost [36]. However, geographic proximity alone implies limited equity. While the average distance to Parkrun events saw a

sharp decline between 2010 and 2019, participation remained stratified by socioeconomic position [81, 86]. Data on exercise referral schemes indicate that leisure centre attendance contributes an average of 55 minutes of moderate-to-vigorous activity per week, with women and older adults demonstrating higher adherence [9]. From a health economic perspective, a free membership programme in London yielded an incremental cost-effectiveness ratio of £20,347 per quality-adjusted life year (QALY), though this estimate was sensitive to assumptions regarding mental health benefits [99]. National statistics show a correlation between facility density and activity levels, but confounding by urban and rural status remains a limitation [67]. Collectively, the evidence suggests modest associations between facility accessibility and anthropometric outcomes. The scarcity of direct evidence linking facility accessibility to disease prevalence underscores the need for future research.

2.4 Greenspace accessibility and health

Research has increasingly examined the health impacts of greenspace using high-resolution exposure metrics and large-scale population cohorts. Longitudinal analyses in Wales linked electronic health records for over 2.3 million adults to satellite-derived vegetation indices and measures of access to public green and blue spaces, finding that greater greenness and proximity predicted substantially lower odds of subsequent common mental disorders, with stronger effects in deprived urban settings [30]. Natural experiment designs during the COVID-19 lockdown similarly demonstrated that adults living within 800 m of accessible greenspace experienced smaller increases in psychological distress, particularly in London [48]. Evidence from UK Biobank indicates that higher neighbourhood greenness, measured via Normalised Difference Vegetation Index (NDVI)[38], was associated both with reduced prevalence of major depressive disorder and with lower BMI, waist circumference, and body-fat percentage, with stronger associations in socioeconomically disadvantaged and highly urban areas [78]. More recent analyses of UK Biobank participants have extended these findings to additional outcomes. Liu et al. (2024) reported that each interquartile-range increase in residential NDVI was associated with a 6% lower incidence of psychiatric disorders among middle-aged and older adults [50]. He et al. (2024) similarly found that participants living in greener neighbourhoods had about a 10% reduced risk of incident delirium [35], while Tang et al. (2025) demonstrated that higher NDVI predicted a 7% lower incidence and slower progression of cardiometabolic multimorbidity [92]. Complementary cohort studies reinforce these findings: children raised in greener neighbourhoods showed lower risks of overweight or obesity by age 10–11 years, and adults in the greenest quartile of neighbourhoods in Norfolk had a 19% lower hazard of incident type-2 diabetes compared with those in the least green areas [23, 104]. Together, these studies highlight consistent benefits of nearby, accessible greenspace across mental health and metabolic outcomes, while revealing important heterogeneity by urbanicity, socioeconomic context, and spatial scale of measurement.

Most of these studies have relied on individual-level cohorts or region-specific analyses, leaving less evidence on how greenness relates to population health across the whole country. In particular, ecological assessments at small-area levels, which can capture ge-

ographic variation in both environmental exposures and disease prevalence, remain comparatively scarce. This gap makes it difficult to evaluate whether the benefits observed in selected cohorts generalise at a national scale or align with patterns of morbidity observed in routine health data.

2.5 Fast food outlet accessibility and health

Fast food outlets are widely discussed as indicators of an obesogenic environment as they sell energy-dense, nutrient-poor foods that are relatively cheap and convenient. In the CARDIA 15 year cohort, evidence shows that eating fast food more than twice per week was associated with about 4·5 kg greater weight gain and higher insulin resistance than eating it less than once per week, after adjustment for lifestyle factors [69]. UK nutritional profiling suggests that a typical takeaway meal can approach or exceed recommended daily intakes for energy, saturated fat, and sodium [42]. These findings support the use of fast food outlet availability as a contextual marker of unhealthy dietary environments.

Neighbourhood studies suggest that where fast food outlets are more common, diets tend to be less healthy and adiposity higher. In a population-based study in Cambridgeshire, adults with the highest combined exposure to takeaway outlets across home, work and commuting environments consumed more takeaway food and had body mass index about 1.2 kg/m² higher than those least exposed [17]. Analyses of UK Biobank participants in Greater London showed that both low household income and a higher neighbourhood proportion of fast-food outlets were independently associated with higher BMI, higher body-fat percentage and greater odds of obesity, with an odds ratio for obesity of 1.51 (95% CI 1.40–1.64) for the most exposed versus the least exposed quartile, and odds rising to 2.43 (95% CI 2.09–2.84) when low income and high exposure co-occurred [18]. In children and adolescents, pooled estimates for outlet access and overweight/obesity are close to null (odds ratio 1.01 [0.97–1.05] for presence and 1.00 [0.99–1.01] for number of outlets), with heterogeneity by subgroup and by whether exposure is defined around homes or schools [44].

Fast food outlet accessibility is also associated with metabolic morbidity. In a large cross-sectional analysis of 347,551 UK Biobank participants across 21 cities, higher residential density of ready-to-eat food outlets within 1 km was associated with greater odds of type 2 diabetes, with odds ratios of 1.129 (95% CI 1.05–1.21) for the highest category of restaurant and cafeteria density and 1.112 (1.02–1.21) for a composite ready-to-eat outlet metric when compared with no nearby outlets, and modestly elevated odds at intermediate densities of hot and cold takeaways [79]. A systematic review and meta-analysis of food environments and adult obesity reported that closer proximity to fast food restaurants was associated with higher odds of obesity, pooled odds ratio 1.15 (95% CI 1.02-1.30), although findings for outlet density were more variable [71].

Studies have also examined cardiovascular outcomes. In a nationwide Dutch cohort of 2,472,004 adults who had lived at the same address for at least 15 years, the incidence of cardiovascular disease and coronary heart disease over one year was higher among urban residents with one or more fast-food outlets within 500 m than among those with none, and increased fast-food outlet density within 1000 m was associated with higher incidence of

cardiovascular disease and coronary heart disease after adjustment for sociodemographic characteristics, comorbidity and neighbourhood income [72]. Ecological studies from the USA report that counties with higher densities of fast-food and similar outlet types tend to have higher prevalence of obesity and diabetes and worse composite health indicators, although these analyses are vulnerable to residual confounding and often rely on coarse spatial units [77].

Existing studies associate greater density or proximity of fast-food outlets with ihigher adiposity and cardiometabolic risk, particularly in deprived settings. However, much of the UK evidence relies on individual-level cohorts or localised studies, which may not capture the population-wide burden of disease attributable to the food environment across diverse geographic contexts. Furthermore, ecological analyses often examine fast-food exposure in isolation, leaving them vulnerable to confounding by other features of the built environment. Consequently, it remains unclear how fast-food accessibility relates to routine health outcomes at a national scale.

3 Method

This study integrates primary care health records with open environmental data to quantify associations between neighbourhood characteristics and non-communicable disease prevalence. The methodology proceeds in four distinct stages. We first define the assembly of environmental exposures and sociodemographic covariates at the small area level. We then describe the derivation of local disease prevalence estimates from national general practice registers. These heterogeneous data streams are subsequently unified within The World Avatar semantic knowledge graph to create an interoperable digital evidence base. The section concludes by detailing the statistical framework employed to characterise dose response relationships and estimate population attributable fractions.

3.1 Environmental features and covariates

All analyses were conducted at the level of Lower Layer Super Output Areas (LSOAs). For each LSOA we assembled a set of environmental exposure indicators capturing the food environment, opportunities for physical activity, housing and energy conditions, and neighbourhood safety, together with demographic and socioeconomic covariates. Table 1 summarises all indicators and data sources.

Firstly, accessibility to facilities supporting physical activity and to fast-food outlets was quantified using network-based car travel time from residential locations within each LSOA to the nearest facility. Specifically, we used the Access to Healthy Assets and Hazards (AHAH) accessibility model, produced by the Geographic Data Service (GeoDS), which estimates driving time along the established road network between population-weighted postcode centroids and the nearest outlet or service[31]. For each LSOA, we summarised postcode-level travel times as the median across constituent postcodes, thereby reducing sensitivity to boundary postcodes and localised extremes in connectivity.

To support interpretation of time-based thresholds in spatial terms, we extended the same

AHAH routing approach [11] to additionally compute and output shortest-path road-network distance for the corresponding nearest-facility route. Distances were derived from the same shortest-path solution as the travel-time estimates and were summarised as LSOA medians across constituent postcodes. These indicators remain area-level proxies and therefore cannot eliminate ecological limitations. However, by using routable networks they better represent potential accessibility constraints than straight-line proximity or single-centroid measures, providing a pragmatic compromise when individual mobility trajectories are not available.

Secondly, green space exposure was represented using Normalised Difference Vegetation Index (NDVI) statistics derived from cloud-free satellite imagery. We used the median NDVI value for each LSOA as a proxy for local greenness. Fuel poverty was captured using sub-regional fuel poverty statistics for England, which provide the estimated proportion of households in fuel poverty in each LSOA under the Low Income Low Energy Efficiency (LILEE) definition. Finally, neighbourhood safety was proxied by the annual rate of police-recorded crime per 1 000 residents in each LSOA. Crime counts were obtained from open police data and aggregated to LSOA level before standardisation by mid-year population.

To reduce confounding we included area-level covariates that capture demographic structure, settlement context, housing conditions, healthcare accessibility, and socioeconomic status. Age structure was represented by the proportions of residents in broad age bands (0–17, 18–64, $\geq$65 years) using official mid-year population estimates by LSOA. Urban-rural status was taken from the 2011 rural–urban classification for LSOAs, dichotomised to distinguish predominantly urban from predominantly rural areas.

Housing quality and thermal performance were proxied by indicators derived from the Energy Performance of Buildings Certificates (EPC) open data for England and Wales. For each LSOA we computed summary measures such as the proportion of dwellings with EPC ratings A–C and the mean Standard Assessment Procedure (SAP) score. Access to primary care was measured as median drive time by car to the nearest general practice, again using network-based accessibility metrics at LSOA level.

Socioeconomic context was represented using the English Indices of Deprivation at LSOA level. In the main models we included the income domain and the education, skills and training domain as continuous rank- or decile-based measures. We additionally included total population and population density as descriptive variables and, where appropriate, as offsets or scaling factors in the regression models.

Table 1: *Environmental exposure, covariate, and health outcome datasets assembled at UK Lower Layer Super Output Area (LSOA) level.*

Category	Indicator	Operational definition	Source dataset and provider
Exposure	Fuel poverty rate	Proportion of households in fuel poverty in each LSOA, based on the Low Income Low Energy Efficiency (LILEE) definition.	Sub-regional fuel poverty statistics for England provided by the Department for Energy Security and Net Zero [25].

(Continued on next page)

Table 1 – Continued

Category	Indicator	Operational definition (LSOA level)	Source dataset and provider
Exposure	Greenness	Median Normalised Difference Vegetation Index (NDVI) value for each LSOA, derived from cloud-free Sentinel-2 satellite imagery and used as a proxy for local green space.	NERC EDS Centre for Environmental Data Analysis, Joint Nature Conservation Committee (2025) [45].
Exposure	Crime records	Street-level crime, outcome, and stop and search information, broken down by police force	Crime and policing open data and related police force statistics [84].
Exposure	Drive time to fast food	Median car travel time (minutes) from the centroid of each postcode within an LSOA to the nearest fast-food outlet, calculated along the OS OpenRoad network using road-specific speed estimates.	Accessibility metrics from the Access to Healthy Assets and Hazards (AHAH) dataset [11, 31].
Exposure	Drive time to leisure centre	Median car travel time (minutes) from the centroid of each postcode within an LSOA to the nearest leisure centre or comparable leisure facility, calculated along the OS OpenRoad network using road-specific speed estimates	Accessibility metrics from the Access to Healthy Assets and Hazards (AHAH) dataset [11, 31].
Covariate	Age structure	Proportion of the LSOA population in broad age bands (0–17, 18–64, $\geq$65 years).	Lower Layer Super Output Area mid-year population estimates; Office for National Statistics [68].
Covariate	Urban-rural classification	Categorical indicator of settlement type (*e.g.* urban vs rural), derived from the 2011 rural–urban classification for LSOAs.	2011 rural–urban classification lookup tables for LSOAs in England and Wales; Office for National Statistics [65, 66].
Covariate	Housing energy efficiency	Summary indicator of dwelling energy performance in each LSOA, such as the proportion of EPC ratings A–C or mean SAP score.	Energy Performance of Buildings Certificates open data; Department for Levelling Up, Housing and Communities [26].
Covariate	GP accessibility	Median car travel time (minutes) from the centroid of each postcode within an LSOA to the nearest general practice (GP) surgery, calculated along the OS OpenRoad network using road-specific speed estimates.	Accessibility metrics from the AHAH dataset [11, 31].
Covariate	Income deprivation	Area-level income deprivation measure at LSOA level (*e.g.* decile or rank), used as a proxy for household income.	English Indices of Deprivation, income domain; Department for Levelling Up, Housing and Communities [58].
Covariate	Education deprivation	Area-level deprivation in education, skills and training at LSOA level.	English Indices of Deprivation, education, skills and training domain; Department for Levelling Up, Housing and Communities [58].

(Continued on next page)

Table 1 – Continued

Category	Indicator	Operational definition (LSOA level)	Source dataset and provider
Weight	Population density	Total resident population and population per square kilometre for each LSOA.	Office for National Statistics small-area population estimates and LSOA boundary data [68].
Health outcome	Disease prevalence (long-term conditions)	Estimated prevalence of selected conditions derived from General Practice disease registers and reallocated from practice to LSOA level using patient residence data [37, 62].	The Quality and Outcomes Framework (QOF) provides practice-level prevalence and exception data for GP practices. These data are linked to patients' residence using NHS Digital patient registration data [63].

3.2 Estimation of small area disease prevalence from GP and QOF data

To estimate disease prevalence at the small-area level, we utilise practice-level Quality and Outcomes Framework (QOF) data, which report both the number of patients on specific disease registers and the total number of registered patients for each general practice. These data are integrated with separate statistics detailing the number of registered patients by their LSOA of residence. The objective is to derive a modelled prevalence proportion for every LSOA and condition by mapping practice-level aggregates to the underlying local geography.

The estimation procedure is fully formalised in Appendix A (Subsection: Estimation of LSOA disease prevalence). Conceptually, the method relies on a proportional allocation process. First, the count of patients from each practice residing in a specific LSOA is reconciled to match the official practice list size. These figures are used to calculate the share of each practice's register that contributes to a given LSOA. Second, these spatial shares act as weights to apportion practice-level disease counts to the LSOA level. The final modelled prevalence is computed by aggregating these allocated cases across all practices serving the area and dividing by the corresponding modelled denominator. This approach yields an ecological measure of local disease burden. The specific reconciliation and attribution steps implemented in this project follow the methodology developed by the House of Commons Library [37].

3.3 Semantic integration in The World Avatar

The integration of environmental and health indicators is established within *The World Avatar* (TWA) dynamic knowledge graph to support reproducible analysis and cross-domain interoperability. Each Lower Layer Super Output Area functions within this framework as a distinct spatial entity identified by its official code where disease prevalence estimates and environmental exposures are attached using ontology terms that distinguish indicator type and provenance.

Figure 2 visualises the semantic architecture structured into three interoperable modules.

The upper Primary Care Electronic Health Records module models clinical entities and disease taxonomies via established public health terminologies. It incorporates Medical Subject Headings (MeSH) [22] maintained by the National Library of Medicine for hierarchical indexing and employs the Ontology of Clinical Health Variables (OCHV) [2] to define clinical observations. These terminologies operate alongside Fast Healthcare Interoperability Resources (FHIR) [10] which constitutes a standard for the exchange of organisational and patient data. The central Geospatial Classifications module anchors these health domains to administrative boundaries. This tier utilises the Ontology for Office for National Statistics (OntoONS) to define statistical geographies including Lower Layer Super Output Areas and relies on the OGC GeoSPARQL standard [8] for geometric representation. The lower Socio-environmental Determinants module integrates contextual factors via external standards and domain-specific models. The framework employs the W3C Sensor, Observation, Sample, and Actuator ontology (SOSA) [41] to model observations alongside Schema ontology [33] for place definitions. Further integration includes the OpenStreetMap Ontology (OSMonto) [21] for amenities and the Ontology of units of Measure (OM) [75] for standardisation. Transportation accessibility is captured by The iCity Transportation Planning Suite of Ontologies (TPSO) [46]. This module is further supported by four domain-specific ontologies developed under The World Avatar project. OntoCrime characterises neighbourhood safety and OntoIMD captures multiple deprivation indices as measures of socioeconomic context, while OntoBuiltEnv and OntoBuiltEnergy provide structured representations of the built environment and its energy performance respectively [20].

A generic measurement pattern is adopted in which each numerical indicator is modelled as an observation entity linking an LSOA to a value alongside its unit and time stamp. Environmental indicators *e.g.* fuel poverty rates and NDVI greenness are instantiated as subclasses of an environmental exposure concept while disease-specific prevalence observations appear as subclasses of a health-condition prevalence concept with explicit links to QOF definitions. Underlying tabular data are exposed to the knowledge graph using ontology-based data access where relational tables are mapped to RDF graphs via declarative mappings. This design enables agents in TWA to retrieve and update environmental and health indicators in a consistent manner that supports cross-study comparisons.

3.4 Exposure response and attributable risk analysis

This section summarises the statistical procedures used to characterise associations between environmental exposures and LSOA level disease prevalence and to quantify the corresponding population level impact. All models and effect measures are defined in detail in the Appendix sections referenced below.

Associations between individual exposures and estimated LSOA prevalence proportions are first described using rank based and partial rank correlations. Spearman correlation is used to capture monotone rather than strictly linear relationships between skewed ecological variables. Partial Spearman correlations adjust for age band composition and urban rural classification with optional population weighting through weighted ridit ranks [16] as set out in Appendix A.2 and Appendix A.2.4.

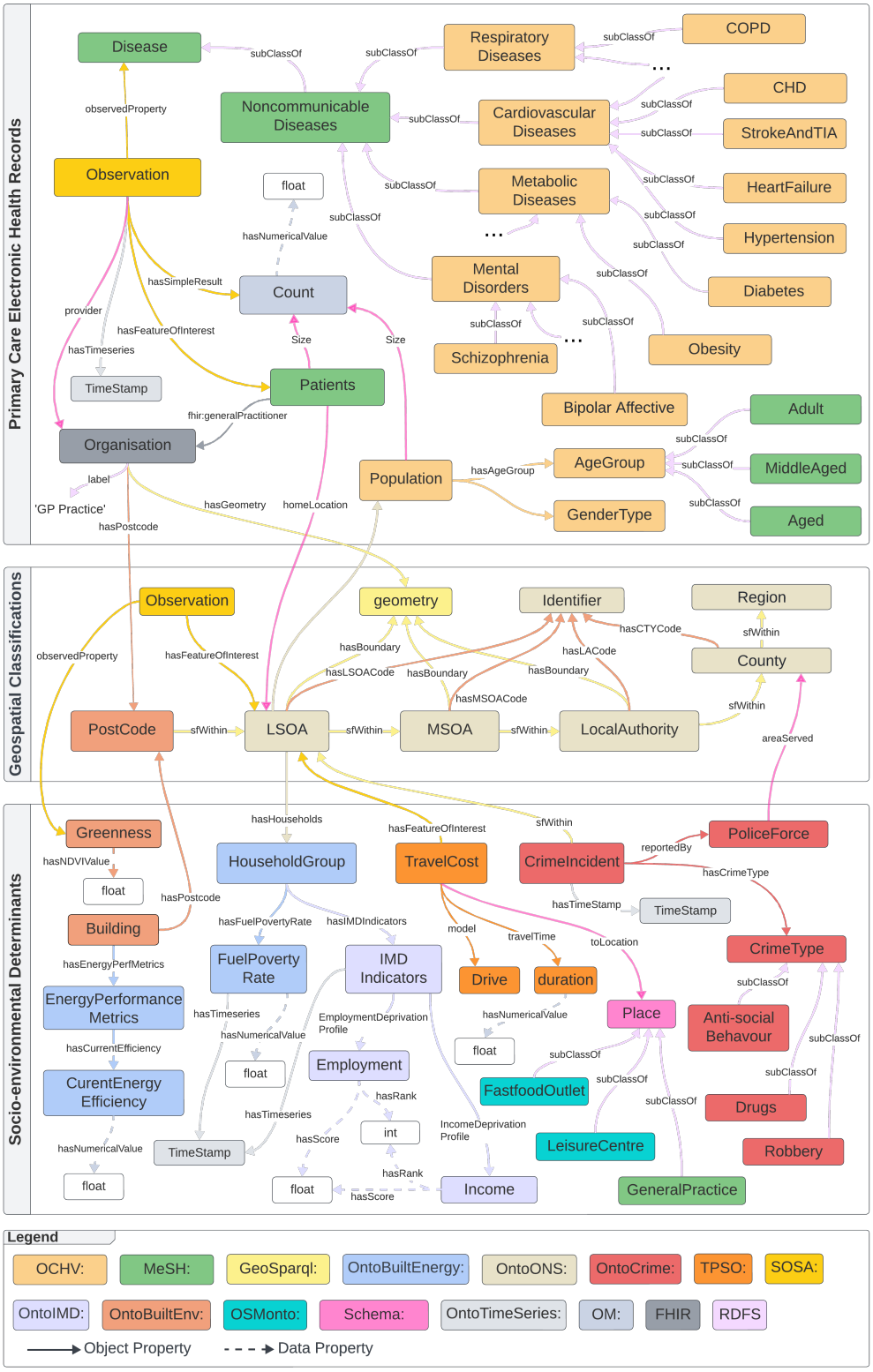


Figure 2: *A multidimensional semantic framework integrating UK primary care electronic health records with socio-environmental determinants of health.*

To obtain adjusted effect estimates on an absolute scale the analysis uses Poisson prevalence models with a log link and an offset for the LSOA denominator. Exposures that are proportions are mapped to a logit transformed scale and enter through restricted cubic spline bases. Each model includes the standard set of ecological covariates namely the age composition parameters the six category urban rural classification and additional scalar covariates such as deprivation and energy efficiency scores. Model based prevalence functions and adjusted prevalence ratios are defined in Appendix A.4.1 and Appendix A.4.3. The main summaries reported in the text are adjusted prevalence ratios for interquartile range contrasts which compare model based prevalences between the first and third quartiles of each exposure while holding covariates at a specified reference profile. For each exposure a likelihood ratio test is used to assess whether spline terms are required beyond a linear effect in the log scale predictor.

Dose response analysis is based on the same Poisson prevalence specification. For each condition a reference covariate profile is fixed in terms of urban rural indicators age band shares and additional covariates. The fitted model is then evaluated on a fine grid of exposure values within the empirical range after stabilisation and logit transformation of the exposure proportion. This yields model based prevalence functions that describe the exposure response pattern at the reference profile as detailed in Appendix A.3. Differences in these curves between selected exposure values correspond to adjusted prevalence ratios on the same scale as those reported for interquartile range contrasts.

Population attributable fractions are derived from counterfactual applications of the fitted Poisson models. For each exposure several percentile based intervention scenarios are considered that cap the exposure distribution at chosen empirical quantiles while leaving covariates and denominators unchanged. Observed and counterfactual expected case counts are obtained by summing the corresponding model based means across LSOAs and the population attributable fraction is computed as the proportional reduction in expected cases under the capped scenario relative to the observed scenario. The construction of these counterfactuals and the definition of the attributable fraction are set out in the Appendix section on population attributable fraction under percentile capping.

Model fit is summarised using the strict McFadden pseudo R^2 which compares the maximised log likelihood of the full Poisson model with that of a null model containing only an intercept and the same offset. The definition and computation of this measure are given in the Appendix section on strict McFadden pseudo R^2. Together these elements provide a coherent exposure response and attributable risk framework that links correlation analysis regression based effect estimation dose response curves and population attributable fractions within a single modelling strategy.

4 Result

The application of the semantic framework successfully harmonised heterogeneous data streams across approximately 34,000 Lower Layer Super Output Areas covering a registered population exceeding 50 million. This integration generated a unified high-resolution evidence base linking primary care prevalence estimates with concurrent environmental exposures. The findings are presented in a sequence that progresses from descriptive

epidemiology to specific exposure response characterisation. We first report the spatial distribution of disease burden at the national scale followed by an evaluation of associations between community health opportunities and metabolic cardiovascular and respiratory profiles. The analysis subsequently quantifies the impact of fuel poverty on these physical health conditions before concluding with an examination of the environmental determinants of severe mental illness.

4.1 Disease prevalence aggregated from general practice records

Figure 3 shows the national distribution of LSOA-level prevalence for metabolic, cardiovascular and chronic respiratory conditions. Within the metabolic group, diabetes and obesity exhibit similar spatial distributions, with overlapping clusters of high-prevalence LSOAs. Within the cardiovascular group, coronary heart disease, stroke or transient ischaemic attack, and peripheral arterial disease also show related spatial patterns. Across all mapped conditions, hypertension, obesity and diabetes are the most prevalent and therefore represent key targets for intervention to alleviate the national non-communicable disease burden.

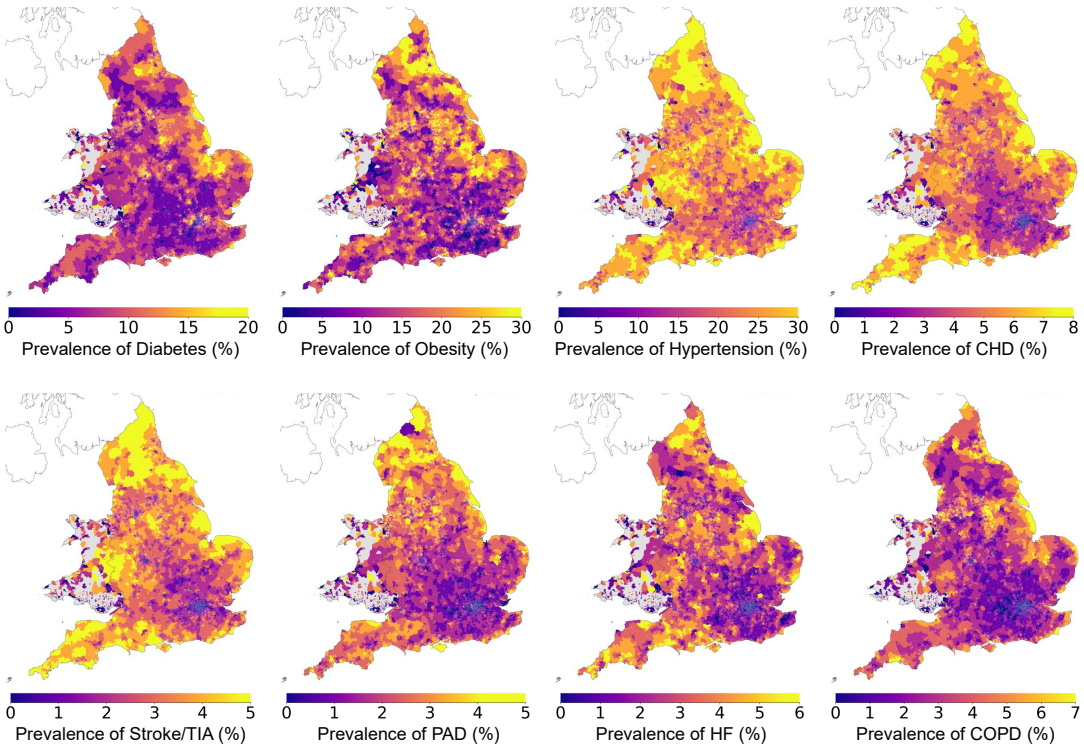


Figure 3: *Aggregated prevalence maps by disease group.* ***Metabolic conditions****: Diabetes, Obesity and Hypertension.* ***Cardiovascular diseases****: Coronary heart disease (CHD), Stroke/transient ischaemic attack (Stroke/TIA), Peripheral arterial disease (PAD), and Heart failure (HF).* ***Chronic respiratory disease****: Chronic obstructive pulmonary disease (COPD). Colour bars report prevalence (%).*

4.2 Limited healthy opportunities and cardiometabolic/respiratory Health

Building on the spatial heterogeneity in disease prevalence shown in Figure 3, this section focuses on three community features that can be interpreted as proxies for everyday health opportunities at the LSOA level. These are neighbourhood greenness, represented by NDVI, the drive time to the nearest fast food outlet, and the drive time to the nearest leisure centre. They are selected as representative examples because they capture three distinct aspects of the local environment that are relevant to physical activity and diet, and because consistent, nationwide data are available from routine sources. The aim is to describe how these features vary across space, how they co-vary with chronic disease prevalence, and how they relate to model based risk estimates.

Figure 4 summarises the spatial distribution of these three exposures and their stratification by rural and urban status. NDVI is systematically higher in rural LSOAs than in urban areas, with rural medians around the upper half of the observed range and urban medians closer to the lower half. In metropolitan cores and surrounding periurban belts, drive time to the nearest fast food outlet is typically short, with most LSOAs within about five minutes, whereas many remote rural LSOAs have drive times above ten minutes. A similar pattern is seen for leisure centres. Urban and periurban LSOAs generally have drive times under ten minutes, while in more isolated rural areas a substantial fraction of LSOAs have drive times above twenty minutes. These gradients indicate that rural areas tend to have greater green space but poorer access to both fast food and formal leisure facilities, whereas dense urban areas combine lower greenness with closer proximity to both fast food outlets and leisure centres.

Associations between these exposures and metabolic and cardiovascular conditions on the correlation scale are shown in Figure 5, using the partial rank correlation framework described in Appendix A.2 and Appendix A.2.4. Greater greenness is negatively associated with the prevalence of obesity, diabetes and hypertension after adjustment for age structure, urban rural classification and other ecological covariates, consistent with lower disease prevalence in greener areas at a given demographic and socioeconomic profile. Shorter drive time to fast food outlets tends to be positively correlated with metabolic outcomes, indicating higher prevalence where fast food is more accessible. In contrast, shorter drive time to leisure centres is generally associated with lower prevalence for the same set of conditions. Across diseases, correlations are largest in magnitude for obesity and diabetes and smaller for cardiovascular conditions such as coronary heart disease and stroke. Correlations are moderate in absolute value, which is expected for ecological data aggregated at the LSOA level and reflects that these environmental features are only one component of the wider risk landscape.

To translate these associations into effect estimates on the risk scale, Figure 6 reports adjusted prevalence ratios from multivariable Poisson models with a log link and LSOA denominators as offsets, as specified in Appendix A.4.1 and Appendix A.3.

Effect sizes are expressed per interquartile range, where the interquartile range is defined as the difference between the seventy fifth and twenty fifth percentiles of the observed exposure distribution. For greenness, an interquartile range increase in NDVI, approxi-

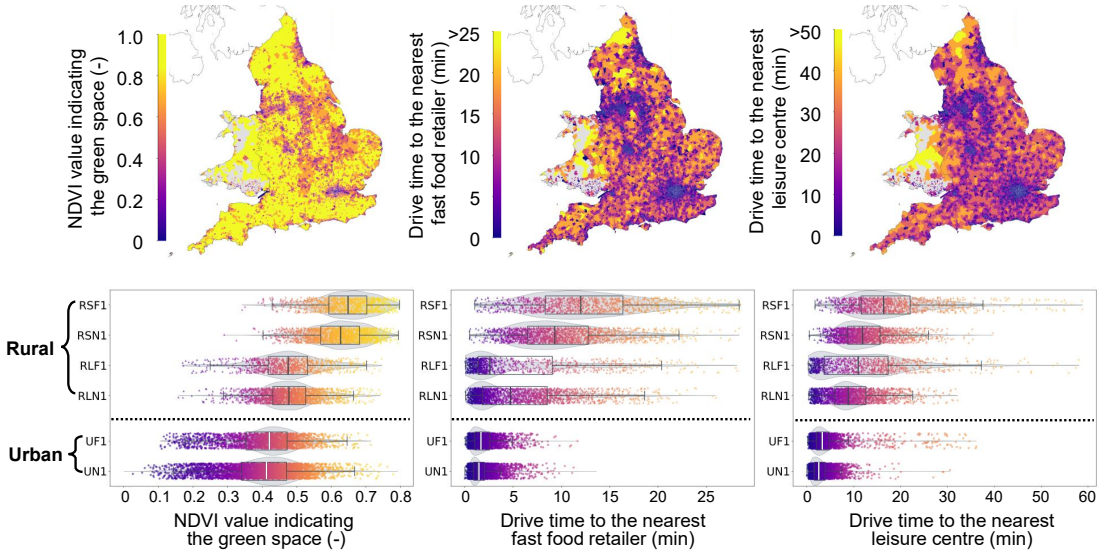


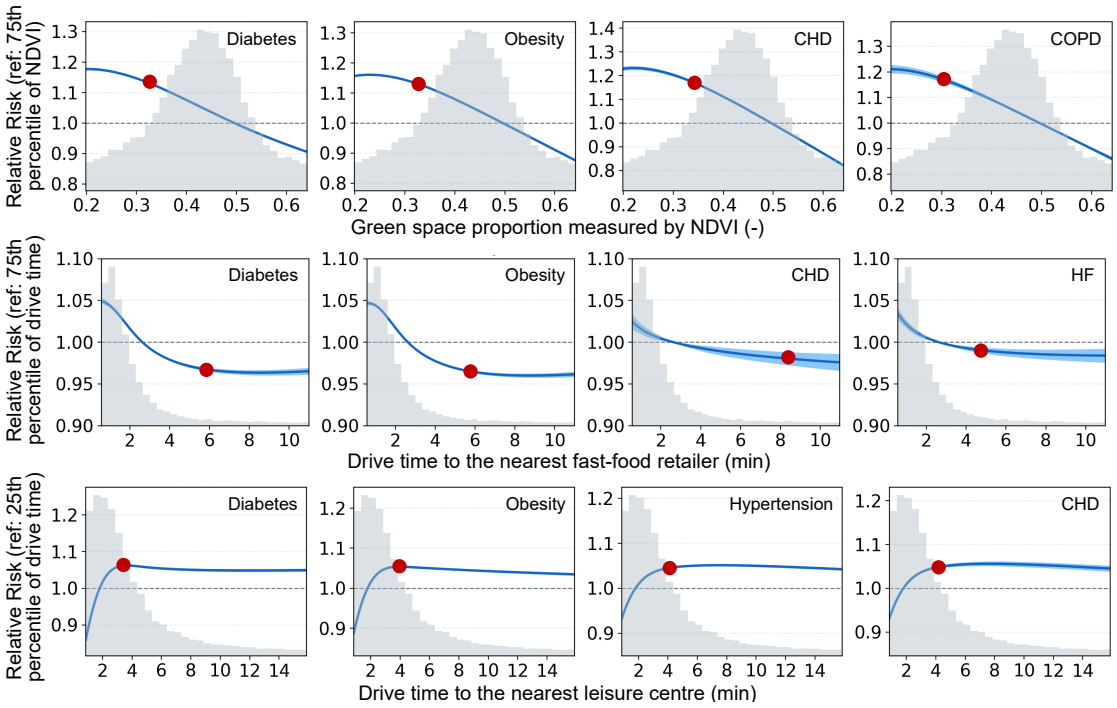
Figure 4: *Spatial distribution and urban-rural contrasts in neighbourhood greenness and accessibility across England. Top row (left to right): LSOA-level greenness (normalised difference vegetation index, NDVI), drive time to the nearest fast-food outlet, and drive time to the nearest leisure centre. The bottom row shows the corresponding exposure distributions stratified by the Rural–Urban Classification (RUC) published by Office for National Statistics*

mately 0.13 NDVI units in this sample, is associated with lower prevalence for several conditions. The model based estimates indicate that such an increase in NDVI corresponds to about a 7.8% lower prevalence of COPD (aPR=0.922), a 7% lower prevalence of diabetes (aPR=0.930) and a 7.3% lower prevalence of obesity (aPR=0.927), after controlling for age composition, urban rural status and additional ecological covariates. Other metabolic and cardiovascular outcomes also show adjusted prevalence ratios below one for higher greenness, although with smaller magnitude and wider confidence intervals. For drive time to fast food outlets, an interquartile range increase in travel time, approximately two minutes, is associated with adjusted prevalence ratios below one for several metabolic conditions, consistent with slightly lower prevalence where fast food is less accessible. For drive time to leisure centres, an interquartile range increase in travel time, approximately four minutes, is associated with adjusted prevalence ratios above one for metabolic and cardiovascular outcomes, indicating higher prevalence where formal leisure facilities are harder to reach. McFadden pseudo R^2 values in panel (d) remain modest in absolute terms but are highest for obesity and diabetes for all three domains, suggesting that these exposures explain a non trivial share of between LSOA variation for metabolic diseases while only a smaller fraction for individual cardiovascular endpoints.

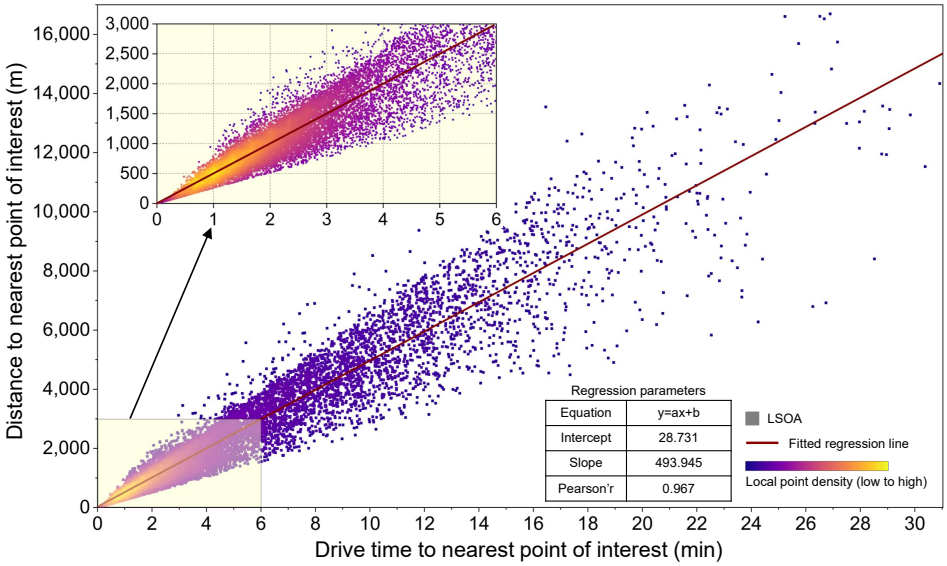
Figure 7 applies the exposure-prevalence association modelling framework described in Appendix A.3, with relative risk interpretation in Appendix A.4.3, to characterise nonlinear association shapes for selected chronic conditions. Relative risks are evaluated across the observed exposure range with covariates held at reference values and expressed relative to health favourable baselines. For greenness, using the 75th percentile of NDVI as

the reference, relative risk for obesity and diabetes is close to 1 around this level, increases at lower NDVI, and flattens at higher NDVI, consistent with diminishing marginal gains in very green settings. For fast food accessibility, with the 75th percentile of drive time as the reference, relative risk decreases as drive time increases from very short values and then approaches a plateau once drive time exceeds about 6 to 8 minutes. For leisure centre accessibility, with the 25th percentile of drive time as the reference, LSOAs with drive times of about 4 minutes or less to the nearest leisure centre show clearly lower relative risks for metabolic and cardiovascular outcomes than areas where leisure centres are more distant.

These profiles provide planning relevant benchmarks that fall within realistic exposure ranges. For greenness, risk begins to decline more clearly once NDVI exceeds about 0.30, suggesting that the protective association becomes more evident beyond this level and is consistent with sustained vegetative cover such as continuous parks, street trees, and larger gardens. For accessibility metrics, Figure 7b links time based thresholds to road network distance using a consistent routing workflow. Drive time estimates follow the accessibility routing approach developed by The Geographic Data Service for postcode-based car travel time to the nearest service along the established road network [31]. Using the same shortest path routing algorithm, with points of interest as sources and postcodes as targets, we extended the workflow to output road network distance alongside the drive time, then summarised postcode level outputs within each LSOA as medians across constituent postcodes. On this basis, Panel (b) of the Figure 7 indicates that a median leisure centre drive time of about 4 minutes corresponds to approximately 2000 m of road network distance, whereas a fast food drive time of about 6 minutes corresponds to approximately 3000 m. These distance equivalents translate model derived pattern features into actionable neighbourhood scale guidance, supporting decisions on leisure facility provision and network connectivity and informing place based planning and licensing approaches that manage the proximity and clustering of fast food outlets.



(a) Exposure-prevalence association for conditions with statistically significant adjusted associations.



(b) LSOA-level point-of-interest accessibility: drive time versus road-network distance.

Figure 7: *(a) Relative risk across the exposure range, referenced to health-favourable baselines. Grey bars show the distribution of LSOAs across exposure bins; red markers indicate potential policy-relevant intervention points (b) LSOA-level median postcode-to-nearest-point-of-interest drive time versus median road-network distance, with both metrics summarised as medians across constituent postcodes; points of interest include fast-food outlets and leisure centres.*

4.3 Cold homes and cardiometabolic/respiratory diseases

As shown in Figure 8, fuel poverty displays marked spatial and contextual heterogeneity. Higher proportions of fuel poor households cluster in northern and south western regions, with additional concentrations in parts of the Midlands. London and other major conurbations show pronounced within city variation, where neighbouring LSOAs can differ by more than 10 percentage points in fuel poverty ratios. The rural urban panel confirms a clear gradient by settlement type, with the highest fuel poverty generally observed in smaller and more remote rural areas and lower values in urban areas nearer to major cities. This pattern indicates that any assessment of health impacts needs to account for settlement context rather than treating fuel poverty as spatially homogeneous.

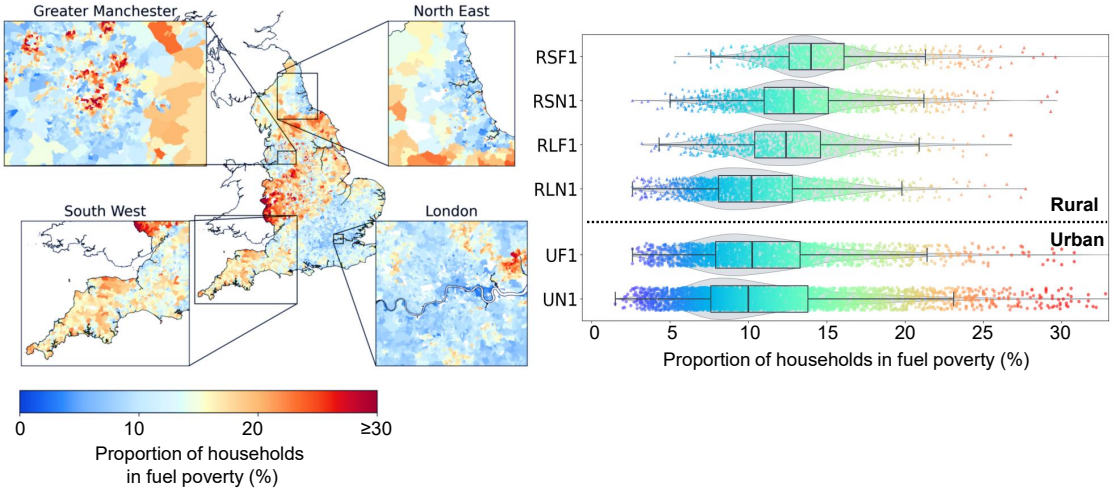


Figure 8: *Spatial distribution and rural–urban stratification of fuel poverty, obtained from the integration of data published by the Department for Energy Security and Net Zero (DESNZ) and the Office for National Statistics (ONS) under the Rural–Urban Classification (RUC). The upper panels show the regional variations in the proportion of households in fuel poverty. The lower panel compares urban and rural areas based on the Rural–Urban Classification (RUC), illustrating distinct distributions of fuel poverty rates across settlement types.*

Figure 9 summarises the rank based associations between fuel poverty and cardiometabolic outcomes using the partial correlation framework described in Appendix A.2 with population weighting as in Appendix A.2.4. Across metabolic and cardiovascular conditions the strongest positive correlations with fuel poverty are observed for chronic obstructive pulmonary disease, peripheral arterial disease, diabetes and obesity, with more modest associations for coronary heart disease, stroke or transient ischaemic attack, hypertension and heart failure. Within each disease group the ordering of correlations is broadly consistent, suggesting that areas with higher fuel poverty tend to carry a higher burden of chronic disease even after accounting for age structure, rural urban classification and other ecological covariates.

The multivariable Poisson prevalence models underpinning the main effect estimates are reported in Figure 10 and follow the adjusted prevalence ratio formulation in Appendix A.4.1. Panel (a) shows that an interquartile range (IQR) increase in fuel poverty (approximately 6 percentage points) is associated with the largest relative increase in COPD prevalence (adjusted prevalence ratio 1.197). Peripheral arterial disease follows with an adjusted prevalence ratio of 1.153. Diabetes and obesity show adjusted prevalence ratios of 1.149 and 1.147. Smaller but statistically precise associations are observed for coronary heart disease (aPR=1.071), hypertension (aPR=1.040), stroke or transient ischaemic attack (aPR=1.037), and heart failure (aPR=1.026). Panel (b) reports McFadden pseudo R^2 values for the same models, which range from 0.151 to 0.406, indicating that the ecological models capture a substantial share of the between area variation in prevalence.

Panel (c) of Figure 10 translates these associations into population attributable fractions using the percentile capping counterfactual framework set out in the Appendix section on population attributable fraction under percentile capping, combined with the baseline relative risk formulation in Appendix A.4.3. Scenario P50 resets every LSOA to the fuel poverty level typical of the healthier half of areas, while scenarios P40, P30, P20 and P10 progressively tighten the reference toward the levels observed in the healthiest decile. Under these counterfactuals the largest preventable burdens arise for chronic obstructive pulmonary disease and peripheral arterial disease, with double digit attributable fractions for chronic obstructive pulmonary disease under the most ambitious P10 scenario and appreciable fractions already evident under P50. For metabolic conditions such as diabetes and obesity the estimated population attributable fractions are smaller in absolute terms but remain non negligible when interpreted at national scale.

Taken together Figures 8, 9 and 10 show that higher fuel poverty aligns with a broad gradient in population health, with more fuel poor areas experiencing higher prevalence of cardiometabolic and chronic respiratory disease. While these ecological analyses cannot establish individual level causation, the consistent patterns across correlation, adjusted prevalence ratios and population attributable fractions suggest that reducing fuel poverty and improving the energy performance of housing are likely to be relevant components of strategies to lessen the burden of chronic disease.

4.4 Neighbourhood exposures and severe mental illness

Population-weighted associations adjusted for age and urban rural status show a consistent positive relation between crime exposure and the prevalence of mental health conditions. The steepest gradients are observed for drug-related offences with a partial correlation of approximately 0.23 and violence and sexual offences at 0.22 which exceed those for burglary public order robbery theft from the person and possession of weapons ranging from 0.13 to 0.21. The stability of these modest effect sizes across panels after trimming extreme values and controlling for demographic structure indicates that the signal is not driven by age composition or settlement type. The heterogeneity by offence class suggests that mechanisms linked to local drug markets and interpersonal violence domains including domestic abuse align more closely with area level mental health burden than acquisitive crime.

Figure 11 further characterises broader environmental determinants associated with mental well-being beyond criminal activity. A significant positive association is observed between fuel poverty and severe mental illness, with a partial correlation of 0.11, reinforcing the link between material deprivation and psychological morbidity. The association profiling for green space indicates a distinct non-linear relationship that contrasts with the monotonic protection observed for physical health. Prevalence of severe mental illness exhibits a U-shaped association with vegetation density, with the lowest prevalence within an intermediate NDVI range of approximately 0.5 to 0.6. This non-linear pattern suggests that while moderate greenness may be associated with protective benefits, both very low vegetation and very high vegetation density are associated with higher severe mental illness prevalence in this ecological context.

These ecological results do not establish causation and may reflect unmeasured confounding *e.g.* deprivation service access and policing intensity or bidirectional influences. Prioritising violence prevention and drug harm reduction alongside mental health service provision may nonetheless yield the greatest mental health gains at neighbourhood scale compared with strategies focused on property crime.

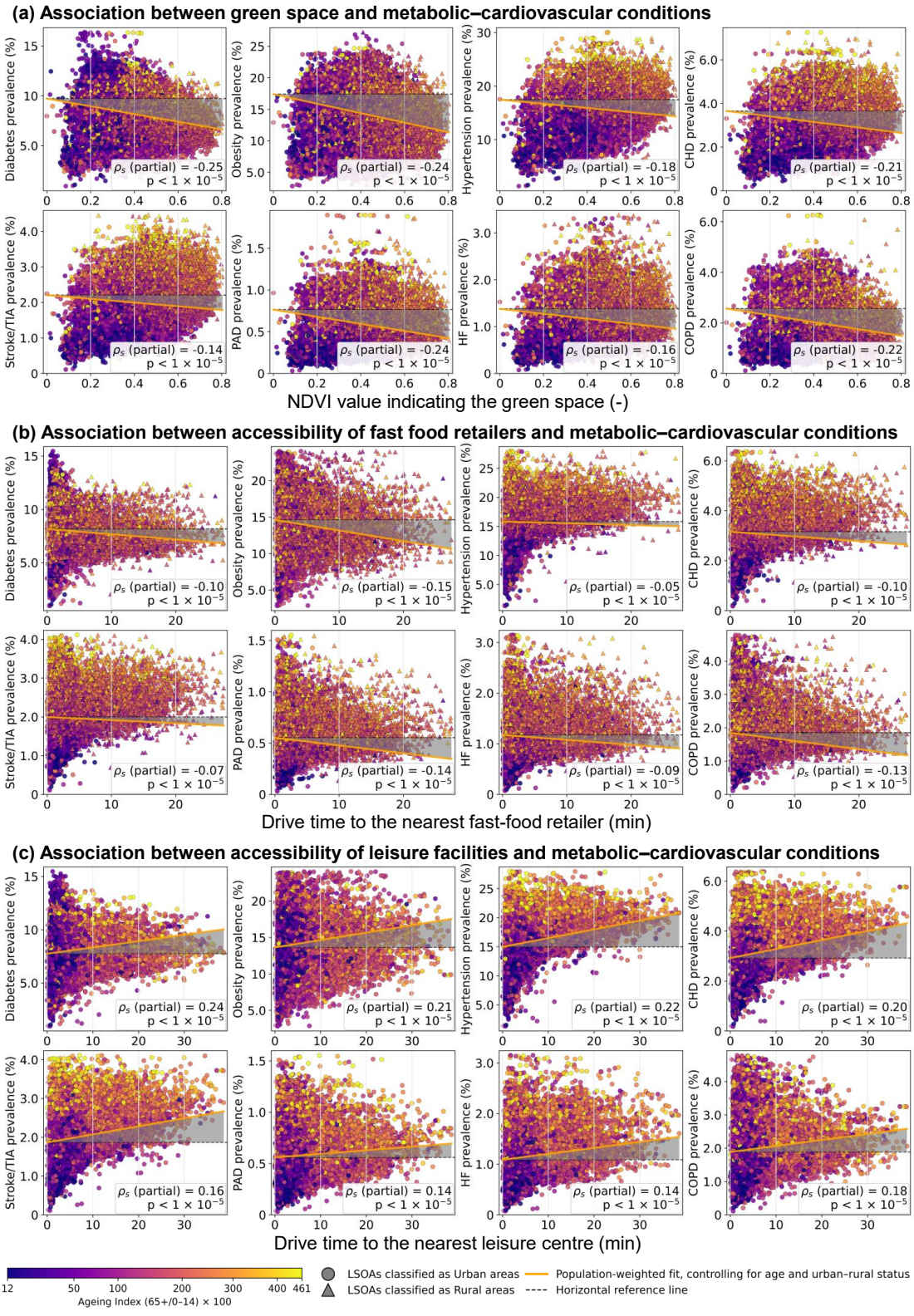


Figure 5: *Associations of neighbourhood greenness and accessibility to health-relevant amenities with cardiometabolic and respiratory conditions.*

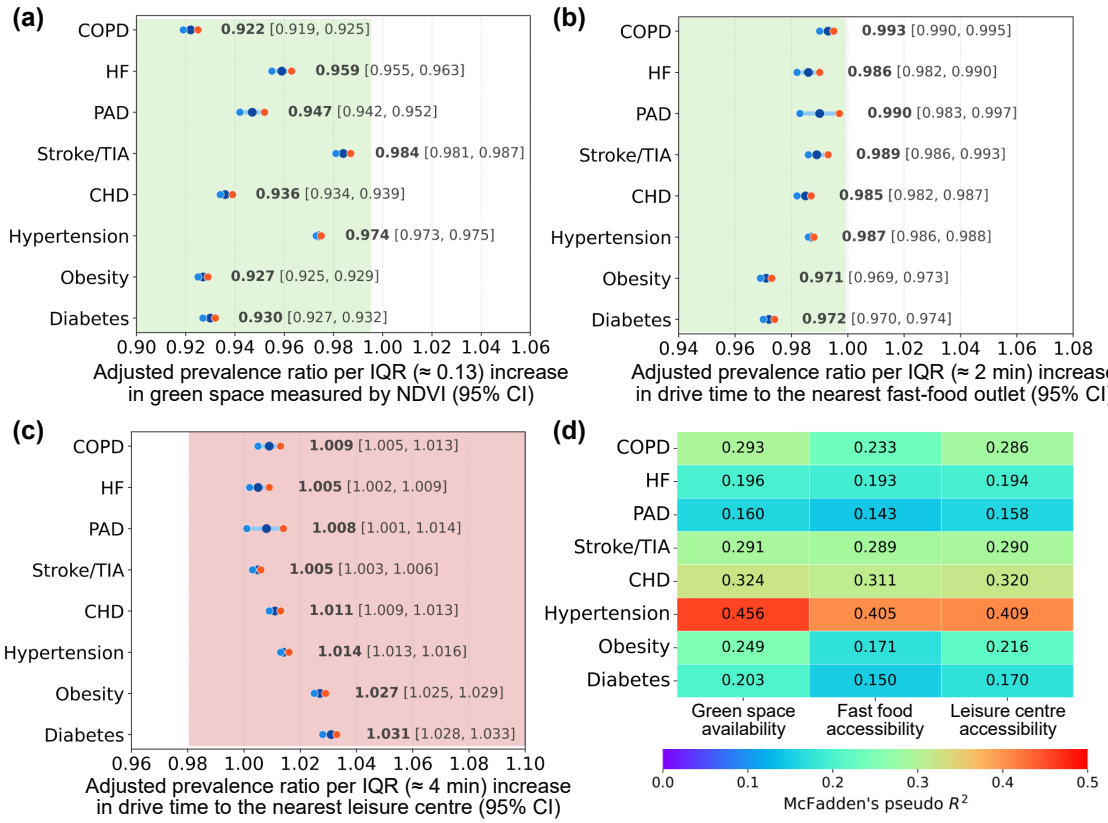


Figure 6: *Risk of cardiometabolic and respiratory conditions in relation to neighbourhood greenness and accessibility to health-relevant amenities. Panels (a),(b) and (c) present adjusted effect estimates with 95% CIs from multivariable models: (a) adjusted prevalence ratio per IQR in NDVI (green-space availability); (b) adjusted prevalence ratio per IQR in drive time to the nearest fast-food outlet; and (c) adjusted prevalence ratio per IQR in drive time to the nearest leisure centre. Values < 1 indicate lower risk and > 1 higher risk. Panel (d) shows McFadden's pseudo-R^2 by outcome and domain.*

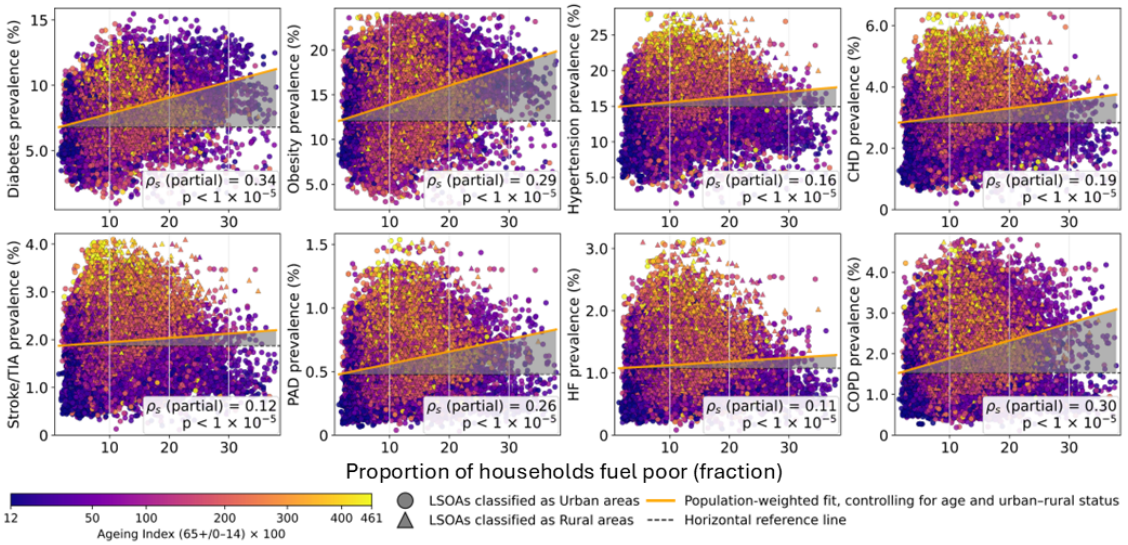


Figure 9: *Associations of fuel poverty ratio with cardiometabolic and respiratory conditions.*

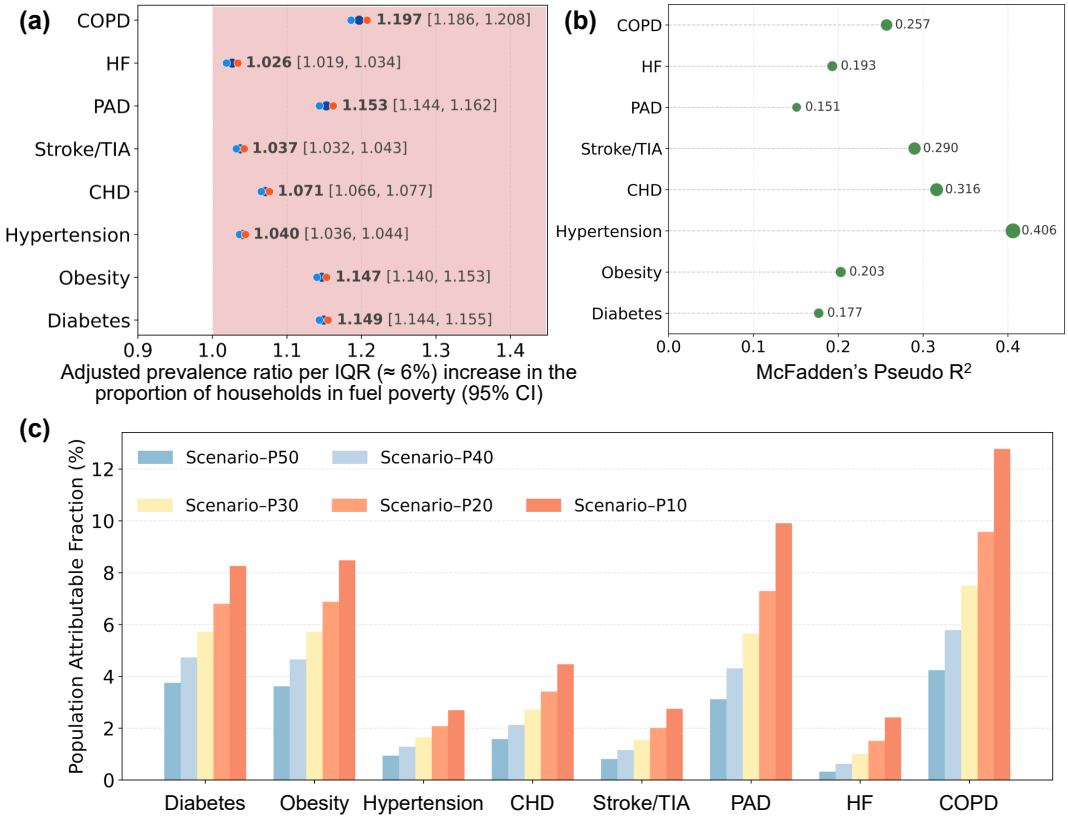


Figure 10: *(a) Adjusted prevalence ratios with 95% confidence intervals. (b) Model performance for each condition measured by McFadden pseudo R^2. (c) Population attributable fractions for counterfactual scenarios that reset area fuel poverty to healthier reference levels.*

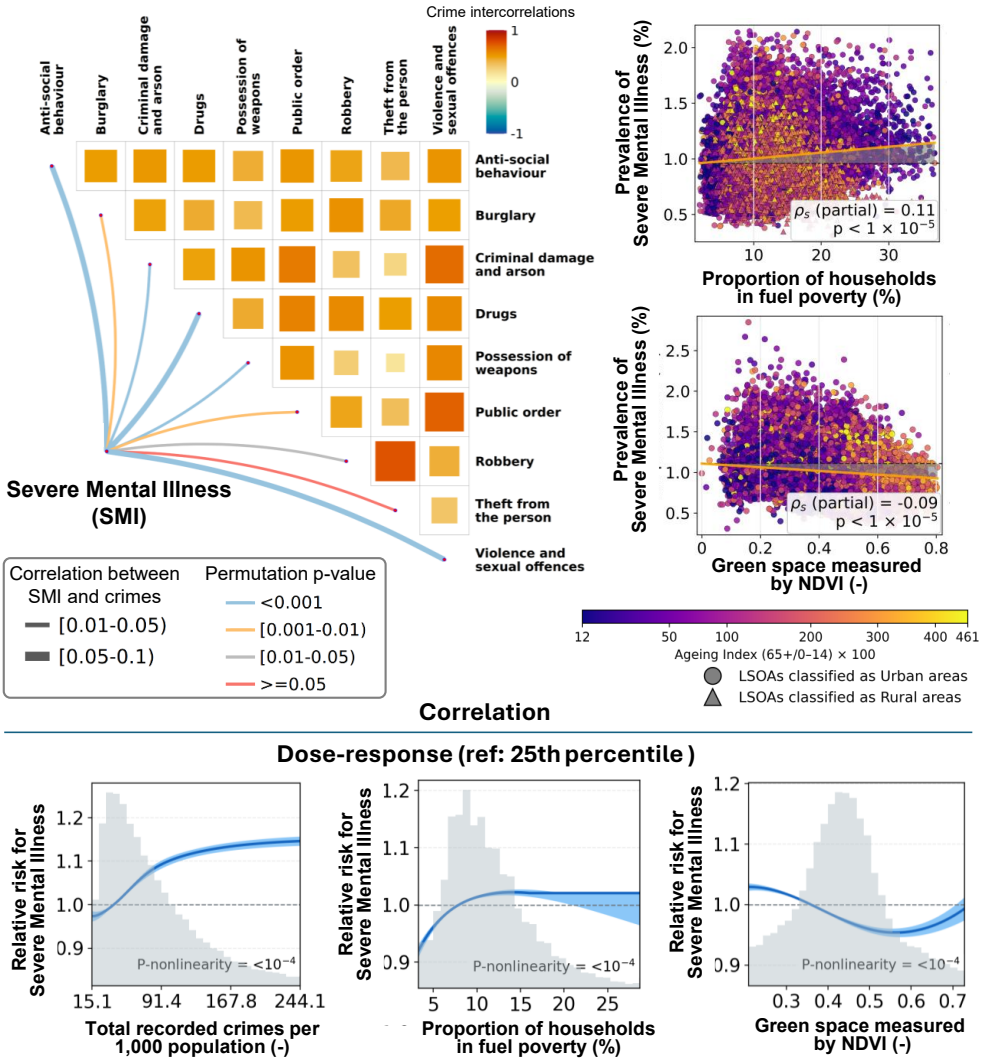


Figure 11: *Multi-domain determinants associated with severe mental illness (SMI) prevalence across LSOAs in England. Top row: Correlation and adjusted association summaries between SMI prevalence and neighbourhood determinants spanning recorded offences (total offences per 1,000 population and selected crime domains), fuel poverty, and green space (NDVI), with adjustment for demographic and socioeconomic covariates as specified in the main model. Bottom row: Exposure-range* association *profiles between SMI prevalence and total recorded offences per 1,000 population, fuel poverty, and green space estimated using restricted cubic spline models. Curves are expressed as adjusted prevalence ratios relative to the 25th percentile of each exposure, with covariates held at reference values. The crime and fuel-poverty curves show strong, monotonic increases in SMI prevalence, while green space exhibits a U-shaped association with the lowest prevalence at intermediate vegetation density (NDVI ≈0.5–0.6).*

5 Discussion

This nationwide small-area study used a semantic knowledge graph to link neighbourhood determinants with non-communicable disease prevalence in the United Kingdom. We combined primary-care disease registers, practice registration patterns, and multiple environmental indicators within one interoperable infrastructure and then estimated adjusted associations while accounting for key covariates. Interpretation requires caution because the analysis is ecological and cross-sectional; nevertheless, the integrated evidence base reveals consistent, policy-relevant patterns that connect modifiable determinants to health outcomes.

First, housing energy deprivation showed consistent adjusted associations with cardiometabolic and respiratory morbidity. Fuel poverty, as a policy-relevant proxy for the ability to heat homes adequately, was associated with the highest estimated attributable burden for chronic obstructive pulmonary disease (COPD) among the health outcomes assessed in adjusted small-area (LSOA) models accounting for age structure, urban-rural status, healthcare accessibility, and socioeconomic conditions. Scenario-based modelling consistent with improvements in home heating efficiency suggests that reducing fuel poverty could potentially avert more than 13% of the population COPD burden. Overall, these patterns are consistent with housing warmth and energy affordability as upstream policy levers for respiratory health, and they support integrating energy and housing policy into chronic disease prevention strategies.

Second, spatial profiling of the built environment identified actionable proximity thresholds that translate directly to place-based intervention. Better access to leisure centres was associated with lower metabolic and cardiovascular burden when available within an approximately 4-minute drive at the reference covariate profile. In contrast, the excess cardiometabolic burden linked to fast-food outlet proximity diminished beyond roughly 6 minutes of drive time. Beyond this point, additional distance was associated with little further change. Framed as determinants-to-outcomes evidence, these benchmarks suggest that the largest differences in model-based burden occur within realistic and tractable ranges of accessibility, rather than only at extremes of the exposure distribution.

Third, mental-health burden showed strong, monotonic associations with neighbourhood harm and insecurity. Areas with higher recorded crime and higher fuel poverty had higher prevalence of severe mental illness even after control for demographic and socioeconomic covariates, with particularly steep gradients for antisocial behaviour, domestic abuse, and drug-related crime. Green space supported metabolic health but exhibited a U-shaped association with severe mental illness, with the lowest prevalence at intermediate vegetation density (NDVI $\approx$0.5–0.6). This pattern is consistent with the possibility that both very low greenness and very high greenness, which may coincide with isolation, limited services, or restricted access, can be linked to higher mental-health burden. Intermediate greenness may reflect environments that better balance access, activity opportunities, and social amenity.

This work contributes methodologically by demonstrating how semantic integration can support holistic, multi-determinant environmental health surveillance using the government data that are currently available. Rather than analysing exposures in isolation, the

framework allows multiple, correlated domains, namely housing energy deprivation, the food environment, physical-activity opportunities, greenness, and neighbourhood safety, to be queried and analysed together on a common geography, while controlling for key covariates. The result is a reproducible, extensible evidence layer that supports consistent comparisons from local to national scale.

The study has several strengths. It uses routinely collected health data and public environmental indicators to construct a nationwide small-area evidence base. Disease prevalence was derived from general practice registers and apportioned to residential areas through established methods, then analysed with adjustment for age composition, urban–rural category, healthcare accessibility, and socioeconomic status. Environmental indicators followed transparent, reproducible definitions based on road-network travel time, satellite-derived greenness indices, fuel poverty statistics, and recorded crime measures.

Association profiles enabled exploration of non-linear shapes and identification of simple quantitative benchmarks, including the approximate 4-minute access threshold for leisure centres, the $\sim$6-minute threshold beyond which fast-food proximity associations attenuate, a minimum greenness level relevant for metabolic health, and an intermediate NDVI band (0.5–0.6) associated with the lowest severe mental illness prevalence.

Important limitations remain. The analysis is ecological and cannot attribute risk at the individual level or rule out residual confounding. Time ordering between exposure and outcome cannot be established from prevalence snapshots, and the scenario-based impact metrics should be interpreted as model-based contrasts conditional on the adjustment set rather than causal effects. People do not live their lives only within LSOA boundaries. Daily mobility may dilute or modify residence-based indicators, and drive-time metrics capture potential access rather than realised behaviour. Apportionment of practice-level registers to small areas may introduce error where registration and residence are misaligned, although such error would usually bias associations towards the null. Environmental indicators represent specific years while disease registers and demographic covariates come from overlapping but not identical periods. Greenness measures describe vegetation density but not quality, safety, accessibility, or patterns of use. Crime data reflect recorded events and can be influenced by reporting and enforcement as well as underlying disorder.

Despite these limitations, the analysis reflects what can currently be achieved with nationally available government data: an integrated, reproducible evidence layer that supports place-based planning and prioritisation and highlights hypotheses for follow-on evaluation. By linking determinants to outcomes in a consistent small-area framework, the results support the view that housing warmth, community safety, and everyday opportunities for activity and diet correspond to measurable differences in chronic disease burden across communities. In practice, fuel poverty reduction, violence and disorder prevention, and improvements to activity-supportive infrastructure and healthier food environments can be viewed as complementary levers for population health.

Future work can extend this framework in several directions as richer public releases and linkage pathways emerge. First, routine ingestion of updated government datasets can support repeated measurements of exposures and disease prevalence over time, strengthening longitudinal inference and enabling evaluation of housing, safety, and planning

interventions. Second, the existing exposure layer can be expanded to incorporate additional behavioural signals (e.g., activity context and multimodal movement traces), improving the stability and comparability of personalised exposure estimates across settings and periods. Third, secure linkage to individual-level cohorts and longitudinal primary-care records can operationalise a digital cohort, in which participants are followed through routine data flows while maintaining a harmonised environmental evidence layer—supporting stronger causal designs (e.g., natural experiments) and more timely assessment of local interventions.

Acknowledgements

This research was supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. Financial support from the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/Y016076/1, the European Union's Horizon Europe research and innovation programme under grants 101058732 (JIDEP), 101074004 (C2IMPRESS), and 101188248 (CLIMATE-ADAPT4EOSC), the Medical Research Council (MC_UU_00006/4) and the National Institute of Health and Care Research (NIHR) Cambridge Biomedical Research Centre (NIHR203312) is also gratefully acknowledged. Markus Kraft gratefully acknowledges the support of the Alexander von Humboldt Foundation. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT GPT-4-turbo in order to enhance the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data and code availability

The codes developed for this work are available under an open-source licence on GitHub in The World Avatar repository https://github.com/cambridge-cares/TheWorldAvatar. The datasets used in this work are freely available for download as per the references in the paper.

Conflicts of interest

There are no conflicts to declare.

A Appendix

A.1 Estimation of LSOA disease prevalence

This section outlines the procedure used to estimate disease prevalence at the Lower Layer Super Output Area (LSOA) level. The objective is to redistribute aggregated case counts from GP practices to local geographical areas by linking clinical records with patient residence data. The calculations were performed using code provided by the House of Commons Library [37].

We use two datasets from NHS Digital to achieve this mapping. The first is the *Quality and Outcomes Framework (QOF)* database [37], which provides total counts of diagnosed cases for specific medical conditions at each GP practice. The second is the *Patients Registered at a GP Practice by LSOA* dataset [63], which reports the number of patients from each practice residing in specific LSOAs. While these datasets are generally aligned, discrepancies occasionally arise where the sum of patients mapped to LSOAs in the residence dataset does not perfectly match the official practice register in the QOF dataset due to administrative inconsistencies. Consequently, a proportional scaling procedure is applied to reconcile these counts before prevalence is calculated.

Let $j = 1,\ldots,P$ index GP practices, $l = 1,\ldots,L$ index LSOAs, and $c = 1,\ldots,C$ index medical conditions. The input data consist of $O_{j,c}$, the observed case count at practice j for condition c; n_j, the total number of patients registered at practice j (both from the QOF database); and $n_{j,l}$, the raw count of patients registered at practice j who reside in LSOA l (from the Patients Registered at a GP Practice by LSOA dataset).

First, to address potential inconsistencies in the raw records, patient counts are reconciled to match practice totals

$$m_{j,l} \;=\; n_{j,l}\,\frac{n_j}{\sum_{l'=1}^{L} n_{j,l'}}, \tag{A.1}$$

where $m_{j,l}$ is the reconciled number of patients from practice j residing in LSOA l. By construction, $\sum_{l=1}^{L} m_{j,l} = n_j$.

Next, we compute the proportion of each practice's registered population residing in each LSOA

$$\varphi_{j,l} \;=\; \frac{m_{j,l}}{n_j}, \tag{A.2}$$

where $\varphi_{j,l}$ denotes the share of practice j's register associated with LSOA l.

Using these proportions, practice-level cases are allocated to LSOAs

$$N_{c,l} \;=\; \sum_{j=1}^{P} \varphi_{j,l}\, O_{j,c}, \tag{A.3}$$

where $N_{c,l}$ is the allocated number of cases of condition c in LSOA l. This allocation assumes that, within each practice, the probability of diagnosis for condition c is proportional to the number of registered patients residing in each LSOA (*i.e.* no within-practice geographic heterogeneity in risk). Because the allocation uses fractional shares $\varphi_{j,l}$, the resulting $N_{c,l}$ need not be an integer and is interpreted as an estimated case count.

The total population for each LSOA is then obtained as

$$d_l = \sum_{j=1}^{P} m_{j,l}, \tag{A.4}$$

where d_l is the modelled registered population residing in LSOA l.

Finally, the LSOA-level prevalence is calculated as

$$P_{c,l} = \frac{N_{c,l}}{d_l}, \tag{A.5}$$

where $P_{c,l}$ denotes the allocated prevalence proportion for condition c in LSOA l implied by the GP-to-LSOA allocation. These LSOA-level population d_l and prevalence estimates $P_{c,l}$ are used in the ecological association analyses that follow.

A.2 Exposure–prevalence association analysis

We quantify associations between area-level exposures and disease prevalence. Rank-based correlations are used for robust association summaries, while the population-weighted linear regressions are used only to produce interpretable adjusted trend lines in figures. We first define the ecological covariates (Section A.2.1), then introduce Spearman's rank correlation (Section A.2.2), extend to partial correlations adjusted for covariates (Section A.2.3), and finally to population-weighted correlations (Section A.2.4).

A.2.1 Ecological covariates

The design matrix K comprises LSOA-level covariates included to adjust for potential confounding in the association between ecological exposure and disease prevalence. The construction and encoding of these covariates are described below.

Age-group composition. Each LSOA is characterised by the proportion of the population falling into six age groups (0–14, 15–44, 45–64, 65–74, 75–84, and ≥ 85).

Let $a_{g,l}$ denote the proportion of the population in age group g in LSOA l, such that $\sum_{g=1}^{6} a_{g,l} = 1$ for each LSOA l. To avoid perfect collinearity, we omit the 15–44 year age group (the modal group in most LSOAs) as the reference category and represent the age structure for LSOA l using the vector

$$a_l = \begin{pmatrix} a_{0\text{–}14,l} \\ a_{45\text{–}64,l} \\ a_{65\text{–}74,l} \\ a_{75\text{–}84,l} \\ a_{\geq 85,l} \end{pmatrix}. \tag{A.6}$$

Urban and rural classification. Urban and rural status is described using a six-level categorical variable based on the 2021 Rural–Urban Classification (see Table A.1).

Let $S_l \in \{\text{RSF1}, \text{RSN1}, \text{RLF1}, \text{RLN1}, \text{UF1}, \text{UN1}\}$ denote the category code for LSOA l. Taking RSF1 as the reference category, we represent urban–rural status for LSOA l using a vector of five binary indicator variables

$$\delta_l = \begin{pmatrix} \mathbb{I}\{S_l = \text{RSN1}\} \\ \mathbb{I}\{S_l = \text{RLF1}\} \\ \mathbb{I}\{S_l = \text{RLN1}\} \\ \mathbb{I}\{S_l = \text{UF1}\} \\ \mathbb{I}\{S_l = \text{UN1}\} \end{pmatrix}, \tag{A.7}$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function. This encoding treats the categories as nominal rather than ordered.

Table A.1: *Six-category urban and rural classification used in the analysis.*

Code	Description
RSF1	Smaller rural and further from a major town or city
RSN1	Smaller rural and nearer to a major town or city
RLF1	Larger rural and further from a major town or city
RLN1	Larger rural and nearer to a major town or city
UF1	Urban and further from a major town or city
UN1	Urban and nearer to a major town or city

Scalar covariates. Additional socioeconomic covariates are drawn from the Indices of Multiple Deprivation (IMD) dataset [58]. For each LSOA l, we include key domain-level deprivation scores as continuous variables, collected in the vector ζ_l

$$\zeta_l = \begin{pmatrix} \text{Income Score}_l \\ \text{Education, Skills and Training Deprivation Score}_l \\ \text{Employment Deprivation Score}_l \\ \text{Wider Barriers Sub-domain Score}_l \\ \text{Outdoors Living Environment Score}_l \end{pmatrix}. \tag{A.8}$$

We use the raw scores rather than ranks to preserve the magnitude of differences in deprivation between areas.

Design matrix of control variables. For implementation in the partial correlation regressions, the LSOA-level covariates defined above are assembled into the design matrix K. The l-th row of K corresponds to the covariate profile of LSOA l, defined as the vector k_l

$$k_l = \begin{pmatrix} 1 \\ a_l \\ \delta_l \\ \zeta_l \end{pmatrix}, \tag{A.9}$$

where the leading 1 is needed to include an intercept in the regression, ensuring that the resulting residuals have mean zero (or weighted mean zero in the weighted case),

which is required for the subsequent correlation calculations. The full design matrix K is constructed by stacking the transposed vectors $k_l^\top$ for all LSOAs

$$K = \begin{pmatrix} k_1^\top \\ \vdots \\ k_L^\top \end{pmatrix}. \tag{A.10}$$

A.2.2 Spearman rank correlation

Let $E_{x,l}$ and $P_{c,l}$ denote the ecological exposure to feature x and the allocated prevalence of condition c in LSOA l, respectively. Let E_x and P_c be vectors of length L representing ecological exposure to feature x and allocated prevalence of condition c for LSOAs $l = 1,\ldots,L$.

The ecological exposure and allocated prevalence are ranked

$$\begin{cases} R_x^{(\mathrm{E})} = \mathrm{rank}(E_x), \\ R_c^{(\mathrm{P})} = \mathrm{rank}(P_c), \end{cases} \tag{A.11}$$

where $\mathrm{rank}(\cdot)$ returns the rank of each entry in its argument, with average ranks assigned to equal values. The vectors $R_x^{(\mathrm{E})}$ and $R_c^{(\mathrm{P})}$ denote the ranks of ecological exposure to feature x and allocated prevalence of condition c, respectively, across LSOAs $l = 1,\ldots,L$.

Spearman's rank correlation coefficient $\rho_{\mathrm{S},x,c}$ is defined as the Pearson correlation between the ranked variables. All correlations are computed across LSOAs.

$$\rho_{\mathrm{S},x,c} = \frac{\sum_{l=1}^{L} \left(R_{x,l}^{(\mathrm{E})} - \bar{R}_x^{(\mathrm{E})}\right)\left(R_{c,l}^{(\mathrm{P})} - \bar{R}_c^{(\mathrm{P})}\right)}{\sqrt{\sum_{l=1}^{L} \left(R_{x,l}^{(\mathrm{E})} - \bar{R}_x^{(\mathrm{E})}\right)^2 \sum_{l=1}^{L} \left(R_{c,l}^{(\mathrm{P})} - \bar{R}_c^{(\mathrm{P})}\right)^2}}, \tag{A.12}$$

where $\bar{R}_x^{(\mathrm{E})}$ and $\bar{R}_c^{(\mathrm{P})}$ denote the sample means of the rank vectors. This definition naturally accommodates tied ranks.

A.2.3 Partial Spearman rank correlation

Partial Spearman rank correlation extends the rank-based analysis to adjust for confounding ecological factors. Let K denote a design matrix of covariates, as detailed in Section A.2.1. We isolate the association between exposure and prevalence by removing the linear effects of the covariates in K.

Specifically, we project the rank vectors onto the column space of K and compute the corresponding residuals

$$\begin{cases} \varepsilon_x^{(\mathrm{E})} = R_x^{(\mathrm{E})} - K(K^\top K)^{-1} K^\top R_x^{(\mathrm{E})}, \\ \varepsilon_c^{(\mathrm{P})} = R_c^{(\mathrm{P})} - K(K^\top K)^{-1} K^\top R_c^{(\mathrm{P})}, \end{cases} \tag{A.13}$$

where $R_x^{(\mathrm{E})}$ and $R_c^{(\mathrm{P})}$ are the column vectors of ranks defined in Equation (A.11). The residual vectors $\varepsilon_x^{(\mathrm{E})}$ and $\varepsilon_c^{(\mathrm{P})}$ represent the components of the exposure and prevalence ranks not explained by the ecological covariates.

The partial Spearman rank correlation is then defined as the Pearson correlation between these residuals

$$\rho_{\mathrm{S},x,c|K} = \frac{\sum_{l=1}^{L} \left(\varepsilon_{x,l}^{(\mathrm{E})} - \bar{\varepsilon}_x^{(\mathrm{E})}\right)\left(\varepsilon_{c,l}^{(\mathrm{P})} - \bar{\varepsilon}_c^{(\mathrm{P})}\right)}{\sqrt{\sum_{l=1}^{L} \left(\varepsilon_{x,l}^{(\mathrm{E})} - \bar{\varepsilon}_x^{(\mathrm{E})}\right)^2 \sum_{l=1}^{L} \left(\varepsilon_{c,l}^{(\mathrm{P})} - \bar{\varepsilon}_c^{(\mathrm{P})}\right)^2}}, \tag{A.14}$$

where $\bar{\varepsilon}_x^{(\mathrm{E})}$ and $\bar{\varepsilon}_c^{(\mathrm{P})}$ denote the sample means of the residual vectors.

A.2.4 Population weighting for partial correlation and visualisation

To account for variation in population size across LSOAs, we extend the partial Spearman rank correlation to incorporate population weights. The weight w_l for LSOA l is defined as the normalised population

$$w_l = \frac{d_l}{\sum_{l'=1}^{L} d_{l'}} \tag{A.15}$$

where d_l is defined in Equation (A.4). By construction, $\sum_{l=1}^{L} w_l = 1$.

The diagonal weight matrix W collects the weights for all L LSOAs

$$W = \mathrm{diag}(w_1, w_2, \ldots, w_L). \tag{A.16}$$

Standard ranks are replaced by weighted Ridit scores [16]. The weighted ranks for ecological exposure and prevalence are defined as cumulative weighted population proportions

$$\begin{cases} R_{x,l}^{(\mathrm{E,w})} = \sum\limits_{l':E_{x,l'}<E_{x,l}} w_{l'} + \frac{1}{2} \sum\limits_{l':E_{x,l'}=E_{x,l}} w_{l'}, \\ R_{c,l}^{(\mathrm{P,w})} = \sum\limits_{l':P_{c,l'}<P_{c,l}} w_{l'} + \frac{1}{2} \sum\limits_{l':P_{c,l'}=P_{c,l}} w_{l'}, \end{cases} \tag{A.17}$$

where the index l' iterates over all LSOAs.

Weighted least squares residuals are then obtained by regressing the weighted rank vectors on the covariate matrix K

$$\begin{cases} \varepsilon_x^{(\mathrm{E,w})} = R_x^{(\mathrm{E,w})} - K(K^\top W K)^{-1} K^\top W R_x^{(\mathrm{E,w})}, \\ \varepsilon_c^{(\mathrm{P,w})} = R_c^{(\mathrm{P,w})} - K(K^\top W K)^{-1} K^\top W R_c^{(\mathrm{P,w})}. \end{cases} \tag{A.18}$$

where $R_x^{(\mathrm{E,w})}$ and $R_c^{(\mathrm{P,w})}$ denote the column vectors of weighted ranks.

The population-weighted partial Spearman rank correlation is then defined as the weighted Pearson correlation between these residuals

$$\rho^{(\mathrm{w})}_{\mathrm{S},x,c|K} = \frac{\sum_{l=1}^{L} w_l \left(\varepsilon^{(\mathrm{E,w})}_{x,l} - \bar{\varepsilon}^{(\mathrm{E,w})}_{x}\right)\left(\varepsilon^{(\mathrm{P,w})}_{c,l} - \bar{\varepsilon}^{(\mathrm{P,w})}_{c}\right)}{\sqrt{\sum_{l=1}^{L} w_l \left(\varepsilon^{(\mathrm{E,w})}_{x,l} - \bar{\varepsilon}^{(\mathrm{E,w})}_{x}\right)^2 \sum_{l=1}^{L} w_l \left(\varepsilon^{(\mathrm{P,w})}_{c,l} - \bar{\varepsilon}^{(\mathrm{P,w})}_{c}\right)^2}}, \tag{A.19}$$

where the weighted means are given by

$$\begin{cases} \bar{\varepsilon}^{(\mathrm{E,w})}_{x} = \sum_{l=1}^{L} w_l \, \varepsilon^{(\mathrm{E,w})}_{x,l}, \\ \bar{\varepsilon}^{(\mathrm{P,w})}_{c} = \sum_{l=1}^{L} w_l \, \varepsilon^{(\mathrm{P,w})}_{c,l}. \end{cases} \tag{A.20}$$

Equation (A.19) was used to calculate the values of the Spearman rank correlation coefficients (labelled ρ_{S}) that appear in the insets on Figures 5, 9, and 11.

A.2.5 Visualisation of adjusted trends

Partial Spearman rank correlation coefficients summarise the strength and direction of monotonic associations between ecological exposures and disease prevalence and are robust to skewness and outliers. Because these coefficients are computed on ranks, they do not yield changes in prevalence per unit change in exposure. For visualisation, we therefore also report covariate-adjusted association trend lines in Figures 5, 9, and 11.

Trend lines are estimated using population-weighted linear regressions of allocated prevalence on the ecological exposure and covariates. Predicted prevalence is evaluated as a function of exposure, with all covariates fixed at their *unweighted* sample means across LSOAs

$$\hat{P}_{x,c}(e_x) \;=\; \gamma_{x,c}\, e_x \;+\; \sum_{h} \gamma_{h,x,c}\, \bar{k}_h, \tag{A.21}$$

where the index h runs over the intercept and all ecological covariates, and

$$\bar{k}_h \;=\; \frac{1}{L} \sum_{l=1}^{L} k_{h,l} \tag{A.22}$$

is the *unweighted* sample mean of the h-th covariate in the vector k_l defined in Equation (A.9). For clarity, the regression fit is population-weighted, but the reference covariate profile used for visualisation is the average across LSOAs (each LSOA contributing equally).

In Equation (A.21), $\hat{P}_{x,c}(e_x)$ denotes the predicted prevalence of condition c at exposure level e_x for feature x (*e.g.* drive time in minutes or NDVI). The coefficient $\gamma_{x,c}$ represents the linear association between exposure and prevalence. The summation term is invariant in e_x and fixes the remaining covariates at $\bar{k}_h$, using exposure–condition–specific coefficients $\gamma_{h,x,c}$. Consequently, the trend lines describe the adjusted exposure–prevalence relationship across the observed range of e_x and provide an interpretable visual complement to the rank-based partial Spearman correlation coefficients used for inference.

A.3 Exposure–prevalence association modelling framework

This section describes the exposure–prevalence modelling framework used to estimate how the magnitude of ecological exposures is associated with the variation in LSOA-level disease prevalence. The framework allows for flexible non-linear relationships, enabling the association to vary non-linearly, plateau or change across the exposure range [27].

A.3.1 Exposure transformation

The ecological exposure metrics differ in scale and units. NDVI and the fuel-poverty rate are proportions and hence take values in $[0,1]$. By contrast, drive-time accessibility (to food retail and leisure facilities) and the crime rate are non-negative and unbounded above, *i.e.* they take values in $[0,\infty)$.

Each exposure $E_{x,l}$ is transformed via

$$z_{x,l} = \begin{cases} \text{logit}(E_{x,l}), & \text{NDVI, fuel poverty,} \\ \log\big(1+E_{x,l}\big), & \text{accessibility, crime,} \end{cases} \tag{A.23}$$

where $\text{logit}(u) = \log\big(u/(1-u)\big)$. These transformations improve numerical robustness and enable the restricted cubic spline to represent flexible non-linear associations across the observed exposure range.

More generally, let $z_x(e_x)$ denote the transformed exposure obtained by applying the stabilisation and transformation described above to a generic exposure level e_x.

A.3.2 Poisson regression model

For each condition c and LSOA l, we model the allocated case count $N_{c,l}$ from Equation (A.3) using a log-linear mean model with an offset for the LSOA population d_l. Because $N_{c,l}$ is an allocated (and generally non-integer) case count, we use a working Poisson likelihood as a convenient estimation device for the conditional mean; inference is based on robust (sandwich/Huber–White) standard errors. Formally, we write the working model as

$$N_{c,l} \sim \text{Poisson}(\mu_{x,c,l}), \tag{A.24}$$

with linear predictor

$$\log(\mu_{x,c,l}) = \log(d_l) + \eta_{x,c,l}(e_{x,l}), \tag{A.25}$$

where $\log(d_l)$ is an offset term and

$$\eta_{x,c,l}(e_{x,l}) \;=\; f_{x,c}(z_x(e_{x,l})) \;+\; \sum_h \beta_{h,x,c}\, k_{h,l}. \tag{A.26}$$

Here, $f_{x,c}(z_x(e_{x,l}))$ is a non-linear exposure contribution, $z_x(e_{x,l})$ is the corresponding transformed exposure, $k_{h,l}$ denotes the h-th adjustment covariate in LSOA l (including the intercept), and $\beta_{h,x,c}$ are the corresponding regression coefficients for exposure x and condition c.

The non-linear exposure term $f_{x,c}(\cdot)$ is represented using a restricted cubic spline (RCS) with $V = 4$ basis functions

$$f_{x,c}(z_x) \;=\; \sum_{v=1}^{V} \beta_{v,x,c}^{(\text{spline})}\, B_{v,x}(z_x), \tag{A.27}$$

where $B_{v,x}(\cdot)$ is the v-th spline basis function for exposure type x and $\beta_{v,x,c}^{(\text{spline})}$ is its associated coefficient. The spline basis is linear below the first knot and above the last knot.

For each exposure variable x and condition c, a design matrix is formed by combining the ecological covariates $k_{h,l}$ with the RCS basis functions $B_{v,x}(z_x)$ evaluated at the transformed exposure values. LSOAs with missing exposure, missing covariates, non-positive population denominators, or missing outcome data are excluded prior to fitting. The covariate coefficients $\beta_{h,x,c}$ and the spline coefficients $\beta_{v,x,c}^{(\text{spline})}$ for $v = 1,\ldots,V$ are estimated jointly by maximum likelihood under the working Poisson likelihood with offset $\log(d_l)$, and robust standard errors are used for uncertainty quantification.

All analyses were implemented in Python. The generalised linear models were fit using `statsmodels` [82, 88], spline design matrices were constructed using `patsy` [85], and data handling and numerical computations used `pandas` [57, 95] and `NumPy` [34, 64]. Statistical functions were obtained from `SciPy` [96, 101]. Figures were produced using `Matplotlib` [40, 94].

A.3.3 Model-based prevalence and number of cases

We use the Poisson regression model in Equation (A.25) to evaluate the model-based prevalence rate as a function of the level of ecological exposure $e_{x,l}$ to feature x in LSOA l

$$\hat{\lambda}_{x,c,l}(e_{x,l}) \;=\; \exp(\hat{\eta}_{x,c,l}(e_{x,l}))\,. \tag{A.28}$$

The expected case count in LSOA l is

$$\hat{\mu}_{x,c,l}(e_{x,l}) \;=\; d_l\,\hat{\lambda}_{x,c,l}(e_{x,l}), \tag{A.29}$$

and the corresponding total number of cases

$$T_{x,c} \;=\; \sum_{l=1}^{L} \hat{\mu}_{x,c,l}(e_{x,l}). \tag{A.30}$$

A.4 Effect measures from the exposure–prevalence association model

This section defines the scalar summaries reported in the main text and figures, all computed from the fitted Poisson exposure–prevalence association models.

A.4.1 Adjusted prevalence ratios

For any two exposure scenarios A and B (defined as two LSOA-level exposure vectors $\{e_{x,l}^{(A)}\}$ and $\{e_{x,l}^{(B)}\}$), we define the adjusted prevalence ratio as

$$\mathrm{aPR}_{x,c}(A,B) \;=\; \frac{\sum_{l=1}^{L} \hat{\mu}_{x,c,l}\big(e_{x,l}^{(A)}\big)}{\sum_{l=1}^{L} \hat{\mu}_{x,c,l}\big(e_{x,l}^{(B)}\big)}. \tag{A.31}$$

Confidence intervals for $\mathrm{aPR}_{x,c}$ were calculated using robust (Huber–White) standard errors [102, 108] and a Wald approximation. The 95% confidence interval is calculated as

$$\Big[\exp\big(\Delta\hat{\eta}_{\mathrm{aPR},x,c} - 1.96\cdot \mathrm{SE}_{\mathrm{robust},x,c}\big),\ \exp\big(\Delta\hat{\eta}_{\mathrm{aPR},x,c} + 1.96\cdot \mathrm{SE}_{\mathrm{robust},x,c}\big)\Big], \tag{A.32}$$

where $\Delta\hat{\eta}_{\mathrm{aPR},x,c}$ is the difference in log total predicted cases between scenarios A and B

$$\Delta\hat{\eta}_{\mathrm{aPR},x,c} = \log\left(\sum_{l=1}^{L} \hat{\mu}_{x,c,l}\big(e_{x,l}^{(A)}\big)\right) - \log\left(\sum_{l=1}^{L} \hat{\mu}_{x,c,l}\big(e_{x,l}^{(B)}\big)\right), \tag{A.33}$$

and $\mathrm{SE}_{\mathrm{robust},x,c}$ is the robust standard error of $\Delta\hat{\eta}_{\mathrm{aPR},x,c}$, computed using the delta method by treating $\Delta\hat{\eta}_{\mathrm{aPR},x,c}$ as a differentiable function of the fitted Poisson regression coefficients (including spline and adjustment terms). We combined the resulting gradient with the Huber–White (sandwich) robust covariance matrix of the fitted coefficients to obtain $\widehat{\mathrm{Var}}(\Delta\hat{\eta}_{\mathrm{aPR},x,c})$, and set

$$\mathrm{SE}_{\mathrm{robust},x,c} \;=\; \sqrt{\widehat{\mathrm{Var}}(\Delta\hat{\eta}_{\mathrm{aPR},x,c})}. \tag{A.34}$$

As a fallback when the Huber–White calculation was numerically unstable, we used the default model-based (non-robust) covariance matrix returned by maximum likelihood.

Model fitting and inference were performed in Python using `statsmodels` [82].

A.4.2 Interquartile range

For empirical analyses, we use the interquartile range (IQR) contrast

$$\mathrm{aPR}_{\mathrm{IQR},x,c} \;=\; \mathrm{aPR}_{x,c}(A,B), \tag{A.35}$$

where scenarios A and B are defined by the constant (LSOA-invariant) exposure vectors

$$\begin{aligned} e_{x,l}^{(A)} &= Q_x(75), \\ e_{x,l}^{(B)} &= Q_x(25). \end{aligned} \tag{A.36}$$

A.4.3 Relative risk with specified baseline

Adjusted prevalence ratios can be interpreted as model-based relative risks (RR) when computed relative to a specified baseline exposure scenario. For any exposure scenario A and baseline scenario B, define

$$\mathrm{RR}_{x,c}(A) \;=\; \mathrm{aPR}_{x,c}(A,B), \tag{A.37}$$

where the baseline exposure scenario B is the constant (LSOA-invariant) exposure vector

$$e_{x,l}^{(B)} \;:=\; \begin{cases} Q_x(75), & \text{for protective exposures,} \\ Q_x(25), & \text{for harmful exposures.} \end{cases} \qquad l = 1,\ldots,L, \tag{A.38}$$

with $Q_x(p)$ denoting the pth percentile of the distribution of exposure x across LSOAs.

A.4.4 Population attributable fraction under percentile capping

The Population Attributable Fraction (PAF) [54] provides a model-based summary of how the fitted exposure–prevalence association translates into differences in the expected number of cases under a specified alternative exposure distribution, with all other covariates held fixed. We report PAF values as *model-implied* quantities: they describe changes in expected cases under the fitted Poisson model and do not, by themselves, establish that changing exposure would causally change prevalence.

The total expected number of cases under the observed exposure distribution is

$$T_{\mathrm{obs},x,c} \;=\; \sum_{l=1}^{L} \hat{\mu}_{x,c,l}(E_{x,l}), \tag{A.39}$$

where $E_{x,l}$ is the observed raw exposure for LSOA l.

We define an alternative scenario in which raw exposures are capped at a percentile threshold θ

$$E_{x,l}^{(\theta)} = \begin{cases} \min\{E_{x,l},\, Q_x(\theta)\}, & \text{for harmful exposures,} \\ \max\{E_{x,l},\, Q_x(100-\theta)\}, & \text{for protective exposures,} \end{cases} \tag{A.40}$$

where $Q_x(\cdot)$ denotes the empirical percentile function of the raw exposure distribution.

Expected cases under this alternative scenario are obtained by transforming $E_{x,l}^{(\theta)}$ through the same mapping $z_x(\cdot)$ used in the model fitting, rebuilding the spline basis terms $B_{v,x}$ for the alternative exposure values, and evaluating the fitted model while keeping the non-exposure covariates fixed at their LSOA-specific values $k_{h,l}$

$$\log\big(\hat{\mu}_{x,c,l}^{(\theta)}\big) = \log(d_l) + \hat{\eta}_{x,c,l}(E_{x,l}^{(\theta)}), \tag{A.41}$$

where $\hat{\eta}_{x,c,l}(\cdot)$ is defined as per (A.26).

The total expected number of cases under the alternative is

$$T_{\theta,x,c} \;=\; \sum_{l=1}^{L} \hat{\mu}_{x,c,l}^{(\theta)}. \tag{A.42}$$

The PAF is then

$$\mathrm{PAF}_{\theta,x,c} \;=\; 1 - \frac{T_{\theta,x,c}}{T_{\mathrm{obs},x,c}}. \tag{A.43}$$

Under the fitted model, this quantity corresponds to the proportionate reduction in expected cases when exposures are set to (or improved to) the chosen percentile threshold, holding non-exposure covariates fixed.

A.5 Strict McFadden pseudo R-squared

To assess goodness-of-fit, we use the strict McFadden pseudo-R^2 [3] for each fitted model.

Let $\ell_{\mathrm{full},x,c}$ denote the maximised log-likelihood of the full Poisson model for exposure x and condition c in Equation (A.25)

$$\ell_{\mathrm{full},x,c} \;=\; \sum_{l=1}^{L} \Big(N_{c,l} \log(\hat{\mu}_{x,c,l}) - \hat{\mu}_{x,c,l} - \log\Gamma(N_{c,l}+1) \Big), \tag{A.44}$$

where $\log\Gamma(N_{c,l}+1)$ is constant with respect to the model parameters.

Next, let $\ell_{\mathrm{null},x,c}$ denote the maximised log-likelihood of the corresponding null model, fitted on the same LSOA sample and using the same offset $\log(d_l)$. The null model

$$\log\left(\hat{\mu}^{(\mathrm{null})}_{x,c,l}\right) \;=\; \log(d_l) + \beta^{(\mathrm{null})}_{0,x,c}, \tag{A.45}$$

is fitted to the number of observed cases $N_{c,l}$ of condition c each LSOA l, where $\beta^{(\mathrm{null})}_{0,x,c}$ is the intercept parameter of the null model, representing a single common log-rate for exposure x and condition c across all LSOAs, so that $\log\left(\hat{\mu}^{(\mathrm{null})}_{x,c,l}/d_l\right) = \beta^{(\mathrm{null})}_{0,x,c}$ for all l.

The strict McFadden pseudo-R^2 is then defined as

$$R^2_{\mathrm{McF},x,c} \;=\; 1 - \frac{\ell_{\mathrm{full},x,c}}{\ell_{\mathrm{null},x,c}}. \tag{A.46}$$

A.6 Traceability of figure quantities

Table A.2: *Summary of equations used to generate quantities shown in figures.*

Figure	Equation	Quantity computed
3	(A.5)	LSOA-level prevalence proportion $P_{c,l} = N_{c,l}/d_l$ used to map disease prevalence.
5	(A.19)	Population-weighted partial Spearman rank correlation $\rho^{(\mathrm{w})}_{\mathrm{S},x,c\mid K}$ (values shown in figure insets).
5	(A.21)	Adjusted trend line / predicted prevalence vs. exposure $\hat{P}_{x,c}(e_x)$, obtained by evaluating covariates at reference values (visualised adjusted association).
6	(A.35)	Adjusted prevalence ratio per IQR contrast $\mathrm{aPR}_{\mathrm{IQR},x,c}$ from the fitted Poisson exposure–prevalence association model (effect size panels).
6	(A.46)	Strict McFadden pseudo-R^2 $R^2_{\mathrm{McF},x,c}$ (model performance panel).
7	(A.37)	Model-based relative risk $\mathrm{RR}_{x,c}(A)$ used for the vertical axis of exposure–prevalence association curves (here A denotes the scenario in which exposure is set to a given level).
9	(A.19)	Population-weighted partial Spearman rank correlation $\rho^{(\mathrm{w})}_{\mathrm{S},x,c\mid K}$ (fuel poverty vs. cardiometabolic outcomes association summaries).
9	(A.21)	Adjusted trend line / predicted prevalence vs. exposure $\hat{P}_{x,c}(e_x)$ (visualised adjusted association).
10	(A.35)	Adjusted prevalence ratio per IQR contrast $\mathrm{aPR}_{\mathrm{IQR},x,c}$ (panel a).
10	(A.46)	Strict McFadden pseudo-R^2 $R^2_{\mathrm{McF},x,c}$ (panel b).
10	(A.43)	Model-implied population attributable fraction $\mathrm{PAF}_{\theta,x,c} = 1 - T_{\theta,x,c}/T_{\mathrm{obs},x,c}$ under percentile-capping scenarios (panel c).
11	(A.19)	Population-weighted partial Spearman rank correlation $\rho^{(\mathrm{w})}_{\mathrm{S},x,c\mid K}$ (SMI vs. exposures association summaries).
11	(A.37)	Model-based relative risk $\mathrm{RR}_{x,c}(A)$ for SMI exposure–prevalence association curves (lower panels; here A denotes the scenario setting exposure to a given level e_x, expressed relative to the baseline defined in Equation (A.38) per caption).

A.7 Nomenclature

Table A.3: *Summary of parameters and symbols.*

Symbol	Description
Indices and sets	
C	Total number of medical conditions.
L	Total number of LSOAs.
P	Total number of GP practices.
V	Number of restricted cubic spline basis functions.
c	Medical condition index ($c = 1,\ldots,C$).
g	Age-group index ($g \in \{0$–$14,\ 15$–$44,\ 45$–$64,\ 65$–$74,\ 75$–$84,\ \geq 85\}$).
h	Covariate index (including the intercept) in k_l and coefficient vectors.
j	GP practice index ($j = 1,\ldots,P$).
l	LSOA index ($l = 1,\ldots,L$); l' denotes a dummy LSOA index in summations.
v	Spline-basis index ($v = 1,\ldots,V$) (restricted cubic spline basis functions).
x	Exposure-feature index.
Observed and constructed population / prevalence quantities	
$N_{c,l}$	Estimated number of cases of condition c in LSOA l, Eq. (A.3).
$O_{j,c}$	Observed QOF case count for condition c at GP practice j.
$P_{c,l}$	Allocated prevalence proportion for condition c in LSOA l, $P_{c,l} = N_{c,l}/d_l$, Eq. (A.5).
d_l	Modelled registered population in LSOA l, $d_l = \sum_{j=1}^{P} m_{j,l}$, Eq. (A.4).
$m_{j,l}$	Reconciled count of patients from practice j residing in LSOA l, Eq. (A.1).
n_j	Total number of patients registered at GP practice j (QOF denominator).
$n_{j,l}$	Raw count of patients registered at practice j who reside in LSOA l (residence dataset).
$\varphi_{j,l}$	Share of practice j's register residing in LSOA l, $\varphi_{j,l} = m_{j,l}/n_j$, Eq. (A.2).

Continued on next page

Table A.3 – continued from previous page

Symbol	Description
Ecological exposures and transformations	
E_x	Column vector collecting $E_{x,l}$ over $l = 1, \ldots, L$.
$E_{x,l}$	Raw ecological exposure value for feature x in LSOA l.
e_x	Generic (scalar) exposure level at which exposure–prevalence functions are evaluated.
$z_x(e_x)$	Transformation mapping applied to a raw exposure level e_x.
$z_{x,l}$	Transformed exposure for feature x in LSOA l, Eq. (A.23).
Covariates and design matrices	
K	$L \times p$ design matrix of ecological covariates; row l equals $k_l^\top$, Eq. (A.10).
S_l	Six-level rural–urban classification code for LSOA l.
$a_{g,l}$	Proportion of the population in age group g within LSOA l; $\sum_g a_{g,l} = 1$.
a_l	Age-composition vector for LSOA l (reference group 15–44 omitted), Eq. (A.6).
$k_{h,l}$	Value of covariate h in LSOA l.
k_l	Column covariate vector for LSOA l (including intercept), Eq. (A.9).
$\bar{k}_h$	Sample mean of covariate h across LSOAs, $\bar{k}_h = \frac{1}{L}\sum_{l=1}^{L} k_{h,l}$, Eq. (A.22).
δ_l	Urban/rural indicator vector for LSOA l (non-reference categories), Eq. (A.7).
ζ_l	Vector of scalar IMD-domain covariates for LSOA l, Eq. (A.8).
Spearman rank correlation coefficients (unadjusted, partial, and weighted)	
$R_x^{(\mathrm{E})}$	Vector of ranks of $E_{x,l}$ across LSOAs, Eq. (A.11).
$R_{x,l}^{(\mathrm{E,w})}$	Weighted Ridit score for exposure $E_{x,l}$, Eq. (A.17).
$R_c^{(\mathrm{P})}$	Vector of ranks of $P_{c,l}$ across LSOAs, Eq. (A.11).
$R_{c,l}^{(\mathrm{P,w})}$	Weighted Ridit score for prevalence $P_{c,l}$, Eq. (A.17).
W	Diagonal weight matrix, $W = \mathrm{diag}(w_1, \ldots, w_L)$, Eq. (A.16).
w_l	Population weight for LSOA l, $w_l = d_l / \sum_{l'=1}^{L} d_{l'}$, Eq. (A.15).
$\varepsilon_x^{(\mathrm{E})}$	Residual vector of exposure ranks after projecting $R_x^{(\mathrm{E})}$ onto K, Eq. (A.13).
$\varepsilon_x^{(\mathrm{E,w})}$	Weighted residuals of exposure Ridit scores after WLS on K, Eq. (A.18).
$\varepsilon_c^{(\mathrm{P})}$	Residual vector of prevalence ranks after projecting $R_c^{(\mathrm{P})}$ onto K, Eq. (A.13).
$\varepsilon_c^{(\mathrm{P,w})}$	Weighted residuals of prevalence Ridit scores after WLS on K, Eq. (A.18).
$\rho_{\mathrm{S},x,c}$	Spearman rank correlation between $E_{x,l}$ and $P_{c,l}$ across LSOAs, Eq. (A.12).
$\rho_{\mathrm{S},x,c\|K}$	Partial Spearman rank correlation adjusted for covariates K, Eq. (A.14).
$\rho_{\mathrm{S},x,c\|K}^{(\mathrm{w})}$	Population-weighted partial Spearman rank correlation, Eq. (A.19).
Adjusted trend visualisation (linear model)	
$\hat{P}_{x,c}(e_x)$	Predicted prevalence vs. exposure used for adjusted trend lines, Eq. (A.21).
$\gamma_{x,c}$	Exposure slope coefficient in the (population-weighted) linear regression used for visualisation, Eq. (A.21).
Poisson exposure–prevalence association model	
$B_{v,x}(z_x)$	v-th restricted cubic spline basis function for exposure x evaluated at z, Eq. (A.27).
$f_{x,c}(z)$	Non-linear exposure contribution on the log scale (restricted cubic spline in z), Eq. (A.27).
$\beta_{h,x,c}$	Poisson regression coefficient for covariate h (including intercept) for exposure x and condition c, Eq. (A.26).
$\beta_{v,x,c}^{(\mathrm{spline})}$	Coefficient of spline basis function $B_{v,x}(\cdot)$, Eq. (A.27).
$\eta_{x,c,l}(e_{x,l})$	Poisson linear predictor (log-rate, excluding offset) at exposure $e_{x,l}$, Eq. (A.26).
$\hat{\lambda}_{x,c,l}(e_{x,l})$	Model-based prevalence rate in LSOA l at exposure $e_{x,l}$, Eq. (A.28).
$\mu_{x,c,l}$	Poisson mean (expected case count) for exposure feature x and condition c in LSOA l, Eq. (A.24).
$\hat{\mu}_{x,c,l}$	Fitted Poisson mean under exposure feature x and condition c, Eq. (A.25).
Effect measures derived from the fitted exposure–prevalence association model	
$E_{x,l}^{(\theta)}$	Percentile-capped (counterfactual) exposure for feature x in LSOA l, Eq. (A.40).
$\mathrm{PAF}_{\theta,x,c}$	Model-implied population attributable fraction under percentile capping, Eq. (A.43).
$Q_x(p)$	Empirical p-th percentile of the raw exposure distribution for feature x.

Continued on next page

Table A.3 – continued from previous page

Symbol	Description
$\mathrm{RR}_{x,c}(A)$	Model-based relative risk for scenario A relative to the baseline scenario, Eq. (A.37).
$\mathrm{SE}_{\mathrm{robust},x,c}$	Robust (Huber–White) standard error used for Wald confidence intervals, Eq. (A.32).
$T_{\mathrm{obs},x,c}$	Total expected cases under observed exposures, Eq. (A.39).
$T_{\theta,x,c}$	Total expected cases under the percentile-capped scenario.
$\mathrm{aPR}_{\mathrm{IQR},x,c}$	Adjusted prevalence ratio for an IQR contrast, Eq. (A.35).
$\mathrm{aPR}_{x,c}(A,B)$	Adjusted prevalence ratio comparing exposure scenarios A vs. B, Eq. (A.31).
$\Delta\hat{\eta}_{\mathrm{aPR},x,c}$	Difference in log total predicted cases used to calculate adjusted prevalence ratio, Eq. (A.33).
$\hat{\mu}^{(\theta)}_{x,c,l}$	Fitted counterfactual Poisson mean under percentile capping, Eq. (A.41).
θ	Percentile threshold used in exposure-capping scenarios.
Model fit summary	
$R^2_{\mathrm{McF},x,c}$	Strict McFadden pseudo-R^2, Eq. (A.46).
$\ell_{\mathrm{full},x,c}$	Maximised log-likelihood of the full Poisson model, Eq. (A.44).
$\ell_{\mathrm{null},x,c}$	Maximised log-likelihood of the intercept-only null model, Eq. (A.45).
$\beta^{(\mathrm{null})}_{0,x,c}$	Intercept parameter in the null model, Eq. (A.45).
Operators and typographic conventions	
$\mathbb{E}[\cdot]$	Expectation operator.
$\mathbb{I}\{\cdot\}$	Indicator function (1 if condition holds; 0 otherwise).
$\mathrm{rank}(\cdot)$	Sample rank operator; ties receive average ranks.
$(\mathrm{E}), (\mathrm{P})$	Superscripts indicating quantities constructed from exposure vs. prevalence ranks.
(null)	Superscript indicating the intercept-only null model.
(w)	Superscript indicating population weighting (via w_l or W).
(θ)	Superscript indicating a percentile-capped counterfactual scenario.
$\top$	Transpose.
$\bar{\cdot}$	Bar accent indicates a mean quantity.
$\hat{\cdot}$	Hat accent indicates an estimated or fitted quantity.
Acronyms	
GP	General Practice.
IMD	Indices of Multiple Deprivation.
LSOA	Lower Layer Super Output Area.
NDVI	Normalised Difference Vegetation Index.
QOF	Quality and Outcomes Framework.
RCS	Restricted cubic spline.

References

[1] T. Althoff, B. Ivanovic, A. C. King, J. L. Hicks, S. L. Delp, and J. Leskovec. Countrywide natural experiment links built environment to physical activity. *Nature*, 2025. doi:10.1038/s41586-025-09321-3.

[2] M. Amith, L. Cui, K. Roberts, H. Xu, and C. Tao. Ontology of consumer health vocabulary: providing a formal and interoperable semantic resource for linking lay language and medical terminology. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, page 1177–1178. IEEE, 2019. doi:10.1109/bibm47256.2019.8983220.

[3] M. Ancukiewicz, D. M. Finkelstein, and D. A. Schoenfeld. Modelling the relationship between continuous covariates and clinical events using isotonic regression. *Statistics in Medicine*, 22(20):3151–3159, 2003. doi:10.1002/sim.1561.

[4] J. Bai, S. Mosbach, C. J. Taylor, D. Karan, K. F. Lee, S. D. Rihm, J. Akroyd, A. A. Lapkin, and M. Kraft. A dynamic knowledge graph approach to distributed self-driving laboratories. *Nature Communications*, 15, 2024. doi:10.1038/s41467-023-44599-9.

[5] G. Baranyi and et al. The impact of neighbourhood crime on mental health: A systematic review and meta-analysis. *Social Science & Medicine*, 282:114106, 2021. doi:10.1016/j.socscimed.2021.114106.

[6] G. Baranyi, M. Cherrie, S. E. Curtis, C. Dibben, and J. Pearce. Neighborhood crime and psychotropic medications: A longitudinal data linkage study of 130,000 Scottish adults. *American Journal of Preventive Medicine*, 58(5):638–647, 2020. doi:10.1016/j.amepre.2019.12.022.

[7] G. Baranyi, M. Cherrie, S. E. Curtis, C. Dibben, and J. Pearce. Changing levels of local crime and mental health: a natural experiment using self-reported and service-use data in Scotland. *Journal of Epidemiology and Community Health*, 74 (10):806–814, 2020. doi:10.1136/jech-2020-213837.

[8] R. Battle and D. Kolas. Enabling the geospatial semantic web with parliament and geosparql. *Semantic Web*, 3(4):355–370, 2012. doi:10.3233/sw-2012-0065.

[9] J. Bell, L. Neubeck, K. Jin, P. Kelly, and C. L. Hanson. Understanding leisure centre-based physical activity after physical activity referral: Evidence from scheme participants and completers in Northumberland UK. *International Journal of Environmental Research and Public Health*, 18(6):2957, 2021. doi:10.3390/ijerph18062957.

[10] D. Bender and K. Sartipi. Hl7 fhir: An agile and restful approach to healthcare information exchange. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, page 326–331. IEEE, 2013. doi:10.1109/cbms.2013.6627810.

[11] C. Berragan, M. Green, and A. Singleton. Access to Healthy Assets & Hazards (AHAH). https://github.com/GeographicDataService/ahah, 2024.

[12] V. Bhavsar, J. Boydell, R. M. Murray, and et al. The association between neighbourhood characteristics and physical victimisation in men and women with mental disorders. *BJPsych Open*, 6(4):e73, 2020. doi:10.1192/bjo.2020.52.

[13] L. Bjerg, E.-M. Dalsgaard, K. Norman, A. A. Isaksen, and A. Sandbæk. Cohort profile: Health in Central Denmark (HICD) cohort - a register-based questionnaire survey on diabetes and related complications in the Central Denmark Region. *BMJ Open*, 12(7):e060410, 2022. doi:10.1136/bmjopen-2021-060410.

[14] S. Boylan, C. Arsenault, M. Barreto, F. A. Bozza, A. Fonseca, E. Forde, L. Hookham, G. S. Humphreys, M. Y. Ichihara, K. Le Doare, X. F. Liu, E. McNamara, J. C. Mugunga, J. F. Oliveira, J. Ouma, N. Postlethwaite, M. Retford, L. F. Reyes, A. D. Morris, and A. Wozencraft. Data challenges for international health emergencies: lessons learned from ten international COVID-19 driver projects. *The Lancet Digital Health*, 6(5):e354–e366, 2024. doi:10.1016/s2589-7500(24)00028-1.

[15] P. Broadbent, R. Thomson, D. Kopasker, G. McCartney, P. Meier, M. Richiardi, M. McKee, and S. V. Katikireddi. The public health implications of the cost-of-living crisis: outlining mechanisms and modelling consequences. *The Lancet Regional Health - Europe*, 27:100585, 2023. doi:10.1016/j.lanepe.2023.100585.

[16] I. D. J. Bross. How to use ridit analysis. *Biometrics*, 14(1):18–38, 1958. doi:10.2307/2527727.

[17] T. Burgoine, N. G. Forouhi, S. J. Griffin, N. J. Wareham, and P. Monsivais. Associations between exposure to takeaway food outlets, takeaway food consumption, and body weight in Cambridgeshire, UK: population based, cross sectional study. *BMJ*, 348, 2014. doi:10.1136/bmj.g1464. URL https://www.bmj.com/content/348/bmj.g1464.

[18] T. Burgoine, C. Sarkar, C. J. Webster, and P. Monsivais. Examining the interaction of fast-food outlet exposure and income on diet and obesity: evidence from 51,361 UK Biobank participants. *International Journal of Behavioral Nutrition and Physical Activity*, 15(1):71, July 2018. doi:10.1186/s12966-018-0699-8.

[19] J. Cartagena-Farias, N. Brimblecombe, and M. Knapp. Evaluating the association between receipt of a winter fuel cash transfer and older people's care needs, quality of life, and housing quality: Evidence from England. *Social Science & Medicine*, 355:117128, 2024. doi:10.1016/j.socscimed.2024.117128.

[20] J. Chen, J. Bai, J. Xu, F. Farazi, S. Mosbach, J. Akroyd, and M. Kraft. Transforming building retrofits: Linking energy, equity, and health insights from The World Avatar. *Advances in Applied Energy*, 19:100230, 2025. doi:10.1016/j.adapen.2025.100230.

[21] M. Codescu, G. Horsinka, O. Kutz, T. Mossakowski, and R. Rau. Osmonto: An ontology of openstreetmap tags. In *Proceedings of the State of the Map Europe (SotM-EU) 2011*, 2011. Available at: https://www.inf.unibz.it/~okutz/resources/osmonto.pdf.

[22] M. H. Coletti and H. L. Bleich. Medical subject headings used to search the biomedical literature. *Journal of the American Medical Informatics Association*, 8 (4):317–323, 2001. doi:10.1136/jamia.2001.0080317.

[23] A. M. Dalton, A. P. Jones, S. J. Sharp, A. J. M. Cooper, S. Griffin, and N. J. Wareham. Residential neighbourhood greenspace is associated with reduced risk of incident diabetes in older people: a prospective cohort study. *BMC Public Health*, 16(1171), 2016. doi:10.1186/s12889-016-3833-z.

[24] A. Davillas, A. Burlinson, and H.-H. Liu. Getting warmer: Fuel poverty, objective and subjective health and well-being. *Energy Economics*, 106:105794, 2022. doi:10.1016/j.eneco.2021.105794.

[25] Department for Energy Security and Net Zero. Annual fuel poverty statistics report: 2025, 2025. URL https://www.gov.uk/government/statistics/annual-fuel-poverty-statistics-report-2025.

[26] Department for Levelling Up, Housing & Communities. Energy Performance of Buildings data, 2025. URL https://epc.opendatacommunities.org/.

[27] U. Ekelund, J. Tarp, J. Steene-Johannessen, B. H. Hansen, B. Jefferis, M. W. Fagerland, P. Whincup, K. M. Diaz, S. P. Hooker, A. Chernofsky, M. G. Larson, N. Spartano, R. S. Vasan, I.-M. Dohrn, M. Hagströmer, C. Edwardson, T. Yates, E. Shiroma, S. A. Anderssen, and I.-M. Lee. Dose-response associations between accelerometry measured physical activity and sedentary time and all cause mortality: systematic review and harmonised meta-analysis. *The BMJ*, 366:l4570, 2019. doi:10.1136/bmj.l4570.

[28] A. Ellaway, K. E. Lamb, N. S. Ferguson, and D. Ogilvie. Associations between access to recreational physical activity facilities and body mass index in Scottish adults. *BMC Public Health*, 16:756, 2016. doi:10.1186/s12889-016-3444-8.

[29] GBD 2021 Chronic Kidney Disease Collaborators. Global, regional, and national burden of chronic kidney disease, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet*, 404(10454):743–772, 2024. doi:10.1016/S0140-6736(24)00650-5.

[30] R. Geary et al. Ambient greenness, access to local green spaces, and subsequent mental health: a 10-year longitudinal dynamic panel study of 2.3 million adults in Wales. *Lancet Planetary Health*, 7(12):e942–e951, 2023. doi:10.1016/S2542-5196(23)00212-7.

[31] M. Green, C. Berragan, and S. Alex. Access to Healthy Assets and Hazards (AHAH), 2025. URL https://data.cdrc.ac.uk/dataset/access-healthy-assets-and-hazards-ahah2.

[32] C. N. B. Grey, S. Jiang, C. Nascimento, and W. Poortinga. The short-term health and psychosocial impacts of domestic energy efficiency investments in low-income areas: a controlled before and after study. *BMC Public Health*, 17:140, 2017. doi:10.1186/s12889-017-4075-4.

[33] R. V. Guha, D. Brickley, and S. Macbeth. Schema.org: evolution of structured data on the web. *Communications of the ACM*, 59(2):44–51, 2016. doi:10.1145/2844544.

[34] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. doi:10.1038/s41586-020-2649-2.

[35] Q. He, M. Sun, Y. Wang, G. Li, H. Zhao, Z. Ma, Z. Feng, T. Li, Q. Han, N. Sun, L. Li, and Y. Shen. Association between residential greenness and incident delirium: A prospective cohort study in the UK Biobank. *Science of The Total Environment*, 937:173341, 2024. doi:10.1016/j.scitotenv.2024.173341.

[36] J. Higgerson, E. Halliday, A. Ortiz-Nuñez, R. Brown, and B. Barr. Impact of free access to leisure facilities and community outreach on inequalities in physical activity: a quasi-experimental study. *Journal of Epidemiology and Community Health*, 72(3):252–258, 2018. doi:10.1136/jech-2017-209882.

[37] House of Commons Library. Local health conditions prevalence estimates based on QOF. https://github.com/houseofcommonslibrary/local-health-data-from-QOF, 2021.

[38] S. Huang, L. Tang, J. P. Hupy, Y. Wang, and G. Shao. A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing. *Journal of Forestry Research*, 32(1):1–6, 2021. doi:10.1007/s11676-020-01155-1.

[39] D. J. Hunter and K. S. Reddy. Noncommunicable diseases. *New England Journal of Medicine*, 369(14):1336–1343, 2013. doi:10.1056/nejmra1109345.

[40] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 9(3):90–95, 2007. doi:10.1109/MCSE.2007.55.

[41] K. Janowicz, A. Haller, S. J. Cox, D. Le Phuoc, and M. Lefrançois. Sosa: A lightweight ontology for sensors, observations, samples, and actuators. *Journal of Web Semantics*, 56:1–10, 2019. doi:10.1016/j.websem.2018.06.003.

[42] A. Jaworowska, T. Blackham, I. G. Davies, and L. Stevenson. Nutritional challenges and health implications of takeaway and fast food. *Nutrition Reviews*, 71 (5):310–318, May 2013. doi:10.1111/nure.12031.

[43] P. Jia. Spatial lifecourse epidemiology. *The Lancet Planetary Health*, 3(2): e57–e59, 2019. doi:10.1016/s2542-5196(18)30245-6.

[44] P. Jia, M. Luo, Y. Li, J.-S. Zheng, Q. Xiao, and J. Luo. Fast-food restaurant, unhealthy eating, and childhood obesity: A systematic review and meta-analysis. *Obesity Reviews*, 22(S1):e12944, 2021. doi:10.1111/obr.12944.

[45] Joint Nature Conservation Committee (JNCC). JNCC Sentinel-2 indices analysis ready data (ARD) normalised difference vegetation index (NDVI) v2. NERC EDS Centre for Environmental Data Analysis, 2025. URL https://catalogue.ceda.ac.uk/uuid/ebb1bd2603cc4efc8ce1d745d03932b5.

[46] M. Katsumi and M. S. Fox. icity transportation planning suite of ontologies (tpso). Technical Report, Enterprise Integration Laboratory, University of Toronto, 2020. Available at https://enterpriseintegrationlab.github.io/icity/iCityOntologyReport_1.2.pdf.

[47] S. M. Khan, X. Liu, S. Nath, E. Korot, L. Faes, S. K. Wagner, P. A. Keane, N. J. Sebire, M. J. Burton, and A. K. Denniston. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *The Lancet Digital Health*, 3:e51–e66, 2021. doi:10.1016/s2589-7500(20)30240-5.

[48] K. O. Lee, K. M. Mai, and S. Park. Green space accessibility helps buffer declined mental health during the covid-19 pandemic: evidence from big data in the united kingdom. *Nature Mental Health*, 1(2):124–134, 2023. doi:10.1038/s44220-023-00018-y.

[49] M. Q. Lim, X. Wang, O. Inderwildi, and M. Kraft. *The World Avatar—A World Model for Facilitating Interoperability*, page 39–53. Springer International Publishing, 2022. doi:10.1007/978-3-030-86215-2_4.

[50] B.-P. Liu, R. R. Huxley, T. Schikowski, K.-J. Hu, Q. Zhao, and C.-X. Jia. Exposure to residential green and blue space and the natural environment is associated with a lower incidence of psychiatric disorders in middle-aged and older adults: findings from the UK Biobank. *BMC Medicine*, 22(1), 2024. doi:10.1186/s12916-023-03239-1.

[51] L. Macdonald and colleagues. Associations between spatial access to physical activity facilities and frequency of physical activity: how do home and workplace neighbourhoods in West Central Scotland compare? *International Journal of Health Geographics*, 18(2), 2019. doi:10.1186/s12942-019-0166-z.

[52] C. D. Maidment, C. R. Jones, T. L. Webb, E. A. Hathway, and J. M. Gilbertson. The impact of household energy efficiency measures on health: A meta-analysis. *Energy Policy*, 65:583–593, 2014. doi:10.1016/j.enpol.2013.10.054.

[53] L. Maitre, J.-B. Guimbaud, C. Warembourg, N. Güil-Oumrait, P. M. Petrone, M. Chadeau-Hyam, M. Vrijheid, X. Basagaña, and J. R. Gonzalez. State-of-the-art methods for exposure-health studies: Results from the exposome data challenge event. *Environment International*, 168:107422, 2022. doi:10.1016/j.envint.2022.107422.

[54] M. A. Mansournia and D. G. Altman. Population attributable fraction. *The BMJ*, 360:k757, 2018. doi:10.1136/bmj.k757.

[55] K. E. Mason, N. Pearce, and S. Cummins. 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(1):e24–e33, 2018. doi:10.1016/S2468-2667(17)30212-8.

[56] K. E. Mason, L. Palla, N. Pearce, J. Phelan, and S. Cummins. Genetic risk of obesity as a modifier of associations between neighbourhood environment and body mass index: an observational study of 335,046 UK Biobank participants. *BMJ Nutrition, Prevention & Health*, 3(2):247–255, 2020. doi:10.1136/bmjnph-2020-000107.

[57] W. McKinney. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56, 2010.

[58] Ministry of Housing, Communities and Local Government. English Indices of Deprivation 2019, 2025. URL https://www.gov.uk/government/statistics/english-indices-of-deprivation-2025.

[59] T. Münzel, O. Hahad, A. Daiber, and J. Lelieveld. The contribution of the exposome to the burden of cardiovascular disease. *Nature Reviews Cardiology*, 20(10): 651–669, 2023. doi:10.1038/s41569-023-00873-3.

[60] J. Newbury, L. Arseneault, A. Caspi, T. E. Moffitt, C. L. Odgers, and H. L. Fisher. Cumulative effects of neighborhood social adversity and personal crime victimization on adolescent psychotic experiences. *Schizophrenia Bulletin*, 44(2):348–358, 2018. doi:10.1093/schbul/sbx060.

[61] J. B. Newbury, L. Arseneault, A. Caspi, T. E. Moffitt, C. L. Odgers, and H. L. Fisher. In the eye of the beholder: Perceptions of neighborhood adversity and psychotic experiences in adolescence. *Development and Psychopathology*, 29(5): 1823–1837, 2017. doi:10.1017/S0954579417001420.

[62] NHS Digital. Quality and Outcomes Framework, 2024–25, 2025. URL https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement-prevalence-and-exceptions-data/2024-25.

[63] NHS Digital. Patients registered at a GP practice, november 2025, 2025. URL https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice/november-2025.

[64] NumPy Developers. NumPy (v2.1.3), 2024. URL https://pypi.org/project/numpy/2.1.3/. Python package.

[65] Office for National Statistics. Rural Urban Classification (2021) of LSOAs in EW, 2021. URL https://www.data.gov.uk/dataset/7f9daee2-ba87-4436-9050-0118a70248c0/rural-urban-classification-2021-of-lsoas-in-ew.

[66] Office for National Statistics. Urban rural classification - Scotland, 2021. URL https://www.data.gov.uk/dataset/f00387c5-7858-4d75-977b-bfdb35300e7f/urban-rural-classification-scotland.

[67] Office for National Statistics. More adults are active in areas with a higher number of sports facilities, 2024. URL https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/moreadultsareactiveinareaswithahighernumberofsportsfacilities/2024-03-07. Published 7 March 2024. Accessed 27 Aug 2025.

[68] Office for National Statistics. Lower layer super output area population estimates (supporting information), 2025. URL https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/lowersuperoutputareamidyearpopulationestimates?

[69] M. A. Pereira, A. I. Kartashov, C. B. Ebbeling, L. Van Horn, M. L. Slattery, D. R. Jacobs, and D. S. Ludwig. Fast-food habits, weight gain, and insulin resistance (the CARDIA study): 15-year prospective analysis. *The Lancet*, 365(9453):36–42, Jan. 2005. ISSN 0140-6736. doi:10.1016/S0140-6736(04)17663-0.

[70] R. Perez-Padilla, L. Wehbe, and GBD 2021 Chronic Respiratory Diseases Collaborators. Global, regional, and national burden of chronic respiratory diseases, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet Respiratory Medicine*, 12(11):949–965, 2024. doi:10.1016/S2213-2600(24)00339-4.

[71] E. Pineda, J. Stockton, S. Scholes, C. Lassale, and J. S. Mindell. Food environment and obesity: a systematic review and meta-analysis. *BMJ Nutrition, Prevention & Health*, Apr. 2024. doi:10.1136/bmjnph-2023-000663. URL https://nutrition.bmj.com/content/early/2024/04/21/bmjnph-2023-000663.

[72] M. Poelman, M. Strak, O. Schmitz, G. Hoek, D. Karssenberg, M. Helbich, A.-M. Ntarladima, M. Bots, B. Brunekreef, R. Grobbee, M. Dijst, and I. Vaartjes. Relations between the residential fast-food environment and the individual risk of cardiovascular diseases in The Netherlands: A nationwide follow-up study. *European Journal of Preventive Cardiology*, 25(13):1397–1405, Sept. 2018. doi:10.1177/2047487318769458.

[73] H. Y. Quek, M. Hofmeister, S. D. Rihm, J. Yan, J. Lai, G. Brownbridge, M. Hillman, S. Mosbach, W. Ang, Y.-K. Tsai, D. N. Tran, W. Tan, Soon Kang, and M. Kraft. Dynamic knowledge graph applications for augmented built environments through "The World Avatar". *Journal of Building Engineering*, 91:109507, 2024. doi:10.1016/j.jobe.2024.109507.

[74] R. Raab, A. Küderle, A. Zakreuskaya, A. D. Stern, J. Klucken, G. Kaissis, D. Rueckert, S. Boll, R. Eils, H. Wagener, and B. M. Eskofier. Federated electronic health records for the European Health Data Space. *The Lancet Digital Health*, 5 (11):e840–e847, 2023. doi:10.1016/s2589-7500(23)00156-5.

[75] H. Rijgersberg, M. Wigham, and J. Top. How semantics can improve engineering processes: A case of units of measure. *Semantic Web*, 4(1):3–13, 2013. doi:10.3233/SW-2012-0069.

[76] J. F. Sallis, E. Cerin, J. Kerr, M. A. Adams, T. Sugiyama, L. B. Christiansen, J. Schipperijn, R. Davey, D. Salvo, L. D. Frank, I. De Bourdeaudhuij, and N. Owen. Built environment, physical activity, and obesity: Findings from the International Physical Activity and Environment Network (IPEN) adult study. *Annual Review of Public Health*, 41:119–139, 2020. doi:10.1146/annurev-publhealth-040218-043657.

[77] M. J. Salois. Obesity and diabetes, the built environment, and the 'local' food economy in the United States, 2007. *Economics & Human Biology*, 10(1):35–42, Jan. 2012. doi:10.1016/j.ehb.2011.04.001.

[78] C. Sarkar. Residential greenness and adiposity: Findings from the UK Biobank. *Environment International*, 106:1–10, 2017. doi:10.1016/j.envint.2017.05.016. URL http://dx.doi.org/10.1016/j.envint.2017.05.016.

[79] C. Sarkar, C. Webster, and J. Gallacher. Are exposures to ready-to-eat food environments associated with type 2 diabetes? a cross-sectional study of 347 551 UK Biobank adult participants. *The Lancet Planetary Health*, 2(10):e438–e450, Oct. 2018. doi:10.1016/S2542-5196(18)30208-0.

[80] M. Schaekermann, T. Spitz, M. Pyles, H. Cole-Lewis, E. Wulczyn, S. R. Pfohl, D. Martin, R. Jaroensri, G. Keeling, Y. Liu, S. Farquhar, Q. Xue, J. Lester, C. Hughes, P. Strachan, F. Tan, P. Bui, C. H. Mermel, L. H. Peng, Y. Matias, G. S. Corrado, D. R. Webster, S. Virmani, C. Semturs, Y. Liu, I. Horn, and P.-H. Cameron Chen. Health equity assessment of machine learning performance (HEAL): a framework and dermatology AI model case study. *eClinicalMedicine*, 70:102479, 2024. doi:10.1016/j.eclinm.2024.102479.

[81] P. P. Schneider, R. A. Smith, A. M. Bullas, H. Quirk, T. Bayley, S. J. Haake, A. Brennan, and E. Goyder. Multiple deprivation and geographic distance to community physical activity events – achieving equitable access to parkrun in England. *Public Health*, 189:48–53, 2020. doi:10.1016/j.puhe.2020.09.002.

[82] S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference*, 2010.

[83] R. A. Sharpe, K. E. Machray, L. E. Fleming, T. Taylor, W. Henley, R. Taylor, and B. W. Wheeler. Household energy efficiency and health: Area-level analysis of hospital admissions in England. *Environment International*, 133:105164, 2019. doi:10.1016/j.envint.2019.105164.

[84] Single Online Home National Digital Team. Police.uk data downloads: Street-level crime, outcome, and stop and search data, 2025. URL https://data.police.uk/data/.

[85] N. J. Smith. Patsy: Describing statistical models in Python (v1.0.1), 2025. URL https://patsy.readthedocs.io/. Python package.

[86] R. A. Smith, P. P. Schneider, R. Cosulich, H. Quirk, A. M. Bullas, S. J. Haake, and E. Goyder. Socioeconomic inequalities in distance to and participation in a community-based running and walking activity: A longitudinal ecological study of parkrun 2010 to 2019. *Health & Place*, 71:102626, 2021. doi:10.1016/j.healthplace.2021.102626.

[87] M. Somerville, I. Mackenzie, P. Owen, and D. Miles. Housing and health: does installing heating in their homes improve the health of children with asthma? *Public Health*, 114(6):434–439, 2000. doi:10.1038/sj.ph.1900687.

[88] Statsmodels Developers. statsmodels (v0.14.5), 2025. URL https://pypi.org/project/statsmodels/0.14.5/. Python package.

[89] B. J. Stear, T. Mohseni Ahooyi, J. A. Simmons, C. Kollar, L. Hartman, K. Beigel, A. Lahiri, S. Vasisht, T. J. Callahan, C. M. Nemarich, J. C. Silverstein, and D. M. Taylor. Petagraph: A large-scale unifying knowledge graph framework for integrating biomolecular and biomedical data. *Scientific Data*, 11:1338, 2024. doi:10.1038/s41597-024-04070-w.

[90] B. A. Swinburn, V. I. Kraak, S. Allender, V. J. Atkins, P. I. Baker, J. R. Bogard, H. Brinsden, A. Calvillo, O. De Schutter, R. Devarajan, M. Ezzati, S. Friel, S. Goenka, R. A. Hammond, G. Hastings, C. Hawkes, M. Herrero, P. S. Hovmand, M. Howden, L. M. Jaacks, A. B. Kapetanaki, M. Kasman, L. A. King, T. Kunej, B. Larijani, T. Leet, T. Lobstein, N. Lonc, V. K. R. Matsudo, S. D. H. Mills, G. Morgan, A. Morshed, M. L. Motu'apuaka, W. Mphatswe, O. Mytton, G. O'Kane, T. Oni, M. Otim, A. Pan, D. W. Patterson, M. Pescud, E. D. Pires, R. Poni, E. Raciot, J. Reynolds, G. Sacks, J. Salles, L. Salmon, J. Sampedro, S. Sang-aroon, B. Sarr, T. Sathish, N. Savona, S. Sengee, A. Sharkey, T. A. Sheldon, I. Shemilt, R. Shrimpton, K. R. Siegel, M. L. Sierra, G. Singh, J. Sorensen, D. Stjepanovic, S. Suggs, S. T., T. Tango, T. T., F. Taylor, R. Tolhurst, P. Trowbridge, M. Van C, V. Van der, M. V., A. L. Vogl, F. Watson, R. Welch, R. Wijesinha-Bettoni, B. Wood, and A. Wood. The global syndemic of obesity, undernutrition, and climate change: The Lancet commission report. *The Lancet*, 393(10173):791–846, 2019. doi:10.1016/S0140-6736(18)32822-8.

[91] Y. R. Tan, M. Hofmeister, S. Z. Phua, G. Brownbridge, K. Rustagi, J. Akroyd, S. Mosbach, A. Bhave, and M. Kraft. Beyond connected digital twins – can digital twins really deliver sustainable cities? *Sustainable Cities and Society*, 131:106596, 2025. doi:10.1016/j.scs.2025.106596.

[92] L. Tang, D. Li, Y. Ma, F. Cui, J. Wang, R. Liu, and Y. Tian. Green environments and cardiometabolic health: exploring incidence and progression through multi-state analysis. *npj Urban Sustainability*, 5(13), 2025. doi:10.1038/s42949-025-00201-3.

[93] The Health Foundation. Relationship between neighbourhood crime and health, 2024. Available at: https://www.health.org.uk/evidence-hub/housing/neighbourhood-and-community/relationship-between-neighbourhood-crime-and-health.

[94] The Matplotlib development team. Matplotlib (v3.10.0), 2024. URL https://pypi.org/project/matplotlib/3.10.0/. Python package.

[95] The pandas development team. pandas (v2.2.3), 2024. URL https://pypi.org/project/pandas/2.2.3/. Python package.

[96] The SciPy community. SciPy (v1.15.1), 2025. URL https://pypi.org/project/scipy/1.15.1/. Python package.

[97] G. Tu, K. Morrissey, R. A. Sharpe, and T. Taylor. Combining self-reported and sensor data to explore the relationship between fuel poverty and health well-being in UK social housing. *Wellbeing, Space and Society*, 3:100070, 2022. doi:10.1016/j.wss.2021.100070.

[98] UK Government. National planning policy framework. https://www.gov.uk/government/publications/national-planning-policy-framework--2, 2024. Accessed 2025-12-15.

[99] T. I. Verhoef, V. Trend, B. Kelly, N. Robinson, P. Fox, and S. Morris. Cost-effectiveness analysis of offering free leisure centre memberships to physically inactive members of the public receiving state benefits: a case study. *BMC Public Health*, 16:616, 2016. doi:10.1186/s12889-016-3300-x.

[100] R. Vermeulen, E. L. Schymanski, A.-L. Barabási, and G. W. Miller. The exposome and health: Where chemistry meets biology. *Science*, 367(6476):392–396, 2020. doi:10.1126/science.aay3164.

[101] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Klöckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G.-L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito,

T. Krauss, U. Upadhyay, Y. O. Halchenko, and Y. Vázquez-Baeza. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3): 261–272, 2020. doi:10.1038/s41592-019-0686-2.

[102] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982. doi:10.2307/1912526.

[103] C. P. Wild. The exposome: from concept to utility. *International Journal of Epidemiology*, 41:24–32, 2012. doi:10.1093/ije/dyr236.

[104] S. Wilding, N. Ziauddeen, D. Smith, P. Roderick, D. Chase, and N. A. Alwan. Are environmental area characteristics at birth associated with overweight and obesity in school-aged children? findings from the slope (studying lifecourse obesity predictors) population-based cohort in the south of england. *BMC Medicine*, 43(18), 2020. doi:10.1186/s12916-020-01513-0.

[105] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 2016. doi:10.1038/sdata.2016.18.

[106] World Health Organization Regional Office for Europe. Implementation framework for phase vii (2019–2024) of the who european healthy cities network: goals, requirements and strategic approaches. Technical report, WHO Regional Office for Europe, Copenhagen, 2019. URL https://www.who.int/docs/librariesprovider2/default-document-library/04-final-phase-vii-implementation-framework_eng.pdf. Accessed 2025-12-15.

[107] J. Zhang, J. Morley, J. Gallifant, C. Oddy, J. T. Teo, H. Ashrafian, B. Delaney, and A. Darzi. Mapping and evaluating national data flows: transparency, privacy, and guiding infrastructural transformation. *The Lancet Digital Health*, 5(10): e737–e748, 2023. doi:10.1016/s2589-7500(23)00157-7.

[108] G. Zou. A modified poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology*, 159(7):702–706, 2004. doi:10.1093/aje/kwh090.