

Linking smoking and caffeine biomarkers to socioeconomic and health outcomes through wastewater-based epidemiology: a tool for hypothesis generation

Barbara Kasprzyk-Hordern

bkh20@bath.ac.uk

University of Bath

Kishore Kumar Jagadeesan

University of Bath <https://orcid.org/0000-0003-0916-256X>

Natalie Sims

RSC

Davey Jones

Bangor University

Matthew Wade

UKHSA <https://orcid.org/0000-0001-9824-7121>

Article

Keywords: Wastewater-based epidemiology, pharmaceuticals, public health, lifestyle, socioeconomic status, smoking, nicotine, caffeine

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-7658391/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Linking smoking and caffeine biomarkers to socioeconomic and health outcomes through wastewater-based epidemiology: a tool for hypothesis generation

Barbara Kasprzyk-Hordern^{1,2*}, Kishore Jagadeesan^{1,2}, Natalie Sims¹, Davey L. Jones^{3,4}, Matthew J. Wade^{2,5}

¹*Department of Chemistry, University of Bath, Claverton Down, Bath, BA2 7AY, UK*

²*Centre of Excellence in Water Based Early Warning Systems for Health Protection, Bath, BA2 7AY, UK*

³*School of Environmental and Natural Sciences, Bangor University, Bangor, Gwynedd, LL57 2UW, UK*

⁴*Verily Life Sciences LLC., Dallas, Texas 75019, United States*

⁵*Chief Data Officer Group, UK Health Security Agency, London, E14 4PU, UK*

**Author for correspondence: B.Kasprzyk-Hordern@bath.ac.uk*

Abstract

Wastewater-based epidemiology (WBE) represents a powerful tool for population health surveillance, however, the systematic integration of wastewater-derived human behavioural markers (WBE_{CI}) with socioeconomic and demographic data remains understudied. The aim of this study was to explore the potential of integrating wastewater chemical markers (WBE_{CI}) with population census data to unravel associations between disease prevalence, lifestyle choices and key risk factors associated with nicotine and caffeinated product use. This was critically evaluated in a multi-city ($n = 10$), national level WBE study in England: 11 WWTPs, 9 sampling days, 2 biological replicates per sample giving 198 data points (day equivalents) per each CI (72 WBE_{CI}s) giving 14256 observations for WBE_{CI}s and 830 observations for ONS_{SEDI}. Liquid chromatography and mass spectrometry were used for WBE_{CI} quantification in wastewater. Multivariate analyses were used to investigate relationships between lifestyle factors and health outcomes. Associations among WBE_{CI}s and ONS_{SEDI} were assessed using Spearman correlation, while principal-component analysis (PCA) and hierarchical cluster analysis (HCA) were applied to explore patterns and clustering. Given the cross-sectional ecological design and aggregated city-level data, multiple linear regression (MLR) was used to identify population-level associations and generate hypotheses rather than infer causality. PCA and HCA identified distinct city clusters characterised by shared patterns of lifestyle exposures, health-related biomarkers, and socioeconomic conditions. Cities with higher deprivation generally exhibited less favourable profiles, including elevated smoking-related WBE_{CI}s and a higher prevalence of pharmaceutical markers associated with chronic diseases. Smoking WBE_{CI}s were positively associated with deprivation, disability, poor self-reported health, and lower education, highlighting the strong links between smoking and health inequalities. Clustering analyses further demonstrated the co-occurrence of deprivation, smoking, pharmaceutical use, and poorer health outcomes, suggesting that wastewater-derived indicators can capture broader patterns of population health disadvantage and support the identification of communities facing similar public health challenges. In contrast, oxidative stress biomarkers were more strongly associated with caffeine consumption and cities with more favourable socioeconomic conditions. Although requiring further investigation, this finding illustrates the potential of WBE to uncover previously unrecognised population-level relationships and generate new epidemiological hypotheses.

Keywords: Wastewater-based epidemiology, pharmaceuticals, public health, lifestyle, socioeconomic status, smoking, nicotine, caffeine.

1. Introduction

Smoking is a significant public health issue and a leading cause of preventable death and disease globally (ONS, 2023a). This is despite significant reductions in smoking prevalence over the past decade (ONS, 2022). According to WHO smoking causes > 8 million deaths each year (WHO, 2022). Smoking prevalence is commonly associated with a variety of socioeconomic factors including education level and household income. Comparison of smoking prevalence and economic activity in the UK indicated that those who were unemployed comprised a higher proportion of current smokers (20.5%), compared with those who were in paid employment (12.7%), and those who were economically inactive (12.7%) (ONS, 2023b). Of those who were classified as being in a routine and manual socio-economic classification, 22.8% were current smokers, compared with 8.3% of managerial and professional occupations.

While smoking is a well-established risk factor for multiple chronic diseases and an important indicator of health inequalities, the relationship between coffee consumption—another common lifestyle factor—and

health outcomes remains incompletely understood. Beneficial effects of drinking coffee on risk of death from all causes (Gunter, Murphy et al. 2017), as well as reduction of oxidative stress (Vargas-Pozada, Ramos-Tovar et al. 2023) are widely reported. However, caffeine has also been considered a risk factor for increased risk of anxiety and depression (Liu, Wang et al. 2024).

This study aims to explore whether wastewater-derived smoking and caffeine biomarkers are associated with city-level socioeconomic and health characteristics by integrating Wastewater-Based Epidemiology (WBE) data with socioeconomic and demographic indices, and to evaluate the potential of WBE as a tool for generating population-level public health hypotheses.

WBE is an innovative approach that enables the monitoring of infectious disease prevalence (Sims and Kasprzyk-Hordern 2020, Wade, Lo Jacomo et al. 2022) and health behaviours such as recreational/illicit drug (González-Mariño, Baz-Lomba et al. 2020) alcohol and tobacco use at the population level (Castiglioni, Senta et al. 2014, Chen, Venkatesan et al. 2019, Mackie, Tscharke et al. 2019, Gracia-Lor, Rousis et al. 2020, Montes, Rodil et al. 2020, Sims and Kasprzyk-Hordern 2020, Zheng, Gartner et al. 2020, Thanh, Vu et al. 2022, Wang, Zheng et al. 2024). WBE offers a population-level ‘fingerprint’ of community health that can be monitored in near real-time, theoretically at any place and time globally. WBE is relatively comprehensive, timely and unbiased in comparison to other approaches (e.g. clinical surveillance, community-wide participatory surveys). Importantly, it captures signals from the vast majority of individuals within a community, providing an inclusive and representative assessment of behavioural and other exposure, and population health. When integrated with Census-derived socioeconomic data, WBE can help identify population-level patterns and potential confounding factors underlying observed exposure–outcome associations, thereby supporting hypothesis generation and interpretation of epidemiological findings.

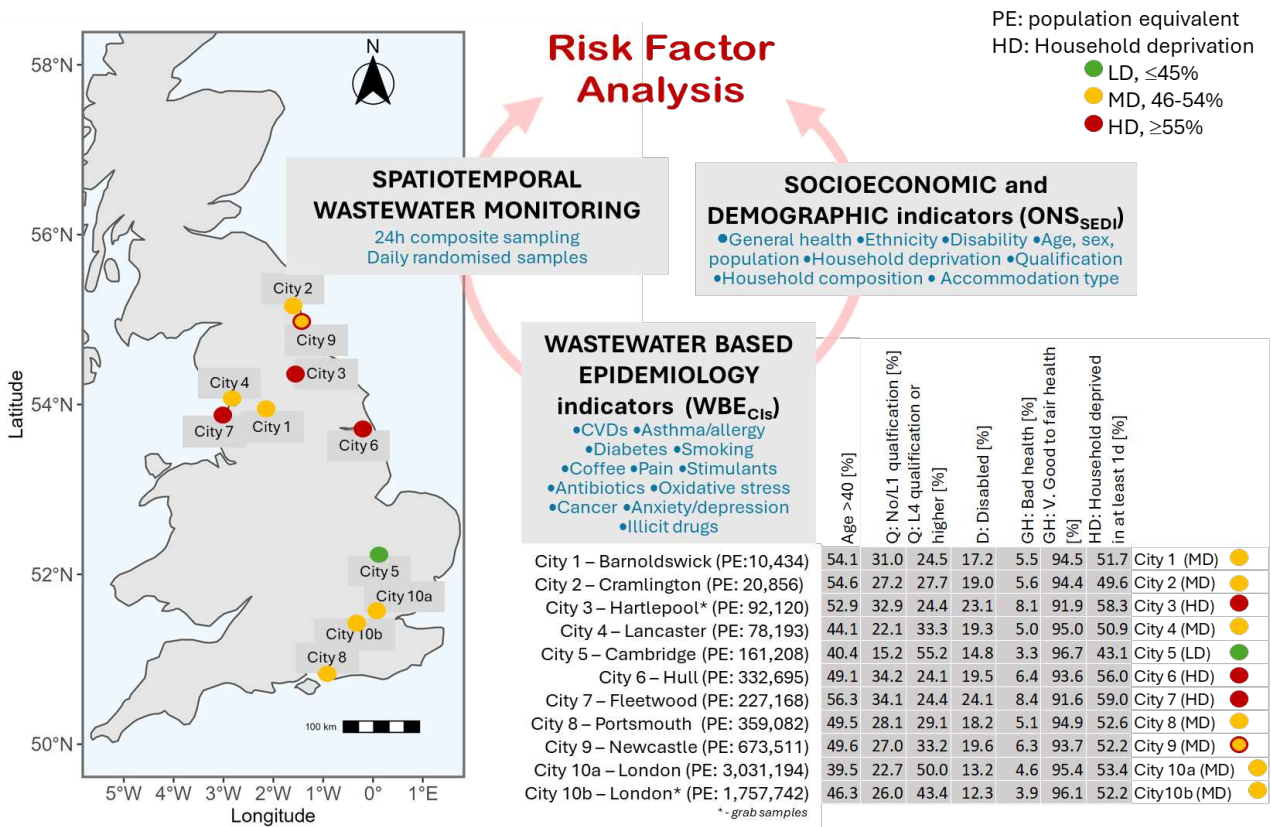


Figure 1. Risk factor analysis combining spatiotemporal wastewater monitoring, WBE_{CIS} and ONS_{SEDI} indicators (CI – chemical indicators, SEDI – socioeconomic and demographic indicators).

This study explores, for the first time, associations between smoking- and caffeine-related WBE_{CIS}, pharmaceutical WBE_{CIS}, and UK Office for National Statistics socioeconomic and demographic indicators (ONS_{SEDI}) in a national multi-city study focussed on 10 geographically distributed towns and cities in England (Fig. 1) representing a combined population of ~7 million. Detailed discussion of spatial distribution in WBE_{CIS} and inter-CI associations is provided elsewhere (Kasprzyk-Hordern, Sims et al. 2023, Kasprzyk-Hordern, Sims et al. 2025). 72 pharmaceutical and health WBE_{CIS} (Tab. 1) were selected as proxies for disease to test exposure-health outcome associations linked with lifestyle choices such as smoking and caffeinated

85 products use. 83 ONS_{SEDI}s (Tab. 1) were selected to explore associations between smoking demography and
86 socioeconomic status. Correlations between nicotine/cotinine, other WBE markers and
87 socioeconomic/demographic indicators have also been reported in a few studies (Choi, Tscharke et al. 2019,
88 Rousis, Li et al. 2022, Shao, Zhang et al. 2022, Yang, Zheng et al. 2022). For example, Choi et al. found
89 significant negative correlations between IRSAD (a measure of advantage and disadvantage, summarising the
90 economic and social conditions of people and households) and nicotine (Choi, Tscharke et al. 2019). Further,
91 it has been shown that populations experiencing disadvantage in income, education, occupation, and marital
92 status showed positive correlations with biomarkers and/or proxies of conditions including cardiac
93 arrhythmias, cardiovascular disease, anxiety disorder and type 2 diabetes (Rousis, Li et al. 2022). Yang et al.
94 found higher metformin use correlated with socially disadvantaged populations (lower education and lower
95 income levels) (Yang, Zheng et al. 2022). Zhang et al. studied allergy prevalence and environmental factors
96 and found that antihistamine use was significantly associated with smoking, green space and pollen but not
97 with air pollution (Zhang, Du et al. 2025). Li et al. (Li et al. 2024) reported associations between the
98 consumption of antimicrobials and socioeconomic status.

99 By combining wastewater-derived biomarkers with city-level socioeconomic and health indicators, this study
100 seeks to demonstrate how WBE can complement traditional epidemiological approaches by identifying
101 population-level patterns and generating hypotheses for future public health research.

102 **Table 1.** Summary of wastewater-based epidemiology chemical indicators (WBE_{CIs}) and socioeconomic and
103 demographic indicators (ONS_{SEDI}) used in this study.

WASTEWATER BASED EPIDEMIOLOGY INDICATORS (WBE _{CI})	SOCIOECONOMIC and DEMOGRAPHIC INDICATORS (ONS _{SEDI})
CVDs: •Irbesartan •Bezafibrate •Atorvastatin •Atenolol •Lisinopril ASTHMA/ALLERGY: •Fexofenadine •Salbutamol DIABETES: •Metformin •Sitagliptin SMOKING: •Nicotine •Cotinine COFFEE/cafeinated products intake: •Caffeine •1,7-dimethylxanthine PAIN treatment: •Acetaminophen •Tramadol •N-desmethyltramadol •O-Desmethyltramadol •Codeine •Dihydrocodeine •Dihydromorphine •Ibuprofen •2-hydroxyibuprofen •Carboxyibuprofen •Naproxen •O-desmethylnaproxen •Diclofenac •4-Hydroxydiclofenac •Pregabalin •N-methylpregabalin •Gabapentin STIMULANTS/DECONGESTANTS: •Ephedrine •Norephedrine ANTIBIOTICS: •Metronidazole •Hydroxy-metronidazole •Sulfapyridine •N-acetyl sulfapyridine •Sulfamethoxazole •N-acetyl sulfamethoxazole •Trimethoprim •Ofloxacin •Sulfasalazine •Flucloxacillin •Clarithromycin •N-desmethyl clarithromycin •Ethambutol •5-Hydroxy-pyrazinoic acid •Isonicotinic acid OXIDATIVE STRESS: •HNE-MA Heartburn and peptic ulcers: •Cimetidine CANCER/AUTOIMMUNE conditions: •Methotrexate KIDNEY function: •Creatinine ANXIETY/DEPRESSION: •Venlafaxine •Desmethylvenlafaxine ILLICIT drugs abuse: - Cocaine, Benzoylcegonine, Cocaethylene, Amphetamine, Methamphetamine, MDMA, MDA, Mephedrone	GENERAL HEALTH: •Very good health •Good health •Fair health •Bad health •Very bad health •Very good to fair health •Bad to very bad health ETHNICITY: •Asian •Black •Mixed •White •Other DISABILITY under the Equality Act: •Disabled •Disabled: Day-to-day activities limited a lot •Disabled: Day-to-day activities limited a little •Not disabled •Not disabled: Has long term physical or mental health condition but day-to-day activities are not limited •Not disabled: No long term physical or mental health conditions AGE, SEX, POPULATION: •Age: ≥40 years •Age: >40 years •Sex: Female •Sex: Male •Population HOUSEHOLD DEPRIVATION: •Household is not deprived in any dimension •Household is deprived in 1 dimension •Household is deprived in 2 dimensions •Household is deprived in 3 dimensions •Household is deprived in 4 dimensions •Household is deprived in at least 1 dimension HIGHEST LEVEL OF QUALIFICATION: •No qualifications •Level 1 and entry level qualifications •Level 2 qualifications •Apprenticeship •Level 3 qualifications •Level 4 qualifications and above HOUSEHOLD COMPOSITION: •One person household (All ≥66 years / Other) •Single family household / All ≥ 66 years / Married or civil partnership couple (No children / Dependent children / All children non-dependent) / Cohabiting couple family (No children / With dependent children / All children non-dependent) / Lone parent family (With dependent children / All children non-dependent) / Other single family household •Other household types •Other household types: With dependent children •Other household types: incl. all full-time students and all aged ≥ 6 years ACCOMMODATION TYPE: •Detached •Semi-detached •Terraced •In a purpose-built block of flats or tenement •Part of a converted or shared house, including bedsits •Part of another converted building, e.g., former school, church or warehouse •In a commercial building, e.g., in an office building, hotel or over a shop •A caravan or other mobile or temporary structure

104

105 **2. Experimental**

106 **2.1. Wastewater treatment sites selected for the study and sampling approaches utilised**

107 Samples were collected in November 2021 from 11 wastewater treatment plants (WWTPs) serving 10 cities
108 during nine sampling campaigns conducted on randomly selected days. In each campaign, all sites were
109 sampled on the same day using 24-hour composite sampling, with subsamples taken at 15-minute or 1-hour
110 intervals using refrigerated autosamplers when available. Two of the 11 sites also provided additional grab
111 samples: WWTP 10 (10 out of 10 campaigns) and WWTP 3 (3 out of 10 campaigns). All samples were
112 transported frozen and processed within 24 hours of collection.

113 The selection of the 11 WWTPs was guided by the following criteria: (i) availability of 24-hour composite
114 sampling (primary requirement), (ii) on-site flow measurement capability (primary requirement), (iii)
115 documented population equivalent (primary requirement), (iv) representation across a wide population range,

116 (v) minimal industrial input (secondary requirement), (vi) broad geographical coverage, and (vii) availability
117 of at least eight 24-hour composite samples per WWTP.

118
119

2.2. Analysis of WBE_{CIS}

120 WBE_{CIS} were analysed as described elsewhere (Kasprzyk-Hordern, Sims et al. 2023).

121 50 mL wastewater samples were pH adjusted to 7.5 ± 0.5 , spiked with an IS (Internal Standard) mix (100
122 ng/compound) and filtered through glass microfibre GF/F (0.7 μm) filters. WBE_{CIS} were extracted with Solid-
123 Phase Extraction (SPE) using HLB (Hydrophilic Lipophilic Balance, Waters, UK) cartridges conditioned with
124 2 mL MeOH followed by 2 mL H₂O at $\sim 5 \text{ mL min}^{-1}$, elution with 4 mL of MeOH, drying under N₂ at 40 °C
125 and resuspension in 500 μL of 80:20 H₂O:MeOH. Two full biological replicates (two samples independently
126 processed and measured) were prepared for every sample analysed.

127 Extracted WBE_{CIS} were analysed using Xevo TQD-ESI triple quadrupole mass spectrometer (Waters, UK)
128 coupled with ACQUITY UPLC™ ultra-performance liquid chromatography instrument (Waters, UK) with an
129 electrospray ionisation source (ESI). Two analytical methods were used.

130 NCD (non-communicable WBE_{CI} disease markers) were analysed using a reversed-phase BEH C18 column
131 (150 x 1.0 mm, 1.7 μm particle size) and TQD-ESI set in both ESI+ and ESI- modes using the following mobile
132 phase conditions: ESI+ mobile phase A: H₂O:MeOH 95:5 + 0.1 % HCOOH and mobile phase B: 100 % MeOH
133 (34 min method time) and ESI- mobile phase A: H₂O:MeOH 80:20 + 1 mM NH₄COOH, mobile phase B: 5:95
134 H₂O: MeOH + 1 mM NH₄COOH (23 min method time). The mobile phase rate was set at 0.04 mL min⁻¹ with
135 an injection volume of 15 μL . The TQD-ESI was set at: source temperature, 150 °C, desolvation temperature,
136 400 °C, cone gas flow, 100 L h⁻¹, desolvation gas flow, 550 L h⁻¹. Nitrogen was used as the nebulising and
137 desolvation gas, and argon as the collision gas.

138 Antimicrobial agents were analysed using a reversed-phase BEH C18 column (150 x 1.0 mm, 1.7 μm particle
139 size) and TQD-ESI set in ESI+. A 19 min gradient elution using H₂O:MeOH (95:5, with 0.1% HCOOH) and
140 MeOH (100%) as mobile phases at a mobile phase flow of 0.2 mL min⁻¹ and injection volume were applied
141 (Holton and Kasprzyk-Hordern 2021).

142 MassLynx V4.1 and TargetLynx (Waters Lab Informatics, UK) were used as the coordinating and the data
143 processing program, respectively. All methods were fully validated and quality assurance process was
144 undertaken according to established protocols (Holton and Kasprzyk-Hordern 2021) and (Kasprzyk-Hordern,
145 Sims et al. 2023).

2.3. Demographic and socioeconomic indicators (ONS_{SEDI}s)

147 Demographic and socioeconomic indicators (ONS_{SEDI}s) for each wastewater treatment plant (WWTP)
148 catchment were estimated using 2021 Census data from the UK Office for National Statistics (ONS)(ONS
149 2022), provided at the Output Area (OA) level. Variables covered the following domains:

150 **Demography:** age, sex, household deprivation, household composition, housing

151 Household deprivation (HD) is classified as: (1) Household is not deprived in any dimension, (2) Household
152 is deprived in one dimension, (3) Household is deprived in two dimensions, (4) Household is deprived in three
153 dimensions, (5) Household is deprived in four dimensions, and is based on four selected household
154 characteristics:

- 155 • Education: A household is classified as deprived in the education dimension if no one has at least level 2
156 education and no one aged 16 to 18 years is a full-time student.
- 157 • Employment: A household is classified as deprived in the employment dimension if any member, not a
158 full-time student, is either unemployed or economically inactive due to long-term sickness or disability.
- 159 • Health: A household is classified as deprived in the health dimension if any person in the household has
160 general health that is bad or very bad or is identified as disabled. People who have assessed their day-to-
161 day activities as limited by long-term physical or mental health conditions or illnesses are considered
162 disabled. This definition of a disabled person meets the harmonised standard for measuring disability and
163 is in line with the Equality Act (2010).
- 164 • Housing: A household is classified as deprived in the housing dimension if the household's
165 accommodation is either overcrowded, in a shared dwelling, or has no central heating.

Household composition (HC) classifies households by the relationships between household members (household composition). One-family households are classified by: (1) the number of dependent children and (2) family type (married, civil partnership or cohabiting couple family, or lone parent family). Other households are classified by: (3) the number of people, (4) the number of dependent children, (5) whether the household consists only of students or only of people aged 66 and over.

Housing: Accommodation type has eight categories: (1) Detached, (2) Semi-detached, (3) Terraced, (4) In a purpose-built block of flats or tenement, (5) Part of a converted or shared house, including bedsits, (6) Part of another converted building, for example, former school, church or warehouse, (7) In a commercial building, for example, in an office building, hotel or over a shop, (8) A caravan or other mobile or temporary structure.

Education: highest level of qualification

Highest level of qualification classifies usual residents aged 16 years and over by their highest level of qualification. There are eight categories:

(1) No qualifications, (2) Level 1 and entry level qualifications: 1 to 4 GCSEs grade A* to C, Any GCSEs at other grades, O levels or CSEs (any grades), 1 AS level, NVQ level 1, Foundation GNVQ, Basic or Essential Skills, (3) Level 2 qualifications: 5 or more GCSEs (A* to C or 9 to 4), O levels (passes), CSEs (grade 1), School Certification, 1 A level, 2 to 3 AS levels, VCEs, Intermediate or Higher Diploma, Welsh Baccalaureate Intermediate Diploma, NVQ level 2, Intermediate GNVQ, City and Guilds Craft, BTEC First or General Diploma, RSA Diploma, (4) Apprenticeship, (5) Level 3 qualifications: 2 or more A levels or VCEs, 4 or more AS levels, Higher School Certificate, Progression or Advanced Diploma, Welsh Baccalaureate Advance Diploma, NVQ level 3; Advanced GNVQ, City and Guilds Advanced Craft, ONC, OND, BTEC National, RSA Advanced Diploma, (6) Level 4 qualifications or above: degree (BA, BSc), higher degree (MA, PhD, PGCE), NVQ level 4 to 5, HNC, HND, RSA Higher Diploma, BTEC Higher level, professional qualifications (for example, teaching, nursing, accountancy) (7) Other: vocational or work-related qualifications, other qualifications achieved in England or Wales, qualifications achieved outside England or Wales (equivalent not stated or unknown), (8) Does not apply.

Ethnic and cultural identity: ethnic group

Health and care: disability and general health

Disability: People who assessed their day-to-day activities as limited by long-term physical or mental health conditions or illnesses are considered disabled. This definition of a disabled person meets the harmonised standard for measuring disability and is in line with the Equality Act (2010). Classification: (1) Disabled under the Equality Act: Day-to-day activities limited a lot, (2) Disabled under the Equality Act: Day-to-day activities limited a little, (3) Not disabled under the Equality Act: Has long-term physical or mental health condition but day-to-day activities are not limited, (4) Not disabled under the Equality Act: No long-term physical or mental health conditions.

General health: A person's assessment of the general state of their health from very good to very bad. This assessment is not based on a person's health over any specified period of time. Classification: (1) Very good health, (2) Good health, (3) Fair health, (4) Bad health, (5) Very bad health.

Catchment boundaries and OA boundaries were processed using R. For each OA, the proportion of its area overlapping with the WWTP catchment (Overlap%) was calculated. If an OA was fully within a catchment, its entire count was included. For partially overlapping OAs, the count for each variable was scaled by Overlap%, assuming a uniform distribution of characteristics within the OA. The resulting adjusted counts were then aggregated to obtain catchment-level estimates. All fractional counts were rounded to the nearest integer prior to aggregation. This spatial weighting method follows the Simple Approach framework, developed and implemented in R. This spatial weighting method, which assumes a uniform distribution of characteristics within an OA, is a robust and widely-used approach for estimating population characteristics in bespoke geographical areas like WWTP catchments where direct census data is unavailable. Full methodological details are available in our previous work (Kasprzyk-Hordern, Jagadeesan et al. 2025). An overview of the implementation is provided in the Supplementary Information.

All above data were normalised either to population size or total number of households (housing and household composition) before further analysis.

2.4. Calculations and statistical analysis

Population (inhabitants; inh) normalised daily mass loads (*PNDLs*) of *CI*s (mg day⁻¹ 1000inh⁻¹) were calculated using Eq. 1:

$$PNDL = \frac{DL}{PE_{WW}} \times 1000 \quad \text{Eq. 1}$$

where: *DL* is the daily mass load of *CI* (mg day⁻¹) in influent wastewater, *PE_{WW}* is the water utility PE estimate.

Daily mass loads (*DLs*) of *CI*s (mg day⁻¹) using Eq. 2:

$$DL = C \times V \quad \text{Eq. 2}$$

where: *C* is the total concentration of *CI*s (mg L⁻¹) in untreated influent wastewater, and *V* is the volume of wastewater received by the WWTP per day (L day⁻¹).

Statistical analysis was undertaken using Origin(Pro), Version 2021 (OriginLab Corp., Northampton, MA, USA). The dataset contained data points for 11 WWTPs, 9 sampling days, 2 biological replicates per sample giving 198 data points (day equivalents) per each *CI* (72 *WBE_{CI}s*) giving 14256 observations for *WBE_{CI}s* and 830 observations for *ONS_{SEDI}*. Population and wastewater flows normalised daily mass loads (*PNDLs*) of *WBE_{CI}* (mg day⁻¹1000inh⁻¹) were used in all statistical analyses. Arithmetic mean values were used. *ONS_{SEDI}* indicators were also normalised to population size and expressed as individuals per 1000 inhabitants (inh 1000inh⁻¹), with the exception of housing and household composition indicators, which were normalised to the total number of households.

Multivariate statistical analyses were conducted to evaluate associations between exposure variables and the outcome measures. Both Pearson and Spearman's rank correlations were used to explore the characteristics of the variables. Normal distribution for all *CI*s was evaluated using graphical methods (calculated using both skewness and kurtosis, Q-Q (quantile – quantile tests) graphs and histograms) and Kolmogorov-Smirnov normality test to verify if the dataset was normally distributed. Not all *CI*s were normally distributed (cimetidine, caffeine, creatinine, lisinopril, sitagliptin, dihydrocodeine, ibuprofen and its metabolites, antibiotics, HNE-MA and some illicit drugs) and therefore a non-parametric Spearman rank correlation was used. All statistical tests were two-sided and a *p*-value of <0.05 was considered statistically significant.

Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) using Origin(Pro), Version 2021 (OriginLab Corp., Northampton, MA, USA) were also applied to explore patterns of similarity among cities and to identify the principal factors contributing to variation in sociodemographic, health, and wastewater biomarker characteristics. Arithmetic mean *WBE_{CI}* values (for 9 sampling days and 2 biological replicates) were applied after testing both arithmetic and geometric mean (the latter is known to be less affected by extreme values in a skewed distribution). All variables (both *WBE* and *ONS* data) were normalised either to population size or total number of households (housing and household composition) before PCA analysis.

Hierarchical PCA approach was employed to construct Census and *WBE*-derived composite indices. This is to minimise redundancy among Census variables and avoid unequal representation of Census and *WBE* domains in subsequent multivariate analyses. Conceptually related variables were first grouped into thematic domains and used to derive composite indices representing deprivation, health, disability and ethnic diversity, and pharmaceuticals usage. Prior to analysis, all variables were standardised to z-scores to ensure comparability across differing measurement scales. Variables within each domain were examined for suitability using pairwise correlation analysis to ensure the presence of meaningful shared variance while avoiding excessive redundancy. Variables exhibiting very weak correlations (*r* < 0.30) and highly correlated variables (*r* > 0.90) were removed where appropriate to minimise redundancy. Separate PCAs were subsequently performed for each domain using the correlation matrix. The first principal component (PC1) was retained as the composite index when it represented a coherent underlying construct, exhibited consistent and interpretable variable loadings (typically >0.40), and explained a substantial proportion of the total variance (>50%). Where necessary, the direction of the component scores was reversed to ensure consistent interpretation, such that higher values represented greater deprivation, higher educational attainment, or better housing quality. Retaining PC1 provided a summary of the domain while reducing dimensionality and minimising redundancy arising from correlations among individual Census variables. Separate PCA models were then conducted for each domain: Deprivation Index, Education Index, Health index, Disability Index, Age Index, Ethnic diversity (see section 4.3). Pharma Index retained bezafibrate, atorvastatin, atenolol, fexofenadine, acetaminophen, tramadol, codeine, naproxen and diclofenac. Prior to interpretation of the PCA results, potential collinearity among the composite variables was assessed using pairwise correlation analysis and variance inflation factors (VIFs). Moderate correlations are expected and desirable for PCA, whereas very

high correlations ($r > 0.90$) or elevated VIF values (>5) may indicate redundancy among variables. High correlation ($r > 0.9$) was observed for Education Index, Health Index and Age Index and therefore Education and Age Index were removed. The final PCA included the Deprivation Index, Health Index, Disability Index, Pharmaceutical Index, Ethnic Diversity Index, Population, Smoking, Caffeine Consumption and Oxidative Stress biomarkers. Principal components were retained based on eigenvalues, the proportion of variance explained, and examination of scree plots. Variable loading were used to identify the factors contributing most strongly to each component, while component scores were used to characterise similarities and differences among cities. The derived Deprivation Index, Health Index, Disability Index, Pharmaceutical Index, Ethnic Diversity Index, Population, Smoking, Caffeine Consumption and Oxidative Stress biomarkers were subsequently included in the final PCA. This approach reduced dimensionality, mitigated biases associated with unequal representation of Census categories, minimised the influence of compositional constraints, and improved the interpretability

Hierarchical cluster analysis (HCA) was performed to identify groups of cities exhibiting similar sociodemographic, health, and wastewater biomarker characteristics. Prior to analysis, all variables were standardised to z-scores to ensure equal weighting and to minimise the influence of differences in measurement scale. Euclidean distance was used as the measure of dissimilarity between cities, and Ward's linkage method was applied to minimise within-cluster variance and maximise separation between clusters. Clustering was conducted using the same set of variables included in the final PCA, comprising the Deprivation Index, Health Index, Disability Index, Pharmaceutical Index, Ethnic Diversity Index, Population, Smoking, Caffeine Consumption, and Oxidative Stress biomarkers. The resulting dendrogram was examined to identify natural groupings of cities and to assess similarities among locations. HCA was used as a complementary multivariate approach to validate patterns observed in the PCA and to provide an objective classification of cities based on their overall sociodemographic and wastewater biomarker profiles.

Multivariate analysis was performed using Origin(Pro), Version 2021 (OriginLab Corp., Northampton, MA, USA) to test specific exposure-outcome pairs adjusted to confounders. The following exposure-outcome pairs were tested: smoking-pain, smoking-CVD, smoking-diabetes, smoking-asthma and caffeine-oxidative stress, caffeine-anxiety/depression. A list of confounders included: age, deprivation, education, qualifications, disability, general health. Confounders for each pair were selected based on causal diagrams (DAGs) and subject-matter knowledge. For each exposure-outcome pair, a separate multivariable regression model was fitted. The exposure and all predefined confounders were included in every model without data-driven selection. Multicollinearity in a regression model was checked using Variance Inflation Factor (VIF). High collinearity ($VIF > 10$) was observed for the initial confounder list due to overlapping socio-economic/health status. Therefore, 'general health' and 'education' predictors were removed. The following confounders were retained: 'disability', 'deprivation' and 'age' giving $VIF \leq 2.5$.

To account for the large number of statistical tests performed and reduce the likelihood of false-positive findings, p-values obtained from the correlation and multiple linear regression analyses were adjusted for multiple comparisons using the Benjamini-Hochberg false discovery rate (FDR) procedure. The FDR approach controls the expected proportion of incorrectly rejected null hypotheses among statistically significant results while maintaining greater statistical power than more conservative family-wise error rate corrections. For each family of tests, the observed p-values were ranked in ascending order and adjusted according to the Benjamini-Hochberg method. Adjusted p-values (q-values) were subsequently calculated and used to assess statistical significance. Associations with q-values < 0.05 were considered statistically significant after correction for multiple testing.

312

313 **3. Results and Discussion**

This study used a cross-sectional ecological design to explore associations between wastewater-based epidemiology community indicators and city-level socioeconomic characteristics derived from Census data across 10 cities (11 WWTPs). Wastewater measurements were collected during a sampling period of one month (9 sampling days), providing a population-level snapshot of community biomarker loads. These data were linked with aggregated Census indicators describing demographic, socioeconomic, and urban characteristics of each city. Given the cross-sectional ecological design and the limited sample size (10 cities), the statistical analyses were applied primarily as exploratory and hypothesis-generating rather than confirmatory. Descriptive statistics, Spearman rank correlations, principal component analysis (PCA), and hierarchical clustering were therefore used to identify broad patterns and similarities among cities while reducing the dimensionality of highly correlated Census variables. Multiple Linear Regression (MLR) was

324 also used as an exploratory approach to assess associations between WBE_{CIS} and city-level socioeconomic
325 characteristics. Given the cross-sectional ecological study design and the use of aggregated city-level data, the
326 models were not intended to establish causal relationships. Instead, MLR was used to identify population-level
327 associations and generate hypotheses for future epidemiological investigations.

328 **3.1. Inter-WBE_{CI} associations: Smoking-WBE_{CIS} correlate with multiple disease-related WBE_{CIS}**
329 **consistent with the role of smoking as a multi-disease risk factor**

330 Smoking is a risk factor for several diseases, including Type 2 diabetes, asthma/allergies and cardiovascular
331 diseases, as well as being associated with increased pharmaceutical consumption such as painkillers or
332 antibiotics. In this study, smoking WBE_{CIS} correlated with wide-ranging groups of pharmaceuticals (Fig S1).
333 These included: CVDs (e.g. significant correlations of nicotine/cotinine with atenolol, $r_{p/s} = 0.44-0.57$, P-
334 Pearson, S-Spearman), asthma (e.g. positive correlations with antihistamines, $r_{p/s} = 0.29-0.47$), type 2 diabetes
335 (positive correlations with metformin, $r_{p/s} = 0.55/0.47$), pain management (positive correlations with analgesics
336 and NSAIDs, $r_{p/s} \leq 0.71$, opioids: tramadol $r_{p/s} = 0.65/0.65$, codeine, $r_{p/s} = 0.70/0.71$) and infectious disease
337 (sulfapyridine/N-acetylsulfapyridine, sulfamethoxazole, trimethoprim; $r_{p/s} \leq 0.50$). Interestingly no
338 correlations were found between nicotine/cotinine and caffeine/dimethylxanthine despite being reported
339 elsewhere (Shao, Cong et al. 2021). No correlations were also observed with HNE-MA (4-hydroxynonenal-
340 mercapturic acid), an oxidative stress marker. Weak to moderate positive correlations were observed with
341 methotrexate, a pharmaceutical used in cancer and autoimmune conditions ($r_{p/s} = 0.27-0.30$). Detailed
342 discussion is provided in SI and in (Kasprzyk-Hordern et al., 2025).

343

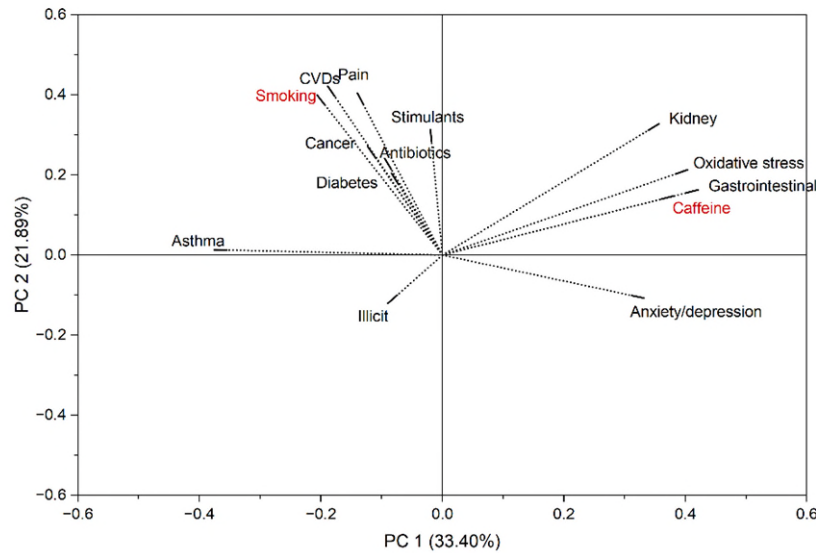


Figure 2. PCA showing associations between smoking-WBE_{CIS} (nicotine/cotinine) and pharma-WBE_{CIS}, proxies for disease. Pearson and Spearman correlation and HCA are available in SI/Tab. S1 and Fig. S1.

PCA was used to visualise key WBE_{CIS} correlations. Two principal components were retained (as seen in Fig. 2) as they were explained by 55% of the variance. PC1 (33.4%) was characterised by positive loadings for ‘caffeine’ (0.381), ‘oxidative stress’ (0.403), ‘gastrointestinal’ (0.421), ‘kidney’ (0.328) and ‘anxiety/depression’ (0.332) variables. Negative loadings were seen for ‘asthma’, ‘diabetes’ (0.376), ‘smoking’ (-0.206), ‘CVDs’ (-0.189), ‘pain’ (-0.140). PC2 (21.9%) was characterised by positive loadings for ‘smoking’ (0.400), ‘CVDs’ (0.422), ‘pain’ (0.405), ‘stimulants’ (0.314), ‘antibiotics’ (0.242), ‘cancer’ (0.270).

PCA confirmed positive correlations between smoking-WBE_{CIS} and pharmaceutical WBE_{CIS} for CVD, pain and infectious disease with weaker correlations observed with class II diabetes, asthma/allergy. PCA also reinforced the lack of correlation (~90° angle) between smoking-WBE_{CIS} and gastrointestinal makers (cimetidine), oxidative stress markers (HNE-MA, caffeine/1,7-dimethylxanthine), and anxiety/depression markers (venlafaxine/desmethylvenlafaxine). As shown in Fig. 2, smoking vectors are located closer to the middle of the plot, indicating that they do not strongly influence PC1 or PC2. This suggests that additional relationships or patterns may exist that are not captured by the leading principal components. To address this, we applied standardization approaches (normalizing values to z-scores), however, these did not improve the PCA outputs. This reinforces the complex relationship between WBE_{CIS} and the need for further work. We also used multivariate HCA for visual representation of key WBE_{CIS} associations between smoking WBE_{CIS} and pharmaceutical proxy WBE_{CIS} (Fig. S1). Consistent with the PCA, two distinct clusters emerged: cluster 1 comprised smoking- and pharmaceutical-WBE_{CIS} proxies for CVD, pain, infectious disease and Type 2 diabetes; cluster 2 comprised gastrointestinal markers (cimetidine), oxidative stress markers (HNE-MA, caffeine/1,7-dimethylxanthine), and anxiety/depression markers (venlafaxine/desmethylvenlafaxine). To explore the above associations further, WBE_{CIS} were compared with ONS_{SEDI}.

3.2. Smoking-WBE_{CIS} and ONS_{SEDI} associations: Smoking-WBE_{CIS} are positively associated with higher deprivation, lower level of education and bad health

Strong correlations between smoking-WBE_{CIS} and ONS_{SEDI} were observed across several indicator categories including deprivation, education, health and housing (Fig. 3).

Smoking and education: Correlations between smoking-WBE_{CIS} and ONS_{SEDI} ‘highest level of education’ clearly showed strong positive associations between cotinine and ‘no qualifications’ ($r_{p/s} = 0.66/0.68$), with negative correlations for ‘level 4 qualifications’ and above ($r_{p/s} = -0.58/-0.61$).

Smoking and demographics: No strong correlations were found between nicotine/cotinine and city size, or gender. However, significant correlations were seen with different age groups, with those aged >50 correlating positively with nicotine/cotinine ($r_{p/s} > 0.50$). Weak correlations were observed between smoking-WBE_{CIS} and ‘ethnic group’ with negative correlations seen between nicotine/cotinine and ‘Asian’, ‘black’, and ‘mixed’ ethnicity (strongest negative correlation seen for cotinine/‘mixed’ $r_{p/s} = -0.44/-0.35$) and positive weak

381 correlations between cotinine and ‘white’ ($r_p = 0.38/0.31$). This is consistent with the conclusions of a recent
382 UK government study (APS, 2024). Regarding ‘household composition’, positive correlations were seen
383 between smoking-WBE_{CIS} and ONS_{SEDI} ‘one person household’ with older demographics ($r_{p/s} \geq 0.46$) and
384 ‘single family household: lone parent family: with dependent children/all households’ ($r_{p/s} \geq 0.32$) versus
385 negative correlations seen for ‘single family household: married or civil partnership couple: with dependent
386 children’ ($r_{p/s} = -0.4$ to -0.6). Regarding ‘accommodation type’, no strong correlations were observed, with
387 semi-detached/terraced houses showing the strongest positive correlations with smoking-WBE_{CIS}.

388 **Smoking and socioeconomic status:** Positive correlations were seen between smoking-WBE_{CIS} and the
389 ONS_{SEDI} ‘deprivation status’ ($r_{p/s} = 0.3$ - 0.6 for cotinine and ‘household deprivation’, one-three dimensions)
390 consistent with the established linkage between smoking prevalence and socio-economic status as captured by
391 the ONS (ONS 2023a). Similar observations were reported by (Choi, Tscharke et al. 2019) who found
392 significant negative correlations between IRSAD (a composite measure of socioeconomic advantage and
393 disadvantage) and nicotine. Smoking-WBE_{CIS} correlated positively with ‘disabled’ parameters (e.g. $r_{p/s}$ for
394 cotinine-‘disabled under the Equality Act’ = $0.47/0.50$). This is consistent with reports that adults with
395 disabilities are more likely to smoke than those not reporting a disability (Emerson 2018).

396 **Smoking and health:** As expected, smoking-WBE_{CIS} showed strong negative correlations with the ONS_{SEDI}
397 ‘very good health’ (cotinine - ‘very good health’ $r_{p/s} = -0.60/-0.59$) progressing to positive correlations from
398 ‘good’, ‘fair’, ‘bad health’ with the strongest positive correlation seen for ‘very bad health’ (cotinine/very bad
399 health’, $r_{p/s} = 0.60/0.71$). This clear gradient from negative to positive correlations across the self-reported
400 health spectrum is consistent with the well-established role of smoking as a major risk factor for multiple
401 diseases.

402

(B) Hierarchical Cluster Analysis

Pearson				Correlation				Spearman			
Nicotine	Cotinine	Caffeine	1,7-dimethylxanthine	SOCIOECONOMIC Indicators				Nicotine	Cotinine	Caffeine	1,7-dimethylxanthine
GENERAL HEALTH [inh/1000inh]											
-0.47	0.66			Very good health				-0.49	0.58	0.2	0.3
	0.33	0.11		Good health					0.40		
0.52	0.59	-0.18	-0.39	Fair health				0.60	0.63	0.21	-0.3
0.57	0.58	-0.14	-0.25	Bad health				0.62	0.66	0.23	-0.27
0.57	0.60	-0.18	-0.24	Very bad health				0.62	0.71	0.21	-0.3
-0.58	-0.58	0.13	0.35	V good to fair health				-0.62	-0.64	-0.2	0.27
0.58	0.59	0.16	0.29	Bad to very bad health				0.62	0.66	0.26	-0.27
ETHNICITY [inh/1000inh]											
	-0.31	-0.41		Asian				-0.35	-0.37	0.21	0.1
	0.22	-0.23		Black				-0.28	-0.30	0.2	0.1
	-0.37	-0.44		Mixed				-0.40	-0.35	0.28	-0.2
	0.05	0.38		White				-0.28	-0.31	0.23	-0.2
	0.2	-0.35		Other				-0.27	-0.33	0.2	0.1
DISABILITY (under the Equality Act) [inh/1000inh]											
0.47	0.47	0.18	0.38	Disabled				0.51	0.50	0.21	-0.3
0.54	0.54	-0.18	-0.25	Disabled: Day-to-day activities limited a lot				0.57	0.55	0.21	-0.3
	0.63	0.66	0.2	Disabled: Day-to-day activities limited a little							
-0.47	-0.46			Not disabled				-0.51	-0.50	-0.2	0.3
				Not disabled: Long term phys./mental health condition					0.52		
-0.41	-0.38			Not disabled: No long term phys./mental health conditions				-0.40	-0.31	0.2	0.1
AGE, SEX, POPULATION											
0.18	0.25			Age: <40 years				-0.12	-0.11	0.19	0.27
	0.25	0.17	0.11	Age: >40 years					0.22	0.17	-0.2
0.25	0.27	0.18	0.17	Sex: Female				0.35	0.35	0.17	-0.2
-0.28	-0.34	-0.18	-0.21	Sex: Male				-0.38	-0.42	-0.23	-0.2
	0.14	0.14	0.14	City size					0.18	0.18	0.18
HOUSEHOLD DEPRIVATION											
-0.42	-0.56			Household is not deprived in any dimension				-0.38	-0.57	0.2	0.2
	0.56			Household is deprived in one dimension					0.49		
0.46	0.56	0.15	-0.29	Household is deprived in two dimensions				0.42	0.55	0.23	-0.3
0.42	0.48	0.13	-0.28	Household is deprived in three dimensions				0.39	0.51	0.23	-0.3
				Household is deprived in four dimensions					0.58	0.58	0.3
0.42	0.48	0.13	-0.28	Household is deprived in at least one dimension				0.3	0.48	0.23	-0.2
AGE [inh/1000inh]											
	0.23	0.28	0.21	Age/Total Age: Aged 4 years and under							
	0.23	0.28	0.21	Age/Total Age: Aged 5 to 9 years					0.23	0.28	0.21
	0.23	0.28	0.21	Age/Total Age: Aged 10 to 14 years					0.23	0.28	0.21
-0.38	-0.54	-0.18	-0.25	Age/Total Age: Aged 15 to 19 years				-0.31	-0.46	-0.2	-0.3
-0.37	-0.56	-0.18	-0.25	Age/Total Age: Aged 20 to 24 years				-0.37	-0.38	-0.2	-0.3
0.37	-0.38	0.12	0.29	Age/Total Age: Aged 25 to 29 years				0.37	-0.36	0.2	-0.3
0.23	0.27	0.17	0.11	Age/Total Age: Aged 30 to 34 years				-0.30	0.45	0.2	-0.3
-0.32	-0.52	-0.18	-0.25	Age/Total Age: Aged 35 to 39 years				-0.38	-0.55	-0.2	-0.3
-0.22	-0.42	-0.11	-0.11	Age/Total Age: Aged 40 to 44 years				-0.31	-0.44	-0.2	-0.3
	0.12	0.12	0.12	Age/Total Age: Aged 45 to 49 years					0.17	0.17	0.17
0.44	0.56			Age/Total Age: Aged 50 to 54 years				0.45	0.56	0.2	-0.3
0.48	0.52	0.18	-0.17	Age/Total Age: Aged 55 to 59 years				0.59	0.54	0.2	-0.3
0.42	0.46			Age/Total Age: Aged 60 to 64 years				0.50	0.44	0.2	-0.3
0.41	0.51			Age/Total Age: Aged 65 to 69 years				0.48	0.52	0.2	-0.3
0.43	0.57			Age/Total Age: Aged 70 to 74 years				0.42	0.59	0.2	-0.3
0.46	0.60			Age/Total Age: Aged 75 to 79 years				0.46	0.64	0.2	-0.3
0.41	0.53			Age/Total Age: Aged 80 to 84 years				0.40	0.54	0.2	-0.3
				Age/Total Age: Aged 85 years and over				0.40	0.64	0.2	-0.3
HIGHEST LEVEL OF QUALIFICATION [inh/1000inh]											
0.46	0.66	0.15	0.35	No qualifications				0.50	0.68	0.26	-0.2
0.38	0.63			Level 1 and entry level qualifications				0.48	0.62	0.2	-0.3
0.38	0.54			Level 2 qualifications				0.44	0.61	0.2	-0.3
0.42	0.58			Apprenticeship				0.40	0.57	0.18	-0.2
	0.24	0.28		Level 3 qualifications							
-0.37	-0.5			Level 4 qualifications and above				-0.44	-0.51	-0.2	-0.3
HOUSEHOLD COMPOSITION [household/all households]											
0.43	0.46			One person household				0.50	0.54	0.2	-0.2
0.43	0.56			One person household: Aged 66 years and over				0.45	0.64	0.2	-0.2
	0.68	0.2		One person household: Other							
0.38	0.36			Single family household				0.35	0.4	0.2	-0.2
-0.2	-0.45			Single family: All aged 66 years and over				0.39	0.41	0.2	-0.2
	0.35	0.2		Married or civil partnership couple				-0.31	-0.48	-0.2	-0.3
-0.40	-0.58			Single family: Married or civil partnership: No children							
	0.34			Single family: Married/civil partner.: Dependent children				-0.35	-0.6	-0.2	-0.3
	0.11	0.17		Single family: Married/civil partner.: All children non-dependent					0.36	0.2	-0.2
	0.34			Single family: Cohabiting couple family							
0.37	0.48			Single family: Cohabiting couple family: No children				-0.18	-0.33	-0.2	-0.3
0.32	0.55			Single family: Cohabiting couple family: With dependent children				0.38	0.46	0.2	-0.2
0.35	0.38	0.24	0.34	Single family: Cohab. couple family: All children non-dependent				0.38	0.54	0.2	-0.2
0.41	0.46	0.23	-0.3	Single family: Lone parent family				0.36	0.8	-0.33	-0.4
	0.16	0.22		Single family: Lone parent family: With dependent children				0.49	0.49	0.23	-0.3
-0.28	-0.23	0.16	0.22	Single family: Lone parent family: All children non-dependent				-0.36	-0.36	0.24	-0.3
-0.43	-0.46			Single family: Other single family household				-0.19	-0.25	-0.2	-0.3
-0.43	-0.46			Other household types				-0.39	-0.37	0.2	-0.2
-0.26	-0.23			Other household types: With dependent children/All households				-0.36	-0.31	0.2	-0.2
-0.44	-0.48			Other household types				-0.41	-0.38	0.2	-0.2
ACCOMMODATION TYPE [accommodation/total accommodation]											
	0.33	0.38		Detached					0.38	0.38	0.38
0.36		-0.28	-0.32	Semi-detached				0.44		-0.28	-0.4
	0.13	0.38	0.38	Terraced					0.35		
-0.21	-0.27			In a purpose-built block of flats or tenement				-0.26	-0.32	-0.2	-0.3
	0.32	0.38		Part of a converted or shared house, including bedsits							
0.29	0.38	0.38	0.36	Part of another converted building, e.g., former school, church				-0.26		0.43	0.5
	0.33	0.38		Commercial building, e.g., office building, hotel, over a shop				0.33			
0.23	0.38	0.38	0.38	A carport or other portable or temporary structure				0.33			

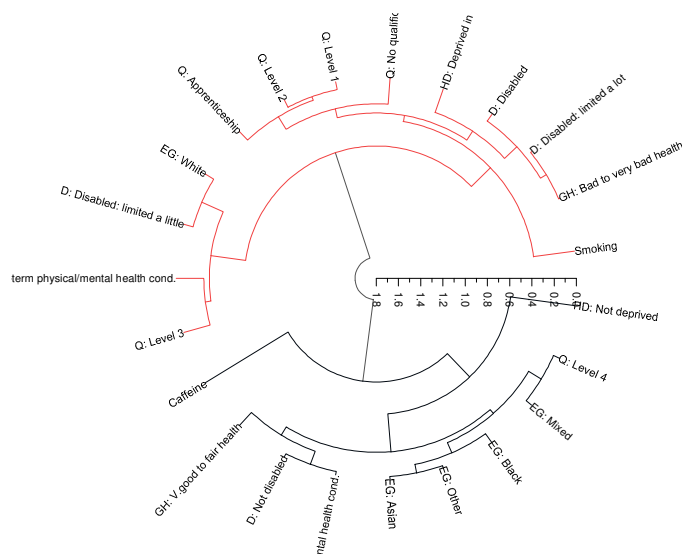


Figure 3. (A) Pearson and Spearman correlation (significant correlations in black font), (B) HCA and (C) PCA showing associations between smoking-WBE_{CIS} (nicotine and cotinine), caffeine-WBE_{CIS} (caffeine and 1,7-DMX) and ONS_{SEDI} (city index description: LD-low deprivation, ≤45%, MD-medium deprivation, 46-54%, HD-high deprivation, ≥55%).

PCA analysis (with two PCs retaining 83% of the variance) confirmed strong positive correlations between smoking-WBE_{CIS} and the ONS_{SEDI}: ‘bad to very bad health’, ‘deprived in more than 1 dimension’, ‘no qualification’ and ‘level 1 qualification’ with the lowest angle seen between nicotine/cotinine and ‘very bad health’ (Fig. 3, Table S2). A strong negative correlation (~180° angle) was seen between nicotine/cotinine and ‘fair to very good health’. As with PCA, HCA revealed strong associations (and two distinct clusters) between smoking-WBE_{CIS} and higher levels of deprivation, lower qualifications, disability and bad health grouped in cluster 1 vs the lowest levels of deprivation, highest levels of qualification, no disability and fair to very good health grouped in cluster 2.

Based on initial results, multivariate analysis was performed to test exposure-outcome pairs: smoking-pain, smoking-CVD, smoking-diabetes, smoking-asthma and smoking antibiotics, adjusted to confounders. Multiple linear regression (MLR) confirmed the models predicting ‘pain’, ‘CVD’ and ‘asthma’ from ‘smoking’ and confounders: ‘disability’, ‘age’, deprivation’ as significant. ‘General health’ and ‘education’ predictors were removed due to high collinearity observed (as discussed in Experimental). MLR predicting ‘pain’ from ‘smoking’, ‘disability’, ‘age’ and ‘deprivation’ was significant, ANOVA: $F(4, 88) = 16.14$, $p < 0.0001$, explaining $R^2 = 0.423$ (adjusted $R^2 = 0.397$). ‘Smoking’ was positively associated with ‘pain’ ($\beta = 49.64$, $SE = 10.16$, $p < 0.001$, 95% CI [29.44, 69.85]). ‘Disability’ and ‘deprivation’ were also associated with ‘pain’ ($\beta = 251.61$, $SE = 114.93$, $p = 0.031$, 95% CI [23.21, 480.02] and $\beta = -258.58$, $SE = 85.69$, $p = 0.003$, 95% CI [-428.87, -88.28], respectively). ‘Age’ was not statistically significant (Tab. S3). Following Benjamini–Hochberg false discovery rate correction for multiple testing, associations with smoking ($q < 0.001$), deprivation ($q = 0.007$), and disability ($q = 0.042$) remained statistically significant.

MLR predicting ‘CVD’ from ‘smoking’, ‘disability’, ‘age’ and ‘deprivation’ was also significant, ANOVA: $F(4, 88) = 11.69$, $p < 0.0001$, explaining $R^2 = 0.347$ (adjusted $R^2 = 0.317$). ‘Smoking’ was positively associated with ‘CVD’ ($\beta = 0.67$, $SE = 0.12$, $p < 0.0001$, 95% CI [0.43, 0.91]) (Tab. S3) and this association remained significant after Benjamini–Hochberg false discovery rate correction ($q < 0.001$). ‘Deprivation’, ‘age’ and ‘disability’ were not statistically significant. MLR predicting ‘asthma’ from ‘smoking’, ‘disability’, ‘age’ and ‘deprivation’ was also significant, ANOVA: $F(4, 88) = 5.86$, $p < 0.0001$, explaining $R^2 = 0.210$ (adjusted $R^2 = 0.175$). ‘Smoking’ was positively associated with ‘asthma’ ($\beta = 0.27$, $SE = 0.06$, $p < 0.0001$, 95% CI [0.15, 0.39]) (Tab. S3). ‘Deprivation’, ‘age’ and ‘disability’ were not statistically significant. Following Benjamini–Hochberg false discovery rate correction, all four associations remained statistically significant ($q < 0.05$). In contrast, associations with deprivation, low educational attainment, and the proportion of the population aged over 40 years were not significant after correction for multiple testing ($q > 0.05$). MLRs predicting ‘antibiotics’, ‘diabetes’, ‘anxiety/depression’ and ‘oxidative stress’ from ‘smoking’, ‘disability’, ‘age’, and ‘deprivation’ were deemed not statistically significant.

The multivariate regression models therefore demonstrate significant associations between the exposure variable ‘smoking’ and the health outcomes ‘pain’, ‘asthma’ and ‘CVDs’, after accounting for potential confounding factors including ‘disability’ and ‘deprivation’. Given the ecological and cross-sectional nature of the data, the models were interpreted as a representation of associations rather than evidence of causality.

444

445 3.3. Caffeine-WBE_{CIS} show weak correlations with other WBE_{CIS} and ONS_{SEDI} indicators

446 We also assessed caffeine as a comparator lifestyle biomarker. Unlike smoking, whose links to health outcomes
447 are well established, caffeine was included to explore the potential of WBE to identify novel associations
448 between population exposures, health outcomes, and socioeconomic factors, therefore supporting hypothesis
449 generation for future epidemiological research.

450 Relatively few statistically significant correlations were seen between caffeine-WBE_{CIS} and pharmaceutical-
451 WBE_{CIS} (Fig. S1). Significant positive correlations were seen between caffeine-WBE_{CIS} and
452 tramadol/metabolites ($r_{p/s} = 0.50 - 0.83$), antibiotics/metabolites (e.g. metronidazole/hydroxymetronidazole, $r_{p/s}$
453 < 0.40), amphetamines (mephedrone ($r_{p/s} = 0.41-0.56$), MDMA ($r_{p/s} = 0.38-0.54$)), ketamine/norketamine ($r_{p/s} =$
454 $0.43-0.59$) and venlafaxine/desmethylvenlafaxine ($r_{p/s} = 0.44-0.64$). However, coffee consumption and health
455 outcomes are still not fully understood. For example, caffeine has been considered a risk factor for increased
456 risk of anxiety and depression (Liu, Wang et al. 2024). Other studies reported beneficial effects of drinking
457 coffee on risk of death from all causes (Gunter, Murphy et al. 2017), as well as reduction of oxidative stress
458 (Vargas-Pozada, Ramos-Tovar et al. 2023). Interestingly, in this study, positive correlations were seen with
459 oxidative stress marker, HNE-MA ($r_{p/s} = 0.41-0.58$), as well as gastrointestinal marker, cimetidine ($r_{p/s} = 0.35-$
460 0.55), suggesting the need for further work. Further discussion can be found in SI.

PCA and HCA analysis (Fig. 2 and Fig. S1) confirmed the statistically significant correlation between caffeine-WBE_{CIS} and gastrointestinal markers (cimetidine), oxidative stress markers (HNE-MA), and a weak association with anxiety/depression markers (venlafaxine/desmethylvenlafaxine). It is worth mentioning that coffee vectors are located closer the middle of the PCA plot, indicating that they do not strongly influence PCA1 or 2, which shows that there might be other relationships or patterns that are not captured by the leading principal components, reinforcing the complex nature of relationships between WBE indicators and the need for further work.

In contrast to smoking-WBE_{CIS}, caffeine-WBE_{CIS} did not show any strong, statistically significant correlations with the tested ONS_{SEDI}, with only moderate positive correlations seen with 'v. good to fair health' (Fig. 3). PCA and HCA analysis (Fig. 3) confirmed associations between caffeine-WBE_{CIS} and the ONS_{SEDI} 'not deprived' and 'fair to very good health', and weaker positive associations with higher levels of education and no disability.

Based on the above observations, multivariate analysis was performed to test exposure-outcome pairs: caffeine-oxidative stress and caffeine-anxiety/depression. MLR predicting 'anxiety/depression' from 'caffeine' and confounders: 'disability', 'age' and 'deprivation' was significant, ANOVA: $F(4, 88) = 15.15$, $p < 0.0001$, explaining $R^2 = 0.408$ (adjusted $R^2 = 0.381$). 'Caffeine' was positively associated with 'anxiety/depression' ($\beta = 0.007$, $SE = 0.0009$, $p < 0.0001$, 95% CI [0.005, 0.009]) (Tab. S3). This association remained significant following Benjamini-Hochberg correction for multiple testing ($q < 0.001$). 'Caffeine' was also positively associated with 'disability' ($\beta = 1.57$, $SE = 0.347$, $p < 0.0001$, 95% CI [0.881, 2.26]) and 'age' ($\beta = 0.76$, $SE = 0.213$, $p < 0.0001$, 95% CI [-1.19, 0.329]), however, it did not remain significant after FDR correction ($q = 0.224$). 'Caffeine's' association with 'deprivation' was not statistically significant.

MLR predicting 'oxidative stress' from 'caffeine' and confounders: 'disability', 'age' and 'deprivation' was significant, ANOVA: $F(4, 88) = 11.70$, $p < 0.0001$, explaining $R^2 = 0.347$ (adjusted $R^2 = 0.318$). 'Caffeine' was positively associated with 'oxidative stress' ($\beta = 0.0014$, $SE = 0.00025$, $p < 0.0001$, 95% CI [0.0019, 0.00049]) (Tab. S3). and this association remained significant following Benjamini-Hochberg false discovery rate correction ($q < 0.001$). Positive association was also observed for 'age' ($\beta = 0.76$, $SE = 0.213$, $p < 0.0001$, 95% CI [-1.19, 0.329]) but these associations did not remain statistically significant after correction for multiple testing ($q > 0.05$). 'Deprivation' and 'disability' were deemed statistically insignificant.

The multivariate regression models therefore demonstrate significant associations between the exposure variable caffeine and the health outcomes 'anxiety/depression' and 'oxidative stress', after accounting for potential confounding factors including 'disability' and 'age'. The observed associations generated the hypothesis that communities with greater caffeine consumption experience a higher burden of anxiety/depression and oxidative stress-related conditions. Future studies incorporating individual-level epidemiological data and longitudinal WBE monitoring are needed to determine whether caffeine consumption represents a causal factor, a behavioural response to underlying health conditions, or a broader marker of lifestyle and socioeconomic influences on population health.

497

498 **3.4. WBE_{CIS} and ONS_{SEDI} associations with lifestyle indicators: smoking and caffeine intake**

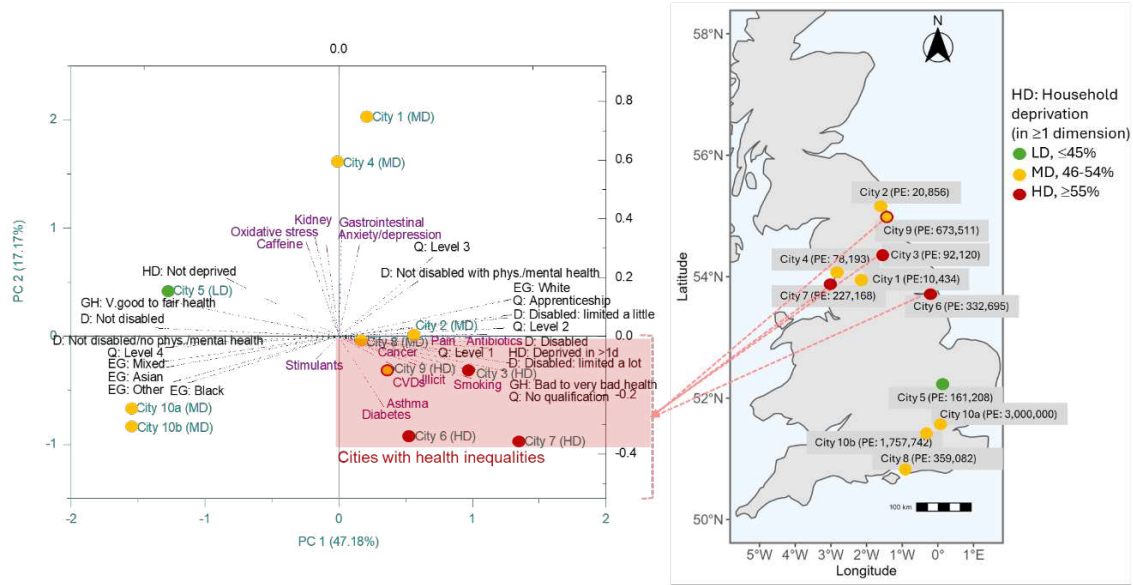
PCA of combined WBE_{CIS} and ONS_{SEDI} is captured in Fig. 4 (Tab. S4) with 65% of the variance explained by the first two components (PC1: 47.0% and PC2: 18.0%) indicating that a low-dimensional representation captures most of the multivariate structure in the data.

PC1 showed positive loadings for several ONS_{SEDI} variables, including 'bad to very bad health' (0.204), 'disabled' (0.217), 'white' (0.202), 'economic status: retired' (0.212), 'apprenticeship/level 2 qualifications' (0.209; 0.200) and WBE_{CI}: 'pain' (0.174), 'antibiotics' (0.161) and 'smoking' (0.160). Negative loadings were associated with: ONS_{SEDI}: 'mixed' ethnicity (-0.214), 'not disabled' (-0.217), 'very good to fair health' (-0.204), 'not deprived' (-0.138), 'higher education (Level 4)' (-0.210) and students ('EA: inactive: student', -0.130). PC2 was characterised by positive loadings for WBE_{CI}: 'caffeine' (0.179), 'oxidative stress biomarkers' (0.161), 'gastrointestinal markers' (0.175), 'anxiety/depression' (0.265), 'kidney' (0.114) and ONS_{SEDI}: 'level 3 education' (0.293), and 'not disabled but with physical/mental health conditions' (0.293), 'not deprived' (0.231) and students (0.244). Negative loadings were observed for ONS_{SEDI} associated with economic inactivity linked with looking after home/family (-0.321), 'black' ethnic group (-0.183), long-term sickness/disability (-0.135), 'no qualifications' (-0.209), and WBE_{CI}: 'diabetes' (-0.210), 'stimulants'

513 (−0.180), ‘CVDs’ (−0.122) and ‘smoking’ (−0.162). PC1 therefore represents socioeconomic, disability and
514 poor health burden. PC2 represents behavioural-psychosocial health characteristics.

515 In summary, PCA showed smoking-WBE_{CIS} clustering together with ‘CVD’, ‘cancer’, ‘illicit drugs’, ‘asthma’,
516 ‘diabetes’ and ‘pain’ WBE_{CIS} as well as with ‘disability’, ‘deprivation’, ‘bad health’ and low/no education
517 ONS_{SEDIS}. Furthermore, these indicators clustered with the cities exhibiting the highest deprivation, lowest
518 education and highest prevalence of weak to bad health (HD Cities 3, 6, 7 and MD City 9: ‘deprived in ≥1
519 dimension’: 58, 56, 59, 52%; ‘bad to very bad health’: 8, 6, 8, 6%; and ‘no/level 1 qualification’: 33, 34, 34,
520 27% for Cities 3, 6, 7, and 9, respectively). Of note, City 9, despite having MD status with household
521 deprivation of 52%, performed worse than other MD cities, in part due to high percentage of ‘bad health’.
522 Interestingly, the least deprived, best educated and healthiest city (LD City 5) clustered with ONS_{SEDIS}: ‘not
523 deprived’ (deprived in ≥1 dimension: 43%), ‘not disabled, ‘v. good to fair health’ (‘bad to very bad health’:
524 3%) as well as with oxidative stress and caffeine-WBE_{CIS}. While WBE correlations do not imply causation,
525 the findings presented here identify significant associations between smoking and several health-related
526 biomarkers, while also indicating that smoking prevalence varies according to socioeconomic and
527 demographic characteristics. These results demonstrate the potential of WBE as an early warning system for
528 monitoring population-level health patterns and associated public health risks. The correlations between the
529 favourable socioeconomic and demographics of City 5 and oxidative stress and caffeine intake represent an
530 exciting opportunity for further exploration of WBE’s potential in public health monitoring.

531 Further exploration of the correlations revealed that pain management and antibiotic usage (infectious disease
532 proxy), high illicit drug use prevalence, and chronic, often lifestyle-driven conditions such as CVDs, diabetes,
533 asthma and cancer clustered closely with disability, deprivation and low levels of education. This was
534 especially notable in HD cities and could serve as an early warning for city-focussed interventions, with HD
535 Cities 3, 6, 7, and 9 and MD Cities 8 and 9 identified as requiring immediate attention. Such data could, for
536 instance, guide local public health bodies in allocating resources for smoking cessation programmes or
537 managing the burden of chronic, lifestyle-driven diseases in the most vulnerable communities.



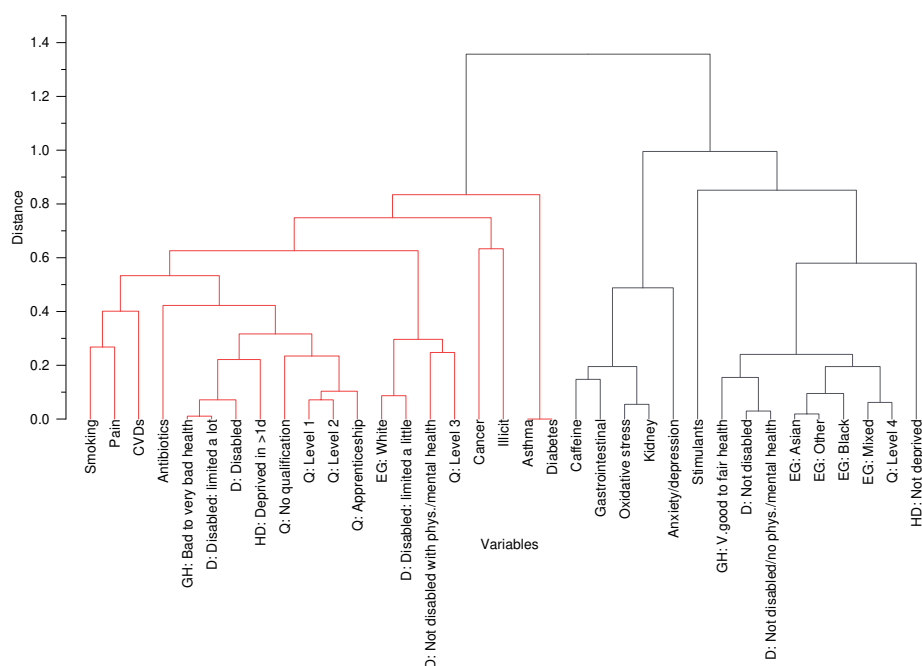


Figure 4. PCA and HCA showing associations between WBE_{CIS} and ONS_{SEDIS}.

Hierarchical PCA was also performed on composite indices (Fig 5 and Table S5): Deprivation Index, Health Index, Disability Index, Pharmaceutical Index, Ethnic Diversity Index, Population, Smoking, Caffeine Consumption and Oxidative Stress biomarkers. This approach reduced dimensionality, mitigated biases associated with unequal representation of Census categories, minimised the influence of compositional constraints, and improved the interpretability. PCA identified two dominant components that together explained 76.3% of the total variance across the study cities (PC1 = 51.6%, PC2 = 24.7%). PC1 was primarily defined by negative loadings for the Health Index (-0.442) and Ethnic Diversity Index (-0.389), and positive loadings for the Disability Index (0.423), Pharmaceutical Index (0.370), Smoking (0.354) and Deprivation Index (0.292). This component therefore represented a gradient from healthier, more ethnically diverse populations with lower disability burden, pharmaceutical usage and smoking prevalence (negative scores) to populations characterised by poorer health outcomes, greater disability burden, higher pharmaceutical consumption, increased smoking prevalence and higher levels of deprivation (positive scores).

A clear separation was observed between City 5 and the highly deprived cities (Cities 3, 6 and 7). City 5 exhibited a strongly negative PC1 score (-2.43), indicating a comparatively favourable health profile characterised by lower disability burden, lower smoking prevalence, reduced pharmaceutical burden and better overall health status. By contrast, City 3 (PC1 = 2.13), City 6 (PC1 = 2.14) and particularly City 7 (PC1 = 2.84) were positioned at the positive end of PC1 and were therefore associated with poorer health outcomes, greater disability burden, higher smoking prevalence, increased pharmaceutical usage and greater socioeconomic deprivation. City 7 represented the most extreme expression of this health-deprivation gradient and consequently exhibited the least favourable overall profile.

PC2 was characterised by strong negative loadings for Oxidative Stress (-0.551) and Caffeine Consumption (-0.511), and positive loadings for Population (0.408), Deprivation Index (0.352) and Ethnic Diversity Index (0.329). This component reflected a gradient contrasting wastewater-derived lifestyle biomarkers with population and deprivation-related characteristics. City 5 exhibited a moderately negative PC2 score (-0.92), suggesting relatively stronger associations with caffeine consumption and oxidative stress biomarkers than cities with positive PC2 scores. This indicates that, despite its favourable health and deprivation profile, City 5 may experience behavioural or lifestyle patterns reflected by elevated caffeine consumption and oxidative stress signatures in wastewater. WWTPs 10a (PC2 = 1.81) and 10b (PC2 = 1.53) representing different geographies of one city of London, combined strongly favourable health profiles (PC1 $\approx$ -3.1) with positive PC2 scores, indicating associations with larger populations, greater ethnic diversity and deprivation-related urban characteristics, while exhibiting lower relative contributions from caffeine and oxidative stress biomarkers. City 9 was positioned towards the positive side of PC1 (1.28) but near the origin on PC2 (0.11), suggesting a tendency towards higher health burden and deprivation, although not to the extent observed for Cities 6 or 7.

Overall, the PCA demonstrates that variation among cities was driven primarily by differences in population health burden, disability, smoking prevalence and pharmaceutical usage (PC1), while a secondary gradient reflected variation in caffeine consumption and oxidative stress biomarkers relative to urban and deprivation-related characteristics (PC2). City 5 emerged as a distinctive low-deprivation city with a favourable health profile but a comparatively stronger association with caffeine consumption and oxidative stress biomarkers, whereas Cities 6 and 7 were characterised by substantially greater health burden and deprivation. City 10 (WWTP 10a and 10b) exhibited the most favourable overall health profiles.

While PCA explained a substantial proportion of the total variance and revealed coherent patterns among cities, the analysis should be regarded as exploratory due to the small number of study sites relative to the number of variables included. The component structure may be sensitive to individual observations, and interpretation is further influenced by the use of composite indices derived from prior dimensionality-reduction procedures. Consequently, the identified gradients should be interpreted with caution and validated using larger datasets and additional study locations.

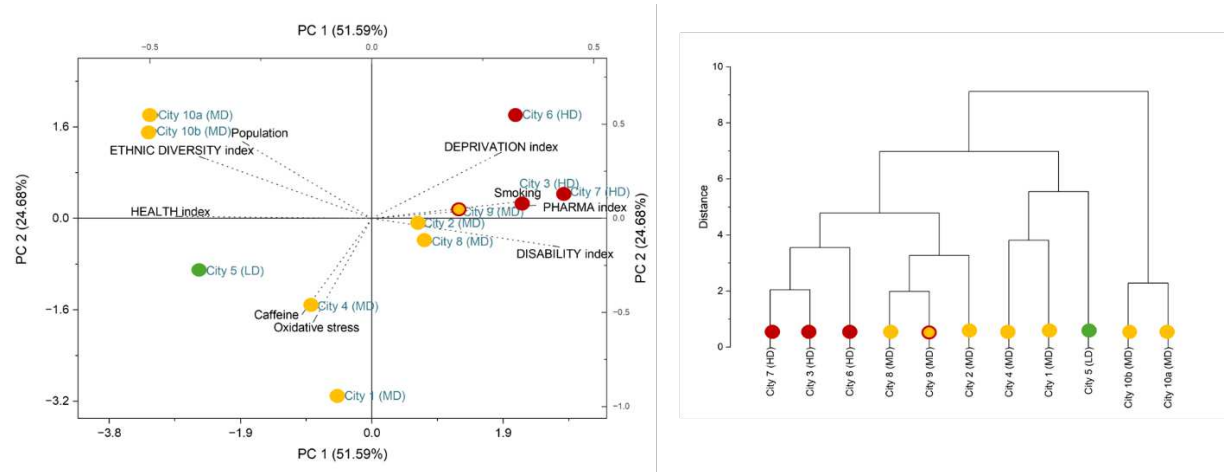


Figure 5. (a) Hierarchical PCA biplot showing relationships between cities and sociodemographic, health and wastewater biomarker indices **(b)** Hierarchical cluster analysis of cities based on standardised sociodemographic, health and wastewater biomarker variables.

Hierarchical cluster analysis (HCA) (Fig. 5) supported the patterns identified by the PCA and revealed distinct groupings of cities based on their combined sociodemographic, health and wastewater biomarker profiles. HD Cities 7, 3, and 6, with highest deprivation index, clustered closely together forming a group characterised by greater health burden, higher disability prevalence, increased smoking and higher pharmaceutical usage. In contrast, MD City 10 and its WWTPs 10a and 10b formed a distinct cluster, reflecting their similar and comparatively favourable health profiles, consistent with their strongly negative PC1 scores in the PCA. MD Cities 8, 9, and 2 clustered together suggesting intermediate characteristics across the health, deprivation and biomarker gradients. MD Cities 1 and 4 formed a separate cluster, reflecting broadly comparable profiles despite the stronger association of City 1 with caffeine and oxidative stress biomarkers. LD City 5 remained separate from the other clusters due to a unique combination of characteristics. Although LD City 5 shared a favourable health profile with City 10, it was distinguished by lower population size, lower deprivation and a relatively stronger association with caffeine consumption and oxidative stress biomarkers. Overall, the HCA identified three broad city groups: a highly deprived and higher health-burden cluster (HD Cities 3, 6 and 7), a healthier urban cluster (City 10: WWTP 10a and 10b), and a group of intermediate cities, while highlighting LD City 5 as a distinct low-deprivation city with a favourable health profile but a unique wastewater biomarker signature. Given the ecological and cross-sectional nature of the data, the model was interpreted as a representation of associations rather than evidence of causality.

4. Limitations

This study highlights the potential of a multi-biomarker approach in WBE for assessing population exposures and health-related outcomes. The findings suggest that the greatest public health value of WBE may be realised when wastewater-derived indicators are integrated with socioeconomic and demographic data, enabling more comprehensive characterisation of population health and supporting the generation of epidemiologically relevant hypotheses. The study has, however, some limitations.

617 WBE considerations:

- 618 1. Normalisation and population factors: To examine inter-WBE_{CI} associations, we applied double
619 normalisation to wastewater flow (affected by weather, season, and water use) and population size (to
620 account for variability across sites). We reduced uncertainty by selecting WWTPs with predominantly
621 domestic discharge. However, because data were derived from heterogeneous pooled community
622 samples, it is not possible to adjust for individual-level demographic, medical, or socioeconomic
623 characteristics, which introduces uncertainty. Nevertheless, wastewater reflects contributions from
624 entire communities, avoiding the group-specific, no response/follow-up biases common in
625 epidemiological cohort studies, aside from inherent intercity differences.
- 626 2. Interpretation of nicotine metabolites: Nicotine metabolites in wastewater may reflect not only tobacco
627 use but also smoking cessation therapies or e-cigarette use. Future studies should therefore include
628 additional tobacco-related biomarkers such as anabasine and anatabine, which are more specific
629 indicators of tobacco combustion.
- 630 3. Enabling full health and lifestyle profiling with WBE: In this work, we measured more than 70
631 WBE_{CIS}, and many more are currently being developed. This expanding suite of biomarkers offers
632 increasing opportunities to characterise community health more comprehensively. Recent
633 advancements have enabled assessments of dietary habits (Elliss, Proctor et al. 2025) and alcohol
634 consumption, further broadening the scope of WBE-based health monitoring.
- 635 4. Spatial granularity: This study operated at the sewerage catchment level, which represents the most
636 cost-effective approach for catchment-level surveillance in the context of early warning. However,
637 sub-catchment monitoring would offer much greater granularity in WBE outputs. Ethical
638 considerations must also be addressed to avoid the stigmatisation of vulnerable communities.
- 639 5. Temporal coverage: This study prioritised spatial coverage across England over temporal sampling.
640 Longitudinal monitoring is still needed to fully characterise seasonal and weather-driven variability
641 (Sims et al. 2023; Ceolotto et al. 2024).

642 Epidemiological concerns:

- 643 1. Correlation versus causation: The analyses presented here are exploratory and identify population-
644 level associations rather than causal relationships. As cities, rather than individuals, were the units of
645 analysis, the findings cannot be used to infer individual behaviours or characteristics, and should be
646 interpreted with caution to avoid ecological fallacy. Socioeconomic indicators are often highly
647 interrelated, making it difficult to disentangle their independent effects, while some associations may
648 reflect broader urban or demographic gradients. Residual confounding is also possible, as important
649 factors such as transient populations and sewer network characteristics were not available. Finally, the
650 relatively small number of cities limited statistical power and multivariable adjustment. Consequently,
651 the findings should be regarded as hypothesis-generating evidence of spatial co-occurrence between
652 wastewater-derived indicators and socioeconomic characteristics, rather than proof of causal
653 relationships.
- 654 2. Limitations Associated with Highly Aggregated Data: The highly aggregated nature of both the
655 wastewater and socioeconomic datasets means that all analyses were conducted at the city level. This
656 may mask within-city variation in behaviours, exposures, health outcomes, and socioeconomic
657 conditions, potentially leading to aggregation bias. The use of city-level averages also limits
658 adjustment for important confounders, such as neighbourhood deprivation, mobility patterns, tourism,
659 and healthcare access. Consequently, the reported associations should be interpreted as broad city-
660 level patterns rather than relationships operating at the individual or neighbourhood level. Future
661 studies using finer spatial scales, such as catchment or neighbourhood-level data, may provide greater
662 insight into the social and environmental drivers of wastewater-derived indicators.

663

664 5. Bringing WBE and epidemiology together

665 1. Integrating Wastewater-Based Epidemiology and Epidemiology

666 As discussed above, the epidemiological interpretation of WBE data is inherently limited by its population-
667 level nature. Associations between WBE_{CIS} and socioeconomic indicators were assessed using aggregated city-
668 level data and should therefore be interpreted as contextual rather than individual-level relationships, with the
669 potential for ecological fallacy remaining an important consideration. In addition, wastewater signals are

670 influenced by multiple social, demographic, environmental, and infrastructural factors that are difficult to fully
671 account for, limiting causal interpretation. These limitations highlight the need for closer integration between
672 WBE and epidemiological methodologies. Future studies would benefit from combining wastewater
673 measurements with independent epidemiological, demographic, clinical, and environmental datasets collected
674 at finer spatial scales. Such approaches could improve interpretation of WBE signals, strengthen causal
675 inference, and help bridge the gap between population-level wastewater observations and the complex
676 determinants of health operating within communities. Ultimately, WBE should be viewed as a complementary
677 epidemiological tool that provides valuable insights into population-level patterns, but whose findings require
678 careful interpretation alongside other sources of public health evidence.

679 2. Wastewater-Based Epidemiology as a Tool for Hypothesis Generation

680 WBE offers a unique opportunity to generate novel epidemiological hypotheses by providing objective,
681 population-level information on community exposures, behaviours, and health-related indicators. Unlike
682 conventional epidemiological approaches that frequently rely on self-reported data, healthcare utilization
683 records, or participant recruitment, WBE captures information from entire populations connected to a
684 wastewater catchment, including individuals who may not engage with healthcare systems or participate in
685 surveys. As a result, WBE can reveal previously unrecognized spatial patterns and associations that may not
686 be detectable through traditional surveillance approaches (Singer 2023). The strength of WBE lies in its ability
687 to identify population-level signals and generate hypotheses regarding the relationships between
688 environmental, social, and health determinants. In the present study, associations between wastewater-derived
689 indicators and socioeconomic characteristics should therefore be interpreted as exploratory observations that
690 highlight potential areas for further investigation. For example, differences in biomarker loads between cities
691 with contrasting socioeconomic profiles may suggest underlying behavioural, environmental, or structural
692 factors that warrant examination through more detailed epidemiological studies. Such associations can help
693 identify research priorities and guide the design of future investigations aimed at understanding the
694 mechanisms responsible for the observed patterns.

695 Importantly, WBE and epidemiology operate at different levels of inference. While epidemiological studies
696 often seek to identify individual-level risk factors and establish causal relationships, WBE inherently measures
697 aggregated community-level signals. Consequently, WBE is particularly well suited to generating population-
698 level hypotheses rather than testing individual-level causal mechanisms. Findings derived from wastewater
699 analyses can subsequently be evaluated using complementary study designs, including cohort studies, case-
700 control studies, surveys, clinical datasets, or spatial epidemiological analyses. In this way, WBE can serve as
701 an early-stage screening tool that highlights potentially important associations before more resource-intensive
702 investigations are undertaken (National Academies of Sciences Engineering and Medicine (U.S.), Committee
703 on Community Wastewater-based Infectious Disease Surveillance 2023).

704 Furthermore, the integration of WBE with conventional epidemiological data provides opportunities for
705 triangulation, whereby evidence from multiple independent sources is used to strengthen interpretation.
706 Associations identified through WBE can be compared with clinical records, prescription data, public health
707 surveillance systems, demographic information, or behavioural surveys to determine whether similar patterns
708 are observed across data sources. This complementary approach can improve understanding of population
709 health trends while reducing reliance on any single surveillance method (Brosky, Prasek et al. 2024).
710 [https://www.who.int/publications/m/item/wastewater-and-environmental-surveillance-for-one-or-more-](https://www.who.int/publications/m/item/wastewater-and-environmental-surveillance-for-one-or-more-pathogens--guidance-on-prioritization--implementation-and-integration)
711 [pathogens--guidance-on-prioritization--implementation-and-integration](https://www.who.int/publications/m/item/wastewater-and-environmental-surveillance-for-one-or-more-pathogens--guidance-on-prioritization--implementation-and-integration)

712 Acknowledgments

713 The work was supported under the C215.3 Wastewater Based Epidemiology Programme within the UK
714 Government Accelerated Capability Environment (ACE), funded by the Environmental Monitoring for Health
715 Protection (EMHP) programme, UK Health Security Agency. We thank Daphne Beniston and Andrew Singer
716 (Accelerated Capability Environment, Homeland Security, UK) for their help organizing the project. We thank
717 Public Health Wales for providing enterovirus primer and probe sequences and the staff at Anglian Water,
718 Northumbrian Water, Southern Water, Thames Water, United Utilities, Yorkshire Water, and Welsh
719 Water/Dŵr Cymru for their support with wastewater sampling. The support of EPSRC (EWS-C19,
720 EP/V028499/1) is greatly appreciated.

721 Credit authorship contribution statement

Barbara Kasprzyk-Hordern: Conceptualisation, Methodology (experimental design, data processing), Formal analysis, Data curation, Writing-original draft, Writing – review editing, Supervision, Project administration, Funding acquisition, Resources. Kishore Jagadeesan: Methodology (socioeconomic and demographic data) Natalie Sims: Methodology (sample preparation, LCMS analysis), Writing – review editing. Davey Jones: Conceptualisation, Writing – review editing. Matthew Wade: Conceptualisation, Writing – review editing.

727

References:

- APS 2024, Annual Population Survey, Cigarette smoking among adults, <https://www.ethnicity-facts-figures.service.gov.uk/health/alcohol-smoking-and-drug-use/adult-smokers/latest/#by-ethnicity-over-time>
- Census 2021 geographies - Office for National Statistics
- Bowes, D. A. (2024). "Towards a Precision Model for Environmental Public Health: Wastewater-based Epidemiology to Assess Population-level Exposures and Related Diseases." *Curr Epidemiol Rep* **11**: 131-139.
- Brosky, H., S. M. Prasek, G. K. Innes, I. L. Pepper, J. Miranda, P. E. Brierley, S. L. Slinski, L. Polashenski, W. Q. Betancourt, K. Gronbach, D. Gomez, R. Neupane, J. Johnson, J. Weiss, H. D. Yaglom, D. M. Engelthaler, C. M. Hepp, K. Crank, D. Gerrity, ... and B. W. Schmitz (2024). "A framework for integrating wastewater-based epidemiology and public health." *Front Public Health* **12**: 1418681.
- Castiglioni, S., I. Senta, A. Borsotti, E. Davoli and E. Zuccato (2014). "A novel approach for monitoring tobacco use in local communities by wastewater analysis." *Tob Control*.
- Chen, J., A. K. Venkatesan and R. U. Halden (2019). "Alcohol and nicotine consumption trends in three US communities determined by wastewater-based epidemiology." *Science of the Total Environment* **656**: 174-183.
- Choi, P. M., B. Tscharke, S. Samanipour, W. D. Hall, C. E. Gartner, J. F. Mueller, K. V. Thomas and J. W. O'Brien (2019). "Social, demographic, and economic correlates of food and chemical consumption measured by wastewater-based epidemiology." *Proceedings of the National Academy of Sciences of the United States of America* **116**(43): 21864-21873.
- Elliss, H., K. Proctor, M. Robertson, J. Bagnall and B. Kasprzyk-Hordern (2025). "A new wide-scope, multi-biomarker wastewater-based epidemiology analytical method to monitor the health and well-being of inhabitants at a metropolitan scale." *ANALYTICAL AND BIOANALYTICAL CHEMISTRY* **417**(26): 5983-6005.
- Emerson, E. (2018). "Smoking among adults with and without disabilities in the UK." *J Public Health (Oxf)* **40**(4): e502-e509.
- González-Mariño, I., J. A. Baz-Lomba, N. A. Alygizakis, M. J. Andrés-Costa, R. Bade, A. Bannwarth, L. P. Barron, F. Been, L. Benaglia, J. D. Berset, L. Bijlsma, I. Bodík, A. Brenner, A. L. Brock, D. A. Burgard, E. Castrignanò, A. Celma, C. E. Christophoridis, A. Covaci, ... and C. Ort (2020). "Spatio-temporal assessment of illicit drug use at large scale: evidence from 7 years of international wastewater monitoring." *Addiction* **115**(1): 109-120.
- Gracia-Lor, E., N. I. Rousis, E. Zuccato and S. Castiglioni (2020). "Monitoring caffeine and nicotine use in a nationwide study in Italy using wastewater-based epidemiology." *Science of the Total Environment* **747**.
- Gunter, M. J., N. Murphy, A. J. Cross, L. Dossus, L. Dartois, G. Fagherazzi, R. Kaaks, T. Kühn, H. Boeing, K. Aleksandrova, A. Tjønneland, A. Olsen, K. Overvad, S. C. Larsen, M. L. Redondo Cornejo, A. Agudo, M. J. Sánchez Pérez, J. M. Altzibar, C. Navarro, ... and E. Riboli (2017). "Coffee Drinking and Mortality in 10 European Countries: A Multinational Cohort Study." *Ann Intern Med* **167**(4): 236-247.
- Holton, E. and B. Kasprzyk-Hordern (2021). "Multiresidue antibiotic-metabolite quantification method using ultra-performance liquid chromatography coupled with tandem mass spectrometry for environmental and public exposure estimation." *Anal Bioanal Chem* **413**(23): 5901-5920.
- Kasprzyk-Hordern, B., K. Jagadeesan, N. Sims, K. Farkas, K. Proctor, J. Bagnall, M. Robertson, D. L. Jones and M. J. Wade (2025). "Multi-biomarker approach for estimating population size in a national-scale wastewater-based epidemiology study." *Water Research* **268**: 122527.
- Kasprzyk-Hordern, B., N. Sims, K. Farkas, K. Jagadeesan, K. Proctor, M. J. Wade and D. L. Jones (2023). "Wastewater-based epidemiology for comprehensive community health diagnostics in a national surveillance study: Mining biochemical markers in wastewater." *J Hazard Mater* **450**: 130989.
- Kasprzyk-Hordern, B., N. Sims, K. Farkas, M. J. Wade and D. L. Jones (2025). "Knowledge discovery by analysis of community wastewater reveals factors driving public health risks." *Water Res* **287**(Pt A): 124370.
- Liu, C., L. Wang, C. Zhang, Z. Hu, J. Tang, J. Xue and W. Lu (2024). "Caffeine intake and anxiety: a meta-analysis." *Front Psychol* **15**: 1270246.

775 Mackie, R. S., B. J. Tschärke, J. W. O'Brien, P. M. Choi, C. E. Gartner, K. V. Thomas and J. F. Mueller (2019).
776 "Trends in nicotine consumption between 2010 and 2017 in an Australian city using the wastewater-based
777 epidemiology approach." Environment International **125**: 184-190.

778 Montes, R., R. Rodil, A. Rico, R. Cela, I. Gonzalez-Marino, F. Hernandez, L. Bijlsma, A. Celma, Y. Pico, V. Andreu,
779 M. L. de Alda, E. Lopez-Garcia, C. Postigo, E. Pocurull, R. M. Marce, M. Rosende, M. Olivares, Y. Valcarcel and
780 J. B. Quintana (2020). "First nation-wide estimation of tobacco consumption in Spain using wastewater-based
781 epidemiology." Science of the Total Environment **741**.

782 National Academies of Sciences Engineering and Medicine (U.S.). Committee on Community Wastewater-
783 based Infectious Disease Surveillance (2023). Wastewater-based disease surveillance for public health action.
784 Consensus study report. Washington, DC, National Academies Press,: 1 online resource (1 PDF file (xvi,152
785 pages)).

786 ONS. (2022). "Census 2021 Bulk - Nomis - Official Census and labour market statistics." Retrieved
787 20/01/2025, 2025, from https://www.nomisweb.co.uk/sources/census_2021_bulk.

788 Rousis, N. I., Z. Li, R. Bade, M. S. McLachlan, J. F. Mueller, J. W. O'Brien, S. Samanipour, B. J. Tschärke, N. S.
789 Thomaidis and K. V. Thomas (2022). "Socioeconomic status and public health in Australia: A wastewater-
790 based study." Environ Int **167**: 107436.

791 Shao, X. T., Z. X. Cong, S. Y. Liu, Z. Wang, X. Y. Zheng and D. G. Wang (2021). "Spatial analysis of metformin
792 use compared with nicotine and caffeine consumption through wastewater-based epidemiology in China."
793 Ecotoxicology and Environmental Safety **208**.

794 Shao, X. T., P. Y. Zhang, S. Y. Liu, J. G. Lin, D. Q. Tan and D. G. Wang (2022). "Assessment of correlations
795 between sildenafil use and comorbidities and lifestyle factors using wastewater-based epidemiology." Water
796 Res **218**: 118446.

797 Sims, N. and B. Kasprzyk-Hordern (2020). "Future perspectives of wastewater-based epidemiology:
798 Monitoring infectious disease spread and resistance to the community level." Environ Int **139**: 105689.

799 Sims, N. and B. Kasprzyk-Hordern (2020). "Future perspectives of wastewater-based epidemiology:
800 Monitoring infectious disease spread and resistance to the community level." Environment International **139**.

801 Singer, A. (2023). "A world of wastewater-based epidemiology."

802 Thanh, B., G. Vu, T. Hue, Q. Zheng, G. Chan, N. Anh and P. Thai (2022). "Assessing changes in nicotine
803 consumption over two years in a population of Hanoi by wastewater analysis with benchmarking
804 biomarkers." SCIENCE OF THE TOTAL ENVIRONMENT **846**.

805 Vargas-Pozada, E. E., E. Ramos-Tovar, C. Acero-Hernández, I. Cardoso-Lezama, S. Galindo-Gómez, V. Tsutsumi
806 and P. Muriel (2023). "The antioxidant and anti-inflammatory activities of caffeine effectively attenuate
807 nonalcoholic steatohepatitis and thioacetamide-induced hepatic injury in male rats." Can J Physiol Pharmacol
808 **101**(3): 147-159.

809 Wade, M. J., A. Lo Jacomo, E. Armenise, M. R. Brown, J. T. Bunce, G. J. Cameron, Z. Fang, K. Farkas, D. F. Gilpin,
810 D. W. Graham, J. M. S. Grimsley, A. Hart, T. Hoffmann, K. J. Jackson, D. L. Jones, C. J. Lilley, J. W. McGrath, J.
811 M. McKinley, C. McSparron, ... and B. Kasprzyk-Hordern (2022). "Understanding and managing uncertainty
812 and variability for wastewater monitoring beyond the pandemic: Lessons learned from the United Kingdom
813 national COVID-19 surveillance programmes." Journal of Hazardous Materials **424**.

814 Wang, Z., Q. Zheng, J. 'Brien, B. Tschärke, G. Chan, K. Thomas, J. Müller and P. Thai (2024). "Analysis of
815 wastewater from 2013 to 2021 detected a recent increase in nicotine use in Queensland, Australia." WATER
816 RESEARCH **250**.

817 Yang, D., Q. D. Zheng, P. K. Thai, F. Ahmed, J. W. O'Brien, J. F. Mueller, K. V. Thomas and B. Tschärke (2022).
818 "A nationwide wastewater-based assessment of metformin consumption across Australia." Environment
819 International **165**.

820 Zhang, L., P. Du, Q. Zheng, M. Zhao, R. Zhang, Z. Wang, Z. Xu, X. Li and P. K. Thai (2025). "Exposure to smoking
821 and greenspace are associated with allergy medicine use - A study of wastewater in 28 cities of China."
822 Environ Int **196**: 109291.

823 Zheng, Q. D., C. Gartner, B. J. Tschärke, J. W. O'Brien, J. F. Gao, F. Ahmed, K. V. Thomas, J. F. Mueller and P.
824 K. Thai (2020). "Long-term trends in tobacco use assessed by wastewater-based epidemiology and its
825 relationship with consumption of nicotine containing products." Environment International **145**.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ACESI20260929.docx](#)