MEDSAT: A Public Health Dataset for England Featuring Medical Prescriptions and Satellite Imagery

Sanja Šćepanović[1,2†*] **Ivica Obadic**[2,3*] **Sagar Joglekar**[1,6‡] **Laura Giustarini**[5]
Cristiano Nattero[5] **Daniele Quercia**[1,4] **Xiao Xiang Zhu**[2,3]
[1]Nokia Bell Labs [2]Technical University of Munich [3]Munich Center for Machine Learning
[4]Centre for Urban Science and Progress, King's College London [5]WASDI [6]Intercom

Abstract

As extreme weather events become more frequent, understanding their impact on human health becomes increasingly crucial. However, the utilization of Earth Observation to effectively analyze the environmental context in relation to health remains limited. This limitation is primarily due to the lack of fine-grained spatial and temporal data in public and population health studies, hindering a comprehensive understanding of health outcomes. Additionally, obtaining appropriate environmental indices across different geographical levels and timeframes poses a challenge. For the years 2019 (pre-COVID) and 2020 (COVID), we collected spatio-temporal indicators for all Lower Layer Super Output Areas in England. These indicators included: i) 111 sociodemographic features linked to health in existing literature, ii) 43 environmental point features (e.g., greenery and air pollution levels), iii) 4 seasonal composite satellite images each with 11 bands, and iv) prescription prevalence associated with five medical conditions (depression, anxiety, diabetes, hypertension, and asthma), opioids and total prescriptions. We combined these indicators into a single MEDSAT dataset, the availability of which presents an opportunity for the machine learning community to develop new techniques specific to public health. These techniques would address challenges such as handling large and complex data volumes, performing effective feature engineering on environmental and sociodemographic factors, capturing spatial and temporal dependencies in the models, addressing imbalanced data distributions, developing novel computer vision methods for health modeling based on satellite imagery, ensuring model explainability, and achieving generalization beyond the specific geographical region.

1 Introduction

The impact of environmental factors on human health has gained significant attention in recent years, particularly in the face of increasing environmental pressures caused by extreme weather events. Understanding these impacts is crucial for effective public and population health interventions. However, existing studies often face challenges in obtaining appropriate health and environmental data.

Fine-grained and comprehensive data on the prevalence of medical conditions is often scarce in public health studies. Traditional approaches rely on infrequent surveys or cohort studies, such as NHANES [23], HSE [46], BRFSS [52], The Swiss National Cohort [60], or the UK Biobank [61]. However,

[*]Shared first authorship.
[†]The work was done during the author's AI4EO Beyond Fellowship at the Technical University of Munich.
[‡]The work was done prior to the author's tenure at Intercom.

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

https://doi.org/10.52202/075280-3400

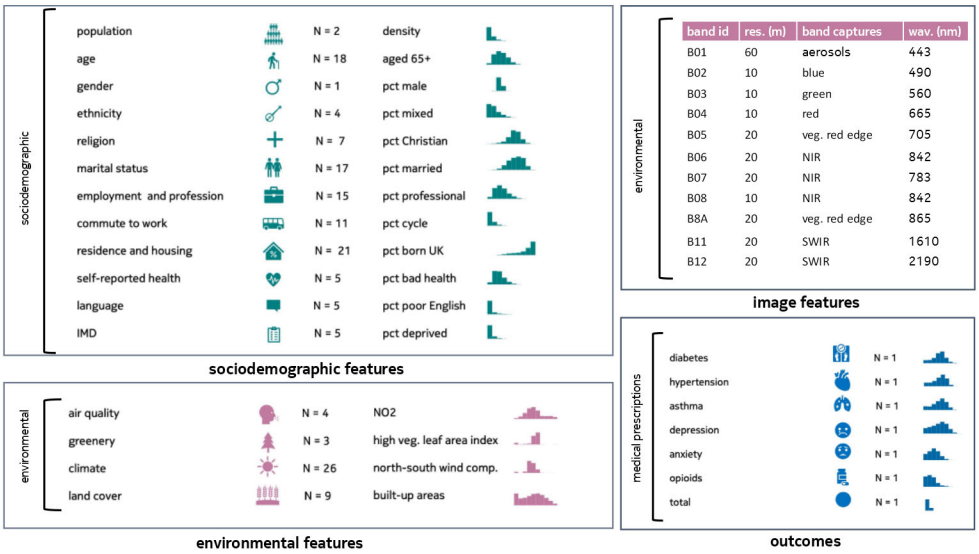

Figure 1: **Structure of MEDSAT dataset (single year):** This figure illustrates the four data components constituting MEDSAT: i) *sociodemographic* features (111), ii) *environmental* point features (43), iii) *image* features (4 seasonal Sentinel-2 composite tiles $\times$ 11 bands each), and iv) *prescription* outcomes (prevalence scores for 7 medical prescription types). The distributions of example variables are shown. Sociodemographic variables are presented as percentages ranging from 0 to 1, while environmental variables have varying ranges (e.g., NO2: (0-3.1 $\times$ 10^{-3}) mol/m^2, $\mu = 2.43 \times 10^{-6}$, $\sigma = 2.0 \times 10^{-4}$). Outcome variables represent yearly prescription quantities per capita and are mostly normally distributed (except opioids and total prescriptions). For instance, diabetes prescriptions range from 0.02 to 104.84 ($\mu = 38.76$, $\sigma = 16.35$). Each Sentinel-2 composite image consists of 11 spectral bands. MEDSAT offers two such yearly snapshots, for 2019 and 2020. A comprehensive description can be found in the Appendix.

survey methods suffer from biases related to sampling, non-response, recall, and question wording. Cohort studies, while aiming to mitigate these biases, are limited in size, expensive, time-consuming, and prone to participant dropout. The All of Us Research Program [58] is an ambitious initiative recruiting over 1 million participants, but its representativeness and long-term engagement remain to be seen. In summary, existing health outcomes data are often limited in scope, granularity, and subject to various biases. Furthermore, despite the increasing availability of finer-resolution measurements for crucial environmental indicators relevant to public and population health, such as greenery, sun radiation, and air pollution, challenges persist in obtaining comprehensive and suitable indices that cover diverse geographical levels and timeframes. While the Earth Observation (EO) community has made significant efforts in capturing detailed satellite imagery with improved resolutions, frequencies, and accessibility, there remains a gap in transforming this vast amount of data into user-friendly indices that can be effectively utilized by non-technical stakeholders and the wider community unfamiliar with EO methods. Even when institutions provide data, such as air quality data from DEFRA [20] in the UK, individuals interested in compiling various environmental information often need to collect it from multiple sources, and there are spatial and temporal limitations to the available data.

In this paper, we present the MEDSAT dataset (Figure 1) consisting of four complementary components and covering two years (2019 and 2020), specifically designed for studying the effects of the environment on population health in small administrative areas in England. Our approach involves an open-source framework that utilizes National Health Services (NHS) practice-level prescription data to extract medical prescription prevalences at fine spatial (i.e., Lower Layer Super Output Area (LSOA)) and temporal (seasonal) scales for the entire population of 57 million. We derived environmental indicators from satellite products, such as Sentinel-5 and OMI, using Google Earth Engine [30]. Additionally, we created cloud-corrected seasonal composite images for 11 spectral bands of multispectral instrument (MSI) Sentinel-2 images in the WASDI platform [67], covering the whole of England, i.e., 130,279 km^2. Our dataset also includes 111 socioeconomic indicators

https://doi.org/10.52202/075280-3400

Table 1: A comparison of MEDSAT with similar datasets, some of which are not publicly available. NHS corresponds to the UK health care system and DHS to the Dutch healthcare system.

dataset	health indicator(s)	indicator source	imagery	env.	soc.	spatial unit	public
SustainBench [69]	BMI, child mortality, water quality, sanitation	surveys	Landsat street view	✗	✗	village (59km2)	✓
Landscape Aesthetics [39]	environment scenicness	crowdsourcing	Sentinel-2	✓	✗	1.6km2	✓
COVID-19 [62]	COVID-19 cases and deaths	WHO	✗	✓	✓	city	✗
Greenery & mortality [7]	mortality	Eurostat	✗	✓	✓	city	✗
Greenery & prescribing [31]	antidepressants **prescriptions**	DHS	✗	✓	✓	municipality (up to 506km2)	✗
Nat. env. & prescribing [27]	mortality, **prescriptions**: cardiovascular antidepressants	NHS	✗	✓	✓	LSOA (up to 18km2)	✗
MEDSAT	**prescriptions**: respiratory (asthma) metabolic (diabetes, hypertension) mental (depression, anxiety), opioids, & total	NHS	Sentinel-2	✓	✓	LSOA (up to 18 km2)	✓

obtained mostly from the UK census. By integrating these diverse datasets for the years 2019 and 2020, we provide researchers with a comprehensive resource for studying spatial and temporal health attributes and identifying regional health disparities.

2 Related Work

Environmental conditions such as air or noise pollution are relevant indicators for various health issues like asthma or heart diseases [22]. However, the benefits of using EO data to monitor the impact of environmental conditions on human health are still limited. This limitation stems from the existing datasets in the literature listed in Table 1 that either focus on a narrow set of conditions derived from surveys that might not be representative of the entire population or do not provide detailed medical prevalence data on a fine-grained spatial level. Often these datasets are not publicly available and fail to include environmental and sociodemographic features relevant to health studies. For example, the SustainBench dataset [69] presents the problem of predicting 4 different health indicators derived from surveys based on Landsat satellite imagery and street-view images. The spatial unit of this study corresponds to a village or a local community covering an area of $\approx$58 km^2 and this dataset does not contain additional environmental and sociodemographic features. By considering a more abstract health indicator, Levering et al. [39] introduced a dataset that associates Sentinel-2 images with crowdsourced data for landscape scenicness used as a proxy for human health and well-being. By analyzing fine-grained geographical regions of 1.6 km^2, the authors discovered plausible associations between a landscape's beauty and its land cover distribution. Targeting a specific condition across the population, Temenos et al. [62] assembles a dataset that relates COVID-19 cases with point features describing environmental data for urban greenness, air quality, sociodemographic features, and health factors. The authors reveal that temperature and mobility trends are among the most important features for predicting COVID-19 cases. Yet, this dataset is not publicly available, does not contain any imagery and the point features represent entire cities, thus describing very coarse spatial units. Further, the impact of the natural environment on mortality rates and prescriptions for various conditions is investigated in [7, 31, 27]. However, the datasets used in these studies are also not publicly available, include only environmental variables related to greenery, and do not contain any imagery. Moreover, the analyses in [7, 31] are performed on larger spatial units like municipalities and cities.

In comparison to these works, our dataset, MEDSAT, enables comprehensive modeling of the population health on a very-fine-grained spatial level as it jointly offers environmental and sociodemographic features as well as EO imagery at LSOA level. Further, it covers prescriptions associated with 5 medical conditions, as well as opioids and total, which allow a detailed understanding of specific condition-related factors, and shed light on the overall population health and well-being.

3 The MEDSAT Dataset

The MEDSAT dataset serves as a comprehensive resource for public and population health studies, encompassing medical prescription quantity per capita as outcomes and a wide array of sociodemographic, environmental and image features across 33K LSOAs in England (Figure 1). In this release,

https://doi.org/10.52202/075280-3400

we provide data snapshots for the years 2019 (pre-COVID) and 2020 (COVID). Sociodemographic variables align with the latest UK census from 2021. Figure 2 visualizes examples of variables present in MEDSAT.

Access the code at `https://github.com/sanja7s/MedSat`,
and the dataset at `https://doi.org/10.14459/2023mp1714817`.
The dataset is released under the CC BY-SA 4.0 license.

3.1 Sociodemographic Features

Our dataset comprises sociodemographic variables sourced from the latest UK Census in 2021 [1]. These variables encompass essential indicators employed in public and population health research, such as gender (percentage of males), age distribution (percentage within 5-year age groups up to 85 and above), deprivation scores (percentage of deprived households in 1-4 dimensions), self-reported health (percentage reporting health on a five-point scale), ethnicity (percentage of individuals with White, Asian, Black, or Mixed backgrounds), and English proficiency (percentage reporting English as their main language). Additional variables indirectly related to health outcomes were incorporated, covering religion, commute means and distance to work, residence and housing, profession, and marital status. Our sociodemographic data do not contain any personally identifiable information because census implements stringent privacy protection measures, including targeted record swapping and cell key perturbation, to ensure confidentiality without compromising aggregated statistics [1].

3.2 Environmental Features

We obtained environmental point features for the MEDSAT dataset using various satellite data products on Google Earth Engine (GEE) [30]. For *air quality*, we used satellite data products such as Sentinel-5P NRTI to derive nitrogen dioxide (NO2) [32], TOMS&OMI for ozone [4], and CAMS for total aerosols and PM2.5 [25]. *Greenery* variables were derived from Sentinel-2 MSI for NDVI and ERA5-ECMWF product for high/low vegetation greenery indices. *Climate* variables, including wind components, air temperature, soil temperature, atmospheric pressure, and incoming solar radiation, were obtained from ERA5-ECMWF. All *land cover* variables were sourced from Google Dynamics World product [11]. Importantly, our open-source code can be easily adapted to extract other similar indices, and at different *spatial* scales (e.g., wards or other countries) and *temporal* scales (e.g., monthly or different years). More information can be found in Appendix section D.2.

3.3 Image Features

In addition to the above-described approach for deriving preprocessed environmental point features, in our dataset we also included the spectral bands provided by the Sentinel-2 mission [3], thus resulting in a more comprehensive set of environmental image features. Concretely, we processed Sentinel-2 images for the years 2019 and 2020 through the WASDI [67] platform by calculating average values for each of the four meteorological seasons [49]. We focused on 11 specific bands (Figure 1) capturing relevant environmental factors, excluding the band B10 that is primarily used for cloud detection and water vapor mapping. To ensure data consistency and quality, we resampled the bands to a uniform resolution of 10 m, applied cloud masks to exclude affected pixels, and computed pixel-per-pixel averages over time under cloud-free conditions. The resulting composite images represent the typical environmental characteristics for each season (over 500 GB per year in total).

To explore the potential of the high-resolution Sentinel-2 imagery for modeling population health on a fine-grained spatiotemporal level, we extracted image features per LSOA for each meteorological season with the procedure depicted in Figure 4 in Appendix. Concretely, we used the shapefiles that describe LSOA's geographical location to crop the LSOA pixels from a seasonal Sentinel-2 image. Next, we extracted basic image features from the cropped LSOA pixels by performing the following 5 aggregations per Sentinel-2 band: *mean, stdev, min, max*, and *median*, thus resulting in 55 image features per LSOA per season.

3.4 Prescription Outcomes

Extracting Prescriptions on an LSOA Level from NHS data. The monthly practice-level prescribing data in England, provided by the National Health Services (NHS) since July 2010 [47],

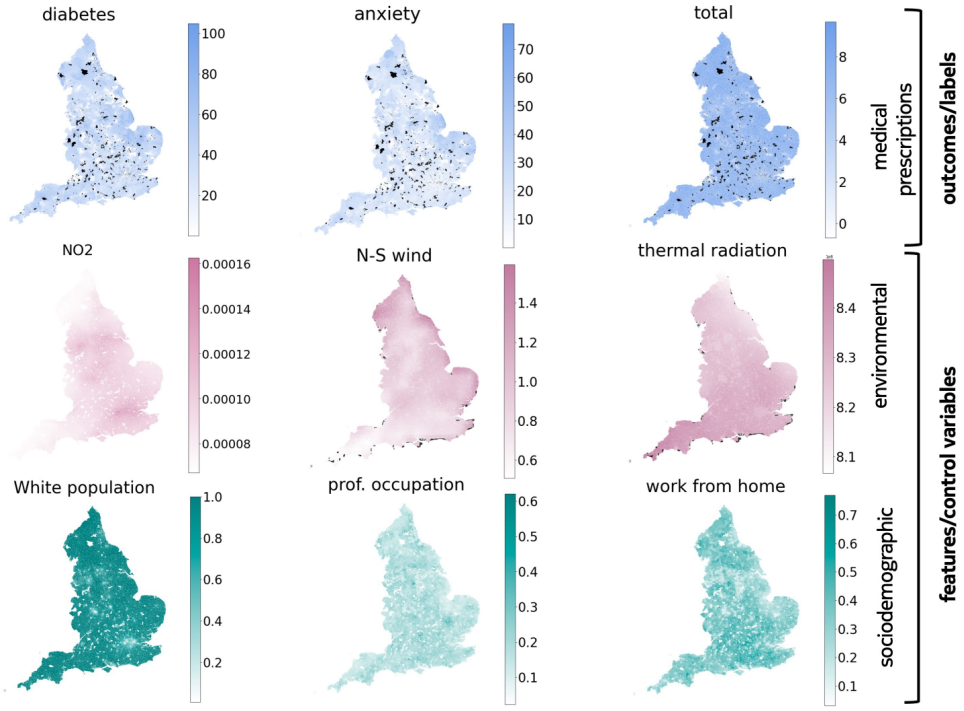

Figure 2: **Visualization of the MEDSAT point features.** The example distributions for the year 2020 of three health outcomes (diabetes, anxiety, and total prescriptions), three environmental variables (NO2, north-south wind component, and incoming thermal radiation), and three sociodemographic variables (percentage of White population, professional occupation, and work-from-home). The missing values are highlighted in black. Depending on the specific analyses intended, the missing value rate will be constrained by the outcomes, standing at 5.7%, and ranging up to 15.2% if all the features are to be used. Notably, we possess sociodemographic and image features data for all LSOAs. Please refer to Appendix for the details.

constitutes the foundation of our analysis. It offers anonymized information about monthly prescriptions across General Practitioner (GP) practices and patient membership to GP practices on an LSOA level. We parse this data to extract the prescribed drugs and summarize the total number of patients per LSOA and compute the fraction of a GP practice's patients associated with a specific LSOA. Further details about the NHS prescriptions data and the applied procedure for calculating the number of patients per LSOA can be found in the Appendix, Section D.4.

Associating Prescriptions with a Condition. To determine prescriptions related to specific medical conditions, our framework utilizes curated lists of drugs, such as the one collated for opioids by previous works [59, 18], or it leverages DrugBank [37] to automatically identify drugs associated with a given condition (see the Appendix DrugBank section D.4.3). DrugBank is an online database that provides comprehensive information on active pharmacological ingredients (APIs) and their corresponding conditions. Each drug name is associated with one or more conditions (i.e., symptoms and diseases), drug categories, and an Anatomical Therapeutic Chemical code assigned by the World Health Organization (WHO) for unique identification purposes. For instance, the drug name Citalopram (`https://go.drugbank.com/drugs/DB00215`) is linked to a range of diseases, including Depression, Anorexia Nervosa, Generalized Anxiety Disorder, and Post Traumatic Stress Disorder. During our crawl, we obtained data on 9,105 drug names from the website, and by filtering out drug names that were not linked to any drug categories, symptoms, or conditions we were left with 3,013 drug names. Next, we generated a curated list of drugs associated with a specific condition by selecting drugs from DrugBank that were linked to that condition (e.g., we associated Citalopram with depression, anxiety, and the other conditions mentioned above).

https://doi.org/10.52202/075280-3400

Estimating the Prescription Prevalence. To estimate the number of prescriptions associated with a specific condition c, we first matched the condition-specific drugs from DrugBank with their British National Formulary (BNF) codes from the NHS prescribing dataset. Next, the number of prescriptions for a specific condition c in area a is computed using the following formula:

$$N_c(a) = \sum_{GP \in a} N_c(GP) \cdot f(GP, a), \tag{1}$$

Here, $N_c(GP)$ represents the total number of prescriptions per GP for drugs associated with the curated list for the condition, and $f(GP, a)$ denotes the fraction of patients of the GP who reside in the area a. To ensure comparability across areas with varying population densities, we computed the metric of "prescriptions quantity per capita," commonly used in medical studies [19, 18], as follows:

$$\tilde{N_c}(a) = \frac{N_c(a)}{n_{pat}(a)}, \tag{2}$$

where $n_{pat}(a)$ corresponds to the total number of patients residing in the area a.

For MEDSAT we calculated medical prescriptions associated with three classes of conditions: i) *metabolic* (diabetes and hypertension), ii) *mental* (depression, and anxiety), and iii) *respiratory* (asthma); as well as *opioids* prescriptions, (which are predominantly prescribed for pain management, but they do have other applications, and have been associated with a consumption crisis in the UK [53, 56]), and *total prescriptions*, as a proxy for general health and well-being. We highlight that, for simplicity, we use condition names to refer to related prescriptions, however we cannot know for each individual prescription what was the exact cause for which it was prescribed. E.g., "depression prescriptions" means *antidepressants* and "anxiety prescriptions" means *anxiolytics*, regardless of actual use. Co-prescriptions across conditions may arise from this method, as it does not ascertain specific prescription reasons, as such details are absent in the NHS dataset. However, the multitude of studies examining prescriptions [59, 18, 8, 41, 34, 33, 31, 44, 35, 63, 12, 66, 29], akin to our approach, attests to its significance as a public health outcome.

Our prescription dataset does not contain any personally identifiable information, as it is derived from publicly available monthly prescription data provided at the level of practices, each serving numerous patients.

4 Results

4.1 Revealing Health Inequalities

In Figure 8 in Appendix, we present the healthcare accessibility disparities across regions. First, interestingly, we find a prevailing pattern where the number of registered patients exceeds the census population in most areas of the country. This aligns with previous investigations by UK authorities [64]. Second, although the correlation ($r = .87$, $p \approx 0$) is strong, certain LSOAs exhibit disproportionate patient-to-population ratios.

Additionally, our analysis highlights broader factors contributing to healthcare inequalities. The residual values, representing deviations from the linear fit of the patient to the population numbers, correlate with deprivation levels ($r = .16$, $p \approx 0$ for mid-deprived areas; $r = .22$, $p \approx 0$ for highly-deprived areas), suggesting a greater burden on healthcare access in socioeconomically disadvantaged regions. Moreover, the residual values exhibit a negative correlation ($r = -.40$, $p \approx 0$) with the percentage of White population, indicating disparities associated with ethnic backgrounds. For more details on this analysis, please refer to Section E.1 in Appendix.

4.2 Predicting Prescriptions

To evaluate the plausibility of our dataset for modeling population health, we applied a classical geostatistical method called Spatial Lag Model (SLM) [5] as well as trained the LightGBM machine learning model [36] to predict the medical prescriptions based on the point features, including image-derived ones, in our dataset. Both models were applied separately for every condition and a combination of the environmental, sociodemographic and image features, to better understand the contribution of different input features in modeling population health.

Table 6 in Appendix displays the SLM results. Collectively, the input features account for a variance ranging from 38% (for diabetes and total prescriptions) up to 63% (for opioids). Individually, sociodemographic features lead the way, explaining between 31% (total) and 52% (opioids) of the variance. They are followed by environmental point features, which account for variances from 13% (diabetes) to 49% (opioids). Image features, though least impactful, still cover a variance from 4% (diabetes) to 28% (opioids).

While the models such as SLM account for the well-known spatial autocorrelation effects [5] present in spatial analysis and modelling (referring to the process that creates clusters of values), machine learning models, such as LightGBM, require a special type of cross-validation that is adapted to account for these effects [55]. A block-buffered cross-validation is a common approach [38], and it is implemented in an `R` package called `blockCV` [65]. We employed this package to calculate spatial folds on the input of our LSOA shapefiles. For the details, please refer to Appendix E.2. The results obtained through spatial cross-validation using LightGBM are detailed in Table 2. In the initial row, it is evident that even fundamental image features exhibit predictive capability, albeit to a limited extent. Conversely, the subsequent two rows highlight that environmental and sociodemographic features offer improved explanatory power for the observed variances, outperforming image-based features. Furthermore, these feature categories display varying degrees of importance across different conditions. Environmental attributes notably enhance the predictive accuracy of depression, opioid prescriptions, asthma, and total prescriptions. Conversely, sociodemographic features prove more effective in accurately forecasting prescriptions for other medical conditions. Notably, the integration of both environmental and sociodemographic characteristics becomes pivotal for a holistic model of population health. This is exemplified by the last row, indicating that using both, the environmental and the sociodemographic features results in the best R^2 scores for all conditions under consideration. Moreover, for both the SLM and the LightGBM model, we observe a consistent pattern that the prescriptions for the mental conditions are predicted with higher accuracy than the ones for the other conditions. In Appendix Section E.2, we present a detailed overview of this experiment's setup and provide a comparison of the LightGBM model with a Feed-Forward Neural Network (FNN), which shows that the LightGBM model consistently outperforms the FNN.

Table 2: **The average R^2 scores resulting from the 5-fold spatial cross-validation of LightGBM.** These scores are computed across various prescription types and combinations of dataset features specifically for the year 2020.

	metabolic		mental		respiratory		
input	**diabetes**	**hypertension**	**depression**	**anxiety**	**asthma**	**opioids**	**total**
Image	0.02 $\pm$0.07	0.15 $\pm$0.12	0.15 $\pm$0.15	0.14 $\pm$0.15	0.14 $\pm$0.13	0.19 $\pm$0.14	0.07 $\pm$0.1
Env.	0.19 $\pm$0.08	0.33 $\pm$0.13	0.42 $\pm$0.15	0.41 $\pm$0.13	0.37 $\pm$0.11	0.52 $\pm$0.12	0.26 $\pm$0.1
Soc.	0.26 $\pm$0.1	0.37 $\pm$0.11	0.41 $\pm$0.15	0.39 $\pm$0.12	0.32 $\pm$0.13	0.47 $\pm$0.14	0.22 $\pm$0.11
Env. + Soc.	0.35 $\pm$0.08	0.44 $\pm$0.1	0.50 $\pm$0.13	0.48 $\pm$0.12	0.43 $\pm$0.1	0.6 $\pm$0.1	0.31 $\pm$0.1

Uncovering Health Factors Understanding the impact of the environment and sociodemographic conditions on human health has been a focus of many studies [14, 2]. Our dataset offers new perspectives on such studies, as it allows the investigation of these relationships on a fine-grained spatial level, and for many conditions simultaneously. To shed light on these perspectives, we apply the SHAP [42] approach on the LightGBM models trained for prescription prediction. In Figure 3 we show the 10 most important features estimated by the SHAP algorithm for the models used to predict diabetes and total prescriptions. The left plot shows that the sociodemographic indicators describing the occupation, commute habits, migration, and ethnicity are highly relevant for predicting diabetes prescriptions. Specifically, the model establishes a linkage between lower prescription rates and LSOAs characterized by a substantial prevalence of professional occupations, a high number of people working from home, an active bicycle commuting trend, and a prominent student population. Conversely, increased prescription levels are associated with LSOAs having a high proportion of individuals of Asian ethnicity. When it comes to the environmental features, the east-west wind component and PM2.5 appear to be relevant indicators as higher diabetes prescriptions align with LSOAs characterized by an eastward wind pattern and heightened air pollution levels, as indicated by the PM2.5 metric. Additionally, bare soil evaporation and ozone stand out as noteworthy contributors to diabetes prescriptions even though these features exhibit no straightforward linear association with their corresponding SHAP values. On the other hand, in the right plot, we note that the environmental features describing canopy evaporation, NO2, east-west wind, and thermal radiation rank among the

https://doi.org/10.52202/075280-3400

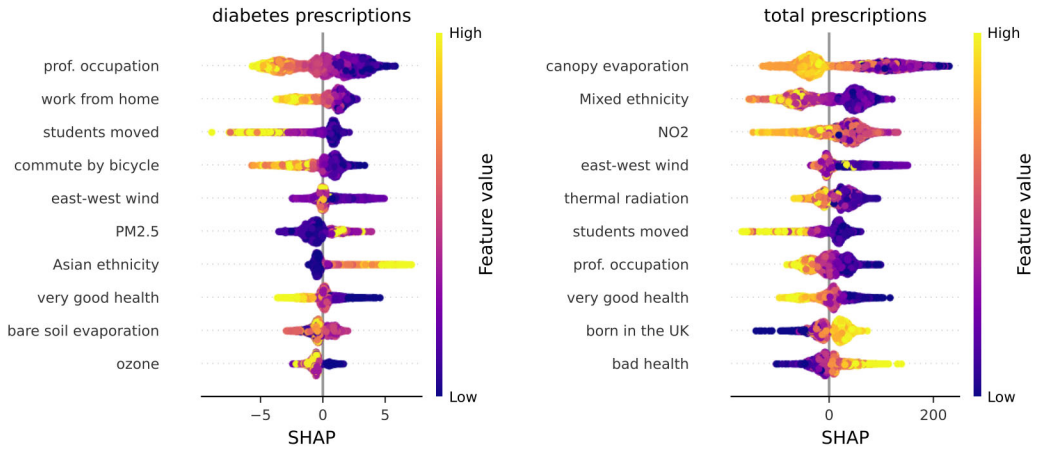

Figure 3: **SHAP summary plots for diabetes (left) and total prescriptions (right)**. The SHAP value for a feature indicates its contribution towards the difference between the prediction for an instance and the average model prediction. These plots reveal the 10 most important features for both conditions and the association between the feature values and their importance. In line with the results shown in Table 2, the sociodemographic indicators appear to be more relevant for modeling diabetes prescriptions than the environmental features while the opposite is observed for the total prescriptions. Although a similar set of features are ranked among the most important for both conditions, some features are particularly relevant for specific conditions, such as work from home and Asian ethnicity for diabetes prescriptions and thermal radiation and Mixed ethnicity for the total prescriptions.

top-5 most relevant features for estimating total prescriptions. Concretely, high values for canopy evaporation, NO2, and thermal radiation are negatively correlated with the total prescriptions while east-west wind displays a similar association as for the diabetes prescriptions. With respect to sociodemographic features, LSOAs characterized by mixed ethnicity are associated with a lower number of total prescriptions compared to the ones where the majority of the people are born in the UK. Moreover, we also see that low prescriptions are again associated with a high percentage of students and professional occupations. Finally, for both conditions, we notice that the positive self-assessment of health is linked to lower prescription values. Examples of SHAP values for the other conditions as well as dependence plots describing the feature interactions are provided in Sections E.2 in the Appendix.

Describing Health of Environment Using Visual Concepts. To shed light on the potential benefits of using the Sentinel-2 imagery for modeling population health, we reveal the learned visual features patterns in Figure 4 by visualizing examples of LSOA Sentinel-2 images for which the LightGBM model trained on the simple image features closely approximates the actual opioids prescriptions. The LSOAs were visualized with the band combination (B11, B06 and B01) as these bands appeared among the most salient image bands according to the SHAP values for the LightGBM model shown in Appendix, Figure 13. First, we note that although the simple image features do not encode the size of an LSOA, they still enable the LightGBM model to associate higher opioid prescriptions with LSOAs covering larger geographical areas. Larger-area LSOAs are rural (because these administrative units are designed to have roughly equal populations), and it is known from previous research that opioid consumption is higher in rural areas [16]. Equally important, we also note that the LSOAs with high prescription values in the first row are characterized by a stronger presence of blue and pink colors occurring near traffic roads than those LSOAs with low opioid prescriptions in the bottom row. Due to the chosen band combination that displays the aerosols (B01) band in blue, this finding points out that the model can relate increased opioid prescriptions for LSOAs exposed to air pollutants. In conclusion, as observed in Table 2 and Table 6 in the Appendix, the environmental and sociodemographic features explain a higher percentage of the variance for prescription predictions. However, this analysis underscores that even basic image features can offer valuable insights into the intricate task of population health modeling. This suggests that leveraging the deep learning methodologies for processing the Sentinel-2 images at the LSOA level has great potential to improve

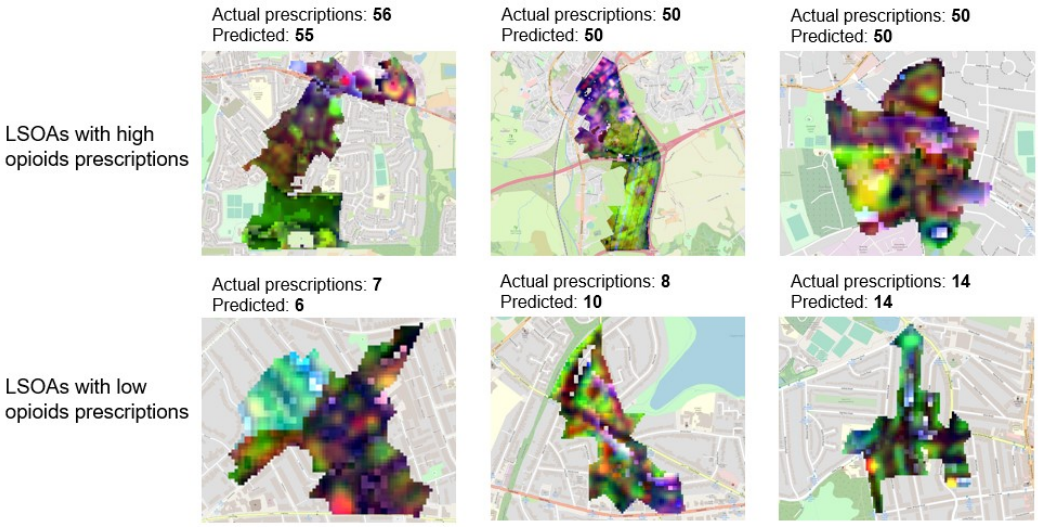

Figure 4: **Visualizing LSOA Instances in the (B11, B06, B01) band combination**. The short-wave infrared band (B11) is shown in red, the near-infrared band (B06) is shown in green and the aerosols band (B01) is shown in blue. The LSOAs with high and low opioid prescriptions are shown in the first and the second row, respectively. Remarkably, LSOAs with high opioid prescriptions cover a larger geographical area than those with low prescriptions (notice the higher zoom-in level), and have greater presence of aerosols (band B01) depicted in blue and purple colors.

prescription predictions by capturing the spatial dependencies inside an LSOA. Moreover, using the recent works in eXplainable Artificial Intelligence (xAI) such as [26] can portray environmental health through rich visual concepts, thus opening possibilities for novel insights about the relevant urban and rural structures influencing population health.

4.3 Temporal Analyses

Using MEDSAT, we analyzed temporal differences in outcome and environmental features between 2019 and 2020. Appendix Figures 15 and 16 illustrate varied distributions for both prescription quantities and environmental features. Notably, in the first COVID year (2020), there was a rise in prescriptions for *anxiety* and *depression* (in line with reports that the pandemic presented enormous challenges to mental health services in UK [13]) and *diabetes* medications in England, while *asthma* and *hypertension* prescriptions decreased [13].

The environmental shifts during the initial COVID year are evident from altered air pollutant distributions. Satellite data showed decreased levels of *NO2*, *ozone*, and *PM2.5* in England for 2020 (as reported erlier in [57]). Land cover changes saw increased *built* areas, likely due to heightened construction activity in the latter half of the year [24, 21], and reduced *trees* cover. Furthermore, 2020 witnessed elevated *temperatures*, *solar radiation*, and both components of *wind* compared to 2019.

5 Impact, Limitations, and Perspectives

We introduced MEDSAT, a unique dataset providing a comprehensive view of medical prescriptions, average yearly environmental indicators, image features, and sociodemographic factors across England for 2019 (pre-COVID) and 2020 (COVID). This resource enables a thorough assessment of health status for various conditions and exploration of their relationships with sociodemographic and environmental factors.

MEDSAT has three significant impacts. First, it has the potential to empower the development of novel machine learning (ML) approaches tailored for spatially-autocorrelated public health data [45], that can augment still predominant traditional statistical models like spatial linear regression [6] and BYM [9], as recent work indicates for xAI models [40]. MEDSAT enables ML research with large and complex data, effective feature engineering, capturing temporal dependencies, addressing

https://doi.org/10.52202/075280-3400

imbalanced data, ensuring interpretability, and achieving generalization across diverse regions. Secondly, MEDSAT can facilitate novel discoveries in public health by revealing influential factors that profoundly affect health outcomes. Through SHAP analyses, we confirmed the established link between diabetes and ethnicity, with higher prevalence among people of Asian descent [15, 28], and the preventive effects of biking and active commuting against diabetes and metabolic conditions [54]. Notably, our data from the initial year of the COVID-19 pandemic highlights the impact of socioeconomic factors. Higher percentages of professional occupations and individuals working from home are associated with lower prevalence of diabetes and total prescriptions, underscoring the influence of deprivation on health outcomes [10, 51, 68, 43]. Our findings not only confirm existing knowledge but also expose less-explored connections between the environment and human health. For instance, our SHAP results demonstrated associations between ozone exposure and mental health prescriptions, as well as between total aerosols and metabolic condition prescriptions. Furthermore, a north-sound wind is linked to a decrease in both types of prescriptions. Thirdly, MEDSAT enables groundbreaking discoveries in population health, particularly regarding health inequalities. Our preliminary analysis uncovers disparities in health accessibility among different economic and ethnic groups. By examining deprivation dimensions such as income, education, and occupational factors, across prescriptions of different types, we can gain a deeper understanding of their contributions to health disparities.

Although MEDSAT is among the most comprehensive publicly-available public and population health datasets to date, it is not without limitations. First, prescription prevalence may not always reflect the true prevalence of the medical condition itself. That is because disparities in healthcare access, privilege, knowledge, and stigmatization can influence prescription rates for certain conditions among different populations [50, 48, 17]. However, it is crucial to note that despite this limitation, our dataset offers a unique opportunity to disentangle these effects, especially when combined with other types of health outcome indicators. Compared to surveys and population samples, which come with their own set of biases, MEDSAT provides a more comprehensive health outcome perspective. Moreover, our method of estimating prescriptions using a probabilistic framework, particularly for the four conditions for which we associated drugs using DrugBank, is imperfect. There exists a possibility that we missed certain drug names, or that medications designed for alternate conditions could potentially be inaccurately included. This limitation of our study arises from our labeling method for prescribed drugs. Drugs are labeled according to associated conditions as sourced from the DrugBank database, without claiming any specific intent behind the prescription from the GP. While we can ascertain that a drug is likely prescribed for a given condition, it is worth noting that drugs can be associated with multiple conditions, both as per DrugBank, and in prescriptions by a GP. This multi-condition association increases the chances of co-prescriptions in our dataset. However, numerous studies on prescription patterns, from antihypertensive [34], to antidepressants [44] to anxiolytics [35] highlight the importance of prescriptions as a health outcome per se and its significance in the field of public health outcomes.We also emphasize that the initial drug list output by DrugBank can be augmented with human expert knowledge in a mixed-method approach to ensure the most accurate results. Our analyses show that the correlations between prescription prevalence scores derived with the automatic and manual methods range from .94 (for anxiety) to .99 for diabetes (see Appendix section D.4.3). Second, our dataset exclusively covers England, representing a single developed country. It is noteworthy that England provides high-quality data on both prescriptions and auxiliary census information, and the methods and insights derived from MEDSAT can serve as a foundation for the development of ML approaches that can subsequently be applied to developing countries once high-quality data becomes available, fostering progress towards tracking SDG about health. Third, there is the risk of stigmatizing certain communities based on our dataset. We believe that listed benefits and opportunities offered by MEDSAT outweigh this risk, and we invite ML and health research communities to employing ethical considerations, fostering inclusivity, and ensuring that the insights gained from MEDSAT are used to enact positive change, promote equity, and reduce health disparities.

Acknowledgments and Disclosure of Funding

Project supported by AI4EO Beyond Fellowship at The International AI for Earth Observation Future Lab at TUM/DLR, ESA Network of Resources Initiative, German Federal Ministry for Economic Affairs and Climate Action in the framework of the "national center of excellence ML4Earth" (grant number: 50EE2201C), Munich Center for Machine Learning, Nokia Bell Labs, and European Union's

Horizon 2020 research and innovation programme under grant agreement No. 869764, awarded to the GoGreenRoutes project.

The authors also thank Sinziana Oncioiu, and Dario Augusto Borges Oliveira for helpful discussions, and Paolo Campanella and Bertrand Robert for their support during the dataset creation.

References

[1] UK Office for National Statistics. `https://www.nomisweb.co.uk/sources/census_2021`, 2022. [Online; accessed 26-February-2022].

[2] Nancy E Adler and Joan M Ostrove. Socioeconomic status and health: what we know and what we don't. *Annals of the New York academy of Sciences*, 896(1):3–15, 1999.

[3] The European Space Agency. Sentinel-2 mission guide. `https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2`, 2023.

[4] Suraiya P Ahmad, Omar Torres, Pawan K Bhartia, Gregory Leptoukh, and SJ Kempler. Aerosol index from toms and omi measurements. In *Proc. of the 86th AMS annual meeting*, 2006.

[5] Luc Anselin. *Spatial econometrics: Methods and Models*, volume 4. Springer Science & Business Media, 1988.

[6] Luc Anselin. Spatial regression. *The SAGE handbook of spatial analysis*, 1:255–276, 2009.

[7] Evelise Pereira Barboza, Marta Cirach, Sasha Khomenko, Tamara Iungman, Natalie Mueller, Jose Barrera-Gómez, David Rojas-Rueda, Michelle Kondo, and Mark Nieuwenhuijsen. Green space and mortality in european cities: a health impact assessment study. *The Lancet Planetary Health*, 5(10):e718–e730, 2021.

[8] Douglas A Becker, Matthew HEM Browning, Olivia McAnirlin, Shuai Yuan, and Marco Helbich. Is green space associated with opioid-related mortality? an ecological study at the us county level. *Urban Forestry & Urban Greening*, 70:127529, 2022.

[9] Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43:1–20, 1991.

[10] Luca Bonacini, Giovanni Gallo, and Sergio Scicchitano. Working from home and income inequality: risks of a 'new normal'with covid-19. *Journal of population economics*, 34(1):303–360, 2021.

[11] Christopher F Brown, Steven P Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie J Pasquarella, Robert Haertel, Simon Ilyushchenko, et al. Dynamic world, near real-time global 10 m land use land cover mapping. *Scientific Data*, 9(1):251, 2022.

[12] Nicholas A Buckley and Peter R McManus. Changes in fatalities due to overdose of anxiolytic and sedative drugs in the uk (1983–1999). *Drug safety*, 27:135–141, 2004.

[13] Wendy Burn and Santosh Mudholkar. Impact of covid-19 on mental health: Update from the united kingdom. *Indian journal of psychiatry*, 62(Suppl 3):S365, 2020.

[14] Bingheng Chen and Haidong Kan. Air pollution and population health: a global challenge. *Environmental health and preventive medicine*, 13:94–101, 2008.

[15] Yiling J Cheng, Alka M Kanaya, Maria Rosario G Araneta, Sharon H Saydah, Henry S Kahn, Edward W Gregg, Wilfred Y Fujimoto, and Giuseppina Imperatore. Prevalence of diabetes by race and ethnicity in the united states, 2011-2016. *Jama*, 322(24):2389–2398, 2019.

[16] Theodore J Cicero, James A Inciardi, and Alvaro Muñoz. Trends in abuse of oxycontin® and other opioid analgesics in the united states: 2002-2004. *The Journal of Pain*, 6(10):662–672, 2005.

 https://doi.org/10.52202/075280-3400

[17] Claudia Cooper, Nicola Spiers, Gill Livingston, Rachel Jenkins, Howard Meltzer, Terry Brugha, Sally McManus, Scott Weich, and Paul Bebbington. Ethnic inequalities in the use of health services for common mental disorders in england. *Social psychiatry and psychiatric epidemiology*, 48:685–692, 2013.

[18] Helen J Curtis, Richard Croker, Alex J Walker, Georgia C Richards, Jane Quinlan, and Ben Goldacre. Opioid prescribing trends and geographical variation in england, 1998–2018: a retrospective database study. *The Lancet Psychiatry*, 6(2):140–150, 2019.

[19] Helen J Curtis and Ben Goldacre. Openprescribing: normalised data and software tool to research trends in english nhs primary care prescribing 1998–2016. *BMJ Open*, 8(2):e019921, 2018.

[20] DEFRA. UK AIR Information Resource. `\url{https://uk-air.defra.gov.uk}`, 2023. [Online; accessed 3-June-2023].

[21] Trading Economic. United kingdom construction pmi. `https://tradingeconomics.com/united-kingdom/construction-pmi`, 2021. [Online; accessed 28-August-2023].

[22] European Environment Agency (EEA). Environmental health impacts. `\url{https://www.eea.europa.eu/en/topics/in-depth/environmental-health-impacts}`, 2023. [Online; accessed 05-June-2023].

[23] Centers for Disease Control, Prevention, et al. National health and nutrition examination survey (nhanes), 2007.

[24] Office for National Statistics. How has uk construction performed over the pandemic? `https://blog.ons.gov.uk/2021/10/19/how-has-uk-construction-performed-over-the-pandemic/`, 2021. [Online; accessed 28-August-2023].

[25] Sebastien Garrigues, Julien Chimot, Melanie Ades, Antje Inness, Johannes Flemming, Zak Kipling, Angela Benedetti, Roberto Ribas, Soheila Jafariserajehlou, Bertrand Fougnie, et al. Monitoring multiple satellite aerosol optical depth (aod) products within the copernicus atmosphere monitoring service (cams) data assimilation system. *Atmospheric Chemistry and Physics*, 22(22):14657–14692, 2022.

[26] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

[27] Christopher J Gidlow, Graham Smith, David Martinez, Richard Wilson, Paul Trinder, Regina Gražulevičienė, and Mark J Nieuwenhuijsen. Research note: Natural environments and prescribing in england. *Landscape and Urban Planning*, 151:103–108, 2016.

[28] Louise M Goff. Ethnicity and type 2 diabetes in the uk. *Diabetic Medicine*, 36(8):927–938, 2019.

[29] Sherif Gonem, Andrew Cumella, and Matthew Richardson. Asthma admission rates and patterns of salbutamol and inhaled corticosteroid prescribing in england from 2013 to 2017. *Thorax*, 74(7):705–706, 2019.

[30] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 2017.

[31] Marco Helbich, Nadja Klein, Hannah Roberts, Paulien Hagedoorn, and Peter P Groenewegen. More green space is related to less antidepressant prescription rates in the netherlands: A bayesian geoadditive quantile regression approach. *Environmental research*, 166:290–297, 2018.

[32] J Irizar, M Melf, P Bartsch, J Koehler, S Weiss, R Greinacher, M Erdmann, V Kirschner, A Perez Albinana, and D Martin. Sentinel-5/uvns. In *International Conference on Space Optics—ICSO 2018*, volume 11180, pages 41–58. SPIE, 2019.

[33] Ruth H Jack, Chris Hollis, Carol Coupland, Richard Morriss, Roger David Knaggs, Debbie Butler, Andrea Cipriani, Samuele Cortese, and Julia Hippisley-Cox. Incidence and prevalence of primary care antidepressant prescribing in children and young people in england, 1998–2017: A population-based cohort study. *PLoS medicine*, 17(7):e1003215, 2020.

[34] Noah Jarari, Narasinga Rao, Jagannadha Rao Peela, Khaled A Ellafi, Srikumar Shakila, Abdul R Said, Nagaraja Kumari Nelapalli, Yupa Min, Kin Darli Tun, Syed Ibrahim Jamallulail, et al. A review on prescribing patterns of antihypertensive drugs. *Clinical hypertension*, 22:1–8, 2015.

[35] A John, AL Marchant, JI McGregor, JOA Tan, HA Hutchings, V Kovess, S Choppin, J Macleod, MS Dennis, and K Lloyd. Recent trends in the incidence of anxiety and prescription of anxiolytics and hypnotics in children and young people: an e-cohort study. *Journal of affective disorders*, 183:134–141, 2015.

[36] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[37] Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, et al. Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, 39(suppl_1):D1035–D1041, 2010.

[38] Kévin Le Rest, David Pinaud, Pascal Monestiez, Joël Chadoeuf, and Vincent Bretagnolle. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global ecology and biogeography*, 23(7):811–820, 2014.

[39] Alex Levering, Diego Marcos, and Devis Tuia. On the relation between landscape beauty and land cover: A case study in the uk at sentinel-2 resolution with interpretable ai. *ISPRS journal of Photogrammetry and Remote Sensing*, 177:194–203, 2021.

[40] Ziqi Li. Extracting spatial effects from machine learning model using local interpretation method: An example of shap and xgboost. *Computers, Environment and Urban Systems*, 96:101845, 2022.

[41] Kumiko M Lippold, Christopher M Jones, Emily O'Malley Olsen, and Brett P Giroir. Racial/ethnic and age group differences in opioid and synthetic opioid–involved overdose deaths among adults aged 18 years in metropolitan areas—united states, 2015–2017. *Morbidity and Mortality Weekly Report*, 68(43):967, 2019.

[42] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[43] Michael Marmot and Jessica Allen. Covid-19: exposing and amplifying inequalities. *J Epidemiol Community Health*, 74(9):681–682, 2020.

[44] Becky Mars, Jon Heron, David Kessler, Neil M Davies, Richard M Martin, Kyla H Thomas, and David Gunnell. Influences on antidepressant prescribing trends in the uk: 1995–2011. *Social psychiatry and psychiatric epidemiology*, 52:193–200, 2017.

[45] Vishwali Mhasawade, Yuan Zhao, and Rumi Chunara. Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence*, 3(8):659–666, 2021.

[46] Jennifer Mindell, Jane P Biddulph, Vasant Hirani, Emanuel Stamatakis, Rachel Craig, Susan Nunn, and Nicola Shelton. Cohort profile: the health survey for england. *International journal of epidemiology*, 41(6):1585–1593, 2012.

[47] NHS. BNF Classifications. `https://digital.nhs.uk/data-and-information/areas-of-interest/prescribing/practice-level-prescribing-in-england-a-summary/practice-level-prescribing-glossary-of-terms`, 2019. [Online; accessed 5-October-2019].

 https://doi.org/10.52202/075280-3400

[48] Barbara I Nicholl, Daniel J Smith, Breda Cullen, Daniel Mackay, Jonathan Evans, Jana Anderson, Donald M Lyall, Chloe Fawns-Ritchie, Andrew M McIntosh, Ian J Deary, et al. Ethnic differences in the association between depression and chronic pain: cross sectional results from uk biobank. *BMC Family Practice*, 16(1):1–10, 2015.

[49] Met Office. When does spring start? Meteorological spring. \url{https://www.metoffice.gov.uk/weather/learn-about/weather/seasons/spring/when-does-spring-start}, 2023. [Online; accessed 3-June-2023].

[50] John R Pamplin II and Lisa M Bates. Evaluating hypothesized explanations for the black-white depression paradox: A critical review of the extant evidence. *Social Science & Medicine*, 281:114085, 2021.

[51] Jay A Patel, FBH Nielsen, Ashni A Badiani, Sahar Assi, VA Unadkat, B Patel, Ramya Ravindrane, and Heather Wardle. Poverty, inequality and covid-19: the forgotten vulnerable. *Public health*, 183:110, 2020.

[52] Carol Pierannunzi, Shaohua Sean Hu, and Lina Balluz. A systematic review of publications assessing reliability and validity of the behavioral risk factor surveillance system (brfss), 2004–2011. *BMC medical research methodology*, 13(1):1–14, 2013.

[53] Mimi Pierce, Jan van Amsterdam, Gerard A Kalkman, Arnt Schellekens, and Wim van den Brink. Is europe facing an opioid crisis like the united states? an analysis of opioid use and related adverse effects in 19 european countries between 2010 and 2018. *European Psychiatry*, 64(1):e47, 2021.

[54] John Pucher, Ralph Buehler, David R Bassett, and Andrew L Dannenberg. Walking and cycling to health: a comparative analysis of city, state, and international data. *American journal of public health*, 100(10):1986–1992, 2010.

[55] Aleksandar Radosavljevic and Robert P Anderson. Making better maxent models of species distributions: complexity, overfitting and evaluation. *Journal of biogeography*, 41(4):629–643, 2014.

[56] Georgia C Richards, Sibtain Anwar, and Jane Quinlan. Averting a uk opioid crisis: getting the public health messages 'right'. *Journal of the Royal Society of Medicine*, 115(5):161–164, 2022.

[57] David Rojas-Rueda and Emily Morales-Zamora. Built environment, transport, and covid-19: a review. *Current environmental health reports*, 8:138–145, 2021.

[58] Pamela L Sankar and Lisa S Parker. The precision medicine initiative's all of us research program: an agenda for research on its ethical, legal, and social issues. *Genetics in Medicine*, 19(7):743–750, 2017.

[59] Rossano Schifanella, Dario Delle Vedove, Alberto Salomone, Paolo Bajardi, and Daniela Paolotti. Spatial heterogeneity and socioeconomic determinants of opioid prescribing in england between 2015 and 2018. *BMC medicine*, 18:1–13, 2020.

[60] Adrian Spoerri, Marcel Zwahlen, Matthias Egger, and Matthias Bopp. The swiss national cohort: a unique database for national and international researchers, 2010.

[61] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[62] Anastasios Temenos, Ioannis N Tzortzis, Maria Kaselimi, Ioannis Rallis, Anastasios Doulamis, and Nikolaos Doulamis. Novel insights in spatial epidemiology utilizing explainable ai (xai) and remote sensing. *Remote Sensing*, 14(13):3074, 2022.

[63] Zoi Tsimtsiou, Mark Ashworth, and Roger Jones. Variations in anxiolytic and hypnotic prescribing by gps: a cross-sectional analysis using data from the uk quality and outcomes framework. *British Journal of General Practice*, 59(563):e191–e198, 2009.

[64] Parliament UK. Population estimates GP registers: why the difference? https://commonslibrary.parliament.uk/population-estimates-gp-registers-why-the-difference, 2016. [Online; accessed 3-June-2023].

[65] Roozbeh Valavi, Jane Elith, José J Lahoz-Monfort, and Gurutzeta Guillera-Arroita. block cv: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2):225–232, 2019.

[66] Alex J Walker, Helen J Curtis, Richard Croker, Seb Bacon, and Ben Goldacre. Measuring the impact of an open web-based prescribing data analysis service on clinical practice: cohort study on nhs england data. *Journal of Medical Internet Research*, 21(1):e10929, 2019.

[67] WASDI platform. Earth Observation tech for everyone. https://www.wasdi.cloud, 2023. [Online; accessed 3-June-2023].

[68] Liam Wright, Andrew Steptoe, and Daisy Fancourt. Are we all in this together? longitudinal assessment of cumulative adversities by socioeconomic position in the first 3 weeks of lockdown in the uk. *J Epidemiol Community Health*, 74(9):683–688, 2020.

[69] Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David B Lobell, and Stefano Ermon. Sustainbench: Benchmarks for monitoring the sustainable development goals with machine learning. *arXiv preprint arXiv:2111.04724*, 2021.

 https://doi.org/10.52202/075280-3400