*Journal of the Royal Statistical Society Series A: Statistics in Society*, 2024, **00**, 1–24
https://doi.org/10.1093/jrsssa/qnae060

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
**Series A**
Statistics in Society

**Original Article**

# Assessments and developments in constructing a National Health Index for policy-making, in the UK

**Anna Freni-Sterrantino[1], Thomas P. Prescott[1], Greg Ceely[2], Myer Glickman[2] and Chris Holmes[1]**

[1]The Alan Turing Institute, London, UK
[2]Health Analysis and Life Events Division, Office for National Statistics, Newport, UK

*Address for correspondence*: Chris Holmes, The Alan Turing Institute, London, UK. Email: cholmes@stats.ox.ac.uk

## Abstract

Composite indicators are useful for summarizing and comparing changes among different communities. The UK Office for National Statistics has created an annual England Health Index (2015–2018) comprised of three main health domains—lives, places, and people—to monitor health over time and across different geographical areas and evaluate the nation's health. We reviewed the conceptual coherence and statistical requirements, focusing on three main steps: correlation analysis at different levels, comparison of the implemented weights, and a sensitivity and uncertainty analysis. Based on the results, we have highlighted features that have improved the statistical requirements of the forthcoming UK Health Index.

**Keywords:** composite indicator, Health Index, robustness assessment, sensitivity analysis, uncertainty, weights

## 1 Introduction

A composite index (CI) is a way to summarize several indicators in one number and provide a tool for policy-making. Besides the known health-related indices like Healthy Life Expectancy (van de Water et al., 1996) or disability-adjusted life years (Hyder et al., 2012; Soerjomataram et al., 2012), in the UK, there has been a long tradition of health-related indices; the first 'Health Index' was developed in 1943 as a surveillance system for population health at the national level, based on mortality and morbidity annual data (Sullivan, 1966). Kaltenthaler et al. (2004), in their systematic review conducted in 2014, evaluated 17 population-level health indexes and found that three were composed for the UK population. The 'Health and material deprivation in Plymouth' (Abbott & Sapsford, 1994), a modification of Townsend's 'Overall Health Index' (Townsend et al., 1988), and the most popular 'Index of Multiple Deprivation' (IMD) (Department of the Environment and the Regions, 2000). However, none of them or any of the other health-population indexes seemed to fulfil the desiderata for a health index (HI): proper health coverage indicators, routinely collected and updated data, indices at local and national level; and statistical coherence. These findings were later confirmed by Ashraf et al. (2019) in a systematic review. They concluded that most of the indices measured population's overall health outcomes, but only few gave focus to specific health topics or the health of specific subpopulations. They urged the development of population health indices that can be constructed systematically and rigorously, with robust processes and sound methodology.

Recently, to fill this gap, the Office for National Statistics ((ONS) of the UK developed an annual (experimental) CI to quantify health in England, to track changes in health across the country and

**Received:** October 17, 2022. **Revised:** May 30, 2024. **Accepted:** June 3, 2024

compare health measures across different population subgroups. The long-term application will include the potential for policy evaluations, decisions, and interventions to be derived by the causal pathways formed by the index indicators.

The HI expands the WHO definition of health: 'a state of complete physical, mental and social well-being and not merely the absence of disease and infirmity (Grad, 2002), to include health determinants that are known to influence people's health. Therefore, the HI is characterized by three main domains: Healthy *People*, Healthy *Lives,* and Healthy *Places*, split across 17 subdomains, for a total of 58 indicators. For example, life expectancy and the standardized number of avoidable deaths define the subdomain 'Mortality' and prevalence at upper-tier local authority (UTLA) level of dementia, musculoskeletal, respiratory, cardiovascular, cancer, and kidney conditions define the subdomain 'Physical health conditions' within the Healthy People domain. Healthy Places is structured over 14 indicators (access to public and private green space, air and noise pollution, road safety, etc.) split in five subdomains: access to green space, Local environment, Access to housing, Access to services, and Crime.

The construction of a new composite indicator is a lengthy process that considers several steps and choices. From the wide literature on composite indicators (Barclay et al., 2019; Freudenberg, 2003; Jacobs et al., 2004), it emerges that there is no gold-standard, with every method having its own drawbacks and advantages (Greco et al., 2019) relative to the purpose of each CI and its future use in policy-making.

In recent years, extensive work was carried out by many institutions, such as Eurostat (Eurostat, 2017), the Organisation for Economic Co-operation and Development (OECD) (Commission, 2008), the Joint Research Centre (JRC) (Saisana & Tarantola, 2002), and specific working groups at the European Commission (JRC, n.d.), to provide statistical guidance on CI construction. The cumulative effort has provided a framework to define CI principles (Nardo et al., 2005), outlining the essential steps, introducing sensitivity, and uncertainty analysis as a core part of composite indicators (Saisana et al., 2005) and advancing composite indicators methodology (Munda & Nardo, 2005).

With no current unanimous approved checklist for evaluating composite indicators, we relied on two main sources to guide us into assessing the HI. The first is based on the COIN step-list from the JRC (JRC, n.d.), which includes observations from the OECD handbook (Commission, 2008). These elements provide a framework that will guide us on the statistical (quantitative) methodological choices and statistical analysis. The second source is based on previous work carried out in an audit format by the JRC composite indicators expert group (Caperna & Becker, 2022; Saisana & Philippas, 2012), where they have evaluated other composite indicators.

In this paper, in an effort to fulfil transparency requirements while illustrating the difficulties faced, choices made, and limitations of a composite health measure, we evaluated the steps taken and arising issues that come into the design of the ONS HI. We highlight areas of improvement or which warrant further investigation, based on our findings, aiming for a statistically and conceptually coherent index, that will be integrated in the future HI release. This paper is structured as follows. We start by describing the beta ONS HI for 2015–2018 structure and steps taken in its construction, in Section 2, and future potential usage in policy decisions under causality. In Section 3, we provide an in-depth correlation analysis which will be useful for the weights system selection that we introduce in Section 4. The index validity is evaluated by sensitivity and uncertainty analysis in Section 5. Finally, we provide a discussion and conclusions, in Section 6.

## 2 The ONS Health Index

The ONS HI is a CI structured in three main domains: 'Healthy People', 'Healthy Lives', and 'Healthy Places'. These domains are based on 17 subdomains, which are in turn based on 58 indicators, collected for the 149 UTLA in England, from 2015 to 2018. See Table 1 for full indicator and subdomain detailed descriptions (see also online supplementary material, Table 1 in Supplementary Material). The choice of the indicators and the definition of the 17 subdomains, and three domains were based on a comprehensive review of contents of existing indices and frameworks; cross-referenced with existing accepted definitions of health; and then consulted on by an expert group with members from the central government, local organizations, think tanks and academia to evaluate the proposal (Ceely, 2020). The methodology was based on the

**Table 1.** Health Index structure: domains, subdomains and indicators

| **Health domains** | | |
|---|---|---|
| **People (Pe)** | **Lives (Li)** | **Places (Pl)** |
| **Pe.1** Mortality: life expectancy, avoidable deaths | **Li.1** Physiological risk factors: diabetes, overweight and obesity in adults, hypertension | **Pl.1** Access to green space: public green space, private outdoor space |
| **Pe.2** Physical health conditions: dementia, musculoskeletal conditions, respiratory conditions, cardiovascular conditions, cancer, kidney disease | **Li.2** Behavioural risk factors: alcohol misuse, drug misuse, smoking, physical activity, healthy eating | **Pl.2** Local environment: air pollution, transport noise, neighbourhood noise, road safety, road traffic volume |
| **Pe.3** Difficulties in daily life: disability that impacts daily activities, difficulty completing activities of daily living (ADLs), frailty | **Li.3** Unemployment: unemployment | **Pl.3** Access to housing: household overcrowding, rough sleeping, housing affordability |
| **Pe.4** Personal well-being: life satisfaction, life worthwhileness, happiness, anxiety | **Li.4** Working conditions: job-related training, low pay, workplace safety | **Pl.4** Access to services: distance to GP services, distance to pharmacies, distance to sports or leisure facilities |
| **Pe.5** Mental health: suicides, depression, self-harm | **Li.5** Risk factors for children: infant mortality, children's social, emotional and mental health, overweight and obesity in children, low birth weight, teenage pregnancy, child poverty, children in state care | **Pl.5** Crime: personal crime |
| | **Li.6** Children and young people's education: young people's education, employment and training, pupil absence, early years development, General Certificate of Secondary Education achievement | |
| | **Li.7** Protective measures: cancer screening, vaccination coverage, sexual health | |

10 steps reported in the COIN guidance promoted by the European Joint Research Center (JRC, n.d.). After collating raw data for the indicators at the UTLA level, the steps taken to construct the HI were

(a) data imputation;
(b) data treatment and normalization;
(c) subdomain weights computation for factor analysis (FA);
(d) arithmetic aggregation with equal weights across subdomains and domains.

The index is computed for each UTLA, aggregated geographically to correspond to English regions, and further aggregated into an overall national figure. The index values are calculated for each year from 2015 to 2018 inclusive, with a normalized value anchored at the baseline year 2015. Full details are provided in Supplementary Material (SM).

The HI starts from a tensor $\mathcal{X}$ of raw data, with elements $x_{cit}$. Here, each $c \in C$ is an UTLA, for the set $C$ of $|C| = 149$ UTLAs; each $i \in I$ is an indicator, for the set $I$ of $|I| = 58$ indicators; and each

$t \in T = \{2015, 2016, 2017, 2018\}$ denotes the year. We are also given a partition of the set of UTLAs, $C$, into a set $R$ of $|R| = 9$ regions, $r \in R$, which are disjoint subsets $r \subseteq C$ of UTLAs.

## 2.1 Data imputation

We first note that $\mathcal{X}$ has missing data, which needs to be imputed. There were two types of missing data: missing for a subset of UTLAs, or for a given year with the indicator values were completely missing for all UTLAs (see online supplementary material, Table 2 in SM). In case of missing values at UTLA levels, these range between 3 and 9 missing values over the 2015–2018 and for 15 indicators (by year 0.04%–0.11%). A negligible size for the HI computations. For the indicators completely missed by year, there were three scenarios. The following indicators 'Difficulty in daily activities', 'Public green space' 'Private green space', data was available only for 2018, and for 'House overcrowding' and 'Noise Pollution' only 2015 and 2016, respectively. To fill 2015, we took the available year.

The second scenario was observed for 'Obesity', 'Physical Activity', 'Eating', and 'Scholarity'; data was completely missed for 2015, but covered 2016–2018. We assigned data from 2016 to cover 2015. Finally, 'Noise complaints' were observed for 2015 and 2018; the two missing central years were interpolated. We have conducted our assessment considering only 2015, which is the year that presents the highest number of missed indicators. Of the total 58 indicators, eight are completely missed. We anticipate that in the sensitivity analysis of all eight missed/imputed indicators, only private and public green spaces have the highest impact. They both shift in absolute value over 10 ranking positions.

## 2.2 Data treatment and normalization

Once the missing data has been imputed, the completed tensor $\mathcal{X} = (x_{cit})$ is decomposed into $|I| = 58$ flattened data sets, $\boldsymbol{X}_i = \{x_{cit} : c \in C, t \in T\}$ for each $i \in I$. Using the data transformations $f_i$ listed in online supplementary material, Supplementary Table 3 for each indicator, $i$, the raw indicator data is transformed to $\boldsymbol{Y}_i = \{y_{cit} = f_i(x_{cit}) : c \in C, t \in T\}$. The assignment of each transformation, $f_i$, to an indicator, $i$, is selected to minimize the absolute values of skewness and kurtosis of $\boldsymbol{Y}_i$, aiming for absolute skewness $\leq 2$ and absolute kurtosis $\leq 3.5$. By minimizing (absolute) skewness and kurtosis, we aim to ensure that the transformed data $\boldsymbol{Y}_i$ is approximately normally distributed. For 18 indicators, the skewness and kurtosis of $\boldsymbol{X}_i$ were optimal, 40 indicators have been transformed and of these 18 have been log-transformed (see online supplementary material, Table 3 in SM).

The normalization step in the ONS HI accounts for time and geography, and allows indicators to be compared on the same scale, weighting by the UTLA populations. The normalization transforms elements $y_{cit}$ of $\mathcal{Y}$ into $z$-scores,

$$z_{cit} = (-1)^{\delta_i}\left[\frac{y_{cit} - \mu_i}{\sigma_i}\right],$$

which then define the elements of the tensor $\mathcal{Z} = (z_{cit})$. For each indicator, $i$, we specify $\delta_i = 0$ or $\delta_i = 1$ to ensure that larger positive values for $z_{cit}$ correspond to improved health, a property which we term as being *health directed*. Note that the mean and standard deviation $\mu_i$ and $\sigma_i$ for each indicator, $i$, are taken to be the population-weighted mean and standard deviation of $y_{cit}$ for the chosen baseline year across UTLAs $c \in C$, fixing $t = 2015$. Finally, given the $z$-scores $z_{cit}$ forming the tensor $\mathcal{Z}$, the ONS HI presents the $z$-scores as HI values,

$$h_{cit} = H(z_{cit}) = 100 + 10z_{cit},$$

which are translated and rescaled $z$-scores, such that $h_{cit} = 100$ means that the transformed value, $y_{cit}$, for indicator $i$ in the UTLA, $c$ in year $t$ is equal to the weighted mean, $\mu_i$.

## 2.3 Subdomain weights computation: a time-series factor analysis

The ONS has chosen to compute weights using a time-series FA. The fundamental assumption of FA is that there is a latent factor that underpins the variables in a group. This translates to this level

of the HI: ONS assumed that there is a single unobserved variable that underpins the indicators within each subdomain. This assumption is plausible in the light of the well-recognized geographical clustering of health and socio-economic (dis)advantage although no single underlying factor has been measured in the literature. Highly correlated indicators within each subdomain could lead to double counting in the index, so FA directly addresses this issue, accounting for the correlation between indicators in their implied weights (Decancq & Lugo, 2013).

To maintain the same weights for all the years considered (2015–2018), a time-series FA was applied. The rationale was to ensure that, by accounting for all the years jointly, they would change with each additional year of data. As such, the weights would need to be calculated for a set time period, e.g. 2015–2019, and these weights would be held constant until a review date. This assured that (i) the indicators selected matched the underlying factor (subdomains) over time; (ii) and then the factor loadings were scaled and used as data-driven weights.

In practice, from the normalized data $\mathcal{Z}_{CT} = (z_{ct})$ are collapsed by year and then rescaled to (0,1), next given $d \in D$, a FA on the indicators $i \in d$ was carried out and the weights were chosen as the first loading factor, taken in absolute value. The weights $w_i$ for indicators $i \in I$ are chosen by running FA for each subdomain, $d \in D$, in turn, allowing for one factor estimated using a maximum-likelihood method. For example, for a subdomain $d = \{i_1, i_2\}$ comprised of two indicators, suppose the factor loadings are 0.5 and 0.75. We would then set the weights $w_{i_1} = 0.4$ and $w_{i_2} = 0.6$. In supplementary material, we address the weights constraints taking into account the different aggregation levels.

### 2.4 Arithmetic aggregation with equal weights across subdomains and domains.

The final step is the arithmetic aggregation of the index, where there are equal weights for subdomains $w_s$ and domains $w_d$, while indicator weights are derived from FA. All the weights have been chosen as positive and summing to one, for all the different aggregation levels. The HI, at the hierarchical levels of indicators, subdomains, domains and overall, is then computed for each year at geographical levels of UTLAs, regions and the nation, where the geographical aggregations at the regional and national levels are population-weighted.

### 2.5 The Health Index ranking distribution

For the year 2015, for each UTLA, we plot each domain's HI values, ordering the UTLAs by the overall HI ranking, in Figure 1. It emerges that Lincolnshire, Leeds, and Staffordshire have all three domain index values concentrated at the same values. In contrast, Westminster (the UTLA with the largest difference in domain indexes) presents Healthy People at 109, similar to Kensington and Chelsea, but Healthy Places at 82. Westminster and Blackpool present similar values for Healthy Places and Healthy People, but their ranking is significantly different. It is interesting to note that Healthy Lives sits within the range defined by Healthy Places and Healthy People. Similar patterns are observed for the following years, as reported in the SM (see online supplementary material, Figures 5–7).

### 2.6 Health Index as policy-making tool

The long-term aim for HI is to contribute and assist in policy-making. It has already seen use by Directors of Public Health in local authorities for monitoring public health and making intervention decisions; and by the Department of Health and Department of Levelling Up, Housing and Communities to inform their evidence of health's determinants. A particularly notable example of its local use is by Northumbria Healthcare NHS Foundation Trust, who have produced a version of the HI at the lower-super-output-area level for their two authorities of Northumberland and North Tyneside (Trust, 2022).

While the CI per se is a way to summarize in a number the health status of a geographical area, internally, it is composed of a mix of reflective indicators (also known as effect indicators) and causal indicators (also known as formative). The first type of indicators is a manifestation of an underlying construct—UTLA health. Thus, a change in the construct will drive a change in the effect indicators. In contrast, causal indicators drive a change in the construct. Policy decisions and interventions rely on actions that promote positive changes by leveraging on causal indicators.


**Figure 1.** The 2015 Health Index ordered by UTLA ranking, jointly with Healthy Lives, Healthy People, and Healthy Places indexes, and bars indicating the minimum and maximum value of the domains.

Hence, indicators can be further exploited to investigate causal structures and answer causal policy questions. Identification of causal pathways was out of the scope of this work, as the methods for the final HI version are not settled yet. However, there are a few observations that are worth consideration in this remit.

By including more years, i.e. longer time series, data could be tested for Granger Causality, a type of causality linked with the concept of predictability. An indicator $Y$ is said to 'Granger cause' $X$, if information about the history of $Y$ improves one's ability to predict the behaviour of $X$, or for the composite indicators as done by Iqbal and Nadeem (2006). The authors exploited 30 years of time series to assess if a causal relationship exists between real economic development and monetary growth in Pakistan. Given the indicators and the vast literature on known causal pathways and associations in health statistics (obesity, cardiovascular, and respiratory disease) and health and environmental exposures, once the ONS HI version is finalized, the effort will be to untangle the potential causal pathways to evaluate policy interventions, to ameliorate the health conditions. Hence, causal diagram pathways will explore healthy lives, places, and people by distinguishing between formative and reflective indicators, adding known causal links and investigating association links and other indicators not included in the CI, such as the IMD, Human Development Index, and GDP.

Thus providing in-depth descriptions of the indicators, infographics, maps, specific for each UTLA, stakeholders will gain enormous insights, similarly to the outcomes of the Building Research Establishment's international Healthy Cities Index (BRE HI). The index was developed to compare how global cities perform relative to one another regarding the urban environment's impact on health and well-being (Pineo et al., 2018). The initial round included 20 cities across Europe, the Middle East, South and North America, South Africa, and Asia, with 10 environmental categories over 58 indicators. As a case study for London and Dubai, a causal pathways framework was defined to explain the relation between urban environment exposures and health outcomes using evidence based on the indicators. The authors reported the benefits of presenting the causal pathways diagram to stakeholders (no specific data analysis). It sparked discussion on responsibilities over the urban environment exposures (noise, pollution, etc.), added insights to local stakeholders into understanding the importance of their respective sectors and raised awareness on the links between built environment and health.

In the future, once more data becomes available, in time lengths, population stratification, and geographical layers, it would be possible to assess causality by leveraging more sophisticated statistical methods. The HI spatio-temporal features could be further explored by creating, for example, synthetic populations to define a matching control area (to mimic a randomized trial) for policy effect evaluation. For example, Ben-Michael et al. (2023) created a synthetic state (a mixture of several other American states) with matching characteristics to California and, using an interrupted time-series approach (Ben-Michael et al., 2023; Freni-Sterrantino et al., 2019), estimated the effects of California's gun control program. As the HI is still in its experimental phases, the causality aspects have been left aside for now. However, future advancement in spatial causality methodologies (Akbari et al., 2023), identification of the causal pathways framework between indicators, external information and policy interventions will constitute the next step in HI development.

## 2.7 A modified ONS Health Index

Before investigating the HI and carrying out further analysis: correlation and sensitivity/robustness analysis, we implemented a slight change to the original HI, as presented above. As pointed out in Commission (2008), a certain coherence in the methods needs to be preserved to create a statistically sound index. This change was done to avoid statistical misinterpretation, as not all the potential combinations of data transformation and subsequent data operations could be properly interpreted, as carried out in the ONS version.

Hence, we have computed a modified ONS HI version. We begin from the imputed matrices $\boldsymbol{X}_i$ for each indicator, $i$. Then, instead of directly selecting and applying transformations $f_i$ to ensure normality, we accounted for kurtosis and skewness using winsorization first and then by transforming. This approach resulted in only five to seven variables per year that have been log-transformed (online supplementary material, Table 4 in SM). We proceed to standardize using a $z$-score (following the ONS), and then aggregated with arithmetic mean and equal weights (see online supplementary material, Table 5 in SM for comparison).

We opted for less strict data transformation, as this would have not changed the aggregation formula interpretation. As it stands at the moment, the ONS data transformations included 40

indicators, with 18 indicators log-transformed. By aggregating all transformed variables using an arithmetic mean, the untransformed variables are effectively aggregated via a mix between a geometric mean (for log-transformed variables) and arithmetic mean (for other variables). As succinctly summarized by Nardo et al. (2005), 'when the weighted variables in a linear aggregation are expressed in logarithms, this is equivalent to the geometric aggregation of the variables without logarithms. The ratio between two weights indicates the percentage improvement in one indicator that would compensate for a one percentage point decline in another indicator. This transformation leads to attributing higher weight for a one unit improvement starting from a low level of performance, compared with an identical improvement starting from a high level of performance'.

We used this modified version as the starting point for the rest of this paper. The $z$-scores (see online supplementary material, Figure 2 in SM) comparison between this modified version and the original ONS shows that several indicators have more outliers below the 25th percentile, but overall, there are no major discrepancies in values. Indeed, this modified version generated a different ranking, that affected the UTLAs in the middle, while the top and bottom UTLAs remain unaffected (see online supplementary material, Table 4 in SM). The biggest shift in ranking is observed for Barking and Dagenham (which moved positively 49 positions), whereas Westminster, Herefordshire, and Shropshire all shifted down the rankings by, respectively, 50, 48, and 51 positions. Overall, 52% of the UTLAs shifted in absolute value of within 10 ranking positions, 38% shifted between 20 and 30 positions and only 9% shifted more than 31 positions. Only Blackpool, Kingston upon Hull, City of Northampton, and Hertfordshire kept the same ranking in comparison to the original ONS HI version. All the analysis was conducted on R version 4.2 (R Core Team, 2022) and COINr package (Becker, 2021).

## 3 Correlation analysis

The core of every CI is the indicators, which have to be selected carefully to represent the dimensions of the phenomenon that we are trying to summarize. Hence, correlation analysis plays a dual crucial role in the composite indicator construction. First, statistical analyses anchored on the correlation—such as principal components analysis, FA, Cronbach's alpha—are all suitable to assess that the selected indicators are appropriately representing the statistical dimensions, i.e. theoretical constructs are supported by the data. Second, it is useful to identify highly correlated indicators (subdomains and domains), to highlight data redundancy and potential structure issues. However, it is possible that the methodological preference to reduce redundancy could conflict with the practical aim of the HI to reflect a sufficiently wide range of indicators to help guide or monitor interventions.

Ideally, each indicator (this is true also for subdomains and domains) should be positively moderately correlated with the others, while high inter-correlations may indicate a multi-collinearity problem and collinear terms should be combined or otherwise eliminated. Negative correlations are an undesirable feature in CI, however, they may occur at different hierarchical levels of the index. For example, if an indicator is negatively correlated, it can be removed. If domains or subdomains show negative correlation then aggregation by geometric or arithmetic mean should be discarded as it would insert an element of trade-off where units that perform well in one domain have their overall performance affected by the poor performance on another domain.

To explain how negative correlations affect the CI, Saisana and Philippas (2012) reviewed the sustainable society index (SSI). The index—similarly composed to the HI—has three main domains: Human, Environmental, and Economic well-being. Human and Environmental well-being show negative correlation, as in many countries Human and Economic well-being go hand in hand, at the expenses of the Environment. Their review suggested that these correlations are a sign of a trade-off, whereby many countries that have poor performance on Environment levels, have good performance on all other categories and vice versa, therefore each domain should be presented as itself in scoreboard and not aggregated. This is what happens to Blackpool and Westminster in Figure 4, where Westminster presents the lowest Places indicator and Blackpool for People, but not for Places. We will explore further the trade-off and correlation and their role in weights definition, but before we provide an extended HI correlation analyses.

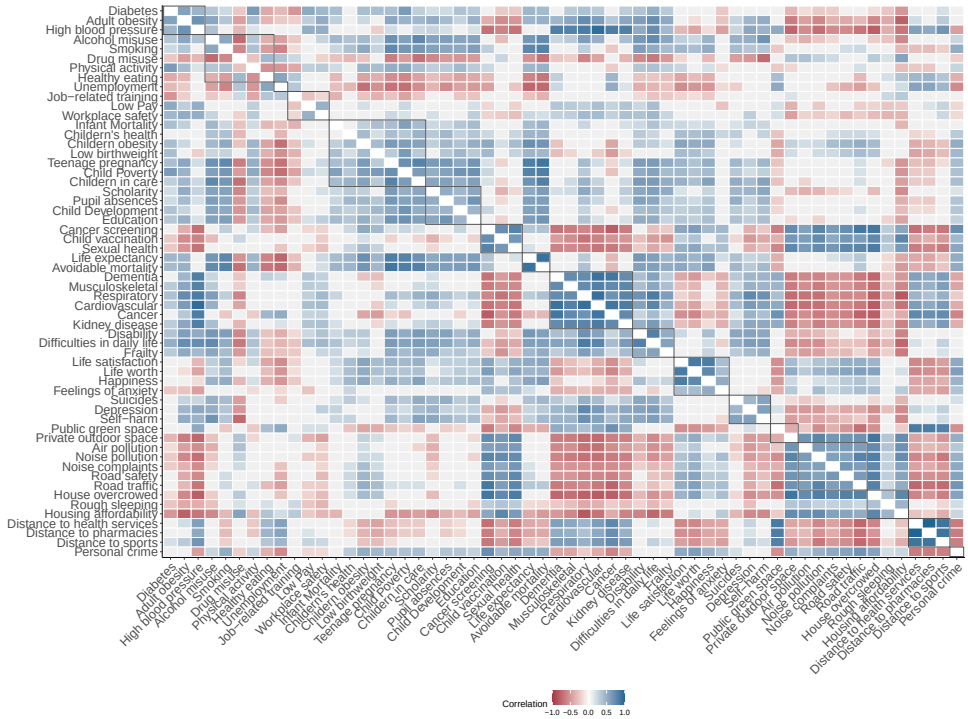


**Figure 2.** Correlation heatmap for indicators grouped in subdomains.

## 3.1 Health Index correlation analysis

We used our modified version of the HI to carry out a correlation analysis (Pearson) at the different levels of aggregation. The correlation analysis provides insights on the potential redundancy of those indicators with high correlation ($\rho \geq 0.9$); negative correlation ($\rho \leq -0.4$) also indicates some conceptual problems. Acceptable correlation values are for weak ($0.3 < \rho \leq 0.4$) and moderate ($0.3 \leq \rho < 0.9$). The ideal situation would be to have indicators positively correlated among them (0.3–0.9), and not highly correlated with other subdomains as this could impact weights and aggregation. In Figure 2, indicators grouped in subdomains are showing overall positive correlations. However, there are some correlations of concern. For example, public and private green spaces that define the subdomain 'Access to green space (Pl.1)' show negative correlations, and in the subdomain 'Access to services (Pl.4)', distance to the nearest pharmacy and general practitioner (GP) are also highly correlated. We suspected that this could be somehow related to the urban/rural UTLA definition. Cardiovascular and respiratory prevalence are highly correlated in subdomain 'Physical health conditions (Pe.2)'. The indicators in 'Behavioural risk factors (Li.2)' and 'Working conditions (Li.4)' present negative and weak correlations. In this heatmap, we see also correlation among the subdomains like blocks. For example, 'Risk factors for children (Li.5)' and 'Children and young's people education (Li.6)' are also correlated, likewise 'Physical health conditions (Pe.2)' and 'Difficulties in daily life (Pe.3)'.

From the subdomain correlation map (see Figure 3), we immediately see that the indicator 'Household overcrowding' is highly correlated with the subdomains on 'Local environment (Pl.2)'. Finally, we correlated (see online supplementary material, Figure 3 in SM) subdomains versus domains, we found that People subdomains are overall well correlated with the other subdomains within their domain. Lives and Places are similar but present some weak correlations: 'Access to services (Pl.4)', and 'Unemployment (L1.3)' and 'Difficulties in daily life (Pe.3)'. This confirms what we have observed in the indicators heatmap.

The panels in Figure 4 show the scatter-plots for the three domains. It can be observed that Healthy Lives and Healthy People have a high Pearson correlation ($\rho = 0.65$), while for Healthy

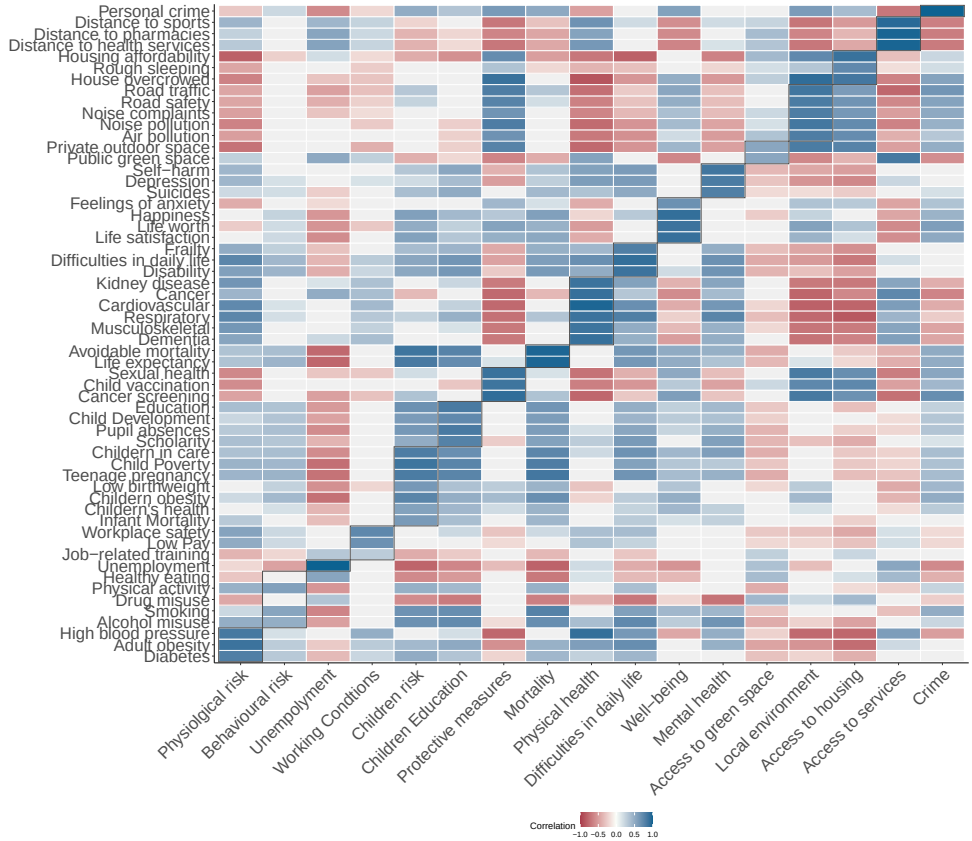


**Figure 3.** Correlation heatmap for indicators and subdomains, grouped by subdomains arithmetic mean.

Lives and Healthy Places ($\rho = -0.12$) and Healthy People and Healthy Places ($\rho = -0.39$) the correlations are negative, a similar situation as described for the SSI by Saisana and Philippas (2012). Once we removed London's UTLAs, characterized by high values of People and Lives and low on Places, the correlation for Lives and People increases ($\rho = 0.72$), null for Lives and Places ($\rho = -0.06$) and diminishes in People and Places ($\rho = -0.25$).

## 4 The choice of a weight system

In this section, we introduce the choice of a weights system that could be employed in the linear aggregation formula that generate the CI. We review the definitions and how the weights can be interpreted. We then proceed on evaluating what role this plays for the correlation at different levels (indicators/subdomains/domains) and we describe the optimized method (Becker et al., 2017) that generates weights that account for correlations.

We also compared the time-series FA-derived weights, currently in use in the ONS HI with that for the ONS HI with weights generated by principal component analysis (PCA). We introduce them here, because we are going to use the PCA weights and the optimized weights as options in our sensitivity and uncertainty analysis.

### 4.1 Weights definitions: compensatory versus noncompensatory

In standard practice (JRC, n.d.; Munda & Nardo, 2005), the composite indicator for time $t$ is defined as

$$z_t = \sum_{c \in C} w_{ct} z_{ct},$$

where $c$ indexes the indicators and $C$ is a set of indicators being composed (which, in the context of the ONS HI, may correspond to subdomains, domains, or the overall index). Thus the composite indicator is a weighted linear aggregation, where weights are (typically) constrained to sum to 1.

In the CI literature (Greco et al., 2019), weights methods are often found to be linear, geometric or multi criteria, or classified into compensatory and noncompensatory approaches. However, the major difference in weight systems boils down to defining weights either as coefficients that address the importance of a variable (indicator/subdomains/domains) or as a trade-off coefficient.

Weights that convey 'importance' should be used in aggregation formulae that do not allow for compensability; that is, where poor performance in some indicators can be compensated by sufficiently high values of other indicators. These definitions are also known as compensatory, because the 'compensation' refers to a willingness to allow high performance on one variable (subdomain/domain) to compensate for low performance on another. The weighted mean (arithmetic/geometric) is a classic example of compensatory approach, where the weight is a de facto trade-off coefficient.

Noncompensatory methods allow the weights to express 'importance', where the greatest weight is placed on the most important 'dimension' (Vansnick, 1990; Vincke, 1992). These approaches have their roots in social choice theory (also known as multicriteria) and more details can be found in Munda (2008). Briefly, in this framework, indicator (subdomain/domain) values rank the countries in different ways and contributed to define the relative performance of each country/option with respect to each of the other countries/option. This indicators-unit ranking, generates an impact matrix and a voting system must be put in place to define the overall ranking. For example, the 'plurality vote' will rank as first the unit (UTLA) that has ranked at first place on the majority of the indicators. However, this approach comes with the price of dealing with preferences and choices on how to select the final ranking given the indicators-ranking (Munda, 2008). Two popular approaches, that take the name after their authors, suggest that a Condorcet approach is necessary when weights are to be understood as importance coefficients, while a Borda approach is desirable when weights are meaningful in the form of trade-offs.

These methods, while valuable, are rather harder to implement as they require an expert panel to grade the indicators in first place, but also lack the immediate facility to explain when compared with a weighted mean. The dual notions of weighting as importance versus weighting as trade-off and their interpretation requires more consideration, to assure that selected weights are in line with the practitioner preferences. In their article, Munda and Nardo (2005, 2009) provide extensive commentary on an interesting misinterpretation around the weights/aggregation combination that gets buried in the CI construction, but is useful to address here.

### 4.2 Weights as 'importance' coefficients, linear aggregation, and correlation

According to the OECD guidelines (Freudenberg, 2003): 'Greater weight should be given to components which are considered to be more significant in the context of the particular composite'. As pointed out by Greco et al. (2019), the popular linear aggregation weights are used as if they were importance coefficients, while they are in fact trade-off coefficients.

Briefly, the authors (Munda & Nardo, 2005, 2009) state that in linear and geometric aggregation the weights play the role of a trade-off ratio that depends on the scale of measurement. If the weight has to be interpreted as a measure of importance, then the weights should be connected with the indicators themselves and not with their quantification; they should be invariant to the units of the indicator. This distinction between weights as trade-off ratio versus importance does not disappear even when all indicators are on the same scale. For a weight to express 'importance', then noncompensablity should be enforced. This issue becomes relevant when CI are composed of different data for multicriteria optimization where improvement in one domain cannot compensate for degradation in another. One way to disentangle this paradox of trade-off weights interpreted as importance weights is proposed by Becker et al. (2017). In order to derive weights as 'explicit importance', we need to evaluate the correlation structure and use it to understand the 'importance' role of the domains/subdomains in the composite indicators, and what the influence of each indicator is on the index, generating optimized weights.

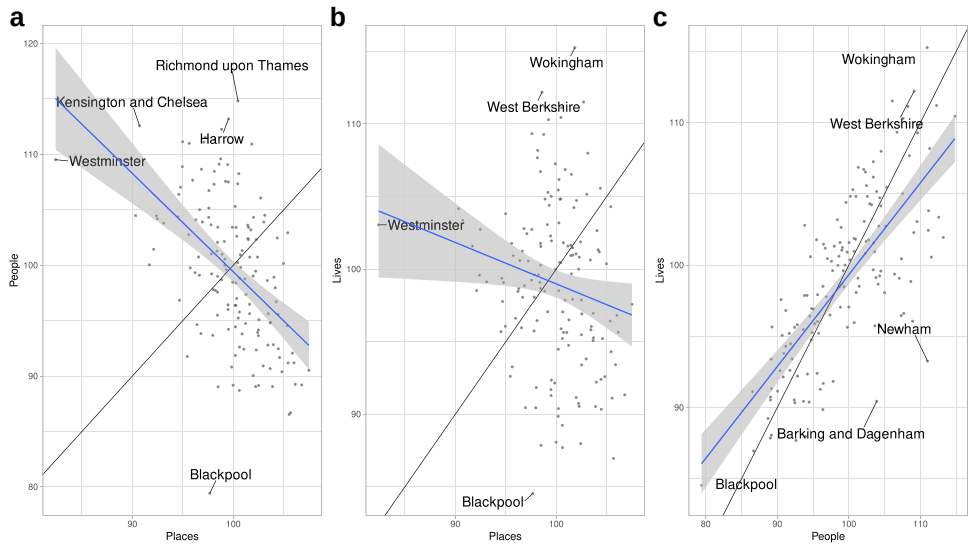


**Figure 4.** UTLA Domains index scatter plots with fitted linear regression (2015): (a) people versus places, (b) lives versus places, and (c) lives versus people.

## 4.3 Optimized compensatory weights

For a weighting system where weights are representing 'explicit importance', then different variances and correlations among indicators (subdomains/domains) mask the weights to represent importance, as shown above in the correlation analysis.

To find weights that reflect importance and not trade-off ratios, conditioned on the correlations, we follow the methods introduced by Becker et al. (2017). We recall that for the HI, domains and subdomains have equal weights, while indicators have data-driven weights derived by FA. If we take the equal weights choice as a way to express equal importance of the three domains, we need to account for each variable's influence on the output, and how weights can be assigned to reflect the desired importance, 'conditioned' on the existing shared information among the domains. Knowing the correlation among domains can help to reduce uncertainty, as strong correlations suggest that the domains should be treated jointly, rather than individually. This can help in reassessing the weights.

A measure of importance, capturing the dependence between the CI and the effect of domains, starts from analysing the correlations ratio $S_d$, also known as the first-order sensitivity index or main effect. We split the correlation ratio in two parts: a correlated part, $S_d^c$, and an uncorrelated part, $S_d^u$, such that

$$S_d = S_d^c + S_d^u,$$

where $d = 1, 2, 3$ indicates the level of aggregation.

A large value for $S_d$, with a relatively low uncorrelated part $S_d^u$ such that $S_d \approx S_d^c$, indicates that the domain contribution to the index variance is only due to the correlation with the other domains (Mara & Tarantola, 2012). However, if $S_d^c$ is negative, this implies conceptual problems with one of the domains, and is not a desirable feature in composite indicators.

The optimized weights have been presented in Becker et al. (2017). It is important to note that, while we have applied this approach at the domain level, the same methodology can be applied to other levels of the hierarchical structure of the HI, i.e. for aggregating indicators into subdomains and subdomains into domains. Briefly, first, we estimate $S_d$ and $S_d^u$ by implementing a series of linear and nonlinear regressions (using splines Wood, 2001). The steps to compute the two summands of the correlation ratio are the following:

(a) Estimate $S_i$ using a nonlinear regression approach
(b) Perform a regression of $x_d$ on $x_{\sim d}$. This can be either linear (using multivariate linear regression), or nonlinear (using a multivariate Gaussian process). Denote this fitted regression as $\hat{x}_d$.
(c) Get the residuals of this regression, $\hat{z}_d = x_d - \hat{x}_d$.
(d) Estimate $S_d^u$ by a nonlinear regression of $y$ on $\hat{z}_d$, using the same approach as in step (a).
(e) The correlated part then is the simple expression $S_d^c = S_d - S_d^u$

Using a simple numerical approach, the weights are estimated that result in the desired importance using an optimization algorithm. If $\tilde{S}_d = \frac{S_d}{\sum_{d=1}^{D} S_d}$ is the normalized correlation of $x_d$, then the targeted normalized correlation ratio is $\tilde{S}_d^\star$, where it is assumed that is $\tilde{S}_d^\star = w_d$ is the weight assumed (in our case equal weights) to reflect the importance. Once these quantities have been computed and provided with equal weights (or any other weight system provided by the user), the optimized weights are the result of minimizing the objective function.

$$w_{opt} = \sum_{d=1}^{D} (\tilde{S}_d^\star - \tilde{S}_d(w))^2.$$

The results of this approach are reported in Figure 5. Our first observation is that the domains have fairly similar linear and nonlinear correlation ratio $S_d$ estimates, indicating that linear estimates would have been sufficient to address the linear correlation among the three domains. Recall that we would like low $S_d^c$ and high $S_d^u$, both positive. What we have obtained is that the correlated part dominates in Healthy Lives, indicating that Healthy Lives has a small impact on the CI as it is mostly imputable to the correlation with the other variable. For Healthy People, both components contribute equally and both positively. Healthy Places has a negative correlated effect, which is similar to the uncorrelated part, but the negative $S_dc$ values imply potential problems in the CI. As we have observed in the correlation analysis, we could have expected that Healthy Places could have some problematic behaviour.

Having unpacked the correlation among the domains, we can use this information to find a new set of weights that truly reflects the importance of each variable in the CI, but that are close to the importance distribution we have specified—in our case, equal importance (each domain 0.33 weight). The optimization algorithm finds optimal weights of 0.45 for Healthy People, 0.16 for Healthy Lives, and 0.73 for Healthy Places. These weights will be used subsequently for a sensitivity and uncertainty analysis.

### 4.4 Principal component analysis derived weights

While there is no objective choice in selecting the weights, we concentrate on a so-called data-driven weighting system derived from PCA or FA. Now, in the context of composite indicator construction, these two methods can be applied at different steps due to their versatile interpretation: to identify dimensions, to cluster indicators and to define weights. While PCA and FA share several methodological aspects, there is a key difference between the two analyses. PCA is a data reduction method based on the correlation matrix, which re-defines a new set of uncorrelated variables as linear combinations of the original variables. In contrast, FA is a measurement model of a latent variable, where the latent factor 'causes' the observed variables. There is a *recommendation* in the CI community (Saisana & Tarantola, 2002) to use the PCA loadings as weights only if the first component accounts at least for the 70% of the total variability. We applied this procedure to derive the weighting systems for subdomains. The 58 indicators are split in 17 subdomains (see Table 1), and for each of these subdomains, we carried out a PCA analysis, for each year.

For most subdomains, over all 4 years, the first PCA component accounted for a range between 51% and 94% of the total variability. Exceptions were observed (see online supplementary material, Table 6 in SM) for 'Mental health (Pe.4)' with variance explained 66%–69%, 'Behavioural risk factors (Li.1)' 53%–55%, 'Working conditions (Li.4)' 50%–55%, 'Risk factors for children (Li.5)' 55%, 'Children and young people's education (Li.6)' 63%–69%, and 'Access to housing (Pl.3)' 65%–69%. We then normalized the loading coefficient and compared them over time, jointly with the weights originally derived from FA for all the years collapsed.

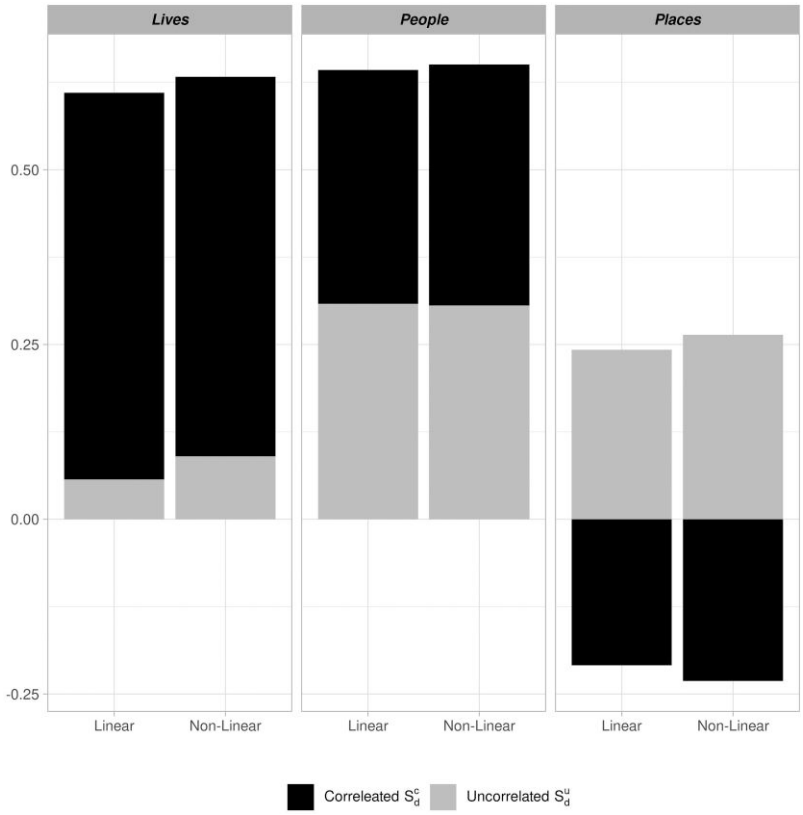


**Figure 5.** Estimates of $S_d$ (full bars), broken down into correlated $S_d^c$ and uncorrelated $S_d^u$, using linear and nonlinear dependence modelling.

We have investigated the PCA weights values over time and compared them with the time-series FA analysis computed for the ONS HI. We have found that these are very similar over time, which is reassuring in terms of the stability of the index weights (see Figure 6). However, when we compared PCA and FA weights, we have found that FA gave higher weights to the following indicators (difference percentage among weights): low pay (12%), self-harm (10%), difficulty completing activities of daily living (5.4%), and drug misuse (6.6%). On the contrary, PCA imposed higher weights to job-related training (7.6%), physical activity (7%), suicides (5.9%), and workplace safety (4.4%).

## 5 Sensitivity and uncertainty analysis

Following the approach introduced by Saisana et al. (2005), Sobol (1993), and Saltelli et al. (2010), we carried out an analysis of the sensitivity and uncertainty of the HI. This analysis is based on a variance-based approach that constructs Monte Carlo estimates of the variability observed due to each step, and due to the interactions between the different steps. For each of the construction steps $\mathbf{q}_i$ we select a potential alternative method. Therefore, indicating the model with $m$, we can compute the global variance as

$$V(m) = \sum_i V_i + \sum_i \sum_{j>i} V_{i,j} + \cdots + V_{1,2,\ldots,k},$$

where

$$V_i = V_{q_i}[E_{\mathbf{q}_{-i}}(m \mid \mathbf{q}_i)],$$
$$V_{i,j} = V_{q_i q_j}[E_{\mathbf{q}_{-ij}}(m \mid \mathbf{q}_i, \mathbf{q}_j)] - V_{q_i}[E_{\mathbf{q}_{-i}}(m \mid \mathbf{q}_i)] - V_{q_j}[E_{\mathbf{q}_{-j}}(m \mid \mathbf{q}_j)],$$

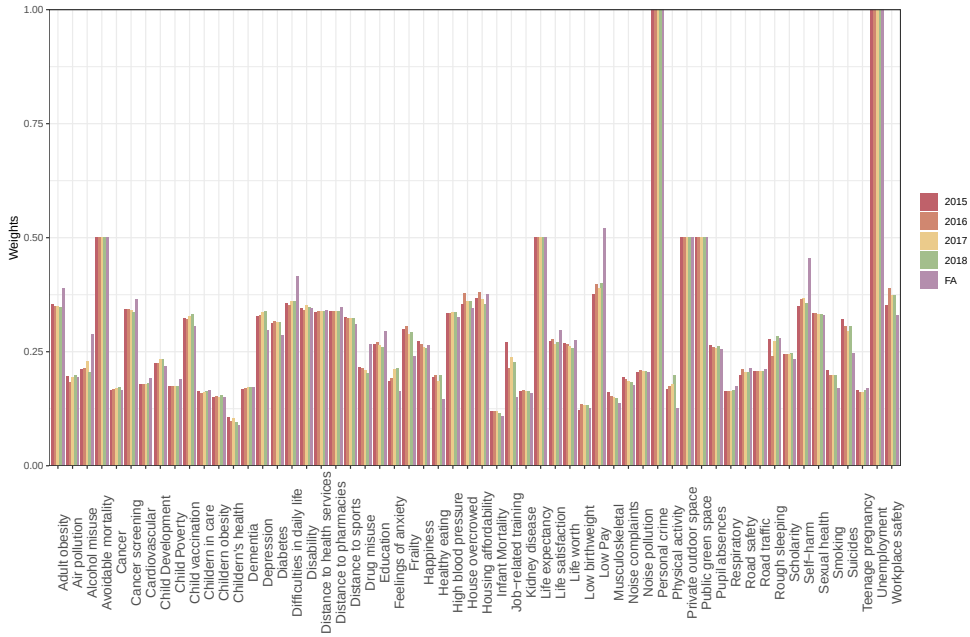


**Figure 6.** Weight comparison between PCA weight per year and time-series factor analysis (FA) weights.

The quantity $V_{q_i}[E_{\mathbf{q}\_i}(m \mid \mathbf{q}_i)]$ and the expectation $E_{\mathbf{q}\_i}$ require the computation of an integral over all factors except $q_i$, including the marginal distributions for these factors. The variance $V_{q_i}$ would imply a further integral over $q_i$ and its marginal distribution.

The sensitivity indices are then $S_i = V_i / V(m)$. These terms measure the contribution of the input $\mathbf{q}_i$ to the total variance and can be interpreted as a fraction of uncertainty.

The first-order sensitivity index, which is the fraction of the output variance caused by each uncertain input assumption alone, is

$$S_i = \frac{V[E(m \mid \mathbf{q}_i)]}{V(m)},$$

this is averaged over variations in other input assumptions, and the total order sensitivity index, (or interaction),

$$S_{Ti} = 1 - \frac{V[E(m \mid \mathbf{q}_{-i})]}{V(m)} = \frac{E[V(m \mid \mathbf{q}_{-i})]}{V(m)}$$

where $\mathbf{q}_{-i}$ is the set of all uncertain inputs except the $i$th quantity, and the quantity $S_{Ti}$ measures the fraction of the output variance caused by $\mathbf{q}_i$ and any interactions with other assumptions. In carrying out the sensitivity analysis, we have selected potential steps $\mathbf{q}_i$ that are coherent with a final linear aggregation.

The steps and the methods to be tested are listed in Table 2. In our analysis, we evaluated (for 2015) the following main outcomes: UTLA ranking by overall index value and UTLA rankings by each domain's index value.

We opted for winsorization to control data kurtosis and skewness, by winsorizing at the second, fifth and tenth values. We allowed for two normalization types: $z$-score centred at 100 and standard deviation at 10; and min–max bounded 1–100. For the weights, we allowed equal weights, PCA derived and optimized weights for domains only, as previously introduced. We ran the computations for 10,000 iterations.

We studied also the absolute mean ranking shift of removing indicators and subdomains, to evaluate the roles played by the hierarchical elements.

**Table 2.** Steps and methods used in the sensitivity analysis

| Steps | Alternatives |
|---|---|
| Data treatment | winsorization (2nd, 5th, 10th points) |
| Normalization | z-score, min–max |
| Weights indicators | equal weights, principal components weights |
| Weights domains | optimized weight |

### 5.1 Results for sensitivity and uncertainty analysis

We carried out the sensitivity analysis on the modified ONS HI and in Figure 7, we note that for the overall index tail rankings are stable, while the middle UTLAs are the ones showing the highest variability with median rankings (green dots) above or below the provided ranking.

We then repeated the analysis for the three domains separately (see online supplementary material, Figure 4 in SM). The estimates are more precise, as the bounds between the 5th and 95th centile are narrower compared with the overall index. We observed that People rankings are quite precise and concentrated and it is possible to see that they are following the HI. Lives and Places are displaying higher variability, with Places acting as the 'wild card'.

The first-order sensitivity and the total order sensitivity have been computed for the overall index and the three domains and we reported them in online supplementary material, Table 7 in SM. We then plotted the main effect $S_i$ and the interactions $S_{T_i}$, see Figure 8. These values can be interpreted as the uncertainty caused by the effect of the $i$th uncertain parameter/assumption on its own. The total order sensitivity index is the uncertainty caused by the effect of the $i$th uncertain parameter/assumption, including its interactions with other inputs. This disentanglement shows that at domain level normalization plays a major role for all of them, with winsorization additionally being quite relevant for Places and weights being relevant for People. For the overall index, weights are the main cause of the variability with normalization and winsorization playing a minor role at interaction levels.

### 5.2 Ranking shifts by removing indicators and subdomains

We assessed the absolute mean differences on the overall rank shift, by removing indicators and subdomains. At the indicator level (see Figure 9), we observed the highest shifts are due to unemployment, access to private and public green space and personal crime. Moderate absolute shifts are observed for job-related training, workplace safety, disability, frailty, suicides, depression, and rough sleeping.

At subdomain levels (see Figure 10), the highest impact is for 'Access to services' and 'Crime', followed by 'Physiological risk' and 'Working conditions (Li.4)'. The observation that Healthy Lives shows the most influence on the overall index values confirms what has already been observed in previous sections, where we note the high correlation between Healthy Lives values and the overall index values.

## 6 Discussion

We have scrutinized the choices made when constructing the ONS HI for England and have evaluated the issues that emerge while assessing each construction step. The resulting HI is easy to explain to wider audiences, and the data collection and the index structure are harmonized to be comparable across time and different geographies. The indicator selection covers the main areas of Health, in line with the WHO definition, and provides access to policymakers to different combinations of indicators and comparisons. The experimental HI fulfils the criteria advocated by Ashraf et al. (2019) and constitutes a starting basis for statistical improvements to improve the future releases.

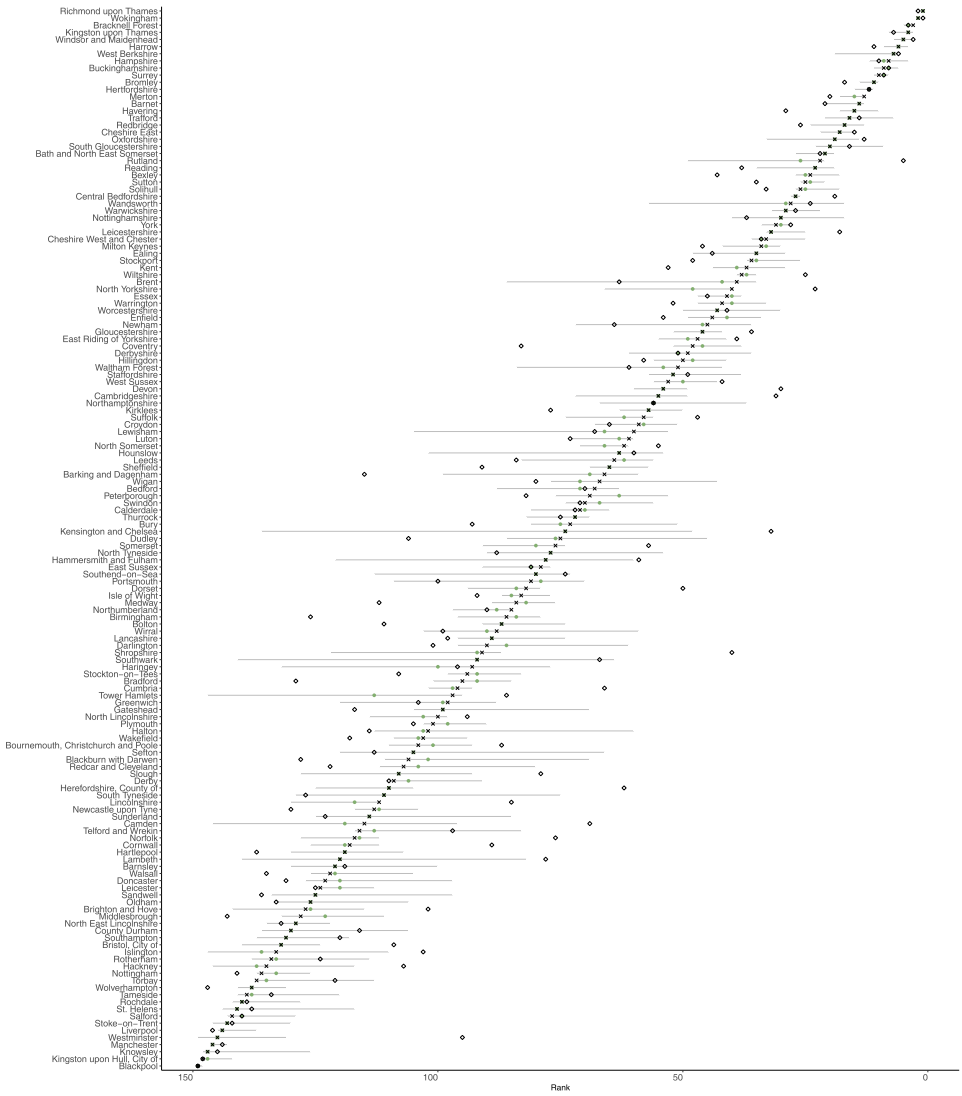


**Figure 7.** Results of UA showing the overall Index for each UTLA, ordered by the modified ranking for 2015 (crosses). With the corresponding 5th and 95th percentiles (bounds) and the median ranking (dots). For comparison, the original ONS UTLA ranking (diamonds).

Our analysis has shown that the weights and normalization steps play a major role in the exhibited variability in the HI, in particular for middle-ranking UTLAs. The steps that generate the most cumbersome decisions to be taken are the choice of the weighting system and the choice of aggregation formula (Greco et al., 2019). However, choices made for both steps need to be taken in the context of the preceding steps. Driven by the desideratum to have an index that is easy to explain, we decided to explore in the sensitivity and uncertainty analysis only those methods that were compatible with the approach taken by the ONS. In our case, we considered the use of different weighting systems and data treatment, while staying consistent with a final linear aggregation formula (i.e. an arithmetic mean). This coherence was also the reason why we recommended to intervene minimally on the data treatment, opting for winsorization and then if still needed, we followed with a transformation to normalize the indicator. The negative correlation exhibited by Healthy Places, the effect on the rank shifting for Healthy Places indicators, and the low ranking correlation with the overall index, could potentially help us to reflect on the choices of the

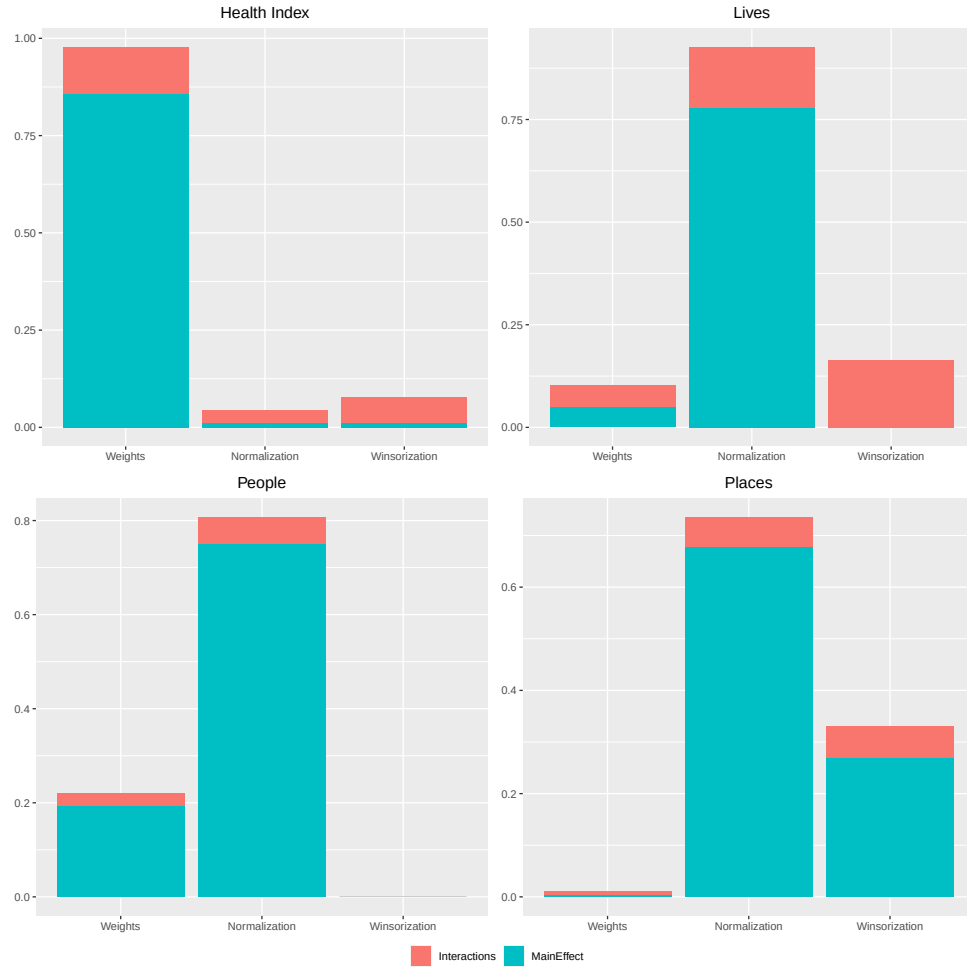


**Figure 8.** Results of the sensitivity order: Main Effect $S_i$ and the interaction $S_{T_i}$.

indicators and potentially revise the indicators selected. However, it is accurate to claim that areas with worse Healthy Places indicators, such as London boroughs (comprising 20% of all UTLAs), score higher values on the other two domains. The reverse is also true, where more rural UTLAs have, for example, lower pollution and good access to private and public green space, but are lower on other indicators.

By exploring the data-derived weights using PCA and comparing them with the initial choices made in the ONS version, we saw some differences but no major discrepancies. This approach also yielded similar results across time. The fact that PCA and FA return similar results, which are then reflected in weights, could be explained by the fact that, overall, the subdomains are composed of a very limited number of indicators. Indeed, the highest number of indicators in a given subdomain is for Li.5, with seven indicators. Therefore, the PCA correlation matrix closely resembles the off-diagonal FA correlation matrix.

The optimized set of weights allowed us to uncover the relationships among the domains. We could also extend this approach to subdomains. We have found that the correlation among the domains could be explored by decomposing the correlation ratio in two parts, and that these estimates can be further used to reflect weights as importance and not as trade-off ratios.

The weights play a major role in the sensitivity and uncertainty analysis, while the ranking uncertainty is smaller at people level only. Once we evaluated the overall index, we observed higher variability for the middle UTLAs. The UTLAs at top and bottom tend to remain stable. When we

**Figure 9.** The absolute mean rank shift for the overall index by removing one indicator at a time.

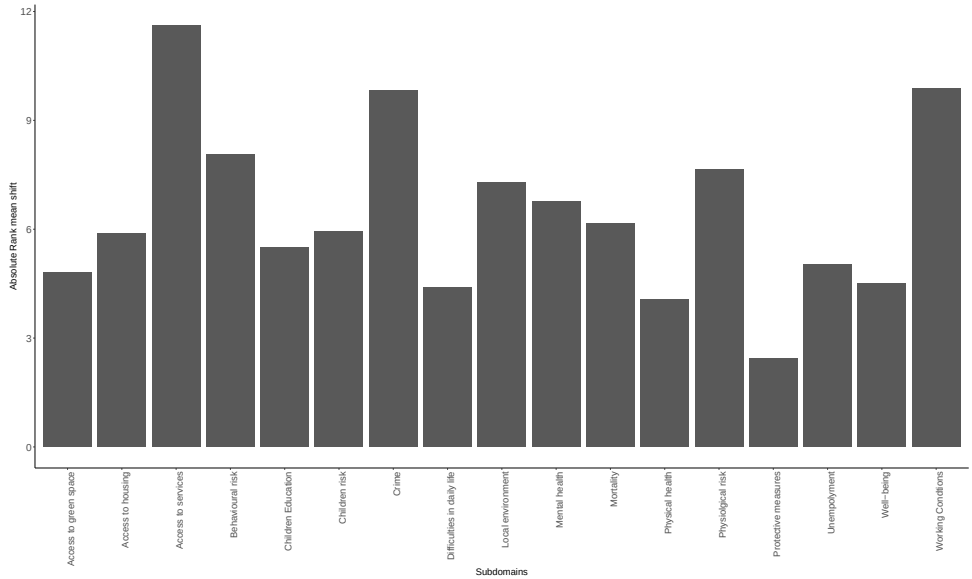


**Figure 10.** The absolute mean rank shift for the overall index by removing one subdomain at a time.

compared the difference between the original ONS ranking and the rankings range based on the modified index, we found these middle UTLAs are most likely to become outliers. This pattern could be a result of using arithmetic mean aggregation. A detailed list of how the proposal has been implemented in the current ONS HI, are reported in online supplementary material, section 4 of the SM.

This information about weights should influence how the HI is used. For example, targeting a policy intervention at the best or worst-performing UTLAs in the HI would be appropriate, because there is greater certainty of these UTLAs' position in the overall distribution. If a policy aimed to improve health for the lower half of the UTLA distribution, the position of those closer to the median is more prone to shift with methodology changes, so it is more likely such a policy would incorrectly target some UTLAs and incorrectly not target some others in need of improvement. Similarly, a UTLA aiming to improve their ranking within the HI should be aware a middling rank is more variable depending on how the HI is constructed.

There are a number of potential aspects of index construction that we have not fully explored in this analysis mainly for the short time series and for the initial scope of this statistical assessment. Given the data spatial feature, we forfeited the spatial correlations in the CI construction steps, nor spatial analysis or effort to evaluate a 'spatial CI' (Fusco et al., 2018; Saib et al., 2015; Siegel et al., 2016; Trogu & Campagna, 2018), that could have provided interesting insights.

Despite the potential bias introduced, we have bypassed the evaluation of alternative imputation methods for the HI. We based our statistical analysis on the linear interpolation and average approaches for missing data. We choose to evaluate other steps in relation to composite constructions as these presented a higher impact in terms of variability at this HI development stage. As per writing, the updated version of the HI 2015–2021 includes an additional geographical layer of Lower Tier Local Authority (approx. 300 areas), and in the future, stratification on population sex and age classes will be the foundation for the HI. Imputation steps will benefit from longer time series, population characteristics and additional geographical layers. The imputation remains an open task for the ONS that will have to be addressed in the near future.

Our approach has not fully explored the indicator's links or association beyond correlation analysis, nor assessed potential causal diagrams. The components of the HI were chosen as either aspects or wider determinants of overall health based on the epidemiological literature, public consultation and the advice of a group of experts from government, the health service, and academia (Ceely, 2023). The design allows users to observe correlations between

individual indicators and the overall index for specific local areas and over time, but cannot in itself prove causality. However, consultations by ONS with public health experts showed that local area HI results do reflect empirically observed health-related issues (ONS internal communication).

In the long term, the HI will provide and guide policy decisions and interventions. However, the indicators carry information that will form causal pathways in addition to the CI. In the CIs literature, the causality topic is often raised but rarely investigated. The debate often centres around defining if the CI—also interpreted as a latent variable—is represented by a reflective or formative model. The first case is similar to causality, which goes from the latent variable to the indicators. The second case assumes that the individual indicators are causing the underlying latent variable: causality goes from indicators to the latent variable. In the latter, changes in indicators will change the HI score for each UTLA (Terzi et al., 2021). It is worth keeping in mind that a change in the indicator does not necessarily lead to a change in the composite indicator and vice versa. For example, countries with high GDP might invest more in technology or more technology might lead to higher GDP (Commission, 2008). However, once longer time series are collected, Granger causality would be tested and other spatial causality tools such as path analysis and Bayesian networks (the probabilistic version of path analysis) could be of some help in studying it.

An aim of the HI is to monitor change in response to interventions. As noted earlier, this argues for a wider range of indicators (to have the potential to show response to intervention targets), although some may be highly correlated. For example, although measures of cardiovascular and respiratory disease tend to be highly correlated, both indicators would be needed separately to monitor response to an intervention targeting one disease type but not the other. Potential causal pathways are inevitably complex and potentially confounded in many cases, and the effect of the same intervention may differ between local areas depending on modifying factors (which may either be HI indicators or unmeasured). The extent to which the HI meets it aims in this respect will be shown by the results of its practical use over time, though testing through simulation of the effects of indicator changes would throw some light. Finally, CIs tend to suffer from the same flaws of observational datasets (Galindo-Rueda, 2019). However, if needed, the microdata that constitutes the spatially aggregated indicators can be unpacked to assess causality, in order to evaluate specific policy interventions.

## 7 Conclusion

In conclusion, the ONS HI (2015–2018) presents a summary of the health of the population of England and fills a gap in policy-making and assessment tools. Our investigation illustrates the methodological choices and trade-offs in the HI as an example of a complex composite measure. We consider both its value and its limitations. The index is based on a hierarchical geographical structure, starting from the UTLA level, rising up to the National level. The composite indicator methodology chosen by the ONS has privileged simplicity, understandability and transparency. The simple arithmetic mean is easy to understand and calculate. Stakeholders can easily see how each individual component contributes to the overall index value.

Hence, the HI provides a detailed and flexible composite measurement that will allow policymakers to assess changes in population health, and to plan interventions by identifying areas and policy domains where interventions can provide significant, quantifiable impact. Future HI editions, with finer geographical granularity and population subgroups, will enrich the understanding of health determinants and guide bespoke interventions and assessments.

## Acknowledgments

A.F.S. is grateful to William Becker for his help and for writing the R-package COINr, to Professor Avi Feller for useful comments and to the Editor and two anonymous referees for insightful comments.

*Conflicts of interest:* None declared.

## Funding

This work was supported by Wave 1 of The UKRI Strategic Priorities Fund under the Engineering and Physical Sciences Research Council Grant EP/W006022/1, particularly the "Shocks and Resilience" cross-theme within that grant & The Alan Turing Institute.

## Data availability

The Health Index datasets (indicator values) which are included are made available by the Alan Turing Institute on behalf of ONS. The underlying data for calculating these indicators' values come from multiple publicly available sources, further information is available at https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandwellbeing/methodologies/-healthindexindicatorsanddefinitions.

Data and code to reproduce the analysis and figures are available at https://github.com/alan-turing-institute/Health-Index-UK.

## Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

## References

Abbott P., & Sapsford R. (1994). Research note: Health and material deprivation in plymouth: An interim replication. *Sociology of Health and Illness*, *16*(2), 252–259. https://doi.org/10.1111/shil.1994.16.issue-2

Akbari K., Winter S., & Tomko M. (2023). Spatial causality: A systematic review on spatial causal inference. *Geographical Analysis*, *55*(1), 56–89. https://doi.org/10.1111/gean.v55.1

Ashraf, K., Ng, C. J., Teo, C. H., & Goh, K. L. (2019). Population indices measuring health outcomes: A scoping review. *Journal of Global Health*, *9*(1), 1–14, 010405. https://doi.org/10.7189/jogh.09.010405

Barclay M., Dixon-Woods M., & Lyratzopoulos G. (2019). The problem with composite indicators. *BMJ Quality and Safety*, *28*(4), 338–344. https://doi.org/10.1136/bmjqs-2018-007798

Becker W. (2021). *Composite indicator development and analysis in R with COINr*. https://bluefoxr.github.io/COINrDoc/.

Becker W., Saisana M., Paruolo P., & Vandecasteele I. (2017). Weights and importance in composite indicators: Closing the gap. *Ecological Indicators*, *80*, 12–22. https://doi.org/10.1016/j.ecolind.2017.03.056

Ben-Michael E., Arbour D., Feller A., Franks A., & Raphael S. (2023). Estimating the effects of a California gun control program with multitask Gaussian processes. *The Annals of Applied Statistics*, *17*(2), 985–1016. https://doi.org/10.1214/22-AOAS1654

Caperna G., & Becker W. (2022). *JRC statistical audit of the European skills index 2022*. JRC Publications.

Ceely G. (2020). *Methods used to develop the Health Index for England: 2015 to 2018* (Technical report). Office of National Statistics.

Ceely G. (2023). *Health Index methods and development: 2015 to 2021* (Technical report). Office of National Statistics.

Commission J. R. C. -E. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. OECD publishing.

Decancq K., & Lugo M. A. (2013). Weights in multidimensional indices of wellbeing: An overview. *Econometric Reviews*, *32*(1), 7–34. https://doi.org/10.1080/07474938.2012.690641

Department of the Environment, T. and the Regions (2000). *Indices of deprivation 2000*, Department of the Environment, Transport and the Regions.

Eurostat (2017). Part 1: Indicator typologies and terminologies. In *Towards a harmonised methodology for statistical indicators*. Publications Office of the European Union.

Freni-Sterrantino A., Ghosh R., Fecht D., Toledano M., Elliott P., Hansell A., & Blangiardo M. (2019). Bayesian spatial modelling for quasi-experimental designs: An interrupted time series study of the opening of municipal waste incinerators in relation to infant mortality and sex ratio. *Environment International*, *128*, 109–115. https://doi.org/10.1016/j.envint.2019.04.009

Freudenberg M. (2003). Composite indicators of country performance. https://www.oecd-ilibrary.org/content/paper/405566708255.

Fusco E., Vidoli F., & Sahoo B. K. (2018). Spatial heterogeneity in composite indicator: A methodological proposal. *Omega*, *77*, 1–14. https://doi.org/10.1016/j.omega.2017.04.007

Galindo-Rueda F. (2019). Oslo manual 2018: Guidelines for collecting, reporting and using data on innovation. In *National bureau of statistics of China, OECD-NBS international training workshop on innovation statistics* (pp. 16–18).

Grad, F. P. (2002). The preamble of the constitution of the world health organization. *Bulletin of the World Health Organization*, *80*, 981–981. https://iris.who.int/handle/10665/268691

Greco S., Ishizaka A., Tasiou M., & Torrisi G. (2019). On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research*, *141*(1), 61–94. https://doi.org/10.1007/s11205-017-1832-9

Hyder A. A., Puvanachandra P., & Morrow R. H. (2012). Measuring the health of populations: Explaining composite indicators. *Journal of Public Health Research*, *1*(3), 222. https://doi.org/10.4081/jphr.2012.e35

Iqbal, J., & Nadeem, K. (2006). Exploring the causal relationship among social, real, monetary and infrastructure development in Pakistan. *Pakistan Economic and Social Review*, *44*, 39–56. https://www.jstor.org/stable/25825283

Jacobs R., Smith P. C., & Goddard M. K. (2004). *Measuring performance: An examination of composite performance indicators: a report for the Department of Health*, Centre of Health Economics, University of York.

JRC (n.d.). Supporting policy with scientific evidence. https://knowledge4policy.ec.europa.eu/publication/your-10-step-pocket-guide-composite-indicators-scoreboards_en.

Kaltenthaler E., Maheswaran R., & Beverley C. (2004). Population-based health indexes: A systematic review. *Health Policy*, *68*(2), 245–255. https://doi.org/10.1016/j.healthpol.2003.10.005

Mara T. A., & Tarantola S. (2012). Variance-based sensitivity indices for models with dependent inputs. *Reliability Engineering and System Safety*, *107*, 115–121. https://doi.org/10.1016/j.ress.2011.08.008

Munda G. (2008). *Social multi-criteria evaluation for a sustainable economy* (*Vol. 17*). Springer.

Munda G., & Nardo M. (2005). Constructing consistent composite indicators: The issue of weights.

Munda G., & Nardo M. (2009). Noncompensatory/nonlinear composite indicators for ranking countries: A defensible setting. *Applied Economics*, *41*(12), 1513–1523. https://doi.org/10.1080/00036840601019364

Nardo, M., Saisana, M., Saltelli, A., & Tarantola, S. (2005). Tools for composite indicators building. *European Comission, Ispra*, *15*(1), 19–20. https://publications.jrc.ec.europa.eu/repository/handle/JRC31473

Pineo H., Zimmermann N., Cosgrave E., Aldridge R. W., Acuto M., & Rutter H. (2018). Promoting a healthy cities agenda through indicators: Development of a global urban environment and Health Index. *Cities and Health*, *2*(1), 27–45. https://doi.org/10.1080/23748834.2018.1429180

R Core Team (2022). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Saib M.-S., Caudeville J., Beauchamp M., Carré F., Ganry O., Trugeon A., & Cicolella A. (2015). Building spatial composite indicators to analyze environmental health inequalities on a regional scale. *Environmental Health*, *14*(1), 1–11. https://doi.org/10.1186/s12940-015-0054-3

Saisana, M., & Philippas, D. (2012). Sustainable Society Index (SSI): Taking societies' pulse along social, environmental and economic issues. *Environmental Impact Assessment Review*, *32*(1), 94–106. https://publications.jrc.ec.europa.eu/repository/handle/JRC76108

Saisana M., Saltelli A., & Tarantola S. (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *168*(2), 307–323. https://doi.org/10.1111/j.1467-985X.2005.00350.x

Saisana M., & Tarantola S. (2002). *State-of-the-art report on current methodologies and practices for composite indicator development* (Vol. 214). Citeseer.

Saltelli A., Annoni P., Azzini I., Campolongo F., Ratto M., & Tarantola S. (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, *181*(2), 259–270. https://doi.org/10.1016/j.cpc.2009.09.018

Siegel M., Koller D., Vogt V., & Sundmacher L. (2016). Developing a composite index of spatial accessibility across different health care sectors: A german example. *Health Policy*, *120*(2), 205–212. https://doi.org/10.1016/j.healthpol.2016.01.001

Sobol, I. M. (1993). Sensitivity analysis for non-linear mathematical models. *Mathematical Modelling and Computational Experiment*, *1*, 407–414. https://doi.org/10.1016/S0378-4754(00)00270-6

Soerjomataram I., Lortet-Tieulent J., Ferlay J., Forman D., Mathers C., Parkin D. M., & Bray F. (2012). Estimating and validating disability-adjusted life years at the global level: A methodological framework for cancer. *BMC Medical Research Methodology*, *12*(1), 1–15. https://doi.org/10.1186/1471-2288-12-125

Sullivan D. F. (1966). Conceptual problems in developing an index of health, number 17. In '1', US Department of Health, Education, and Welfare, Public Health Service.

Terzi S., Otoiu A., Grimaccia E., Mazziotta M., & Pareto A. (2021). Open issues in composite indicators: A starting point and a reference on some state-of-the-art issues.

Townsend P., Phillimore P., & Beattie A. (1988). *Indicator of health and deprivation: Inequality in the north* (pp. 30–40). Croom Helm.

Trogu D., & Campagna M. (2018). Towards spatial composite indicators: A case study on sardinian landscape. *Sustainability*, *10*(5), 1369. https://doi.org/10.3390/su10051369

Trust N. F. (2022). Northumbria local Health Index. https://www.northumbria.nhs.uk/about-us/reducing-health-inequalities/northumbria-local-health-index-tool#2162dd22.

van de Water H. P., Perenboom R. J., & Boshuizen H. C. (1996). Policy relevance of the health expectancy indicator; an inventory in European Union countries. *Health Policy*, *36*(2), 117–129. https://doi.org/10.1016/0168-8510(95)00803-9

Vansnick J.-C. (1990). Measurement theory and decision aid. In *Readings in multiple criteria decision aid* (pp. 81–100). Springer.

Vincke P. (1992). *Multicriteria decision-aid*. John Wiley & Sons.

Wood, S. N. (2001). mgcv: Gams and generalized ridge regression for R. *R News*, *1*(2), 20–25. https://journal.r-project.org/articles/RN-2001-015/RN-2001-015.pdf