2022 Resident Populations: NHSE North East and Yorkshire Analytics Process Document

# Overview

This document describes how the NHS England North East and Yorkshire Analytics Team have calculated resident population profiles broken down by age, sex and deprivation using the ONS 2022 mid-year Lower Super Output Area (LSOA) population estimates.

# Description of Process

## Part One: Aggregating Populations into 5-year age bands

Mid-year 2022 LSOA population estimates were downloaded from the [ONS website](https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/lowersuperoutputareamidyearpopulationestimates) as an Excel file. The populations were then aggregated into 5-year age bands (0-4, 5-9…90+) for males and females. The ‘pivot\_longer’ function from the R ‘tidyr’ package was used to create a new Excel output which had the following columns:

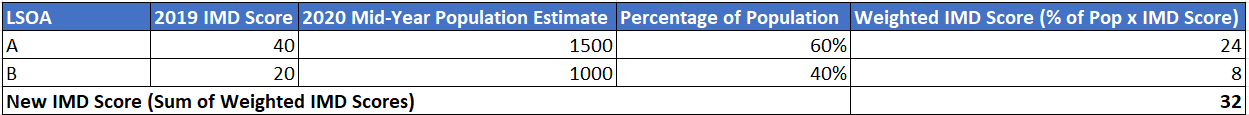
* LSOA 2021 Code
* LSOA 2021 Name
* Mapped 2021 Local Authority Code for the LSOA
* Mapped 2021 Local Authority Name for the LSOA
* 5-year age bands
* Sex
* Population

## Part Two: Converting 2019 IMD Scores to 2021 LSOA Codes

Each LSOA is assigned an Index of Multiple Deprivation (IMD) score as part of the 2019 English Indices of Deprivation publication. However, these are published using the 2011 LSOA codes, whereas the 2022 mid-year population estimates are published using updated 2021 LSOA codes.

To convert the 2019 LSOA IMD scores to the 2021 LSOA configurations, an ONS [best-fit 2011 to 2021 LSOA mapping file](https://geoportal.statistics.gov.uk/datasets/ons::lsoa-2011-to-lsoa-2021-to-local-authority-district-2022-lookup-for-england-and-wales-version-2-1/about) was used. In cases where LSOAs in the 2011 configuration had merged to create a new LSOA for the 2021 configurations, a weighted average of their IMD scores was calculated. Weightings were based on the 2020 mid-year population estimates.

The table below provides a hypothetical worked example. LSOA A has an IMD score of 40, and LSOA B an IMD score of 20. A simple average of the two would give a new IMD score of 30. However, as LSOA A has a larger 2020 mid-year population estimate, it’s score of 40 is given a greater weighting than LSOA B. This leads to the new IMD score being calculated as 32.



After the new IMD scores were calculated, LSOAs were placed into national quintiles and deciles, with a value of 1 indicating the most deprived national quintile/decile.

Deciles and quintiles were calculated using the [NTILE function](https://learn.microsoft.com/en-us/sql/t-sql/functions/ntile-transact-sql?view=sql-server-ver16) in SQL.

## Part Three: Calculating Aggregated Populations

The final stage in the process involved aggregating the LSOA populations and IMD scores up to create outputs for larger geographies. For each of the below geographies, outputs were created showing the population broken down by age, sex and IMD quintile:

* England
* NHS Region
* Integrated Care Board
* Local Authority
* Sub-ICB Location

LSOAs were aggregated up to these higher geographies using an [ONS LSOA (2021) to Sub ICB Locations to Integrated Care Boards Lookup in England](https://geoportal.statistics.gov.uk/datasets/8905a9ad35284b78945c3f3eb30498a2_0/explore) mapping file. An NHS region column was added by mapping ICBs to NHS regions.

## Software Used

Part one of the process was done using the ‘R’ data science tool. For parts two and three, SQL Server Manager Studio (SSMS) was used within the NHS National Commissioning Data Repository (NCDR) environment.

# Outputs

This GitHub repository contains the following outputs:

Input Files

* Links to all the input files used for produce the analysis

SQL and R Script

* R script for aggregating LSOA populations into 5-year age and sex bands.
* SQL file for calculating LSOA 2019 IMD scores mapped to the 2021 configurations.

Outputs

* Excel file containing LSOA populations broken down by 5-year age and sex bands for males and females.
* Excel file containing LSOA 2019 IMD scores mapped to the 2021 LSOA configurations.
* Excel file containing England, NHS Region, ICB, Local Authority and sub-ICB populations broken down by 5-year age, sex and IMD quintile bands. This also includes a ‘Total’ IMD score, which can be used if you just want to view the data by age and sex.
* Excel file showing the number of people who would be in each age and sex band if the England population was 100,000. This can be used to calculate directly age-sex standardised rates per 100,000 population with England acting as the reference population.

The SQL scripts for creating the outputs by the various populations are not included, but these can be shared upon request.

# Application

The outputs of this analysis can be used to profile populations by age, sex and deprivation. Questions that could be asked of the data include:

* What percentage of the population live in areas of high or low deprivation?
* How does the age and sex profile of the population compare to England or other geographies (i.e. is it younger or older)?
* How do the age and sex profiles of areas of high and low deprivation compare to each other?

When profiling populations by age and sex, an effective visualisation of the data could be a population pyramid.

The populations could also be used to calculate age-sex standardised activity rates by IMD quintile. As more deprived populations typically have younger age profiles, accounting for how age might impact a metric (for instance, rates of admissions to hospital) is an important methodological step when making comparisons between areas of high and low deprivation.

It is important to note that another source of population data is GP registered populations, which are published monthly by NHS Digital.

For analysis produced by the NHS, GP registered populations will typically be the most appropriate to use, as these form the basis of ICB allocations. However, NHS Digital does not publish GP Registered populations broken down by age, sex and deprivation (monthly GP registered populations are broken down by age and sex, and once a quarter total GP registered populations are broken down by LSOA, but there are no publications which combine all three views).

For this reason, any analysis requiring a breakdown by all three factors (for instance, age-sex standardised rates by IMD Quintile) will need to be based on resident rather than GP registered populations.

# Caveats

The below are caveats which ONS publish alongside the LSOA mid-year population estimates broken down by single year of age:

## Rounding

The estimates are presented unrounded. Unrounded estimates are published to enable and encourage further calculations and analysis. However, the estimates should not be taken to be accurate to the level of detail provided. More information on the accuracy of the estimates is available in the Quality and Methodology document (QMI). It is ONS policy to publish population estimates rounded to at least the nearest hundred persons. These unit-level estimates are provided to enable and encourage further calculations and analysis. However, the estimates cannot be guaranteed to be as exact as the level of detail implied by unit-level data. We request that data are rounded to the nearest 100 if quoted in any publication, documentation or similar.

## Small Counts

The estimates are produced using a variety of data sources and statistical models, including some statistical disclosure control methods, and small estimates should not be taken to refer to particular individuals.

## Usual residence definitions

The estimated resident population of an area includes all those people who usually live there, regardless of nationality. Arriving international migrants are included in the usually resident population if they remain in the UK for at least a year. Emigrants are excluded if they remain outside the UK for at least a year. This is consistent with the United Nations definition of a long-term migrant. Armed forces stationed outside of the UK are excluded. Students are taken to be usually resident at their term time address.

## Supporting information

These data are made available as supporting information to provide users with 'building bricks' for producing best-fit population estimates for higher levels of geography, and to allow the creation of population estimates for bespoke age groups. Direct use of these OA estimates by single year of age is not recommended as the estimates cannot be guaranteed to be accurate at this level of detail.

## Accuracy

These estimates (including estimates for Lower layer Super Output Areas (LSOAs), Middle layer Super Output Areas (MSOAs), electoral wards, Westminster Parliamentary constituencies, Clinical Commissioning Groups (CCGs) and National Parks) were initially intended for publication by five year age group and sex. The more detailed estimates provided here are intended to enable and encourage further analysis and use of the estimates. Particular caution should be exercised in using estimates at a greater level of disaggregation - for example, for Output Areas, or for single year of age groups, as these would not be expected to have the same level of accuracy as the aggregated estimates.

## Geography

These population estimates reflect boundaries in place as of March 2024. The main products are the estimates for Super Output Areas (SOAs), which are based on the 2021 Census and rolled forward annually using a ratio change methodology. This approach uses the change in the population recorded in the Personal Demographics Service as an indicator of the change in the true population. Estimates for LSOAs by broad ages and MSOAs by quinary age groups (five-year age groups) hold Accredited Official Statistics status. Estimates at greater levels of disaggregation, for example by single year of age, are provided as supporting information only.

The remainder of these population estimates products relate to a range of different geographic areas and are derived directly from the SOA figures. First, estimates for LSOAs are broken down to Output Area (OA) level using an apportionment approach. These OA estimates are then aggregated to produce estimates for electoral wards and Westminster Parliamentary constituencies on a best-fit basis. Estimates for National Parks are also calculated from the OA-level data. Electoral wards, Westminster Parliamentary constituencies and National Parks all hold Official Statistics in Development status. Estimates for health geographies are aggregated directly from LSOAs and hold Accredited Official Statistics status.

## Specific notes for LSOAs

This workbook uses LSOA codes used by ONS from Census 2021.