Office for Health Improvement & Disparities

Atlas of health variation in head and neck cancer in England

Introduction to the data and methods

Published 22 October 2024

Contents

Introduction to the data and methods 3
Statistical significance method 4
Statistical significance interpretation 4
The maps 6
The column chart 7
The regional dot plot 7
The deprivation bar chart 8
The box plot time series 8
Box plot interpretation 9
Box plot time series table 9
Statistical methods 11
Standardisation 11
Confidence intervals 11
References 12

Introduction to the data and methods

The indicators included in this atlas have been chosen to enable local areas to identify unwarranted variation and to determine where local actions might be needed.

For each indicator the atlas presents maps and charts at the smallest appropriate geography, often sub integrated care board (Sub-ICB) or local authority, showing variation in terms of magnitude and geographical location within England, usually for the latest available time period. Time series analyses in the form of repeated box and whisker plots reveals trends in the level and spread of local area values. The standard atlas outputs are explained in this document.

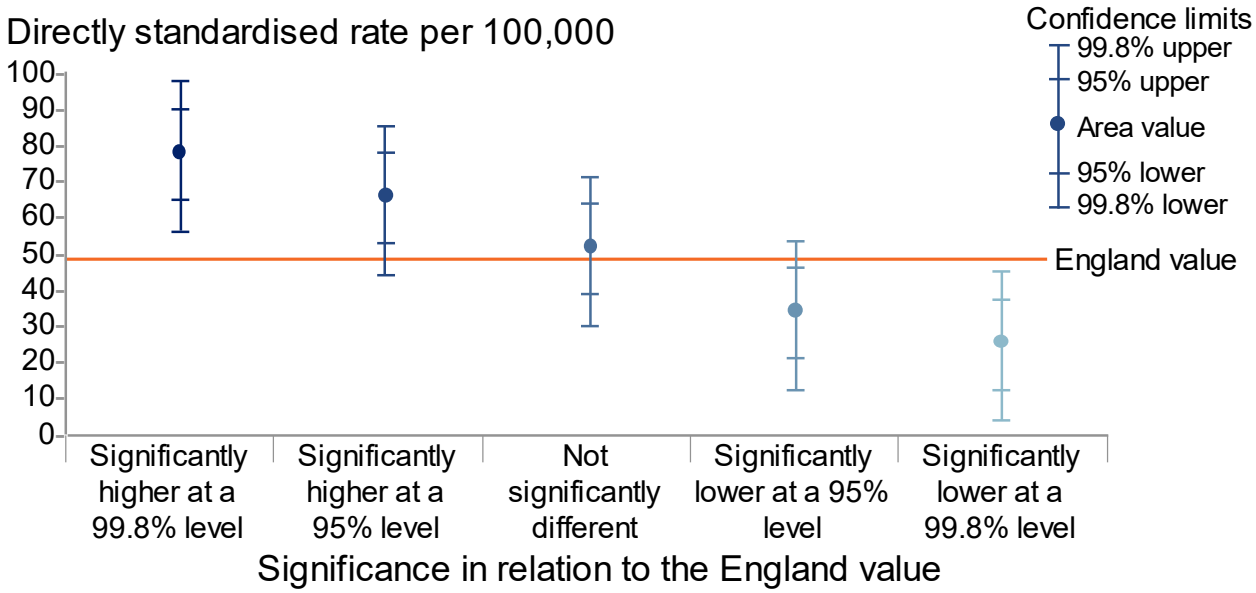
A data sheet with all indicator values and a metadata document with details of the statistical methodology, data extraction coding schemes and data sources for each indicator are available on the Atlas of Variation web page.

Statistical significance method

Many of the atlas outputs use a statistical significance method to compare each area value and confidence interval values to the England value. Each area value is categorised by whether the area value is significantly higher or lower or not significantly different than the England value. The significance is calculated by comparing the confidence intervals of the area value to the England value. Within the atlas two sets of confidence intervals are calculated, 95% and 99.8%, for each area.

The following chart example (Figure 1) shows how different confidence intervals are benchmarked against the England value:

Figure 1



Statistical significance interpretation

The significance band does not indicate whether a high or low value represents good or bad performance, simply whether the indicator value is significantly higher or lower than the England value, and the degree of statistical confidence that the difference is not due to random variation.

In the significance map and charts, the England value is used as the statistical benchmark against which local areas are compared. It is important to note that this does not imply that the England value is the optimal or aspirational level but gives an indication of how each area compares with the England value.

Area values that are not significantly different from the England value are said to display ‘random’ variation alone. Area values that are higher or lower than the England value at the 95% significance level are deemed statistically significantly different. However, as so many area values are being simultaneously tested against the England value, the likelihood of finding values that are significantly different from the England value is raised by chance alone. For this reason, a more stringent 99.8% significance level is also applied. There is much greater certainty that area values found to be different from the England value at the 99.8% significance level are due to a systematic non-random variation that requires investigation.

If many area values are significantly different from the England value at the 99.8% level this could be the result of small confidence intervals which can result from a large denominator. Alternatively, this may be due to overdispersion, characterised by many areas having values at the extremes of the distribution. Overdispersion can occur when there are further factors influencing the values that are not accounted for when calculating the confidence intervals, such as demographic risk factors, case mix or localised service configuration. It is important to consider whether all known warranted factors have been adjusted for when assessing whether the observed variation is unwarranted.

The maps

The standard map presentation in the atlas is two maps shown side by side for each indicator. Typically, this is a quintile map alongside a statistical significance map. Sometimes the atlas displays maps using other methods, such as benchmarking to national thresholds or targets.

The quintile map splits the geographical areas into five equal size groups with 20% of areas in each group. The categories are shaded from darkest (highest values) to lightest (lowest values) on the quintile maps. The legend for the quintile map may overlap with another quintile group. This is because legend values are rounded.

The following image (Figure 2) shows the statistical significance map shades used for each area by whether it is significantly higher or lower or not significantly different than the England value. The two darkest shades are significantly higher than the England value and the two lightest shades are significantly lower than the England value. The middle shade is used for areas not significantly different to the England value.

Figure 2

Shade	Quintile	Significance band
Darkest blue	Highest 20%	Significantly higher than England at the 99.8% level
Second darkest blue	Second highest 20%	Significantly higher than England at the 95% level
Middle blue	Middle 20%	Not significantly different from England
Second lightest blue	Second lowest 20%	Significantly lower than England at the 95% level
Light blue	Lowest 20%	Significantly lower than England at the 99.8% level

When two maps are presented, they will show different approaches to categorising data. It can be useful for local areas to compare the categories they fall into on both maps presented in the atlas.

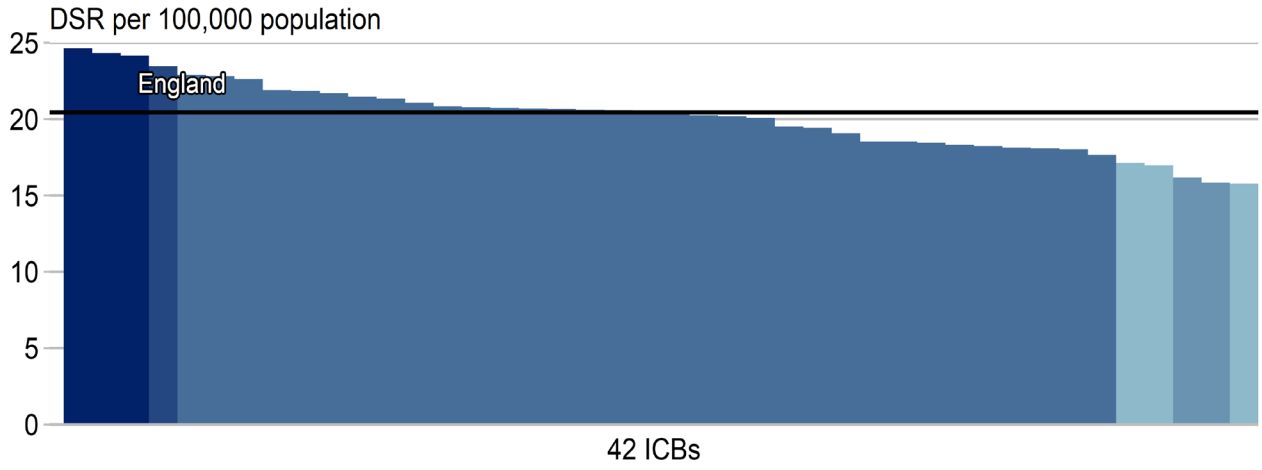
For example, grouping data in quintiles does not take into account the distribution of data. Each quintile can have very different ranges between the highest and lowest values. This should be taken into consideration when comparing areas falling within different categories on the quintile map.

The column chart

The column chart shows the local area values compared to the England value, which is displayed as a horizontal line. The atlas displays the column chart with the same shades as the significance map, if there is no significance map the chart will display the same categories as the quintile or other map presented in the atlas.

Due to the number of areas the column chart does not display the confidence intervals as separate lines on the columns. The chart (see example in Figure 3) uses the significance map colours to show how the area confidence intervals compared to the England rate. The confidence intervals for each indicator and area are available to download in the data sheet.

Figure 3



The regional dot plot

The regional dot plot presents the area values grouped by their NHS England region (sometimes referred to as NHS regions) or statistical region. The plot shows the distribution of area values within each region. The value and colours displayed are the same as the significance map.

The number of Sub-ICBs in an NHS England region varies. NHS England - London only contains 5 Sub-ICBs whereas the NHS England - Midlands contains 19 Sub-ICBs. In NHS England regions with more Sub-ICBs, the potential for the dot plot to display a wider spread is greater and so consideration of context is recommended.

The deprivation bar chart

The deprivation bar charts present the indicators by the values for each deprivation decile or quintile based on the Index of Multiple Deprivation (IMD) 2019. The analysis is conducted at Lower Super Output Area (LSOA) rather than Sub-ICB or local authority level. LSOAs are much smaller geographic areas, with average population sizes between 1,000 to 3,000, this enables deprivation analysis to be conducted at a much more granular level. The indicator values within each LSOA have been allocated to a deprivation decile or quintile and then the value for each decile or quintile is calculated.

The box plot time series

A time series box and whisker plot is presented for each indicator where data is available for a number of time periods. The purpose is to give an impression of the level and spread, or distribution, of the area values. The plots show the indicator median value and spread of area values across England at consecutive time points. Importantly, the tables accompanying the plots show whether there has been any statistically significant change in the median, or in the degree of variation of values over time.

The plots use a methodology which is unrelated to the method determining the significance map and bar chart shading, they do not represent statistical significance.

The ‘box’ runs from the upper quartile (75th percentile area value) to the lower quartile (25th percentile area value) and represents the middle 50% of area values. The height of the box is known as the interquartile range (IQR) and is calculated as upper quartile value minus the lower quartile value.

Inside the box is a horizontal line, which shows the median. The median is the middle area value of the dataset with half of area values above and half below it. The median is different from the value for England, the more skewed the distribution of data the greater the difference between the median and the England value.

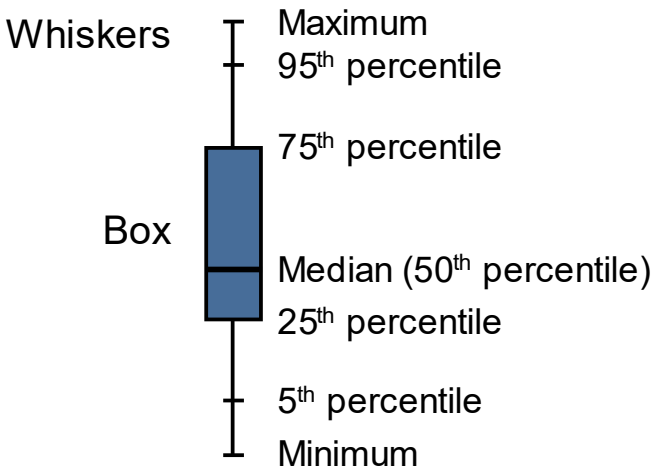
The ‘whiskers’ extend out from either end of the box and show the highest (maximum) and lowest (minimum) values within the dataset. The 95th percentile and the 5th percentile are also represented by tick marks on the ‘whiskers’.

Box plot interpretation

A box and whisker plot presents information about the shape or spread of the area values and whether they have a symmetric or skewed distribution. A dataset with a normal distribution is symmetric (non-skewed) around the mean (average), the mean and the median are equal, and each half of the distribution is a mirror-image of the other half. In a distribution that is skewed there is a lack of symmetry between the upper and lower halves of the dataset, the median and the 'box' is not centrally located between the maximum and minimum.

The following example (Figure 4) shows the different aspects displayed on the box and whisker plot.

Figure 4



Box plot time series table

Presented below the box plot is a table of statistics summarising trend.

Max–min (range): The difference between the highest and lowest value. However, extreme outliers can heavily influence this statistic and consequently mislead about the extent of variability across the dataset. It may be more helpful to use the 95th to 5th percentile.

95th–5th percentile: This shows the range of the data between the 95th and the 5th percentile of the dataset; can be used instead of the max-min range if there are extreme outliers in the dataset.

75th–25th percentile: Known as the interquartile range (IQR). These percentiles are the upper and lower limits of the middle 50% of area values. If the IQR is small it indicates that the central 50% of data values are close to the median; if the IQR is large it indicates that the data is spread out from the median and there is more dispersion in the middle 50% of values

Median: The middle value identified by arranging the values in ascending order from the lowest to the highest. If there is an even number of values the median will be the average of the 2 central area values. The median is not as influenced by outliers as a mean value.

The significance column of the table is a summary of whether each of these ranges is narrowing or widening, or the median is increasing or decreasing, and whether the trends are statistically significant at the 95% level. The statistical significance was determined using a 2-tailed t-test on the slope of a linear regression line fitted to the values in the table over time, where the null hypothesis is that the slope equals zero. The significance test is only performed for indicators with three or more time periods. This regression line and the detailed results of the t-test are not presented in this atlas.

Statistical methods

Indicators are based on established methods for calculating statistics within public health intelligence. This section provides further information on the methods typically used within the atlas to enable areas to be compared robustly.

Standardisation

Differences in the rates of disease can be strongly related to the population age structure. Populations with higher proportions of older people can be expected to have higher incidence or mortality rates. To identify variation that is not related to age structure, a technique called standardisation is used. This enables the level of testing to be compared between populations with different age structures producing a more level playing field. There are 2 main methods of calculating age-standardised rates:

• direct standardisation

• indirect standardisation

Only direct standardisation has been used within this atlas and so only this method is discussed here. Directly age-standardised rates adjust for the differences in age distribution in a population by weighting the observed rate in each age band by a standard population (in this case the European Standard Population) and are usually expressed, for example, as a rate per 100,000 population.

Confidence intervals

Confidence intervals are used to represent the level of uncertainty of an area value. Statistical uncertainties usually arise because indicators are based on a subset from the population of interest or over a defined time period, both of which may not be representative of the whole population. A smaller confidence interval indicates that the value is more reliable, and a larger confidence interval indicates that the value is less reliable. Confidence intervals are used within the atlas to compare the local area value against the England benchmark value in a similar way to performing an exact statistical test. Confidence intervals were used to determine the shading in the significance maps and charts. The two main methods of calculating confidence intervals in this atlas are:

• the Wilson score method for proportions[1] [2]

• the Dobson and Byar's methods for rates[2] [3]

References

[1] Wilson EB (1927 published online 08 May 2012) 'Probable Inference, the Law of Succession, and Statistical Inference' Journal of the American Statistical Association, 1927, volume 22, issue 158, pages 209 to 212 [Accessed 30 May 2024]

[2] Department of Health and Social Care 'Public Health Technical Guidance: APHO Technical Briefing 3 – Commonly used public health statistics and their confidence intervals' [Accessed 11 October 2024]

[3] Breslow NE, and Day N E (1987) Statistical methods in cancer research. Volume II--The design and analysis of cohort studies, IARC scientific publications, volume 82, pages 1 to 406

Office for Health Improvement and Disparities