Public Health
England

Protecting and improving the nation's health

Technical Guide

RAG Rating Indicator Values

Introduction

This document sets out Public Health England's standard approach to the use of RAG ratings for indicator values in relation to comparator or benchmark values.

RAG (Red-Amber-Green) ratings, also known as 'traffic lighting,' are used to summarise indicator values, where green denotes a 'favourable' value, red an 'unfavourable' value and amber a 'neutral' value.

These colours are used in different visualisations. One method is in a matrix, typically referred to as a 'tartan rug.' This matrix usually has columns for each area or organisation, and rows for each indicator. Each cell then has a background colour of red, amber or green which is determined by the indicator value, with a contrasting text colour.

The same methodology is also used to highlight favourable and unfavourable values in other presentations, for example on spine charts[1] or in key messages for profiles.

There are different fundamental approaches to assigning RAG ratings:

1. arbitrary or subjective criteria or targets
2. statistical significance

Approach 1 is common in performance management. For example, some performance monitoring systems can be set to 'trigger' at +/– 10%, say, or when a fixed target is missed, without regard to the statistical properties of the indicator being measured.

[1] https://fingertips.phe.org.uk/documents/PHDS Guidance - Spine Charts.pdf

This document focuses on the second approach: definitions should be based on statistical comparisons with a comparator; indicator values should only be given red or green RAG ratings if they are statistically significantly higher or lower than the comparator.

However, there are circumstances where additional criteria might be applied in addition to statistical significance (see Additional criteria below).

Terminology and colouring conventions

Usually, when comparing a value to a comparator, red corresponds to a value that is significantly 'worse,' green to a value that is significantly 'better,' and amber indicates that there is no significant difference.

When the polarity of the indicator (that is higher values are 'worse' or higher values are 'better') is unequivocal the terms 'better' and 'worse' should be used, and the colours red, amber and green are appropriate. Metadata should define clearly the polarity of the indicator, as well as detailing the data sources and methodology. This approach should be used whenever it is appropriate as it makes results a lot easier to interpret.

Where it is inappropriate to label high or low values as 'better' or 'worse,' for example proportion of ethnic minorities, or the proportion of pregnancies terminated, the terms 'higher' and 'lower' should be used with neutral colouring such as shades of blue, from light to dark. Labelling must not imply that high values of such indicators, for example, are 'worse.'[2]

Comparative methods

There are a number of methods that are commonly used for making the comparisons required for RAG rating. These include use of:

- confidence intervals (CIs)
- statistical process control (SPC)
- hypothesis testing

Confidence intervals and SPC are derived from summary statistics (for example, mean values, standard deviations, counts) whereas often hypothesis (statistical) tests may require calculations from raw data. However there are cases when using confidence intervals is inappropriate, cases when SPC is not applicable and cases where neither is appropriate.

[2] Some PHE profile publications have used neutral colouring for all indicators but this is confusing and not recommended.

Whether each approach is appropriate will be determined by the following factors:

1. the type of comparator
2. the type of indicator

The table below summarises which methods are likely to be appropriate for different indicator and comparator types. Note that there will be exceptions to these generalisations. The types are explained in the sections that follow.

	Type of comparator								
Type of indicator	**Fixed reference value**			**Reference population value**			**Reference sample value**		
Counts/rates	🟩 CIs	🟩 SPC	🟩 Test	🟧 CIs	🟧 SPC	🟩 Test	🟥 CIs	🟥 SPC	🟩 Test
Proportions	🟩 CIs	🟩 SPC	🟩 Test	🟧 CIs	🟧 SPC	🟩 Test	🟥 CIs	🟥 SPC	🟩 Test
Normal variates	🟩 CIs	🟩 SPC	🟩 Test	🟧 CIs	🟧 SPC	🟩 Test	🟥 CIs	🟥 SPC	🟩 Test
Others (for example, life expectancy, synthetic estimates etc.)	🟩 CIs	🟥 SPC	🟥 Test	🟧 CIs	🟥 SPC	🟥 Test	🟥 CIs	🟥 SPC	🟥 Test

Key

Red: method probably not applicable or appropriate
Amber: method not ideal, but acceptable under many circumstances
Green: method probably appropriate

Types of comparator

Broadly there are three types of comparator:

1. A fixed reference value (for example, 95% target for screening coverage)
2. A reference population indicator value (for example, England life expectancy at birth)
3. A reference sample comparator (for example, comparison of mortality rates for a geography with a baseline time period, or comparison of areas with peers)

A key characteristic of these different comparator types is whether the comparator value is treated as being a fixed parameter value (as in 1), or whether it has inherent uncertainty, that is, its own confidence interval (as in 3).

A reference population comparator (2) is typically treated as being a fixed value, particularly when the reference population is much larger than that of the area being compared, so its confidence interval is very much narrower. Most of the uncertainty derives from the small sample value, so it is convenient to ignore the uncertainty in the reference population value. This is reasonable, when necessary, although it should be described carefully. In particular, it is generally not appropriate to state that, with regard to an indicator, one sub-geography is significantly better or worse than England, when that geography lies within England. Rather the indicator should be described as being significantly higher or lower than the mean (or median, for example, as appropriate) for England.

Values of indicators being compared should be independent of each other. For the example in the previous paragraph, the England value is not completely independent of the sub-geography being compared, but if the sub-geography is small, its influence on the England value is very small and so can be ignored. Another example is overlapping 3 or 5-year periods, where sequential time periods share two-thirds or four-fifths of the data they are derived from: these can never be compared using methods that assume independence – they require more complex methods and should be avoided: compare only distinct time periods, for example, 2001-2005 can be compared with 2006-2010, but 2005-09 cannot.

However, in some cases where samples are not independent a simple alternative approach can be used. For example, if observations relate to a fixed group of subjects (or organisations, for example) in two separate time periods then it is not usually appropriate to assume that the observations are independent. In this case it may be appropriate to compare the increase (or reduction) in the indicator value between the two periods for each subject. This then becomes a comparison with a fixed reference value of zero.

In the following discussion types 1 and 2 (as defined above) are considered jointly as relating to a fixed reference value, and type 3 separately as relating to comparison of two variables.

Types of indicator

Different methods are required according to the type of value that an indicator can take. Generally, indicators can be:

- counts and rates (for example, hospital admissions, admission rates or mortality rates)
- proportions (for example, prevalence of a condition or percentage low birthweight)
- normal variates (any indicator can typically be treated as a normal variate if it is an average of a sufficiently large sample of events/cases)
- others (for example, indicators whose underlying distributional properties are complex, for example, synthetic estimates based on survey data, life expectancy, and so on)

Indicator types are often assumed to have a particular statistical distribution. For example, continuous indicators may be normally distributed, and even if the underlying distribution is non-normal, they can be treated as approximately normally-distributed if they are based on a large enough sample, or if the number of cases is large (the central limit theorem of distribution theory states that any distribution tends to the normal distribution as the sample becomes large). For many public health indicators, the distributional properties are simple and known, for example, counts (and rates that are counts divided by a population) are commonly assumed to be Poisson-distributed, and proportions are assumed to be binomially distributed.[3] APHO Technical Briefing 3: Common Public Health Statistics and their Confidence Intervals[4] describes the methods used to calculate confidence intervals for these types of indicator, based on the appropriate underlying distributions. It is important to understand, and verify, such assumptions, as they have an impact on the assessment of statistical significance, and so may affect an indicator's RAG rating. Appendix 1 deals with each type of indicator in a little more detail.

Some indicators are derived from more complex distributions. For example, life expectancy has no simple underpinning distribution, as it is the result of combining death rates for different age groups using a life table. Confidence intervals can be calculated by making a normal approximation, but life expectancies for different areas cannot be compared using significance tests or SPC approaches. The same is true for synthetic estimates based on models which combine multiple predictor variables, each of which has its own uncertainty. Significance tests however may be possible using more advanced techniques such as multiple regression or bootstrapping/simulation.

Confidence intervals

Confidence intervals are associated with an estimate obtained from data, such as an indicator value. A confidence interval is defined in terms of a confidence level (often denoted as $(1-\alpha) \times 100\%$, where α is the 'significance level') and calculated using observed data. The interval is described by a lower confidence limit (LCL) and an upper confidence limit (UCL).

RAG ratings are defined as red or green if a fixed reference value falls above the UCL or below the LCL respectively, and amber otherwise.

In the case of a comparator which has uncertainty itself, confidence intervals cannot simply be used. If two confidence intervals do not overlap then it is accurate to state that the indicators differ at the appropriate significance level. However, where confidence intervals do overlap it is does not follow that there is no evidence that the indicator values differ significantly. In such a case hypothesis testing can often be used (see example under Hypothesis testing below).

[3] Departures from these assumed distributions should be checked though, as these could lead to bias and/or over- or under-dispersion.

[4] https://fingertips.phe.org.uk/documents/APHO Tech Briefing 3 Common PH Stats and CIs.pdf

Definitions of confidence intervals for many public health indicators are given in APHO Technical Briefing 3: Common Public Health Statistics and their Confidence Intervals.[5]

Statistical process control

Alternative, but generally equivalent methods are frequently available through statistical process control (SPC). Rather than comparing the reference value with confidence intervals around an indicator value, SPC compares an indicator value with 'control limits' around a reference value. Indicators within control limits are considered to exhibit 'common cause' variation only and are classified as amber.

Indicator values that fall outside of control limits are considered to be a result of 'special cause' variation (that is, they are different as the result of underlying causes, such as socio-economic or other determinants) and are classified as red or green.

Control limits are often defined as multiples of the standard deviation (σ) about the reference value (μ).[6]

Control limits vary according to some characteristics, such as population size. Control limits and data points in such a case can be plotted against this characteristic in the form of a funnel plot.

Further details of SPC methods are given in APHO Technical Briefing 2: Statistical process control methods in public health intelligence.[7]

Hypothesis testing

Hypothesis testing can be used to test for significance relative to a fixed reference or a variable comparator, with RAG ratings being set as red or green if the observed indicator value falls in the upper or lower critical region, and amber if no significant difference is found. However in the case of the fixed reference confidence interval and SPC methods are generally equivalent to hypothesis tests, and are often simpler to implement and so are not described here. Instead the focus here is on comparisons with a variable comparator.

Appendix 1 provides more information about hypothesis tests for different indicator types.

[5] https://fingertips.phe.org.uk/documents/APHO Tech Briefing 3 Common PH Stats and CIs.pdf

[6] Occasionally five rating bands are defined in terms of whether the indicator is less than the reference $\mu-3\sigma$, between $\mu-3\sigma$ and $\mu-2\sigma$, between $\mu-2\sigma$ and $\mu+2\sigma$, between $\mu+2\sigma$ and $\mu+3\sigma$, or greater than $\mu+3\sigma$.

[7] https://fingertips.phe.org.uk/documents/APHO Tech Briefing 2 SPC Methods.pdf

Additional criteria

In the introduction, we state that RAG ratings “should be based on statistical comparisons with a comparator; indicator values should only be given red or green RAG ratings if they are statistically significantly higher or lower than the comparator.” While this is true (that is, red or green flags should not normally be used unless the differences are statistically significant), there are times when additional criteria might be applied.

Values of indicators based on common events in large populations can be highly statistically significant even when the absolute differences from the comparator are very small. For example, comparing total admission rates for whole Clinical Commissioning Groups or local authorities, where a rate may appear almost on the centre line of a spine chart, but is significantly higher or lower. In fact, sometimes the values for almost all areas being compared are significantly different from the comparator. This is not incorrect – it is a statement that they are not varying randomly: they vary as a result of systematic influences, for example, socioeconomic conditions or other determinants.

However, it may be desirable only to focus on a subset of those significant values, for example, to focus attention on the largest differences. Setting such criteria is essentially subjective, unless there is scientific justification for ignoring smaller differences: in clinical trials, the clinician has to set a level of difference which is ‘clinically significant,’ that is, the minimum difference that would matter for patient outcomes or decisions on treatment. It is hard to see the equivalent in measuring population health outcomes, but it would be along the lines of ‘even if 1% excess deaths is statistically significant, no-one will change policy as a result of such a small change.’ Whether this is justifiable is dubious. However, having spine charts with red and green dots hovering around the centre line, or having a RAG chart with almost identical values coloured red and green respectively, may be considered confusing.

One possible use of such additional criteria is when targets are involved: for example, if the comparator is a target range, say, ‘the local area should be within +/–5% of the national average’ – even if the local area is significantly worse than the national average, if it is within the target range it may be inappropriate to flag it as red.

If, for any reason, a minimum threshold difference is set, for example, only differences that differ from the reference value by more than 5% and are statistically significantly different are flagged red or green, then this criterion should be stated clearly and the reason for it should be documented.

Checklist

In defining RAG ratings there are a number of considerations. These include:

1. The ‘polarity’ of the indicator: is higher or lower better?
2. What type of comparator is being used?
3. What is the indicator type?
4. Are all assumptions and definitions clearly stated?
5. Are all definitions and parameters appropriately documented?
6. Is the method used appropriate? (For example, is a distributional assumption or an assumption of independence valid.)

APPENDIX 1 – Technical details

Normal variates

When comparing indicators that are assumed to be independent and normally distributed with values $\overline{X}_1$ and $\overline{X}_2$ and standard deviations σ_1 and σ_2 respectively it is possible to use the Wald test. The Wald statistic $\frac{\overline{X}_1-\overline{X}_2}{\sqrt{\sigma_1^2+\sigma_2^2}}$ is compared with the standard normal distribution.

Thereby it is possible to determine whether any difference in the indicator value is statistically significant at the α significance level.

For example, if $\overline{X}_1$ = 10.0 and $\overline{X}_2$ = 10.5 and σ_1= 0.1 and σ_2 = 0.2 then the Wald statistic is calculated as –2.24 which is significant at the α = 0.05 level. However if confidence intervals are calculated, the confidence interval for $\overline{X}_1$ is (9.80, 10.20) and the confidence interval for $\overline{X}_2$ is (10.11, 10.89). That is, the confidence intervals overlap. This illustrates the danger of using confidence intervals to assess difference with a variable comparator.

A refinement of the Wald test is possible where $\overline{X}_1$ is the mean of N_1 normally distributed observations and $\overline{X}_2$ is the mean of N_2normally distributed observations. In this case the Wald statistic corresponds to a Welsh *t*-test statistic, and rather than being compared to the standard normal distribution it should, more accurately, be compared with the *t*-distribution with $\frac{(\sigma_1^2+\sigma_2^2)^2}{\sigma_1^4/(N_1-1)+\sigma_2^4/(N_2-1)}$ degrees of freedom. (Note that σ_1 and σ_2 are the standard deviations (or standard errors) of $\overline{X}_1$ and $\overline{X}_2$ and not the standard deviations of the underlying observations.) When N_1 and N_2 are large the *t*-distribution is well approximated by the standard normal distribution.

Alternative non-parametric methods such as the Mann-Whitney U test and the Kolmogorov-Smirnov test are available where distributions cannot reasonably be assumed to be normal. Such methods are not described in detail within this guidance document.

Counts and rates

Hypothesis tests for counts and rates are also possible. More general techniques, such as Poisson regression, can be applied to compare two groups and it is also possible to use a Likelihood Ratio Test (LRT) which makes parametric assumptions. Alternatively, non-parametric methods can be used, as described above.

Proportions

Proportions can be compared between two (or more) groups using logistic regression. It is also possible to compare proportions between two groups by tabulating results into a 2x2 table in terms of outcome by group. Such a table can then be analysed using a chi-square test for homogeneity with Yates' continuity correction. Alternatively, where there counts are relatively low, an exact test may be possible.

Power and type I error

Indicator values are random variables. That is, an indicator reflects 'common cause', or random variation, as well as 'special cause', or systematic variation. The power of a test, or RAG rating, reflects the likelihood that an indicator will be red or green given that there is some special cause variation. This depends on many things, including the test being used, the size of the special cause effect, the number of observations underlying the indicator and the variability of these underlying observations. Ideally power will be high, and indicators will be correctly classified as red or green in such cases, rather than being flagged as amber.

However, if there is no special cause variation, and the only difference between an indicator and its comparator is random, there is still a risk of the indicator being misclassified as red or green. This is known as a type I error, or a 'false positive'.

With 95% confidence there is a 5% risk that each such indicator is misclassified; that is, one in twenty red or green ratings could be expected even where all the underlying distributions are essentially amber.

This problem of type I errors is particularly important to consider when a number of RAG ratings are being produced, for example on a 'tartan rug.' Such a scenario is often referred to as 'multiple testing'. Methods, such as Bonferroni correction, are possible which effectively shrink the size of critical regions representing red and green ratings. The effect of this is unfortunately also to reduce power. Alternatively, it is important that users of RAG ratings understand the risk of type I errors and the danger of over-interpreting relatively small numbers of red or green indicators.

APPENDIX 2 – Colouring conventions

There are two particular issues to bear in mind when choosing colours for RAG ratings:

1. accessibility guidelines for colour blindness: the choice of red, amber and green should be checked using a tool such as Vischeck.[8]
2. monochrome printing: again, the choice of colours should be checked to ensure they are distinguishable when printed in black and white.

The example colours here are used in Fingertips. The red is clearly distinguishable from the others for those with colour blindness and for black and white printing (as long as a key is provided) as it is much darker than the amber or the green, but the amber and green are harder to differentiate.

Colour	RGB – screen colours	Nearest Pantone	Display
Red	R=192, G=0, B=0	485 2X	🟥
Amber	R=255, G=192, B=0	116 2X	🟧
Green	R=146, G=208, B=80	367	🟩

PHE Technical Guides

This document forms part of a suite of PHE technical guides that are available on the Fingertips website: https://fingertips.phe.org.uk/profile/guidance

[8] http://www.vischeck.com/vischeck/