

Guide to benchmark reports

Benchmark reports are produced for most surveys to show how the survey results for each trust participating in a particular survey compares with the results from all other trusts.

This guide is divided into five sections:

Section one: describes the benchmark reports

Section two: describes how to use the benchmark reports and the limitations of the data

Section three: describes how to understand the data

Section four: provides guidance on using the benchmark reports to make comparisons between trusts

Section five: describes how the data in the benchmark reports is calculated

1.) Description of the reports

The graphs included in the reports display the scores for a trust, compared with national benchmarks. Each bar represents the range of results for each question across all trusts that took part in the survey. In the graphs, the bar is divided into three sections:

- the red section (left hand end) shows the scores for the 20% of trusts with the lowest scores
- the green section (right hand end) shows the scores for the 20% of trusts with the highest scores
- the orange section (middle section) represents the range of scores for the remaining 60% of trusts

The score for a trust is represented by a white diamond. If the diamond is in the green section of the bar, for example, it means that the trust is among the top 20% of trusts in England for that question. The line on either side of the diamond shows the amount of uncertainty surrounding the trust's score, as a result of random fluctuation. These are known as lower and upper confidence intervals. Please see section three below for more detailed information about confidence intervals.

An example of a set of graphs from a benchmark report for trust X can be seen in chart one below. For the first question (When you had important questions to ask the doctor, did you get answers that you could understand?) it can be seen that nationally, scores varied between 75 (the lowest score) and 91 (the highest score). Trust X scored 79 for this particular question. The lower confidence interval is 75 and the upper confidence interval is 83. The threshold score for the 20th percentile is 79 and for the 80th percentile is 84.

Chart 1

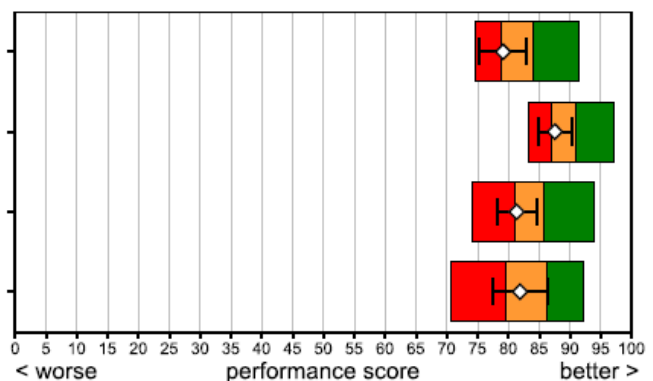
Doctors

When you had important questions to ask a doctor, did you get answers that you could understand?

Did you have confidence and trust in the doctors treating you?

Did doctors talk in front of you as if you weren't there?

As far as you know, did doctors wash or clean their hands between touching patients?



2.) How the benchmark reports should be used

Benchmark reports should be used to identify how a trust is performing in relation to all other trusts that took part in the survey. From this, areas for improvement can be identified.

Limitations of the data

Because the average scores for each trust are estimates based on a sample of patients rather than *all* patients at the trust, it is very often impossible to separate the performance of trusts. That is, in many cases the differences between trusts' mean scores will not be *statistically significant*. This means that if we were to repeat the survey with a different sample of patients, we would not be confident that the results would show the same differences. As such, data used in the benchmark reports is fundamentally **not** suitable for generating league tables.

Also, it should be noted that the data only show performance relative to other trusts: there are no absolute thresholds for 'good' or 'bad' performance. Thus, a trust may score lowly relative to others on a certain question whilst still performing very well on the whole. This is particularly true on questions where the majority of trusts score very highly.

3.) Understanding the data

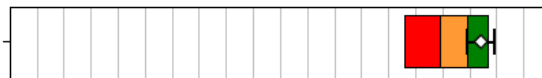
Since the score is based on a sample of patients in a trust rather than *all* patients, the score may not be exactly the same as if everyone had been surveyed and had responded. To account for this, a 'confidence interval' is calculated. For each trust, then, the benchmark report shows three values for each question – an average ('mean') score as well as its lower ('lci') and upper ('uci') confidence limits.

A confidence interval is calculated as an indication of the range within which the 'true' score would lie if all patients had been surveyed. The confidence interval gives upper and lower limits of a range within which you have a stated level of confidence that the 'true' average lies. These are commonly quoted as 95% confidence intervals, which is the level used in the benchmark reports. They are constructed so that you can be 95% certain that the 'true' average lies between these limits.

For example, chart 2, below, show a trust's score for a question asking if respondents had enough time to discuss their health or medical condition with a doctor or nurse. Trust X has an average score of 87, with a lower confidence limit of 84 and an upper confidence limit of 90. This means that we can be 95% confident that the 'true' trust score lies between 84 and 90.

Chart 2

Did you have enough time to discuss your health or medical problem with the doctor or nurse?



The width of the confidence interval gives some indication of how cautious we should be; a very wide interval may indicate that more data should be collected before any firm conclusions are made. In the example above, the confidence intervals are relatively small. In chart three below, the confidence intervals are much wider: the trust has a score of 89, with a lower confidence interval of 83 and an upper confidence interval of 95. The confidence intervals are wider here as fewer people responded to this question (as it was only answered by respondents who were given medication to take home).

Chart 3

Did a member of staff explain the purpose of the medications you were to take at home in a way you could understand?

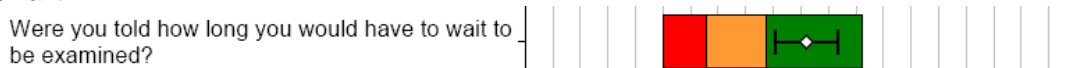


When considering how a trust performs, it is very important to consider the confidence interval surrounding the score. In chart three above, it can be seen that the trust's average score falls into

the orange section of the graph – however, the lower confidence limit falls into the red section of the graph and the upper confidence limit falls into the green section of the graph. This means that you should be more cautious about the trust’s result because, if the survey was repeated with a different random sample of patients, it is possible that their average score would be in a different place and would therefore show as a different colour.

Only if a trust’s average score *and* confidence intervals are completely in one section of the graph can you be confident that that is how a trust is performing. An example is below (chart 4) where it can be seen that the score for trust X and its lower and upper confidence intervals are all shown in the green. This means that we can be confident that this trust’s ‘true’ score is in the top 20% of observed scores for this question.

Chart 4



In summary:

- If a trust’s average score *and* both of its confidence limits appear in green, you can be confident that the trust’s ‘true’ score is in the top 20% of all observed scores.
- If a trust’s average score *and* both of its confidence limits appear in red, you can be confident that the trust’s ‘true’ score is in the bottom 20% of all observed scores.
- If a trust’s average score is in one colour, but either of its confidence limits are shown as falling into another colour (for example, see chart three above) this means that you should be more cautious about the trust’s result because, if the survey was repeated with a different random sample of patients, it is possible their average score would be in a different place and would therefore show as a different colour.
- For example, if a trust’s average score is in red but their upper confidence limit is in orange or green, we **cannot** say that the trust’s ‘true’ score is in the bottom 20%.

This final point has important implications. By definition, one in five trusts will have an average score that is in the bottom 20% of all trusts. For many of these trusts, however, we cannot be confident that if we repeated the survey with a different sample they would again come out as being in the bottom 20%. The only instances where we *can* be confident that a trust’s ‘true’ average lies beneath this threshold are where both their average score *and* their upper confidence limit appear in red.

4.) Comparing scores between trusts

The confidence intervals make it possible to determine if the results from two trusts are significantly different. If the ranges for two trusts overlap then there is no significant difference between the trusts: we cannot be confident that the difference in the average scores does not simply result from random variation. If there is no overlap in the scores of two trusts, then we can be confident that the results for the two trusts are genuinely different.

For example, if trust A had an average score of 70 for a particular question, with a lower confidence limit of 60 and an upper confidence limit of 80; and trust B had a score of 80, with a lower confidence limit of 70 and an upper confidence limit of 90, then the two averages scores **are not** significantly different as the confidence intervals overlap. This is illustrated in table 1 (below). By contrast, if trust C had an average score of 70 for a question, with a lower confidence limit of 66 and an upper confidence limit of 74; and trust D had a score of 80, with a lower confidence limit of 76 and an upper confidence limit of 84, then the two scores **are** significantly different as the confidence intervals **do not** overlap. This is illustrated in table 1.

Tables 1 (left) and 2 (right):

Trust	qX_lcl	qX_mean	qX_ucl
A	60.00	70.00	80.00
B	70.00	80.00	90.00
Difference is not statistically significant.			

Trust	qX_lcl	qX_mean	qX_ucl
C	66.00	70.00	74.00
D	76.00	80.00	84.00
Difference is statistically significant.			

Chart five (below) shows the ranges between the upper and lower confidence limits for all trusts for a question asking if it was possible to find a convenient place to park in the hospital car park. This is shown as a line for each trust that shows the area in which we can be 95% confident that their 'true' score lies. As this question is only answered by those who travelled to hospital by car, the confidence intervals are quite wide for most trusts.

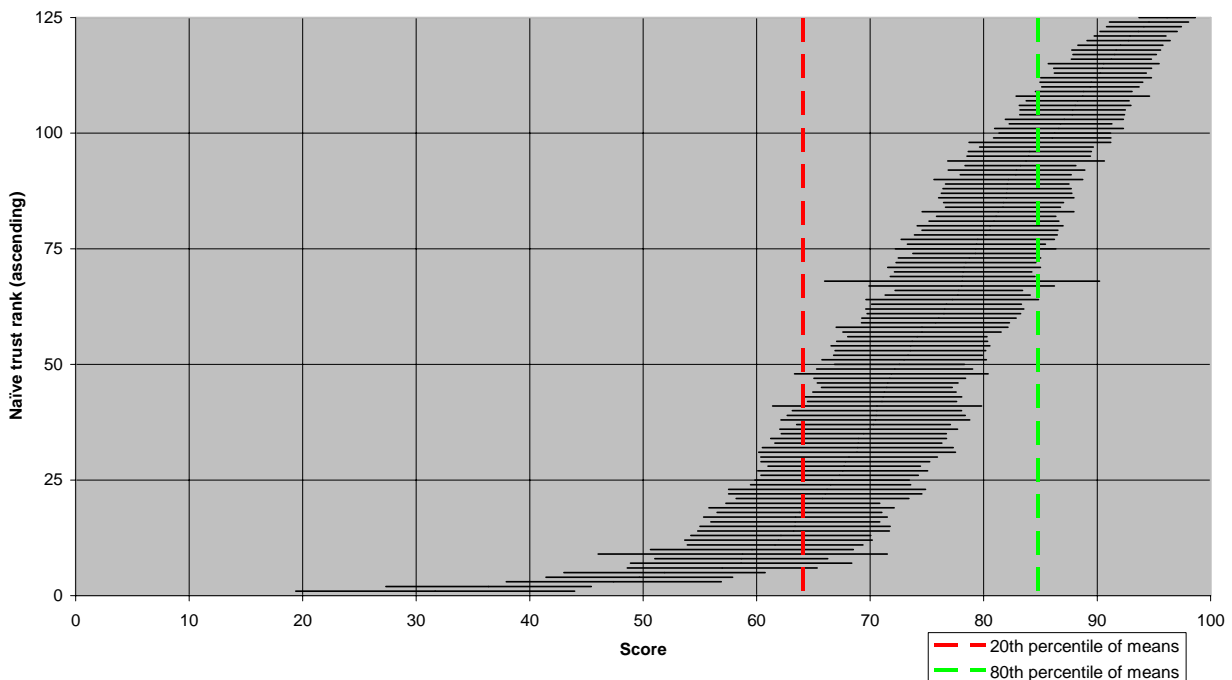
The red line indicates the 20th percentile score, below which a value may be considered to be within the lowest 20% of trusts' average scores for this question. The green line indicates the 80th percentile, above which values may be considered to be within the highest 20% of trusts' average for this question.

It can be seen that when confidence intervals are considered, the scores for many trusts do not differ significantly from each other – each trust's confidence interval overlaps with at least four others, and the majority overlap with a great many other trusts. In other words, trusts are typically quite similar once confidence intervals are taken into account.

Where a trust's score (including the confidence intervals) does **not** cross the red or green threshold lines you can be confident that that is how a trust is performing. For example, in the below chart, there are trusts for which the score (including the confidence intervals) is clearly over the green line indicating the highest 20% of scores.

Chart 5

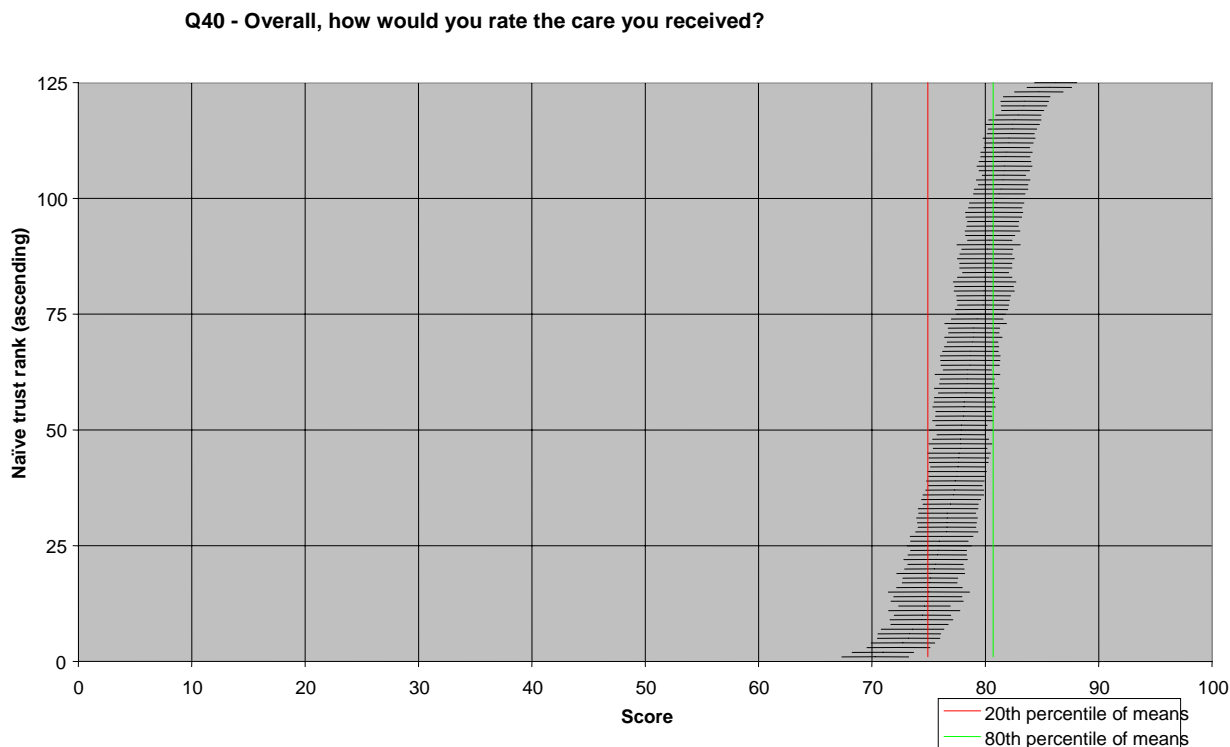
Q3 - Was it possible to find a convenient place to park in the hospital car park?



Even when the majority of respondents answer a question, it may be hard to differentiate trust scores. Chart 6 shows the confidence intervals for all trusts on a question asking respondents how they would rate their overall care. In this chart, we can see that the majority of trusts are very

similar *once the confidence intervals are taken into account*. For example, the trust with the fifth lowest average score for this question is not significantly different from 56 of the 120 trusts with higher average scores. This strongly demonstrates why the data should not be used to construct league tables; league tables cannot fairly account for confidence intervals and, without these, differences are implied where there are none.

Chart 6



5.) How the data was calculated

The data in the benchmark reports is calculated by converting responses to particular questions into scores. These were calculated by converting each respondent's answer to a question into a score (from 0 to 100) then averaging these to arrive at a single score for the trust, for each question. The higher the score, the better a trust is performing. An example of a scored question is shown below. A 'scored' questionnaire is available for each survey on the Care Quality Commission website which shows how each question is scored.

Q: When you had important questions to ask a doctor, did you get answers that you could understand?

100 1 Yes, always

50 2 Yes, sometimes

0 3 No

- 4 I had no need to ask

In most cases, the scores are allocated such that the most positive possible response corresponds to a score of 100 and the least positive to a score of 0, with intermediary options assigned scores at equal intervals. Note that this approach is equivalent to that typically used with Likert scales.

Please also note that it is not appropriate to score all questions within the questionnaire for benchmarking purposes. This is because not all of the questions assess the trusts in any way (for example, the question “Who referred you to see a specialist?”), or they may be ‘filter questions’ designed to filter out respondents to whom following questions do not apply (for example ‘During your stay in hospital, did you have an operation or procedure?’).

Format of the Data

Results shown in the benchmark reports are based on ‘standardised’ data. We know that the views of a respondent can reflect not only their experience of NHS services, but can also relate to certain demographic characteristics, such as their age and sex. For example, older respondents tend to report more positive experiences than younger respondents, and women tend to report less positive experiences than do men. Because the mix of patients varies across trusts (for example, one trust may serve a considerably older population than another), this could potentially lead to the results for a trust appearing better or worse than they would if they had a slightly different profile of patients. To account for this we ‘standardise’ the data. Standardising data adjusts for these differences and enables the results for trusts to be compared more fairly than could be achieved using non-standardised data. More detailed information for each survey is available on request to the Care Quality Commission survey team by contacting:

patient.survey@cqc.org.uk

**Care Quality Commission
May 2009**