

NHS Patient Survey Programme

2019 Community mental health survey

Technical details for analysing trust-level results

Published November 2019

CQC publication

Contents

1. Introduction	2
2. Selecting data for reporting	2
3. The CQC organisation search tool.....	3
4. The trust benchmark reports	3
5. Interpreting the data	4
6. Further information.....	5
Appendix A: Scoring for the survey results	6
Appendix B: Calculating the trust score and weight	8
Appendix C: Calculation of the expected ranges	13
Appendix D: Calculation of standard errors	16

1. Introduction

This document outlines the methods used by the Care Quality Commission to score and analyse the trust level results for the 2019 Community Mental Health Survey, as available on the Care Quality Commission website, and in the benchmark reports for each trust.

On the CQC website the survey data is shown in a simplified way, identifying whether a trust performed 'better' or 'worse' or 'about the same' as the majority of other trusts for each question. This analysis is done using a statistic called the 'expected range' (see **section 5.3**). On publication of the survey, an A to Z list of trust names will be available at the link below, containing further links to the survey data for all NHS trusts that took part in the survey: www.cqc.org.uk/cmhsurvey.

The CQC webpage also contains a statistical release document containing England level results, alongside relevant national policy and comparisons with the results from previous iterations from the survey. Further information on the survey is available in the Quality and Methodology document.

A benchmark report is available for each trust on the [NHS surveys website](#). Results displayed in the benchmark report are a graphical representation of the results displayed for the public on the CQC website. The tables in the back of each benchmark report also highlight any statistically significant changes in the trust score between 2018 and 2019.

2. Selecting data for reporting

Scores are assigned to responses to questions that are of an evaluative nature: in other words, those questions where results can be used to assess the performance of a trust (see **section 5.1** for more detail). Questions that are not presented in this way tend to be those included solely for 'filtering' respondents past any questions that may not be relevant to them (such as, 'In the last 12 months, have you been receiving any medicines for your mental health needs?') or those used for descriptive or information purposes (such as, 'When was the last time you saw someone from NHS mental health services?').

The scores for each question are grouped on the website, and in the benchmark reports for each trust, according to the sections of the questionnaire as completed by respondents. For example, the Community Mental Health Survey includes sections on 'your health and social care workers', 'organising your care' and 'planning your care' amongst others. The only exception to this is Q3 which is included in the 'overall views of care and services' section.

Alongside both the question and section scores on the website are one of three statements:

- Better
- About the same
- Worse

This analysis is done using a statistic called the 'expected range' (see **section 5.3**)

3. The CQC organisation search tool

The organisation search tool contains information from various areas within the Care Quality Commission's functions. The presentation of the survey data was designed using feedback from people who use the data, so that as well as meeting their needs, it presents the groupings of the trust results in a simple and fair way. It shows where we are confident that a trust's score is better or worse than we would expect, when compared with most other trusts.

The survey data can be found from the A to Z link available at: www.cqc.org.uk/cmhsurvey or by searching for a provider from the CQC home page, then clicking on 'Surveys'.

4. The trust benchmark reports

Benchmark reports should be used by NHS trusts to identify how they are performing in relation to all other trusts that took part in the survey. Tables at the back of the report show if a score has significantly increased or decreased compared with the previous survey in 2018. From this, areas for improvement can be identified. The reports are available from the [NHS surveys website](#).

The graphs included in the reports display the scores for a trust, compared with the full range of results from all other trusts that took part in the survey. In the graphs, the bar is divided into three sections:

- If a trust score lies in the grey section of the graph, the trust result is 'about the same' as most other trusts in the survey
- If a trust scores lies in the orange section of the graph, the trust result is 'worse' than expected when compared with most other trusts in the survey
- If a score lies in the green section of the graph, the trust result is 'better' than expected when compared with most other trusts in the survey.

A black diamond represents the score for this trust. No chart is shown for questions answered by fewer than 30 people because the uncertainty around the result would be too great.

5. Interpreting the data

5.1 Scoring

Questions are scored on a scale from 0 to 10. Details of the scoring for this survey are available in **Appendix A**.

The scores represent the extent to which the service user's experience could be improved. A response that was assigned a score of zero refers to the most negative experience we can measure. Whereas a response that was assigned a score of 10 refers to the most positive experience we can measure.

Where a number of options lay between the most negative and positive responses, they were placed at equal intervals along the scale. Where options were provided that did not have any bearing on the trust's performance in terms of peoples' experience, the responses were classified as not applicable and a score was not given. Where respondents stated they could not remember or did not know the answer to a question, no score is given.

5.2 Standardisation

Results are based on standardised data. We know that the views of a respondent can reflect not only their experiences of NHS services, but can also relate to certain demographic characteristics, such as their age and gender. The mix of patients varies across trusts and this could lead to bias, resulting in a trust appearing better or worse than they would if they had a slightly different profile of service users. 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.

The 2019 Community Mental Health Survey is standardised by **age** and **gender**.

5.3 Expected range

The better/about the same/worse categories are based on the expected range that is calculated for each question. This is the range within which we would expect a particular trust to score if it performed about the same as most other trusts in the survey. The range takes into account the number of respondents from each trust as well as the scores for all other trusts, and allows us to identify which scores we can confidently say are better or worse than the majority of other trusts (see **Appendix C** for more details). Analysing the survey information in such a way allows for fairer conclusions to be made in terms of each trust's performance. This approach presents the findings in a way that takes account of all necessary factors yet is presented in a simple manner.

As the expected range calculation takes into account the number of respondents at each trust who answer a question, it is not necessary to present confidence intervals around each score for the purposes of comparing across all trusts.

5.4 Comparing scores across or within trusts

The expected range statistic is used to arrive at a judgement of how a trust is performing compared with all other trusts that took part in the survey. However, if you want to use the scored data in another way, to compare scores between different trusts, you will need to undertake an appropriate statistical test to ensure that any changes are statistically significant.

5.5 Conclusions made on performance

It should be noted that the data only show performance relative to other trusts; we have not set out absolute thresholds for 'good' or 'bad' performance. Thus, a trust may have a low score for a specific question, while still performing very well on the whole. This is particularly true on questions where the majority of trusts exhibit a high score.

A separate report, which explores how overall results between trusts vary across the country, known as an Outlier report, will be available on CQC site www.cqc.org.uk/cmhsurvey once the pre-election reporting restrictions for the 2019 general election have been lifted. This report focuses on identifying significantly higher levels of better or worse patient experience across the entire survey, rather than considering performance on individual questions.

6. Further information

England and trust level results can be found on the CQC website. Also available is a Quality and Methodology report which provides information about the survey development and methodology:

www.cqc.org.uk/cmhsurvey

The results from previous community mental health surveys are available at the link below. Please note that due to redevelopment work, results from the 2019 survey are only comparable with 2014, 2015, 2016, 2017 and 2018:

<https://nhssurveys.org/surveys/survey/05-community-mental-health/>

More information on the NHS Patient Survey Programme, including results from other surveys and a schedule of current and forthcoming surveys can be found at: www.cqc.org.uk/publications/surveys/surveys

Appendix A: scoring for the 2019 community mental health survey results

The following describes the scoring system applied to the evaluative questions in the survey. Taking question four as an example (Figure A1), it asks respondents if they were given enough time to discuss their needs and treatment. The option of 'No' was allocated a score of 0, as this suggests that the respondents experience needs to be improved. A score of 10 was assigned to the option 'Yes, definitely', as it reflects a positive experience. The remaining option, 'Yes, to some extent', was assigned a score of 5 as the respondent did not feel fully listened to. Hence it was placed on the midpoint of the scale.

If the respondent did not know, this was classified as a 'not applicable' response, as this option was not a direct measure of the trust.

Figure A1 Scoring example:

4. Were you given enough time to discuss your needs and treatment?	
Yes, definitely	10
Yes, to some extent	5
No	0
Don't know/can't remember	Not applicable

Where a number of options lay between the negative and positive responses, they were placed at equal intervals along the scale. For example, question nine asks how well the person who is in charge of organising their care organises the care and services they receive (Figure A2). The following response options were available:

- Very well
- Quite well
- Not very well
- Not at all well

A score of 10 was assigned to the option 'Very well', as this represents best outcome in terms of peoples' experiences. A response of 'not at all well' was given a score of 0. The remaining two answers were assigned a score that reflected their position in terms of quality of experience, spread evenly across the scale.

Figure A2 Scoring example:

9. How well does this person organise the care and services you need?	
Very well	10
Quite well	6.7
Not very well	3.3
Not at all well	0

Find a link to the scored questionnaire [here](#).

Details of the method used to calculate the scores for each trust and for individual questions are available in **Appendix B**. This also includes an explanation of the technique used to identify scores that are better, worse or about the same as most other trusts.

All analysis is carried out on a cleaned data set. Cleaning refers to the editing process that is undertaken on the survey data. A document describing this can be [here](#).

As part of the cleaning process, responses are removed from any trust that has fewer than 30 respondents to a question. This is because the uncertainty around the result is too high, and very low numbers would risk respondents being recognised from their responses.

Appendix B: calculating the trust score and weight

Calculating trust scores

The scores for each question and section were calculated using the method described below.

Weights were calculated to adjust for any variation between trusts that resulted from differences in the age and gender groupings of respondents. A weight was calculated for each respondent by dividing the England proportion of respondents (based on all respondents to the survey) in their age/gender group by the corresponding trust proportion. The reason for weighting the data was that respondents may answer questions differently, depending on certain characteristics. If a trust had a large population of young people or women, their performance might be judged more negatively than if there was a more consistent distribution of age and gender of respondents.

Weighting survey responses

The first stage of the analysis involved calculating the England age and gender proportion. It must be noted that the term 'England proportion' within this context refers to the respondent population rather than the entire population of England as it was obtained from pooling the survey data from all trusts.

The questionnaire asked respondents to state their year of birth. The approximate age of each respondent was then calculated by subtracting the figure given from the survey year. The respondents were then grouped according to the categories shown in **Figure B1**.

If a respondent did not fill in their year of birth or gender on the questionnaire, this information was inputted from the sample file. If information on a respondent's age and/or gender was missing from both the questionnaire and the sample file, the respondent was excluded from the analysis as it is not possible to assign a weight.

As shown in Figure B1, the proportion of respondents who were men aged 18-35 is 0.053; the proportion of respondents who were women aged 18-35 is 0.099 etc.

Figure B1 England Proportions

Gender	Age Group	England proportion 2019
Men	18-35	0.053
	36-50	0.091
	51-65	0.117
	66+	0.176
Women	18-35	0.099
	36-50	0.102
	51-65	0.137
	66+	0.224

Note: All proportions are given to three decimals places for this example. The analysis included these figures to 14 decimal places and can be provided on request from the CQC surveys team at patient.survey@cqc.org.uk.

These proportions were then calculated for each trust using the same procedure.

The next step was to calculate the weighting for each individual. Age/gender weightings were calculated for each respondent by dividing the England proportion of respondents in their age/gender group by the corresponding trust proportion.

If, for example, a lower proportion of men aged 18-35 within Trust A responded to the survey, in comparison with the England proportion, then this group would be under-represented in the final scores for the trust. Dividing the England proportion by the trust proportion results in a weighting greater than one for members of this group (**Figure B2**). This increases the influence of responses made by respondents within that group in the final score, thus counteracting the low representation.

Figure B2 Proportion and Weighting for Trust A

Gender	Age Group	England proportion	Trust A Proportion	Trust A Weight (England/Trust A)
Men	18-35	0.053	0.049	1.082
	36-50	0.091	0.092	0.989
	51-65	0.117	0.135	0.867
	66+	0.176	0.164	1.073
Women	18-35	0.099	0.118	0.839
	36-50	0.102	0.101	1.010
	51-65	0.137	0.128	1.070
	66+	0.224	0.213	1.052

Likewise, if a considerably higher proportion of men aged 18-35 within Trust B responded to the survey (**Figure B3**), then this group would be over-represented in the final scores for the trust. To counteract this, dividing the England proportion by the proportion for Trust B results in a weighting of less than one for this group.

Figure B3 Proportion and Weighting for Trust B

Gender	Age Group	England proportion	Trust B Proportion	Trust B Weight (England/Trust B)
Men	18-35	0.053	0.062	0.862
	36-50	0.091	0.080	1.145
	51-65	0.117	0.123	0.955
	66+	0.176	0.177	0.997
Women	18-35	0.099	0.106	0.938
	36-50	0.102	0.114	0.899
	51-65	0.137	0.141	0.975
	66+	0.224	0.201	1.117

To prevent the possibility of excessive weight being given to respondents in an extremely under-represented group, the maximum value for any weight was set at five. There was no minimum weight for respondents as applying very small weights to over-represented groups does not have the same potential to give excessive impact to the responses of small numbers of individual respondents.

Calculating question scores

The trust score for each question displayed on the website and in the benchmark reports, was calculated by applying the weighting for each respondent to the scores allocated to each response.

The below is a working example of this process for the 'health and social care workers' section of the questionnaire which for simplicity uses three respondents.

The responses given by each respondent were entered into a dataset using the 0-10 scale described in section 5.1 and outlined in **Appendix A**. Each row corresponded to an individual respondent, and each column related to a survey question. For those questions that the respondent did not answer (or received a "not applicable" score for), the relevant cell remained empty. Alongside these were the weightings allocated to each respondent.

Figure B4 Scoring for the ‘health and social care workers’ section, Trust B

Respondent	Scores		Weight
	Q4	Q5	
1	5	.	1.145
2	10	10	0.955
3	5	0	0.997

Respondents’ scores for each question were then multiplied individually by the relevant weighting, in order to obtain the numerators for the trust scores.

Figure B5 Numerators for the ‘Health and Social Care workers’ section, Trust B

Respondent	Scores		Weight
	Q4	Q5	
1	5.725	.	1.145
2	9.550	9.55	0.955
3	4.985	0	0.997

Obtaining the denominators for each domain score

A second dataset was then created. This contained a column for each question, and again with each row corresponding to an individual respondent. A value of one was entered for the questions where a response had been given by the respondent, and all questions that had been left unanswered or allocated a scoring of not applicable were set to missing.

Figure B6 Values for non-missing responses, ‘health and social care workers’ section, Trust B

Respondent	Scores		Weight
	Q4	Q5	
1	1	.	1.145
2	1	1	0.955
3	1	1	0.997

The denominators were calculated by multiplying each of the cells within the second dataset by the weighting allocated to each respondent. This resulted in a figure for each question that the respondent had answered. Again, the cells relating to the questions that the respondent did not answer (or received a not applicable score for) remained set to missing.

Figure B7 Denominators for the ‘Health and Social Care workers’ section, Trust B

Respondent	Scores		Weight
	Q4	Q5	
1	1.145	.	1.145
2	0.955	0.955	0.955
3	0.997	0.997	0.997

The weighted mean score for each trust, for each question, was calculated by dividing the sum of the weighted scores for a question (i.e. numerators), by the weighted sum of all eligible respondents to the question (i.e. denominators) for each trust.

Using the example data for Trust B, we first calculated weighted mean scores for each of the two questions that contributed to the ‘health and social care workers’ section of the questionnaire.

$$\text{Q4: } \frac{5.725 + 9.550 + 4.985}{1.145 + 0.955 + 0.997} = 6.542$$

$$\text{Q5: } \frac{9.550 + 0.000}{0.955 + 0.997} = 4.892$$

Calculating section scores

A simple arithmetical mean of each trust’s question scores was then taken to give the score for each section. Continuing the example from above, Trust B’s score for the ‘health and social care workers’ section of the Community Mental Health Survey would be calculated as:

$$(6.542 + 4.892) / 2 = 5.717$$

Appendix C: calculation of the expected ranges

Z statistics (or Z scores) are standardised scores derived from normally distributed data, where the value of the Z score translates directly to a p-value. That p-value then translates to what level of confidence you have in saying that a value is significantly different from the mean of your data (or your target value).

A standard Z score for a given item is calculated as:

$$z_i = \frac{y_i - \theta_0}{s_i} \quad (1)$$

where: s_i is the standard error of the trust score¹,
 y_i is the trust score
 θ_0 is the mean score for all trusts

Under this banding scheme, a trust with a Z score of < -1.96 is labeled as 'worse' (significantly below average; $p < 0.025$ that the trust score is below the England average), $-1.96 < Z < 1.96$ as 'about the same'; and $Z > 1.96$ as 'better' (significantly above average; $p < 0.025$ that the trust score is above the England average) than what would be expected based on the distribution of trust scores for England.

However, for measures where there is a high level of precision in the estimates (the survey sample sizes average around 400 to 500 per trust), the standard Z score may give a disproportionately high number of trusts in the significantly above/below average bands (because s_i is generally so small). This is compounded by the fact that all the factors that may affect a trust's score cannot be controlled. For example, if trust scores are closely related to economic deprivation then there may be significant variation between trusts due to this factor, not necessarily due to factors within the trusts' control. In this situation, the data are said to be over dispersed. That problem can be partially overcome by the use of an additive random effects model to calculate the Z score (we refer to this modified Z score as the Z_D score). Under that model, we accept that there is natural variation between trust scores and this variation is then taken into account, by adding this to the trust's local standard error in the denominator of (1). In effect, rather than comparing each trust simply to one target value for England, we are comparing them to an England distribution.

The Z_D score for each question and section was calculated as the trust score minus the England mean score, divided by the standard error of the trust score plus the variance of the scores between trusts. This method of calculating a Z_D score differs from the standard method of calculating a Z score in that it recognizes that there is likely to be natural variation between

¹ Calculated using the method in Appendix D.

trusts which one should expect and accept. Rather than comparing each trust to one point only (i.e. the England mean score), it compares each trust to a distribution of acceptable scores. This is achieved by adding some of the variance of the scores between trusts to the denominator.

The steps taken to calculate Z_D scores, based on the method presented in Spiegelhalter et al. (2012)², are outlined below.

Winsorising Z-scores

The first step when calculating Z_D is to Winsorise the standard Z scores (from (1)). Winsorising consists of shrinking in the extreme Z-scores to some selected percentile, using the following method:

1. Rank cases according to their naive Z-scores.
2. Identify Z_q and $Z_{(1-q)}$, the 100q% most extreme top and bottom naive Z-scores. For this work, we used a value of $q=0.1$
3. Set the lowest 100q% of Z-scores to Z_q , and the highest 100q% of Z-scores to $Z_{(1-q)}$. These are the Winsorised statistics.

This retains the same number of Z-scores but discounts the influence of outliers.

Estimation of over-dispersion

An over dispersion factor $\hat{\phi}$ is estimated for each indicator which allows us to say whether the data for that indicator are over dispersed or not:

$$\hat{\phi} = \frac{1}{I} \sum_{i=1}^I z_i^2 \quad (2)$$

Where I is the sample size (number of trusts) and z_i is the Z score for the i th trust given by (1). The Winsorised Z scores are used in estimating $\hat{\phi}$.

An additive random effects model

If $I \hat{\phi}$ is greater than $(I - 1)$ then we need to estimate the expected variation between trusts. We take this as the standard deviation of the distribution of θ_i (trust means) for trusts, which are on target, we give this value the symbol $\hat{\tau}$, which is estimated using the following formula:

² Spiegelhalter D., Sherlaw-Johnson, C., Bardsley, M., Blunt, I., Wood, C., & Grigg, O. (2012). Statistical methods for healthcare regulation: Rating, screening and surveillance. *Journal of the Royal Statistical Society (Series A)*, 175(1), 1-47.

$$\hat{\tau}^2 = \frac{I\hat{\phi} - (I-1)}{\sum_i w_i - \sum_i w_i^2 / \sum_i w_i} \quad (3)$$

where $w_i = 1 / s_i^2$ and $\hat{\phi}$ is from (2). Once $\hat{\tau}$ has been estimated, the Z_D score is calculated as:

$$Z_i^D = \frac{y_i - \theta_0}{\sqrt{s_i^2 + \hat{\tau}^2}} \quad (4)$$

Appendix D: calculation of standard errors

In order to calculate statistical bandings from the data, it is necessary for CQC to have both trusts' scores for each question and section and the associated standard error. Since each section is based on an aggregation of question mean scores that are based on question responses, a standard error needs to be calculated using an appropriate methodology.

For the patient experience surveys, the z-scores are scores calculated for section and question scores, which combines relevant questions making up each section into one overall score, and uses the pooled variance of the question scores.

Assumptions and notation

The following notation will be used in formulae:

X_{ijk} is the score for respondent j in trust i to question k

Q is the number of questions within section d

w_{ij} is the standardization weight calculated for respondent j in trust i

Y_{ik} is the overall trust i score for question k

Y_{id} is the overall score for section d for trust i

Associated with the subject or respondent is a weight w_{ij} corresponding to how well the respondent's age/gender is represented in the survey compared with the population of interest.

Calculating mean scores

Given the notation described above, it follows that the overall score for trust i on question k is given as:

$$Y_{ik} = \frac{\sum_j w_{ij} X_{ijk}}{\sum_j w_{ij}}$$

The overall score for section d for trust i is then the average of the trust-level question means within section d . This is given as:

$$Y_{id} = \frac{\sum_{k=1}^Q Y_{ikd}}{Q}$$

Calculating standard errors

For questions, the variance of the trust score is estimated with the Taylor series linearization method (see e.g. Lee & Forthofer, 2006; Lumley, 2004). The standard error of the trust score, s_i , is the square root of the Taylor series estimate of variance.

The variance within trust i on question k is given by:

$$\hat{\sigma}_{ik}^2 = \frac{\sum_j w_{ij} \left(X_{ijk} - Y_{ik} \right)^2}{\sum_j w_{ij}}$$

This assumes independence between respondents.

For ease of calculation, and as the sample size is large, we have used the biased estimate for variance.

The variance of the trust level average question score is then given by:

$$\begin{aligned} V_{ik} &= \text{Var}(Y_{ik}) = \text{Var}\left(\frac{\sum_j w_{ij} X_{ijk}}{\sum_j w_{ij}} \right) \\ &= \frac{\text{Var}\left(\sum_j w_{ij} X_{ijk} \right)}{\left(\sum_j w_{ij} \right)^2} \\ &= \frac{\hat{\sigma}_{ik}^2 \sum_j w_{ij}^2}{\left(\sum_j w_{ij} \right)^2} \end{aligned}$$

Covariances between pairs of questions (here, k and m) can be calculated in a similar way:

$$\text{COV}_{ik.im} = \text{Cov}(Y_{ik}, Y_{im}) = \frac{\hat{\sigma}_{ikm} \sum_j w_{ij}^2}{\left(\sum_j w_{ij} \right)^2}$$

Where $\hat{\sigma}_{ikm} = \frac{\sum_j w_{ij} (X_{ijk} - Y_{ik})(X_{ijm} - Y_{im})}{\sum_j w_{ij}}$

Note: w_{ij} is set to zero in cases where patient j in trust i did not answer both questions k and m .

The trust level variance for the section score d for trust i is given by:

$$V_{id} = Var(Y_{id}) = \frac{1}{Q^2} \left\{ \sum_{k=1}^Q V_{ik} + 2 \sum_{k=2}^Q \sum_{m=1}^{k-1} COV_{ik,im} \right\}$$

The standard error of the section score is then:

$$SE_{id} = \sqrt{V_{id}}$$

This simple case can be extended to cover sections of greater length.