Methods, Reasoning and Scope

Statement of Methodology for the Overall Patient Experience Scores (Statistics)

Prepared by NHS England Clinical Programmes and Patient Insight Analytical Unit

Contents

Introduction

This document explains the methods used to calculate the overall patient experience scores statistics. It explains why those particular methods are used, including an explanation of the key decision points. The document also explains the scope and purpose of the statistics and explains how they relate to other measures of feedback from service users and the public.

The document has the following sections:

- Scope, purpose and context
- Background: The NHS Patient Survey Programme
- Key methodological issues
- Full methodology, including mathematical notation
- Annex 1: List of survey questions used in the scoring
- Annex 2: Method for calculating confidence intervals

The underlying data used to calculate these statistics are available to registered users of the UK Data Archive[1]. From this document, it should be possible for researchers and others to replicate our calculations from these raw datasets.

It's worth noting that in some sections of this document (i.e. methodology and mathematical notations) explanations refer to five domains, culminating results for the overall measure. This applies for the Outpatient, Inpatient and Accident & Emergency surveys and is illustrative of the general approach. The Community Mental Health Survey has four domains; therefore in this case some of the examples used in this document, referring predominantly to the Inpatient survey as an example, may be slightly different.

[1] www.data-archive.ac.uk

Scope, purpose and context

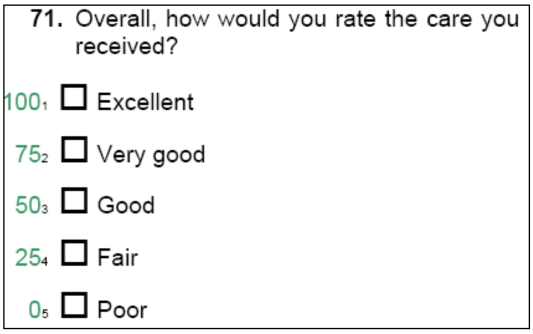
It is important that the NHS, and other public services, take account of feedback and views from individuals in our society (both in responding to individual comment, and also in addressing collective issues and concerns). As the NHS is one of the largest employers in the world, and with an annual cost of over £100bn, it is unsurprising that a wide range of feedback mechanisms have emerged; gathering views from staff, patients, taxpayers, the general public as well as particular interest groups, such as those with long term conditions.

To understand the “overall patient experience scores”, it is helpful to first draw a distinction between three different types of ‘public feedback’:

1. Public perceptions/ public opinion
2. Service user satisfaction/ perception
3. Service user reported experience

The NHS is a taxpayer funded service, used by virtually everyone at some point in their life, so it is important to consider broader public opinion of it. Many sources of public opinion data exist, including the British Social Attitudes Survey[2] and the regular Ipsos MORI poll of public perceptions of the NHS[3] . These sources are useful for public accountability purposes, but they do not directly measure the experience of users. In responding to these surveys, members of the public are likely to have regard to other broader issues, for example their own political views. This is true, even when the respondents are categorised into those who have or have not used the service recently.

If we wish to gauge the views of those who use the NHS, we might consider direct satisfaction measures, for example a survey question that asks:



71. Overall, how would you rate the care you received?

100$_1$ □ Excellent

75$_2$ □ Very good

50$_3$ □ Good

25$_4$ □ Fair

0$_5$ □ Poor

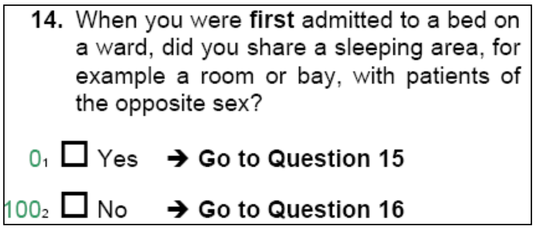
Again, these measures are useful in some settings. The above question is taken from the National Patient Survey Programme (NPSP), which is the same source that we use for the overall patient experience scores. The result of this ‘satisfaction question’ is reported in the Care Quality Commission’s summary of results.

However, user satisfaction measures are not always helpful in holding a service to account, or in taking steps to improve it. If you find that users are not satisfied, it is

[2] http://natcen.ac.uk/our-research/research/british-social-attitudes/

[3] https://www.ipsos.com/ipsos-mori/en-uk/public-perceptions-nhs-and-social-care-survey-winter-2016-wave

not immediately clear what you need to do to improve. You could simply work earnestly to improve all aspects of service. Or, you could try to find out why users are dissatisfied and then focus measurement and improvement on these specific experiences, for example asking questions like:



14. When you were **first** admitted to a bed on a ward, did you share a sleeping area, for example a room or bay, with patients of the opposite sex?

0_1 □ Yes ➔ **Go to Question 15**

100_2 □ No ➔ **Go to Question 16**

This leads us to the idea of measuring service-user reported experience. We continue to focus on feedback and responses directly from service users, but instead focus much more on objective, measurable actual experiences (e.g. did you wait less than an hour?, were you given information about your medicines?, did you get a copy of the letter to your GP?)

Measures of this type are more useful in responding to issues or addressing areas for improvement, because they can be measured objectively and consistently over time. The overall patient experience scores use questions of this type, taken from the existing NPSP. The methodology is designed to overcome the two main drawbacks of this type of survey question:

1. Questions can only focus on one aspect of service at a time, so they don't immediately give you the 'big picture'
2. There is a risk that the range of questions will not properly capture the full range of patient or user concerns, or that questions will focus on issues that are not those of primary interest to service users.

The overall patient experience scores have been developed along the lines suggested by extensive literature on the development of patient centred care. It takes results from individual survey questions, from individual respondents, and seeks to aggregate up these raw data into a set of overall scores that capture all aspects of service of interest to people who use the NHS.

The technical methodology is therefore one of averaging and totalling data to form a set of overall composite measures that allow valid comparison over time. The key decision points in the methodology are about how to aggregate the results, and which survey questions or aspects of care to include.

Background: The NHS Patient Survey Programme

The NHS Patient Survey Programme is a rolling programme of surveys, with different NHS service areas surveyed each year. For example, in 2016 surveys were carried out for Adult Inpatients, Community Mental Health Services and the Emergency Department. Surveys typically have around 60,000 to 80,000 respondents, from an initial sample frame of 1,250 patients per NHS organisation. Each respondent answers around 50 questions.

Results from these surveys, including summaries of which question results show significant change, are reported by the Care Quality Commission in a separate publication. An example publication is provided in the following footnote[4].

The Overall Patient Experience Measure was introduced in 2011 under the responsibility of the Department of Health. Responsibility for production of the measure was transferred to NHS England in April 2013.

In understanding NHS England's overall patient experience scores, it is important to realise that it does not attempt to provide a full narrative summary of the underlying survey data. We are not aiming to compete with the Care Quality Commission's summary of the data, but to provide a complementary product. We use the data collected through the NHS Patient Survey Programme as an 'administrative source'; we take data that have been collected for another purpose and use them to calculate a statistic of interest.

[4] www.cqc.org.uk/public/reports-surveys-and-reviews/surveys/inpatient-survey-2015

Decision points in the methodology

In this document so far, we have explained our goal of producing an overall measure of patient experience, based on objectively defined survey questions, but then aggregated together into an overall indicator that defines 'the whole of patient experience'. In this next section, we describe some of the key methodological decision points.

Decision point 1: "Let the data speak" by using factor analysis or construct a measure using a set of pre-defined domain headings

We aim to summarise the existing NHS Patient Survey Programme data in a way that captures the full range of service aspects impacting on patients' views of the NHS – ideally into a small number of overall scores. One way to do this would be to use a statistical technique on the raw survey data, to condense down the information into a smaller set of statistics or measures, arising naturally from the data. The standard technique for this is called factor analysis.

The factor analysis approach was tested via commissioned research through Sheffield University's School for Health and Related Research (ScHARR). Unfortunately, factor analysis tended to condense the data into a set of 'factors' that reflected the structure of the surveys (so, for example, the survey has a section headed 'doctors' and the analysis returns a factor relating to 'doctors'). There is a risk that our summary measure simply tells us what the surveys are measuring, rather than what is important to patients.

The testing took into consideration whether these factors derived directly from the surveys would be 'better' at explaining patient satisfaction than the headings used now. In general, analytically defined factors were no 'better' than the existing domains.

Therefore the decision was taken to use a 'defined domain' approach.

Decision 2: Choosing the overall headings (domains)

The aim was to find a short list of headings that describe all aspects of service that are relevant and important to patients. One phrase used to describe this is 'a framework for patient centred care'. There is extensive literature on the development of patient centred care. The starting point for development of this work in the NHS was the American book "Through the Patient's Eyes"[5]. A review of this book published in the New England Journal of Medicine helpfully summarises the methodology[6]

The book describes a scientific process used to assess the elements of patient focussed care, using focus groups, site visits, survey data and literature reviews. The book concludes that there are patient-focussed solutions to be found in seven

[5] Gerteis, M., Edgman-Levitan, S., Daley, L., Delbanco, T.L. (eds) (1993) Through the patient's eyes. Jossey-Bass, San Francisco

[6] www.nejm.org/doi/full/10.1056/NEJM199403243301225

separate areas of patient care. These seven areas formed the basis for the five domain headings.

The list of seven headings was adapted to provide an appropriate measure for an NHS setting. Statisticians worked with policy officials to determine which headings were relevant to the NHS and whether there were any other aspects of NHS care that were not reflected in the US literature. The most important change was to add a heading for ‘access and waiting’ as this was clearly a policy concern for UK patients. Other headings were combined together to reflect policy in England, for example grouping ‘Information’ with ‘Choice’.

Condensing the seven scientific headings into five domains to reflect current priorities was a policy driven process, informed by reasoned analytical judgement about whether the domains were coherent and made sense. The domains were validated further using analytical work from ScHARR. One consideration was whether the defined headings were broadly coherent, in the sense that they measure one consistent aspect of care each. It is possible to test this analytically by examining whether different aspects of the ‘domain’ tend to vary in the same direction. For example, if we were interested in measurement of people, ‘physical size’ would probably be a coherent concept, because people who are larger in one regard tend to be larger in another. A concept of ‘height and hair colour’ would not be coherent, because these two things tend to vary independently. The ScHARR analysis confirmed that the five domains were broadly coherent, although in some areas there was some evidence that they covered more than one aspect of care, even though these tended to vary in the same direction (for example ‘Better Information More Choice’ measured information and choice separately). The formal analytical technique used was a ‘cronbach-alpha’ test.

The five domains are:

- Access & waiting
- Safe, high quality co-ordinate care
- Better information, more choice
- Building closer relationships
- Clean, comfortable, friendly place to be

Decision 3: How to use pre-existing survey data to measure the five pre-existing domain headings

Data is available for each of the survey questions, as reported in the CQC publication. This key decision here was to identify which survey questions related to which of the pre-defined domain headings. It would be necessary to ensure that the range of questions under each domain heading was coherent and provided a rounded measure.

This decision process was fairly straightforward, because only a limited number of survey questions were available. Some of the survey questions were contextual questions for which it was not possible to calculate a score (for example, “who referred you to see a specialist”). These questions were excluded from the analysis.

The remaining questions could be assigned to the five domains by direct logical deduction (for example, “How long did you wait…” is clearly a question about access & waiting).

Statisticians and policy officials then considered the remaining list, and formed subjective judgements about which questions were most useful in defining the relevant domain. In some cases, this involved removing questions that were similar to others, for example privacy whilst discussing your condition was considered similar to privacy whilst being examined or treated.

This was a subjective process, carried out when these indicators were first developed. If the process were to be repeated now, it is likely that NHS England would consult users more widely and take account of their views. However, these statistics are used to measure change over time and there is merit in retaining an unchanged means of measurement, to allow valid comparisons, unless there is a strong reason to revise it. In some areas there may be a need to revisit the list of questions, for example the choice to exclude questions on mixed sex accommodation as part of the measure (this is a policy area that is now of increased importance to users).

Alternative methods for measuring the patient experience were explored for example finding ways to report more frequently on results. As part of this work, it was likely that the choice of these lists of questions would be considered with users. Any change to the published series would be subject to formal consultation with users.

A full list of survey questions used to calculate these scores is given in Annex 1

Decision 4: There was a pre-existing scoring mechanism. Should it be accepted?

Questions in these surveys present patients with a number of multiple choice options (for example, yes/no). When this method was devised, there was a pre-existing scoring schema for each question. For example, ‘yes always’ would typically be given a score of 100, whilst ‘no, never’ would typically be given a score of 0. By default, intermediate categories are scored with equal steps up the scale (so in a question with three categories ‘yes, most of the time’ would be scored at 50). In effect, respondents choose a response from an ordered list, which is then converted into a number or score that assumes, for example, that ‘yes a bit’ is exactly half as good as ‘yes, totally’.

This is a standard social surveys technique for scoring multiple choice responses, but analytically there is a need to consider whether a more sophisticated technique would be more appropriate.

Notionally, values could be assigned to the multiple choice options by somehow determining the real value placed on them by the respondents, or by broader society. This approach was attempted by designing a statistical model to measure how much bearing these response categories had on the patient’s overall satisfaction rating – using the five response categories for the ‘overall satisfaction’ question in the survey.

(For example, the model tests whether patients who say ‘yes I did get enough pain relief’ are more likely to say their overall experience was ‘excellent’, and if so, how much more likely).

This was exploratory work and concluded that this more sophisticated method should not be used, for three reasons:

1. It was hugely complicated to work through
2. The overall results for patient experience were hardly changed at all
3. There were resource considerations. Applying a more sophisticated technique requires more resource to be diverted from other aspects of the analysis

> A decision was taken to use the pre-existing scoring scheme (which is summarised in Annex 1), with some minor adjustments to ensure consistency over time (for example ensuring that a two hour wait in outpatients always received the same score)

Decision 5: Which variables should be used to standardise responses (adjust them to take account of differences in responses for different groups)

These data are based on responses from patients and, of course, each patient is representative of themselves and their own views. However, despite the objective nature of many of these questions, views depend to an extent on the personal characteristics of the respondent such as their age, gender and ethnicity.

For example, for inpatient data, the pattern of responses is also very different for those admitted as an emergency case rather than from a waiting list (elective).

Typically, older patients tend to give more positive responses (sometimes referred to as a ‘gratitude bias’) and so do male patients. These differences are more marked in questions with a subjective element (for example ‘how clean was the ward’?).

However, there is a balance to be struck in deciding whether these differences arise as a natural consequence of that patient’s demographic characteristics, or whether patients in that group receive a different level of experience (which is what these statistics seek to measure). We need to be careful, for example, not to ‘standardise out’ a bias against patients with low educational attainment.

There is also a limit to the number of different elements that can be effectively standardise by. For example, in the inpatient survey patients are categorised into four age bands, two genders and two admission methods, giving sixteen categories. Standardising by another variable, such as the five level ‘education’ question, would give 80 categories and would start to generate problems with small numbers and unpredictable weightings.

Analysis showed that scores did not change very much if standardised by ethnic group or educational attainment. By conclusion there was a clear case for standardising data by admission method, and that on balance age and gender should also be standardised. This approach allows valid comparisons to be made between Trusts with a different demographic mix of patients.

When considering the national figures, this argument for an adjustment no longer applies. In theory, one might go back to the original un-weighted scores and aggregate them across all Trusts. This would have the advantage of giving equal weight to responses from each patient, but there are strong arguments against it:

- Experience suggests that users in the NHS may find it difficult to see how the national figures relate to the published (weighted) trust level figures
- To correct for this, there may be a need to separate the set of unweighted Trust level figures to show how the national figures are derived. This might generate confusion.
- The overall national figures do not vary greatly if we include or exclude a weighting step. The notional argument for giving patients equal weight is therefore not a strong one.

On balance, the arguments favour using the standardised results.

Decision 6: How should questions be aggregated to handle situations where some questions are only answered by a subset of patients

This question was the subject of extensive analysis by the Department of Health during implementation and the CQC's predecessor organisation the Healthcare Commission. Two main methods were explored:

Method 1

Add up the (weighted) scores within the domain to form a numerator. This involves summing over both questions and patients. Then count the number of 'none missing' responses in this selection to form a denominator. Divide.

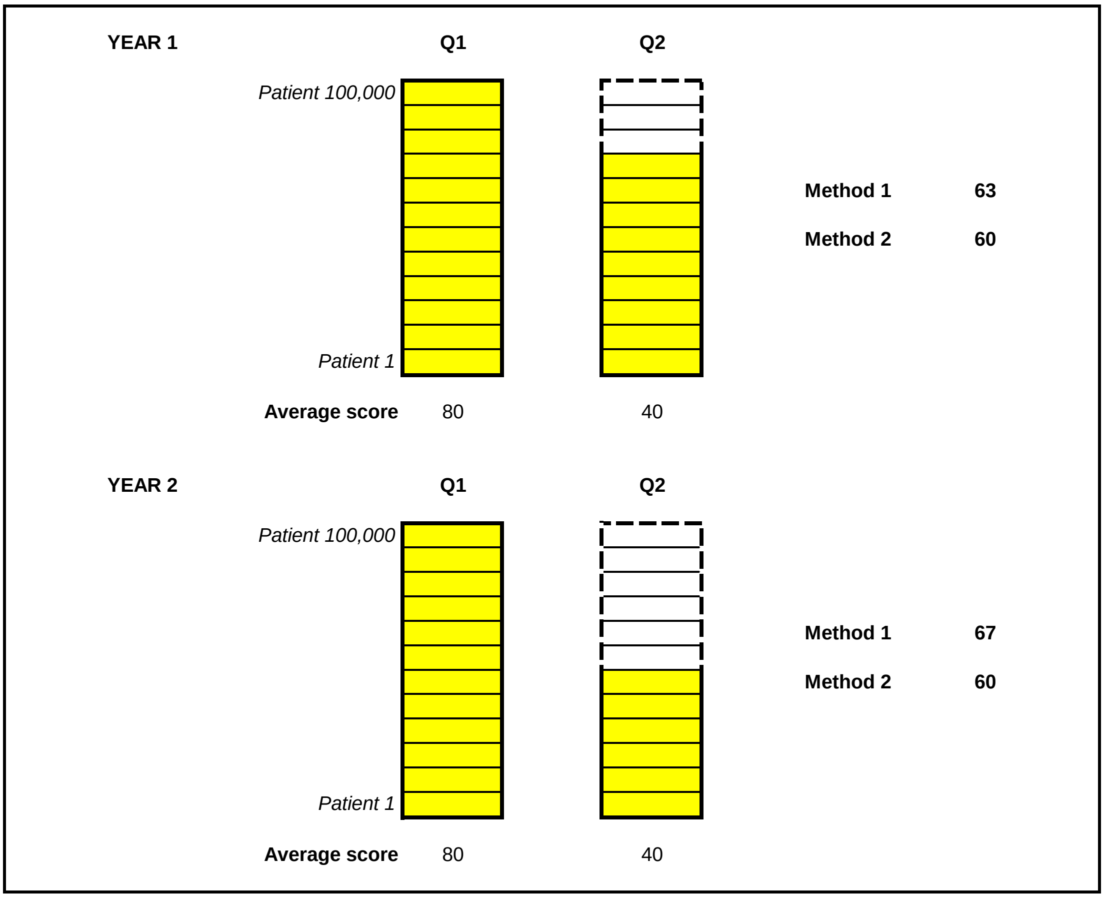
Method 2

Add up the (weighted) scores across patients only (treat each question separately). Calculate an 'average question score' for each question and sum these average question scores across domains by a simple average.

If all patients answered all questions, the results would be identical.

Below is an illustration how the two methods give different results with a simple example. Assume there is a domain that has only two questions. The first question is answered by all respondents in two consecutive surveys. The second question is answered by 75 percent of patients in year 1, and only 50 percent of patients in year 2.

Assume the simple average score for each question remains the same in the consecutive surveys (Q1 = 80, Q2 = 40).



If question 2 applies only to a subset of patients (for example emergency cases only) one might ideally want to give it slightly lower weight. (This is covered in the next ‘decision point’, below). This is a disadvantage of method 2, which automatically gives the two questions equal weight (the 80 and the 40 are averaged to give 60).

Method 1 gives more weight to question 1 and as a result gives slightly higher scores overall. These higher scores might be more representative of the ‘average’ patient. However, the change in response rates for question 2 has generated an artificial change in overall scores under method 1. This is an undesirable feature of these statistics, and for that reason Method 2 is preferable.

The approach taken was to calculate average question scores first, then average those averages to form domains

Decision 7: Should different weights be applied to different questions

Further consideration was given to the idea of allocating different weight to different survey questions. This could be applied artificially as part of ‘Method 2’. This was rejected because there was no robust, objective evidence base on which to generate a weighting. Investigation took place to see which questions were most strongly associated with positive overall satisfaction levels, but in effect this reduced the experience measure to a disguised measure of overall satisfaction. Any attempt to weight question scores by number of responses tends to complicate the methodology and make the indicator less stable over time. It is vital for these statistics to provide valid comparisons over time, so such instability is highly undesirable.

Decision 8: How to aggregate Trust level scores to national level scores (Trust weighting issues)

There are three possible ways to work out national figures:

1. Use the whole dataset and ignore distinctions between trusts (i.e. treat the country as one giant trust – the 'one nation' method)
2. Work out individual Trust level scores and then take a simple average
3. Calculate a weighted average of Trust level scores using some measure of trust size

Since the initial sample size for all Trusts is the same (1,250) the first two methods give similar results, and the second has the advantage of showing a clear and simple relationship between Trust level scores and national scores.

In principle the third method would give a more accurate national picture by giving a higher weight to larger Trusts, but the difference is not large. This approach would introduce an extra layer of complexity to the calculation and would require judgements to be made about the most appropriate measure of size to use for Trusts. Whichever measure of size is selected (beds, admissions, patient episodes etc) this method would require links to other data because there are no direct measures of Trust size within the patient survey datasets. Different measures of size would give different results, and this introduces a degree of subjectivity into the methodology. In addition, this approach would tend to make the overall national measure unstable when there is organisational change in the NHS (for example Trust mergers).

It is important to note that because all Trusts have the same sample size, patients at smaller Trusts are disproportionately represented in the national figures, but this provides a sensible balance between transparent and simple methodology and analytical rigour.

Decision 9: How to handle minor data corrections (e.g. male respondents at Birmingham Women's Hospital)

There is a risk of distortion in the figures from 'oddities', such as male patients at Birmingham Women's Hospital. If these are ignored they would attract very large weights in the calculations. This is addressed in two ways, firstly by examining the data before analysis to identify and correct any simple anomalies (and where appropriate making judgements to re-code these patients as female). Furthermore, cap the standardisation weights at 5 to avoid placing undue weight on a small number of individuals.

Full methodology (with mathematical notation)

This section takes into account the decision points reported in the previous section, and summarises the resulting methodology, where appropriate with mathematical notation to define formulae etc.

The raw dataset comprises of survey responses from individual patients. Each row in the dataset corresponds to one returned survey, and the columns of the dataset contain demographic details such as age, gender etc together with the individual question responses.

This section does not describe the calculation of confidence intervals, as this is covered separately in Annex 2.

Step 1: Basic cleaning and scoring

- All question responses are converted to scores using the scoring scheme in Annex 1
- Missing values for age and gender are replaced with values from the sample file, where available
- Records without a valid age or gender or admission type are removed from the file
- Records without any valid question responses are removed from the file
- Ages are grouped into bands: 18 to 35, 36 to 50, 51 to 65, 66 plus
- Survey filters are followed, and inappropriate routing corrected (e.g. respondent answers Q42, which then says ‘go to Q46’, but respondent has also answered Q43-45. In this example, the answers to Q43-45 would be removed).

Step 2: Standardisation by age, gender and (for inpatients only) admission method

Each individual is assigned to a group based on their age-band, gender and (for inpatients only) whether they were admitted as emergency or elective. Totals for each such group (or strata) are calculated for each NHS Trust, and also nationally. A weight is then calculated for each individual as follows:

$$W_i = \frac{\sum n_{\text{national},i}}{\sum n_{\text{trust},i}}$$

Where n is the number of valid records at Trust or national level **in the same stratum as patient i**.

In statistical terms, this is straightforward direct standardisation, weighting individuals within the Trust to standardise to the national mix of age, gender and admission method.

In cases where the weight is greater than 5, it is reduced to 5 to avoid distortions from a few individuals.

Step 3: Question scores at Trust level

An average score is then worked out for each of the relevant questions, within each Trust. The Trust level average (mean) score for question j and Trust k is given by:

$$\overline{x_{jk}} = \frac{\sum_i W_i x_{ijk}}{\sum_i W_i}$$

Here, the x_{ijk} represent the question scores from each individual patient and the W_i represents the weight for that individual. The weight is a standardisation weight, as defined above, and it flows directly from the age, gender, and admission method of patient i.

Note at this point how the method handles missing values. The relevant x_{ijk} would be 'missing' and therefore excluded from the numerator. To compensate, the W_i for those individuals is set to zero, so they do not contribute to the denominator either. Thus, the question score is a weighted average of responses received, regardless of how many responses there were.

Step 4: Using question scores to calculate domain scores and the overall measure

Trust level question scores are aggregated up to domain scores and to national level scores by simple linear calculations (in most cases, just by working out the average). It doesn't matter whether the average is applied across questions first, and then to trusts, or the other way around. To maintain consistency with Annex 2, the methodology presented here shows the calculation of trust level domain scores first, then aggregate to national scores later.

Each domain is calculated by directly averaging relevant questions as listed in Annex 1. Let d_{lk}denote the lth domain score in Trust k. Here, the values of l range from 1 to 6, with 6 denoting the overall score.

For example, the domain scores for inpatients are defined as follows:

$$d_{1k} = \frac{1}{3}(\overline{x_{1k}} + \overline{x_{2k}} + \overline{x_{3k}})$$

$$d_{2k} = \frac{1}{3}(\overline{x_{4k}} + \overline{x_{5k}} + \overline{x_{6k}})$$

$$d_{3k} = \frac{1}{3}(\overline{x_{7k}} + \overline{x_{8k}} + \overline{x_{9k}})$$

$$d_{4k} = \frac{1}{4}(\overline{x_{10,k}} + \overline{x_{11,k}} + \overline{x_{12,k}} + \overline{x_{13,k}})$$

$$d_{5k} = \frac{1}{6}\left(\frac{1}{2}(\overline{x_{14,k}} + \overline{x_{15,k}}) + \overline{x_{16,k}} + \overline{x_{17,k}} + \overline{x_{18,k}} + \overline{x_{19,k}} + \overline{x_{20,k}}\right)$$

For simplicity the question scores have been renumbered so that questions 1, 2 and 3 feed into domain 1 and so on. Also note that in domain 5 the first two questions are averaged first to form a single score for ‘noise at night’. This is for historical reasons, to allow comparison with data from 2001 when a single survey question was asked.

Similar formulae are used for surveys in other NHS service areas. This one is shown as an example because of the exception around ‘noise at night’. There are no similar exceptions for other surveys.

The overall measure is given by:

$$d_{6k} = \frac{1}{5}(d_{1k} + d_{2k} + d_{3k} + d_{4k} + d_{5k})$$

Step 5: Aggregating to national totals

There is no need to apply differential weights when calculating national scores – The national (England) level scores are calculated as a simple average (mean) of Trust level scores:

$$\overline{x_j} = \frac{\Sigma\, \overline{x_{jk}}}{K}$$

for a single question and;

$$\overline{d_l} = \frac{\Sigma\, \overline{d_{lk}}}{K}$$

for a domain, where K is the number of Trusts (typically around 150).

Annex 1: Full list of survey questions, and scoring regime

The scoring mechanism below has a range of 0 to 100 for each question, in line with the pre-existing scoring mechanism. In 2011, CQC changed the scale for scoring questions in the National Patient Survey Programme to a range from 0 to 10[7]. The change, reflected in data publications by CQC, is solely presentational. To retain clear historical comparability, the Overall Patient Experience Scores continue to use the range 0 to 100 but it is important to note that the scores are exactly the same, CQC have simply divided them by ten. Unscored response options (scored as 'M') are interpreted as missing.

The questionnaires are regularly revised to ensure they are up to date and in line with current policy and practice. The questions listed below are correct for the survey year referenced.

Adult Inpatient Survey (2017 questionnaire)

	Scoring (Response = score)
Access and waiting	
Was your admission date changed by the hospital? [Waiting list only]	1=100 2=67 3=33 4=0
How do you feel about the length of time you were on the waiting list before your admission to hospital? [Waiting list only]	1=100 2=50 3=0
From the time you arrived at the hospital, did you feel that you had to wait a long time to get to a bed on a ward? [All]	1=0 2=50 3=100
Safe, high quality, coordinated care	
Sometimes in a hospital, a member of staff will say one thing and another will say something quite different. Did this happen to you? [All]	1=0 2=50 3=100
On the day you left hospital, was your discharge delayed for any reason? + What was the main reason for the delay? [All/All delayed]	1=*see main reason* 2=100 ………………..... *Main reason:* 1=0 2=0 3=0 4=M
Did a member of staff tell you about any danger signals you should watch for after you went home? [All] (this question is usually included in the OPES calculation, but was not included this year due to printing errors)	1=100 2=50 3=0 4=M

[7] The change was made to emphasise that the scores are not percentages.

Better information, more choice	
Were you involved as much as you wanted to be in decisions made about your care and treatment? [All]	1=100 2=50 3=0
Did a member of staff explain the purpose of the medicines you were to take at home in a way you could understand? [All given medication]	1=100 2=50 3=0 4,5=M
Did a member of staff tell you about medication side effects to watch for when you went home? [All given new medication and wanting an explanation]	1=100 2=50 3=0 4=M
Building closer relationships	
When you had important questions to ask a doctor, did you get answers that you could understand? [All wanting an explanation]	1=100 2=50 3=0 4=M
Did doctors talk in front of you as if you weren't there? [All]	1=0 2=50 3=100
When you had important questions to ask a nurse, did you get answers that you could understand? [All wanting an explanation]	1=100 2=50 3=0 4=M
Did nurses talk in front of you as if you weren't there? [All]	1=0 2=50 3=100
Clean, comfortable, friendly place to be	
Were you ever bothered by noise at night from other patients? + Were you ever bothered by noise at night from hospital staff? [All]	1=0 2=100 (both Qs) The scores are then averaged.
In your opinion, how clean was the hospital room or ward that you were in? [All]	1=100 2=67 3=33 4=0
How would you rate the hospital food? [All who had food]	1=100 2=67 3=33 4=0 5=M
Were you given enough privacy when being examined or treated? [All]	1=100 2=50 3=0
Do you think the hospital staff did everything they could to help to control your pain? [All who were in pain]	1=100 2=50 3=0

Overall, did you feel you were treated with respect and dignity while you were in hospital? [All]	1=100 2=50 3=0

Outpatient Department Survey (2011 questionnaire)

	Scoring (Response = score)
Access and waiting	
How long after the stated appointment time did the appointment start? [All]	1=100 2=83 3=67 4=50 5=33 6=17 7=0 8=M
From the time you were first told you needed an appointment to the time you went to the Outpatients Department, how long did you wait for an appointment? [All]	1=100 2=75 3=75 4=50 5=33 6=17 7=0 8,9=M
Safe, high quality, coordinated care	
Sometimes in a hospital or clinic, a member of staff will say one thing and another will say something quite different. Did this happen to you? [All]	1=0 2=50 3=100
Did a member of staff tell you about what danger signals regarding your illness or treatment to watch for after you went home? [All]	1=100 2=50 3=0 4=M
Did you have confidence and trust in the doctor examining and treating you? [All who saw doctor]	1=100 2=50 3=0
Did you have confidence and trust in him/her? [All who saw other healthcare professional]	1=100 2=50 3=0
Did the doctor seem aware of your medical history? [All who saw doctor]	1=100 2=50 3=0 4=M
Better information, more choice	
Were you involved as much as you wanted to be in decisions about your care and treatment? [All]	1=100 2=50 3=0
Did a member of staff explain the purpose of the medications you were to take at home in a way you could understand? [All prescribed new medication]	1=100 2=50 3=0 4=M
Did a member of staff tell you about medication side effects to watch for? [All prescribed new medication]	1=100 2=50 3=0 4=M

While you were in the Outpatients Department, how much information about your condition or treatment was given to you? [All]	1=50 2=100 3=50 4=0
Before the treatment, did a member of staff explain any risks and/or benefits in a way you could understand? [All who had treatment]	1=100 2=50 3=0 4=M
Building closer relationships	
If you had important questions to ask the doctor, did you get answers that you could understand? [All who saw doctor]	1=100 2=50 3=0 4=M 5=0
If you had important questions to ask him/her, did you get answers that you could understand? [All who saw other healthcare professional]	1=100 2=50 3=0 4=M 5=0
Did doctors and/or other staff talk in front of you as if you weren't there? [All]	1=0 2=50 3=100
Did you have enough time to discuss your health or medical problem with the doctor? [All who saw doctor]	1=100 2=50 3=0
Did the doctor listen to what you had to say? [All who saw doctor]	1=100 2=50 3=0
Clean, comfortable, friendly place to be	
In your opinion, how clean was the Outpatients Department? [All]	1=100 2=67 3=33 4=0 5=M
Were you told how long you would have to wait? [All waiting longer than 15 minutes for their appointment to start]	1=100 2=100 3=50 4=0 5=M
Overall, did you feel you were treated with respect and dignity while you were in the outpatient department? [All]	1=100 2=50 3=0

Emergency Department survey (2016 questionnaire)

	Scoring (Response = score)
Access and waiting	
Sometimes, people will first talk to a nurse or doctor and be examined later. From the time you arrived, how long did you wait before being examined by a doctor or nurse? [All]	1=100 2=80 3=60 4=40 5=20 6=0 7,8=M
Overall, how long did your visit to the emergency department last? [All]	1=100 2=100 3=80 4=60 5=40 6=20 7=0 8=0 9=M
How long did you wait before you first spoke to a nurse of doctor? [All]	1=100 2=67 3=33 4=0 5=M
Safe, high quality, coordinated care	
Sometimes, a member of staff will say one thing and another will say something quite different. Did this happen to you in the emergency department? [All]	1=0 2=50 3=100
Did a member of staff tell you about what danger signals regarding your illness or treatment to watch for after you went home? [All]	1=100 2=50 3=0 4=M
Did you have confidence and trust in the doctors and nurses examining and treating you? [All]	1=100 2=50 3=0
Better information, more choice	
Were you involved as much as you wanted to be in decisions about your care and treatment? [All]	1=100 2=50 3=0 4=M
Did a member of staff explain the purpose of the medications you were to take at home in a way you could understand? [All who were prescribed new medication]	1=100 2=50 3=0 4=M
Did a member of staff tell you about medication side effects to watch for? [All who were prescribed new medication]	1=100 2=50 3=0 4=M

While you were in the emergency department, how much information about your condition or treatment was given to you? [All]	1=50 2=100 3=50 4=0
Building closer relationships	
Did doctors or nurses talk to each other about you as if you weren't there? [All]	1=0 2=50 3=100
Did you have enough time to discuss your health or medical problem with the doctor or nurse? [All]	1=100 2=50 3=0 4=M
While you were in the emergency department, did a doctor or nurse explain your condition and treatment in a way you could understand? [All]	1=100 2=50 3=0 4=M
Did the doctors and nurses listen to what you had to say? [All]	1=100 2=50 3=0
If you had any anxieties or fears about your condition or treatment, did a doctor or nurse discuss them with you? [All]	1=100 2=50 3=0 4=M
Clean, comfortable, friendly place to be	
In your opinion, how clean was the emergency department? [All]	1=100 2=67 3=33 4=0 5=M
Were you given enough privacy when being examined or treated? [All]	1=100 2=50 3=0
Overall, did you feel you were treated with respect and dignity while you were in the emergency department? [All]	1=100 2=50 3=0
Do you think the hospital staff did everything they could to help control your pain? [All]	1=100 2=50 3=0 4=M

Community Mental Health survey (2018 questionnaire)

	Scoring (Response = score)
Access and waiting	
Do you know how to contact this person [the person in charge of organising the respondent's care and services] if you have a concern about your care? [All who were told who was in charge or their care and services]	1 = 100 2 = 0 3 = M
Do you know who to contact out of office hours if you have a crisis? [All]	1 = 100 2 = 0 3 = M
Safe, high quality, coordinated care	
How well does this person [the person in charge of organising the respondent's care & services] organise the care and services you need? [All who were told who was in charge or their care and services]	1 = 100 2 = 67 3 = 33 4 = 0
In the last 12 months have you had a formal meeting with someone from NHS mental health services to discuss how your care is working? [All]	1 = 100 2= 0 3= M
In the last 12 months, has an NHS mental health worker checked with you about how you are getting on with your medicines? (That is, have your medicines been reviewed?) [all who had been receiving medicines for 12 months or longer]	1 = 100 2 = 0 3 = M
In the last 12 months, did NHS mental health services give you any help or advice with finding support for physical health needs (this might be an injury, a disability, or a condition such as diabetes, epilepsy, etc)? [All]	1 = 100 2 = 50 3 = 0 4,5,6 = M
Better information, more choice	
Have you agreed with someone from NHS mental health services what care you will receive? [All]	1=100 2=50 3=0
Were you involved as much as you wanted to be in agreeing what care you will receive? [All who had agreed with NHS mental health services what care they would receive]	1 = 100 2 = 50 3 = 0 4,5 = M
Does this agreement on what care you will receive take your personal circumstances into account? [All who had agreed with NHS mental health services what care they would receive]	1 = 100 2 = 50 3 = 0 4 = M
Were you involved as much as you wanted to be in decisions about which medicines you receive? [All who received medicines in the previous 12 months]	1 = 100 2 = 50 3 = 0 4,5 = M

Were you involved as much as you wanted to be in deciding what NHS therapies to use? [All who in the previous 12 months received NHS therapies that did not involve medicines]	1 = 100 2 = 50 3 = 0 4 = M 5 = M
Building closer relationships	
Were you given enough time to discuss your needs and treatment? [All]	1 = 100 2 = 50 3 = 0 4 = M
Did the person or people you saw understand how your mental health needs affect other areas of your life? [All]	1 = 100 2 = 50 3 = 0 4 = M
Have you been told who is in charge of organising your care and services? (This person can be anyone providing your care, and may be called a “care coordinator” or “lead professional”.) [All]	1 = 100 2 = 0 3 = M
Overall in the last 12 months, did you feel that you were treated with respect and dignity by NHS mental health services? [All]	1 = 100 2 = 50 3 = 0

Annex 2: Confidence intervals for the overall patient experience scores

This is a technical annex that explains how NHS England calculates confidence intervals for the overall patient experience scores, and for the five domain scores that make up the overall measure. The annex is designed for users with an understanding of (mathematical) statistical concepts, rather than the general reader. We calculate confidence intervals for scores at Trust level and at national (England) level.

The method described here uses a variety of different statistical principles to directly calculate the variance of the statistics. We have verified the methodology by comparing the resulting confidence intervals with those produced by Picker Europe using a bootstrapping methodology. The two methods give very similar results.

The method for calculating confidence intervals closely follows the methodology for calculating the scores themselves. We build up the required variance using a series of steps as follows:

Step 1: Question scores at individual level

The first step in calculating scores is to convert questionnaire responses into scores out of 100 using a standard scoring schema, and then to weight those scores at individual level to standardise by age, gender and whether the patient was an emergency or elective case.

There are typically around 400 responses for a single question within a single NHS Trust and we consider these to be a simple random sample from the population of patients at that Trust. The Trust level mean score for question j and Trust k is given by:

$$\overline{x_{jk}} = \frac{\sum_i W_i x_{ijk}}{\sum_i W_i}$$

Here, the x_{ijk} represent the question scores from each individual patient and the W_i represents the weight for that individual. The weight is a standardisation weight, and flows directly from the age, gender, and admission method of patient i.

The sample variance of the individual question scores (x_{ijk}) for question j at Trust k is given by:

$$s_{jk}^2 = \frac{\sum_i W_i \left(x_{ijk} - \overline{x_{jk}}\right)^2}{\sum_i W_i}$$

Note that for ease of calculation, we have used the biased estimate for variance (The $\sum_i W_i$ here replaces the usual 'n' in variance calculations. The equivalent formula for '$n - 1$' involves sums of squared weights etc.) As the sample size is very large, using this biased estimator does not substantively alter the variance.

The sample standard error of the Trust level average question score is then given simply by:

$$SE_{jk}^2 = \frac{s_{jk}^2}{\sum_i W_i} = \frac{\sum_i W_i(x_{ijk} - \overline{x_{Jk}})^2}{(\sum_i W_i)^2}$$

Covariances between pairs of questions can be calculated in a similar way:

$$\mathrm{Cov}\left(\overline{x_{J_1k}}, \overline{x_{J_2k}}\right) = \frac{\sum_i \left(W_\mathrm{i}\left(x_{ij_1k} - \overline{x_{J_1k}}\right).\left(x_{ij_2k} - \overline{x_{J_2k}}\right)\right)}{\sum_i W_i}$$

Note, the $\mathrm{W_i}$ here would be set to zero in cases where patient i had not answered both questions for which we wish to calculate covariance, and hence the covariance would not be calculated.

Step 2: Using question scores to calculate domain scores and the overall measure

In practice, NHS England work out national level scores for each question next (by averaging the $\overline{x_{Jk}}$ across trusts k) then we combine the scores into domains. It is algebraically equivalent to work our domain scores at Trust level first, and then to average domains across Trusts. This approach makes it slightly easier to describe the method for calculating variances. We present the analysis here as if we start by working out domain scores for each Trust, and then aggregate up to England level results.

Each of the domains is defined as a linear function of the average question scores. Let d_{lk} denote the lth domain score in Trust k. Here, the values of l range from 1 to 6, with 6 denoting the overall score.

The domain scores for Trust k are defined as follows:

$$d_{1k} = \frac{1}{3}(\overline{x_{1k}} + \overline{x_{2k}} + \overline{x_{3k}})$$

$$d_{2k} = \frac{1}{3}(\overline{x_{4k}} + \overline{x_{5k}} + \overline{x_{6k}})$$

$$d_{3k} = \frac{1}{3}(\overline{x_{7k}} + \overline{x_{8k}} + \overline{x_{9k}})$$

$$d_{4k} = \frac{1}{4}(\overline{x_{10,k}} + \overline{x_{11,k}} + \overline{x_{12,k}} + \overline{x_{13,k}})$$

$$d_{5k} = \frac{1}{6}\left(\frac{1}{2}(\overline{x_{14,k}} + \overline{x_{15,k}}) + \overline{x_{16,k}} + \overline{x_{17,k}} + \overline{x_{18,k}} + \overline{x_{19,k}} + \overline{x_{20,k}}\right)$$

For simplicity we have re-numbered the question scores so that questions 1, 2 and 3 feed into domain 1 and so on. Also note that in domain 5 the first two questions are averaged first to form a single score for ‘noise at night’. This was for historical reasons, to allow comparison with data from 2001 when a single survey question was asked.

The overall measure is given by:

$$d_{6k} = \frac{1}{5}(d_{1k} + d_{2k} + d_{3k} + d_{4k} + d_{5k}).$$

It is easy to see that each of the domains, and the overall score, can be considered as some linear combination of question scores:

$$d_{lk} = \sum V_p . \overline{x_{pk}}$$

where V_p denotes some set of linear weights, and p ranges as appropriate to define the formula. It is also clear, from even a very brief examination of the data, that the $\overline{x_{jk}}$ are correlated with each other and therefore cannot be considered independent.

The variance of the domain scores is therefore given by the general formula:

$$s_{lk}^2 = \sum_p V_p^2 . s_{pk}^2 + 2 \sum_p V_{p_1} . V_{p_2} \mathrm{Cov}(\overline{x_{p_1 k}}, \overline{x_{p_2 k}}).$$

Applying this formula gives us variance values for each domain, and the overall score, for each Trust k.

Step 3: Aggregating to national level

The national (England) level scores are calculated as a simple average (mean) of Trust level scores:

$$\overline{x_j} = \frac{\sum \overline{x_{jk}}}{K}$$

and;

$$\overline{d_l} = \frac{\sum \overline{d_{lk}}}{K}$$

where K is the number of Trusts (typically around 150).

To calculate the variance at England level, we consider each of the Trust level scores to represent the result of one ‘experiment’ to measure the true national average. Each experiment might have a different mean (because Trusts vary), but the underlying variability is assumed to be roughly the same for all Trusts. This is a description of a standard situation in which a pooled variance calculation is appropriate.

The standard formula for pooled sample variance is:

$$s_p^2 = \frac{\sum_{i=1}^{k}\left((n_i - 1).s_i^2\right)}{\sum_{i=1}^{k}(n_i - 1)}$$

Note that this gives a value for the 'average' variance at Trust level. To obtain an estimate of sample standard error at national level, we divide again by 'big N', to give:

$$SE_p{}^2 = \frac{\sum_{i=1}^{k}\left((n_i - 1).s_i^2\right)}{\left(\sum_{i=1}^{k}(n_i - 1)\right)^2}$$

or in our notation, for domain l:

$$SE_l^2 = \frac{\sum_{k=1}^{K}\left((J_k - 1).s_{lk}^2\right)}{(\sum_{k=1}^{K}(J_k - 1))^2}$$

Where $\mathrm{J_k}$ is the overall number of respondents in Trust k. This could equally be calculated for an individual question score.

There is one final note on this formula. The number of respondents J_k will not necessarily equal the number of useable responses for each individual question. Some respondents will not answer all questions. It would be possible to construct a more complicated formula, by going back to original questions scores and calculating the overall variance directly as a linear combination of 150 * 20 values (150 Trusts, 20 questions, in a very large linear expression). However, this would be computationally complex and our testing has shown that this does not make a material difference to the results of significance testing.

We observe three points that validate this simplifying assumption:

1. The resulting variances give a close match to those produced by boot-strapping
2. They are used as relative weights in the above formula, and it is reasonable to assume that the relative weights between Trusts remains reasonably constant across questions
3. The range of weights is not large, Trusts typically have roughly the same number of respondents overall

Confidence intervals

The sample sizes in these surveys are very large, and the central limit theorem applies. We can assume that the domain scores and overall patient experience measure follow a normal distribution. The 95% confidence intervals are therefore given by:

$$CI(\overline{x_{jk}}) = \pm 1.96 SE_{jk}$$

$$CI(\overline{x_j}) = \pm 1.96\, SE_j$$

$$CI(d_{lk}) = \pm 1.96\, SE_{lk}$$

$$CI(d_l) = \pm 1.96\, SE_l$$

These four values give us confidence intervals for question scores at Trust level, question scores at national level, domain scores at Trust level and domain scores at national level.

Statistical testing

For comparisons across years, we have used the pooled t-test procedure. This is appropriate in instances where we can assume that the two populations have the same standard deviation.

The generic formula for the pooled estimate of the variance is given by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

where s_1^2 and s_2^2 are the sample variances for the two samples, and n_1 and n_2 are the number of observations in each sample.

In our cases, the two samples will be of question or domain scores (including the OPES score itself) for either a single Trust or all of England. Although for any given question, the number of observations will be different, for this calculation we use the number of survey respondents at the Trust (or for all of England). Experimentation has shown that using each of the different possible choices of n does not affect the outcome of significance testing.

The t-statistic formula is

$$t = \frac{\overline{x_1} - \overline{x_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $\overline{x_1}$ and $\overline{x_2}$ are the two means to be compared. The number of degrees of freedom is $n_1 + n_2 - 2$. A one-tailed test at the 95% confidence level is conducted to test whether an increase or decrease from one year to the next is statistically significant (i.e. if there’s a less than 5% chance that the increase or decrease could have happened by chance).