Care Quality Commission

CQC Insight

NHS Trusts and Community Interest Companies that provide specialist mental health services

Statistical methodology

The purpose of this document is to provide a detailed description of the statistical methods used to support CQC Insight for NHS mental health services.

Contents

1. Introduction

This document describes, in some detail, the statistical methods we have used to analyse the data that supports the CQC Insight risk assessment model.

This analysis is relevant to the majority of indicators, but not all, as some are already analysed by external organisations. Indicators that are analysed externally are summarised in section 6.

Our general approach for the model is to assess risk by comparing a trust's observed outcomes with others. Where appropriate, we account for the relative sizes of trusts and, in several cases, their variable case mix.

We use a cross-sectional analysis, which assesses variation by comparing practice outcomes over a fixed period of time. Previous values or trends are not accounted for.

2. Analysis of cross-sectional data using z-scores

2.1 Z-scores

With cross-sectional data we measure the deviation of observed values from an expected or target value. Where we can transform the data into a standard normal distribution we generate z-scores which reflect the number of standard deviations from the mean.

If the trust value for an indicator is *y*, and it has an expected or target value *t*, we can express the deviation of the indicator from the expected value as a z-score, defined as:

$$z = \frac{y - t}{s_0}$$

where s_0 is the standard deviation of *y* if the trust's observed outcomes were randomly distributed about *t*.

Here z is referred to as the unadjusted z-score. Under a null hypothesis that a trust's true level of outcomes is exactly the same as the expected value, z has mean 0 and standard deviation 1, and if we assume normality, then p-values 0.025 and 0.001 correspond to z = ± 1.96 and z = ± 3.10 respectively, which corresponds very closely to 2 and 3 standard deviations from the mean.

The default expected values against which a trust is compared are calculated by comparing rates observed for an individual trust against the national rates. However, for some items we standardise by case mix (for example, by age and sex) in order to compare observed outcomes against what you would expect if the rate for each patient was the same as for similar patients over the whole country. Often the raw data are not normally distributed, in which case we use one of the following appropriate transformations.

2.1.1 Z-scores from proportions

Assume an observed proportion y = $^{r}/_{n}$, with an expected or target proportion p. The observed proportion is transformed to render it more normally distributed by applying an arcsine transformation to the square root of the observed proportion:

$$Y = arcsin\sqrt{\frac{r}{n}}$$

The expected value can be approximated by:

$$T = arcsin\sqrt{p}$$

and the standard deviation *(s)* is approximated by:

$$s = \frac{1}{2\sqrt{n}}$$

Hence the transformed unadjusted z-score:

$$z = \frac{Y - T}{s} = 2\sqrt{n}\left(arcsin\sqrt{\frac{r}{n}} - \arcsin\sqrt{p}\right)$$

2.1.2 Z-scores from standardised ratios

This method is used when comparing an observed value against an expected value derived using indirect standardisation.

We assume a standardised ratio $y\ = {}^{O}/_{E}$ based on an observed count O and an expected count E.

A square root transformation is applied to the standardised ratio (y):

$$Y = \sqrt{\frac{O}{E}}$$

which has an expected value approximately equal to one.

Under appropriate Poisson assumptions, the standard deviation approximates to:

$$s = \frac{1}{2\sqrt{E}}$$

Thus, the transformed unadjusted z-score is given by:

$$z = \frac{Y-1}{s} = 2\left(\sqrt{O} - \sqrt{E}\right)$$

2.1.3 Z-scores from ratios of counts

We assume a ratio indicator of the form $y = {}^{O_1}/_{O_2}$, where $O_1$and $O_2$are both counts, and an average or target ratio *t*.

In order to deal with zero/low counts we add 0.5 to all observations, and, noting that a log transformation reduces positive skewness, the transformed indicator becomes:

$$Y = log_e\left(\frac{O_1 + 0.5}{O_2 + 0.5}\right)$$

with an expected value approximately equal to

$$\mathrm{T} = log_e(t)$$

and a standard deviation:

$$s = \sqrt{\frac{O_1}{(O_1 + 0.5)^2} + \frac{O_2}{(O_2 + 0.5)^2}}$$

Thus the transformed, unadjusted z-score becomes:

$$z = \frac{Y-T}{s} = \frac{log_e[(O_1 + 0.5)/(O_2 + 0.5)] - log_e(t)}{\sqrt{O_1/(O_1 + 0.5)^2 + O_2/(O_2 + 0.5)^2}}$$

If either O_1 or O_2 is much bigger than the other, say when one represents a population, it will have a negligible impact on the score.

2.1.4 Z-scores from percentages

We assume an indicator that consists of an observed percentage p, where the numerator and denominator are not available. We can then use the mean percentage across all providers and the standard deviation of the percentages to calculate the Z-scores.

The Z-score for a provider is then given by:

$$\mathrm{Z} = \frac{(\mathrm{p} - \bar{\mathrm{p}})}{\mathrm{s}}$$

Where p is the percentage for the provider, $\bar{p}$ is the mean percentage across all providers (or the target), and s is the standard deviation of the percentages across all providers.

2.1.5 Low numbers z-scores (working z-scores)

Where data do not meet numerator and denominator requirements, “working z-scores” are more appropriate as it is not possible to generate sufficiently robust z-scores using methods based on the normal distribution (i.e. those described in sections 2.1.1 to 2.1.4).

Calculating “working z-scores” is a two-step process. The first step is to determine which alternative statistical distribution is the best fit to the data.

Different types of data will have a different subset of potentially applicable statistical distributions (Table 1).

Table 1: Statistical distributions tested for various data types

Statistical distribution					
Data type	Binomial*	Poisson	Zero-inflated Poisson	Negative Binomial	Zero-inflated Negative Binomial
Raw count	✓	✓	✓	✓	✓
Ratio of counts	×	✓	✓	×	×
Standardised ratio	×	✓	✓	×	×
Proportion	✓	✓	✓	✓	✓

*Note: The Binomial distribution is used for proportions if the denominator in any time period is less than 30 and on average $0.1<p<0.9$. The remaining applicable distributions are tested if the conditions for the Binomial distribution are not satisfied.

The distributions are tested using regression models, where the dependent variable is the numerator and the independent variable is the denominator. If there is no denominator we use a null model (i.e. a model without any independent variables).

For each model, a measure of model fit called Akaike’s Information Criterion (AIC) is obtained. The model with the lowest AIC is selected as being the better fit to the data.

The second step is to use the best-fit distribution to has generate a p-values for each observation in the dataset and convert these to z-scores using the normal distribution. This is done as follows:

i. Use the mid-p-value method to calculate the probability of falling below the observed data point by chance alone, based on the best-fit distribution. Note: this might be 1 minus the usual p-value, which is a tail measure.

   The mid-p-value for O, as defined by Speigelhalter et al (2012), is calculated as follows:

$$p = Prob(Y < O) + \frac{1}{2} Prob(Y = O)$$

   where Y has the best-fit distribution. For computational simplicity it is easier to write it in a different form. Let F(.) be the cumulative distribution function of the best-fit distribution and P(.) the associated probability mass function. An equivalent expression for p is

$$p = F(O) - \frac{1}{2} P(O)$$

ii. The next and final step is to convert p to the corresponding working Z-score from a standard Normal distribution. Let ϕ(.) be the standard Normal cumulative distribution function. Then

$$Z_3 = \Phi^{-1}(p)$$

The z-scores produced using this method then undergo the same testing and possible adjustment for over-dispersion (2.2) as would z-scores produced using our standard methods.

2.2 Over-dispersion

Many z-scores are likely to be over-dispersed, that is their true variances are greater than one, which may be because of insufficient benchmarking or the presence of common-cause factors that render the Poisson model inadequate. The consequence is that analyses may pick up statistically significant differences that are not of practical importance. When considering an outcome based on an ‘average’ or ‘expected’ level, it may then be reasonable to accept as inevitable a degree of between-trust variability and we therefore seek to identify trusts that deviate from this distribution, rather than deviating from a single measure. In order to do this we must estimate the degree of over-dispersion (see section 2.2.2). When estimating over-dispersion it may be better to do so using techniques that avoid undue influence of outlying trusts, such as winsorisation (see section 2.2.1).

The significance of observed deviations then takes into account both the precision with which the indicator is measured within each trust (i.e. the sample size), and the estimated between-trust variability.

2.2.1 Winsorisation

Winsorisation is the process of transforming outliers in statistical data. In this context it involves shrinking in extreme unadjusted z-scores to the value of a selected percentile. This is done by:

1. Ranking trusts according to their unadjusted z-scores.
2. Identifying z_q and z_{1-q}, the $100q\%$ most extreme high and low unadjusted z-scores, where *q* may be, for example, 0.1.
3. Setting the lowest $100q\%$ of unadjusted z-scores to z_q and the highest $100q\%$ of z-scores to z_{1-q}. These are the winsorised statistics.

This process retains the same number of Z-scores, but protects our estimation of over-dispersion from the influence of actual outliers.

2.2.2 Estimating over-dispersion

In calculating an adjusted z-score for an indicator, we estimate the over-dispersion factor phi (ϕ) as follows:

$$\hat{\phi} = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_i^2$$

where n is the number of trusts for a data item and $\hat{z}_i$ is the winsorised z-score for the ith trust.

Under a null hypothesis that all units only exhibit random variability around the expected value, which is derived from the data, $n\hat{\phi}$ has an approximate χ^2_{n-1} distribution. This can therefore be used as a standard test of heterogeneity.

2.2.3 Calculating adjusted Z-scores

We then use the resulting over-dispersion factor to calculate an adjusted z-score for each observation.

The over-dispersion model we use is an additive random effects model. This model assumes that each trust has its own true underlying level t_i, and that for ’on-target’ providers t_i is distributed with mean t_0 and standard deviation,τ. In other words the null hypothesis is represented by a distribution rather than a single point.

A standard method of moments estimate for τ^2 is:

$$\hat{\tau}^2 = \frac{n\hat{\phi} - (n-1)}{\sum_{i=1}^{n} w_i - \left(\sum_{i=1}^{n} w_i^2 / \sum_{j=1}^{n} w_j\right)}$$

Where $w_i = {}^{1}\!/_{s^2}$ and $n\hat{\phi}$ is the test for heterogeneity. (s is as calculated in section 2.1 with the appropriate transformation.)

If $n\hat{\phi} \geq (n-1)$ then the adjusted Z-scores are given by:

$$z_i^* = \frac{(z_i - t_0)}{\sqrt{s_i^2 + \hat{\tau}^2}}$$

where z_i is equal to the raw z-score value.

Otherwise, if $n\hat{\phi} < (n-1)$, τ^2 is set to zero, complete homogeneity is assumed and no adjustments are necessary.

3. Cross-sectional analysis of raw counts data

3.1 Poisson events

In some instances, for example, when monitoring never events, observations may be sufficiently infrequent that it is not possible to generate sufficiently robust z-scores.
Where there is no evidence of over-dispersion, we can assume the events are Poisson distributed and establish levels of risk based on the probability any observed outcome could have happened by chance (the p-value).

Suppose X is a random variable representing the number of events reported at a trust over a given period of time and that λ represents the expected number of events, based on national reported rates. If n events are observed, then a p-value can be expressed as:

$$p(X > n) = 1 - p(X \leq n) + 0.5\, p(X = n) = 1 - \sum_{i=0}^{n} \frac{\lambda^i}{i!} e^{-\lambda} + \frac{\lambda^n}{2n!} e^{-\lambda}$$

(where the latter term of the formula is used as a *mid-P-value*).
These p-values can then be used in correspondence with given thresholds of significance to define levels of risk.

3.2 Negative binomial distributions

3.2.1 Fitting a model to the data

For some events the Poisson assumptions may not provide a good fit to the data, in which case a negative binomial distribution may be more appropriate. Also, for some indicators there may be a disproportionate number of zeroes among the trust-level values, necessitating a zero-inflated model.

Some of our count-based indicators are correlated with the volumes of patients seen by trusts, and so our assessment of risk needs to identify trusts that have unusually high counts compared with what would be expected given their patient volumes (measured in either bed-days or total patient contacts, both scaled by 100,000). Other indicators consist of negative and positive comments, the numbers of which are often correlated. Here, our assessment of risk needs to identify trusts with unusually high counts of negative comments compared with what would be expected given their count of positive comments.

The negative binomial distribution is expressed as:

$$f(y_i|x_i) = \frac{\Gamma(y_i + \theta)}{y_i!\,\Gamma(\theta)}\left(\frac{\theta}{\theta + \mu_i}\right)^{\theta}\left(\frac{\mu_i}{\theta + \mu_i}\right)^{y_i}, y_i = 0, 1, 2, \ldots$$

Where $\mu_i$is the conditional mean, and θ is a positive gamma distribution parameter used to determine the conditional variance.

The zero-inflated negative binomial model has two parts – a negative binomial count model as above, and a logistic regression model for predicting excess zeroes. Additional zeroes occur with the probability φ_i as determined by:

$$\varphi_i = f(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

where f is the logistic function and t is typically a linear function of one or more explanatory variables.

For a given count indicator, we determine which probability distribution – Poisson, negative binomial, zero-inflated Poisson, or zero-inflated negative binomial – is the best fit to the data, by modelling the raw count as a function of the most appropriate comparator – bed-days, total patient contacts, or count of positive comments.

If Y is the count indicator and X the comparator, our model seeks to determine:

$$\Pr(Y = y|X = x)$$

with estimated parameters $\hat{\mu}$ and $\hat{\theta}$. If there are no zeros, then we fit only the negative binomial model. If there are zeros, we fit each of the chosen distributions in turn. The measure of the fit of each model is expressed as a log-likelihood.

The ratio of the log-likelihoods of the two models is approximately chi-square distributed with degrees of freedom equal to $df2$ – $df1$, where, $df2$ is the degrees of freedom for the zero-inflated model (which is more complex) and $df1$ is the degrees of freedom for the ordinary negative binomial model. A statistically significant likelihood ratio test indicates that one model is a better fit than the other.

3.2.2 Identifying extreme values

Given the model of best fit, we iteratively establish levels of risk based on the probability that the most extreme outcome could have happened by chance (the p-value). Each iteration of the model comprises a series of steps, as follows:

i. First we condition on Y being positive, such that:

$$\Pr(Y = y|X = x, Y > 0) = \frac{\Pr(Y = y|X = x)}{\Pr(Y > 0|X = x)} = \frac{\{1 - p_0(x)\}f(y|x)}{1 - f(0|x)}$$

This conditioning means that we do not need to be overly concerned about the point-mass at zero for the zero-inflated models as we are primarily interested in non-zero counts.

ii. Next, we find the trust-level p-values. For example, for trust j:

$$p_j = \sum_{y \geq y_j} \Pr(Y = y|X = x_j, Y > 0)$$

iii. We find the smallest p-value across all trusts (p_{min}), and use this to calculate the group-level p-value (p_g), which accounts for multiple comparisons:

$$p_g = 1 - (1 - p_{min})^M$$

This is the probability that none of a sample size of M is more extreme than p_{min}. It is a measure of how much of an outlier p_{min} actually is.

iv. If p_g is small ($\leq$ 0.20) then we conclude that the count associated with $p_{min}$is too high to have come from the fitted model. Trusts meeting both the $p = p_{min}$ and $p_g \leq 0.20$ criteria typically have very small p-values (much less than 0.05), and are therefore assigned to the “Elevated Risk” category. This trust is then removed from the dataset, and steps i. through iv. are repeated.

If p_g is large ($>$ 0.20) then the trust with $p = p_{min}$ is not an outlier, and we stop iterating the model. Of the trusts remaining in the dataset at this point, any with p-values less than or equal to 0.05 are assigned to the “Risk” category.

4. Aggregate scoring methods

4.1 Electronic Staff Record (ESR) data

From the data provided a total of 6 individual items relating to staffing were created, each of which was allocated into one of the following categories:

- Sickness
- Staff Ratios
- Support/Supervision.

A list of items is shown in table 2.

Table 2: Electronic Staff Record item list and parameters

Area	Common area parameters	Item Code	Item Name	Parameters
Sickness	Abs Rate Person Dim.User Person Type LIKE 'Employee%'	MHWEL137	Proportion of days sick in the last 12 months for Medical and Dental staff	Abs Rate Staff Group Dim.Staff Group LIKE 'Medical and Dental'
	(((Abs Rate Staff Group Dim.Staff Group = 'Medical and Dental') AND (Abs Rate Assignment Dim.Contracted Wte <= 1.2)) OR ((Abs Rate Staff Group Dim.Staff Group <> 'Medical and Dental') AND (Abs Rate Assignment Dim.Contracted Wte BETWEEN 0.05 and 1)))	MHWELL138	Proportion of days sick in the last 12 months for Nursing and Midwifery staff	Abs Rate Staff Group Dim.Staff Group LIKE 'Nursing and Midwifery Registered'
		MHWEL139	Proportion of days sick in the last 12 months for other clinical staff	Abs Rate Staff Group Dim.Staff Group IN ('Add Prof Scientific and Technic', 'Additional Clinical Services', Allied Health Professionals', 'Healthcare Scientists')
		MHWEL140	Proportion of days sick in the last 12 months for non-clinical staff	Abs Rate Staff Group Dim.Staff Group IN ('Estates and Ancilliary', 'Administrative and Clerical', 'Students')
Support	Abs Rate Person Dim.User Person Type LIKE 'Employee%' Primary area of work = 'Psychiatry' AND secondary area of work NOT IN ('Child and Adolescent Psychiatry', 'Medical Psychotherapy') (((Wfc Staff Group Dim.Staff Group = 'Medical and Dental') AND (Wfc Fact.Contracted WTE for Assignment <= 1.2)) OR ((Wfc Staff Group Dim.Staff Group <> 'Medical and Dental') AND (Wfc Fact.Contracted WTE for Assignment BETWEEN 0.05 and 1)))	MHESR01	Proportion of registered nursing staff	Wfc Staff Group Dim.Staff Group = 'Nursing and Midwifery Registered' AND Wfc Staff Group Dim.Job Role NOT IN ('Midwife', 'Midwife - Consultant', 'Midwife - Manager', 'Midwife - Specialist Practitioner', 'Student Midwife', 'Community Nurse', 'Community Practitioner') (Numerator)
				(Wfc Staff Group Dim.Staff Group = 'Additional Clinical Services' AND Wfc Staff Group Dim.Job Role IN ('Healthcare Assistant', 'Health Care Support Worker', 'Helper/Assistant')) OR (Wfc Staff Group Dim.Staff Group = 'Nursing and Midwifery Registered' AND Wfc Staff Group Dim.Job Role NOT IN ('Midwife', 'Midwife - Consultant', 'Midwife - Manager', 'Midwife - Specialist Practitioner', 'Student Midwife', 'Community Nurse', 'Community Practitioner')) (Denominator)
Staff Ratios	Abs Rate Person Dim.User Person Type LIKE 'Employee%'	MHESR02	Ratio of occupied beds to all nursing staff	Wfc Staff Group Dim.Staff Group = 'Nursing and Midwifery Registered' AND Wfc Staff

	Primary area of work = ‘Psychiatry’ AND secondary area of work NOT IN (‘Child and Adolescent Psychiatry’, ‘Medical Psychotherapy’) (((Wfc Staff Group Dim.Staff Group = 'Medical and Dental') AND (Wfc Fact.Contracted WTE for Assignment <= 1.2)) OR ((Wfc Staff Group Dim.Staff Group <> 'Medical and Dental') AND (Wfc Fact.Contracted WTE for Assignment BETWEEN 0.05 and 1)))			Group Dim.Job Role NOT IN ('Midwife', 'Midwife - Consultant', 'Midwife - Manager', 'Midwife - Specialist Practitioner', 'Student Midwife', 'Community Nurse', 'Community Practitioner')

5. Analyses carried out by external organisations

The previous sections describe the analysis that has been carried out by CQC where appropriate.

Several indicators are already analysed by external organisations, and in such cases we report the results of the external analysis.

Indicators that are analysed externally are shown in tables 3 below.

Table 3: Community Mental Health Survey indicators analysed externally

Indicators	Source of information	Summary of analysis
Community Mental Health Survey	CQC/Picker Institute Europe	The Picker Institute Europe run this on behalf of CQC, and calculate modified z-scores for trusts which take into account not just the distribution of trust-level results but also the sample size within the trust. We have used these modified z-scores to assign positive and negative bandings: Much worse than expected: Z-score ≤ -3.09 Worse than expected: Z-score ≤ -1.96 About the same: -1.96 > Z-score < 1.96 Better than expected: Z-score ≥ 1.96 Much better than expected: Z-score ≥ 3.09

6. Further reading

Cross-sectional analyses using z-scores and funnel plots

Spiegelhalter D J. Funnel plots for comparing institutional performance. *Stat Med* 2005; **24**: 1185-1202.