Care Quality Commission

Quality and Risk Profiles

Statistical guidance

Outcome-based risk estimates in QRPs produced for NHS providers

March 2013

Contents

Introduction

This guide explains the analytical methods that we use in our statistical model that calculates risk estimates of non-compliance with the essential regulatory outcomes. It is a technical guide and assumes that those reading it have some statistical knowledge.

The model uses a vast amount of information and brings together many different types of data. For each type of data, the analysis method we use is designed to measure the difference between the observed result and an expected level of performance on a common scale.

For each data item (i.e. indicator), we calculate a Z-score for every observation (i.e. provider), which is a statistical measure of how far the observed value differs from what we would expect.

For each provider, we then aggregate all the Z-scores for items mapped against each outcome to calculate an estimated likelihood of non-compliance with each outcome.

We explain these steps further in the following sections.

1. Z-scores: a common scale for measuring deviation

For each data item of information (i.e. indicator) we calculate a Z-score for every observation – this measures how far the observed value (e.g. the performance of a provider) differs from what we would expect.

We assume an indicator ***y*** with an expected or target value ***t***, and express the deviation of the indicator from the expected value as a Z-score, defined as:

$$Z = \frac{(y - t)}{s^0} \qquad (1)$$

where $\boldsymbol{s^0}$ is the standard deviation of ***y*** if the provider were to exactly meet the expected value.

Here ***Z*** is referred to as the unadjusted Z-score. Under a null hypothesis that a provider's performance 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 expected level of performance against which a provider is compared can be calculated in several ways. For some items, we compare providers against the national average of all providers. In other cases, an expected level of performance (target) has been set for providers in government policies.

We recognise that performance on some data items is affected by factors beyond a provider's control. In such cases, we have ways of adjusting the analysis to present the fairest reflection of performance. There are two main ways in which we achieve this. We can either standardise the 'raw' data for the provider (for example by age and sex) or we may set our expectation for that provider as the average performance of a group of other organisations with similar local circumstances (referred to as the 'benchmark group').

Details of the calculation of the unadjusted Z-scores for different types of data are as follows.

1.1 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 both the square root of the observed proportion and to the target:

$$Y = \arcsin\left(\sqrt{\frac{r}{n}}\right)$$

$$T = \arcsin\left(\sqrt{p}\right)$$

The standard deviation *(**s**)* is given by:

$$s = \sqrt{\frac{1}{(4 * n)}}$$

where ***n*** is the denominator.

Hence the transformed unadjusted Z-score is:

$$Z = \frac{(Y - T)}{s} = \frac{\arcsin\left(\sqrt{r/n}\right) - \arcsin\left(\sqrt{p}\right)}{\sqrt{1/(4 * n)}} \qquad (2)$$

1.2 Standardised ratios

Assume a standardised ratio ***y* = *O/E*** based on an observed count ***O*** and an expected count ***E***. We assume an expected or target ratio ***t***.

A square root transformation is applied to both the standardised ratio *(**y**)*, and the target *(**t**)*:

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

$$T = \sqrt{t}$$

The standard deviation *(**s**)* is given by:

$$s = \sqrt{\frac{1}{(4 * E)}}$$

Thus, the transformed unadjusted Z-score is:

$$Z = \frac{(Y - T)}{s} = \frac{\sqrt{O/E} - \sqrt{t}}{\sqrt{1/(4 * E)}} \qquad (3)$$

1.3 Ratios of counts

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

We apply a logarithmic transform to the ratios to ensure that is it irrelevant which item is in the numerator or denominator. When either $\boldsymbol{O_1}$ or $\boldsymbol{O_2}$ is much bigger than the other, say when one is a population, it will have a negligible impact on the Z-score.

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 indicator and expected values become:

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

$$T = \log_e(t)$$

The standard deviation *(**s**)* is given by:

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

Thus, the transformed unadjusted Z-score is:

$$Z = \frac{(Y - T)}{s} = \frac{\log_e\left[(O_1 + 0.5)/(O_2 + 0.5)\right] - \log_e(t)}{\sqrt{\left(O_1/(O_1 + 0.5)^2\right) + \left(O_2/(O_2 + 0.5)^2\right)}} \qquad (4)$$

1.4 Ordinal data

We assume that the categories are due to grouping of a 'latent' normally distributed quantity, so that the labels given to the categories are completely irrelevant, and all that matters is the proportion in each category and the order of the categories (i.e. worst category to best category). First, the 'cut-offs' in a standard normal distribution ***N(0, 1)*** are found that would give the observed proportions: for example, if a three-category response had observed proportions 70%, 20%, and 10% in the three categories, this would represent cut-offs of 0.52 and 1.28, which divide a standard normal distribution into the required proportions. We then assign a Z-score to each category corresponding to the mean ***N(0, 1)*** response within that category, so that in the above example we would assign the first category a mean of a ***N(0, 1)*** variable constrained to lie between negative infinity (***-Inf***) and 0.52, which turns out to be -0.5, using the argument:

If Z **~ N(0, 1)** then:

$$Exp(Z \mid a < Z < b) = \frac{-(\phi(b) - \phi(a))}{\Phi(b) - \Phi(a)} \qquad (5)$$

Where ***ϕ(x)*** is the standard normal probability density function, and ***Φ(x)*** is the standard normal cumulative distribution function.

For the three-category example above, the comparative results are:

Category label	%	Z-scores using 'latent' method
0	70	-0.50
1	20	0.86
2	10	1.75

In some cases we do not want to base the categorical Z-score on the proportion of observations in each category. Instead, we may wish to set a particular category as the target value. In this case we calculate the Z-scores, but the distribution is then shifted to ensure that the required target group has a Z value of zero. This is achieved by adjusting the calculated Z value for each group by the difference between the Z value for the target group and zero.

Using the same example as before, and assuming that the target group is category 1, the following results would be produced:

Category label	%	Z-scores using 'latent' method
0	70	-1.36
1	20	0
2	10	0.89

1.5 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:

$$Z = \frac{(p - \overline{p})}{s} \qquad (6)$$

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

2. Winsorisation and over-dispersion

Data items may show substantially more variability than would be expected by chance alone. Over-dispersion is more likely to occur when an item is based on large numbers of observations, as these items have a greater precision. However, this is not always the case.

The consequence is that analyses may pick up statistically significant differences that are not of practical importance. When considering an outcome based on 'average' performance, it may then be reasonable to accept as inevitable a degree of between-provider variability in performance and we therefore seek to identify providers 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). When estimating over-dispersion it may be better to do so using techniques that avoid undue influence of outlying providers, such as Winsorisation (see Section 2.1).

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

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 providers according to their unadjusted Z-scores.
2. Identifying Z_q and Z_{1-q}, the **100_q%** most extreme high and low unadjusted Z-scores, where *q* may be, for example, 10% (or 0.1).
3. Setting the lowest **100_q%** of unadjusted Z-scores to Z_q and the highest **100_q%** of Z-scores to Z_{1-q}. These are the winsorised statistics (now called ***Z'***).

Where the percentile rank (p_Z) of each provider's score is given by:

$$p_Z = \frac{100}{N}\left(n - \frac{1}{2}\right)$$

where ***N*** is the total number of providers, and ***n*** is the ranked position of a given provider. Then the winsorised Z-score ***(Z')*** is calculated as:

$$Z' = \forall Z\left(\begin{bmatrix} p_Z < 100q \\ 100q \le p_Z \le 100 - 100q \\ p_Z > 100 - 100q \end{bmatrix} \rightarrow \begin{bmatrix} Z_q \\ Z \\ Z_{1-q} \end{bmatrix}\right)$$

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

2.2 Estimating over-dispersion

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

$$\hat{\phi} = \frac{1}{I}\sum_{i=1}^{I} Z_i^2 \qquad (7)$$

where I is the number of providers for a data item and Z_i is the winsorised Z-score (from Z') for the ith provider for the item (as calculated for the relevant data type in Section 1).

A standard test of heterogeneity is:

$I * \hat{\varphi}$

which has an approximate χ_I^2 distribution under a null hypothesis that all units only exhibit random variability around the average performance.

2.3 Calculating adjusted Z-scores

We then use the resulting over-dispersion factor to calculate an adjusted Z-score for each observation. We use an additive random effects model for adjusting for over-dispersion here.

The additive model assumes that each provider has its own true underlying level t_i, and that for non-standard 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. τ^2 can be estimated using a standard method of moments (1).

$$\hat{\tau}^2 = \frac{I * \hat{\phi} - (I - 1)}{\sum_{i=1}^{I} w_i - \left(\sum_{i=1}^{I} w_i^2 / \sum_{i=1}^{I} w_i \right)} \qquad (8)$$

Where $w_i = 1/s_i^2$ and $I * \hat{\varphi}$ is the test for heterogeneity: if $I * \hat{\varphi} < (I - 1)$, then $\hat{\tau}^2$ is set to zero and complete homogeneity is assumed. Otherwise the adjusted Z-scores are given by:

$$Z = \frac{(Y' - t_0)}{\sqrt{s^2 + \hat{\tau}^2}} \qquad (9)$$

Where Y' is equal to the winsorised data value corresponding to Z' in Section 2.1. As a final step, Z-scores are constrained to be between -3.0 and +3.0.

3. Utility factors

Each item of information is assigned three values in relation to the following utility factors, all of which are on a three-point scale (1 = low, 3 = high).

Strength (or Causality – CS) – this is how closely the item of information relates to the outcome(s) to which it has been mapped.

Patient Experience (PE) – the degree to which an item impacts on, or reflects the experiences of patients.

Data Quality (DQ) – how confident we are in the information (e.g. the source).

4. Qualitative information

As well as quantitative information such as survey results, mortality rates etc, the QRP also makes use of a large body of qualitative information collected locally about providers. This includes information from Learning Disability Partnership Boards (LDPBs), overview and scrutiny committees (OSCs), local involvement networks (LINks), foundation trust boards of governors, NHS Choices, Children's Services Inspections, local Safeguarding Children Boards, Audit Commission PbR coding audits and other third party groups. It also includes information supplied by our inspectors following their engagement activities locally and information from follow-up reports by the regulator.

4.1 Processing qualitative data

To be included in the model, the qualitative information has to be converted into a numeric representation.

Each statement (engagement form, report or commentary) is processed by an analyst. The statement is broken down into one or more parts (known as comments), which may be a sentence, a paragraph, or more.

Where possible, each comment is coded to one or more outcomes, and is graded as being “positive”, “negative”, or “neutral”. Positive comments are those that are positive about the performance of a provider, etc.

Each comment is also assigned a score for the three utility factors described in Section 3 (*CS*, *PE*, and *DQ*), which allows us to give different weights to comments.

A pseudo Z-score *(Z)* is then calculated for each comment using the following equation:

$$Z = \frac{CS \times PE \times DQ}{8} \times \begin{cases} -1 & \text{when the comment is positive} \\ 1 & \text{when the comment is negative} \end{cases} \qquad (10)$$

The weighting factor of 1/8 is based on a comment that is average across all utility factors (i.e. 2 x 2 x 2) receiving a weighting of unity. Z-scores are constrained to be between -3.0 and +3.0.

5. Data item risk bands

Each adjusted item level Z-score (from (9)) is placed in a risk band as follows:

Adjusted Item level Z-score	Risk Band
$Z \leq -2$	Much better than expected
$-2<Z\leq-1.6$	Better than expected
$-1.6<Z\leq-1.2$	Tending towards better than expected
$-1.2<Z<1.2$	Similar to expected
$1.2\leq Z<1.6$	Tending towards worse than expected
$1.6\leq Z<2$	Worse than expected
$Z\geq2$	Much worse than expected

6. Site level data items

Some data items are measured at location or site level rather than provider level. The risk model operates at provider level and so the site level data need to be combined prior to use in the risk model.

7. Aggregation to produce outcome-level risk estimates

Once we have calculated an adjusted Z-score (or pseudo-Z-score in the case of qualitative information) for all items of data, we aggregate these together to produce a final likelihood estimate, based on all the data for a provider for a particular outcome.

We calculate the aggregated estimate (now called ***Z****) as follows:

$$Z^{*} = \frac{\sum_{i=1}^{m} r_i^{-1} x_i Z_i \left[(2CS_i + PE_i)/6\right] + \sum_{i=m+1}^{n} v_i Z_i}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} r_i^{-1} r_j^{-1} \times \left[(2CS_i + PE_i)/6\right] \times \left[(2CS_j + PE_j)/6\right] \times C_{ij} + \sum_{i=m+1}^{n} v_i^2}} \qquad (11)$$

where: $r_i^{-1} = \dfrac{1}{\sum_{i=1}^{m}\sum_{j=1}^{m} \max(C_{ij}, 0)}$ and $v_i = 1$

and:

m = number of quantitative data items available for the provider

n-m = number of qualitative data items available for the provider

r_i^{-1} = reciprocal sum of the correlations of item ***i*** with other items in the outcome (note that the correlations are based on using items across all outcomes)

r_j^{-1} = reciprocal sum of the correlations of item ***j*** with other items in the outcome (note that the correlations are based on using items across all outcomes)

x_i = the number of replicates of the data item (where $x_i = 1$ if no site-level data)

C_{ij} = correlation coefficient between item ***i*** and item ***j***, including 1 when ***i=j***

We divide (2CS + PE) by six because it will produce a weighting of unity when an item scores an average of 2 on both utility factors.

(continued)

Each aggregate Z-score (**Z***) is placed in a risk band, based on its value. Risk bands are as in the table below.

AGGREGATE Z-SCORE (Z*)	RISK BAND
$Z^* < -1.6$	Low Green
$-1.6 \leq Z^* < -1.2$	High Green
$-1.2 \leq Z^* < 0$	Low Yellow
$0 \leq Z^* < 1.2$	High Yellow
$1.2 \leq Z^* < 1.6$	Low Amber
$1.6 \leq Z^* < 2.0$	High Amber
$2.0 \leq Z^* < 2.3$	Low Red
$Z^* \geq 2.3$	High Red