HQIP

Healthcare Quality
Improvement Partnership

Detection and management of outliers for National Clinical Audits

Guidance prepared by National Clinical Audit Advisory Group/HQIP (2011).
Updated by HQIP in consultation with CQC, NHS England, NAGCAE,
NHS Improvement (May 2017)

Title	Detection and Management of Outliers for National Clinical Audits.
Author	Original 2011: Department of Health/Healthcare Quality Improvement Partnership/National Advisory Group on Clinical Audit and Enquiries (NAGCAE) Updated April 2016: Healthcare Quality Improvement Partnership (HQIP)
Publication Date	May 2017
Target Audience	Chief Executives, Medical Directors, Clinical Directors, Commissioners, Provider Organisations, Suppliers of National Clinical Audits, Specialist Societies, Regulatory Bodies.
Circulation List	Chief Executives, Medical Directors, Clinical Directors, Commissioners, Provider Organisations, Suppliers of National Clinical Audits, Specialist Societies, Regulatory Bodies.
Description	This document updates 2011 guidance on the detection and management of "outliers" identified through national clinical audits. This guidance pertains to the National Clinical Audit and Patient Outcomes Programme (NCAPOP) and Non-NCAPOP audits.
Contact Details	HQIP, communications@hqip.org.uk, 020 7997 7370 www.hqip.org.uk

DETECTION AND MANAGEMENT OF OUTLIERS

These recommendations apply to:

- comparisons of providers (general practices, hospitals, trusts, clinical networks, but not individual practitioners[1]) using batches of data collected over an appropriate defined period (and not continuous monitoring); and
- both outcome and process measures of performance (referred to as 'performance indicators').

The statistical analyses involved should be carried out by people with appropriate statistical expertise and experience.

These recommendations are based on original advice provided by an expert group of statisticians (Appendix 1). This document has been updated in 2017 by the Healthcare Quality Improvement Partnership (HQIP), in consultation with the Care Quality Commission (CQC), NHS Improvement, NHS England and the National Advisory Group on Clinical Audit and Enquiries (NAGCAE), to reflect changes in NHS organisational structures, to confirm expectations of the supply of outlier information to the CQC and to clarify responsibilities and actions.

1. Choice of performance indicator

Performance indicators must provide a *valid* measure of a provider's quality of care in that there is a clear relationship between the indicator and quality of care, and relate to frequently occurring events to provide sufficient statistical power.

2. Choice of target (expected performance)

The expected performance may be based either on external sources (research evidence, clinical judgment, audit data from elsewhere) or on internal sources (such as average performance of all providers, though may exclude the provider in question or outliers).

3. Data quality

Three aspects of data quality must be considered and reported:

- Case ascertainment: number of patients included compared to number eligible, derived from external data sources; impact on the generalisability (representativeness) of the results.
- Data completeness: in particular performance indicator data and data on patient characteristics required for case-mix adjustment.

[1] This is covered by the Clinical Outcomes Publication Technical Manual, HQIP, (2016) http://www.hqip.org.uk/resources/clinical-outcomes-publication-technical-manual/

- Data accuracy: tested using consistency and range checks, and if possible external sources.

4. Case-mix (risk) adjustment

Comparison of providers must take account of differences in the mix of patients between providers by adjusting for known, measurable factors that are associated with the performance indicator. These are likely to include age, sex, disease severity and co-morbidity. Other possibilities include socio-economic status and ethnicity.

Adjustment should be carried out using an up-to-date statistical model. The model should have been rigorously tested with regard to its power of discrimination (such as the area under the Receiver Operating Characteristic) and its calibration (such as, goodness-of-fit) and, together with details of the model, both attributes should be publicly reported. Judgment as to the adequacy of a model will depend on the performance indicator selected and the clinical context so universal, absolute values cannot be provided.

5. Detection of a potential outlier

Statistically derived limits around the target (expected) performance should be used to define if a provider is a potential outlier: more than two standard deviations from the target is deemed an ‘alert’; more than three standard deviations is deemed an ‘alarm’.

Note that these are definitions of statistically significant differences from expected performance are differences that may not be clinically significant if based on large numbers of patients.

6. Management of a potential outlier

Management of a potential outlier involves several people:
National Clinical Audit (NCA) supplier: the team responsible for managing and running the audit nationally.
NCA supplier lead: person responsible for the audit, often chair of the Board of Management of the audit.
Provider lead clinician: clinician contact for NCA in provider organisation
Provider medical director and chief executive will need to be involved.

The following table indicates the stages that may be needed in managing a potential outlier, the actions that need to be taken, the people involved and the time scale. It aims to be both feasible for those involved, fair to providers identified as outliers and sufficiently rapid so as not to unduly delay the disclosure of comparative information to the public.

7. Involvement of the regulator

The CQC are included in the guidance so as to provide them with assurance that organisations are engaging appropriately in the process. The CQC recognises that alert level outliers may be identified as a result of chance alone. For alarm level outliers the CQC expect to see evidence of appropriate initial and substantive action plans. The CQC will consider the data as part of its monitoring process. The CQC will not usually take regulatory action if organisations are responding appropriately to each stage of the outlier management process at alert and alarm level.

Stage	What action?	Who?	Within how many working days
1	Providers with a performance indicator 'alert' or 'alarm' require careful scrutiny of the data handling and analyses performed to determine whether there is: 'No case to answer' • potential outlier status not confirmed • data and results revised in NCA records • details formally recorded 'Case to answer' • potential outlier status *Proceed to stage 2*	NCA Supplier	10
2	The Lead Clinician in the provider organisation informed about the potential outlier status and requested to identify any data errors or justifiable explanation/s. All relevant data and analyses should be made available to the Lead Clinician. A copy of the request should be sent to the provider organisation CEO and Medical Director.	NCA Supplier lead	5
3	Lead Clinician to provide written response to NCA supplier.	Provider Lead Clinician	25
4	Review of Lead Clinician's response to determine: 'No case to answer' • It is confirmed that the data originally supplied by the provider contained inaccuracies. Re-analysis of accurate data no longer indicates outlier status. • Data and results should be revised in NCA records. Details of the provider's response and the review result recorded. •Lead Clinician notified in writing copying in provider organisation CEO and Medical Director. 'Case to answer' • It is confirmed that although the data originally supplied by the provider were inaccurate, analysis still indicates outlier status; or • It is confirmed that the originally supplied data were accurate, thus confirming the initial designation of outlier status. *proceed to stage 5*	NCA Supplier	20

5	Contact Lead Clinician by telephone, prior to sending written confirmation of alert or alarm status to CEO copied to Lead Clinician and Medical Director. All relevant data and statistical analyses, including previous response from the Lead Clinician, made available to the Medical Director and CEO. In case of 'alarm' status, NCA supplier to inform CQC[2] and Provider CEO advised to inform commissioners, NHS Improvement[3] and relevant royal colleges. CEO informed that the NCA supplier will be publishing information of comparative performance that will identify providers.	NCA Supplier Lead	5
6	Acknowledgement of receipt of the letter confirming that a local investigation will be undertaken with independent assurance of the validity of this exercise for alarm level outliers, copying in the CQC[4].	Provider CEO	10
7	If no acknowledgement received, a reminder letter should be sent to the CEO, copied to CQC. If not received within 5 working days, CQC[5] and NHS Improvement[6] notified of non-compliance.	NCA Supplier	5
8	Public disclosure of comparative information that identifies providers (e.g. annual report of NCA, data publication online).	NCA Supplier	

[2] Via clinicalaudits@cqc.org.uk

[3] Via nhsi.medicaldirectorate@nhs.net

[4] Via clinicalaudits@cqc.org.uk

[5] Via clinicalaudits@cqc.org.uk

[6] Via nhsi.medicaldirectorate@nhs.net

Appendix 1

Statistical principles for identifying poor performance in national clinical audits

Advice of an expert group prepared for the National Clinical Audit Advisory Group

27 June 2010

I INTRODUCTION

1. The Department of Health has requested the National Clinical Audit Advisory Group (NCAAG) to produce statistical guidance on how potentially outlying performance of healthcare providers can be identified. A growing number of national clinical audits publish quantitative data that allow comparisons of processes and outcomes. These audits will flag up providers that have results which do not seem to be in line with what can be expected compared to other providers or to existing benchmarks.

2. An expert group has produced guidance on the statistical principles that build on a range of statistical approaches that have been used for this purpose over the years.

3. This document complements one on the procedures that should be followed once a provider with potential outlying performance has been identified (i.e. what action should be taken, who should take it, and when) (Appendix 2). Although the identification of outlying performance and the subsequent handling of it are separate activities, there is a mutual interaction. For example, the threshold levels distinguishing acceptable from outlying performance that are being used in the statistical analysis depend on what is going to happen when a provider has been identified as a potential outlier (see paragraph 17). Furthermore, data quality as well as differences in case mix that are not fully adjusted need to be taken into account when potential outlying performance is being investigated. This requires an understanding of the underlying statistical analyses.

4. The document is targeted at stakeholders of national clinical audits (clinicians, providers, commissioners, policy makers, patients and the public) who have a basic understanding of statistical principles. Rather than addressing detailed statistical issues, it sets out fundamental principles that need to be followed and provides advice on how that could be implemented. Methodological information related to more advanced topics is made available through references of key papers.

5. This document is not comprehensive and sets out only one of many possible options. However, it was our intention to be relevant to as many as possible of the scenarios that are encountered within national clinical audits.

6. The identification of outlying performance touches on many fundamental statistical principles. There is an increasing body of research that aims to further develop the available methodology in this area. It is therefore crucial in our view that the analyses are carried out by individuals with appropriate statistical expertise and experience.

II CHOICE OF THE PERFORMANCE INDICATOR

7. In this document, we consider *performance indicators* that are quantitative measures of either processes or outcomes of care. This indicator should be carefully chosen. First, the s*tatistical power* should be considered (i.e. the probability that a provider with truly outlying performance will be detected) which depends on the number of patients (or any other "unit of interest") per provider as well as on the frequency of events if the performance indicator is derived from a discrete outcome or on a measure of variability if it is derived from a continuous outcome. Second, the *validity* of the outcome should be assessed. Important considerations are the extent to which the outcome is *attributable to care provided by the unit* and the *clarity about the relationship between indicator and good and poor quality of care*. A further issue is the *objectivity* of the indicator or the extent to which there is the potential that the indicator can be manipulated by the provider (e.g. "gaming"). Third, the anticipated adequacy of adjustment for potential differences in important risk factors needs to be considered as this will determine the *fairness* of the comparison.

8. Performance indicators can be derived from *different types of data*. Roughly speaking, one can distinguish indicators based on discrete outcomes expressed as proportions or counts or based on summary measures of continuous measures expressed as means or medians as well as on various derived measures. The statistical approach to determine the limits of the acceptable range will need to take the type of data into account, especially when numbers of cases or events are low. In that case, appropriate methods for small samples should be used.

III DESIGN

9. Two different types of design can be distinguished. The first is a comparison of providers based on data collected over a *given period of time* or including a *given number of patients*. The second is based on *sequential monitoring techniques* (e.g. CUSUM methods) that allow an update of the assessment of performance of a provider after each case accrued. In this document, we only consider the analysis of data collected over a given period of time or a given number of patients as those are most commonly used in ongoing national clinical audits.

10. There is an obvious trade off between the statistical power of the analysis and the timeliness of detecting outlying performance. If the reporting time

period is short, the number of included patients may be small and truly outlying providers may not be detected because of a lack of statistical power. On the other hand, if the reporting time period is relatively long outlying providers may remain undetected for a considerable amount of time.

IV DEFINITION OF AN OUTLIER

11. A provider will be identified as an outlier if the value of the performance indicator is outside the range of *acceptable performance*. This range will be determined by a *target* and a *range of values* around that target that are defined on the basis of statistical principles.

IVa Choosing the target

12. The *target* can be set on the basis of external criteria (e.g. historical data, data collected elsewhere, a clinical practice guideline, or clinical judgement) or on the basis of internal criteria derived from the audit data under consideration. An example of the latter is the use of the average over all cases among all providers included in an audit.

13. If an internal target is used, one could consider using a *cross-validation approach* and compare each provider against a target derived from all other providers. An advantage of using cross-validation is that it removes the influence of a provider on the target against which it is being compared. If the number of providers is relatively large, cross-validation has only a marginal impact on the target that is used for each provider. Conversely, if cross-validation is used when only a small number of providers are being compared, the influence on the target could be substantial.

14. Alternatively, one may want to avoid the influence of outlying providers on an internal target by *resetting the top and bottom values of a performance indicator to a specified percentile* (e.g. 5%), a process sometimes referred to as “winsorising”. A further option for defining an internal target is to take the *average observed in providers that were recognised centres of excellence* on the basis of pre-defined criteria.

IVb Limits of acceptable performance

15. The definition of the *limits of acceptable performance* should be based on statistical criteria. Statistical process control processes developed in an industrial context typically define a range of values that are within three standard deviations from the target value as “in control” (i.e. acceptable). This would correspond to statistically testing whether a performance indicator is different from the target at a two-sided significance level of 0.002. In practice, this would imply that 99.8% of all providers are expected to be within the acceptable range, if all providers are in control (i.e. performing according to the target). *Alternative significance levels* can be used. For example, a two-sided significance level of 0.05 would define

all values within two standard deviations from the target as acceptable. If all providers are in control, 95% would be within these limits.

16. We recommend that as a starting point the two-sided significance levels of 0.05 and 0.002 should be used to define limits of acceptable performance. These limits could be considered as the thresholds for an "*alert*" or an "*alarm*", respectively. As a consequence, the use of other significance levels will require an explicit justification (see paragraph 17).

17. The final choice of the actual significance levels needs to take into account the *relative weight of the two potential errors*: erroneously identifying a provider as an outlier (a false-positive result or Type I error); and erroneously considering a provider's performance as acceptable (a false-negative result or Type II error). Further relevant determinants of the limits are data quality, the adequacy of the risk adjustment, and the issue of multiple testing (see paragraph 31).

18. It is important to note that the limits of acceptable performance defined in this way depend on the number of cases per provider. Especially when the number of cases is large, differences that are *statistically significant* may not always be *clinically significant*. For this reason, it has been suggested to use separate thresholds: one to demonstrate evidence for safety and one to demonstrate evidence for danger. [1]

V ASSESSMENT OF DATA QUALITY

19. A report comparing performance indicators among providers should explicitly describe data quality. A number of measures of data quality should be made available. First, *case ascertainment* should be given as a proportion of included cases out of all eligible cases. The number of eligible cases should be derived from external sources. Second, *completeness* of critical data fields should be presented as a proportion of non-missing values of fields containing information on process or outcome measure under consideration as well as on case mix factors that are likely to be included in the risk adjustment models. Third, *accuracy* of critical data fields should be investigated. This can be done "internally" through consistency and plausible range checks within the available data sets and – if feasible – "externally" through comparison with another data source.

20. The *generalisability* (i.e. representativeness) should be assessed by comparing characteristics of the included cases against those that are not included or against all eligible cases. The potential of using existing data – if possible linked at patient level – should be explored for this purpose.

[1] Demonstrating safety through in-hospital mortality analysis following elective abdominal aortic aneurysm repair in England. Br J Surg 2008;95:64-71.

VI RISK ADJUSTMENT

21. The process of identifying outliers should always include adjustments for potential variations in risk due to case mix. The development of the risk adjustment approach depends on what *outcome* (or process) is being considered, the *time frame*, and the *population*. As a result, the risk adjustment approach should always be "tailor-made" and match the specific requirements of the comparison that is being carried out.

22. Most risk adjustment methods rely on *stratification* or *statistical modelling*. Stratification implies that the comparison is being carried out within strata that are homogeneous according to pre-specified risk factors. A risk adjusted result is then produced by pooling the estimate from the two or more strata into a single pooled estimate. The advantage of stratification is that it is relatively straightforward to implement and comprehend, but it has two important drawbacks. First, there is a potential of information loss as continuous variables have to be categorised. Second, there is the problem of low numbers within strata especially when multiple risk factors are being considered. For these two reasons, we recommend statistical modelling.

23. A statistical risk adjustment model should aim to include all patient and disease characteristics that are available *before the start of the care process* and that are accepted as potential risk factors for the outcome under consideration. Important factors that should be considered for inclusion are age, sex, disease severity, and co-morbidity. Depending on the specific clinical context other candidate risk factors are socio-economic deprivation and ethnicity, but it is important to realise that adjustment for these risk factors may mask established inequalities.

24. The risk adjustment model can be either based on *existing statistical models* that are generally accepted as appropriate for the purpose of risk adjustment or *newly developed* within the data under consideration. Irrespective of whether an existing or newly developed risk adjustment model is being used, its performance should be examined. Parameters of the model's *goodness-of-fit* and *discrimination* or *explanatory power should* be presented and its appropriateness should be discussed. It is not possible to set minimum criteria for the risk model's performance as these will depend on the type of indicator that is being used as well as the specific clinical context.

25. We recommend that as a first step the risk adjusted performance estimates are based on "*indirect standardisation*". This implies in its simplest form that for discrete outcomes the observed number of events for a provider is divided by the number expected on the basis of the statistical model. A ratio of one would indicate that the outcomes are as expected. A risk adjusted performance estimate on the same scale as the original indicator can be calculated by multiplying this ratio by the average over all providers. For continuous outcomes the differences between the observed result and that expected (i.e. "residual") is calculated for a

provider. In this case, a difference of zero would indicate that the outcomes are as expected. A risk adjusted estimate that can be directly compared with the unadjusted results is calculated by adding this difference to the average over all included providers.

26. An important argument to use indirect standardisation is that it allows an explicit comparison of the unadjusted and the adjusted results provided that the risk adjustment model is well *calibrated* (i.e. the observed and expected results are equal when averaged over all cases). A direct comparison of unadjusted and adjusted results is helpful as it provides an opportunity to evaluate the direction and size of the impact of risk adjustment. In addition, the unadjusted and adjusted results, if derived from a well calibrated model, can be compared against the same limits of acceptable performance (see paragraph 16).

27. Risk adjustment will always be *incomplete* as it will never be possible to fully measure all relevant case mix factors or represent them adequately in a risk adjustment model. It is therefore important to accept that there will always be "*residual confounding"*. In other words, it should always be highlighted that even risk adjustment differences in case mix can never be excluded as possible explanations for outlying performance.

VII PRESENTATION OF PROVIDER COMPARISONS

28. A *"funnel plot"*, a form of control chart, provides an attractive graphical format for the presentation of the performance indicators. In a funnel plot, the result for each provider is presented as a function of its precision. The target as well as the limits of acceptable performance can be pre-specified as they do not depend on the actual results. The precision parameter corresponds to the number of cases (or an equivalent measure of volume) for each provider.[2]

29. Potential outlying providers can be detected as those with results outside the funnel limits (see paragraph 11). An important advantage of funnel plots is that they clearly demonstrate how the limits of acceptable performance depend on the number of cases. It is therefore easy to appreciate the difference between a provider's performance and the target and its position with respect to the acceptable range.

30. We recommend that for each provider measures of case ascertainment, data completeness (separate for performance indicator and for case mix variables), and data accuracy are also available to inform the interpretation of these plots.

VIII Further issues

Multiple comparisons

[2] Funnel plots for comparing institutional performance. Stat Med 2005;24:1185-202.

31. It has been argued that the limits of acceptable performance should be adjusted to take into account that many providers are being compared simultaneously. If we use *limits of acceptable performance* based two-sided significance levels of 0.05 and 0.002 as explained earlier (see paragraph 16), we may identify providers with results outside these limits simply by chance alone (i.e. a "false-positive result").

32. One can adjust the limits of acceptable performance for multiple comparisons by using lower significance levels. One classic approach would be to divide the proposed significance levels of 0.05 and 0.002 by the number of providers (i.e. the *Bonferroni correction*). This ensures that the probability of one or more false-positive results among all included providers is not greater than 0.05 or 0.002, respectively. However, this criterion is too strict as it will strongly reduce the statistical power to detect true outliers (see paragraph 13). Alternatively, an approach based on the "*false discovery rate*" can be used. The false discovery rate can be considered as the probability that a provider is not an outlier if its p-value is found to be lower than the defined significance level. The false discovery rate will produce more false-positive results than the Bonferroni correction but less than if no adjustment of multiple comparisons is being applied. [3]

33. A fundamental issue in this context is *whether an adjustment for multiple comparisons is justified* in the first place. One could consider that each provider is to be compared to the target on an individual basis and that the consequences of this comparison only relate to this provider itself. If that is accepted, no adjustments for multiple comparisons should be made.

Overdispersion

34. Many comparisons of healthcare providers have demonstrated a greater level of variability among providers than can be explained by chance and the existence of a few outlying units. This phenomenon is often referred to as *overdispersion*. Important explanations for overdispersion are the limitations of the available risk adjustment methods and the variable data quality.[4]

35. A number of options for dealing with overdispersion have been proposed that could be incorporated when the limits for acceptable performance are set: for example by improving the risk adjustment through the analysis within groups of providers that are expected to be more homogeneous, or the use of an interval as a target. It is also possible to estimate the level of overdispersion and to adjust the limits by estimating an *overdispersion factor* and inflating the limits of acceptable performance around the target by this factor (i.e. *multiplicative* adjustment). Alternatively, one can estimate the between-provider variance (the "random effects") and add this to the variance expected if there were no differences between the

[3] Use of the false discovery rate when comparing multiple health care providers. J Clin Epidemiol 2008;61:232-40.

[4] Handling over-dispersion of performance indicators. Qual Saf Health Care 2005;14:347-51.

providers. It should be emphasised that these adjustments for overdispersion reflect the limitations of the risk adjustment and data quality. Attempts should continue to explain the excess variability, to improve the risk adjustment, and to improve data quality.

Multilevel or random-effects models

36. An important advantage of *multilevel models* is that they explicitly incorporate the overdispersion (i.e. between-provider variance). Furthermore, these models provide a flexible framework for incorporating determinants of outcome at different levels of the hierarchal structure. This implies that provider characteristics as well as characteristics of groups of providers can be simultaneously included in the model and their impact on outcome investigated.[5] A further extension of these models is their use from a Bayesian viewpoint.[6]

37. Multilevel models also allow the estimation of risk adjusted estimates which are "*shrunken*" towards the overall mean. The level of *shrinkage* is stronger for the providers with fewer cases. These shrunken estimates compensate for the *regression-to-the-mean* effect which is especially relevant if the number of cases per provider varies substantially. It has been recognised that multilevel models are more conservative than conventional approaches based on fixed-effects models. As a result the chance of a false positive result is lower but chance of false-negative higher.[7]

Imputation of missing values

38. Data sets from national clinical audits will inevitably contain a number of cases for whom not all data are available. Imputation techniques could be used to deal with missing values for the case mix factors that are candidates to be included the risk adjustment model. Multiple imputation of missing data has the potential to increase the statistical power (cases with missing values can be retained in the analysis) and to reduce bias (mechanism of missingness may depend on case mix).[8]

Members of Expert Group

Dr Jan van der Meulen, Professor of Clinical Epidemiology, Department of Health Services Research & Policy, London School of Hygiene and Tropical Medicine (Chair)

Dr David Cromwell, Royal College of Surgeons of England; (previously Senior Lecturer in Health Services Research, Department of Health Services Research & Policy, London School of Hygiene and Tropical Medicine)

Dr David Harrison, Senior Statistician, Intensive Care National Audit & Research Centre

Dr Chris Rogers, Senior Research Fellow, Clinical Trials and Evaluation Unit, Bristol Heart Institute, University of Bristol

[5] Statistical and clinical aspects of hospital outcomes profiling. Statist Science 2007;2:206-

226.

[6] Identifying outliers in Bayesian hierarchical models: a simulation-based approach. Bayesian Analysis 2007;2:409-44.

[7] The use of fixed- and random-effects models for classifying hospitals as mortality outliers: a Monte Carlo assessment. Med Decis Making 2003;23:526-39.

[8] Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls BMJ 2009;*338*:b2393