Analysing indicators of performance, satisfaction, or safety using empirical logit transformation

Sarah Stevens,[1,2] Jose M Valderas,[3] Tim Doran,[4] Rafael Perera,[1,2] Evangelos Kontopantelis[2,5]

[1]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK
[2]National Institute for Health Research School for Primary Care Research, Oxford, UK
[3]APEx Collaboration for Academic Primary Care, Institute for Health Services Research, University of Exeter Medical School, University of Exeter, Exeter, UK
[4]Department of Health Sciences, University of York, UK
[5]Centre for Health Informatics, Institute of Population Health, University of Manchester, Manchester M13 9GB, UK

Correspondence to: E Kontopantelis e.kontopantelis@manchester.ac.uk

Cite this as: *BMJ* 2016;352:i1114
http://dx.doi.org/10.1136/bmj.i1114

Accepted: 20 January 2016

Performance, satisfaction, and safety indicators are commonly measured on a percentage scale. Such indicators are often subject to ceiling or floor effects and performance may be inherently non-linear. For example, improving from 85% to 95% might be more difficult and need more effort than improving from 55% to 65%. As such, analysis of these indicators is not always straightforward and standard linear analysis could be problematic. We present the most common approach to dealing with this problem: a logit transformation of the score, following which standard linear analysis can be conducted on the transformed score. We also demonstrate how estimates can be back-transformed to percentages for easier communication of findings. In this paper, we discuss the benefits of this method, use algebra to describe the relevant steps in the transformation process, provide guidance on interpretation, and provide a tool for analysis.

SUMMARY BOX

- Performance, satisfaction, or safety indicators in healthcare are commonly measured on a percentage scale
- Standard linear analysis could be problematic owing to ceiling or floor effects or non-linearity
- A logit transformation of the score is the most common solution
- Estimates can be back-transformed to percentages for a more intuitive interpretation

In recent years, efforts to improve the quality and safety of healthcare have resulted in the introduction of systems for monitoring the performance of healthcare providers and the satisfaction and safety of patients. New quality and performance indicators have been created, to which financial and reputational rewards for providers are often attached. Although performance indicators are measured for each patient, they are often only reported in aggregate form (eg, at the practice or hospital level). Therefore, an indicator that begins as a binary outcome (that is, the target is either met or not met for each patient),[1] becomes a proportion (that is, the percentage of patients for whom the quality target is met). Such summary indicators are usually analysed by linear models. This is appropriate in many scenarios where the scores retain linear properties, for example, in the analysis of referral rates and their predictors.[2]

However, for aggregate analyses of performance indicators, two particular problems can emerge. Firstly, it is common for individual indicators within a set to vary in intrinsic difficulty (eg, recording blood pressure is easier than controlling blood pressure) or vary in the size of associated incentives. This frequently results in healthcare providers achieving targets for 100% of patients for easier indicators[3] and, less often, for 0% of patients for more difficult indicators, or for indicators with smaller incentives. Maximum (100%) and minimum (0%) scores are more common when patient groups are small (the problem of small denominators). These "ceiling" and "floor" effects can cause problems in analyses of data at the patient level, but also make the use of aggregate performance scores in linear models problematic. This is a particular problem for prediction modelling (eg, in interrupted time series designs)[4] where predictions might fall outside the 0-100% range.

Secondly, there is inherent non-linearity in performance indicators, because the effort required by a health worker is not uniform across patients. For example, some patients might attend clinic appointments infrequently while others might be persistent in refusing a measurement or treatment. Similarly, satisfaction is subjective and different levels of effort are needed to satisfy different patients, whereas in terms of safety, risk management is inexact and some patients might be more difficult to manage clinically. Therefore, it is generally more difficult to achieve an improvement from 85% to 95%, than from 55% to 65%. Analogously, an improvement from 0% to 10% should pose very little difficulty. Box 1 presents some examples of performance, satisfaction, and safety[5] indicators.

One potential solution to these issues is for researchers to dichotomise the indicator by classifying healthcare providers simply as high or low achievers (in terms of performance, satisfaction, or safety), based on a specified threshold of achievement. For example, assume that a healthcare provider has met a target (=1) if the relevant performance score is over 85%, and not met the target (=0) otherwise. Analyses are then possible by use of logistic models, and odds ratios would be used to quantify effects. However, odds ratios are intuitively difficult to conceptualise and are frequently interpreted as

Box 1: Examples of ceiling or floor effects

Quality and Outcomes Framework performance indicator DM22 (2006-07 to 2012-13)

Measured the percentage of patients with diabetes who have a record of estimated glomerular filtration rate or serum creatinine testing in the previous 15 months. In 2006-07 and with 8365 general practices reporting the indicator, one practice had a score of 0% and 700 a score of 100%. The mean score was 96.4%.

GP patient satisfaction survey, from 8307 general practices in 2008

Survey on how easy it was for patients to get through on the phone at own doctor's surgery (no *v* yes): 81 practices scored 100%, when the mean score was 87.4%. Survey also asked about the ability for patients to get appointment within two days (no *v* yes): 93 practices scored 100%, when the mean score was 85.7%

Investigation of prescribing safety using the Clinical Practice Research Datalink, in 2013[5]

Investigation looked at the proportion of women with a breast cancer diagnosis who were prescribed oral or transdermal oestrogens: 319 (61%) of 523 practices had a prevalence of 0%, when the mean prevalence was 1.1%. For patients prescribed repeated amiodarone without a thyroid function test within the recommended time period, 43 (9%) of 505 practices had a prevalence of 100%, when the mean prevalence was 42%.

relative risks. Although such an approach could be acceptable in a scenario where few providers are low scoring (where the rare event approximation stands), more generally odd ratios overestimate the relative risk and their interpretation tends to be flawed.[6] In addition, such a simplification discards a considerable amount of information and reduces statistical power, while the choice of the threshold value used to dichotomise performance may be arbitrary.

A more suitable approach is based on a transformation that uses the logit function to linearise a performance, satisfaction, or safety indicator. This transformation effectively takes a scale that ranges from 0 to 1 (or 0% to 100%), and expands the scale so that it ranges from minus (–) infinity to plus (+) infinity. The indicator can then be analysed by standard linear models or similar methods, in a frequentist or Bayesian framework. Given the particular challenges associated with this approach, in this paper we provide guidance for researchers in conducting and interpreting analyses of logit transformed performance scores of quality indicators.

For practical examples, we draw on our experience of analysing family practice performance under the United Kingdom's Quality and Outcomes Framework (QOF).[7-9] The framework is a financial incentive scheme introduced by the UK government in 2004, which rewards practices on the basis of their performance on more than 100 quality indicators related to the clinical management of chronic disease, practice organisation and, patient experience.[10] For the clinical indicators, which are regularly reviewed and could be withdrawn,[11] practices are assessed on the basis of the percentage of eligible patients for whom each target is met (eg, the percentage of patients with coronary heart disease who give a blood pressure recording of ≤150/90 mm Hg). The main research questions in relation to performance on the QOF indicators relate to how practice performance varies between practices with different characteristics and patient profiles, and how performance changes over time.

We have focused on performance indicators in primary care to provide practical examples and place the methods into context. However, the methods are relevant for the analysis of any percentage score that aggregates binary or continuous information from a lower level unit (eg, patient) to a higher level unit (eg, general practice population) and where non-linearity is present. Examples of satisfaction and safety include the percentage of people who are happy with access to their preferred general practitioner or who are prescribed a drug that puts them at risk.

Approach

Logit transformation

The first step is to assemble practice scores (in the case of the QOF, the proportion of eligible patients for whom a given target has been achieved) and model these scores on the logit (log-odds) scale. There are two main options to do this and to achieve an expansion in the scale from [0, 1] to [± infinity]: simple and empirical transformation.

In the simple logit transformation, the score p ($0 \leq P \leq 1$) is transformed into a log odds: $\text{logit}(p)=\ln(p/(1-p))$ (fig 1). For example, a difference in the untransformed score (p) from 0.97 to 0.98 (that is, 97% to 98%) represents the same difference in transformed score ($\text{logit}(p)=0.41$) as a difference from 0.55 to 0.65 (that is, 55% to 65%). The transformed score can then be modelled with standard linear models and the analysis of transformed scores also ensures that predicted achievement scores lie between 0% and 100%. The main drawback of the simple logit transformation is that achievement scores of 0% and 100% become minus and plus infinity, respectively, following transformation. As a consequence, these observations will be interpreted as missing values by statistical software packages and removed from analyses. If there are a large number of scores at the ceiling or floor values, this effectively renders simple logit transformation ineffective.

The empirical logit transformation offers an improvement over the simple logit transformation at the ceiling and floor points by making a separate transformation at these values.[12 13] For scores where p is strictly greater than 0 and less than 1, the simple logit transformation is applied as above. For scores where p is equal to 0 or 1, the empirical logit transformation is given by formula below, where n is the number of observations over which p is calculated:[14 15]

$$\text{Logit}(p) = \ln\left[\left(p+\frac{0.5}{n}\right)/\left(1-p+\frac{0.5}{n}\right)\right]$$

Details of equation

In the QOF setting, n would be the number of patients for which an indicator is evaluated (the denominator for the indicator), for example, the number of patients diagnosed with diabetes. Not only does this transformation overcome the problems described above, but it also has an additional benefit; scoring 100% on an indicator evaluated on a large number of people (n) is rewarded with a higher transformed score

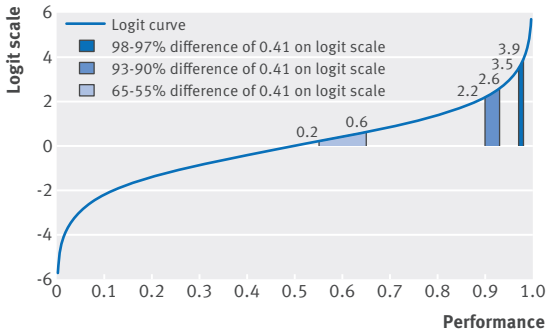

Fig 1 | Simple logit transformation of a performance proportion (*p*)

compared with scoring 100% based on a smaller *n*. The effect is to further expand the scale of quality scores, allowing for even greater discrimination between practices (fig 2). For example, a 100% score on five patients for a practice would be transformed to a score of 2.40 on the logit scale, whereas the same score on 10 patients for another practice would correspond to a score of 3.04 on the logit scale.

Although the denominator adjustment made by empirical logit transformation (where p=0 or 1) could also be useful for p values strictly greater than 0 and less than 1, the justification is less clear and interpretation is problematic (eg, a score of p=0.8 (80%) and a denominator of n=35 would be equivalent to a score of p=0.85 (85%) and a denominator of n=5 on the empirical logit scale). Use of empirical logit transformation across the range of values in [0, 1] may be attractive, where a score of 100% on one patient would be transformed to 1.1; under simple logit transformation, a score of 99% on 99 patients would be transformed to much higher score of 4.6. Therefore, it might be reasonable to use empirical logit transformation across all scores for consistency, which would effectively act as an adjustment for the different effort needed to meet the same score across varying denominators. However, an alternative and perhaps simpler approach would be to apply both methods of logit transformation as described and then use the denominator as an additional predictor in a multiple regression model to control for effort in the analysis stage.

Two other aspects should be considered. Firstly, on the logit scale, a unit with a very high score will lose

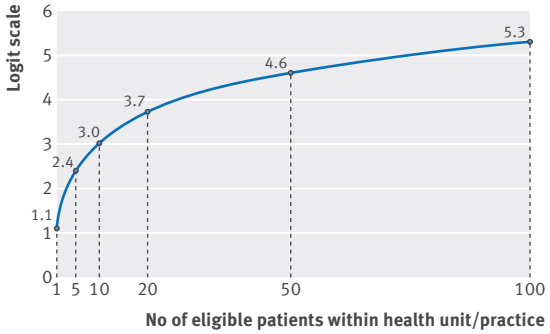

Fig 2 | Empirical logit transformation when performance score (*p*) is 100%, over various denominator sizes (*n*)

much more from the next failure than it will gain from a success (with the picture reversed for a unit with a very low score). Secondly, empirical logit transformation assumes that all data are available or that the reported denominators across the higher level units are comparable to the true denominators. In other words, the units should report a representative sample of their data, and the proportion reported needs to be similar across units. If that is not the case, units that report a smaller proportion are penalised under the empirical logit.

Back-transformation

The interpretation of regression coefficients on the logit scale is intuitively difficult, and hence a back-transformation of the effects to proportions (percentages) is desirable. However, the non-linear nature of the transformed scores and the resulting effects complicates the back-transformation to a linear proportion scale. As shown in figure 1, a fixed size effect on the logit scale corresponds to a smaller effect on the untransformed scale at the extremes, and hence the back-transformed effect depends on the underlying achievement score. Therefore, an anchor achievement score must be chosen on which the back-transformation is to be based. Figure 3 explains this principle formulaically.

To demonstrate this effect in practice, consider recent research on the quality of diabetes care (measured as a percentage achievement score) and the prevalence of disease at the practice level.[9] A 1% higher prevalence of diabetes at the practice level was found in regression analyses to be associated with a 0.031 lower achievement score on the logit scale. As shown in table 1, the effect of differences in prevalence on the untransformed achievement score differs, depending on the anchor achievement rate selected. Assuming that we want to quantify the effect of a 1% increase in the prevalence rate of diabetes on the back-transformed scale, we observe a larger effect for practices whose underlying achievement score is 0.5 (50%). The same difference in prevalence has a much smaller effect on achievement for practices with a median achievement score of 0.9245 (92.45%) or with very low achievement scores.

Choice of anchor score

As demonstrated in table 1, while the choice of a specific anchor score over another has no bearing on the statistical significance of results, it does affect the relative clinical or practical significance of the factor of interest. The anchor value should not be arbitrary, but rather it should be based on a plausible value for the performance, satisfaction, or safety score, for example, it can be based on its mean or median. Use of the median or mean achievement score is intuitively sensible if researchers want to describe the relation between achievement and other factors in the average or typical case. However, if researchers are examining these factors with a view to developing interventions for improvement, then an anchor score reflecting poor performance is a sensible choice, assuming any intervention is aimed

Assuming that prevalence (x%) is a predictor of achievement, the achievement score (p) for a practice on the logit scale can be expressed though a simple regression as:
$\text{Logit}(p) = a + \beta x$, where a is the intercept and β is the regression coefficient quantifying the association between x and p.
Achievement on the logit scale for the same practice, if prevalence increases by 1%, becomes:
$\text{Logit}(p_{new}) = a + \beta(x + 1)$
The difference gives the change in achievement (on the logit scale) for a 1% increase in prevalence (beta values in a linear regression model): $\text{Logit}(p_{new}) - \text{Logit}(p) = \beta$
Achievement on the logit scale is related to a particular value of percentage achievement (as defined by the logit transformation):

$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right) \quad (1)$$

Similarly, $$\text{Logit}(p_{new}) = \ln\left(\frac{p_{new}}{1-p_{new}}\right)$$

To work backwards to get a change in percentage achievement, we assume that percentage achievement at prevalence (x+1)% is equal to the percentage achievement at prevalence x% plus a constant c, the change in percentage achievement per 1% increase in prevalence.

$$\left.\begin{array}{l} p_{new} = p + c \\ \text{Logit}(p_{new}) - \text{Logit}(p) = \beta \end{array}\right\}^{(1)} \Rightarrow \ln\left(\frac{p+c}{1-(p+c)}\right) - \ln\left(\frac{p}{1-p}\right) = \beta$$

Solving for c, we obtain:

$$c = \left[exp\,(-\beta)\left(\frac{1-p}{p}\right) + 1\right]^{-1} - p$$

To obtain c, we need to assume an anchor value for p, and the average achievement score across clinical units or practices is as good an assumption as any.

Fig 3 | **Back-transformation explained**

at improving performance, satisfaction, or safety in a low achieving setting only.

We would therefore argue that choice of an anchor score is largely at the discretion of researchers, who should use the research aims to inform their choice. However, the mean or median scores should be suitable in most scenarios and a priori justification would be needed for alternative anchor choices. Another approach could be to present back-transformed results obtained using several different anchor scores to stimulate discussion around this issue, although attention needs to be given to the interpretation of the group of results. It should also be noted that transformed scores, like percentage scores, do not account for the difficulty in meeting a specific indicator and that investigators should be careful with comparisons across indicators of varying difficulty levels. In these cases, the anchor score can be chosen to reflect the inherent difficulty for an indicator, although the relation between the anchor score and difficulty is not intuitive.

To aid researchers with use of these methods, we have made available an Excel workbook with the transformation and back-transformation formulas, given different anchor scores (available from the corresponding author on request or from his personal website (www.statanalysis.co.uk/files/logit_transformation.xlsx)).

Table 1 | **Quantification of back-transformed effects of a 1% increase in diabetes prevalence, at various anchors**

Increase in prevalence (%)	Effect of predictor (prevalence) on logit scale (95% CI)	Anchor achievement score (p)	Back-transformed effect of predictor on achievement score (absolute difference c) (95% CI)
1	−0.031 (−0.041 to −0.021)	0.9500	−0.0015 (−0.0020 to −0.0010)
		0.9245 (median)	−0.0022 (−0.0029 to −0.0015)
		0.7500	−0.0059 (−0.0078 to −0.0040)
		0.5000	−0.0077 (−0.0102 to −0.0052)
		0.2500	−0.0058 (−0.0076 to −0.0039)
		0.0500	−0.0015 (−0.0019 to −0.0010)

Discussion

We have demonstrated the use of empirical logit transformation for the analysis of performance, satisfaction, or safety indicators that are subject to ceiling or floor effects. We have argued the benefits of this method, algebraically described the processes, provided guidance on interpretation, and have made available a simple tool to aid researchers in using the method. These methods have broad applicability in health services research, but can also be applied in other settings, for example, citizen satisfaction with urban services[16] or hotel websites.[17]

We thank Jill Stokes, researcher at the University of Manchester, who provided the prescription safety examples; and Ben Amies, a GP registrar, who provided feedback on the paper.

Contributorship: SS wrote the manuscript, with help from EK. JMV, TD, and RP critically edited the manuscript. SS is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. In this article, we present examples on applications of these methods to a range of research questions and studies as published in major clinical journals, including *The BMJ*. SS is an early career statistician who has recently been involved with analysing performance measures in a primary care setting. TD is a clinical researcher with interests in quality of care and experience in applying these methods. JMV is a clinician with expertise on the measurement of quality of care and in psychometric methods. RP is a medical statistician whose research program focuses, not exclusively, on monitoring in primary care. EK is a biostatistician and health services researcher who has used the reported methods to answer research questions pertaining to incentivisation in primary care.

Funding: UK Medical Research Council Health eResearch Centre grant MR/K006665/1 supported the time and facilities of EK.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Kontopantelis E, Reeves D, Valderas JM, Campbell S, Doran T. Recorded quality of primary care for patients with diabetes in England before and after the introduction of a financial incentive scheme: a longitudinal observational study. *BMJ Qual Saf* 2013;22:53-64. doi:10.1136/bmjqs-2012-001033.
2 Hippisley-Cox J, Hardy C, Pringle M, Fielding K, Carlisle R, Chilvers C. The effect of deprivation on variations in general practitioners' referral rates: a cross sectional study of computerised data on new medical and surgical outpatient referrals in Nottinghamshire. *BMJ* 1997;314:1458-61. doi:10.1136/bmj.314.7092.1458.
3 Doran T, Fullwood C, Gravelle H, et al. Pay-for-performance programs in family practices in the United Kingdom. *N Engl J Med* 2006;355:375-84. doi:10.1056/NEJMsa055505.
4 Kontopantelis E, Doran T, Springate DA, Buchan I, Reeves D. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. *BMJ* 2015;350:h2750. doi:10.1136/bmj.h2750.
5 Stocks SJ, Kontopantelis E, Akbarov A, Rodgers S, Avery AJ, Ashcroft DM. Examining variations in prescribing safety in UK general practice: cross sectional study using the Clinical Practice Research Datalink. *BMJ* 2015;351:h5501. doi:10.1136/bmj.h5501.
6 Davies HTO, Crombie IK, Tavakoli M. When can odds ratios mislead?*BMJ* 1998;316:989-91. doi:10.1136/bmj.316.7136.989.
7 Campbell SM, Reeves D, Kontopantelis E, Sibbald B, Roland M. Effects of pay for performance on the quality of primary care in England. *N Engl J Med* 2009;361:368-78. doi:10.1056/NEJMsa0807651.

8 Doran T, Kontopantelis E, Valderas JM, et al. Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework [correction in: *BMJ* 2013;347:f5939]. *BMJ* 2011;342:d3590. doi:10.1136/bmj.d3590.
9 Ricci-Cabello I, Stevens S, Kontopantelis E, et al. Impact of the prevalence of concordant and discordant conditions on the quality of diabetes care in family practices in England. *Ann Fam Med* 2015;13:514-22. doi:10.1370/afm.1848.
10 Roland M. Linking physicians' pay to the quality of care--a major experiment in the United kingdom. *N Engl J Med* 2004;351:1448-54. doi:10.1056/NEJMhpr041294.
11 Reeves D, Doran T, Valderas JM, et al. How to identify when a performance indicator has run its course. *BMJ* 2010;340:c1717. doi:10.1136/bmj.c1717.
12 Cox DR, Snell EJ. *Analysis of binary data*.2nd ed. Chapman and Hall, 1989.
13 Abraham B, Unnikrishnan Nair N, eds. *Quality improvement through statistical methods*.Birkhäuser, 1998doi:10.1007/978-1-4612-1776-3.
14 Berkson J. Maximum Likelihood and Minimum X2 Estimates of the Logistic Function. *J Am Stat Assoc* 1955;50:130-62.
15 Anscombe FJ. On Estimating Binomial Response Relations. *Biometrika* 1956;43:461-4doi:10.1093/biomet/43.3-4.461.
16 Stipak B. Citizen Satisfaction with Urban Services - Potential Misuse as a Performance Indicator. *Public Adm Rev* 1979;39:46-52doi:10.2307/3110378.
17 Chung T, Law R. Developing a performance indicator for hotel websites. *Int J Hospit Manag* 2003;22:119-25doi:10.1016/S0278-4319(02)00076-2.