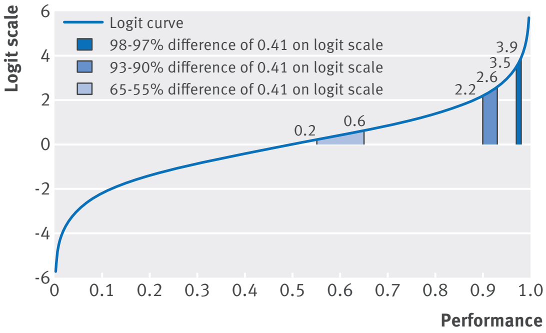

Fig 1 | Simple logit transformation of a performance proportion (*p*)

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 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.

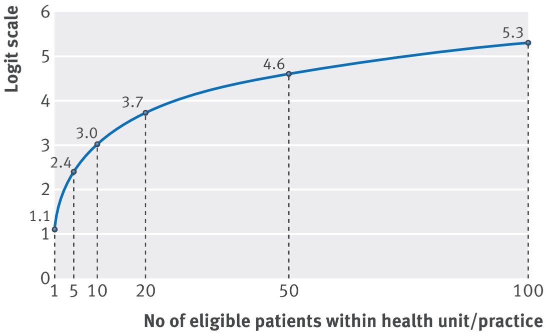

Fig 2 | Empirical logit transformation when performance score (*p*) is 100%, over various denominator sizes (*n*)

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