BJR

Received: 26 January 2016 | Revised: 2 March 2016 | Accepted: 10 March 2016

http://dx.doi.org/10.1259/bjr.20160094

Cite this article as:
Keeble C, Baxter PD, Gislason-Lee AJ, Treadgold LA, Davies AG. Methods for the analysis of ordinal response data in medical image quality assessment. *Br J Radiol* 2016; **89**: 20160094.

REVIEW ARTICLE

Methods for the analysis of ordinal response data in medical image quality assessment

[1,2]CLAIRE KEEBLE, MSc, GradStat, [1]PAUL D BAXTER, BSc, PhD, [2]AMBER J GISLASON-LEE, BSc, MSc, [2]LAURA A TREADGOLD, BSc, PhD and [2]ANDREW G DAVIES, BSc, MSc

[1]Division of Epidemiology and Biostatistics, University of Leeds, Leeds, UK
[2]Division of Biomedical Imaging, University of Leeds, Leeds, UK

Address correspondence to: Miss Claire Keeble
E-mail: *c.m.keeble@leeds.ac.uk*

ABSTRACT

The assessment of image quality in medical imaging often requires observers to rate images for some metric or detectability task. These subjective results are used in optimization, radiation dose reduction or system comparison studies and may be compared to objective measures from a computer vision algorithm performing the same task. One popular scoring approach is to use a Likert scale, then assign consecutive numbers to the categories. The mean of these response values is then taken and used for comparison with the objective or second subjective response. Agreement is often assessed using correlation coefficients. We highlight a number of weaknesses in this common approach, including inappropriate analyses of ordinal data and the inability to properly account for correlations caused by repeated images or observers. We suggest alternative data collection and analysis techniques such as amendments to the scale and multilevel proportional odds models. We detail the suitability of each approach depending upon the data structure and demonstrate each method using a medical imaging example. Whilst others have raised some of these issues, we evaluated the entire study from data collection to analysis, suggested sources for software and further reading, and provided a checklist plus flowchart for use with any ordinal data. We hope that raised awareness of the limitations of the current approaches will encourage greater method consideration and the utilization of a more appropriate analysis. More accurate comparisons between measures in medical imaging will lead to a more robust contribution to the imaging literature and ultimately improved patient care.

INTRODUCTION

Qualitative ordinal scores are often used for a range of activities in medical imaging. One common use of such ordinal scores is in the quality assessment of an image. Clinical image quality is measured in this way for a number of reasons, including the assessment of a change in imaging technique,[1–3] to compare imaging systems,[4] to assess a change in computer enhancement or processing (for instance the use of a new reconstruction algorithm in a CT scanner),[5,6] to measure image quality when optimizing the radiographic settings in radiography,[7,8] to assess the performance of a machine vision algorithm[9] or to compare methods of image quality assessment.[10] Viewing sessions are used to collect the subjective scores and usually display either one image at a time (absolute visual grading analysis) or show two images to compare (relative visual grading analysis),[2] with several observers often rating the image set.[11,12] Scores are usually collected using a three- or five-point ordinal[2,6–8,13] scale, labelled with words rather than numbers,[14] for example, "poor", "fair", "good", "very good" and "excellent",[15,16] or a Likert scale.[17] Recently annoyance or impairment related scales, rather than preference scales, have been used.[12] Alternative question formats include binary responses, which require the observer to select the "better" of two images regarding a feature of interest or to answer a simple yes/no response to the suitability of an image for a given task.[18]

It is not unusual for the analysis of such ordinal data to be oversimplistic or inappropriate. Once the non-numerical scale data have been collected, it is common practice to assign each ordinal category[13] a number, typically "1, poor"; "2, fair"; "3, good"; "4, very good" and "5, excellent".[19] The analyses which follow often utilize methods developed for numerical data, such as the arithmetic mean.[20] This measure can be referred to as the mean opinion score (MOS)[7,21–23] or the visual grading analysis score (VGAS).[7,10,24] Comparison of the summary values across different areas of interest are then used to answer the research question. For example, subjective scores for two

systems or computer processing methods[6,7] are often compared using *t*-tests or analysis of variance,[25] such as in the comparison of a newer and established X-ray system with respect to image quality. When the subjective ordinal measure is compared with an objective measure, for instance with measurements from a phantom image[10] or computer vision algorithm, the comparator variable is often on a continuous scale. Examples include the signal-to-noise ratio. In these cases, methods such as Spearman's and Pearson's correlation coefficients have been used[10,26,27] and their magnitude reported. Other common techniques include the use of Cohen's kappa statistic[28] to test for the agreement between observers or measures[2] or visual grading characteristics (VGCs) analysis; based upon receiver operator characteristic (ROC) curves and the area under the curve.[6,8]

This work highlights limitations with these data collection and analysis steps and their adverse consequences when interpreting study results, which have subsequent clinical implications. We suggest a number of ways in which improvements can be made through the study design and data analysis, and each approach is demonstrated using a medical imaging example. A checklist and associated flowchart are provided for guidance. Improving the methods for data collection, and using a more sophisticated and appropriate analysis, can provide more accurate results and demonstrate differences between groups that would otherwise not have be seen.[25]

LIMITATIONS OF COMMON APPROACHES

There are a number of problems statistically with the approaches currently adopted to perform image quality assessment on imaging systems; each of which are described here.

Questionnaire design

The scale labels

Using three- or five-point scales, with words rather than numbers, forms some assumptions. For example, it assumes the words excellent and good have the same meaning across observers; however, some observers may interpret excellent to be more positive than others.[29] Of course, inter-[30] and intra-observer[31] variability is a problem across all scales, but with these scales there is the additional variability associated with observer interpretation of the scale labels, which cannot be quantified and accounted for during the analysis. Observers may also interpret these words differently through the course of the viewing, depending upon the question posed. Additionally, it assumes symmetry in the scale such that excellent is as positive as poor is negative, whereas this may not be the case for all observers.[29]

In the following commonly used five-point scale: 1, poor; 2, fair; 3, good; 4, very good and 5, excellent,[19] there is not an obvious neutral response, although the centre option good could be seen as neutral from its scale position. The word fair may be interpreted as a neutral response, but its position on the scale, 2 of 5, may suggest it is more negative than positive. There is disagreement whether observers are more influenced by the wording or position of a category,[29,32] but some scales are formed with the intention that the mid-point represents a neutral response. Of course, whether to include a neutral response (usually by using an odd number of categories) is another area of debate, since it may be used whenever the observer cannot make a decision.[33–35] There are instances where the observer would like to rate an image between two categories, such as between fair and good, but this flexibility is not possible. The optimal number of points to use along a scale is another area of disagreement,[36–39] as are whether to include an option such as "don't know",[40] if a scale should be equally balanced between positive and negative responses[29] and if scale labels should be used in conjunction with scale numbers.[32]

If words are used to label the five points on the scale, the numbers are hidden from the observer and hence are meaningless. Assigning numbers arbitrarily is unhelpful and assumes the categories are equally spread in the decision space, as recognized by some authors.[2] For example, it assumes that the difference between excellent and very good is the same as between very good and good. This may be a correct assumption for some observers but not for others.[29] Assigning the numbers 1–2–3–4–5 may be as meaningful as assigning the numbers 1–24–56–789–1253 for example.

Analysis

Summary of the scale responses

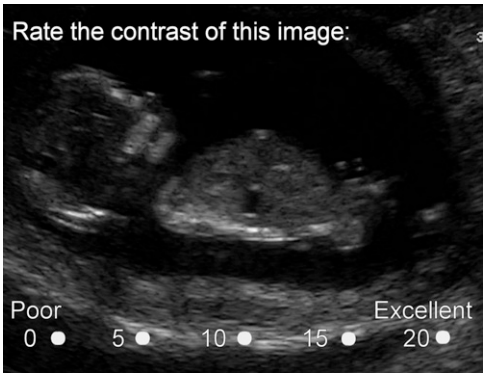
The data are collected using an ordinal scale,[13] yet summarized using a method designed for continuous data; the arithmetic mean.[20] This approach would be more suitable had the data been collected using an interval rather than ordinal scale[13] as correctly reported by some other imaging authors.[2,8,24,41]

Taking the mean of the responses, which can only take integer values between one and five, can also result in a value which is non-integer. If a mean response for a particular image is, for example, 3.4, there is no predefined word to interpret this average response from observers. It is known to be more positive than good but not as positive as very good, yet there is no definitive answer for its interpretation and hence it lacks meaning. The mean is also dependent upon the arbitrary coding given, therefore would differ if the scale 1–2–3–4–5 was used compared with 1–24–56–789–1253. More generally, reducing the image quality to a single score may be questionable,[41] since there are variations in contrast, resolution and noise. A simple mean value, or similar, may be an oversimplification of the information in the image.

Repeated images, patients or observers

Often studies use repeats, such as several images from the same patient, the same observer to rate multiple images or the same image altered in some way (this may include image degradation or enhancement).[11,12,18] Alternatively, observers may answer several questions regarding different aspects of the same image. If an observer likes or dislikes a particular image, their ratings may be similarly high or low for all questions relating to that image. It is also therefore expected that an image would be rated in a more similar way to an enhanced or degraded version of itself than an equally enhanced or degraded version of another image. It follows that images from the same patient would have more similarities with one another than with images from a different patient. Factors such as the patient characteristics or

Figure 1. Example question with numbered categories for the contrast of a crown rump length ultrasound image.



machine settings may lead to generally poor images from one patient yet excellent images from another.[42] Additionally, data are usually collected from multiple observers, each with their own opinions and standards. Observers may respond consistently between images, but they may not necessarily be consistent with one another.[30] This is particularly applicable if one observer rates more harshly than another, resulting in a similar ordering of images from best to worst, but with a shift along the ordinal scale.[13]

Many approaches do not allow for these similarities between repeated images, patients or observers, but instead assume that each response is independent.[6,8,24,41] This includes methods which summarize the ratings as a single figure such as the MOS or VGAS. This results in the standard errors being underestimated and hence possibly differing conclusions.[43] Approaches such as the VGC curve,[24] while able to compare two methods, may also struggle to incorporate additional variables or repeated data, owing to the increased complexity required of the methodology stemming from the assumption of two underlying normally distributed variables[44] and the basis of the method lying in ROC curves.[24]

Conditions and confounders

Conditions vary between the images, including the patient characteristics and machine settings, and these conditions can affect whether the image is fit for purpose.[45] There may also be confounders present, which could lead to confounding bias if not accounted for.[46] Confounding is the effect of an extraneous variable which wholly or partially accounts for an apparent association or which masks an underlying true association.[47] Confounders can be controlled for in the study design by matching the comparison groups on confounding variables, or by random allocation to one of the groups which leads to the assumption that the groups are comparable with respect to confounders. Alternatively, confounding variables can be included in the analysis to remove confounding bias. Many of the current approaches do not allow conditions or confounders to be incorporated into the analysis, hence could result in bias and lead to inaccurate findings. This includes those methods which reduce the data to a single figure (MOS or VGAS) or those which directly compare two methods (VGC).

Comparison of the subjective and objective measures

The final stage of the analysis usually compares either the often continuous objective measure with the subjective measure which originated from an ordinal scale, or two objective measures. Since both Spearman's and Pearson's correlation coefficients[10,26,27] are usually calculated, there seems to be uncertainty as to whether a parametric or non-parametric test is required.[48] Confirmation of the nature of both the subjective and objective measures should be used to determine the most suitable type of test, with all test assumptions verified. For example, Pearson's correlation assumes that the variables have a number of characteristics—the variables are interval or ratio measurements, are approximately normally distributed, have a linear relationship with one another and have minimal outliers—and homoscedasticity (equal variance). Spearman's correlation can be used if these assumptions are violated or if the data are ordinal and assumes a monotonic relationship between the variables (as one variable increases, the other increases or decreases, but not both).

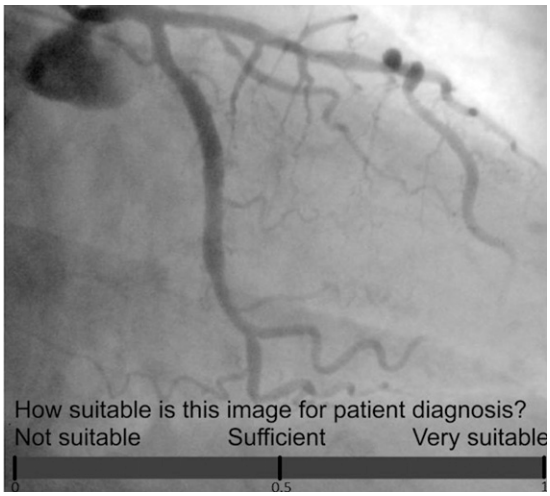
These correlation coefficients also do not allow for the likely complexity of the data structure such as its hierarchical nature[49] or any confounding factors, resulting in an oversimplification of the association between the objective and subjective measures.

Another method which may be used is an interrater agreement score such as Cohen's kappa statistic,[28] with a chosen interpretation of the 0–1 scale.[2] Although this offers a useful approach for comparison between observers, it is limited when most ratings for the observers gather at one level and is unhelpful when there is total agreement between observers for all images (since there is a division by zero).[2] It is also unable to account for any confounders or any repeats in the data. Finally, *t*-tests or analysis of variance are sometimes used,[25] but these assume the data to be interval, whereas the data in image assessment are usually ordinal.[25]

SUGGESTED IMPROVEMENTS

There are a number of ways in which these analyses may be improved upon. Six suggestions are given below, presented in the order in which they would appear in a study, along with their limitations and examples of how these approaches can be

Figure 2. Example question with a continuous scale for the suitability of a still image frame from a left coronary angiogram for stenosis identification.



applied to medical imaging data. Changes are suggested for the data collection phase or to the way in which the ordinal data are analysed, and these will be summarized in a checklist for best practice and a flowchart for guidance.

Questionnaire design

Numbered categories

One very simple amendment to the study design to improve on the current data collection would be to present the observer with categories labelled with numbers rather than words. The observer is instantly aware of the presumed equality of the gaps between categories and can interpret the response options accordingly. Words may still be used to indicate the extremes of the categories and possibly the mid-point, but fewer terms focus the attention of the observer towards the numbers and their equal spaces. The numbers need not be the integers 1–5, but could instead be 0, 5, 10, 15, 20, to prompt greater observer consideration of the scale and the equality of the gaps between the categories. However, it has been found that assigning different numbers affects responses,[50] so careful consideration of the scale is required.

This suggested interval approach would address the problems associated with the assumed equally spaced categories along the scale[13] but would not allow for repeated images, patients or observers within the analysis nor any conditions or confounders thought to affect the study findings. Therefore, this approach would be most suitable for simple data sets only.

Example An example of this type of scale can be seen in Figure 1 where the responses are numbered from 0 to 20. Only the first and last categories have words assigned to show the direction of the scale, but all five categories are given a number to show the differences between the categories.

Continuous scale

Another simple approach to circumvent some of the problems associated with the analysis of ordinal data would be to replace the ordinal scale with a more flexible numerical continuous scale. Values from 0 to 1 or 1 to 100 could be used, along which any point may be selected by the observer. These data could be analysed as a continuous outcome, using traditional analyses such as linear regression,[26,51] calculation of the mean or any other appropriate summary.[20] For comparison with previous studies, this continuous scale could be split into three or five categories, for example, 1–20, 21–40, 41–60, 61–80, 81–100, and analysed as ordinal data. Alternatively, a cut-off point could be specified, such as the mid-point where a scale changes from "disagree" to "agree", resulting in a binary response which could be analysed using standard approaches to binary outcomes, including logistic regression.[26,51]

This approach of amending the scale would again address any problems associated with the assumed equality of gaps between response categories but would unfortunately not allow for repeated images from repeated patients, viewed by multiple observers, nor any confounders or conditions in the study. Therefore, this approach would only be suitable for simple data sets.

Example Figure 2 shows an example of an image with a continuous scale which could be analysed as continuous data or, if necessary, dichotomized to a suitable/unsuitable scale or split into three or five categories for comparison with previous ordinal work.

Analysis

Summary of the scale response

If using scales with words, rather than assign numbers and summarize using the mean, instead select the median or modal response.[26] By choosing the mid-point or the category which appears most frequently, the summary value can be interpreted using the wording assigned to the original category. For ordinal data, the median is usually recommended.[13]

Example A study uses the five-point scale poor, fair, good, very good and excellent to collect data from 60 observers about the

contrast of an X-ray image, where excellent is considered to be high contrast. The results are as shown in Table 1.

The modal response is simply the most selected category. Therefore for Image 1, good is the most frequently selected category, and for Image 2, it is fair, suggesting Image 1 has better contrast than Image 2. The median is the middle number after ordering the responses and hence for Image 1 is good and for Image 2 is fair.

Regression analysis

Regression[51] would allow more study information to be incorporated in the analysis such that any confounders or conditions can be accounted for, resulting in a more accurate summary of the data than the frequently used MOS.

Different regression models are available depending on the nature of the scale used to record responses from the observers. Logistic regression[51] is useful for binary (often yes/no) responses and linear regression[51] is useful for continuous scales (for example, any point between zero and one). Ordinal regression models may be most useful in medical imaging as the data are frequently recorded using ordered categories.[52] Each of these forms of regression allow multiple independent variables and hence can incorporate confounders or any other variables thought to be important for the study results.

This approach is suitable for confounders but does not allow for a hierarchical structure of images, patients and observers.

Example A study is conducted which collects responses from observers regarding the suitability of an image for a given task. Images within the viewing session are at a range of different contrasts to determine a contrast value at which images become unsuitable. However, the images also contain some noise, which is known to affect the suitability judgment from the observers and can cause the contrast to appear lower. According to the definition,[47] noise is a confounder since it affects the perceived contrast plus the suitability of an image for the identified task. To account for the effect of noise, it can be included in the regression model used for analysis.[53]

Let there be a measure for the suitability of the image taken from the observers, along with a measurement for noise and for contrast; these may be with respect to a reference image. The association of interest is between the contrast value and the response from the observer. Let the regression model have response as the dependent variable, and both contrast and noise as the independent variables. This enables the contrast and noise variables to predict the response and, consequently, reduce the confounding bias from noise. Any other recorded confounders can be added to the model in the same way to reduce bias.[53] Linear regression can be used when the responses are collected using a continuous scale, and logistic or ordinal regression can be used when the data are collected using two or more categories, respectively.[52]

Multilevel proportional odds model

Proportional odds models,[54] also referred to as ordered logit models or ordered logistic regression models, are an extension of logistic models, which allow for an outcome with more than two ordered categories.[52] Therefore, these models are ideal for responses in medical imaging which are often recorded using a five-point scale.

Multilevel models,[55] also known as mixed, random-effects, nested or hierarchical models, can be used for continuous responses and allow the analysis to account for repeated images, patients and observers within a data set. The "levels" correspond to responses that are given for each repeat or variant of a given image, from each of the patients who have provided images, rated by each of the observers participating in the viewing session. Defining these levels in the model allow it to account for the similarities between these repeats which lead to responses which are not truly independent. The levels suggest the research question does not relate to the variables which define the levels but rather the wider population from which they were drawn.[56] In medical imaging, this may be that the particular images or patients within a viewing are not of particular interest, but the wider populations of images and patients are. Hence, these variables are often referred to as random effects, or nuisance parameters.[56] The levels also suggest that the responses are expected to differ between different categories of a level, but these differences cannot be explained *via* the measured variables.[56] For example, responses regarding the quality of images taken from one patient may be consistently higher than responses from images taken from another patient, and these differences may be due to underlying patient characteristics.[45] Any variables included in the model which do not form the levels are referred to as fixed effects.[56]

Multilevel models are therefore a type of model which can allow for the similarity between images taken from a particular patient or a group of patients from the same hospital but allow for differences from one image or one patient to another. For further details on fixed and random effects, an introductory tutorial is given by Winter,[57] which includes examples using the lme4 package[58] in the statistical software R (R Core Team, Vienna, Austria).[59] Here, a random effect is described as something expected to have a non-systematic, idiosyncratic or random influence on the data, while fixed effects are expected to have a systematic and predictable influence on the data.[57] Fixed

Table 1. The contrast study results

Contrast	Poor	Fair	Good	Very good	Excellent
Image 1	3	8	34	9	6
Image 2	14	23	8	8	7

effects can be thought to "exhaust" the population of interest or levels of a factor, such as including both levels of sex (male/female) or all levels of machine setting (such as high/medium/low).[57] Random effects are usually a sample from the population of interest,[57] such as some images from a database or some observers from those trained to look at the images of interest.

Multilevel proportional odds models[60] offer the most flexibility of the suggestions given here and can be implemented using a range of general statistical software[59,61,62] or specialized multilevel software.[63] They combine the advantages of both the proportional odds model which allows the outcome to consist of ordered categories, with the ability of multilevel models to account for repeats within the study design. However, the proportional odds assumption must be met for it to be suitable.[60] If not satisfied, the multinomial multilevel model can be used,[64] but this can be difficult to interpret and therefore a statistician is recommended. Further examples of this approach in medical imaging can be found in the literature.[44]

Example—multilevel model Let there be a study conducted into the observer annoyance[12] induced by a coronary angiogram. The study includes three images, each shown at four different simulated X-ray dose levels using image-degrading software. The original image is shown along with simulated reductions to 80%, 60% and 40% of the original dose. The question of interest concerns the level of dose reduction tolerable by observers and consequently the level at which they become annoyed by the image. Observers are asked to respond using a continuous scale.

It is expected that a particular image may annoy an observer more than another, regardless of the dose level. In coronary angiography, there may have been a bad projection angle used, the patient might have a large body mass index or the radiopaque dye might not be injected properly; all factors which may cause observer annoyance.[45]

Therefore, there may be similarities in the responses from the same image at different simulated dose levels. Let the data set include the (repeated) image number relating to the original image, alongside the four dose levels. Analysis using a multilevel model according to the image number accounts for the repeated use of the same image after degradation. The same method can be used to account for repeated observers.

Example—multilevel proportional odds model Let there be a viewing session which records responses using a 5-point scale from 10 observers. Each observer rates 36 images; 2 images from 6 patients, each at 3 simulated noise levels. The question of interest relates to the simulated noise within an image and the image quality. However, it is known that observers may respond differently to one another, for example, with some rating more harshly than others; that patients have different characteristics which may affect the quality of the image; and that some original images may be of higher quality than others.

A multilevel proportional odds model can be used, with the levels defined to be the images within the patients, looked at by the observers, and with the five response categories as the model outcome. Thus, the model has incorporated both the hierarchical structure of the data set and the repeated values, plus the ordinal nature of the responses. Any confounders or conditions thought to affect image quality can also be included in the model as extra independent variables.[44]

Comparison of the subjective and objective measures

The nature of the subjective and objective measures, such as whether they are continuous, categorical or normally distributed, for example, should be considered so that comparisons can be made while satisfying the assumptions of any tests or methods used. Resources are available which give the description of the two measures and suggest a suitable method to compare them.[65,66] Methods include *t*-tests, regression models and Mann–Whitney tests.[26] The choice of method will also depend upon the structure of the data. For example, if there is a hierarchical structure or if the data are affected by confounders, a regression model which can allow for these features would be recommended. If the data have a simple structure and there are no repeats within it, then an appropriate test or correlation may be suitable. The requirements and assumptions of any approach should be obtained and verified before the comparison is completed.

Example Let there be a study conducted which collects responses from observers regarding image quality using a five-point scale, to compare machines from two manufacturers. The question of interest is whether there is a difference in image quality between the different manufacturers, since one is considerably cheaper than the other. The viewing session displays 32 images; 2 images from each of 8 patients from 1 machine, and 2 images from another 8 patients on the other. 10 observers are enrolled to the study and each observer rates all the images, resulting in 320 responses from the observers on an ordinal scale. Factors affecting image quality from the patients and machines are also recorded as confounders.

For an ordinal subjective measure and a binary manufacturer choice with continuous confounders, an ordinal regression model should be used. The details of the ordinal regression model depend upon the nature of the data gathered; in this instance, hierarchical data with repeated patients and observers. Therefore, a multilevel proportional odds model is required.

COMPARISON OF THE MEAN OPINION SCORE AND MULTILEVEL PROPORTIONAL ODDS MODEL USING IMAGE QUALITY DATA

Let there be a study comparing two image-processing methods in radiography. Processing type A is a form of image processing currently used in capturing images for diagnostic purposes, and processing type B is an alternative method for image enhancement. The research question is whether type B can produce images which are as useful for diagnostic purposes as type A.

18 raw images were taken from 5 patients (3 images from 2 patients and 4 images from 3), and the 2 image-processing

Table 2. The raw ordinal response data

Processing	Score 1	Score 2	Score 3	Score 4	Score 5
Processing type A	66	51	157	91	67
Processing type B	47	43	145	101	96

techniques were applied. 24 observers were asked to rate the image quality of the resulting 36 processed images (18 from each processing type) on a 1–5 scale (worst–best). The results are shown in Table 2. Both MOS analysis and a multilevel proportional odds model will be applied to the data, for comparison of the results.

Mean opinion score analysis

The average score for processing type A and B must be calculated using the data in Table 2 and a difference between the two sought.

Processing type A MOS

$$= \frac{(66 \times 1) + (51 \times 2) + (157 \times 3) + (91 \times 4) + (67 \times 5)}{(66 + 51 + 157 + 91 + 67)}$$
$$= \frac{1338}{432}$$
$$= 3.097$$

Processing type B MOS

$$= \frac{(47 \times 1) + (43 \times 2) + (145 \times 3) + (101 \times 4) + (96 \times 5)}{(47 + 43 + 145 + 101 + 96)}$$
$$= \frac{1452}{432}$$
$$= 3.361$$

The average score is therefore slightly higher for type B than type A, but there appears to be no sizeable difference between the processing types in terms of image quality.

Multilevel proportional odds model analysis

The data in Table 2, in conjunction with observer and patient information, can be used to form a multilevel proportional odds model. Observers are considered to be repeated, as they each view more than 1 image (a total of 36), as are patients who each provide 3 or 4 images. Since the same raw images are used for both processing types, all machine settings except the processing type remain constant; hence, these variables are not considered to be confounders and do not need to be included in the model. For measures such as dose and patient size to be classified as confounders, they would need to affect image quality, which they do, but also the processing type, which in this case they do not. Processing type can be put into the model as usual in regression modelling (as a fixed effect), but the observer and patient variables must be entered into the model as random effects, since they are repeated measures with the 24 observers each having viewed the 36 images taken from the 5 patients. The analysis was completed using R statistical software[59] but can also be carried out using other software packages.[61–63]

The fixed effects are shown in Table 3, where the estimate for processing type is highly significant ($p = 3.03 \times 10^{-6}$), showing processing type B to have generally higher quality scores than type A (positive estimate).

Table 4 displays the random effects, showing a measure of the variability for observer and patient. Patients show more variability than observers, suggesting greater differences between patients than between observers in relation to the quality score.

The research question concerned differences in image quality between processing types A and B. The MOS analysis reported mean values just above three for each processing type (slightly higher for B than A). However, the estimate for processing type in the multilevel proportional odds model was highly significant and showed processing type B to give significantly higher image quality ratings than processing type A. It may be that patient characteristics such as patient thickness affected the image quality and that some observers scored the image quality more generously than others. These differences between patients/observers and similarities within patients/observers were not accounted for during the MOS analysis.

Therefore, taking into account the repeated nature of both observers and patients, the conclusion from the study is clearer. The same data set was used for both methods, but less variables were included in the MOS analysis.

DISCUSSION

We have highlighted some weaknesses in the methods currently used to analyse data collected from medical imaging viewing studies, where method assumptions are not always known, checked or adhered to. We have suggested simple amendments to the data collection, as well as more sophisticated analysis models to include conditions within the data which have not previously routinely been accounted for. We encourage

Table 3. Output from the multilevel proportional odds model: fixed effects

Fixed effects	Estimate	Standard error	p-value
Processing type B	0.615	0.132	3.03×10^{-6}

Table 4. Output from the multilevel proportional odds model: random effects

Random effects	Variance	Standard deviation
Observer	1.625	1.275
Patient	2.168	1.472

researchers to implement a multilevel proportional odds model where appropriate but suggest the consideration of the other approaches given too. A flowchart is shown in Figure 3 which guides researchers through data collection suggestions when data have not yet been acquired and, subsequently, recommends an analysis approach given the structure of the data. A checklist is also provided in Table 5 detailing best practice during the study planning phase, data collection, analysis and results reporting. Imaging examples have been used to demonstrate each of the suggestions, and these approaches can be implemented using packages in statistical software such as R[59] or Stata® (StataCorp, College Station, TX).[61] Some statistical knowledge is required, so the guidance of a statistician may be necessary, and any model assumptions must be verified to ensure the model is valid before the results are interpreted.

We have not discussed here the issue of sample size and statistical power of a study. Calculations exist for simple models and tests, and there are some general guides available,[67,68] but these become less common as the complexity of the analysis increases, and simulation[67,68] is often recommended instead. Although it is desirable to have a large number of observers and a large number of images in a study, in practice these numbers will largely be determined by the availability of images and observers and restricted by time constraints of observers to perform the study. It is also likely that these two factors are inversely associated, and a compromise must be achieved to maintain reasonable statistical power.[68] Care must be taken, however, to ensure that the study is not compromised by too narrow a selection of observers or images. In addition, there are modelling requirements to abide by, such as the ratio of the number of parameters and number of observations, to avoid overfitting a model.[44] For example, in logistic regression, the number of observations should be at least 10 times the number of parameters.[44,67]

The focus here is on evaluating image quality using visual assessment and a given criteria, often referred to as visual grading. However, ordinal data are also used for assessing agreement between observers, devices and methods or for assessing the agreement with an accepted reference standard. In each of these three scenarios, different analyses will be required. When testing agreement between two or more ratings, analyses such as polychoric correlation or the weighted kappa statistic, an extension of Cohen's kappa statistic,[28] may be suitable, whereas approaches such as ROC curves[6,8] are more suited to agreement with a given value such as in diagnostic accuracy. The inclusion of observers as random effects in a regression model has been suggested here for the assessment of image quality which does not require a ground truth,[44] but the information relating to the observers themselves will be minimal and hence method choice will be affected by the variable(s) of interest. Whichever method is selected and whatever be the purpose of the analysis, all assumptions should be checked and adhered to.

Figure 3. Method flowchart tool.

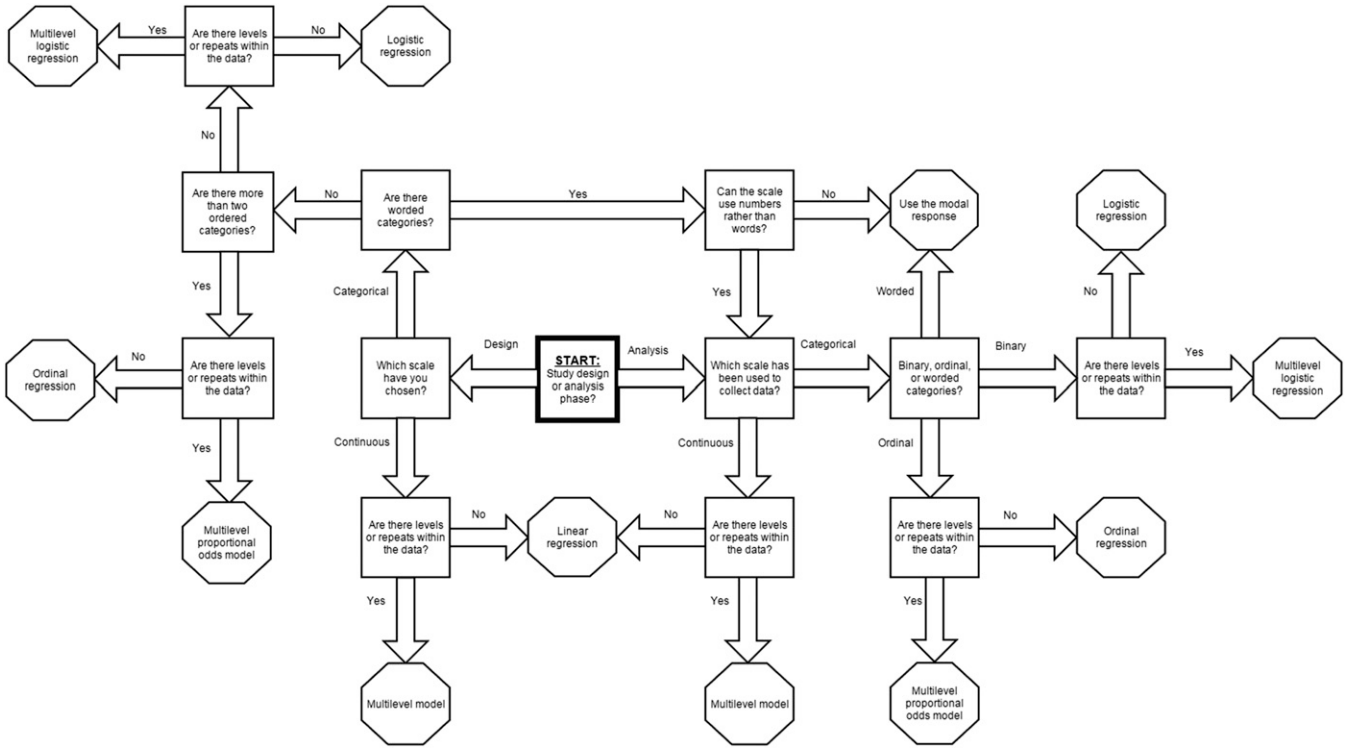


Table 5. Medical imaging viewing checklist

Area	Item	Recommendation	Check
Planning			
Observer selection	1	Determine a comfortable study length and select observers with suitable experience	□
Patient selection	2	Choose patients who are suitable for the research question. Unless necessary, do not select multiple images from one patient	□
Image selection	3	Choose appropriate images to address the research question. Unless comparing two approaches which can utilise the same image set, *e.g.* image processing, select each image only once	□
Data collection			
Scale	4	Where possible, collect the data using a continuous scale, labelled at each end to show scale direction and in the centre as a reference point	□
Confounders	5	Collect data on any variable thought to be a confounding factor. This may relate to the image, observer or viewing conditions	□
Repeats	6	Record any repeats present in the viewing, such as repeated images, patients providing multiple images, or observers providing a response to more than one image	□
Data analysis			
Scale	7	Decide upon the outcome of interest and amend the scale accordingly. The continuous scale adopted allows for a continuous outcome, a binary (two category) outcome or an ordinal (more than two category) outcome. Categorise the scale accordingly and justify the scale chosen	□
Analysis	8	Select an appropriate means of analysis using the flowchart (Figure 3)	□
Reporting			
Detail and justify	9	Ensure all aspects of the study design, data collection and analysis have been included in the report and all choices justified. Include assumptions for any analysis conducted	□
Interpret	10	Present all statistical findings and a full interpretation of the results	□

Although the focus here has been on image quality, the true importance if using radiation is regarding patient dose, which is assumed to be positively associated with image quality and which is the reason for many of the viewing studies.[69] Regression models in general allow the inclusion of continuous variables relating to factors such as dose and hence the effect of these parameters can be assessed.[44] A method has been suggested which successfully uses ordinal logistic regression with random effects to quantify the potential for dose reduction using post-processing and which supports the analytical suggestions here.[70] Although other authors have highlighted some of the issues discussed here,[8,24,25] with some also proposing alternative methods of analysis, not all can be used in all scenarios associated with subjective image analysis. For example, some authors suggest[24] and others use[6,8] VGC analysis which is advantageous over using the MOS since it does not make assumptions regarding the distribution of the data nor does it average ordinal data. However, VGC analysis is based upon ROC curves and is unable to account for repeated measures in the data or to return information regarding the importance of confounding variables. It is also affected by the interobserver variability.[24] The approach suggested here, namely multilevel modelling, provides results relating to these additional variables and permits repeated images, observers and patients. Consideration of these factors during the analyses is highlighted by other authors,[41,44] with one of these publications[44] also recommending ordinal logistic regression. Additional recent publications agreeing with our conclusions include an evaluation[69] of several regression models, which recommends ordinal logistic regression for ordinal data from visual grading experiments in medical imaging, as well as an approach for quantifying potential dose reduction using ordinal data which also uses ordinal logistic regression.[70] Although other authors have drawn conclusions supporting our message here, many focus on the analysis,[41] whereas we

have covered the entire study design, from questionnaire format to data analysis.

CONCLUSION

Greater insight can be gained through improved experimental design and appropriate analytical methods for ordinal data in image quality assessment. We have highlighted a number of limitations in common approaches and provided a checklist and accompanying flowchart for guidance on how to approach different situations. These suggested improvements can be used not only in future studies and to contribute to the medical imaging literature but the suggestions relating to data analysis can also be used to reanalyse previous studies to verify older findings. More informative results with less bias will lead to greater knowledge in medical imaging which should impact positively on future patient care.

FUNDING

This work has been supported by the EU-funded PANORAMA project, funded by grants from Belgium, Italy, France, Netherlands, UK and the ENIAC Joint Undertaking. The funding source had no involvement in the study design, in the collection, analysis and interpretation of data, in the writing of the report or the decision to submit the article for publication.

REFERENCES

1. Wanyonyi SZ, Napolitano R, Ohuma EO, Salomon LJ, Papageorghiou AT. Image-scoring system for crown-rump length measurement. *Ultrasound Obstetrics Gynecol* 2014; **44**: 649–54. doi: http://dx.doi.org/10.1002/uog.13376
2. Geijer H, Beckman K, Jonsson B, Andersson T, Persliden J. Digital radiography of scoliosis with a scanning method: initial evaluation. *Radiology* 2001; **218**: 402–10. doi: http://dx.doi.org/10.1148/radiology.218.2.r01ja32402
3. Gorham S, Brennan PC. Impact of focal spot size on radiologic image quality: a visual grading analysis. *Radiography* 2010; **16**: 304–13. doi: http://dx.doi.org/10.1016/j.radi.2010.02.007
4. Davies AG, Cowen AR, Kengyelics SM, Moore J, Sivananthan UM. Do flat detector cardiac X-ray systems convey advantages over image-intensifier-based systems? Study comparing X-ray dose and image quality. *Eur Radiol* 2007; **17**: 1787–94. doi: http://dx.doi.org/10.1007/s00330-006-0458-0
5. Seeram E, Seeram D. Image postprocessing in digital radiology—a primer for technologists. *J Med Imaging Radiat Sci* 2008; **39**: 23–41. doi: http://dx.doi.org/10.1016/j.jmir.2008.01.004
6. Leander P, Soderberg M, Falt T, Gunnarsson M, Albertsson I. Post-processing image filtration enabling dose reduction in standard abdominal CT. *Radiat Prot Dosimetry* 2010; **139**: 180–5. doi: http://dx.doi.org/10.1093/rpd/ncq086
7. Wiltz HJ, Petersen U, Axelsson B. Reduction of absorbed dose in storage phosphor urography by significant lowering of tube voltage and adjustment of image display parameters. *Acta Radiologica* 2005; **46**: 391–5. doi: http://dx.doi.org/10.1080/02841850510021184
8. Martin L, Ruddlesden R, Makepeace C, Robinson L, Mistry T, Starritt H. Paediatric X-ray radiation dose reduction and image quality analysis. *J Radiological Prot* 2013; **33**: 621–33. doi: http://dx.doi.org/10.1088/0952-4746/33/3/621
9. Szeliski R. *Computer vision: algorithms and applications*. London, UK: Springer-Verlag; 2011.
10. Moore CS, Wood TJ, Beavis AW, Saunderson JR. Correlation of the clinical and physical image quality in chest radiography for average adults with a computed radiography imaging system. *Br J Radiol* 2013; **86**: 1–12. doi: http://dx.doi.org/10.1259/bjr.20130077
11. Geijer H, Geijer M, Forsberg L, Kheddache S, Sund P. Comparison of color LCD and medical-grade monochrome LCD displays in diagnostic radiology. *J Digital Imaging* 2007; **20**: 114–21. doi: http://dx.doi.org/10.1007/s10278-007-9028-5
12. Umbaugh SE. *Computer imaging: digital imaging analysis and processing*. Boca Raton, FL: CRC Press; 2005.
13. Stevens SS. On the theory of scales of measurement. *Science* 1946; **103**: 677–80. doi: http://dx.doi.org/10.1126/science.103.2684.677
14. 2011 Census questionnaire for England. UK: Office for National Statistics. [Updated 2015, cited 25 August 2015.] Available from: http://www.ons.gov.uk/
15. Di Palo MT. Rating satisfaction research: is it poor, fair, good, very good, or excellent? *Arthritis Care Res* 1997; **10**: 422–30.
16. Kane RL, Radosevich DM. *Conducting health outcomes research*. Sudbury, MA: Jones & Bartlett Learning; 2011.
17. Likert R. A technique for the measurement of attitudes. *Arch Psychol* 1932; **140**: 1–55.
18. Keeble C, Wolstenhulme S, Davies AG, Evans JA. Is there agreement on what makes a good ultrasound image? *Ultrasound* 2013; **21**: 118–23. doi: http://dx.doi.org/10.1177/1742271X13493088
19. Jones AM, Rice N, Bago d'Uva T, Balia S. *Applied health economics*. Oxon, UK: Routledge; 2013.
20. Medhi J. *Statistical methods: an introductory text*. New Delhi, India: New Age International Limited Publishers; 1992.
21. International Telecommunication Union. *Series P: telephone transmission quality. Methods for objective and subjective assessment of quality*. Geneva, Switzerland: ITU; 1996.
22. Bankman IS. *Handbook of medical image processing and analysis*. San Diego, CA: Academic Press; 2008.
23. Sandborg M, Tingberg A, Dance DR, Lanhede B, Almn A, McVey G, et al. Demonstration of correlations between clinical and physical image quality measures in chest and lumbar spine screen-film radiography. *Br J Radiol* 2001; **74**: 520–528. doi: http://dx.doi.org/10.1259/bjr.74.882.740520
24. Bath M, Mansson LG. Visual grading characteristics (VGC) analysis: a nonparametric rank-invariant statistical method for image quality evaluation. *Br J Radiol* 2007; **80**: 169–76.
25. Zarb F, McEntee MF, Rainford L. Visual grading characteristics and ordinal regression analysis during optimisation of CT head examinations. *Insights Into Imaging* 2015; **6**: 393–401. doi: http://dx.doi.org/10.1007/s13244-014-0374-9
26. Harris M, Taylor G. *Medical statistics made easy*. Banbury, UK: Scion Publishing Ltd; 2014.
27. Myers JL, Well AD. *Research design and statistical analysis*. New York, NY: Lawrence Erlbaum Associates; 2003.
28. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; **20**: 37–46. doi: http://dx.doi.org/10.1177/001316446002000104

29. Friedmann HH, Amoo T. Rating the rating scales. *J Marketing Management* 1999; **9**: 114–23.
30. Caro TM, Roper R, Young M, Dank GR. Inter-observer reliability. *Behaviour* 1979; **69**: 303–15. doi: http://dx.doi.org/10.1163/156853979X00520
31. Wever JJ, Blankensteijn JD, van Rijn JC, Broeders IA, Eikelboom BC, Mali WP. Inter- and intraobserver variability of CT measurements obtained after endovascular repair of abdominal aortic aneurysms. *AJR Am J Roentgenol* 2000; **175**: 1279–82.
32. Wildt AR, Mazis MB. Determinants of scale response: label *versus* position. *J Mark Res* 1978; **15**: 261–7. doi: http://dx.doi.org/10.2307/3151256
33. Garland R. The mid-point on a rating scale: is it desirable? *Mark Bull* 1991; **2**: 66–70.
34. Kalton G, Roberts J, Holt D. The effects of offering a middle response option with opinion questions. *J R Stat Soc Series D* 1980; **29**: 65–78.
35. Guy RF, Norvell M. The neutral point on a Likert scale. *J Psychol* 1977; **95**: 199–204. doi: http://dx.doi.org/10.1080/00223980.1977.9915880
36. Friedman HH, Friedman EM. A comparison of six overall evaluation rating scales. *J Int MarkRes* 1986; **22**: 129–38.
37. Churchill GA Jr, Peter JP. Research design effects on the reliability of rating scales: a meta analysis. *J Mark Res* 1984; **21**: 360–75. doi: http://dx.doi.org/10.2307/3151463
38. Cox EP. The optimal number of response alternatives for a scale: a review. *J Mark Res* 1980; **17**: 407–22.
39. Jacoby J, Matell MS. Three-point Likert scales are good enough. *J Mark Res* 1971; **8**: 495–500. doi: http://dx.doi.org/10.2307/3150242
40. Tull DS, Hawkins DI. *Marketing research: measurement and method*. New York, NY: Macmillan Publishing; 1993.
41. Bath M. Evaluating imaging systems: practical applications. *Radiat Prot Dosimetry* 2010; **139**: 26–36. doi: http://dx.doi.org/10.1093/rpd/ncq007
42. Ho JS, Barlow CE, Reinhardt DB, Wade WA, Cannaday JJ. Effect of increasing body mass index on image quality and positive predictive value of 100-kV coronary computed tomographic angiography. *Am J Cardiol* 2010; **106**: 1182–6. doi: http://dx.doi.org/10.1016/j.amjcard.2010.06.032
43. Park S, Lake ET. Multilevel modeling of a clustered continuous outcome: nurses work hours and burnout. *Nurs Res* 2005; **54**: 406–13. doi: http://dx.doi.org/10.1097/00006199-200511000-00007
44. Smedby O, Fredrikson M. Visual grading regression: analysing data from visual grading experiments with regression models. *Br J Radiol* 2010; **83**: 767–75. doi: http://dx.doi.org/10.1259/bjr/35254923
45. Bushberg JT, Seibert JA, Leidholdt EM Jr, Boone JM. *The essential physics of medical imaging*. Philadelphia, PA: Lippincott Williams & Wilkins; 2011.
46. McNamee R. Confounding and confounders. *Occup Environ Med* 2003; **60**: 227–34. doi: http://dx.doi.org/10.1136/oem.60.3.227
47. Schlesselman JJ. *Case-control studies: design, conduct, analysis*. New York, NY: Oxford University Press; 1982.
48. Sheskin DJ. *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, FL: Chapman & Hall; 2011.
49. Gelman A, Hill J. *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press; 2006.
50. Schwarz N, Knauper B, Hipler HJ, Noelle-Neumann E, Clark L. Numeric values may change the meaning of scale labels. *Public Opin Q* 1991; **55**: 570–82.
51. Freedman DA. *Statistical models: theory and practice*. New York, NY: Cambridge University Press; 2009.
52. Bender R, Grouven U. Ordinal logistic regression in medical research. *J R Coll Physicians Lond* 1997; **31**: 546–51.
53. Hosmer DW Jr, Lemeshow S, Sturdivant RX. *Applied logistic regression*. Hoboken, NJ: John Wiley & Sons; 2013.
54. Gameroff MJ. Using the proportional odds model for health-related outcomes: why, when, and how with various SAS procedures. SAS Institute Inc., New York, NY. [Updated 2005, cited 25 August 2015.] Available from: http://www2.sas.com/proceedings/sugi30/205-30.pdf
55. Snijders TA, Bosker R. *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. London, UK: SAGE Publications Ltd; 2011.
56. Snijders TA. Fixed and random effects. In: *Encyclopedia of statistics in behavioral science*. Everitt BS, Howell DC, eds. Chichester, UK: Wiley; 2005.
57. Winter B. A very basic tutorial for performing linear mixed effects analyses (Tutorial 2). 2013 [updated 19 May 2014, cited 24 September 2015]. Available from: http://www.bodowinter.com/tutorial/bw_LME_tutorial2.pdf
58. Bates D, Maechler M, Bolker B, Walker S. lme4: linear mixed-effects models using Eigen and S4. R package version 1.1-8. 2015. Available from: http://CRAN.R-project.org/package=lme4
59. R Core Team. *R: a language and environment for statistical computing*. Vienna, Austria. R Foundation for Statistical Computing; 2015.
60. De Silva DB, Sooriyarachchi MR. Generalized linear multilevel models for ordinal categorical responses: methods and application to medical data. *Sri Lankan J Appl Stat* 2012; **12**: 83–99.
61. StataCorp. *Stata statistical software: release 14*. College Station, TX: StataCorp LP; 2015.
62. IBM Corp. *IBM SPSS Statistics for Windows, version 22.0*. Armonk, NY: IBM Corp.; 2013.
63. Rasbash J, Charlton C, Browne WJ, Healy M, Cameron B. *MLwiN version 2.02*. Bristol, UK: University of Bristol, Centre for Multilevel Modelling; 2005.
64. Leyland AH, Goldstein H. *Multilevel modelling of health statistics*. Chichester, UK: Wiley; 2001.
65. What statistical test do I need? Sheffield, UK: Mathematics and Statistics Help (MASH), University of Sheffield, Sheffield, UK. [Updated 2013, cited 25 August 2015.] Available from: http://www.mash.dept.shef.ac.uk/Resources/MASHWhatStatisticalTestHandout.pdf
66. Marenco A. When to use a particular statistical test. California State University. [Updated 2007, cited 25 August 2015.] Available from: http://www.csun.edu/~amarenco/Fcs%20682/When%20to%20use%20what%20test.pdfs
67. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Med* 2004; **66**: 411–21.
68. Snijders TAB. Power and sample size in multilevel linear models. In: Everitt BS, Howell DC, eds. *Encyclopedia of statistics in behavioral science*. Volume 3. Chicester, UK: Wiley; 2005. pp. 1570–3.
69. Saffari ES, Love A, Fredrikson M, Smedby O. Regression models for analyzing radiological visual grading studies—an empirical comparison. *BMC Med Imaging* 2015; **15**: 49. doi: http://dx.doi.org/10.1186/s12880-015-0083-y
70. Smedby O, Fredrikson M, De Geer J, Borgen L, Sandborg M. Quantifying the potential for dose reduction with visual grading regression. *Br J Radiol* 2013; **86**: 31197714. doi: http://dx.doi.org/10.1259/bjr/31197714