RESEARCH ARTICLE

Using ROC Curves to Choose Minimally Important Change Thresholds when Sensitivity and Specificity Are Valued Equally: The Forgotten Lesson of Pythagoras. Theoretical Considerations and an Example Application of Change in Health Status

Robert Froud[1,2]*, **Gary Abel**[3]

1. Clinical Trials Unit, Warwick Medical School, University of Warwick, Gibbet Hill Road, Coventry, United Kingdom, **2.** Norge Helsehøyskole, Campus Kristiania, Prinsens Gate 7–9, Oslo, Norway, **3.** Cambridge Centre for Health Services Research, University of Cambridge, Robinson Way Cambridge, Cambridgeshire, United Kingdom

*r.froud@warwick.ac.uk

Citation: Froud R, Abel G (2014) Using ROC Curves to Choose Minimally Important Change Thresholds when Sensitivity and Specificity Are Valued Equally: The Forgotten Lesson of Pythagoras. Theoretical Considerations and an Example Application of Change in Health Status. PLoS ONE 9(12): e114468. doi:10.1371/journal.pone.0114468

Editor: Joan A. Caylà, Public Health Agency of Barcelona, Spain

Received: June 2, 2014

Accepted: November 7, 2014

Published: December 4, 2014

Data Availability: The authors confirm that, for approved reasons, some access restrictions apply to the data underlying the findings. Data are from the UK BEAM trial and are not owned by the authors. Data requests may be sent to Prof. David Torgerson, Department of Health Sciences, University of York (david.torgerson@york.ac.uk) and Prof. Martin Underwood, Clinical Trials Unit, University of Warwick (m.underwood@warwick.ac.uk).

Funding: The authors received no specific funding for this work.

Competing Interests: GA declares that he has no competing interests. RF is also a director and shareholder of a company that provides electronic measurement services to health services researchers; notwithstanding this, he declares that he has no conflicts of interest. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

Abstract

Background: Receiver Operator Characteristic (ROC) curves are being used to identify Minimally Important Change (MIC) thresholds on scales that measure a change in health status. In quasi-continuous patient reported outcome measures, such as those that measure changes in chronic diseases with variable clinical trajectories, sensitivity and specificity are often valued equally. Notwithstanding methodologists agreeing that these should be valued equally, different approaches have been taken to estimating MIC thresholds using ROC curves.

Aims and objectives: We aimed to compare the different approaches used with a new approach, exploring the extent to which the methods choose different thresholds, and considering the effect of differences on conclusions in responder analyses.

Methods: Using graphical methods, hypothetical data, and data from a large randomised controlled trial of manual therapy for low back pain, we compared two existing approaches with a new approach that is based on the addition of the sums of squares of 1-sensitivity and 1-specificity.

Results: There can be divergence in the thresholds chosen by different estimators. The cut-point selected by different estimators is dependent on the relationship between the cut-points in ROC space and the different contours described by the estimators. In particular, asymmetry and the number of possible cut-points affects threshold selection.

Conclusion: Choice of MIC estimator is important. Different methods for choosing cut-points can lead to materially different MIC thresholds and thus affect results of responder analyses and trial conclusions. An estimator based on the smallest sum of squares of 1-sensitivity and 1-specificity is preferable when sensitivity and specificity are valued equally. Unlike other methods currently in use, the cut-point chosen by the sum of squares method always and efficiently chooses the cut-point closest to the top-left corner of ROC space, regardless of the shape of the ROC curve.

Introduction

Initially developed during World War II for use in interpreting radar signals, receiver operator characteristic (ROC) curves are commonly used in medical research to evaluate screening tests, identify thresholds to facilitate decision-making about patients, and to quantify the responsiveness of quasi-continuous patient-reported outcomes measures (PROMs). [1, 2] Our focus in this paper is on their use to estimate minimally important change (MIC) thresholds on PROMs. [3]

MIC is usually defined as smallest magnitude of change that can be considered important (at the level of the individual), and in the absence of troublesome side-effects and excessive costs, mandates a change in a patient's management. [4, 5] In 1986, Deyo and Centor suggested that PROMs could be viewed as diagnostic tests; in the sense that they can be thought to be diagnostic of improvement. [6] Using an external criterion – for example a health transition question or an accepted gold standard diagnostic test for improvement – continuous, or quasi-continuous, PROM change scores can be plotted on a ROC curve, facilitating a choice of cut-point associated with optimal sensitivity and specificity. [6] When estimating MIC in chronic conditions with variable clinical trajectories, such as low back pain, sensitivity and specificity are assumed to be of equal value. [7–10]

Epidemiologists have taken different approaches to calculating the optimum MIC cut-point using ROC curves even though they agree that sensitivity and specificity should be valued equally. [7, 11, 12] For example, Farrar *et al*, [12] choose the point closest to the intersection of a -45° tangent line (*i.e.* passing from (0,1) to (1,0), intersecting the ROC curve). As Farrar points out, mathematically this is equivalent to the point at which sensitivity and specificity are closest together. [12] In contrast, an approach by researchers from the EMGO institute, is to choose the point closest to the top-left corner of the ROC curve, which they suggest is found by choosing the cut-point that minimises the sum of $1-$ sensitivity and $1-$ specificity. [7, 13] The Farrar and EMGO approaches can often produce points that differ, and this may lead to different conclusions about responder thresholds and results of responder analyses. We suggest that a third approach may be optimal and preferable. We assert that the commonly practised EMGO method does not in general select the point closest to the top-left corner of the ROC curve, and that the $1-$ sensitivity and $1-$ specificity terms should first be squared before summing, in accordance with Pythagoras' theorem. In this paper, we first present contour diagrams highlighting the nature of the divergence of these three different approaches, and then using both hypothetical and real trial data we explore the extent of the possible divergence between these approaches.

Methods

Equations 1, 2, and 3 show the different estimators being considered. In all cases we choose the point in ROC space that minimises the quantity q, where q is defined differently depending on which estimator is used. In equation 1, the Farrar method is used to estimate q_f by minimising the gap between sensitivity (*sens*) and specificity (*spec*) for all possible cut-points on the instrument, that are contained within ROC space. In equation 2, the EMGO method is used to estimate q_e by minimising the sum of $1-sens$ and $1-spec$. Equation 3 is the method we propose, which estimates q_s by minimising the sums of squares of $1-sens$ and $1-spec$, and which, as follows from simple Euclidean geometry, is closest to the top-left corner (1,0) of ROC space.

$$q_f = \min\{|sens - spec|\} \tag{1}$$

$$q_e = \min\{2 - sens - spec\} \tag{2}$$

$$q_s = \min\{(1 - sens)^2 + (1 - spec)^2\} \tag{3}$$

For each of these estimators, we first constructed diagrams showing contours of the quantities being minimised, to illustrate the how methods favour different areas of ROC space. Using these diagrams we illustrated how the three different methods can give rise to similar or different estimates of MIC in the situation where the number of possible cut-points is large, *i.e.* tending toward a truly continuous nature. Next, this was done using three hypothetical examples of well-behaved ROC curves, by which we mean the second derivative of the curve is always negative (*i.e* the gradient is always decreasing, moving from left to right). And finally, we selected real examples using data from the UK Back pain Exercise And Manipulation (UK BEAM) trial to explore the extent of differences in practice, where ROC curves are seldom well-behaved. [14]

The UK BEAM trial

UK BEAM was a multi-centre and multi-arm randomised controlled trial (RCT) (ISRCTN32683578) in which 1,334 participants with non-specific low back pain were randomised. The trial methods are described in detail elsewhere; [14] but briefly, balanced randomisation was used to allocate participants to one of four different complex interventions: *best care, best care plus an exercise programme, best care plus a spinal manipulation package, or a combined treatment package of best care, exercise, and spinal manipulation*. The primary outcome measure used in the trial was the Roland Morris Disability Questionnaire (RMDQ). The RMDQ is a 24-item validated questionnaire measuring disability and is one of the most commonly cited and used PROMs in studies of back pain. [15–17] Scores on the RMDQ range from 0 to 24, where higher scores indicate greater disability. The score is derived from the sum of the specific item statements that a participant agrees with; an example statement is 'Because of my back, I use a handrail to get upstairs'. A secondary outcome in the trial was the modified von Korff Scales, which measure chronic pain and disability due to LBP. [18] The scores are calculated from responses to three questions each for pain and disability; an example question from this questionnaire is 'In the past 4 weeks how much has your back pain interfered with your daily activities?'. Participants then respond using a numerical rating that is graded from 0 to 10, where zero is none and 10 is '*unable to do any at all*', '*extreme*', or '*as bad as it could be*', depending on the question. The average of the three numerical rating scales for each domain are taken and multiplied by 10. [18] The outcome measures in UK BEAM were completed at baseline, one month, six months, and then finally at one year. At each follow-up time point, participants also completed a health transition question, which was included with the questionnaires. This is a single question assessing global perception of change, by asking participants whether they have experienced improvement or deterioration in their low back pain since beginning treatment. [19] It has seven possible responses ranging from '*completely recovered*' to '*vastly worsened*'.

Use of data from the UK BEAM trial

We dichotomised the health transition scale as suggested by Lauridsen *et al* (and as is typical), categorising patients as improved if they were in the top two categories of recovery, using this as our external criterion for generating ROC curves. [20] We then purposively sampled RMDQ data, and data from the modified von Korff disability scale, across all trial arms, and from the Best Care arm in particular, at three-months, and one-year follow-up time points.

Ethics statement

The trial protocol was approved by the Northern and Yorkshire multi-centre research ethics committee and 41 local research ethics committees. No additional ethics approval was required to reuse the anonymous trial data in this analysis.

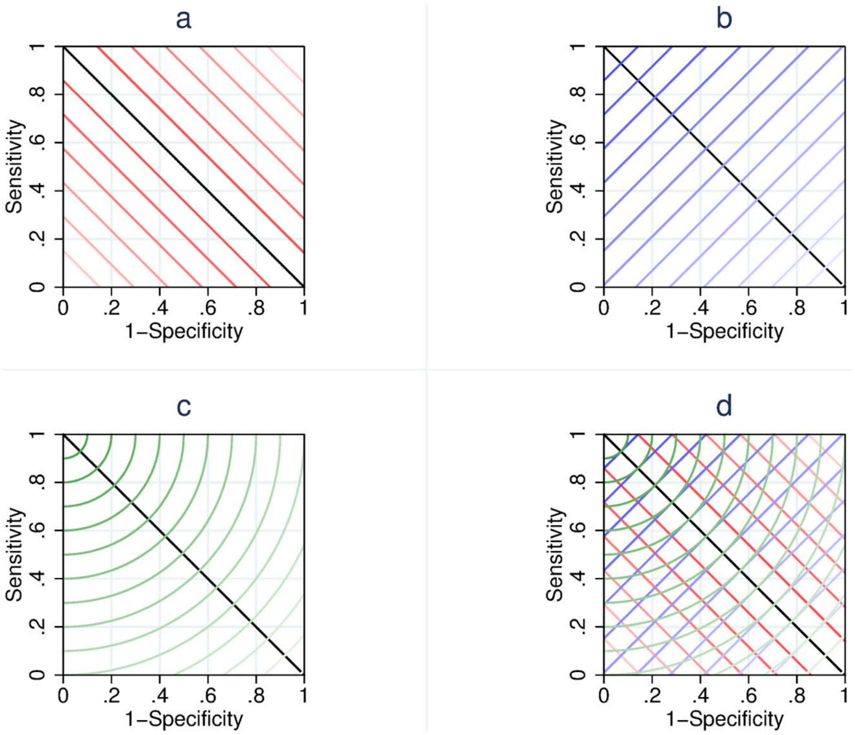

Figure 1. Contour diagrams. The figure shows the contour diagrams we constructed for the Farrar estimator (Figure 1a), the EMGO estimator 2 (Figure 1b), and the sums of squares estimator (Figure 1c), and contours resulting from all three estimators in the same space (Figure 1d).

doi:10.1371/journal.pone.0114468.g001

Results and Discussion

Figure 1 shows the contours of the quantities being minimised by each of the estimators we explored; estimator 1, the Farrar method (Figure 1a); estimator 2, the EMGO method (Figure 1b); estimator 3, the sum of squares method (Figure 1c), and contours resulting from all three estimators in the same space (Figure 1d). In each of these figures the contour lines describe the points within the ROC space that are valued equally by the estimator. The intensity of the line indicates the magnitude of the quantity being minimised, with darker lines indicating preferential selection for MIC.

In Figure 1a, the contours show how the Farrar estimator chooses the cut-point that falls closest to the -45° tangent line (*i.e.* the black line passing from (1,0) to (0,1)). The cut-point falling closest to this line, *i.e* that point which is intersected by the darkest (reddest) contour line, will be chosen using this method. Figure 1b shows how, in contrast, using the the EMGO method the point which is chosen is that point which is intersected by the darkest (bluest) diagonal line that runs orthogonally to the -45° tangent line. Finally, our sums of squares method is illustrated in Figure 1c where the contours are formed of circle arc segments centred around (1,0) and the point which is intersected by the darkest (greenest) contour is selected. We note that, in each case, an infinite number of lines exist and we have rendered only some of these for the sake of clarity.

These contour plots show that the three methods are not equivalent. In particular they demonstrate that different points of the contours drawn by EMGO method (Figure 1b) are not equidistant from the top-left corner of ROC space, but that each possible point on contours described by the sum of squares method is equidistant from the top-left corner. Exactly which point on the ROC curve is chosen depends on the positions of the cut-points in ROC space. This is more clearly demonstrated using our hypothetical plots in Figure 2.

Hypothetical examples

Figure 2 shows three hypothetical examples of well-behaved (*vide supra*) ROC curves when the number of cut-points is large (*i.e.* tending toward a truly continuous nature). Figure 2a shows a case in which all three estimators provide the same cut-point; we note for later discussion that away from the point of convergence, divergence happens quickly between the sums of squares method and the EMGO method. This example is a special case in which the ROC curve is symmetrical. When symmetry is present (and the gradient of the ROC curve monotonically decreases) all three estimators will choose the same point (even when the curve is not continuous). Conversely, if there is asymmetry skewing the curve left or right, as is the case in Figure 2b or 2c, the methods choose different points. The mechanism for this becomes clear if one considers the contours and minimum quantities chosen by each of the three methods, which are also shown in Figure 2. It is quite clear that in Figures 2b and 2c that the points chosen by both the EMGO and Farrar methods are further from the top-left corner of the ROC space than the point chosen by the sum of squares method (in so far as the blue and red points are further from the top left corner than the green contour). In the well-behaved continuous situation, such as is shown here, the point selected by the sum of squares method will always be between the points selected by the other two methods.

Results from UK BEAM trial data

In real applications, ROC curves often consist of a finite number of possible cut-points, rather than being continuous, and are noisy, rather than well behaved. These issues add further complexities to the selection of MIC using the different approaches. Figure 3a, comprises more typical quasi-continuous data from the RMDQ score change in the Best Care arm of the UK BEAM trial at the one-year follow-up time point. The differing points of minimisation are shown to fall on the different contours described by the different estimators. The sum of squares method corresponds to a cut-point of five points (left/green) on the 24-point scale, the Farrar method corresponds to a cut-point of four points (centre/red), and the EMGO method corresponds to a cut-point of three points (right/blue). For this example, we have provided sensitivity and specificity data in Table 1.

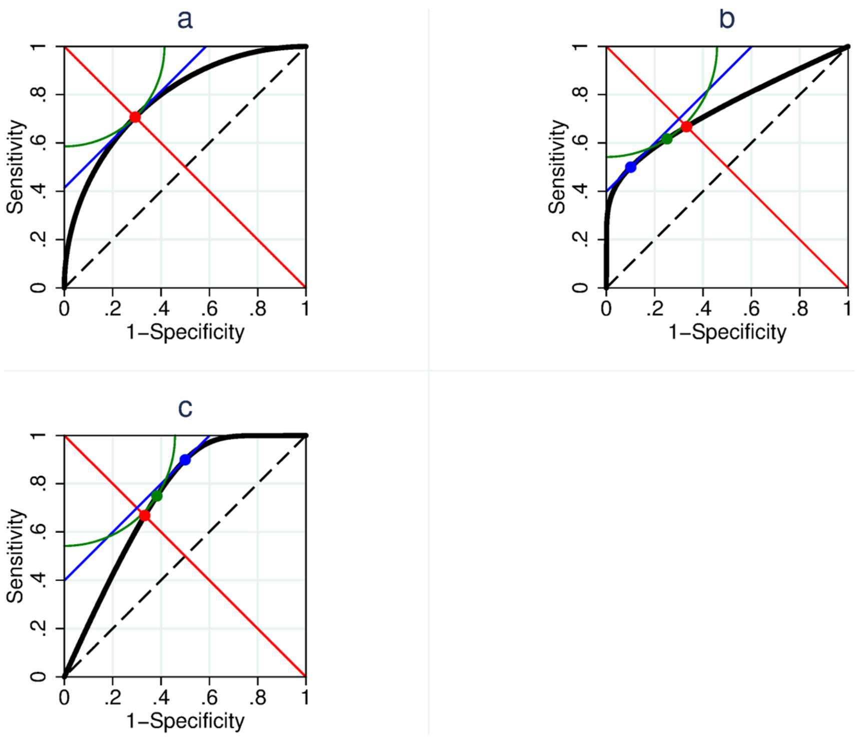

Figure 2. Hypothetical plots. The figure shows three hypothetical examples of well-behaved ROC curves when the number of cut-points is large (*i.e.* tending toward a continuous nature). Panel 2a shows a case of symmetry and a monotonically decreasing gradient where all three estimators provide the same cut-point. Panels 2b and 2c shows how the methods choose different points when there is asymmetry skewing the curve left or right.

doi:10.1371/journal.pone.0114468.g002

In Figure 3b, the plot was formed using modified von Korff disability data from all four trial arms in UK BEAM at three-months. Even though, on the face of it, this curve appears to be reasonably symmetrical, because there are many potential cut-points, even relatively subtle differences are minimised differently by the contours described by the different estimators. In this case, the sum of squares method favours a cut-point of 14.333 (*NB* thirds are common in modified von Korff scores, which is based on an average of three ratings; right/green), the Farrar method a cut-point of 17.667 (centre/red), and the EMGO method a cut-point of 21.000 points (left/blue).

Figure 3c shows modified von Korff disability data from the Best Care arm at three-months. In this example, the sum of squares method favours a cut-point of 14.333 (centre/green), the Farrar method favours a cut-point of 17.667 (left/red), and the EMGO method favours a cut-point of 11.000 (right/blue).

Finally, Figure 3d shows modified von Korff disability data from all four trial arms arm at one-year. Both the sum of squares method and the Farrar method estimates a cut-point of 21.000 (left/green), but the EMGO method favours a cut-point of 14.333 (right/blue).

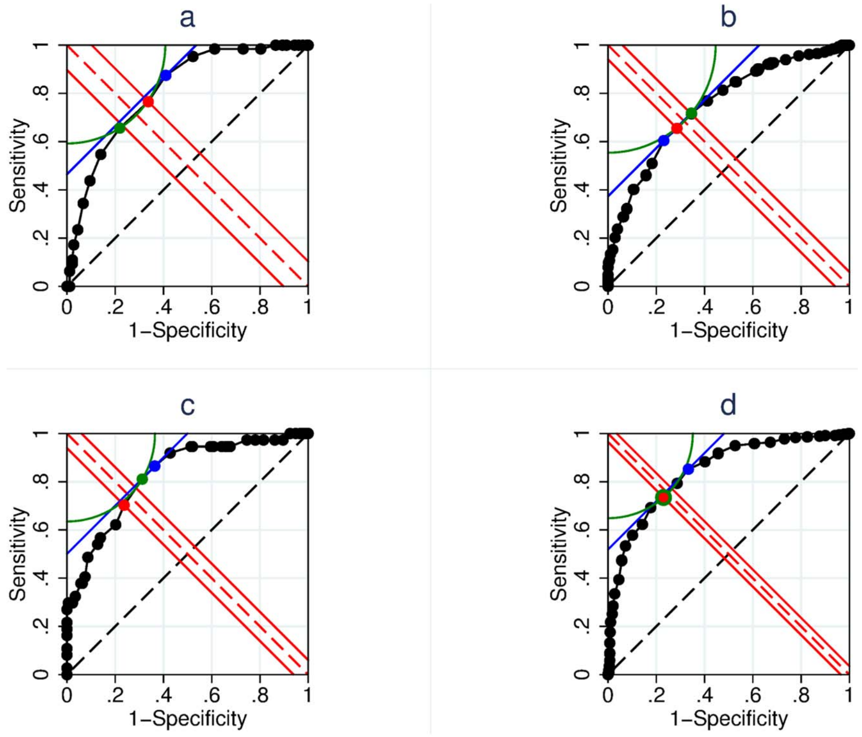

Figure 3. ROC curves constructed using UK BEAM trial data. The figure shows ROC curves constructed using real trial data. Panel 3a shows the Roland Morris Disability Questionnaire changes in the Best Care arm of the UK BEAM trial, at one year. The three estimators choose different points (Farrar=red, EMGO=blue, and sums of squares=green). Panel 3b shows the modified von Korff disability changes across all four arms of the trial, at three months. Panel 3c shows modified von Korff disability data from the Best Care arm at three months. Panel 3d shows modified von Korff disability data from all four trial arms at one-year.

doi:10.1371/journal.pone.0114468.g003

Considerations

The results demonstrate that the three estimators can lead to different cut-points and that, in practice, the differences can be magnitudes of several points. This has implications for interpreting improvement in individual patients, and for interpreting the number or proportion of improved patients in clinical trial arms. [21] In responder analyses, (*i.e.* analyses involving comparisons of the number of individual improvements in trial arms), the choice of method could affect trial results and therefore trial conclusions. We note that when conducting responder analyses the minimal detectable change of the outcome measure must also be examined, but we consider a detailed discussion of this to be outside the scope of this paper and on this point refer the reader to other material. [5, 8]

The Farrar method finds the point at which sensitivity and specificity are closest together. One problem with this approach is that while ROC curves are monotonic in nature, it is possible for the shape of curve to be such that it crosses the 45 degree tangent line at a point where sensitivity and specificity are closest together, but with a combination of sensitivity and specificity that is less appealing

Table 1. Sensitivity and specificity data for the ROC curve in Figure 3a.

Cut-point	Sensitivity (%)	Specificity (%)	|Farrar|x100%	|EMGO|x100%	|Sum of Squares|x100%
−11	100	0	100	100	100
−10	100	0.560	99.44	99.44	99.44
−9	100	2.250	97.75	97.75	97.75
−7	100	3.930	96.07	96.07	96.07
−5	100	5.620	94.38	94.38	94.38
−4	100	8.990	91.01	91.01	91.01
−3	100	10.67	89.33	89.33	89.33
−2	100	13.48	86.52	86.52	86.52
−1	98.44	19.66	78.78	81.90	80.36
0	98.44	26.97	71.47	74.59	73.05
1	98.44	38.76	59.68	62.80	61.26
2	95.31	47.75	47.56	56.94	52.46
3[†]	87.50	58.99	28.51	53.51	42.87
4[‡]	76.56	66.29	10.27	57.15	41.06
5[*]	65.63	78.09	12.46	56.28	40.76
6	54.69	85.96	31.27	59.35	47.44
7	43.75	90.45	46.70	65.80	57.05
8	34.38	93.26	58.88	72.36	65.97
9	23.44	95.51	72.07	81.05	76.69
10	17.19	97.19	80.00	85.62	82.86
11	10.94	97.75	86.81	91.31	89.09
12	9.380	97.75	88.37	92.87	90.65
13	6.250	98.88	92.63	94.87	93.76
15	0	98.88	98.88	101.1	100.0

[†] Cut-point chosen by the EMGO method.
[‡] Cut-point chosen by the Farrar method.
[*] Cut-point chosen by the sum of squares method.

doi:10.1371/journal.pone.0114468.t001

(*i.e.* further from the top-left corner) than another point on the ROC curve. While the EMGO approach avoids this weakness, it can produce cut-points that are less sensible in terms of disproportionately minimising either sensitivity or specificity, when the minimising cut-point is at a lateral extreme of a contour described by its estimator. The EMGO cut-point is sometimes described as the point for which the sum of the percentages of misclassified patients is lowest, [7] and the sum of $1-\mathit{sens}$ and $1-\mathit{spec}$ as the proportion of misclassification. [10] However, we suggest that it is not correct to describe the approach in this way, as the percentage of misclassified patients is dependent on the prevalence of improvement and is calculated as one minus sensitivity, multiplied by the total number of improved patients, plus the specificity multiplied by the total number of non-improved patients.

Our sums of squares approach always selects the cut-point closest to the top-left corner (1,0) of ROC space. Furthermore, even in the situation where a ROC curve is approximately symmetrical and all three methods should choose the same point, the EMGO method will be more sensitive to noise than the either the sum of squares or Farrar method, and as such should be avoided. This can be seen from Figure 2a, where at the point of convergence, the contour described by our method and the contour described by the EMGO method are both adjacent to the ROC curve. Following the path of the contours, one can see that the contour described by our method diverges away from the curve, in contrast to the EMGO method that sits on the curve, in what could be a noisy zone. For this reason the estimator we propose is more efficient.

Finally, we acknowledge that our proposed approach and its utility is obvious from simple Euclidean geometry and as such this presented work should be unnecessary. Nevertheless, since currently non-optimal approaches are in widespread use, we considered that this study was necessary to provide clarification and to highlight the extent to which different approaches can affect results. We urge epidemiologists to more carefully consider cut-point selection when using ROC curves to make decisions about MIC thresholds. To help with this, we provide a Stata module that produces estimates based on all three of the approaches discussed, along with a help file, which is held in the RePEc Statistical Software Components archive, and may be installed at the Stata prompt by typing *ssc install rocmic.* [22]

Acknowledgments

Thanks are due to David Torgerson and Martin Underwood for allowing us to re-use the UK BEAM trial data in this work. Thanks are also due to University of Warwick, Campus Kristiania, and University of Cambridge for jointly covering the associated open access publication charges.

Author Contributions

Conceived and designed the experiments: RF GA. Performed the experiments: RF GA. Analyzed the data: RF GA. Wrote the first draft of the manuscript: RF. Commented on successive drafts: GA. Created the figures: GA. Wrote initial rocmic Stata module and help file: RF. Contributed to modifications of the updated code: GA.

References

1. **Collinson P** (1998) Of bombers, radiologists, and cardiologists: Time to ROC. Heart 80: 215–217.
2. **de Vet HC, Bouter LM, Bezemer PD, Beurskens AJ** (2001) Reproducibility and responsiveness of evaluative outcome measures. Int J Technol Assess Health Care 17: 479–487.
3. **de Vet HCW, Terluin B, Knol DL, Roorda LD, Mokkink LB, et al.** (2010) Three ways to quantify uncertainty in individually applied ''minimally important change'' values. J Clin Epidemiol 63: 37–45.
4. **Jaeschke R, Singer J, Guyatt GH** (1989) Measurement of health status. ascertaining the minimal clinically important difference. Control Clin Trials 10: 407–15.

5. **Froud R, Underwood M, Eldridge S** (2012) Improving the reporting and interpretation of clinical trial outcomes. Br J Gen Pract 62: e729–e731.

6. **Deyo RA, Centor RM** (1986) Assessing the responsiveness of functional scales to clinical change: An analogy to diagnostic test performance. J Chronic Dis 39: 897–906.

7. **de Vet H, Ostelo R, Terwee C, van der Roer N, Knol D, et al.** (2007) Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Quality of life research 16: 131–142.

8. **de Vet H, Terwee C, Ostelo R, Beckerman H, Knol D, et al.** (2006) Minimal changes in health status questionnaires: Distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes 4: 54.

9. **Terwee C** (2003) On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. Quality of life research 12: 349–362.

10. **de Vet H, Terwee C, Mokkink L, Knol D** (2011) Measurement in Medicine. Cambridge University Press.

11. **Ostelo RW, de Vet H** (2005) Clinically important outcomes in low back pain. Best Practice & Research Clinical Rheumatology 19: 593.

12. **Farrar JT, Young JJP, LaMoreaux L, Werth JL, Poole RM** (2001) Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 94: 149–58.

13. **de Vet H, Terluin B, Knol D, Roorda L, Mokkink B, et al.** (2010) There are three different ways to quantify the uncertainty when minimally important change (MIC) values are applied to individual patients. J Clin Epidemiol 63(1): 37–45.

14. **UKBEAM** (2004) United Kingdom back pain exercise and manipulation (UK BEAM) randomised trial: Effectiveness of physical treatments for back pain in primary care. British Medical Journal 329: 1377–1381.

15. **Roland M, Morris R** (1983) A study of the natural history of back pain. part I: development of a reliable and sensitive measure of disability in low-back pain. Spine 8: 141–4.

16. **Muller U, Duetz MS, Roeder C, Greenough CG** (2004) Condition-specific outcome measures for low back pain. Part I: Validation. Eur Spine J 13: 301–13.

17. **Froud R** (2010) Improving interpretation of patient-reported outcomes in low back pain trials. Ph.D. thesis, Queen Mary University of London.

18. **Underwood MR, Barnett AG, Vickers MR** (1999) Evaluation of two time-specific back pain outcome measures. Spine 24: 1104–12.

19. **Beurskens A, de Vet H, Koke A** (1996) Responsiveness of functional status in low back pain: A comparison of different instruments. Pain 65: 71–76.

20. **Lauridsen HH, Hartvigsen J, Korsholm L, Grunnet-Nilsson N, Manniche C** (2007) Choice of external criteria in back pain research: Does it matter? recommendations based on analysis of responsiveness. Pain 131: 112–20.

21. **Froud R, Eldridge S, Lall R, Underwood M** (2009) Estimating number needed to treat from continuous outcomes in randomised controlled trials: Methodological challenges and worked example using data from the UK back pain exercise and manipulation (BEAM) trial. BMC Med Res Meth 9: 35.

22. **Froud R, Abel G** (2014) ROCMIC v2.0.0: Stata module to estimate minimally important change (MIC) thresholds for continuous clinical outcome measures using ROC curves. Statistical Software Components, Boston College Department of Economics. RePEc:boc:bocode:s457052. Available: http://ideas.repec.org/c/boc/bocode/s457052.html. Accessed 2014 November 14.