Quantifying the variability of optimal cutpoints and reference values for diagnostic measures

Gerrit Hirschfeld, Boris Zernikow

Aim: Empirical studies in medicine – and most other fields of study - yield results that are uncertain to a certain degree. Medical research on interventions has made tremendous leaps forward by quantifying and reporting this uncertainty using p-values and confidence intervals. In contrast to this, most diagnostic studies that determine optimal cutpoints or reference values for diagnostic measures ignore that their outcomes, i.e. the specific cutpoints or normal ranges they recommend, are subject to chance variability. **Methods:** In this paper we use a simple simulation approach to quantify the variability of optimal cutpoints for two published studies. The first determined an optimal cutpoint for Becks Depression Inventory (BDI) in adults. The second determined reference values for Quantitative Sensory Testing (QST) in children. **Results:** We find that frequently employed cutpoints to interpret BDI scores and QST results are highly variable. For the BDI we find that replication of this study may identify values between 14 and 21 as optimal cutpoints. The lower cutpoint results in a misclassification of 15% of the healthy adults as depressed, the upper cutpoint results in a misclassification rate of 2%. For the QST we find that the upper end of the normal range HPT varies between 46.9 and 50.2 degrees Celsius. **Conclusions:** Based on our results we argue that researchers should be required to estimate and report the variability of reference values and optimal cutpoints for diagnostic tools. This may improve the harmonization of findings across studies and provides a rationale for planning future studies.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.635v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2014, publ: 25 Nov 2014

Quantifying the variability of optimal cutpoints and reference values for diagnostic measures

G Hirschfeld (PhD)[1,], B Zernikow (MD, PhD) [2]

[1] University of Applied Sciences Osnabrück

[2] German Paediatric Pain Centre, Children's Hospital Datteln, Germany;

And Chair for Children's Pain Therapy and Paediatric Palliative Care, Witten/Herdecke University, Germany;

Corresponding author: Prof. Dr. Gerrit Hirschfeld; Faculty of Business Management and Social Sciences, University of Applied Sciences Osnabrück, Caprivistr. 30 A, 49076 Osnabrück, Germany; eMail: hirschfeld@hs-osnabrueck.de

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.635v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2014, publ: 25 Nov 2014

Introduction

Optimal cutpoints and reference values – that are defined by the limits of the normal range - are used throughout medicine to aid the interpretation of diagnostic test results. These standards are of major importance for clinicians trying to tailor treatments to individual patients, and researchers trying to quantify the rate of patients who suffer from a specific disease or significantly improved after treatment. In line with this, cutpoints play a central role in planning (Sackett & Haynes, 2002) and evaluating diagnostic research (Whiting et al., 2011). However, many researchers who perform studies aimed at empirically determining cutpoints are unaware of the fact that their outcomes are susceptible to chance variability. As a result it is very hard for researchers to integrate conflicting findings about optimal cutpoints so that highly specific cutpoints are being developed that are supposedly applicable only to a very specific group of patients. The domain of pain measurement provides a good example for this. A large number of high-impact articles have determined optimal cutpoints for mild, moderate, and severe pain. Whenever studies identified a different set of cutpoints as optimal in a new group of patients it was assumed that this new group of patients had some peculiar characteristics that necessitate the use of specific cutpoints. However, it was recently shown that the methods used to determine these diverging cutpoints results in considerable chance fluctuations that may be the more likely explanation for the different results across studies and samples (Hirschfeld & Zernikow, 2013b). In the following we give a short overview of the methods that are used to define cutpoints before we use a simulation-based approach to quantify this variability for two diagnostic tools.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.635v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2014, publ: 25 Nov 2014

How are optimal cutpoints developed?

Since there are seldom a-priori ways to define normal vs. abnormal results on a continuous scale, the vast majority of empirical studies use either anchor- or distribution-based methods to define the limits of the normal range (Revicki et al., 2008). Anchor-based methods utilize data from both patients and healthy participants. For many patient reported outcomes (e.g. depression ratings) optimal cutpoints are determined based on a comparison of patients with an established diagnosis and healthy participants. A number of techniques based on statistical tests, e.g. the cutpoint that yields the best prediction of survival as indexed by the lowest p-value (Altman et al., 1994), or Receiver Operating Characteristics (ROC), e.g. the cutpoint that maximizes the sum of sensitivity and specificity. It is well known that a post-hoc choice of optimal cutpoints tends to overestimate the diagnostic utility when applied to other samples (Leeflang et al., 2008; Hirschfeld & do Brasil, 2014; Smith et al., 2014).

Distribution-based methods utilize data from healthy participants and use some statistical parameter (standard error, effect size, etc) to determine the range of "normal" values. For example many laboratory tests (e.g. plasma levels) are interpreted with regard to the normal range of results from healthy participants. The most widely used method to define the upper (lower) limits of this "normal" range is adding (subtracting) 1.96 standard deviations to (from) the mean. This ensures that about 95% of the healthy participants will receive a normal test-result. A recent study found that this procedure yields different reference values for central markers of laboratory medicine (Giannoni et al., 2014). Instead of rejecting the whole idea of

cutpoints all together (Vickers & Lilja, 2009) we believe that many of these problems can be averted once the variability of cutpoints is taken into account.

The aim of the present study is to quantify the level of variability that is inherent in typical studies that determine optimal cutpoints or reference values. Towards this end we use a specific simulation based method to analyze data because this allows to calculate these estimates from data that is typically reported in diagnostic studies. Furthermore we can use the same technique to investigate the impact of sample size on the variability of the cutpoint estimates.

Materials and Methods

Example 1: How variable are ROC-based optimal cutpoints?

In our first example we take a closer look at a published study that uses ROC-methodology to establish cutpoints for the BDI for use in community samples (Arnau et al., 2001). The study collected BDI ratings in 335 primary-care patients, 31 of whom suffered from major depression. The BDI ratings in controls were significantly lower (M = 6.7; SD = 7.1) than in patients (M = 28.0; SD = 9.7). The optimal cutpoint in this study – a value of 18 - was defined as the cutpoint that yielded the highest Youden-index, i.e. sum of specificity and sensitivity.

In order to quantify the variability of this optimal cutpoints, we perform a simulation study using the published summary statistics for the two groups. First, we assume that the distribution of BDI scores in the sample of patients and controls is identical to the distribution in the population (fig. 1A). We then draw a sample of 304 controls (without depression) and 31 patients (with major depression) from the respective distributions. We then determine the cutpoint that optimally differentiates between the two groups according to the Youden-criterion.

In order to inspect the variability of optimal cutpoints, we draw different samples of controls and patients. As shown by two example samples (fig. 1B, fig. 1C) several different cutpoints are selected as optimal. After 10,000 repetitions we can inspect a distribution of optimal cutpoints (fig. 1D). This shows that several different cutpoints between 14 and 21 are identified as optimal in different runs of this simulation. Using the lower of these cutpoints to classify the healthy participants would result in a prevalence for depression of 15.19%, while using the lower would miss-classify only 2.22%. The last panel (fig. 1E) shows the 5% percentile, median, and 95% percentiles of this distribution for various sample sizes (50 and 5,000) but with the same distributional properties and prevalence. This demonstrates that the variability of the optimal cutpoints depends on sample size and even at large sample sizes of 5,000 participants cutpoints between 15 and 18 are identified as optimal.

---

insert figure 1 about here

Example 2: How variable are normal ranges?

In the second example we aim to quantify the variability of a published study that aimed to determine reference values for QST in children (Blankenburg et al., 2010). We will only inspect the data on the heat pain threshold (HPT), but the same principle applies to all other subtests as well. In the study 123 children were tested but due to the stratification by gender and age the cutpoints for HPT were based on 32 boys between 13 and 16 years. The study found that

the mean HPT in this group was 42.60 with a standard deviation of 4.1, resulting in a normal range from 34.6 to 50.6 degrees Celsius.

Again we perform a simulation study based on the given summary data to quantify the variability of these estimates. First, we assume that the sample distribution is identical to the population distribution (fig. 2A). We then draw a sample of 32 participants from this population and calculate the limits of the normal range as the mean of the sample minus/plus 1.96 times the standard deviation. In order to yield a distribution of these limits, we draw different samples from this population. As shown by the two examples different cutpoints emerge in each of these samples (fig. 2B, fig. 2C). After 10,000 replications the distribution of cutpoints (fig. 2D) shows that the lower limits for the HPT varies between 34.1 and 37.4 degrees Celsius and the upper limit for the HPT varies between 46.9 and 50.2 degrees Celsius. Again this strongly depends on sample size (fig. 2E). While even with 500 participants, the variability for the lower limit encompasses a whole degree Celsius (34.1 to 35.1), large samples (N = 5,000) yield very little variability in the estimate for the reference range (34.4 and 34.7).

---

insert figure 2 about here

Discussion

The present paper argues that diagnostic studies need to take into account the variability of optimal cutpoints and reference values. The two examples demonstrate that studies aiming to

determine reference values or optimal cutpoints yield highly variable results. The sample sizes in our examples may seem very low, but a review of diagnostic-accuracy studies found that the median sample size was 118 (Bachmann et al., 2006). Furthermore, we have seen that even considerably larger sample sizes may still yield variable estimates for cutpoints. In what follows, we compare the present to previous approaches, describe the limits of the current approach and provide an outlook about the advantages.

It is important to note that several lines of methodological research have developed methods to address specific problems that result from the variability of cutpoints. First, several authors in the field of ROC methods have described the bias or optimism that is introduced by post-hoc choice of cutpoints (Ewald, 2006; Leeflang et al., 2008; Smith et al., 2014). Second, Crawford (Crawford & Howell, 1998) criticized the practice in neuropsychology to compare patients against small groups of controls without accounting for the variability in the controls. He also showed that this tends to overestimate the abnormality of patients, as indexed by z-scores and developed methods to calculate confidence intervals for patients' z-scores. Third, Altman (Altman et al., 1994) criticized the "minimum p-value approach" to determine optimal cutpoints for prognostic markers in cancer. He was able to show that the minimum-p-value approach grossly overestimates the true diagnostic utility of tools and described methods to control for overestimation. While these approaches all highlight the problems for studies that try to quantify the diagnostic utility of a diagnostic measure when optimal cutpoints are chosen, they offer no specific advice for the growing number of studies that are aimed to determine a specific threshold. Other than our own studies, we are not aware of any primary research studies that report on the variability of the determined cutpoints. We believe that reporting the variability of

the primary outcome of such studies will be greatly improve the ability of the research community to integrate findings and in the long-run harmonize the cutpoints that are being used.

Limitations

We would like to highlight two limitations of the simulation-based approach used to analyze the summary data from published studies. First, it is only viable for some methods to define optimal cutpoints and not for others. While the necessary summary statistics are almost always given for ROC-based methods, most studies do not report the data necessary to replicate the minimum p-value approach. In order to quantify the variability of cutpoints in these cases, a permutation-based approach based on bootstrapping can be used but only if the raw data is available (Hirschfeld et al., 2014). Second, it is unclear how accurate the variability estimates are if the assumptions are not met, i.e. data is not distributed normally. While this assumption is frequently made, it is often not warranted. It may be that this assumption overestimates the true variability of the optimal cutpoint. For example, if controls show positive skew and patients negative skew, the two groups may show a smaller overlap than assumed by a symmetrical normal distribution. As a result we believe that the simulation-based approach taken here should only be used when raw data is not available, e.g. when one performs a meta-analysis of cutpoints. Researchers with access to data on the participant level should use bootstrapping.

Conclusion

To conclude, we have shown that the variability of optimal cutpoints and reference ranges is a relevant problem in so far that the range of values that may be identified as optimal cutpoints. We furthermore believe that this problem is not specific to the two studies scrutinized here but is pertinent many more studies in applied diagnostic research. While this variability

highlights the limitations on the generalizability of individual studies, and the utility of specific cutpoints, it may have positive impact on the field of diagnostic research as a whole. First, it may facilitate the harmonization of research-findings. For some widely used questionnaires and laboratory tests many conflicting cutpoints have been proposed (Giannoni et al., 2014). As we have found in several studies on the classification of pain-intensity, these disparate results may be integrated once the variability of cutpoints is taken into account (Hirschfeld & Zernikow, 2013a,b). Importantly, using the simulation-based approach described here enables a post-hoc quantification of cutpoint variability that only uses published summary statistics. Second, knowledge about the variability of optimal cutpoints provides a rationale for the planning of future studies. As we have demonstrated in the two examples very large 4-digits-samples may be necessary to estimate the cutpoints and reference values with adequate precision. Since the simulation-based method or bootstrapping are easily implemented in freely available statistical software we strongly recommend to make the reporting of the variability of optimal cutpoints mandatory for studies that empirically define reference values and optimal cutpoints.

References

Altman DG, Lausen B, Sauerbrei W, Schumacher M. 1994. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute* 86:829–835.

Arnau RC, Meagher MW, Norris MP, Bramson R. 2001. Psychometric evaluation of the Beck Depression Inventory-II with primary care medical patients. *Health Psychology* 20:112–119.

Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. 2006. Sample sizes of studies on diagnostic accuracy: literature survey. *BMJ (Clinical Research Ed.)* 332:1127–1129.

Blankenburg M, Boekens H, Hechler T, Maier C, Krumova E, Scherens A, Magerl W, Aksu F, Zernikow B. 2010. Reference values for quantitative sensory testing in children and adolescents: developmental and gender differences of somatosensory perception. *Pain* 149:76–88.

Crawford JR, Howell DC. 1998. Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist* 12:482–486.

Ewald B. 2006. Post hoc choice of cut points introduced bias to diagnostic research. *Journal of Clinical Epidemiology* 59:798–801.

Giannoni A, Baruah R, Leong T, Rehman MB, Pastormerlo LE, Harrell FE, Coats AJ, Francis DP. 2014. Do Optimal Prognostic Thresholds in Continuous Physiological Variables Really Exist? *PloS one* 9:e81699.

Hirschfeld G, do Brasil PEAA. 2014. A simulation study into the performance of "optimal" diagnostic thresholds in the population:"Large" effect sizes are not enough. *Journal of Clinical Epidemiology*.

Hirschfeld G, Wager J, Schmidt P, Zernikow B. 2014. Minimally Clinically Significant Differences for Adolescents With Chronic Pain—Variability of ROC-Based Cut Points. *The Journal of Pain* 15:32–39.

Hirschfeld G, Zernikow B. 2013a. Cut points for mild, moderate, and severe pain on the VAS for children and adolescents: What can be learned from 10 million ANOVAs? *PAIN* 154:2626–2632.

Hirschfeld G, Zernikow B. 2013b. Variability of "optimal" cut points for mild, moderate, and severe pain: Neglected problems when comparing groups. *Pain* 154:154–159.

Leeflang MM, Moons KGM, Reitsma J, Zwinderman A. 2008. Bias in Sensitivity and Specificity Caused by Data-Driven Selection of Optimal Cutoff Values: Mechanisms, Magnitude, and Solutions. *Clin Chem* 54:729–737.

Revicki D, Hays RD, Cella D, Sloan J. 2008. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of clinical epidemiology* 61:102–109.

Sackett DL, Haynes RB. 2002. The architecture of diagnostic research. *BMJ : British Medical Journal* 324:539–541.

Smith GCS, Seaman SR, Wood AM, Royston P, White IR. 2014. Correcting for Optimistic Prediction in Small Data Sets. *American Journal of Epidemiology* 180:318–324.

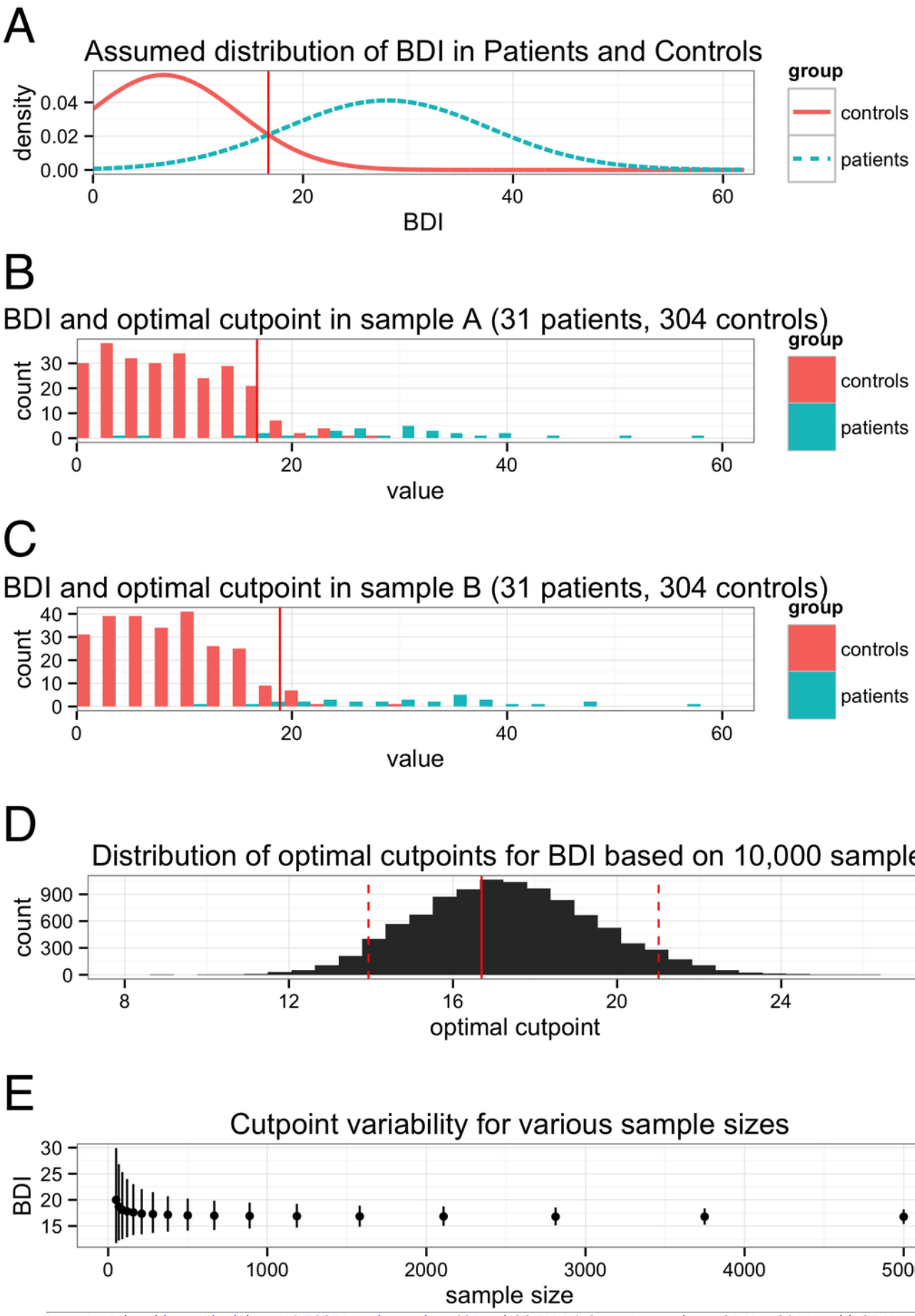
Vickers AJ, Lilja H. 2009. Cutpoints in clinical chemistry: time for fundamental reassessment. *Clinical chemistry* 55:15.

Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MMG, Sterne JAC, Bossuyt PMM. 2011. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine* 155:529–536.

1

Variability of optimal cutpoints for Beck's Depression inventory

(A)Distribution of BDI in patients and controls; (B) BDI in a random sample of patients and controls; (C) BDI in a random sample of patients and controls; (D) Distribution of ROC-based cutpoints in 10,000 samples; (E) Relation between sample size and variability Broken lines represent 5% and 95% quartiles of the cutpoint estimates.



2

Variability of the normal-range for Heat Pain Thresholds

(A) Distribution of HPT in the population; (B) HPT in a random sample of children; (C) HPT in a different random sample of children; (D) Distribution of upper and lower limits defining the normal range for HPT in 10,000 samples. (E) Relation between sample size and variability of limits. Broken lines represent 5% and 95% quartiles of the cutpoint estimates.