ELSEVIER

Journal of Clinical Epidemiology 67 (2014) 449–453

Journal of Clinical Epidemiology

A simulation study into the performance of ‘‘optimal’’ diagnostic thresholds in the population:‘‘Large’’ effect sizes are not enough

Gerrit Hirschfeld[a,b,*], Pedro Emmanuel Alvarenga Americano do Brasil[c]

[a]*German Pediatric Pain Center, Children’s Hospital, Dr.-Friedrich-Steiner Str. 5, 45711 Datteln, Germany*
[b]*Children’s Pain Therapy and Paediatric Palliative Care, Witten/Herdecke University, 45711 Datteln, Germany*
[c]*Instituto de Pesquisa Clínica Evandro Chagas, Fundação Oswaldo Cruz, Av. Brasil 4365, CEP 21040-360, Rio de Janeiro, Brazil*

Accepted 31 July 2013; Published online 18 January 2014

Abstract

Objectives: Many diagnostic studies are aimed at defining ‘‘optimal’’ thresholds. Here, we evaluate the performance of empirically defined optimal thresholds (1) in the sample in which they were defined and (2) in the population from which the sample was drawn.

Study Design and Setting: We simulated test results for 120,000 samples varying the number of people without a disease (*n* between 20 and 500), number of people with a disease (*m* between 20 and 500), the magnitude of the difference between group means [effect size (ES) between 0.5 and 4], and distributions (normal and log-normal). The thresholds associated with the maximal Youden index were defined as optimal. Performance was defined as the percentage of correct classifications in the sample and when applied to the whole population.

Results: At the population level, the thresholds defined for the four ESs (0.5, 0.8, 2, and 4) yielded a median of 59%, 65%, 83%, and 97% correct classifications, respectively. At the sample level, the samples with similar characteristics yielded widely varying estimates of the performance that were systematically higher than at the population level.

Conclusion: Researchers need to be careful defining cut points for mean differences that are traditionally considered ‘‘large’’ (ES = 0.8). The diagnostic utility of optimal thresholds needs to be assessed in prospective studies.

Keywords: Sensitivity and specificity; Optimal threshold; Diagnostic techniques and procedures; Epidemiologic research design; Computer simulation; Youden index

1. Introduction

Medical decision making relies in large part on interpreting diagnostic test results [1]. Because most diagnostic tests measure a continuous outcome, although treatment decisions often take a yes or no form, decision thresholds are used to aid the interpretation of the results and determine the course of action. Given the importance of decision thresholds for medical practice, much of diagnostic research is concerned with identifying ‘‘optimal’’ thresholds for tests that are known to yield different results for participants with (in the following ‘‘patients’’) and without (‘‘controls’’) a target condition. However, even if earlier research has identified large mean differences between patients and controls, this does not guarantee that an optimal threshold can be found [2]. Furthermore, even if an optimal threshold performs well in a specific sample, the performance will be worse when applied to the whole population [3]. Although these issues are in principle well known, it seems that most applied researchers underestimate their relevance for typical studies in which optimal thresholds are determined.

A straightforward method to define optimal thresholds is to systematically apply all possible thresholds to the collected sample and chose the threshold that yields the highest Youden index (sum of sensitivity and specificity minus 1). Thresholds generated using this method are by definition optimal for a given sample. However, when these are used in other samples or the whole population, the accuracy is lower, and alternative thresholds may yield more correct classifications [3,4]. The aim of the present article was to develop realistic expectations about the performance of optimal thresholds at the population level. Toward this end, we evaluated two Youden-based methods in scenarios that are

Conflict of interest: The authors declare that they have no competing interests.

* Corresponding author. Tel.: +49-2363-975-185; fax: +49-2363-975-181.

E-mail address: gerrit.hirschfeld@gmail.com (G. Hirschfeld).

0895-4356/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jclinepi.2013.07.018

What is new?

- It is well known that "optimal" thresholds for diagnostic tests yield overly optimistic estimates for the performance in other samples.
- In this simulation study we assess how well thresholds defined as optimal in one sample perform in the population from which the sample was drawn.
- "Optimal" thresholds for tests that yield "large" mean-differences between patients and controls, result in many misclassifications at the population level.
- The performance of "optimal" thresholds varies widely across samples with similar characteristics and is higher than the performance at the population level.
- Authors, reviewers and editors need to be more skeptical about performances of "optimal" thresholds that are higher than those that can realistically be expected, based on the mean-differences between groups.

typical for diagnostic studies [5]. The first data-driven method uses cut point associated with the largest Youden index. The second robust method derives cut points assuming that data are sampled from normal distributions [6]. Furthermore, we studied the impact of the sample composition (numbers of patients and controls) and mean differences between patients and controls, both for normally distributed and log-normally distributed test results. Although the composition of the sample is within the control of the investigator, the magnitude of the mean differences between the groups can only be estimated from earlier studies. The results of the present simulation study can be used to estimate the feasibility of defining diagnostic thresholds.

2. Methods

We performed a simulation study for which we generated test results for two samples: controls and patients in 120 different scenarios. In all scenarios, healthy controls had a mean of 100 and both groups' test results had a standard deviation (SD) of 10. Scenarios differed in the number of controls (n = 20, 50, 100, 200, or 500) and patients (n = 20, 100, or 500) tested, thereby varying the ratio of patients to controls, the mean for patients' test results (means = 105, 108, 120, 140), and the underlying distribution (normal or log-normal). The means for patients result in effect sizes (ES = $[\text{mean}_{\text{controls}} - \text{mean}_{\text{patients}}]/\text{SD}_{\text{controls}}$) of 0.5, 0.8, 2, and 4, and in the terminology of receiver operating characteristic (ROC) analysis, an area under the curve (AUC) of 0.64, 0.76, 0.92, and 0.99. For each scenario, 1,000 samples were generated.

For each generated sample, we empirically determined the optimal threshold by calculating the Youden index (J) for all possible thresholds and selecting the threshold associated with the maximal J [7]. For each sample, we recorded the selected thresholds and the accuracy of classifications within the sample ([true positives + true negatives]/size of the sample). As the samples were drawn from populations with known distributions, we were also able to calculate the overall accuracy of classifications. As this depends on the prevalence in the population, we assumed that patients and controls are tested equally often. Descriptive statistics [median and interquartile range (IQR) 25–75%] were used to describe the accuracy in the 1,000 samples from similar scenarios. To test whether robust estimation methods yield qualitatively different results, we replicated the same analysis determining cut points by assuming a normal distribution of the patient and control population. We estimated the mean and SD for patients and controls and determined the cut point that gives the largest Youden index for these two distributions. The analysis was performed with R (R Development Core Team; www.r-project.org). The code necessary to reproduce the data generation and analysis is available as an online Appendix at www.jclinepi.com.

3. Results

3.1. Data-driven Youden method

We will first describe the results in the sample (Fig. 1A) before turning to the results at the population level (Fig. 1B). In the sample, the performance of optimal thresholds depended strongly on the magnitude of the mean difference. The medians for the accuracy for the four mean differences (ES = 0.5, 0.8, 2, and 4) were 63%, 68%, 86%, and 99% correct classifications, respectively. With regard to the sample size, larger samples yielded smaller percentages of correct classifications at the sample level. The accuracy varied considerably between the samples from one scenario. For example, when 100 patients and 500 controls with large mean differences (ES = 0.8) are tested, the IQR was still 10 percentage points.

At the population level, the estimates for the accuracy are systematically smaller and much less variable. The median accuracies for the four mean differences were 59%, 65%, 84%, and 98% correct classifications. In contrast to the results at the sample level, larger samples did not yield more accurate classifications at the population level, and the IQR for the accuracy was very small, for example, the IQR for 100 patients and 500 controls was only 1 percentage point. The pattern of results was similar for normal and log-normal data (Fig. S1 at www.jclinepi.com).