Opinion Paper

Per Hyltoft Petersen*

Performance criteria based on true and false classification and clinical outcomes. Influence of analytical performance on diagnostic outcome using a single clinical component

DOI 10.1515/cclm-2014-1138
Received November 19, 2014; accepted March 12, 2015; previously published online April 17, 2015

Abstract

Background: In the general classical model for diagnoses based on a single analytic component, distributions of healthy and diseased are compared and several investigations of varying analytical performance on the percentage of misclassifications have been published. A new concept based on an alternative type of diagnosing, based on sharp decision limits has been introduced in diagnostic guidelines, but only a few publications on investigation of analytical performance have been seen.

Methods: The two diagnostic models (bimodal and unimodal) based on natural logarithmic Gaussian distributions are simulated.

Results: In the bimodal model it is possible to evaluate the influence of prevalence of disease in combination with varying analytical performances. In the unimodal model the prevalence is pre-decided by the chosen decision limit. In this model the influence of analytical performance is investigated for diagnosing diabetes using haemoglobin A1c (HbA1c), and for patients with high and low risk for coronary heart disease defined by serum-cholesterol concentrations.

Conclusions: For HbA1c, the guidelines and recommendations define a maximum inter-laboratory coefficient of variation of 3.5%, but this is in DCCT units (without a true zero-point), so after transformation to IFCC units (which are proportional) it was 5.2%, which allows for analytical bias as high as approximately ±9%. Consequently, analytical quality specifications should be separated as maximum bias and imprecision.

Keywords: analytical bias; analytical imprecision; analytical quality specifications; bimodal models; diagnostic misclassifications; unimodal models; use of prevalence; weighting of diseased group.

Introduction

The basic concept of clinical diagnosing based on a single analytical component was invented by Galen and Gambino [1], who described the concepts of sensitivity and specificity, with related formulas for the two groups involved in the diagnostic process, diseased and healthy, as two separate distributions with some minor overlapping. Later, an alternative concept based on a new type of diagnostic guidelines was introduced for the total distribution of the analyte, where the diagnosis was defined by a decision limit defined by national or international scientific groups, based on the collected evidence on clinical outcome [2].

The description of the first model is a common evaluation of the influence of analytical performance on the traditional diagnostic process using samples from the two distributions of healthy and diseased. This model is general in nature and useful for all components used in the diagnostic process. The second model is defined for specific diseases diagnosed by use of a single analyte and illustrated here for the use of blood-haemoglobin A1c (HbA1c) in the diagnosis of diabetes and of serum-cholesterol in the classification of patients as high and low risk for coronary heart disease. In both cases, the distributions are shown according to assumed or estimated
logarithmic Gaussian (In-Gaussian) distributions and the evaluations of influence of analytical performance are achieved by changing the concentrations with adding varying values of analytical bias, which will have the same effect as moving the cut-off downwards by transferring the results upwards. Likewise, the effect of analytical imprecision is investigated by introducing increasing values of coefficients of variation, \( CV_A \), to the constant biological variation, \( CV_B \), by \((CV_A^2+CV_B^2)^{1/2}\), where \( CV_B \) is either the within-subject biological variation, \( CV_{\rho} \) or total biological variation, \( CV_{\sigma}=CV_{\rho}+CV_{\eta} \) where \( CV_{\eta} \) is the between-subject biological variation. In all these computations the amount of misclassifications are summed up for varying amounts of analytical bias and imprecision.

The purpose of this article is to describe the two theoretical models as tools for evaluation of analytical quality specifications for the two dominating methods of diagnostic performance (bi- and uni-modal models), and in the uni-modal model to demonstrate two different practical examples: one based on the traditional method of estimating a reference interval, and the other of using the decision limit for defining the healthy and diseased among all individuals, as it is the only possibility.

### Bimodal model

This is a bimodal description of the two groups determined as diseased and healthy (non-diseased). The distributions of values from healthy as well as from patients from specified diseases are In-Gaussian as already Gräsbeck et al. demonstrated for vitamin B12 in 1962 [3]. This is often forgotten by scientists who simply describe their distributions as Gaussian so it is necessary to transform these distributions to In-Gaussian (e.g., by the formulas from Fokkema et al. [4]). The method of simulating analytical errors in non-parametric bimodal models was introduced by Groth [5] and compared with a parametric model [6], which was further expanded in, e.g., Nørregaard-Hansen et al. [7].

In the present general example the two distributions are presented as In-Gaussian with mean values as, e.g., \( \mu_{\text{H}}=4.372 \) (healthy \( \sim 79 \) units) and \( \mu_{\text{D}}=4.678 \) (diseased \( \sim 108 \) units) and with the standard deviations \( \sigma_{\text{H}}=0.10 \) and \( \sigma_{\text{D}}=0.15 \), respectively. Sensitivity and specificity as function of In-values of analytical concentration (varying cut-off) can be shown as two S-shaped curves with specificity increasing from low values and sensitivity decreasing from 1.0 (100%) with increasing concentration. The concentration for which sensitivity=specificity is often chosen as the optimum cut-off, here 4.5. If fractions of false positive (FP) and false negative (FN) are used instead, it is easier to introduce the prevalence of disease [7, 8], so the curves for FP will decrease and FN will increase for increasing concentrations and if the prevalence is, e.g., 20% (from 200 healthy and 50 diseased), then the fraction of FP will have a maximum of 0.8 and FN a maximum of 0.2 as calculated from the total sum of the two populations (Figure 1A). In order to demonstrate the importance of prevalence for choosing the cut-off point and the effect on analytical quality specifications, the optimum cut-off from the sensitivity=specificity is chosen so the sum of FP and FN (the combined fraction of misclassifications) will show up as a minimum at a higher concentration than the first cut-off (Figure 1B). If the cut-off of 4.5 is conserved it can be used for evaluation of the effects of analytical bias by defining the 4.5 as zero bias and reverse the figure in a mirror image around the 4.5, so the minimum sum of 6% misclassifications is located at a bias of \(-0.08\) (approx. \(-8\%\)) (Figure 1C).

We can now introduce the effect of analytical imprecision by adding varying values of analytical variation \( \sigma \), where a \( CV_{\sigma}=0.06 \) (6%) is transformed to a ln-standard deviation \(\sigma_{\ln}\equiv0.0599\) (which cannot be distinguished from the 0.06) so the \( \sigma_{\text{combined}}=(\sigma_{\text{A}}^2+\sigma_{\text{I}}^2)^{1/2}=0.116 \) and 0.161 for healthy and diseased, respectively. The effect of \( \sigma_{\text{A}}=0.06 \) results in a new minimum at \(-0.11\) and an increase in misclassifications to 7.7% (Figure 1D). If we assume that a value of 9% misclassifications is acceptable, then a bias between \(-0.02\) and \(-0.18\) is tolerable as long as the imprecision is zero (Figure 1E), but is reduced to the interval from \(-0.05\) to \(-0.17\) when the imprecision is 0.06 (Figure 1F). It is also possible to introduce different weighting factors to FP and FN [9], e.g., for the level of pain or the cost in Euro, but then it is necessary to change the units of the ordinate.

### Unimodal model

This is a unimodal description of a sample from the total population where the diagnostic decision is based on a single concentration value distinguishing the diseased from the non-diseased. This decision limit for a chosen analytical component in diagnosis of a certain disease is decided as the optimum based on observed clinical outcome.

Evaluation of the effects of analytical performance is achieved after the same principles as for bimodal valuation by calculating the percentage of misclassifications due to introduction of errors from analytical bias and imprecision. However, as the diagnosis is based solely on the decision limit of the analytical component there is in principle no FP and no FN in the error-free performance.
Such an investigation was performed for plasma-glucose when the new World Health Organisation (WHO) and American Diabetes Organisation (ADA) recommendations on diagnosis of diabetes mellitus were introduced [10].

Two examples on guideline-driven medical decision limits are discussed; one for the diagnosis of diabetes based on HbA1c and one for cholesterol in the classification of high and low risk for coronary heart disease [11].
The first is based on the assumption of a reference population but this is not possible for cholesterol. In both guidelines, it is recommended to perform a confirmatory second sampling and measurement if the initial sampling reveals a concentration above the decision limit. This process has a considerable influence on the percentage of FP diagnoses which is reduced as the diagnostic decision is based on two independent measurements above the decision limit.

**Diabetes and HbA\textsubscript{1c}**

According to clinical guidelines for diagnosing diabetes the decision limit is 6.5% HbA\textsubscript{1c} in the The Diabetes Control and Complications Trial Research Group (DCCT) units [12], which is the same as 48 mmol/mol in International Federation of Clinical Chemistry (IFCC) units, and the IFCC units are chosen here because they are proportional, whereas there is no true zero concentration in DCCT units [11]. A traditional reference interval for HbA\textsubscript{1c} was accomplished before the guidelines for diagnosis of diabetes was based on HbA\textsubscript{1c} [13]. The reference distribution is estimated to be ln-Gaussian with a mean of 3.63 (38 mmol/mol) and standard deviation 0.086 (CV\textsubscript{I}) corresponding to a reference interval from 31.9 to 44.7 mmol/mol [13], and CV\textsubscript{I} is 1.94% approximately 0.0194 in natural logarithms [14]. The cut-off of 48 mmol/mol corresponds to the natural logarithm 3.87.

The standard deviation of ln-set-points is (0.086\textsuperscript{2}−0.0194\textsuperscript{2})\textsuperscript{½}=0.084 (CV\textsubscript{I}) and the distribution of set-points can be plotted together with the distribution of the within-subject values for a person (CV\textsubscript{S}=0.0194) with mean=3.87 (48 mmol/mol) (Figure 2A). This within-subject variation also indicates the probability of exceeding the cut-off, shown for CV\textsubscript{A}=0.0 and 5.0% (Figure 2B). The probability is 50% at the cut-off (48 mmol/mol), and if we define that also the second independent results should exceed the cut-off [11], then the probability reduces to 0.5\textsuperscript{2}=0.25. For each small interval in the set-point distribution the chance of exceeding the cut-off can be estimated by multiplication with the probability, so the FP of each set-point can be illustrated in a frequency distribution showing the origin concentrations of the FP values, as described for the two situations of CV\textsubscript{A}=0.0 and 5.0% in Figure 2C. By accumulation over all set-points the total percentage of FP from the reference population is estimated. If increasing values of analytical imprecision is added the frequency curve will be broader and the small set-point intervals will get higher probability for exceeding the cut-off, so the cumulated percentage of FP will increase [11]. The effect of analytical bias is different, as a positive bias will increase the measured concentrations of HbA\textsubscript{1c} resulting in movement to higher concentrations of the set-points, whereas the defined cut-off is unchanged, so thereby the percentage FP will increase [11]. This effect gives the same result as moving the cut-off to the left, which is illustrated in Figure 2D.

The analytical imprecision has a considerable effect on FP when only one sampling and measurement is performed, as the value increases from 0.38% to 1.04% for imprecision=5% (and bias=0.0%) whereas a bias-interval from −5% to +5% changes the percentage from 0.06% to 1.83% (with CV\textsubscript{A}=0.0%). The effect of two samples reduces the FP to 0.15% in the error-free situation and the effect of bias from −5% to +5% to 0.02%−0.86% whereas the imprecision of 5% reduces it considerably to 0.19% by the confirmatory second measurement. In general the effect of imprecision is negligible for two samplings with measurements and the effect of bias more than halved, but approximate with a doubling of FP for each +2% bias [11]. The very low percentage of FP results is due to the use of a traditional reference interval in the calculations with painstaking selection of the healthy reference population [13] whereas a new investigation of a general population revealed about 2% undiagnosed diabetics [15]. The use of a reference interval to estimate FP results in diagnosis based on decision point demonstrates the difference between the two models. For the model of reference intervals, all the individuals with measured values above the cut-off are considered FP, whereas, in the decision value model, only individuals from the set-point distribution below the decision limit is considered FP.

The recommendation for analytical performance of HbA\textsubscript{1c} is 2% for intra-laboratory CV and 3.5% for inter-laboratory CV [16]. Now this is given in NGSP/DCCT units which corresponds to 5.2% around 48 mmol/mol in IFCC units, so in reality a bias of up to ±9% is acceptable from this recommendation, which means that the FP could vary from 0.0% to 2.8% when two measurements above cut-off is recommended.

**Cholesterol in the classification of high and low risk for coronary heart disease**

For cholesterol there is no possibility for estimation of a traditional reference interval, so the distribution of all cholesterol results in a sample of measured serum-cholesterol values in a large clinical chemical hospital...
laboratory from Klee et al. [17] is used to estimate the total population and transform the units to natural logarithms of mmol/L, giving the total population with a ln-mean of 1.660 (approx. 5.25 mmol/L) and standard deviation 0.1391 (CVT). The cut-off is 1.826 (6.21 mmol/L), so the population is divided at this concentration value, where the lower part represents the low risk group and the upper part is removed (Figure 3).

The evaluation of influence of bias and imprecision in percentage FP is performed after the same lines as for HbA1c and CVI = 6% in accordance with Ricós et al. [18].

The results show a reduction of FP from 4.47% to 1.21% by measuring two independent samples and the effect of imprecision is practically reduced to zero, whereas the effect of ±5% bias gives FP from 0.13% to 5.30% [11]. The consequence of the Clinical Laboratory Improvement Amendments (CLIA) 88 recommendation accepting errors up to ±10% can result in FP from 0% to 13.61%.

This cholesterol example is based on a distribution of measurements performed in a routine laboratory and not on a thorough investigation designed for the purpose [17], so the evaluation of analytical quality specifications is theoretical, in order to demonstrate how the model and formula can be used. Further, an investigation of the influence of analytical quality on the percentage of FN
decisions might be performed on a model comparable to the model used for the effect on FP. Influence of imprecision and possible bias in the distributions of the basic distributions has not been considered, but may be eliminated before new calculations of same type.

Discussion

The two diagnostic models described, are based on different views on health and disease, as the traditional bimodal model assumes an overlap between healthy and diseased individuals, whereas the unimodal model makes a sharp distinction on a single concentration in the total distribution of all individuals.

For the bimodal model it is possible to evaluate the influence of prevalence and even to introduce different weighting factors to FP and FN [9], e.g., for the level of pain or the cost in Euro, but then it is necessary to change the units of the ordinate. For the unimodal model the recommended cut-off defines the prevalence, and will vary if the decision limit is changed. A strong element for the unimodal model is the recommendation of two independent samplings and measurements, which reduces the influence of analytical imprecision considerably and also decreases the effect of analytical bias on the percentage of FP.

The recommendations for analytical quality in measurements of components according to the unimodal model are much too vague as they do not distinguish between analytical bias and imprecision, but define specifications for intra- and inter-laboratory CV-values [16], which may allow for considerable sizes of analytical bias which is of vital importance for the percentage of FP, so the specifications should be given for analytical bias and imprecision separately, as also described in [2] and [11].

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Financial support: None declared.

Employment or leadership: None declared.

Honorarium: None declared.

Competing interests: The funding organisation(s) played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

References

8. Hyltoft Petersen P, Rosleff F, Rasmussen J, Hobolth N. Studies on the required analytical quality of TSH measurements in...