A Gray Zone Assigned to Inconclusive Results of Quantitative Diagnostic Tests: Application to the Use of Brain Natriuretic Peptide for Diagnosis of Heart Failure in Acute Dyspneic Patients

Joël Coste,1* Patrick Jourdain,2 and Jacques Pouchot1,3

Background: Most quantitative diagnostic tests do not perfectly differentiate between persons with and without a given disease. We present a simple method to construct a 3-zone partition for quantitative tests results, including positive and negative zones and a gray zone between, and we describe its use in the diagnosis of heart failure by brain natriuretic peptide (BNP) measurement in acute dyspneic patients.

Methods: We conducted a prospective cohort study of 699 consecutive patients with acute dyspnea who were treated at the emergency department of 3 participating hospitals. Heart failure (acute or decompensated) was assessed independently at discharge by cardiologists blind to the results of BNP measurements.

Results: The discriminatory performance of BNP was insufficient to provide a single cutoff value that could be used to correctly diagnose heart failure in clinical practice. Also, the discriminatory performance differed between patients with and without a history of chronic heart failure. The gray zone of inconclusive results was 167–472 ng/L for those without and 0–334 ng/L for those with such a history. Diagnosis of the current episode of heart failure by BNP results and history of heart failure was not enhanced by data from any other sources, including electrocardiography.

Conclusions: The gray zone approach applied to the diagnosis of heart failure by BNP might allow sensible cutoff values to be determined for clinical practice according to relevant subgroups of patients. The gray zone approach might be usefully applied to many other quantitative tests and clinical diagnostic or screening problems.

© 2006 American Association for Clinical Chemistry

Quantitative diagnostic tests usually require the determination and use of a cutoff value to obtain a binary response to differentiate diseased from nondiseased individuals. Therefore, methods for optimizing the selection of cutoff values have been developed and refined (1–3). Skepticism exists toward this approach, however, as crystallized by Feinstein, who showed that binary models are inappropriate to clinical reality (4) and proposed a 3-zone partition, including a middle “uncertain, indeterminate, inconclusive” zone of intermediate values, corresponding to a “prediction not precise enough for diagnostic decision”. We extended this approach, initially developed for categorical and ordinal tests, to quantitative diagnostic and screening tests (5). For a given clinical context, this approach allows a 3-zone partition of test results including the gray zone. It requires only the knowledge of the distributions of test values in diseased and nondiseased individuals and the specification of (a) a reasonable value (or range of values) of pretest probability of the suspected disease (diagnostic hypothesis) in the clinical setting or practice where the application of the test is considered, and (b) values of “working diagnosis” posttest probabilities for confirming (“ruling in”) and excluding (“ruling out”) the diagnostic hypothesis in this practice. “Desirable likelihood ratios” (test properties that relate pretest and posttest probabilities) are easily deduced by use of the Bayes theorem and can be used in a straightforward way to determine the upper and lower bounds of the gray zone. The upper bound of the gray zone is defined by the
value of the test associated with the minimal positive likelihood ratio (LR+) that allows ruling in the diagnostic hypothesis, and the lower bound by the maximal negative likelihood ratio (LR−) that allows ruling out the diagnostic hypothesis (Fig. 1). The likelihood ratio measures the extent to which the odds of a disease are altered according to a result of the test: posttest odds = pretest odds × likelihood ratio; and the odds of having the disease are given by probability/(1 − probability). In the case of a binary (or dichotomized) test, there are 2 likelihood ratios, one corresponding to a negative and the other to a positive result (1). The likelihood ratio for a positive test (LR+) is sensitivity/(1 − specificity), and the likelihood ratio for a negative test (LR−) is (1 − sensitivity)/specificity.

We tested the value of the gray zone approach to the diagnosis of heart failure in dyspneic patients by use of a brain natriuretic peptide (BNP) assay. BNP is widely used by clinicians to confirm or exclude acute or decompensated heart failure, but international guidelines do not propose a clear strategy for using this marker. Indeed, controversy persists regarding both the cutoff(s) to be used for the exclusion and/or the confirmation of heart failure and the subgroups of patients to whom these cutoffs apply (6–9). The gray zone approach may help

---

*Nonstandard abbreviations: BNP, brain natriuretic peptide; LR, likelihood ratio; CI, confidence interval; ECG, electrocardiogram.
overcome the current diagnostic standstill by identifying BNP cutoffs for the diagnosis of heart failure.

Materials and Methods

Participants and Clinical and Laboratory Data

Study participants were a total of 699 consecutive patients with acute dyspnea who were referred to the emergency department in 1 of 3 participating hospitals (CHR René Dubos, Pontoise; CHU Beaumont, Clichy; and CHU Ambroise Paré, Boulogne in the Paris suburbs) from September 23, 2003 to February 23, 2005. All patients underwent standardized clinical examination on admission, and blood samples were collected before any therapeutic intervention. Physical examination looked for jugular venous distension, lower limb edema, and pulmonary rales. After patients rested for 15 min, blood was obtained via a peripheral venous catheter as part of the usual laboratory work-up. Study samples were collected into vials containing potassium EDTA acid (1 g/L). Aliquots (2.5 mL) of whole blood were immediately assayed for BNP with the triage BNP test (Biosite Diagnostics Inc.). The mean confidence limit of the analytical sensitivity was <5 ng/L [95% confidence interval (CI), 0.2–4.8 ng/L] (10). Chest x-ray and an electrocardiogram (ECG) were performed in all patients, and 92% of the patients underwent echocardiography during hospitalization. No study patients had severe renal dysfunction. Clinicians caring for the patients were blind to BNP results throughout the diagnostic and therapeutic process. BNP diagnostic results were compared with a gold standard, the final patient diagnosis at discharge, which was consensually established by 2 independent cardiologists who used all usual techniques (except BNP results) to rule in or rule out the diagnosis of heart failure, as recommended by the European Society of Cardiology guidelines (11).

Statistical Analysis

We used logistic regression modeling according to Coughlin et al. (12) to analyze heterogeneity between subgroup diagnostic differentiation based on BNP. Briefly, a logistic model was constructed in which the dependent variable was the presence or absence of heart failure as defined by the gold standard, and the BNP value was included as an explanatory variable, along with variables used to define the subgroups of interest and terms of interaction between BNP and subgroup variables. Significant interaction (P <0.05) indicated subgroups that should be differentiated on the basis of BNP assay results.

Gray Zone Construction

The gray zone was constructed to ensure that posttest probabilities >0.95 and <0.05 would confirm and exclude the presence of heart failure, respectively [these values were suggested by a French group of experts on heart failure: the Working Group on Heart Failure of the French Society of Cardiology (13)]. Using the pretest probability estimates of heart failure for the studied sample, we computed the minimal positive likelihood ratio (i.e., the ratio of the minimum acceptable posttest odds of disease given a positive test result divided by the pretest odds of disease) and the maximal negative likelihood ratio (i.e., the ratio of the maximum acceptable posttest odds of disease given a negative test result divided by the pretest odds of disease) ensuring critical posttest probabilities of 0.95 and 0.05, respectively. These values were used to identify the cutoff points delimiting the gray zones g\text{sup} associated with the minimal value of LR+, and g\text{low}, associated with the maximal value of LR—. Because LRs generally do not increase or decrease monotonically with the BNP value, g\text{sup} was defined as the smallest observed BNP value associated with a positive likelihood ratio greater than the minimal positive likelihood ratio, and g\text{low} as the largest observed BNP value associated with a negative likelihood ratio lower than the minimal positive likelihood ratio. Nonparametric bootstrap 95% CIs were estimated (by the percentile method) for g\text{sup} and g\text{low}, defined as described above, and for the proportion of results that fell within the gray zone (5000 replications were performed). To evaluate the stability of the gray zone limits, we performed 2-way sensitivity analysis, simultaneous varying of pretest (disease prevalence) and posttest probabilities, and analysis of the effect on LRs.

Results

Main characteristics of patients (all but 11 were Caucasians) are shown in Table 1. A final diagnosis of acute or decompensated heart failure was established for 60% of the patients. The distributions of BNP concentrations according to the final diagnosis are shown in Fig. 2. Age, sex, individual clinical manifestations, abnormal ECG, and the number of heart failure criteria (the sum of clinical manifestations and ECG result) were not found to be independently associated with a final diagnosis of heart failure (all univariate relationships disappeared after adjusting for BNP), nor were they found to interact significantly with BNP in the logistic model. Conversely, a history of chronic heart failure was an independent predictor of heart failure and interacted (negatively) with BNP: a history of chronic heart failure increased the risk of heart failure diagnosis almost 10-fold but significantly decreased the risk of heart failure associated with any given BNP concentration (Table 2). Second-order polynomial effects of BNP were introduced into the logistic regression model but were found to be nonsignificant, suggesting that the relationship between BNP and the risk of heart failure was linear (on a logistic scale) in both subgroups. Note that the goodness-of-fit of the final logistic model was excellent (Table 2), suggesting that no important predictor had been ignored.

Therefore, the gray zones were constructed separately for the patients without (75%) and with a history of chronic heart failure (Figs. 3 and 4). For patients without
We propose a simple method to construct a 3-zone partition for quantitative tests results, including a gray zone between positive and negative conclusions about the condition tested. This method can be used both to display graphically the discriminatory performance of 1 (or several) quantitative test(s) in a variety of contexts and also to scrutinize its (their) components of variability (5). The major value of this method is that it avoids the binary constraint of a “black or white” decision, which is often inappropriate to clinical or screening practice, in which existing quantitative tests almost never perfectly differentiate between individuals with and without the target condition, and unique cutoff values cannot be found that allow the tests results to simultaneously confirm and exclude diagnostic hypotheses. Note that test results falling in the gray zone are not uninformative and may lead to a search for further evidence, which can transform the test result from decisive (sufficient to initiate a therapy) to contributory (requiring further testing to reach a diagnosis).

In practice, the estimation of the gray zone limits (and proportion of cases within it) is straightforward, simply requiring computing LR+ and LR− for each value of the test. The precision of the estimates can also easily be documented with the bootstrap method. However, in the “good” (but rare) case in which distributions of test values for diseased and nondiseased individuals hardly overlap, particular attention should be paid to the sample sizes (diseased and nondiseased groups) to make sure that enough observations are obtained in the “tails” of (overlapping) distributions defining the gray zone. Otherwise, the study sampling method is nonspecific and the recommendations of Reid et al. (14) must be followed to avoid the various biases (spectrum bias, verification bias, review bias) affecting the representativeness of diseased and nondiseased individuals and the evaluation of the performance of diagnostic tests.

Because the gray zone is constructed in terms of desirable LR$s, it may change according to the clinical or...
screening context, which should therefore be analyzed beforehand. A simple 2-way sensitivity analysis, with simultaneous varying of pre- and posttest probabilities according to reasonable scenarios or hypotheses and study of the effect on LRs and the resulting limits of the gray zone, could be useful to test the robustness of those limits. The rule of thumb proposed by the Evidence-Based Medicine group, (15), i.e., to consider LR+/H1 > 10 and LR−/H1 < 0.1 as indicating conclusive tests, could also be used as a first approximation.

The application of the gray-zone approach to the diagnosis of heart failure allowed some clarification of the controversial issue of BNP cutoffs. Although this test clearly outperformed all clinical and ECG criteria, its discriminatory performance was not sufficient to provide a single cutoff value that would correctly identify heart failure in clinical practice. Therefore 2 cutoffs appear necessary, one to be used to rule out the diagnostic hypothesis (e.g., that could be used in an emergency department to decide not to admit a patient with mild dyspnea) and another to rule in the diagnosis and undertake a specific treatment (16). Patients in the gray zone would require further urgent exploration, typically echocardiography, which could therefore be limited to these patients. Our study also showed that a history of chronic

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>SE of the coefficient</th>
<th>Adjusted odds ratio = exp (coefficient)</th>
<th>95% CI of the odds ratio</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNP (ng/L)</td>
<td>0.010</td>
<td>0.001</td>
<td>1.010</td>
<td>1.008–1.011</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>History of chronic heart failure</td>
<td>2.288</td>
<td>0.536</td>
<td>9.85</td>
<td>3.45–28.15</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>BNP × history of chronic heart failure</td>
<td>−0.005</td>
<td>0.001</td>
<td>0.995</td>
<td>0.993–0.998</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Hosmer and Lemeshow goodness-of-fit statistics: $\chi^2 = 9.17$ (8 d.f.), $P = 0.33$, Harrell's c = 0.952.

Fig. 3. Distributions (histograms) and gray zones of BNP concentrations (ng/L) for the diagnosis of heart failure patients.
For patients without a history of chronic heart failure, the gray zone is constructed using LR+ = 18.25 and LR− = 0.05 (see text). □, no heart failure; ■, heart failure.

Fig. 4. Distributions (histograms) and gray zones of BNP concentrations (ng/L) for the diagnosis of heart failure patients.
For patients with a history of chronic heart failure, the gray zone is constructed using LR+ = 3.35 and LR− = 0.01 (see text). □, no heart failure; ■, heart failure.
heart failure decreased the probability of heart failure associated with any given BNP concentration (compared to those without a such a history) and required the use of very different cutoffs. This result, which has been reported recently (17), suggests that BNP cannot be used alone to rule out the diagnosis of heart failure in patients with a history of heart attacks and therefore invalidates previous attempts to quantify heart failure probability as a function of BNP value alone (18). It is also noteworthy that we found no differences according to age or sex; this finding contrasts with recent suggestions by Maisel et al. (19), whose study results might have been confounded by the history of chronic heart failure, which was not addressed. Our results, together with those of Möckel et al. (20), who used a different approach (CART regression), again illustrate that diagnostic tests do not perform equally in all situations; appropriate interpretation depends not only on disease severity, but also on history and various other characteristics of patients (21).

Our study sample was typical of emergency department recruitment here in France; the problem of identifying heart failure was very satisfactorily modeled, and the gray zone limits proved to be relatively insensitive both to disease prevalence and posttest probability requisites. Nevertheless, the cutoffs for BNP values determined in this study require further validation before being firmly advocated (and tested for clinical usefulness and cost-effectiveness) in preference to classical 1-cutoff binary decisions in various settings and populations. In particular, the apparent conclusion that the gray zone should be constructed only for patients with or without a history of chronic heart failure because of the large effect of such a history on BNP diagnostic performance (overwhelming the effects of most previously reported influences on BNP values, including age, sex, ECG, and clinical manifestations) requires confirmation in other settings. Further (and larger) studies are also needed to enhance the accuracy of estimates of the gray zone limits; here our estimates were only approximate, especially in the subgroup of patients with a history of chronic heart failure. Note that the gray zone limits may possibly change according to the method used to measure BNP, so that the limits calculated in this study (using the triage system) cannot be adopted without verification by laboratories using other systems [e.g., microparticle enzyme immunoassay, IRMA (Shionogi & Co), and ADVIA (Bayer)]. The analytical and clinical performance of N-terminal prohormone BNP differs from that of BNP (8, 22–24), so its use may also lead to different gray zone values (and even possibly a different subgroup categorization).

The gray zone approach does not solve all the problems associated with decisions based on diagnostic and screening quantitative tests results, and we do not suggest that it replaces the use of LR and Bayesian computation or derived nomograms. In particular, quantitative test users may find it necessary to use LR in various instances to take into consideration precisely “how positive” and “how negative” the results are, avoiding a potentially restrictive categorization (25). In many situations, however, the direct handling of LR and probabilities may be unfeasible or even risky (26–29), and the gray zone approach may help decision-makers in practice, as do the “SpPIn” and “SnNOut” rules (1) when they are wisely and critically applied (30). The gray zone approach and its graphical representations may also provide useful support during the development, evaluation, and publication of the performances of quantitative tests. These representations may facilitate early identification of suitable cutoff points and prevent the development of controversies regarding their use, as in the case of BNP discussed in this study.

We are grateful to all of the cardiologists and physicians in the participating hospitals for their cooperation and help. We also acknowledge the contributions of the anonymous reviewers.

References


