Braz J Infect Dis. 30 (2026) 104605

Contents lists available at ScienceDirect

Brazilian Journal of Infectious Diseases

journal homepage: www.elsevier.com/locate/bjid

Brief Communication

Endocarditis risk stratification with scores: what about reproducibility? The case of NOVA and DENOVA scores for *Enterococcus faecalis* bacteremia

Pierre Danneels [a,*], Floris Chabrun [b], Lucia Grandière-Pérez [c], Ali Touré [d], Vincent Dubée [a]

[a] University Hospital, Infectious Diseases and Tropical Medicine Department, Angers, France
[b] University Hospital, Biochemistry and Molecular Biology Laboratory, Angers, France
[c] General Hospital, Infectious Diseases and Tropical Medicine Department, Le Mans, France
[d] Pole Santé Sud, Emergency Department, Le Mans, France

ARTICLE INFO

Editor: Dr. Luciano Goldani.

Keywords:
E. faecalis
Endocarditis
Bacteremia
Reproducibility

ABSTRACT

Objectives: NOVA and DENOVA scores were developed to guide endocarditis risk assessment in *Enterococcus faecalis* Bacteremia (EfB), but some of their criteria may be open to interpretation. We aimed to evaluate their inter-rater reliability and feasibility.
Methods: Thirty-two physicians from four specialties involved in the management of endocarditis independently evaluated eight EfB patient records using the NOVA and DENOVA scores. Each score was applied eight times per case. Inter-rater reliability was measured with Krippendorff's alpha, and agreement with Fleiss' Kappa. Completion time was also recorded.
Results: No record received identical scores from all raters. NOVA showed low inter-rater reliability ($\alpha = 0.37$), while DENOVA reached moderate levels ($\alpha = 0.49$). High agreement was found for extreme score values, but agreement dropped markedly for intermediate values. Among score items, Auscultation of murmur (A) and Valve disease (V) had the highest reliability ($\alpha > 0.8$), while Duration of symptoms (D) and Origin of infection (O) had the lowest ($\alpha < 0.2$). Completion times were similar between NOVA and DENOVA but varied by specialty.
Conclusion: The reproducibility of these scores is limited, especially near critical thresholds, highlighting the need to complement scoring tools with clinical judgment in EfB.

Introduction

Gram-positive cocci bacteremia is frequently associated with endocarditis: 5 %–20 % of cases for *Staphylococcus aureus*; 8 %–26 % for *Enterococcus faecalis*; and a significant variability ranging from 1 % to 48 % depending on the *Streptococcus* spp. species.[1,2] Hence, there is a need for reliable and simple tools to rapidly identify patients with bacteremia at higher risk of endocarditis.

In their narrative review, Rasmussen et al. present an algorithm based on an endocarditis risk stratification system for investigating patients with cocci gram-positive bacteremia.[1] This risk assessment uses scores developed to help clinicians target patients requiring echocardiography: the VIRSTA, PREDICT and POSITIVE scores for *S. aureus*; NOVA and DENOVA scores for *E. faecalis*; or the HANDOC score for non-β-hemolytic *Streptococcus* spp.[3,4] These scores have been designed and optimized to maximize their diagnostic performance. In the case of *E. faecalis* bacteremia, the DENOVA score shows a better performance (Sensitivity 95 %–100 % and Specificity 84 %–85 % for a threshold ≥ 3) than the NOVA score (Sensitivity 97 %–99 % and Specificity 23 %–56 % for a threshold ≥ 4).[4,5]

However, these scores were established from data collected retrospectively and some of their criteria may be open to interpretation. Reproducibility and feasibility studies, which have not yet been carried out, are therefore needed to address these limitations.

The objective of our study is therefore to investigate the reproducibility and feasibility of the NOVA and DENOVA scores.

Methods

Definitions

The NOVA score includes 4 criteria composing its acronym: N (5-points), O (4-points), V (2-points), and A (1-point). The DENOVA score adds two new criteria (D and E), each criterion is worth only 1-point.

* Corresponding author.
E-mail address: pierre.danneels@chu-angers.fr (P. Danneels).

https://doi.org/10.1016/j.bjid.2025.104605
Received 17 June 2025; Accepted 29 November 2025
Available online 12 December 2025

The criteria used were defined as follows: for D, Duration of symptoms consistent with endocarditis ≥ 7-days before the first positive blood culture; for E, clinical examination or imaging result compatible with septic Embolization; for N: Number of positive blood cultures for *E. faecalis* ≥ 2; for O: Absence of focal infection susceptible to be the Origin of bacteremia; V: Heart Valve disease predisposing to a moderate or high risk of infective endocarditis including native valve disease, previous endocarditis, or the presence of a valve prosthesis; A: Auscultation of a heart murmur.

Score feasibility and reproducibility

In one of the centers included in the DENOVA validation study, 27 patients had *E. faecalis* bacteremia in 2019.[5] Eight of them (P1–P8) were intentionally selected to cover a gradient of endocarditis risk (DENOVA scores 0–5). According to the ESC 2023 modified Duke criteria, two patients (P1 and P6) had definite and two (P3 and P4) possible endocarditis.[6]

Thirty-two clinicians (8 infectious diseases physicians, 8 cardiologists, 8 internal medicine physicians, and 8 medical residents) participated as raters. They had full access to the retrospective medical records, including clinical notes, microbiological results, and imaging reports. Each rater assessed four cases: two using NOVA and two using DENOVA. Because NOVA is fully embedded within DENOVA, the same items were scored twice, resulting in 64 evaluations for DENOVA and 128 for NOVA, with DENOVA and NOVA scores determined 8- and 16-times per medical record, respectively.

The data, collected by questionnaire, were the response to each criterion and the time spent reading the record and determining the score (completion time).

Study endpoints and specific statistical analysis

The primary endpoint was the inter-rater reliability. Inter-rater reliability was assessed using the Krippendorff's alpha coefficient test: $\alpha = 1$ indicates perfect reliability and $\alpha = 0$ the absence of reliability; the score was considered similarly interpretable by different raters if $\alpha \geq 0.8$, and still acceptable if $\alpha \geq 0.67$.[7] The mean concordance for each criterion was calculated by averaging the percentages of agreement for each criterion of each patient; 95 % Confidence Intervals of concordance were computed using 1000 iteration bootstrapping with random resampling with replacement of patients and operators. The analysis of agreement for each value was assessed using the Fleiss' Kappa test.[8] The times were compared using the Kruskal-Wallis test.

To our knowledge, formal power calculations are not available for Krippendorff's α or Fleiss' κ. However, a post-hoc power calculation for the *t*-test comparing completion times indicated a power of 0.88.

Results

Score reproducibility

The distribution of scores obtained for each record with NOVA and DENOVA is shown in Fig. 1. No record was given the same score by the 8 evaluators. The score was positive or negative for all evaluators only for half of the patients (8/16 for NOVA and 4/8 for DENOVA).

Using the Krippendorff test, inter-rater reliability was significant for both scores but with a low α-value for NOVA ($\alpha = 0.37$ [0.24–0.49]; $p < 0.001$) and medium for DENOVA ($\alpha = 0.49$ [0.24–0.73]; $p < 0.01$). When considering each criterion independently, criteria V and A showed strong inter-rater reliability ($\alpha = 0.86$ and 0.83; concordance 95 %). E and N had moderate reliability ($\alpha = 0.34$ and 0.41; concordance 77 % and 75 %). In contrast, D and O had the lowest reliability ($\alpha = 0.16$ and 0.13; concordance 55 % and 62 %). Confidence intervals and p-values are detailed in Table 1.

When assessing the agreement according to the values of the score by the Fleiss Kappa test, a moderate agreement was observed for very high or very low score values, but it was not significant for intermediate values: 6–9 for NOVA and 1–3 for DENOVA.

Score feasibility

No significant difference was observed between the median completion times for NOVA and DENOVA scores (4′38 vs. 5′32, $p = 0.21$). Median completion time differed between evaluator groups: 4′11 (extremes: 1′46 to 9′06) for cardiologists, 4′32 (2′36 to 10′05) for infectious diseases physicians, 5′09 (1′15 to 12′55) for internal medicine physicians, and 6′37 (3′00 to 14′20) for medical residents ($p < 0.01$).

Table 1
Inter-rater reliability according to the criteria using the Krippendorff test.

Criterion	Mean concordance (95 % CI)	Alpha (95 % CI)	p-value
D	55 % (45–94)	0.16 (0–0.38)	0.14
E	77 % (61–95)	0.34 (0–0.75)	0.1
N	75 % (59–95)	0.41 (0–0.86)	0.06
O	62 % (46–81)	0.13 (0–0.38)	0.28
V	95 % (78–100)	0.86 (0.51–1)	<0.001
A	95 % (88–100)	0.83 (0.53–1)	<0.001

The mean concordance was calculated by averaging the percentages of agreement for each criterion of each patient (the value is between 50 % and 100 %); 95 % Confidence Intervals of concordance were computed using 1000-iteration bootstrapping with random resampling with replacement of patients and operators. Inter-rater reliability was assessed for each criterion using the Krippendorff's alpha coefficient test ($\alpha = 1$ indicates perfect reliability and $\alpha = 0$ the absence of reliability; the score was considered similarly interpretable by different raters if $\alpha \geq 0.8$, and still acceptable if $\alpha \geq 0.67$).

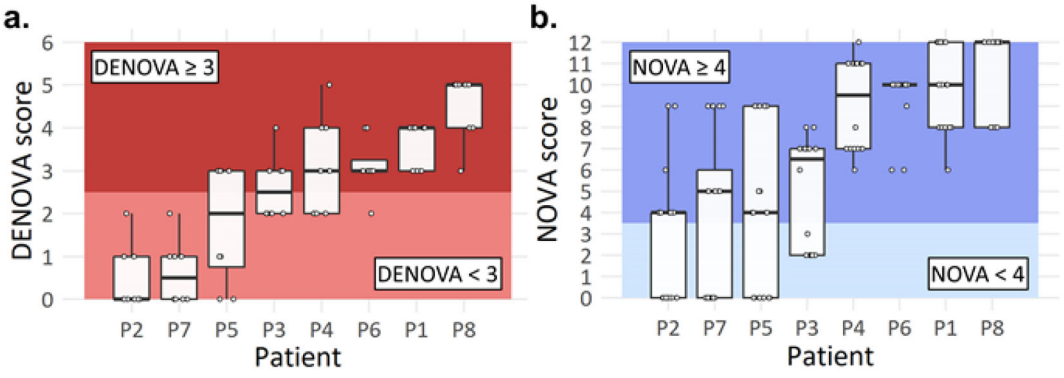

Fig. 1. Score obtained for each patient record (P1 to P8) with the NOVA (a) and DENOVA (b) scores. Each white circle corresponds to the answer to a questionnaire; the line on the y-axis represents the score cutoff.

This time did not differ whether the score was positive or negative for both NOVA (respectively 4′33 vs. 5′15, $p = 0.4$) and DENOVA (6′03 vs. 4′55, $p = 0.21$).

Discussion

Both scores showed overall poor inter-rater reproducibility, particularly for values close to the positivity threshold. It was also observed that some raters occasionally disagreed with the conclusion of the score they obtained (data not reported). Beyond diagnostic performance and reproducibility, we also considered feasibility, using completion time as a pragmatic, albeit imperfect, indicator. The similar durations observed for NOVA and DENOVA suggest that the two additional items in DENOVA do not increase the practical burden for clinicians. As in other domains, we confirm that it is therefore imperative to assess both score reproducibility and ease of use when developing a new scoring system for endocarditis risk assessment, especially when the criteria involved are susceptible to personal interpretation.[9]

In addition, DENOVA and NOVA scores have other limitations. Endocarditis diagnosis has no gold standard and relies on major and minor criteria – some of which are already part of the DENOVA score (number of positive blood cultures, predisposing cardiac lesion, embolization) – and the major criterion of cardiac imaging is not sought in all patients in the studies.[6] Moreover, using retrospective data for the elaboration and validation of the score is also an issue: firstly, it is easier to evaluate the criteria afterwards with the entire medical record, and secondly, determining the exact time of assessment is often ambiguous, as the question of TEE remains unresolved for many patients. These limitations are sources of important biases which may overestimate the score performance.[10]

In our study, raters assessed retrospective data. This design made it possible to explore both interpretation variability and uncertainty, while allowing inclusion of a larger number of independent evaluators. Nevertheless, it may not reproduce real-time clinical reasoning. In addition, the small number of clinical scenarios may not fully reflect the diversity of *E. faecalis* bacteremia presentations, whereas the limited number of raters per case could have affected the precision of inter-rater comparisons.

Further studies, particularly with prospective data collection, are therefore needed to better define the interest of the endocarditis scores in real life, and to compare their performance with clinical judgement. The objective remains to avoid unnecessary TEE in patients at low risk of endocarditis.

Conclusion

When assessing the risk of endocarditis in patients with *E. faecalis* bacteremia, low inter-rater reproducibility suggests that scores should not replace clinical judgment, especially in situations of intermediate risk.

Ethical approval

The study was approved by an institutional review board, the Ethical Committee of Research in Tropical and Infectious Diseases (CER-MIT 2022–0106).

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank all the physicians who responded to the questionnaires, Prof. Xavier Duval and Prof Mathieu Nacher for critical reading of the manuscript.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Rasmussen M, Gilje P, Fagman E, Berge A. Bacteremia with gram positive bacteria – when and how do I need to look for endocarditis? *Clin Microbiol Infect*. 2024;30:306–311.
2. Østergaard L, Bruun NE, Voldstedlund M, Arpi M, Andersen CØ, Schønheyder HC, et al. Prevalence of infective endocarditis in patients with positive blood cultures: a Danish nationwide study. *Eur Heart J*. 2019;40:3237-44.
3. Bouza E, Kestler M, Beca T, Mariscal G, Rodríguez-Créixems M, Bermejo J, et al. The NOVA score: a proposal to reduce the need for transesophageal echocardiography in patients with enterococcal bacteremia. *Clin Infect Dis*. 2015;60:528-35.
4. Berge A, Krantz A, Östlund H, Nauclér P, Rasmussen M. The DENOVA score efficiently identifies patients with monomicrobial Enterococcus faecalis bacteremia where echocardiography is not necessary. *Infection*. 2019;47:45-50.
5. Danneels P, Chabrun F, Picard L, Martinet P, Rezig S, Lorleac'h A, et al. Enterococcus faecalis endocarditis risk assessment in patients with bacteremia: external validation of the DENOVA score. *J Infect*. 2023;87:571-3.
6. Delgado V, Ajmone Marsan N, de Waha S, Bonaros N, Brida M, Burri H, et al. 2023 ESC guidelines for the management of endocarditis. *Eur Heart J*. 2023;44:3948-4042.
7. Krippendorff K. Computing Krippendorff's Alpha-Reliability. Departmental Papers (ASC) [Internet] 2011;Available from: https://repository.upenn.edu/asc_papers/43.
8. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76:378-82.
9. Barth J, de Boer WEL, Busse JW, Hoving JL, Kedzia S, Couban R, et al. Inter-rater agreement in evaluation of disability: systematic review of reproducibility studies. *BMJ*. 2017;356:j14.
10. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;282:1061-6.