Brief report

Arch Argent Pediatr. 2026;e202510882

Entropy reduction and F-score: Applications of information theory in diagnostic test evaluation

Eduardo Cuestas[1,2,3], María E. Cieri[2], L. Johana Escobar Zuluaga[2], María M. Ruiz Brünner[2]

ABSTRACT

This paper proposes an approach based on information theory and machine learning, applying the F-score and entropy reduction to analyze the flow of information among the patient, the diagnostic test used to evaluate the disease, and the evaluators who interpret the results, thereby enriching the critical assessment of diagnostic tests.

The F-score, by integrating positive predictive value and sensitivity, offers a synthetic measure that is less dependent on prevalence. Entropy reduction is a comprehensive metric that quantifies the decrease in uncertainty and informational gain of a diagnostic test, allowing accurate comparisons between different tests using a single measure of accuracy.

For practical application, we designed a specific calculator that integrates these indicators. Its implementation would facilitate the interpretation of diagnostic studies and improve clinical decision-making.

Keywords: *predictive value of tests; uncertainty; information theory; entropy; routine diagnostic tests.*

doi: http://dx.doi.org/10.5546/aap.2025-10882.eng

To cite: Cuestas E, Cieri ME, Escobar Zuluaga J, Ruiz Brünner MM. Entropy reduction and F-score: Applications of information theory in diagnostic test evaluation. *Arch Argent Pediatr.* 2026;e202510882. Online ahead of print 26-MAR-2026.

[1] Pediatrics and Neonatology Service, Hospital Privado Universitario de Córdoba, Córdoba, Argentina; [2] Instituto Universitario de Ciencias Biomédicas de Córdoba (IUCBC), Centro de Investigación en Medicina Traslacional Severo R. Amuchástegui (CIMETSA); Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Córdoba, Argentina; [3] 2° Chair of Pediatric Clinic, Faculty of Medical Sciences, Universidad Nacional de Córdoba, Argentina.

Correspondence to *Eduardo Cuestas: eduardo.cuestas@iucbc.edu.ar*

Funding*: None.*

Conflict of interest*: None.*

Received*: 9-2-2025*
Accepted*: 12-18-2025*

INTRODUCTION

The proper use of medical literature in evidence-based medicine reduces medical errors.[1] Assessing the diagnostic accuracy of complementary tests is an essential step in improving patient safety and quality of care. Traditional metrics — sensitivity, specificity, and positive (PPV) and negative (NPV) predictive values — are fundamental to diagnostic evaluation. However, they have limitations when comparing tests in different clinical contexts.[2] Sensitivity and specificity lose their usefulness once the result is obtained, at which point the pediatrician must consider the post-test probability, represented by the PPV and NPV. These, in turn, depend on prevalence, making it difficult to compare them across populations with different epidemiological profiles.[3]

Although Claude Shannon's information theory (1948) has been successfully applied in many fields,[4] its use in medical diagnostic evaluation remains limited.[5,6]

The concepts of uncertainty reduction (entropy) and precision-exhaustiveness[3,7] provide solid tools for making a source of information more predictable.[8] Uncertainty reduction measures how much a test clarifies the presence or absence of disease. At the same time, the F-score summarizes the balance between sensitivity (precision) and PPV (exhaustiveness), with relative independence from prevalence.

The objective of this study was to apply and adapt tools derived from information theory and machine learning—the F-score and entropy reduction—to evaluate and compare diagnostic tests. It also presents a specific calculator for their implementation to enrich critical analysis and strengthen evidence-based decision-making.

METHODS

First, we used the F-score—traditionally used in computer science—to evaluate medical diagnostic tests.[3] This indicator was adapted to estimate the test's performance in detecting sick patients, reducing the influence of prevalence by combining sensitivity with PPV.

Mathematically, it is expressed as the harmonic mean of both metrics:

$$\text{F-score} = 2 \cdot \frac{\text{PPV} \cdot \text{Sensitivity}}{\text{PPV} + \text{Sensitivity}}$$

The F score reflects the balance between sensitivity, i.e., the ability to detect patients correctly, and positive predictive value, which indicates the probability that a positive result corresponds to a true case. High values indicate good overall diagnostic performance in a single metric.

Shannon entropy *(H)* measures the uncertainty associated with a probability distribution. In medicine, this translates into diagnostic uncertainty before and after a test. A useful test should reduce entropy, i.e., reduce uncertainty about whether a patient has a disease.

General formula for Shannon entropy:[8]

$$H(x) = -\Sigma p_i \cdot \log 2(p_i)$$

For binary events (sick/healthy), the maximum occurs when $p = 0.5$ (maximum uncertainty).

Calculation of diagnostic entropy reduction:

Start with a 2 × 2 table containing the test data (TP = true positives, FP = false positives, FN = false negatives, TN = true negatives, and N = total cases) and calculate:

1. Pre-test entropy:
   Pre-test H = [FP+TN] / N.(log2(N)-log2(FP+TN)) + [TP+FN]\/N.(log2(N)-log2(TP+FN))
2. Post-test entropy for positive results (positive n):
   Post-test H (+) = [VP]\/positive n.(log2(positive n)-log2(VP)) +[FP]\/positive n.(log2(positive n)-log2(FP))
3. Post-test entropy for negative results (n negative):
   Post-test H (-) = [FN]\/n negative.(log2(n negative)-log2(FN)) +[VN]\/n negative.(log2(n negative)-log2(VN))
4. Entropy reduction:
   Δ H= pre-test H - [positive n\/N .post-test H (+) + negative n\/N . Post-test H (-)
5. Interpretation: the greater the Δ, the greater the ability to reduce diagnostic uncertainty. If the test were perfect (100% sensitivity and 100% specificity), the uncertainty would be reduced by 100%.

To illustrate the practical application, we analyzed an example using data taken from the article "Urine test strip for the diagnosis of urinary tract infections in febrile infants aged 2 to 6 months".[9] These data were used to calculate sensitivities, specificities, PPV, and NPV, along with true and false positives/negatives, likelihood ratios, entropy reduction, F-score, and diagnostic accuracy. This study is exempt from institutional ethical review because it involves only pre-existing data from public sources and poses no risk of direct or indirect identification of the subjects.

RESULTS

According to the data in *Table 1*, leukocyte esterase is the most robust test for diagnosing urinary tract infection, achieving the greatest reduction in entropy (48.3%), reflecting the greatest information gain and the most marked decrease in uncertainty between the pre- and post-test situations. This indicator integrates all components of the 2 × 2 table, weighted by their relative frequencies, and simultaneously shows the ability to both confirm and rule out disease.

In this sense, entropy reduction is the parameter that best captures the actual information gained in the relative comparison of diagnostic tests in clinical practice, as it directly reflects the decrease in uncertainty. Likewise, leukocyte esterase achieves the highest F-score (68.9%), resulting from an optimal balance between sensitivity (91.2%) and positive predictive value (55.3%), making it less dependent on prevalence and, therefore, more reliable.

In contrast, nitrites, although exhibiting greater specificity (98.9%) and positive predictive value (80.8%), have a clearly lower sensitivity (36.7%), while leukocyte count shows intermediate performance.

Figure 1 shows the variation of the PPV (solid lines) and the F-score (dashed lines) according to prevalence, for sensitivities and specificities of 0.90, 0.80, 0.70, and 0.60. As prevalence increases, the PPV increases markedly, following a non-linear rational relationship that demonstrates its strong dependence on prevalence at all levels of diagnostic performance.

Even with a sensitivity and specificity of 0.80, the PPV decreases substantially when prevalence falls, going from very high values at high prevalences to much lower values at intermediate and low prevalences. In this scenario, the PPV drops from approximately 0.97 at a prevalence of 0.90 to around 0.73 when prevalence falls to 0.40, implying an absolute decrease of 0.24 points and a relative reduction of 24.7%.

In contrast, with the same diagnostic performance, the F-score shows a much more attenuated variation over the same range of prevalences. Its curves (dotted lines) adopt a smoothly polynomial shape, more stable and less susceptible to changes in prevalence. With a sensitivity and specificity of 0.80, the F-score decreases from approximately 0.88 at a prevalence of 0.90 to 0.76 when the prevalence is 0.40, representing an absolute decrease of 0.12 points and a relative reduction of 13.6%, approximately half that observed in the PPV. This quantitative difference shows that, although the F-score is not completely independent of prevalence, its variation in response to changes in prevalence is considerably less than that of the PPV, especially in tests with moderate to high

Table 1. Comparison of the diagnostic value of leukocyte esterase, nitrite, and urine sediment tests vs. urine culture[1]

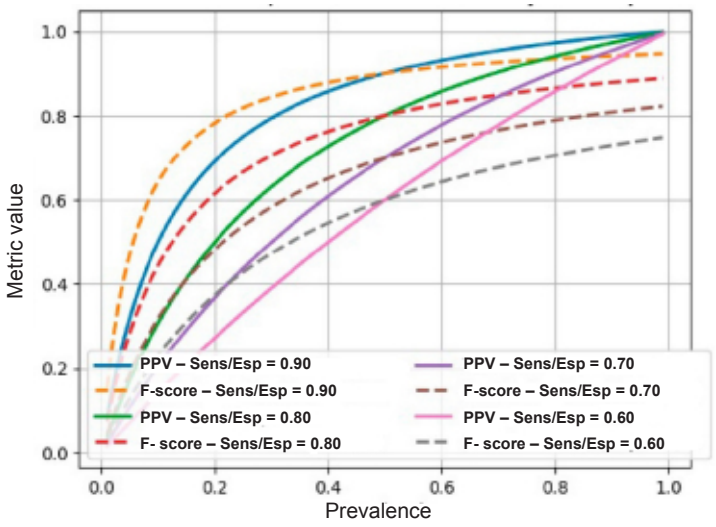
Metric	Leukocyte esterase	Nitrites	Leukocyte count ≥5 × HPF
True positives, n	810	373	759
False negatives, n	78	645	121
True negatives, n	6,217	7950	3489
False positives, n	655	89	742
Sensitivity, %	91.2	36.7	86.2
Specificity, %	90.5	98.9	82.5
Positive predictive value, %	55.3	80.8	50.6
Negative predictive value, %	98.8	92.5	96.6
Positive likelihood ratio	9.57	33.15	4.92
Negative likelihood ratio	0.10	0.64	0.17
Entropy reduction, %	48.3	20.9	33.1
F-score, %	68.9	13.8	63.7
Accuracy, %	90.5	91.9	83.1

HPF: high power field.

[1]From: Hunt KM, et al. Urine Dipstick for the Diagnosis of Urinary Tract Infection in Febrile Infants Aged 2 to 6 Months. Pediatrics. *2025;155:e2024068671.*

Calculations were performed using the free online Diagnostic Test Metrics calculator developed by the authors, available at: https://ejcuestas.github.io/diagnostic-calculator/

Figure 1. Influence of prevalence on positive predictive value (PPV) and F-score in diagnostic tests with equivalent sensitivity and specificity*



**Simulations were performed in Python (3.14.2) using NumPy to generate simulated populations and explore a continuous range of prevalences between 0.01 and 0.99, maintaining constant sensitivity and specificity in each scenario (0.90; 0.80; 0.70; 0.60). The Monte Carlo method was used with 100000 iterations per condition to ensure numerical stability. The positive predictive value (PPV) was calculated using the standard Bayesian formula, and the F-score was derived from the PPV and sensitivity according to the canonical formula.*

The resulting curves were plotted using Matplotlib. No fitting or inference methods were applied, as the objective was to deterministically illustrate the comparative behavior of both metrics.

diagnostic performance.

In summary, leukocyte esterase, by achieving the optimal balance between F-score and entropy reduction, is the most informative and consistent test for clinical practice.

To facilitate its application in clinical practice, we have specifically developed an online calculator that allows readers to directly and simultaneously calculate the F-score and entropy reduction by entering the basic data (TP, FP, TN, FN) into the 2 × 2 contingency table, available at: https://ejcuestas.github.io/diagnostic-calculator/

DISCUSSION

Analysis of data from 2 × 2 diagnostic tables using metrics derived from information theory and machine learning, such as entropy reduction and F-score, shows the potential of these tools to strengthen critical evaluation of evidence.

Entropy reduction is a comprehensive metric that quantifies the decrease in uncertainty and informational gain of a diagnostic test, allowing accurate comparisons between different diagnostic tools using a single measure of accuracy.[6]

Similarly, the F-score facilitates a more nuanced comparison between diagnostic tests by transcending the limitations of PPV and offering a measure of overall predictive performance, relatively independent of prevalence, reflecting the balance between correctly identifying true cases and minimizing misclassifications, surpassing the value of PPV and sensitivity considered in isolation.[5]

Our results reveal that leukocyte esterase provided the greatest reduction in entropy (48.3%) and the best F-score (68.9%). This finding underscores the diagnostic superiority of this method when evaluated using metrics that integrate accuracy (sensitivity) and completeness (PPV).[4]

This perspective represents a step toward personalizing clinical decisions. Understanding the reduction in uncertainty at each diagnostic step enables more informed decision-making. This approach could facilitate personalized decision-making in real time in different settings, representing an improvement over standard practice, which is often based on population risk assessment based on patient history and demographics.[4,8]

Our findings constitute a statistical-mathematical model that requires further clinical validation. Although the application of Shannon entropy and the F-score is a very promising use of information theory in diagnostic analysis,

its clinical advantages must be confirmed by prospective studies evaluating its performance in real-world practice.

The main limitation is that these metrics, like traditional indicators, do not always reflect the clinical consequences of false positives and false negatives. This consideration is subject to clinical judgment and pediatric experience, where the cost of a false negative may exceed that of a false positive, or vice versa, depending on the specific clinical scenario.[10]

In conclusion, the results indicate that both the F-score and entropy reduction can be valuable tools for the comparative analysis of diagnostic tests and for the design of future studies and systematic reviews. Their application in different populations will allow for the corroboration of the stability and validity of these indicators in various clinical contexts. ■

REFERENCES

1. Cosby K, Yang D, Fineberg HV. Assessing Diagnostic Performance. *NEJM Evid*. 2024;3(2):EVIDra2300232. doi:10.1056/EVIDra2300232.
2. Van Rijsbergen CJ. Information Retrieval. 2nd ed. Michigan: Butterworths; 1979.
3. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep*. 2022;12(1):5979. doi: 10.1038/s41598-022-09954-8.
4. Casagrande A, Fabris F, Girometti R. Fifty years of Shannon information theory in assessing the accuracy and agreement of diagnostic tests. *Med Biol Eng Comput*. 2022;60(4):941-55. doi:10.1007/s11517-021-02494-9.
5. Krause P. Information Theory and Medical Decision Making. *Stud Health Technol Inform*. 2019;263:23-34. doi: 10.3233/SHTI190108.
6. Benish WA. A Review of the Application of Information Theory to Clinical Diagnostic Testing. *Entropy (Basel)*. 2020;22(1):97. doi:10.3390/e22010097.
7. Mosquera C, Ferrer L, Milone DH, Luna D, Ferrante E. Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance. *Eur Radiol*. 2024;34(12):7895-903. doi: 10.1007/s00330-024-10834-0.
8. He S, Chong P, Yoon BJ, Chung PH, Chen D, Marzouk S, et al. Entropy removal of medical diagnostics. *Sci Rep*. 2024;14(1):1181. doi: 10.1038/s41598-024-51268-4.
9. Hunt KM, Green RS, Sartori LF, Aronson PL, Chamberlain JM, Florin TA, et al. Urine Dipstick for the Diagnosis of Urinary Tract Infection in Febrile Infants Aged 2 to 6 Months. *Pediatrics*. 2025;155(4):e2024068671. doi: 10.1542/peds.2024-068671.
10. Taylor RA, Sangal RB, Smith ME, Haimovich AD, Rodman A, Iscoe MS, et al. Leveraging artificial intelligence to reduce diagnostic errors in emergency medicine: Challenges, opportunities, and future directions. *Acad Emerg Med*. 2025;32(3):327-39. doi: 10.1111/acem.15066.