

Methods note · Observation and data

Rare events and the denominator of a flag

Amina Okafor · 13 October 2026

Fictional author and institute · Editorial demonstration, not peer reviewed

ABSTRACT

High sensitivity does not establish that most threshold flags are correct. This educational methods note uses a synthetic waveform archive to distinguish sensitivity, false-positive rate, precision, false-alarm ratio, and overall accuracy. Natural-frequency tables demonstrate how a rare target can yield more incorrect than correct flags, while a second population shows how precision changes with the base rate even when conditional rates are held fixed. A stricter invented threshold makes the retrieval and review-burden trade-off explicit. Primary reasoning and classifier-evaluation studies accompany forecast-verification definitions. The note also addresses event units, correlated windows, threshold tuning, and human review. It validates no detector, recommends no operational threshold, and explicitly excludes medical, navigation, emergency, and other safety applications.

Keywords: base rates · false positives · precision and recall · threshold evaluation

A source-based editorial text, not a report of original Still North research. Examples and imagery are illustrative.

A flag is evidence, not confirmation

A threshold can detect most target events and still produce many flags that do not correspond to an event. This is not a contradiction. When events are rare, a small false-positive fraction applied to the much larger non-event population can rival or exceed the number of correct detections. The missing information in a claim such as “90% detection” is often the denominator, not another decimal place.

This educational methods note compares natural-frequency reasoning, contingency-table definitions, and research on imbalanced classifier evaluation. Its case uses a wholly synthetic waveform motif with known labels. It does not concern medical testing, navigation, ice travel, emergency warnings, or any other operational safety system. The invented rates are not performance claims for an actual instrument or algorithm, and no practical alarm threshold is recommended.

Keep four counts and three conditional questions separate

A labelled test set produces four outcomes: correctly flagged targets, missed targets, incorrectly flagged non-targets, and correctly unflagged non-targets. Sensitivity, also called recall or hit rate, divides correctly flagged targets by all actual targets. The false-positive rate divides incorrectly flagged non-targets by all actual non-targets. Precision, or positive predictive value, divides correctly flagged targets by all flags.

The Joint Working Group on Forecast Verification Research’s guidance uses contingency tables to distin-

guish hits, misses, false alarms, and correct negatives. Its terminology also illustrates a trap: the false-alarm ratio uses all positive forecasts as its denominator, while the false-alarm rate uses observed non-events. These are different quantities. A report should give a definition rather than assume that readers interpret “false alarms” consistently.

Precision answers the reader’s likely question after seeing a flag: among cases flagged by this procedure in this population, what fraction are targets? Sensitivity answers a different question: among targets, what fraction were flagged? Reversing those conditionals is a fundamental error. Neither a polished interface nor an impressive overall accuracy can bridge it.

Fictional worked case: a synthetic archive

Suppose 100,000 independently generated waveform windows form an invented teaching archive. Exactly 1,000 contain the target motif, so the base rate is 1%. A fixed rule flags 90% of target windows and 1% of non-target windows. For this example, treat the resulting counts as exact. There is no uncertain labelling process and no ambiguity about whether a motif is present.

The rule detects 900 of 1,000 targets, but only 900 of 1,890 flags are correct. Precision is approximately 47.6%. The false-positive rate is 990 divided by 99,000, or 1%. The false-alarm ratio is 990 divided by 1,890, approximately 52.4%. Saying that the rule has “only 1% false alarms” without naming the denominator would invite a very different interpretation from the fraction of flags that are wrong.

Rule output	Target present	Target absent	Total
Flagged	900	990	1,890
Not flagged	100	98,010	98,110
Total	1,000	99,000	100,000

Overall accuracy is 98.91%, since 98,910 windows are classified correctly. Yet a rule that never flags anything would achieve 99% accuracy by exploiting the rarity of targets, while missing every one. This does not make the never-flag rule useful. It shows why accuracy alone can reward the wrong behaviour when the purpose is finding rare events.

Hold performance fixed and change the population

Now imagine a second synthetic archive with a 10% target base rate. For arithmetic comparison only, hold sensitivity at 90% and the false-positive rate at 1%. Among 100,000 windows there are 9,000 correct flags and 900 incorrect flags. Precision becomes 9,000 divided by 9,900, approximately 90.9%. The rule's assumed conditional rates have not improved; the mix of cases has changed.

In a real transfer, those rates may also change because background signals, event strength, or annotation differ. The calculation isolates the base-rate effect rather than asserting stable performance across environments. A demonstration built from equal numbers of target and non-target examples can help compare discrimination, but its observed precision should not be advertised as the expected precision in a rare-event archive without an appropriate adjustment and supporting assumptions.

The general expression is sensitivity multiplied by the target base rate, divided by that quantity plus the false-positive rate multiplied by the non-target base rate. Natural frequencies make the same calculation easier to inspect. Gigerenzer and Hoffrage's original experiments studied how presenting such problems in frequency formats affects Bayesian reasoning. Their findings support attention to information format, not a claim that one table will guarantee understanding for every audience.

A stricter threshold changes both kinds of error

Return to the 1% archive. Suppose another invented threshold has 70% sensitivity and a 0.2% false-positive rate. It produces 700 correct flags, 198 incorrect flags, and 300 missed targets. Precision is approximately 78.0%, and the review queue falls from 1,890 windows to 898. But 200 additional targets are missed compared with the first threshold.

The stricter threshold is not unconditionally better. It answers a different trade-off between retrieval and review burden. One archive task might prioritise finding nearly every motif for later analysis; another might require a small high-precision sample. The purpose and consequences determine what trade-off is acceptable. The toy calculation cannot determine that choice, and it must not be repurposed as a safety-system design rule.

Saito and Rehmsmeier examine how precision-recall and receiver-operating-characteristic displays behave with imbalanced datasets. Their simulations and re-analysis show why an attractive ROC summary can obscure the practical number of false positives. ROC analysis remains informative about discrimination; precision-recall analysis makes the composition of retrieved cases more visible. Neither display removes the need to state the base rate and evaluation population.

Define the unit before counting success

A window is not always an event. A single long motif can occupy ten adjacent windows, producing ten flags but one underlying occurrence. Conversely, several short motifs can fall in one window. A method evaluated per window should not quietly be described as detecting the same fraction of independent events. Specify matching rules, time tolerance, and whether the unit is a frame, window, episode, or site.

Correlation also changes uncertainty. Adjacent windows from one recording can share background conditions. Splitting them randomly between development and test sets can make testing resemble material already seen during tuning. A separate recording period, deployment, or site can offer a more meaningful challenge, depending on the intended transfer. Such separation should follow the scientific question rather than a ritual percentage split.

Threshold selection needs separation from final evaluation too. Trying many cutoffs on a test set and reporting the best one makes that set part of development. Preserve a suitable final evaluation or account for the selection process. Report the four counts and uncertainty in the estimated rates, especially when target examples are few. A large total sample does not compensate for only a handful of positive cases.

Human review does not erase the first stage

A flagging rule may be used to create a queue for human interpretation. Review can reduce incorrect final labels, but the whole workflow must then be evaluated. If reviewers inspect only flags, they cannot discover all missed targets. A suitably designed sample of unflagged material can inform that problem. The final report should distinguish automatic flags, reviewed detections, and any population estimate inferred from them.

Reviewers also need the context required for a sound label. A cropped waveform may conceal an alternative explanation visible in neighbouring time. Agreement between reviewers is informative but not a perfect reference truth. The fictional example assumes flawless labels precisely to keep the denominator problem clear. Actual annotation uncertainty needs its own account rather than being absorbed into a claim of algorithmic performance.

Limits and implication

The calculations assume exact rates, fixed populations, independent windows, and perfect target definitions. Real observation tasks can violate all four assumptions, and the rarity of an event can change over time. This note supplies no validated detector, acceptable error rate, or action protocol. Its tables should remain an educational illustration.

The practical communication lesson is nonetheless robust: report what was counted, how often the target occurred, the four outcomes, and the conditional quantity relevant to the reader's question. A threshold flag can be useful without being confirmation. Showing its denominator makes that usefulness assessable instead of leaving a high percentage to imply more than it proves.

References

1. Gigerenzer, G., and Hoffrage, U. (1995). How to Improve Bayesian Reasoning Without Instruction: Frequency Formats. *Psychological Review*, 102, 684–704.
2. Saito, T., and Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10, e0118432.
3. Joint Working Group on Forecast Verification Research. *Forecast Verification: Issues, Methods, and Frequently Asked Questions*, sections on dichotomous forecasts and validity.