

Replaying Beat-Interval Estimates in MIT-BIH Record 100 for Verifiable Waveform Reasoning

Caleb Foster, Ryan Cooper, Dylan Brooks

Abstract

Verifiable physiological reasoning begins with a measurable claim whose inputs, units, and transformations can be replayed. This secondary analysis examines that requirement using reference annotations from MIT-BIH Arrhythmia Database record 100. Of 2,274 annotation entries, 2,273 identify beats; restricting intervals to consecutive normal-labeled beats yields 2,204 intervals. Their median is 797.222 milliseconds. An ordinary least squares projection of interval duration on recording time gives a slope of 0.108 milliseconds per minute, with a Newey-West 95 percent interval from -0.453 to 0.669. This small linear association does not establish physiological stationarity or clinical normality. The analysis uses annotations rather than newly detected waveform events and evaluates no foundation model or language agent. It develops a concrete measurement ledger connecting a published waveform-reasoning position with the evidence that a future model would need to preserve, inspect, and verify.

A position paper becomes a measurable question

Wang and colleagues advocate physiological waveform reasoning supported by localized signal evidence and a closed-loop planning, action, and verification process (X. Wang et al., 2026). This position distinguishes a persuasive explanation from a claim that can be checked against a signal. The present article operationalizes one small part of that distinction. Before asking whether a model can reason about a recording, it asks whether a simple timing statement can be reconstructed from a versioned source without ambiguity about the events, units, or retained observations.

The chosen statement concerns intervals between adjacent reference beats labeled normal. It is intentionally narrower than a diagnosis or an assessment of rhythm. A statement about a median interval can be true while concealing short episodes, annotation mistakes, or a change in another signal property. A linear time association can also be close to zero while the series varies nonlinearly. The purpose is not to make the statement clinically comprehensive, but to make its scope precise enough that another analyst can determine exactly what has and has not been measured.

This framing matters for language-based systems because an explanation can silently broaden its evidence. A model shown a sequence of normal-labeled intervals might describe the entire recording as normal, although abnormal labels were excluded by construction. Another model might call an imprecise slope evidence of stability. The proposed verification ledger should retain these distinctions. It should identify the claim type, annotation rule, time span, measurement unit, aggregation, and statistical uncertainty before a higher-level physiological interpretation is considered.

Source and retained observations

The analysis uses the public MIT-BIH Arrhythmia Database version 1.0.0 and the downloaded header and reference annotation file for record 100 (Moody and Mark, 2005). The header specifies a sampling frequency of 360 samples per second and a recording duration of approximately 30.093 minutes. The annotation file contains both beat and non-beat entries. The single non-beat entry is excluded before intervals are constructed. Raw waveform amplitudes are not processed, and the result should not be described as an independent evaluation of beat detection or waveform morphology.

The database's role as a shared reference for arrhythmia analysis is established in its original account (Moody and Mark, 2001). A shared reference provides reproducible event locations, but it does not turn

one recording into a representative clinical sample. The present unit is one interval within one selected record. Record 100 was chosen explicitly for a small replayable demonstration, not sampled to estimate a population distribution. The number of intervals therefore cannot be used as the number of independent patients or clinical episodes supporting a general physiological conclusion.

There are 2,239 normal-labeled beat annotations, 33 atrial-premature labels, and one ventricular-premature label. Intervals are formed after retaining the beat entries, and only pairs whose current and preceding beat are both labeled normal are kept. This rule yields 2,204 intervals. It excludes intervals adjacent to a non-normal label even when the other beat is normal. No additional duration threshold is used to trim the result. This prevents an outcome-dependent exclusion rule from making the retained timing distribution appear artificially narrow.

Transformation and an explicitly limited model

For consecutive retained beat locations s and s_{previous} , interval duration is calculated as 1,000 times their sample difference divided by 360. Recording time is the current beat's sample location divided by 360 and then by 60. These transformations produce milliseconds and minutes respectively. The rule can be replayed directly from the source annotations. Retaining sample indices alongside converted units also makes it possible to inspect rounding and to map a reported interval back to the reference event locations.

The fitted descriptive specification is $RR_{\text{ms}} = \text{intercept} + \text{slope} \times \text{elapsed_minutes} + \text{residual}$. It is a within-record linear projection, not a physiological process model. The specification does not include treatment, patient characteristics, signal morphology, or latent state. It is useful because its narrow target makes unsupported interpretations visible. Any statement that the coefficient represents an intervention effect, a change in disease, or a comparison between patients would require information absent from this model.

Consecutive intervals can be serially dependent. The analysis therefore uses Newey-West heteroskedasticity and autocorrelation consistent covariance with twenty lags and the finite-sample correction available in the fitted implementation (Newey and West, 1987). This is a declared working choice for the interval sequence. It does not establish independence across all remaining lags or provide population inference from one record. The covariance calculation affects the uncertainty attached to the fitted projection, not the observed event locations or the definition of an interval.

Results of the annotation-based calculation

The mean retained interval is 795.012 milliseconds and its standard deviation is 35.961 milliseconds. The observed minimum and maximum are 652.778 and 888.889 milliseconds. These describe the selected normal-to-normal pairs. The median, 797.222 milliseconds, is a location summary rather than a statement that each interval is close to that value. Because the original sample grid is retained, the many displayed decimal places should be understood as transformed grid locations or aggregates, not newly acquired temporal precision.

Quantity	Estimate	Unit	Scope
Retained pairs	2204	intervals	Consecutive N-to-N beats
Median interval	797.222	ms	Selected pairs only
Linear time slope	0.108	ms/min	Within-record projection
Slope 95% interval	-0.453 to 0.669	ms/min	HAC, twenty lags

The fitted intercept is 793.393 milliseconds and the time-slope standard error is 0.286 milliseconds per minute. The within-file R-squared is approximately 0.000686. The confidence interval contains zero. More specifically, the declared linear specification does not precisely distinguish the sign of the time association at this confidence level. This is different from demonstrating no time variation. The very small R-squared also does not establish that the recording is uninformative: it describes only how much variation this single linear time term explains in the retained intervals.

No patient-level standard error is available because there is no patient sample in this calculation. Repeating the numerical fit on the same annotations would reproduce the projection, not create new physiological evidence. Resampling intervals independently would also discard their time structure. A broader uncertainty study would need more recordings, an explicit population, and a design that retains subject and acquisition dependencies. Those are future requirements, not properties implied by the present interval count.

What the calculation cannot verify

An annotation-derived interval is not an independently measured waveform event. A beat detector such as the real-time QRS method of Pan and Tompkins addresses a different step: locating events from the signal itself (Pan and Tompkins, 1985). A full waveform-reasoning assessment would need to compare the detector's event positions with the reference, record missed or extra events, and inspect the corresponding signal spans. The present analysis uses the reference positions as given. Its reproducibility therefore establishes the arithmetic and selection rule, not their complete physiological validity.

The restriction to normal-to-normal pairs creates a meaningful but selected outcome. It removes intervals associated with the non-normal labels and does not describe the prevalence or consequences of those labels. Normal is an annotation category here, not a statement about the participant's overall health. Likewise, a near-zero linear time coefficient cannot rule out transient episodes, nonlinear trends, or an artifact shared by the annotation process and the recording. Such distinctions should be preserved in any generated explanation of the results.

Heteroskedasticity-consistent inference and time-dependence corrections solve particular statistical problems rather than validate a clinical interpretation (White, 1980). A future analysis could compare covariance choices and time specifications, but it should not search those alternatives until a preferred conclusion appears. Preserving the initial question and reporting the alternatives would be more informative than presenting the narrowest interval. The important output is a transparent relationship between the recorded observations, the statistical target, and the interpretation actually licensed by that target.

A verification ledger for a future reasoning system

The minimum ledger for the median claim contains the database version, record identifier, source hashes, beat-label inclusion rule, sampling frequency, retained sample pairs, and aggregation code. The slope claim adds the time transformation, regression specification, covariance method, and lag choice. These are different layers of evidence. A source hash establishes which file was used; it does not establish that the annotation was correct. A successful replay establishes computational consistency; it does not establish that the chosen statistic answers the clinical question.

A future agent should be tested with plausible but incorrect claims about these same records. Examples include confusing seconds with milliseconds, counting non-beat annotations as beats, calling all selected pairs independent patients, or interpreting the confidence interval as proof of normality. These are proposed evaluation cases, not attacks conducted in this study. They make the target behavior concrete:

the system should identify the missing or contradictory evidence and either correct its statement or decline to extend it beyond the observed quantities.

Calibration has a role only after the system's correctness event is defined (Guo et al., 2017). Confidence in a numerical replay is different from confidence in a physiological interpretation. The two should not be merged into one unspecified reliability probability. A calibration study would need independently adjudicated claims spanning both correct and incorrect transformations. It would also need to report what happens on unfamiliar records, annotation conventions, or acquisition conditions. The current single-record audit supplies a traceable example, not that calibration dataset.

Adaptation and the option to stop

Tong and colleagues propose a Hadamard tensor-ring formulation for efficient low-rank fine-tuning (H. Tong et al., 2026). A prospective use in waveform reasoning would be to adapt a language component that interprets a measurement ledger. The adaptation target should preserve units, event provenance, and the distinction between arithmetic and physiological validation. This study has trained no such component. In particular, its interval regression does not measure a tensor-ring method's accuracy, speed, or clinical usefulness.

Low-rank adaptation provides an established comparison for parameter-efficient updating (Hu et al., 2022). Any future comparison should use the same ledger tasks and hold out records or acquisition settings, rather than place different windows of one recording on both sides of the evaluation. The resources spent producing verified training labels and checking generated interpretations belong in the budget. Fewer trainable parameters are not a substitute for correct interpretation of a time unit or an exclusion rule.

Selective prediction offers a framework for evaluating a system that can defer when its evidence is insufficient (Geifman and El-Yaniv, 2017). In this setting, the appropriate action might be to request the raw waveform, inspect a flagged interval, or decline a diagnostic inference. Deferral should be evaluated with its workload and consequence, not counted automatically as success. A system that always stops can be safe but uninformative. A useful assessment would compare the claims it answers, the unsupported claims it avoids, and the additional checking it requires.

Conclusion

A future reasoning benchmark could turn the present ledger into several deliberately different questions. One would ask for the selected interval median; another would ask whether the same data establish whole-record normality. The correct responses require different evidentiary thresholds even though they share one source. A model that repeats the right number while accepting the broader unsupported conclusion would fail the second task. This comparison would evaluate interpretation rather than reward numerical copying alone.

The benchmark should also preserve cases in which a requested claim cannot be answered from annotations. Questions about waveform amplitude, morphology, or detector disagreement would require additional signal evidence. Recording that missing requirement is a useful output, rather than an incomplete version of an answer the model should invent. Human assessment could then check whether the system requests the relevant evidence and whether its explanation identifies the precise gap. Those tests are proposed extensions; none was performed in the completed interval analysis.

The completed calculation demonstrates a small, replayable annotation-based timing claim: 2,204 selected intervals, a median of 797.222 milliseconds, and an imprecise linear association with recording time. It

also exposes the limits of that claim. Neither the annotation rule nor the regression establishes whole-record normality, clinical stationarity, or the performance of a reasoning model. Verifiable waveform reasoning requires these limits to remain available alongside the numerical result. The next useful experiment would test whether a model preserves this evidence boundary when it is asked to move from a measured interval to a physiological explanation.

References

- Wang, Xiaoda, et al. "Position: Beyond Prediction: Toward Verifiable Physiological Waveform Reasoning with Foundation Models and Agentic LLMs." *Proceedings of the 43rd International Conference on Machine Learning*, PMLR, vol. 306, 2026, pp. 172441-172459. <https://proceedings.mlr.press/v306/wang26lz.html>.
- Tong, H., Xu, G., & Chen, Y. (2026, May). Hadamard Tensor Ring for Efficient Low-Rank Fine-Tuning. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
- Moody, George, and Roger Mark. "MIT-BIH Arrhythmia Database." Version 1.0.0, *PhysioNet*, 2005. <https://doi.org/10.13026/C2F305>.
- Moody, G.B. and R.G. Mark. "The impact of the MIT-BIH Arrhythmia Database." *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, 2001, pp. 45-50. <https://doi.org/10.1109/51.932724>.
- Newey, Whitney K., and Kenneth D. West. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, vol. 55, no. 3, 1987, pp. 703-708. <https://doi.org/10.2307/1913610>.
- Pan, Jiapu and Willis J. Tompkins. "A Real-Time QRS Detection Algorithm." *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 3, 1985, pp. 230-236. <https://doi.org/10.1109/tbme.1985.325532>.
- White, Halbert. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. <https://doi.org/10.2307/1912934>.
- Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. "On Calibration of Modern Neural Networks." *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. <https://proceedings.mlr.press/v70/guo17a.html>.
- Hu, Edward J., et al. "LoRA: Low-Rank Adaptation of Large Language Models." *International Conference on Learning Representations*, 2022. <https://openreview.net/forum?id=nZeVKeeFYf9>.
- Geifman, Yonatan, and Ran El-Yaniv. "Selective Classification for Deep Neural Networks." *Advances in Neural Information Processing Systems*, vol. 30, 2017. <https://arxiv.org/abs/1705.08500>.