

Recoverable Structure and Measurement Uncertainty in Fetal ECG-to-Doppler Translation

Caleb Foster, Ryan Mitchell, Dylan Cooper

Abstract

Translating electrical recordings into a Doppler waveform raises two separate questions: how well a model predicts the recorded target, and what its errors reveal about fetal physiology. This methodological evidence synthesis examines that distinction using a recent fetal-maternal ECG-to-Doppler framework. The motivating study combines dilated convolutions and attention and reports evaluation on 885 synchronized segments from 39 pregnancies. Its signal-level results justify examining cross-modal predictability, while leaving the interpretation of residual waveforms dependent on measurement, synchronization, model capacity, and population transfer. We propose an evaluation sequence that begins with the measurement definition, preserves pregnancy-level separation during model selection, and reports temporal agreement alongside spectral and heart-rate error. Maternal-input ablations are treated as tests of predictive information rather than direct estimates of causal coupling. The article conducts no new clinical analysis or model training. Its contribution is an evidence framework for deciding which physiological claims a waveform reconstruction can support and which require additional observations.

Keywords: fetal ECG; Doppler envelope; cross-modal translation; measurement agreement; structured validation; residual interpretation.

1. Two modalities and one prediction target

Electrical and mechanical measurements describe related processes, but they are not duplicate measurements of a single scalar quantity. An ECG records electrical activity through a particular electrode configuration and processing pipeline. A Doppler envelope is derived from an ultrasound measurement with its own acquisition geometry and signal processing. Their temporal relationship can make translation useful without making every feature of one recoverable from the other. The scientific value of reconstruction therefore depends on specifying which target is being predicted and what interpretation is intended for the components that remain unexplained.

The Cross-Modal Generative Framework for Signal Translation from Fetal-Maternal Electrocardiograms to Fetal Doppler Waveforms uses dilated convolutions, maternal-input cross-modal attention, and self-attention to model this relationship (Su et al., 2026). The reported dataset contains 885 synchronized segments from 39 pregnancies. These details define the scale and architecture of the motivating study, not a new experiment in this article. The central methodological question is whether a residual between a generated envelope and a recorded envelope can be interpreted as a specifically mechanical component. That interpretation requires assumptions beyond the existence of a predictive model with a favorable reconstruction score.

Earlier fetal ECG signal-processing work describes the difficulties of separating fetal signals from maternal activity and other recording contributions (Sameni and Clifford, 2010). Those difficulties matter twice in a translation task. They affect the quality of the electrical input and can influence which recordings are retained for analysis. A model evaluated on successfully extracted fetal segments may face a different population from a system receiving all acquired recordings. The evaluation record should therefore retain extraction failures and exclusion criteria. An apparently precise reconstruction conditional on a usable input does not describe performance when that input cannot be obtained reliably.

2. The envelope is a constructed measurement

The target should be documented before the model architecture is compared. Relevant fields include the Doppler acquisition configuration, envelope-generation procedure, sampling rate, normalization, filtering, and segment boundaries. A generated signal can reproduce properties introduced by these processing choices as well as properties of the underlying physiology. If the target is normalized separately within each segment, for example, amplitude agreement has a different meaning from agreement on an absolute instrument scale. This is a conditional example of a design choice, not an assertion that the motivating study used that particular normalization rule.

Synchronization similarly contributes to the observed error. A clock offset or a filtering delay can create a consistent discrepancy even when the model captures the principal signal pattern. Conversely, aligning every predicted segment to its reference after generation can remove timing errors that a real-time application would experience. An evaluation should distinguish prespecified acquisition alignment from target-dependent postprocessing. If both aligned and unaligned scores are useful, they can be reported separately. Their difference then becomes evidence about timing rather than an invisible step that makes a reconstruction look more faithful.

Acquisition conditions can also affect the meaning of a segment. Motion, electrode contact, transducer placement, and recording quality are not interchangeable nuisance variables. Some primarily affect the input, others the target, and some can affect both. A useful quality annotation should identify the affected channel and the reason for exclusion or uncertainty. A single global quality label can conceal whether a model failed to translate informative input or was asked to reproduce a corrupted reference. Independent quality review on a subset would help establish this distinction without using the model's own reconstruction error as the definition of a poor recording.

3. Reconstruction metrics answer different questions

The motivating study reports a power-spectral-density mean squared error of 49.9 plus or minus 15.8 square decibels and a heart-rate error of 4.71 plus or minus 0.77 beats per minute. These are reported study results, not estimates recalculated here. The two metrics examine different properties. A spectral comparison emphasizes the distribution of power across frequency, while a heart-rate comparison summarizes periodicity. Neither necessarily establishes agreement in local timing, envelope shape, or an individual beat's mechanical features. A small improvement in one measure should not be presented as an improvement in every physiologically relevant aspect of the waveform.

Spectral estimation itself needs a specification. Welch's method formalizes estimating power spectra through averaging modified periodograms (Welch, 1967). Window length, overlap, frequency resolution, detrending, and conversion to logarithmic units can influence a spectral comparison. A reproducible study should report these settings and whether predictions and references undergo identical processing. Low error after heavy smoothing may describe agreement in a coarse frequency representation rather than faithful reconstruction of short events. The appropriate interpretation follows from the estimator and time scale, not from the familiar name of the metric.

Agreement analysis provides a complementary perspective when the output is a derived scalar measured on comparable units. Bland and Altman distinguish agreement from correlation in comparing measurement methods (Bland and Altman, 1986). For a proposed derived quantity, plotting signed differences against measurement magnitude can reveal bias and changing variability that a correlation obscures. Extending this idea to repeated fetal observations requires respecting within-pregnancy dependence. It also requires choosing a genuinely comparable quantity. An electrical feature and a

Doppler feature should not be placed in a method-agreement analysis merely because both originate from cardiovascular recordings.

A layered outcome set would separate segment-level spectral error, beat-timing error, shape error under a declared alignment, and agreement in prespecified derived measurements. The primary outcome should correspond to the intended scientific question. Secondary outcomes can then explain failure modes rather than provide alternative opportunities to declare success. Empty or unusable segments should remain visible in coverage reporting. A model that reconstructs only a selected subset can have low average error while leaving the most difficult acquisition conditions unaddressed. The reported denominator must identify which of these interpretations the score represents.

4. Pregnancy-level transfer and repeated segments

The number of segments measures available signal material, while the number of pregnancies measures a different source of variation. Segments from one pregnancy may share anatomy, equipment, acquisition session, and stable signal characteristics. Randomly assigning adjacent or overlapping segments to training and testing can therefore estimate performance under substantial shared information. A claim about a previously unseen pregnancy requires a split that withholds that pregnancy throughout the pipeline. The available study-level counts alone do not establish the split used; this article does not assume that the motivating study committed a particular leakage error.

Structured cross-validation guidance supports matching the withholding unit to the intended generalization target (Roberts et al., 2017). In this setting, separate targets could include a new segment from the same session, a later recording from the same pregnancy, and a new pregnancy from another acquisition setting. These targets should be named explicitly. Their results need not follow a simple ranking because the distribution of signal quality and physiological states can differ across the held-out sets. Reporting only one pooled score would make it difficult to determine which type of transfer has actually been evaluated.

All learned transformations belong inside the relevant training boundary. Normalization statistics, quality thresholds selected from data, alignment models, and feature selection can carry information from a held-out pregnancy even when the final network has not seen its labeled segments. Hyperparameter selection requires a separate development boundary. Work on model-selection bias shows why choosing settings on the final cross-validation result can make the estimated performance optimistic (Varma and Simon, 2006). A nested pregnancy-level procedure is one possible design, although the limited number of independent pregnancies can make its variability substantial. That uncertainty should be reported rather than hidden through repeated reuse of the evaluation set.

Repeated measurements also change the analysis of error. A mixed-effects model can represent a participant-level component alongside segment-level variation (Laird and Ware, 1982). A prospective error analysis might relate reconstruction error to recording quality while allowing a pregnancy-specific intercept. That model would not make the pregnancies representative of an external clinical population or fix every form of dependence. It would state a covariance structure that can be inspected and challenged. No mixed model is fitted in this article, and no confidence interval is invented from the segment and pregnancy counts alone.

5. Maternal input and the interpretation of attention

Adding maternal ECG may improve prediction for several reasons. It could provide information related to shared physiological timing, recording conditions, or other correlates of the target. A comparison with and without that input measures its predictive contribution under the specified training and evaluation procedure. It does not isolate a causal effect of maternal cardiovascular activity on a fetal waveform. The distinction becomes especially important if channels share acquisition artifacts or preprocessing. An ablation can change the information available to a model without corresponding to a biological intervention that selectively removes one pathway.

Attention weights describe an internal computational operation. Research on attention and explanation in NLP challenges treating those weights as a sufficient explanation of a prediction (Jain and Wallace, 2019). The relevant lesson here is methodological rather than a direct transfer of an NLP result to fetal physiology. A cross-modal attention map should not be interpreted automatically as a measure of physiological coupling. Testing its role would require comparisons such as controlled input perturbations, alternative attention settings, and reference measures of the relationship being claimed. The design must distinguish a useful computational diagnostic from evidence about a biological mechanism.

Several proposed controls could help separate competing explanations. A deliberately mismatched maternal channel could test whether subject-specific information matters beyond a generic signal prior, provided the mismatch preserves an appropriate comparison distribution. Timing perturbations could examine sensitivity to synchronization. Quality-stratified analysis could identify a benefit concentrated in cleaner or noisier inputs. Each control changes more than one property unless carefully designed, so its interpretation should remain narrow. These are suggestions for future experiments. No mismatching, perturbation, or attention-intervention study was performed for this manuscript.

6. What remains in a residual

Let the recorded Doppler envelope be Y and the fitted prediction be $f(X)$, where X contains the permitted electrical inputs. The residual Y minus $f(X)$ contains everything not reproduced by this particular fitted function on this particular measurement scale. It can include information absent from X , noise in Y , errors in X , imperfect synchronization, finite-data estimation error, and limitations of the model class or training procedure. Calling the entire residual a purely mechanical contribution assigns all of those possibilities to one biological explanation. The prediction equation alone cannot justify that assignment.

A stronger interpretation would require evidence that the principal alternative contributions are small or have been separately characterized. Repeated reference measurements can help estimate aspects of measurement variability. Independent synchronization checks can constrain timing error. Comparisons among adequately trained model classes can probe some capacity limitations, although agreement among models does not prove that they have reached the recoverability limit. Additional physiological measurements could test whether residual features align with a specified mechanical process. The relevant goal is a decomposition supported by observations, rather than a residual renamed after the mechanism investigators hope to study.

Recoverability is also distribution-dependent. A feature can be predictable in a cohort because it covaries with available inputs, yet become less predictable when that relationship changes. Dataset-shift studies show why predictive uncertainty should be evaluated under changed conditions rather than assumed to transfer with average accuracy (Ovadia et al., 2019). Here, an external acquisition site or a different signal-quality distribution would provide a meaningful challenge if it matches the intended use. A

reconstruction method should retain uncertainty about both its output and the applicability of the relationships learned from the original cohort.

7. Coverage, deferral, and a testable reporting record

A translation system may need to decline when the electrical input is inadequate or the requested interpretation exceeds the available evidence. Selective prediction offers a language for evaluating the tradeoff between the fraction of answered cases and the error on those cases (Geifman and El-Yaniv, 2017). Applying that idea to waveform generation requires defining an accepted segment and an acceptable error event; a classification result does not automatically provide a guarantee for continuous waveforms. Reporting coverage across acquisition conditions would help reveal whether low error is achieved by excluding a practically important subset.

The minimum record proposed here includes synchronized input and target definitions, preprocessing, pregnancy and session identifiers, split manifests, tuning boundaries, exclusion reasons, metric code, and uncertainty at the appropriate sampling level. It also distinguishes reported source results from new calculations and proposed tests. For a physiological residual claim, the record should name the competing measurement and modeling explanations and identify what observation would discriminate among them. This format makes the unresolved scientific question visible instead of allowing a visually convincing generated waveform to stand in for a mechanistic demonstration.

Cross-modal fetal signal translation is valuable as a way to study conditional predictability and to formulate better questions about complementary measurements. The evidence becomes stronger when reconstruction quality, transfer, and physiological attribution are evaluated as separate claims. The motivating framework supplies a concrete architecture and initial study setting for that program. Further progress depends on preserving the measurement chain and testing the missing links, especially those between a residual waveform and a specific biological explanation. This approach gives both successful reconstructions and informative failures a clear scientific role.

References

- Su, Tongli, et al. "Cross-Modal Generative Framework for Signal Translation from Fetal-Maternal Electrocardiograms to Fetal Doppler Waveforms." arXiv preprint arXiv:2607.08073 (2026).
- Sameni, Reza, and Gari D. Clifford. "A Review of Fetal ECG Signal Processing; Issues and Promising Directions." *The Open Pacing, Electrophysiology & Therapy Journal*, vol. 3, 2010, pp. 4-20. <https://doi.org/10.2174/1876536X01003010004>.
- Welch, Peter D. "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging over Short, Modified Periodograms." *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, 1967, pp. 70-73. <https://doi.org/10.1109/TAU.1967.1161901>.
- Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. [https://doi.org/10.1016/S0140-6736\(86\)90837-8](https://doi.org/10.1016/S0140-6736(86)90837-8).
- Jain, Sarthak, and Byron C. Wallace. "Attention is not Explanation." *Proceedings of NAACL-HLT*, 2019, pp. 3543-3556. <https://doi.org/10.18653/v1/N19-1357>.
- Roberts, David R., et al. "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure." *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. <https://doi.org/10.1111/ecog.02881>.
- Varma, Sudhir and Richard Simon. "Bias in error estimation when using cross-validation for model selection." *BMC Bioinformatics*, vol. 7, no. 1, 2006. <https://doi.org/10.1186/1471-2105-7-91>.

- Laird, Nan M. and James H. Ware. "Random-Effects Models for Longitudinal Data." *Biometrics*, vol. 38, no. 4, 1982, pp. 963. <https://doi.org/10.2307/2529876>.
- Ovadia, Yaniv, et al. "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift." *Advances in Neural Information Processing Systems*, vol. 32, 2019. <https://arxiv.org/abs/1906.02530>.
- 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>.