Resp-GRASP: Clinically-Grounded Reasoning for Respiratory Signal Interpretation via Guardrailed RAG with Personalized Baselines

Excel Widjaja [1 *] **Celina Li** [1] **Josh Chatterjee** [1 °] **Arshia Sharda** [1 °] **Kiran Nijjer** [2 †]

Abstract

Although continuous monitoring allows early identification of physiological deterioration five days prior to actual symptomatic COPD events, there exist two major limitations: predictions in a "black box" format which cannot be reviewed by clinicians, and population-based threshold settings resulting in false positives for abnormal measurements. This paper introduces Resp-GRASP, a six-step guardrailed reasoning pipeline that solves both limitations using double-layer z-score hierarchy that favors personal trends over population statistics, as well as MedGemma-27B that generates clinical reports while citing evidence-based ATS/ERS guidelines. Resp-GRASP achieves selective prediction in a multi-step setting, aborting 18.9% of recordings with low signal discriminability to ensure reliable inference. The primary finding of this study is the six-fold reduction in false positive rates ($15.8\% \rightarrow 2.6\%$) while nearly doubling sensitivity ($53.3\% \rightarrow 93.3\%$) with McNemar test $p = 0.031$ in comparison with approaches based on population thresholds. We introduce Semantic Directional CRC (SD-CRC), a step-level reasoning verification metric that checks whether each step's clinical narrative directionally agrees with its quantitative z-score evidence after stripping retrieved passages to prevent evidence contamination from inflating scores. Resp-GRASP has an accuracy of 85.7% SD-CRC ($p < 0.0001$).

[*]Lead Author [°]Equal Contribution [†]Senior Author [1]Independent [2]Stanford University. Correspondence to: Excel Widjaja <2citiesmen@gmail.com>, Celina Li <celina.li374@gmail.com>.

Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

About 16 million people in the USA have Chronic Obstructive Pulmonary Disease (COPD). Every year, the exacerbation of this disorder requires expenditures for treatment higher than $32 billion (Ford et al., 2015). Patients with three or more severe exacerbations face a fourfold increase in total mortality (HR 4.13, 95% CI 1.80–9.41) compared to those without exacerbations (Soler-Cataluña et al., 2005). The signs of the deterioration can manifest themselves in about 2 to 5 days (Wilkinson et al., 2004) before the first symptoms.

Despite these opportunities for early intervention, current monitoring systems remain unreliable because they treat prediction as a one-shot, black-box decision: the pipelines are not aligned with underlying signal quality or uncertainty, do not perform abstention mechanisms under low signal quality, and are based on population-level thresholds that ignore patient-specific variations. Together, this often leads to high false positive rates. Explainability has been found by clinicians to be one of the basic requirements for trust in machine learning predictions (Tonekaboni et al., 2019), yet existing models provide neither interpretable reasoning nor mechanisms to verify whether their outputs are actually supported by the underlying physiological data. In continuous physiological monitoring, these challenges are especially relevant due to noisy data, prompt or context sensitive variability, and the need for sequential evidence aggregation and stopping rules.

To address these limitations, Resp-GRASP implements a guardrailed pipeline that performs stepwise selective prediction, aborting inference when signal quality is insufficient rather than forcing unreliable outputs. This abstention mechanism employs selective prediction in a real-time monitoring setting, where coverage should be traded off with reliability. The six-step guardrailed structure and signal-quality-conditioned abort criteria address the need for sequential evidence aggregation and principled stopping rules essential to continuous physiological monitoring. Moreover, Resp-GRASP also uses a dual hierarchy of z-scores, which combines population-level statistics with personalized baselines in order to provide a principled approach to distribution shift and personalization, without needing

held-out calibration data. Finally, this work leverages the guardrailed RAG structure provided by C-GRASP (Cheng et al., 2026) to achieve 69.6% of CRC for predicting the heart rate variability.

Resp-GRASP's main contributions include (1) a dual z-score hierarchy that enforces patient-adaptive normalization with formal conflict resolution, creating an uncertainty-aware normalization scheme under distribution changes between population and patient data; (2) the SD-CRC mechanism with evidence stripping to ensure that AI can correctly interpret biological signals, while simultaneously solving evidence contamination issues within RAG auditing through a verification process for multi-stage pipelines; and (3) a breath-by-breath approach that generates formatted reports according to ATS/ERS and NEWS2 guidelines, supported by a selective prediction pipeline that utilizes stepwise abstention and uncertainty propagation.

2. Related Work

2.1. Clinical Language Models and Physiological RAG

The clinical language models Med-PaLM 2 (Singhal et al., 2023), Meditron (Chen et al., 2023), and MedGemma (Sellergren et al., 2025), which create accurate clinical texts, do work on the basis of narrative input rather than physiological input signals and provide single-pass output that lacks auditable reasoning steps. C-GRASP (Cheng et al., 2026) established a guardrailed RAG pipeline for HRV interpretation with dual z-score normalization and an eight-step reasoning structure. However, C-GRASP's design is specific to cardiac autonomic signals, lacks a formal rule for resolving baseline conflicts, and does not address evidence contamination. Resp-GRASP extends this architecture by introducing respiratory-specific artifact detectors, a formal dual z-score hierarchy, and SD-CRC with evidence stripping.

2.2. Selective Prediction and Uncertainty in Agentic Pipelines

Selective Prediction proves that reliable algorithms must abstain rather than output unreliable predictions (Geifman & El-Yaniv, 2017), with conformal risk control extending this notion to generation (Angelopoulos & Bates, 2022). Resp-GRASP employs the idea in a clinical setting by halting the process whenever $S \geq 0.7$, i.e., if the model predicts an incorrect clinical report with confidence, then the system will simply refuse to give any response at all. While abstention in the conformal setup is based on held-out calibration sets, Resp-GRASP does not rely on any such set, because the real-world application domain does not afford the luxury of exchangeability and the possibility of distribution shift between patients is a reality. In addition to being a selective predictor, Resp-GRASP also handles compositional uncertainty in the pipeline through cumulative confidence penalty and SD-CRC, both of which are currently unavailable within any framework designed for RAG evaluation.

3. Methods

Resp-GRASP operates as a three-stage pipeline: (A) Feature Extraction, (B) RAG-Enhanced Reasoning, and (C) Evaluation (Figure 1).

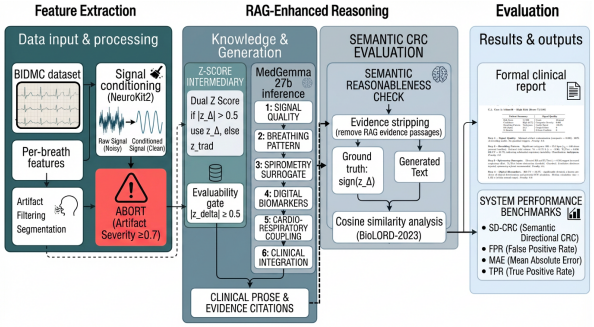

Figure 1. Resp-GRASP three-stage pipeline: (A) Feature Extraction with NeuroKit2 segmentation and four artifact detectors, (B) RAG-Enhanced Reasoning over six guardrailed steps using MedGemma-27B grounded in 395 RAG chunks, and (C) Evaluation via SD-CRC with evidence stripping.

3.1. Feature Extraction and Artifact Detection

NeuroKit2 (Makowski et al., 2021) segments 8-minute impedance pneumography recordings (125 Hz) into individual breaths via bandpass filtering (0.1–1.0 Hz), zero-crossing detection, and peak identification, yielding 80–180 per-breath feature vectors: {Ti, Te, Ttot, Vt, Ti/Ttot, Te/Ttot, PIF, PEF, RR}. Four artifact detectors target cough (amplitude $> 3\sigma$, duration < 500 ms), speech (zero-crossing rate > 5 Hz), movement (windowed variance $> 2\times$ mean), and sensor failure (dropout, saturation, flatline). Severity scores combine as $S = 0.4s_{\text{cough}} + 0.3s_{\text{speech}} + 0.2s_{\text{move}} + 0.1s_{\text{sensor}}$, mapping to four grades (minimal/mild/moderate/severe). Severe grades abort the pipeline.

This abort mechanism accomplishes selective prediction: the system abstains from clinical report generation when signal quality is insufficient to support reliable inference, preferring a null output over a confidently wrong one. Of 53 recordings, 10 (18.9%) were aborted at Step 1. Remaining recordings receive graded confidence penalties proportional to artifact severity, enabling uncertainty-calibrated output for moderate-quality signals rather than forcing a narrow pass/fail decision.

3.2. Dual Z-Score Baseline Hierarchy

For each feature x, we compute a population z-score against published norms (e.g., RR: $\mu = 18.0, \sigma = 3.0$ bpm) (Charlton et al., 2021) and a patient-specific z-score against the patient's own baseline (minutes 0–5):

$$z_{\text{trad}} = \frac{x - \mu_{\text{pop}}}{\sigma_{\text{pop}}}, \qquad z_{\Delta} = \frac{x - \mu_{\text{patient}}}{\sigma_{\text{patient}}} \tag{1}$$

A patient with resting RR of 20 bpm who increases to 22 bpm shows a small z_{trad} (22 is near the population mean) but a meaningful z_{Δ} (elevated relative to their personal baseline of 20). When the two disagree in direction, a priority rule resolves the conflict:

$$z_{\text{primary}}(x) = \begin{cases} z_{\Delta}(x) & \text{if } |z_{\Delta}| > 0.5 \\ z_{\text{trad}}(x) & \text{otherwise} \end{cases} \tag{2}$$

3.3. Six-Step Guardrailed Reasoning Pipeline

MedGemma-27B-Thinking (4-bit quantization) generates clinical narratives grounded in 395 RAG chunks drawn from a fixed library of 14 publicly available sources spanning respiratory monitoring guidelines, artifact detection, clinical deterioration prediction, and cardiorespiratory physiology, embedded via BioLORD-2023 (Remy et al., 2024). The output is a structured JSON report containing artifact breakdown, z-scores with hierarchy resolution, risk score, confidence, narrative with citations, recommendations, and limitation disclosures.

Table 1. Six-step reasoning pipeline with guardrail correspondences.

Step	Name	Clinical Task	Guardrail
1	Signal Quality	Grade artifacts, usable data	Abort $S \geq 0.7$
2	Breathing Pattern	Classify eupnea/tachypnea	—
3	Spirometry Surr.	Te/Ttot flow limitation proxy	-0.1 conf.
4	Digital Biomarkers	RR-CV, exhalation time	—
5	Cardio-Resp.	HR–RR correlation, RSA	—
6	Clinical Integr.	NEWS2 risk, confidence, recs.	Agg. penalties

3.4. Semantic Directional CRC (SD-CRC)

SD-CRC verifies that each step's narrative agrees with its quantitative evidence: a strongly positive z_{Δ} should yield language describing tachypnea, not bradypnea. SD-CRC uses embedding-based similarity and strips evidence before scoring.

Evidence stripping isolates model reasoning R by removing retrieved RAG passages from the generated narrative via exact substring lookup prior to embedding. This is done because without stripping, retrieved passages would inflate faithfulness scores artificially: retrieval is relevance-based, and retrieved text already directionally agrees with ground truth regardless of whether the model genuinely reasoned from patient data.

For each evaluable pair (m, j) with $|z_{\Delta}| \geq 0.5$, we strip retrieved passages from the narrative to isolate model reasoning R, embed via BioLORD-2023 (ϕ : text $\rightarrow \mathbb{R}^{768}$), and compare cosine similarity against positive and negative reference phrase sets P_m^+, P_m^-:

$$\text{sim}^{\pm} = \max_{p \in P_m^{\pm}} \cos\big(\phi(R), \phi(p)\big) \tag{3}$$

$$\text{consistent}(m, j) = \mathbb{1}\big[\text{sign}(\text{sim}^+ - \text{sim}^-) = \text{sign}(z_{\Delta}(m, j))\big] \tag{4}$$

$$\text{SD-CRC} = \frac{1}{|\Omega|} \sum_{(m,j) \in \Omega} \text{consistent}(m, j) \tag{5}$$

Table 2. SD-CRC directional reference phrases. † = excluded from BIDMC evaluation due to insufficient within-subject variability.

Feature	Positive (↑)	Negative (↓)
`rr`	tachypnea, rapid breathing	bradypnea, slow breathing
`vt`	deep breathing, hyperpnea	shallow breathing, restrictive
`rr_cv`	irregular, high variability	regular, stable pattern
`ti_ttot`†	prolonged inspiration	shortened inspiration
`te_ttot`†	air trapping, flow limitation	normal expiratory flow

The evaluable set Ω is computed via a two-stage gate before any LLM generation:

$$\Omega = \Big\{(m, j) : \tfrac{1}{J} \sum_{j} \mathbb{1}[|z_{\Delta}| \geq \tau] \geq \alpha \wedge |z_{\Delta}(m, j)| \geq \tau\Big\} \tag{6}$$

with $\tau = 0.5$ and $\alpha = 0.35$. On BIDMC this retains respiratory rate, tidal volume, and RR-CV; Ti/Ttot and Te/Ttot are excluded due to insufficient within-subject variability ($< 4\%$ evaluable), a dataset rather than design constraint.

3.5. Mechanistic Probing

To distinguish comprehension from decoding failures, we extract hidden states $\mathbf{h}_l \in \mathbb{R}^d$ from MedGemma-27B at selected layers and train three linear probes (5-fold CV): (1) logistic regression predicting CRC correctness, (2) ridge regression predicting $|z_{\Delta}|$, and (3) logistic regression predicting z-score direction on balanced feature subsets.

3.6. Cohort

We evaluate on the BIDMC dataset (Pimentel et al., 2017), comprising 53 adult ICU patients (ages 19–90+, median age

64) with 8-minute simultaneous impedance pneumography, ECG, and PPG recordings sampled at 125 Hz, and expert-annotated ground truth ($\kappa = 0.92$). Of 53 recordings, 43 completed the full pipeline; the remaining 10 were aborted at Step 1 due to severe artifact contamination. No model training is performed. This evaluation therefore establishes proof-of-concept feasibility rather than predictive clinical validity; prospective validation on longitudinal COPD cohorts with labeled exacerbation events is required before clinical deployment.

4. Results

4.1. Signal Processing Accuracy

Resp-GRASP achieves MAE $= 0.435$ bpm vs. NeuroKit2's 0.364 bpm ($N = 43$). The 0.071 bpm overhead is clinically negligible (Figure 2).

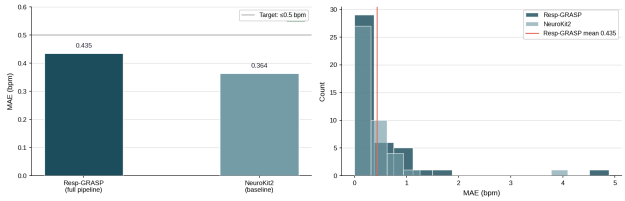

Figure 2. Respiratory rate MAE comparison (left); Per-record error distribution (right).

4.2. Personalized Baselines

The personalized baseline hierarchy is Resp-GRASP's strongest empirical result. The hierarchical strategy recovers sensitivity from population baselines ($53.3\% \rightarrow 93.3\%$, McNemar $p = 0.031$) with no cold-start requirement, representing the deployable contribution of this work. Once sufficient patient monitoring history is established, the patient-only strategy demonstrates the ceiling this architecture can reach: FPR drops six-fold ($15.8\% \rightarrow 2.6\%$) alongside the same sensitivity gain, achieving MCC $= +0.907$. The hierarchy provides the deployable middle ground between these two operating points (Figure 3).

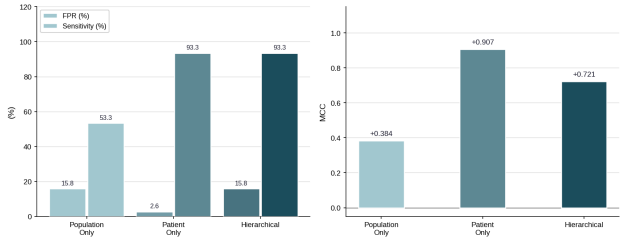

Figure 3. FPR and sensitivity by strategy (left), Matthews Correlation Coefficient (right). Hierarchical strategy recovers sensitivity ($53.3\% \rightarrow 93.3\%$, McNemar $p = 0.031$) while patient-only achieves the lowest FPR (2.6%).

4.3. Reasoning Consistency

85.7% SD-CRC ($t = 8.74$, $p < 0.0001$): roughly 6 of every 7 evaluable feature-patient pairs show a clinical narrative that directionally matches the quantitative z-score evidence—when the system says "tachypnea," the underlying RR z-score is positive; when it says "shallow breathing," tidal volume is reduced. Per-feature: respiratory_rate 100% ($n = 14$), tidal_volume 92% ($n = 26$), rr_cv 75% ($n = 28$). Number fidelity was 89.8% (415/462). Binning by $|z_\Delta|$ (Figure 4) reveals a step-function: 77% at 0.5–1.0 ($n = 30$), 95% at 1.0–2.0 ($n = 19$), 94% at 2.0–5.0 ($n = 16$), 100% at > 5.0 ($n = 3$).

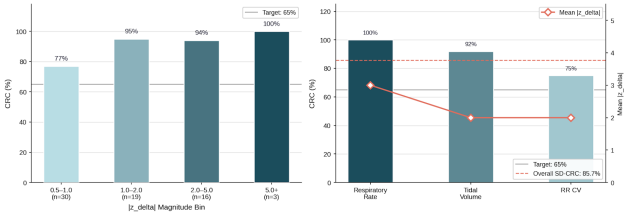

Figure 4. CRC accuracy by z-score magnitude (left). Per-feature CRC vs. mean $|z_\Delta|$ (right). Reasoning failures concentrate at low discriminability ($|z_\Delta| < 1.0$), a clinically safer error profile than uniform inaccuracy.

4.4. Mechanistic Probing

Probe 1 (self-uncertainty): 83.7% at layer 40, below the 86.8% majority-class baseline. Probe 2 (discriminability): $R^2 = 0.342$ at layer 16. Probe 3 (direction): 91.7% for respiratory_rate, 100% for tidal_volume. Evidently, the model encodes both magnitude and direction internally but fails to consistently translate them into output text (Figure 5).

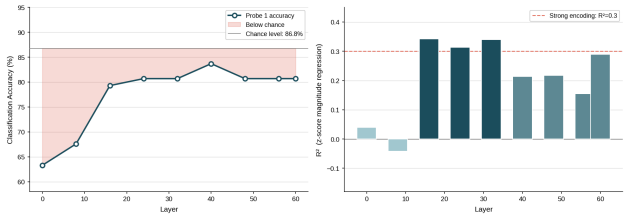

Figure 5. Mechanistic probing: self-uncertainty (left) and discriminability encoding (right) across layers.

4.5. Ablation

Table 3 shows component contributions (small sample sizes; directional trends only). Removing the hierarchy causes the largest CRC drop ($86.5\% \rightarrow 80.6\%$), confirming it as the most important component for reasoning alignment. Removing guardrails inflates confidence to 0.850, the base maximum before penalties are applied.

Table 3. Ablation study. n = eligible pairs for CRC evaluation.

Config	CRC%	n	Fidel.%	MCC	Conf.
Full System	86.5	16	90.3	+0.707	0.712
w/o RAG	88.9	9	90.4	+0.707	0.708
w/o Guardrails	85.9	13	90.6	+0.714	0.850
w/o Hierarchy	80.6	12	89.7	+0.707	0.712
w/o Artifacts	91.1	15	89.1	+0.601	0.750

5. Discussion

5.1. Guardrail Aborts as a Form of Selective Prediction

Of the 53 recordings, ten (18.9%) were aborted due to heavy artifacts by the Step 1 guardrail. The abort acts as a safety constraint in clinical AI applications, where a false confident decision may mislead treatment, whereas an aborted decision triggers further investigation. Overall, in the 43 recordings left after the abort, the mean interpretable signal per recording was 67.2% with a graded severity distribution (13 minimal, 16 mild, 14 moderate). This quality-based information representation leads to more accurate decision-making performance via confidence estimation.

The same behavior can also be explained by selective prediction. Instead of unconditional predictions, Resp-GRASP employs a structured abstention mechanism, which is implemented through guardrails, confidence penalties, and abort conditions in the pipeline. The abstention policy, which did not generate any outputs in 18.9% of aborted recordings and thus avoided unreliable predictions, instantiates the concept of selective prediction in a clinical monitoring setting. The coverage cost (10 aborted recordings) is the price of this improved reliability: a rational trade-off of abstention and outcome faithfulness that population-threshold mechanisms often fail to describe.

5.2. Personalization Decreases False Positive Rate

The move from population thresholds to patient baselines decreases the FPR from 15.8% to 2.6%. This reduction is especially significant because of alert fatigue, a well-known factor preventing continuous monitoring from being implemented in clinical practice (Ancker et al., 2017).

In addition to improving sensitivity, which rises from 53.3% to 93.3% (McNemar $p = 0.031$), the hierarchy's ability to overcome the limitations of patient baselines, which miss nearly half of true deteriorations because low-baseline patients (12–14 bpm) can relatively increase while remaining within norms, may be important for practical applications. By aligning alerts with patient-specific physiology, Resp-GRASP reduces unnecessary disruptions and increases the likelihood that clinicians will act on the system's outputs.

5.3. Failures in Reasoning May Be Clinically Appropriate

The discriminability-CRC step function produced 77% for $|z|$ between 0.5 and 1.0 versus 95–100% for $|z|$ greater than 1.0, implying that the most critical points for the algorithm correspond to the points with larger deviations, such as significant tachypnea and large decreases in tidal volume. Thus, Resp-GRASP tends to make mistakes at smaller z-scores: at points close to the baseline that do not trigger immediate attention.

It is evident that such a non-linear dependency cannot be described by any correlation measures ($r = 0.179$). As it becomes apparent through magnitude binning, this indicates that reasoning accuracy depends on the thresholds rather than a linear function of the signal strength. In the case of an 85% accurate system, 15% of its errors will be evenly distributed across all ranges of severity levels, including deviations with large values of z that warrant urgent intervention. Instead, Resp-GRASP tends to make mistakes at smaller z-scores, indicating that they are made at points close to baseline and do not trigger immediate attention. This is a clinically safer error profile than uniform accuracy, as most errors occur in the low-discriminability cases in which it would be reasonable not to have an intervention and to rely more on clinical reasoning.

5.4. The Decoding-Comprehension Dissociation

MedGemma encodes both the size and direction of the z-score ($R^2 = 0.342$; 91.7–100%, respectively), but it achieves only a CRC of 85.7% in output. This dissociation resembles general findings in LLM numeracy, where models linearly encode the magnitude but fail to reason correctly (Zhu et al., 2025). In such cases, errors arise not due to the limitations of the representational ability of the network, but rather to decoding instability, where internally consistent representations are not reliably expressed in the generated text.

Uncovering this dissociation leads to practical implications for hallucination mitigation strategies that would not involve costly measures, such as retraining MedGemma. Instead, a light-weight classifier can be used to find any potential directional inconsistencies using layer-16 representation prior to the completion of the generation, thus enabling re-sampling or output verification. Therefore, constrained decoding using hidden state probes is a realistic way of closing the gap between the output consistency of 85.7% and the already encoded 91.7–100% direction accuracy.

5.5. Evidence Contamination in RAG Evaluation

The evidence stripping of SD-CRC resolves a general issue. In RAG systems that use multi-step reasoning, the passages

retrieved are those that agree with the context, such that their directional language agrees with the ground truth irrespective of the system's reasoning. Without stripping, such systems that repeat evidence verbatim would achieve near perfect scores. This contamination is not addressed by RAGAS (Es et al., 2024), ARES (Saad-Falcon et al., 2024), and RAGChecker (Ru et al., 2024). Resp-GRASP breaks down reasoning into auditable steps and evaluates each step's reasoning after stripping evidence, ensuring that the scoring reflects a reasoning task utilizing patient-specific information rather than the replication of supporting text. In addition to clinical applications, evidence stripped SD-CRC represents a generalizable testing strategy for any multi-step agentic process where contextual information retrieved might contaminate faithfulness scoring. This is an important omission from the current benchmarking strategies for reliable agentic behavior beyond respiratory health.

5.6. Ablation Results

Ablation results should be interpreted cautiously given small sample sizes (9–16 eligible pairs per configuration), but the hierarchy seems to be the most important component: removing it produces the largest CRC drop ($86.5\% \rightarrow 80.6\%$) and is the single intervention that degrades end-to-end reasoning quality the most. This suggests that the hierarchy's primary contribution is stabilizing the alignment between feature deviations and downstream reasoning rather than improving classification thresholds. Removing artifact detection produces the largest MCC drop ($+0.707 \rightarrow +0.601$), confirming that artifact severity grades contribute meaningfully to clinical risk assessment. Removing guardrails inflates mean confidence to 0.850 (the base maximum before penalties), confirming that guardrails function as a calibration mechanism. Without it, the pipeline reports maximum confidence regardless of signal quality, a failure mode similar to overconfident models that lack proper uncertainty reporting. The remaining ablations (w/o RAG, w/o artifacts on CRC) were too underpowered to support further interpretive claims.

5.7. Limitations

Proof of concept only. BIDMC lacks longitudinal exacerbation outcome labels, so performance claims reflect internal consistency rather than predictive validity. **Small dataset.** 53 recordings limit statistical power, particularly for ablation (9–16 pairs) and probing (68 pairs, 9 incorrect). **RAG corpus gaps.** Steps 2–4 retrieve below the relevance threshold, suggesting incomplete evidence integration for intermediate reasoning steps. **Feature exclusion.** Ti/Ttot and Te/Ttot were excluded from SD-CRC due to insufficient within-subject variability on BIDMC (3.8% evaluable).

6. Conclusion

Resp-GRASP's core contribution is showing that personalized baselines plus guardrailed stepwise reasoning produce a respiratory monitoring system that is simultaneously more sensitive (53.3% $\rightarrow$ 93.3%) and far less noisy (FPR 15.8% $\rightarrow$ 2.6%) than population-threshold approaches, while generating clinical reports that clinicians can actually audit step-by-step. SD-CRC with evidence stripping provides a tool for honestly evaluating whether such systems reason from their data or merely echo their citations. Mechanistic probing shows that remaining reasoning failures may be addressable through constrained decoding rather than model retraining. As a proof of concept on real ICU respiratory physiology, these findings motivate but do not replace prospective validation on longitudinal COPD cohorts with exacerbation outcome labels.

Impact Statement

Resp-GRASP significantly reduces false-positive alerts and enables earlier intervention, potentially reducing hospital costs. Ethically, it replaces "black-box" predictions with transparent, documented reasoning, resulting in reduced clinician alert fatigue and ensuring that medical interventions are grounded in personalized data rather than often misleading population averages. The guardrail mechanisms establishes a concrete precedent for selective abstention in clinical AI: systems that decline to output rather than output unreliably.

References

Ancker, J. S., Edwards, A., Nosal, S., Hauser, D., Mauer, E., and Kaushal, R. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Medical Informatics and Decision Making*, 17(1):36, 2017.

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2022.

Charlton, P. H., Bonnici, T., Tarassenko, L., Clifton, D. A., Beale, R., Watkinson, P. J., and Alastruey, J. An impedance pneumography signal quality index: Design, assessment and application to respiratory rate monitoring. *Biomedical Signal Processing and Control*, 65:102339, 2021.

Chen, Z., Hernandez-Cano, A., Romanou, A., Bonber, A., Munoz-Ferrandis, K., Estrella, R., et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

Cheng, C.-L., Lin, T.-C., and Chang, C.-K. C-grasp:

Clinically-grounded reasoning for affective signal processing. *arXiv preprint arXiv:2601.10342*, 2026.

Es, S., James, J., Espinosa-Anke, L., and Schockaert, S. Ragas: Automated evaluation of retrieval augmented generation. In *EACL*, 2024.

Ford, E. S., Murphy, L. B., Khavjou, O., Giles, W. H., Holt, J. B., and Croft, J. B. Total and state-specific medical and absenteeism costs of copd among adults aged 18 years and older in the united states for 2010 and projections through 2020. *Chest*, 147(1):31–45, 2015.

Geifman, Y. and El-Yaniv, R. Selective prediction in deep neural networks. In *NeurIPS*, 2017.

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., and Chen, S. H. A. Neurokit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53 (4):1689–1696, 2021.

Pimentel, M. A. F., Johnson, A. E. W., Charlton, P. H., Birrenkott, D., Watkinson, P. J., Tarassenko, L., and Clifton, D. A. Toward a robust estimation of respiratory rate from pulse oximeters. *IEEE Transactions on Biomedical Engineering*, 64(8):1914–1923, 2017.

Remy, F., Demuynck, K., and Demeester, T. Biolord-2023: Semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, 31(9): 1844–1855, 2024.

Ru, D. et al. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *NeurIPS Datasets and Benchmarks Track*, 2024.

Saad-Falcon, J., Khattab, O., Potts, C., and Zaharia, M. Ares: An automated evaluation framework for retrieval-augmented generation systems. In *NAACL*, 2024.

Sellergren, A. et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Singhal, K. et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

Soler-Cataluña, J. J., Martínez-García, M. Á., Román Sánchez, P., Salcedo, E., Navarro, M., and Ochando, R. Severe acute exacerbations and mortality in patients with chronic obstructive pulmonary disease. *Thorax*, 60(11):925–931, 2005.

Tonekaboni, S., Joshi, S., McCradden, M. D., and Goldenberg, A. What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Proceedings of the 4th Machine Learning for Healthcare Conference (MLHC)*, volume 106 of *PMLR*, pp. 359–380, 2019.

Wilkinson, T. M. A., Donaldson, G. C., Hurst, J. R., Seemungal, T. A. R., and Wedzicha, J. A. Early therapy improves outcomes of exacerbations of chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine*, 169(12):1298–1303, 2004.

Zhu, F., Dai, D., and Sui, Z. Language models encode the value of numbers linearly. In *COLING*, 2025.

A. RAG Corpus Sources

1. Celli, B. R. and MacNee, W. Standards for the diagnosis and treatment of patients with COPD: A summary of the ATS/ERS position paper. *European Respiratory Journal*, 23(6):932–946, 2004.

2. Global Initiative for Chronic Obstructive Lung Disease. Global Strategy for Prevention, Diagnosis and Management of COPD: 2024 Report. 2024.

3. Royal College of Physicians. National Early Warning Score (NEWS) 2: Standardising the Assessment of Acute-Illness Severity in the NHS. RCP, London, 2017.

4. Peter H. Charlton, Timothy Bonnici, Lionel Tarassenko, David A. Clifton, Richard Beale, Peter J. Watkinson, and Jordi Alastruey. An impedance pneumography signal quality index. *Biomedical Signal Processing and Control*, 65:102339, 2021.

5. Jonathan Moeyersons et al. Artefact detection in impedance pneumography signals using machine learning. *Sensors*, 21(8):2613, 2021.

6. Carlo Massaroni et al. Contact-based methods for measuring respiratory rate. *Sensors*, 19(4):908, 2019.

7. Andrea Nicolò et al. The importance of respiratory rate monitoring: From healthcare to sport and exercise. *Sensors*, 20(21):6396, 2020.

8. Peter H. Charlton et al. Breathing rate estimation from the electrocardiogram and photoplethysmogram: A review. *IEEE Reviews in Biomedical Engineering*, 11:2–20, 2018.

9. Samuel Bawua et al. A systematic review of methods used to measure respiratory rate in hospitalized patients. *Annals of Noninvasive Electrocardiology*, 26(5):e12885, 2021.

10. Daniel Garrido et al. Respiratory rate variability as a prognostic factor in hospitalized patients transferred to the ICU. *Cureus*, 10(1):e2100, 2018.

11. Kohei Mochizuki et al. Importance of respiratory rate for the prediction of clinical deterioration after emergency department discharge. *Acute Medicine & Surgery*, 4(2):172–178, 2017.

12. Ata Mahmoodpoor et al. Prognostic value of National Early Warning Score and Modified Early Warning Score on ICU readmission and mortality. *Frontiers in Medicine*, 9:938005, 2022.

13. Alfredo J. Garcia et al. Cardiorespiratory coupling in health and disease. *Autonomic Neuroscience*, 175(1–2):26–37, 2013.

14. Nicholas S. Hill et al. Respiratory monitoring in the ICU: A consensus of 16. *Critical Care*, 16(5):219, 2013.

B. Metrics Glossary

Table 4. Definition, scope, and interpretation of evaluation metrics reported in the main text.

Metric	Computed On	Interpretation	Notes
MAE (bpm)	43 complete recordings	Lower is better; ≤ 0.5 bpm is clinically negligible	Predicted vs. expert-annotated RR
SD-CRC (%)	Evaluable pairs Ω ($\lvert z_\Delta \rvert \geq 0.5$)	Higher is better; 85.7% means 6 of 7 pairs directionally consistent	Evidence stripped before scoring to isolate model reasoning
Number Fidelity (%)	462 value-mentions across 43 runs	Higher is better	Checks quoted numbers match computed values within 5% relative tolerance
FPR (%)	38 stable recordings	Lower is better; 15.8% $\approx$ 1-in-6, 2.6% $\approx$ 1-in-40	Fraction of stable patients incorrectly flagged
Sensitivity (%)	15 spiked recordings	Higher is better	Detection of synthetic RR spikes, not labeled clinical exacerbations
MCC	All 53 recordings	-1 to $+1$; 0 = chance; accounts for class imbalance	Preferred over accuracy given unequal class sizes ($N_{\text{stable}} = 38$, $N_{\text{spike}} = 15$)
McNemar p	Paired recording decisions	$p < 0.05$ indicates significantly different error pattern from population-only	Exact binomial test on paired classifiers
Confidence Score	43 complete runs	Base = 0.850; penalties reduce toward 0; mean full system = 0.712	Reflects signal reliability and surrogate use, not classification accuracy
R^2 (Probe 2)	68 probing pairs, layer 16	Higher = stronger internal encoding of discriminability	Ridge regression predicting $\lvert z_\Delta \rvert$ from hidden states

C. Representative Clinical Reports

We present six Resp-GRASP clinical reports spanning the full risk spectrum (0–71/100), selected to illustrate the system's behavior across signal quality grades, z-score conflict scenarios, breathing patterns, and confidence levels. Each report is generated end-to-end from an 8-minute BIDMC impedance pneumography recording with no manual editing. Reports are condensed from full JSON output for readability; complete outputs are available in the supplementary repository.

C.1. Case 1: bidmc06 – High Risk (Score 71/100)

Patient Summary		Signal Quality	
Risk Score	71/100	Grade	Minimal
Confidence	High (0.75)	Comp. Severity	0.000
Breathing Pattern	Tachypnea	Usable Signal	100.0%
RR (bpm)	25.2	Cough Events	0
N Breaths	151	Z-Score Conflicts	0

Step 1 – Signal Quality. Minimal artifact contamination (composite = 0.000). 100% of recording usable. No guardrail triggers. *Penalty: 0.0.*

Step 2 – Breathing Pattern. Significant tachypnea: RR = 25.2 bpm ($z_\Delta = 4.66$ above personal baseline). Reduced tidal volume: Vt = 0.171 L ($z = -3.90$). Ti/Ttot = 0.496. RR-CV = 33.7%, indicating substantial respiratory instability. Classification: **tachypnea**. *Penalty: 0.0.*

Step 3 – Spirometry Surrogate. Elevated RR and Ti/Ttot ($z = 0.58$) suggest increased respiratory effort. Te/Ttot below obstruction threshold. *Guardrail: Limitation disclosure injected; spirometry referral recommended. Penalty: 0.1.*

Step 4 – Digital Biomarkers. RR-CV = 33.7% – significantly elevated, a known predictor of clinical deterioration and potential ICU admission. Median exhalation time = 1.152 s (within normal range). *Penalty: 0.0.*

Step 5 – Cardio-Respiratory Coupling. HR–RR correlation $r = 0.024$: **diminished** coupling. Mean HR = 81.5 bpm. Diminished RSA may reflect autonomic dysregulation. *Penalty: 0.0.*

Step 6 – Clinical Integration. Risk score 71/100, driven by RR component of 3/3 NEWS2 points. Score exceeds threshold for urgent clinical response ($\geq$7). Supplemental oxygen adds 2 NEWS2 points.

Recommendations. (1) Urgent pulmonology consult within 24 hours. (2) Increase monitoring frequency to every 4 hours. (3) Review bronchodilator therapy. (4) Formal spirometry recommended.

Limitations. Spirometry surrogate used.

C.2. Case 2: bidmc07 – High Risk with Z-Score Conflicts (Score 62/100)

Patient Summary		Signal Quality	
Risk Score	62/100	Grade	Minimal
Confidence	High (0.75)	Comp. Severity	0.000
Breathing Pattern	Tachypnea	Usable Signal	100.0%
RR (bpm)	24.0	Cough Events	0
N Breaths	154	Z-Score Conflicts	2

Z-Score Conflict Detail.

- `ti_ttot`: $z_\Delta = -0.65$, $z_{\text{trad}} = +0.58 \rightarrow$ Hierarchy selects z_Δ (patient trend: decreasing)
- `te_ttot`: $z_\Delta = +0.65$, $z_{\text{trad}} = -0.58 \rightarrow$ Hierarchy selects z_Δ (patient trend: increasing)

Step 2 – Breathing Pattern. RR = 24.0 bpm ($z_\Delta = 8.45$). Vt = 0.221 ($z_\Delta = -6.00$). Tachypnea with markedly shallow respirations. RR-CV = 30.0%.

Step 5 – Cardio-Respiratory Coupling. $r = 0.107$: diminished. Mean HR = 90.4 bpm.

Step 6 – Clinical Integration. Risk 62/100. Despite conflicting z-scores for timing parameters, the overall assessment indicates significant respiratory compromise requiring immediate evaluation. NEWS2 threshold for urgent response exceeded.

Recommendations. Urgent pulmonology consult. Increase monitoring frequency. Review bronchodilator therapy. Formal spirometry.

C.3. Case 3: bidmc05 – Moderate Risk with Maximum Z-Score Conflicts (Score 40/100)

Patient Summary		Signal Quality	
Risk Score	40/100	Grade	Moderate
Confidence	Medium (0.60)	Comp. Severity	0.476
Breathing Pattern	Eupnea	Usable Signal	62.0%
RR (bpm)	12.4	Cough Events	63
N Breaths	78	Z-Score Conflicts	5

Z-Score Conflict Detail. This patient exhibits the highest conflict count in the dataset, demonstrating the hierarchy's conflict resolution in action:

- `respiratory_rate`: $z_\Delta = +0.57$, $z_{\text{trad}} = -1.12 \rightarrow z_\Delta$ selected ($|z_\Delta| > 0.5$)
- `ti`: $z_\Delta = -0.47$, $z_{\text{trad}} = +2.40 \rightarrow z_{\text{trad}}$ selected ($|z_\Delta| \leq 0.5$)
- `te`: $z_\Delta = -0.49$, $z_{\text{trad}} = +0.66 \rightarrow z_{\text{trad}}$ selected ($|z_\Delta| \leq 0.5$)
- `ti_ttot`: $z_\Delta = -0.39$, $z_{\text{trad}} = +0.53 \rightarrow z_{\text{trad}}$ selected ($|z_\Delta| \leq 0.5$)
- `te_ttot`: $z_\Delta = +0.39$, $z_{\text{trad}} = -0.53 \rightarrow z_{\text{trad}}$ selected ($|z_\Delta| \leq 0.5$)

For respiratory rate, the patient's value increased relative to their own baseline ($z_\Delta = +0.57$) but is *below* the population mean ($z_{\text{trad}} = -1.12$). The hierarchy correctly prioritizes the patient-specific upward trend. For the four timing features, within-subject deviations are insufficient ($|z_\Delta| \leq 0.5$), so population norms serve as the default.

Step 1 – Signal Quality. Moderate artifacts: 63 cough events (severity = 1.0), speech contamination over 3 windows, 5 movement episodes. Only 62.0% usable. *Guardrail: Confidence penalty 0.15 applied.*

Step 4 – Digital Biomarkers. RR-CV = 25.2% – elevated, indicating respiratory instability. Median exhalation time within normal range.

Step 6 – Clinical Integration. Risk 40/100 (medium). Despite eupnea classification, supplemental oxygen requirement contributes 2 NEWS2 points. Multiple z-score conflicts noted. Confidence reduced to Medium (0.60) due to moderate artifacts and spirometry surrogate.

Recommendations. Continue monitoring with trend attention. Review medication adherence. Schedule spirometry within 1 week.

C.4. Case 4: bidmc04 – Moderate Artifacts with Preserved Coupling (Score 13/100)

Patient Summary		Signal Quality	
Risk Score	13/100	Grade	Moderate
Confidence	Medium (0.60)	Comp. Severity	0.533
Breathing Pattern	Tachypnea	Usable Signal	57.4%
RR (bpm)	20.5	Cough Events	109
N Breaths	137	Z-Score Conflicts	0

Step 1 – Signal Quality. 109 cough events (cough severity = 1.0), speech contamination over 17 windows. Composite = 0.533 (moderate). Only 57.4% usable – narrowly below the severe abort threshold (0.7). *Penalty: 0.15.*

Step 2 – Breathing Pattern. RR = 20.5 bpm ($z_\Delta = 2.46$). Vt = 0.176 L ($z_\Delta = -1.01$): shallow breaths despite increased frequency. RR-CV = 17.8%.

Step 5 – Cardio-Respiratory Coupling. $r = -0.338$: **preserved** coupling. This is the only case among the presented reports with preserved RSA, suggesting intact autonomic regulation despite tachypnea – a finding that moderates the clinical concern and demonstrates that the system does not uniformly flag diminished coupling.

Step 6 – Clinical Integration. Low risk (13/100). Despite tachypnea, preserved coupling and stable biomarkers support routine monitoring. Confidence reduced to Medium due to moderate artifacts.

C.5. Case 5: bidmc02 – Low Risk with Z-Score Conflicts (Score 0/100)

Patient Summary		Signal Quality	
Risk Score	0/100	Grade	Minimal
Confidence	High (0.75)	Comp. Severity	0.000
Breathing Pattern	Eupnea	Usable Signal	100.0%
RR (bpm)	15.0	Cough Events	0
N Breaths	116	Z-Score Conflicts	3

Z-Score Conflict Detail. Three conflicts demonstrate that even stable patients can exhibit population-vs-patient discrepancies:

- `tidal_volume`: $z_\Delta = -0.55$, $z_{\text{trad}} = +2.55 \rightarrow z_\Delta$ selected. Population norm sees this patient's Vt as highly elevated ($z_{\text{trad}} = +2.55$), but relative to their own baseline, Vt has actually *decreased* ($z_\Delta = -0.55$). The hierarchy correctly identifies the clinically relevant trend.
- `pif`: $z_\Delta = -0.82$, $z_{\text{trad}} = +0.60 \rightarrow z_\Delta$ selected.
- `pef`: $z_\Delta = -0.93$, $z_{\text{trad}} = +1.30 \rightarrow z_\Delta$ selected.

This case exemplifies the personalization problem: a patient with naturally high tidal volume would be flagged by population thresholds ($z_{\text{trad}} = +2.55$) despite being *stable or declining from their own baseline*. The hierarchy prevents this false positive.

Step 2 – Breathing Pattern. Eupnea at 15.0 bpm. Vt = 1.010 L ($z_\Delta = -0.55$): reduced from personal baseline but elevated relative to population. Ti/Ttot = 0.508. RR-CV = 13.0%.

Step 6 – Clinical Integration. Risk 0/100. Despite three z-score conflicts, all are resolved by prioritizing patient-specific trends, yielding a coherent low-risk assessment with high confidence.

Recommendations. Stable respiratory pattern. Continue routine monitoring. Formal spirometry recommended.

C.6. Case 6: bidmc03 – Low Risk, Clean Baseline (Score 0/100)

Patient Summary		Signal Quality	
Risk Score	0/100	Grade	Minimal
Confidence	High (0.75)	Comp. Severity	0.112
Breathing Pattern	Eupnea	Usable Signal	91.0%
RR (bpm)	17.6	Speech Windows	5
N Breaths	138	Z-Score Conflicts	0

Step 1 – Signal Quality. Minimal artifacts. Minor speech contamination (5 windows) and 7 movement episodes, but composite remains low (0.112). 91.0% usable. *Penalty: 0.0.*

Step 2 – Breathing Pattern. Eupnea: RR = 17.6 bpm ($z_\Delta = -0.07$, essentially at personal baseline). Vt reduced ($z_\Delta = -0.52$), suggesting slightly shallow breathing. Ti/Ttot = 0.504. RR-CV = 6.3% – low variability indicating stable, regular breathing.

Step 4 – Digital Biomarkers. RR-CV = 6.3% (within normal range). Median exhalation time = 1.700 s (normal). No indicators of obstructive or restrictive patterns.

Step 5 – Cardio-Respiratory Coupling. $r = 0.017$: diminished coupling. Mean HR = 76.3 bpm.

Step 6 – Clinical Integration. Risk 0/100. All parameters within or near normal ranges. No z-score conflicts. Routine monitoring sufficient. This case represents the system's baseline behavior with a physiologically stable patient and clean signal.

Recommendations. Stable respiratory pattern. Continue routine monitoring. Formal spirometry recommended for definitive airflow assessment.

Limitations. Spirometry surrogate used.

C.7. Summary of Presented Cases

Table 5. Overview of six representative Resp-GRASP clinical reports, spanning the full range of risk scores, signal quality grades, and z-score conflict scenarios observed in the BIDMC dataset.

Patient	Risk	Conf.	Grade	Usable%	Pattern	Coupling	Conflicts
bidmc06	71	High (0.75)	Minimal	100.0	Tachypnea	Diminished	0
bidmc07	62	High (0.75)	Minimal	100.0	Tachypnea	Diminished	2
bidmc05	40	Med (0.60)	Moderate	62.0	Eupnea	Diminished	5
bidmc04	13	Med (0.60)	Moderate	57.4	Tachypnea	Preserved	0
bidmc02	0	High (0.75)	Minimal	100.0	Eupnea	Diminished	3
bidmc03	0	High (0.75)	Minimal	91.0	Eupnea	Diminished	0

These cases demonstrate several key system behaviors: (1) the guardrail abort threshold correctly allows moderate-artifact recordings (bidmc04: $S = 0.533$, bidmc05: $S = 0.476$) to proceed with reduced confidence rather than aborting, reserving abort for truly uninterpretable signals ($S \geq 0.7$); (2) the z-score hierarchy resolves population-vs-patient conflicts in clinically appropriate directions, as illustrated by bidmc02 where population norms would have generated a false positive for tidal volume ($z_{\text{trad}} = +2.55$) that the patient-specific trend correctly overrides ($z_{\Delta} = -0.55$); (3) the system produces differentiated coupling assessments (preserved in bidmc04 vs. diminished in all others), confirming that Step 5 does not uniformly default to a single classification; and (4) recommendations scale appropriately with risk, from routine monitoring (bidmc02, bidmc03) through trend attention and medication review (bidmc05) to urgent pulmonology consult (bidmc06, bidmc07).