

TeleSAFE: Telemetry Stewardship Alarm Fatigue rEduction

Presenting: Claire Shin^{1,4}, Sara Parvaresh Rizi^{2,4}, Luca Gagliardi^{3,4}

Included: Marta Mhanna^{2,4}, Layla Browman^{1,4}, Jacob Udell⁶

¹ Western University, ² University of Toronto, ³ McMaster University, ⁴ Peter Munk Cardiac Center

Table of Contents



INTRO

Clinical impacts, current
snapshot, gaps, research
question and purpose



METHODS / RESEARCH

Data collection,
analysis, SOP, more



CONCLUSION/ Q&A

Future direction, clinical
impacts, and more!

01. Introduction:

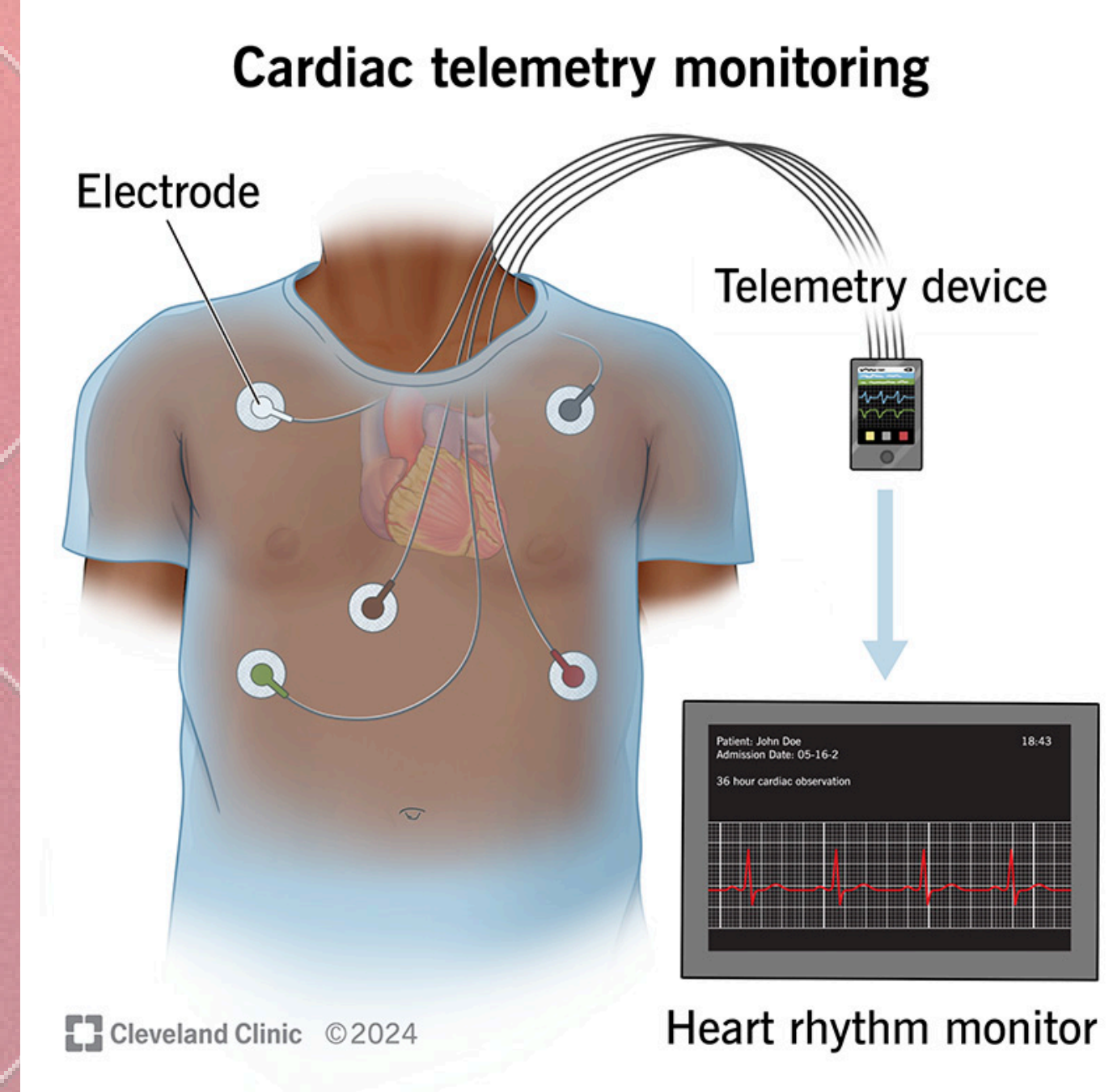
Clinical impacts, current snapshot & gaps, research question, and purpose

What is Telemetry?

Cardiac telemetry: continuous, remote monitoring of a patient's heart patterns with electrode patches on the chest. Data transmits to central station where staff can find irregularities.

Non-actionable alarm: No documented intervention within 60 min. “Intervention” can include: new medication order, stat lab, rhythm intervention, cardiology escalation, RRT, code.

- *Why 60 min?* A convention we will test with a sensitivity analysis at 30 and 120 min.





Statistics

42 telemetry units

available for Cardiology and
Internal Medicine patients in TGH
and Princess Margaret

01

76% inappropriate

telemetry days according to
established practice standards [3]

02

80-99% non-actionable

or false alarms that are
contributing to alarm fatigue and
cognitive overload [2]

03

\$528,241 saved

(on estimate) by eliminating
inappropriate telemetry days [1]

04



Current Gaps

01

Lack of Dynamic Reassessment
Rigid 24/48/72-hour renewal intervals

Black-Box AI

02

Limited transparency reduces clinician trust in current models (ex. CHARTWatch, MoMe Arc, Qaly) [4,5,6,7]

03

Workflow Disconnect

Discontinuation is infrequent as it is unclear who reassesses—or when to discontinue



Research Question

Can an explainable AI-assisted telemetry stewardship model safely identify low-risk patients for earlier telemetry discontinuation without increasing adverse cardiac events?



Research Purpose

To help clinicians make safer, more *transparent*, and more efficient decisions—reducing unnecessary monitoring, false alarms, and staff burden while preserving patient safety.

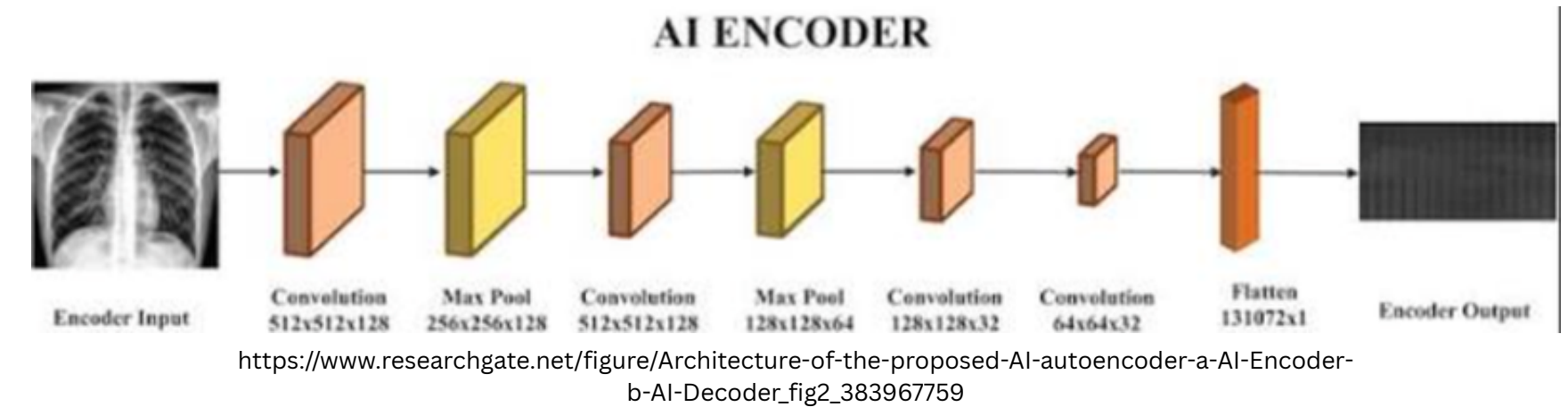
02. Methods: Requirements, objectives, and constraints to validation.

Background on Concepts

01

Encoder

a neural network part that changes raw input data, like text or images, into a compact internal format



02

Ablation study

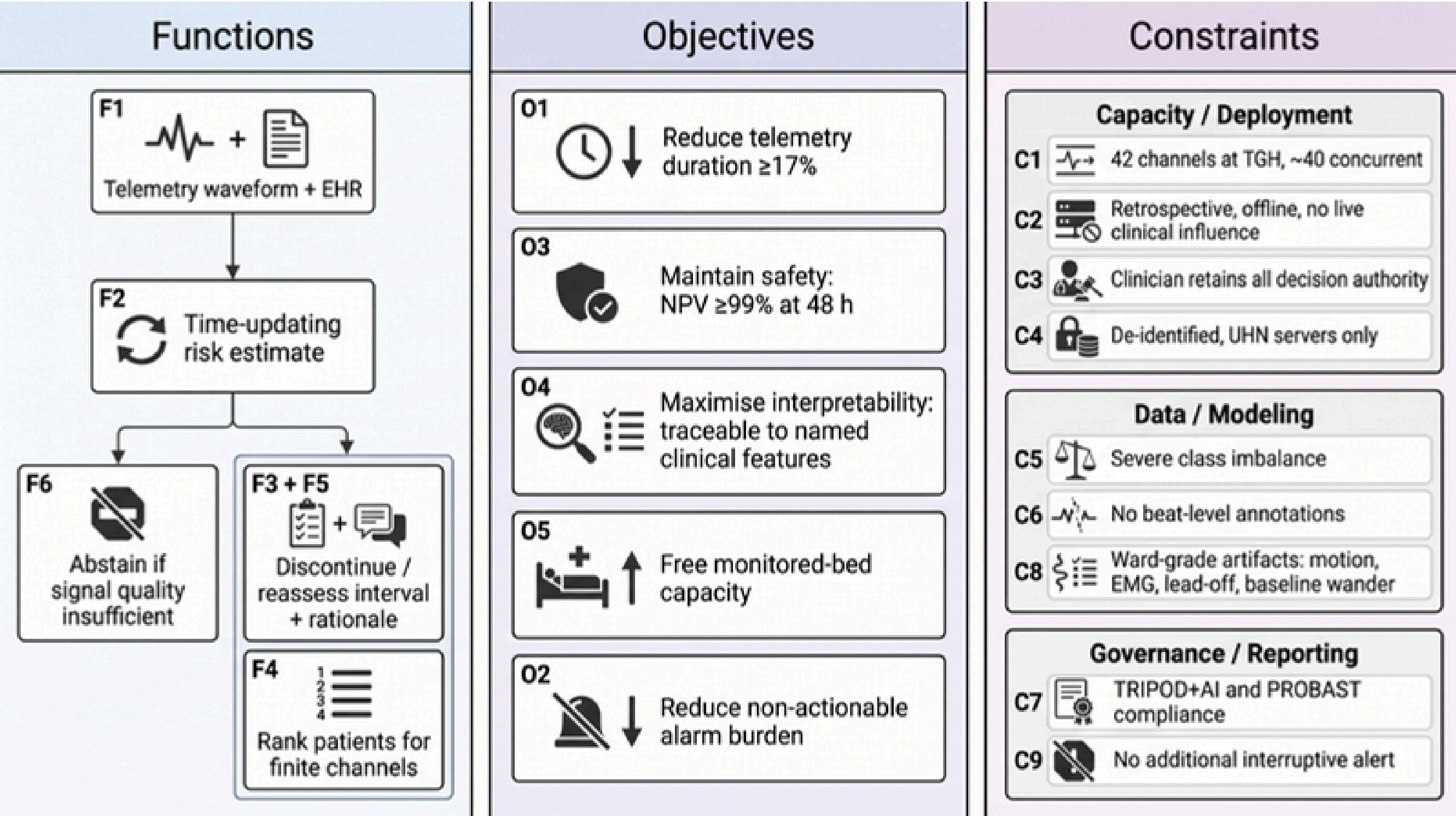
a test where parts of a model are removed/turned off to see how performance changes. This helps see if that piece helps the system or does nothing at all.

03

Negative Predictive Value (NPV)

a statistical measure to evaluate diagnostic tests, being the **chance that a person does not have a disease when they get a negative test result.**

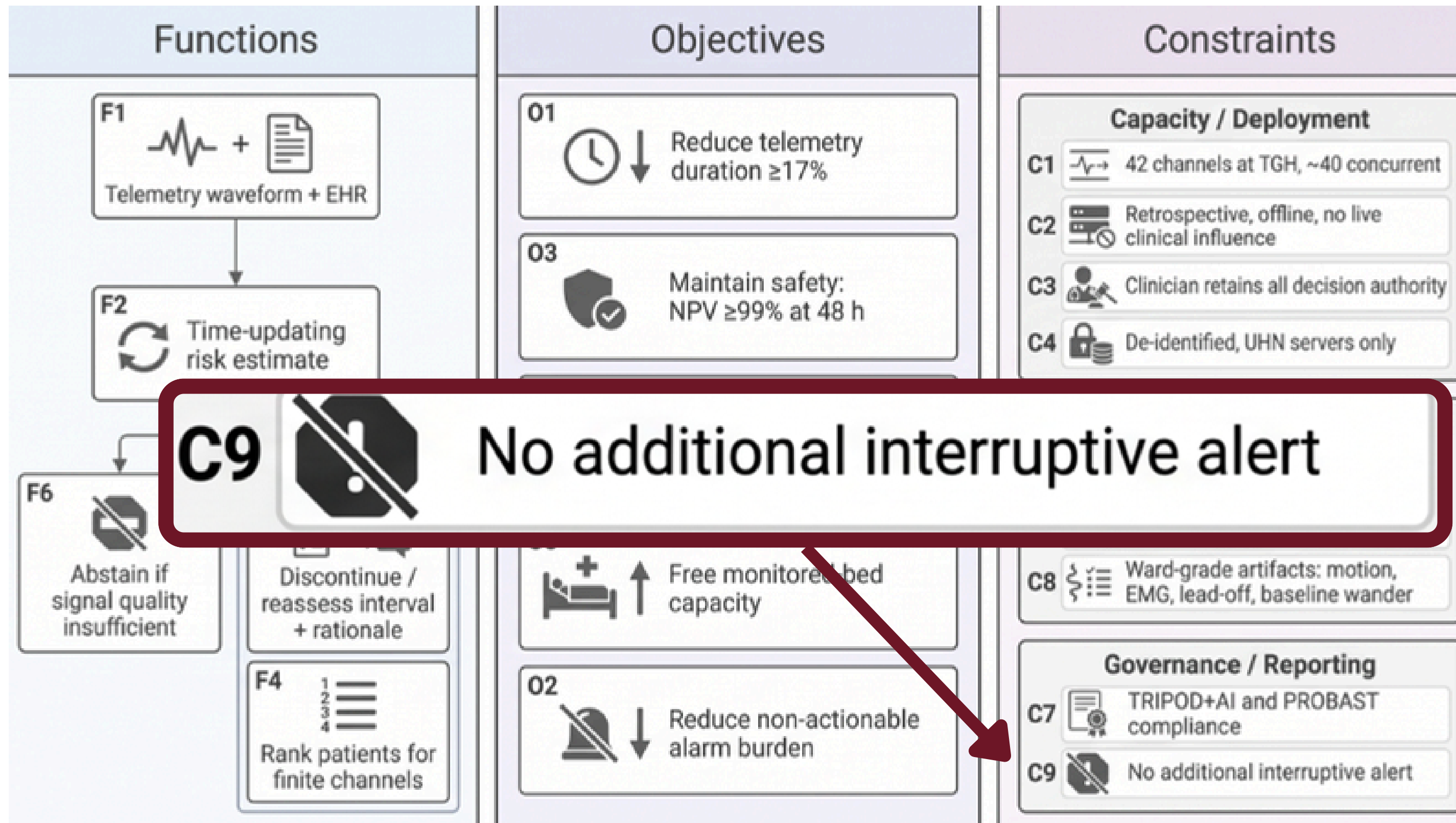
What the system has to do



TLDR

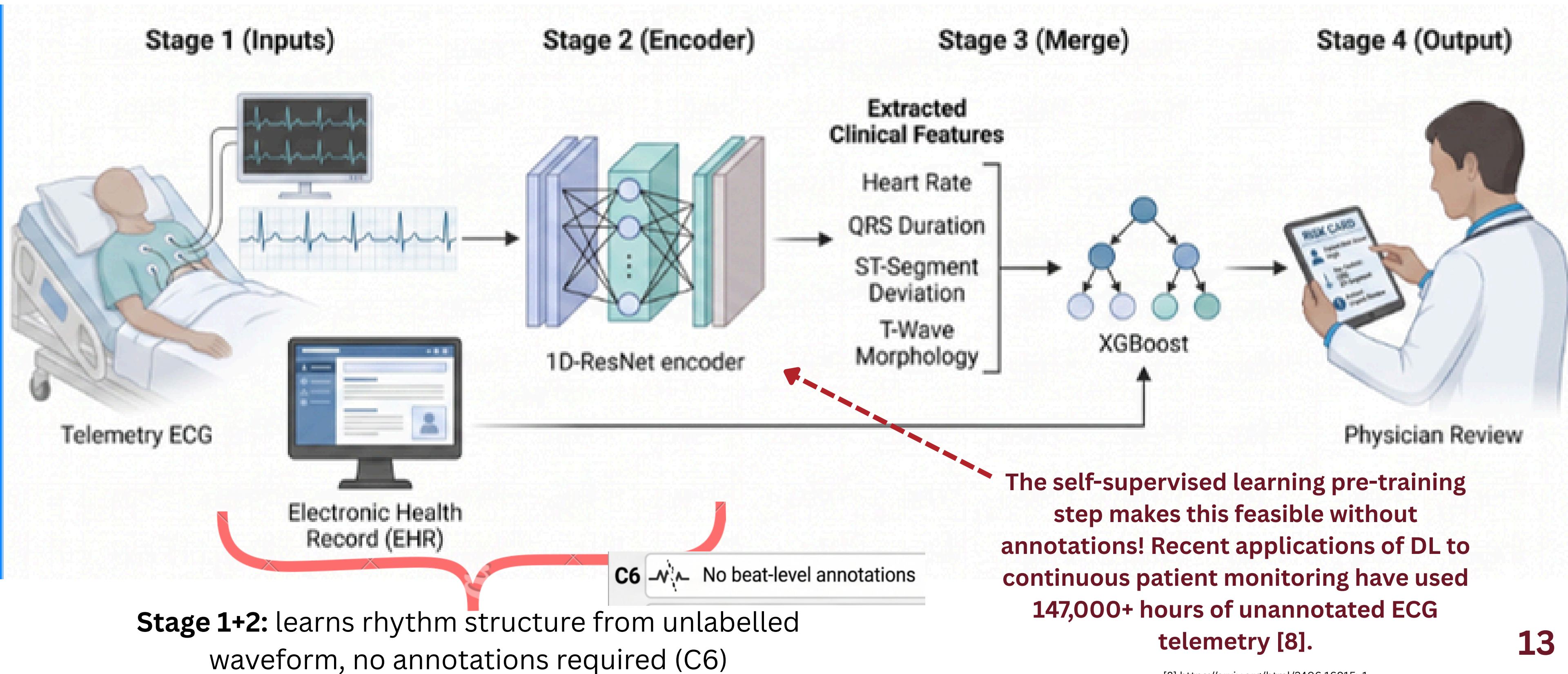
- MUST:**
- 1.recommend when tele can safely stop
 - 2.explain every rec in clinical terms
 - 3.rank patients when channels run short
- MUST NOT:**
- act on its own (clinician decides)
 - add another interruptive alert
- FIXED LIMITS:**
- 42 monitoring channels at TGH
 - adverse events are rare (~2%)
 - NO labelled ward tele exists

Attention to C9, being no interrupts

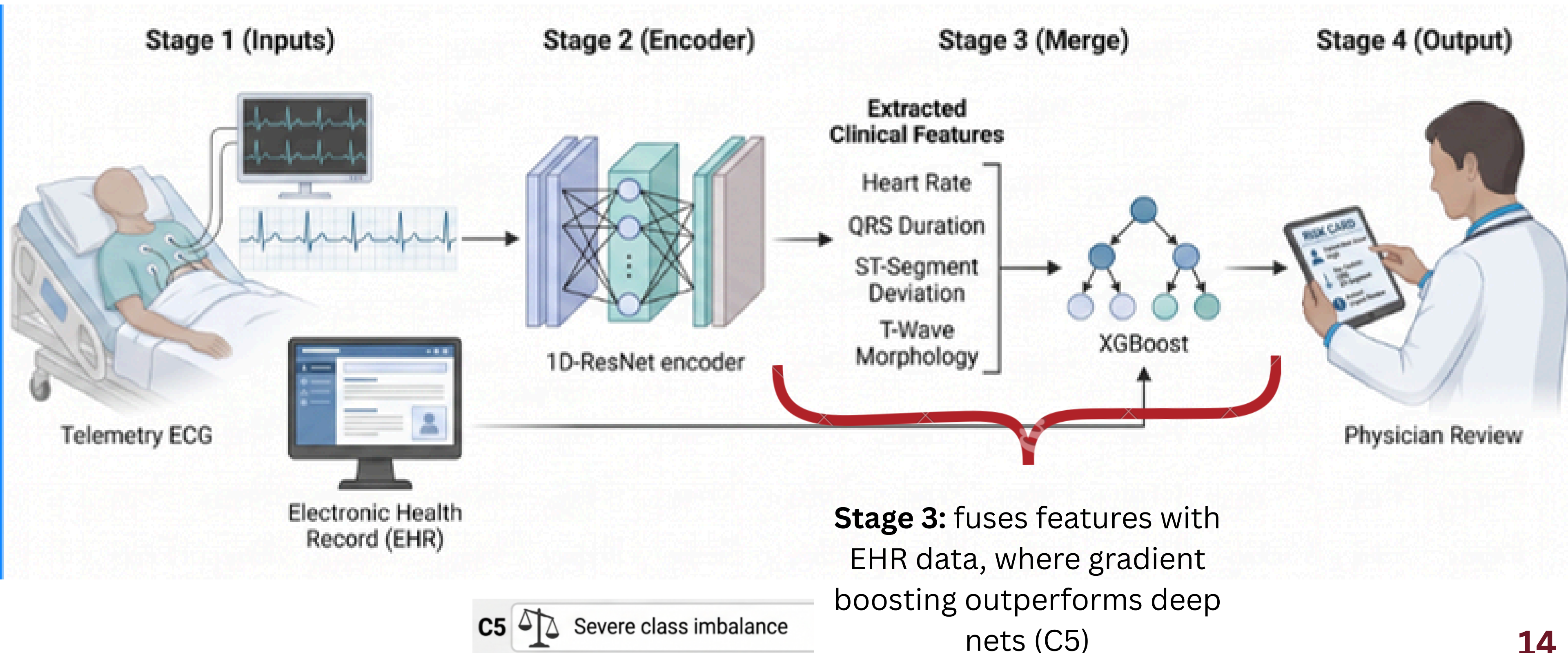


- C9 constrains us **THE MOST, and it is usually most ignored.**
- Targeted EHR telemetry alerts have been trialled: ~31% clinician acceptance [15].
- A tool with lots of pop-ups is just a **NEW** alarm.

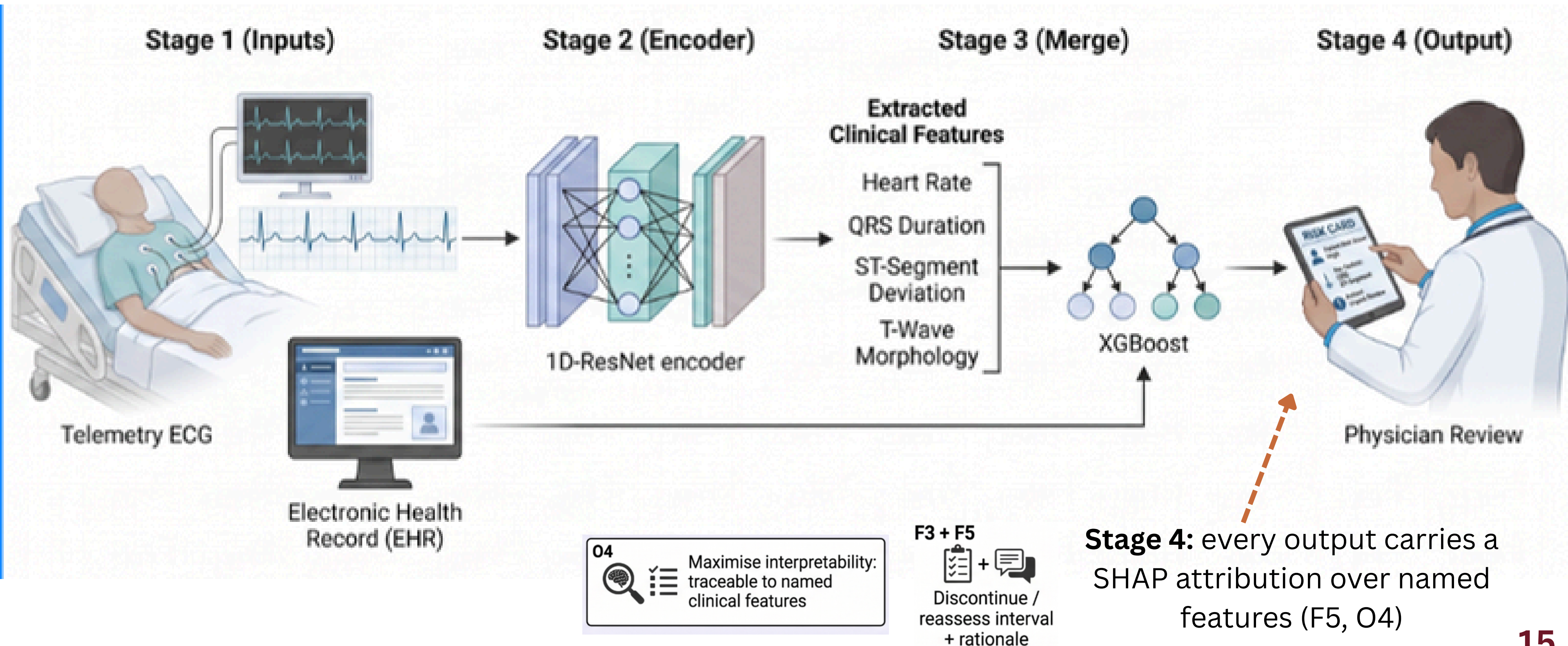
Two-stage hybrid (because tele gives 2 types of data) Waveform encoder -> tabular classifier



Two-stage hybrid (because tele gives 2 types of data) Waveform encoder -> tabular classifier



Two-stage hybrid (because tele gives 2 types of data) Waveform encoder -> tabular classifier



Explainable is a requirement, not a feature

Stage 4 (Output)



Physician Review

Bed 5B-12

TELESAFE: LOW RISK

Recommend: reassess for discontinuation
Next reassessment: 12 h

- Sinus rhythm, stable 41 h
- No PVC runs > 3 beats in 24 h
- K+ 4.1, Mg 0.9 — in range
- K+ 4.1, Mg 0.9
- Day 2 post-PCI

Signal quality: good (94% analysable)

☐ Accept ☐ Keep telemetry ☐ Why?

A mocked-up clinician card

- SHAP over raw latent dimensions is *not* clinically explainable
 - E.g. "embedding 47 contributed -0.42" is not actionable
- Constrain the encoder to emit named clinical concepts (e.g. longest pause, rate, rhythm stability). Connect risk to features a cardiologist already reasons with (O4)

O4



Maximise interpretability:
traceable to named
clinical features



What we *could* test without patient data

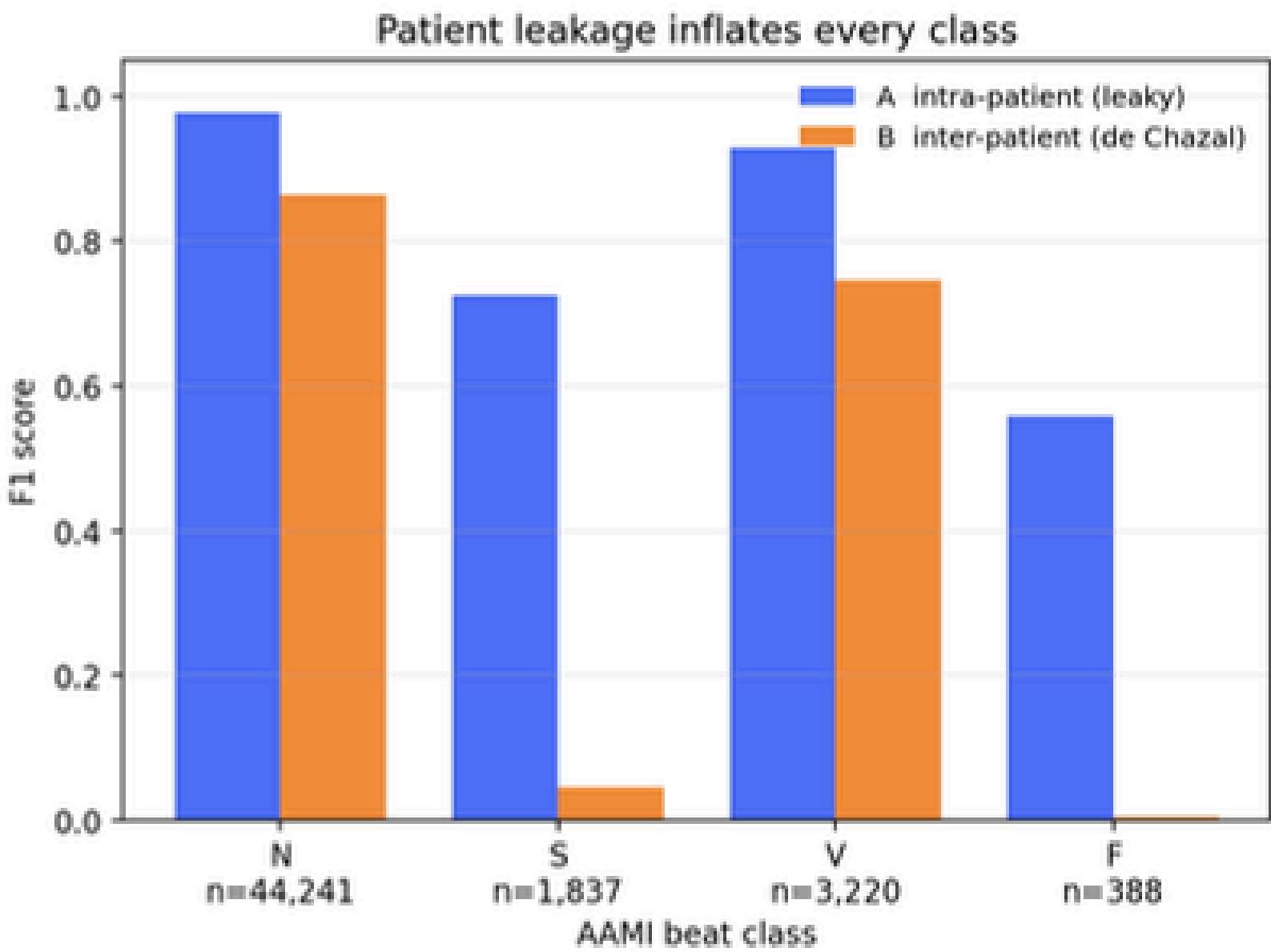
Still waiting on data, so cannot validate clinical claim yet. However, we verified the components independelty against requirements, using free data/modelling.

	What we tested	How we tested it
F2  Time-updating risk estimate	Can it detect arrhythmia in patients it's never seen? Is the two-stage design necessary?	Public ECG database (MIT-BIH, PhysioNet) [12,13] Ablation
03  Maintain safety: NPV ≥99% at 48 h	Is our own safety threshold meaningful	Statistical analysis
05  Free monitored-bed capacity C1  42 channels at TGH, ~40 concurrent	Is the capacity claim real?	Queueing/capacity model (Erland-B)

What we tested	How we tested it
Can it detect arrhythmia in patients it's never seen?	Public ECG database (MIT-BIH, PhysioNet)



Initial model test on MIT-BIH open-source (clean) ECG dataset



	Random split	Patient-disjoint
Accuracy	0.957	0.770
Macro F1	0.655	0.332
Weighted F1	0.963	0.819

Same model, only split changed.

Shuffle beats randomly so a patient in both training and testing (“leaky”):

96% accurate

Split so no patient appears in both (“de Chazal” [14]): **77% accurate**

[14] doi: 10.1109/TBME.2004.827359

What we tested	How we tested it
Can it detect arrhythmia in patients it's never seen?	Public ECG database (MIT-BIH, PhysioNet)



Initial model test on MIT-BIH open-source (clean) ECG dataset



If a patient's beats are in both sets, the model doesn't learn to recognise the arrhythmia. It learns to recognise the patient.

Flattering (96%) vs. honest (77%) comparison 🤔🤔

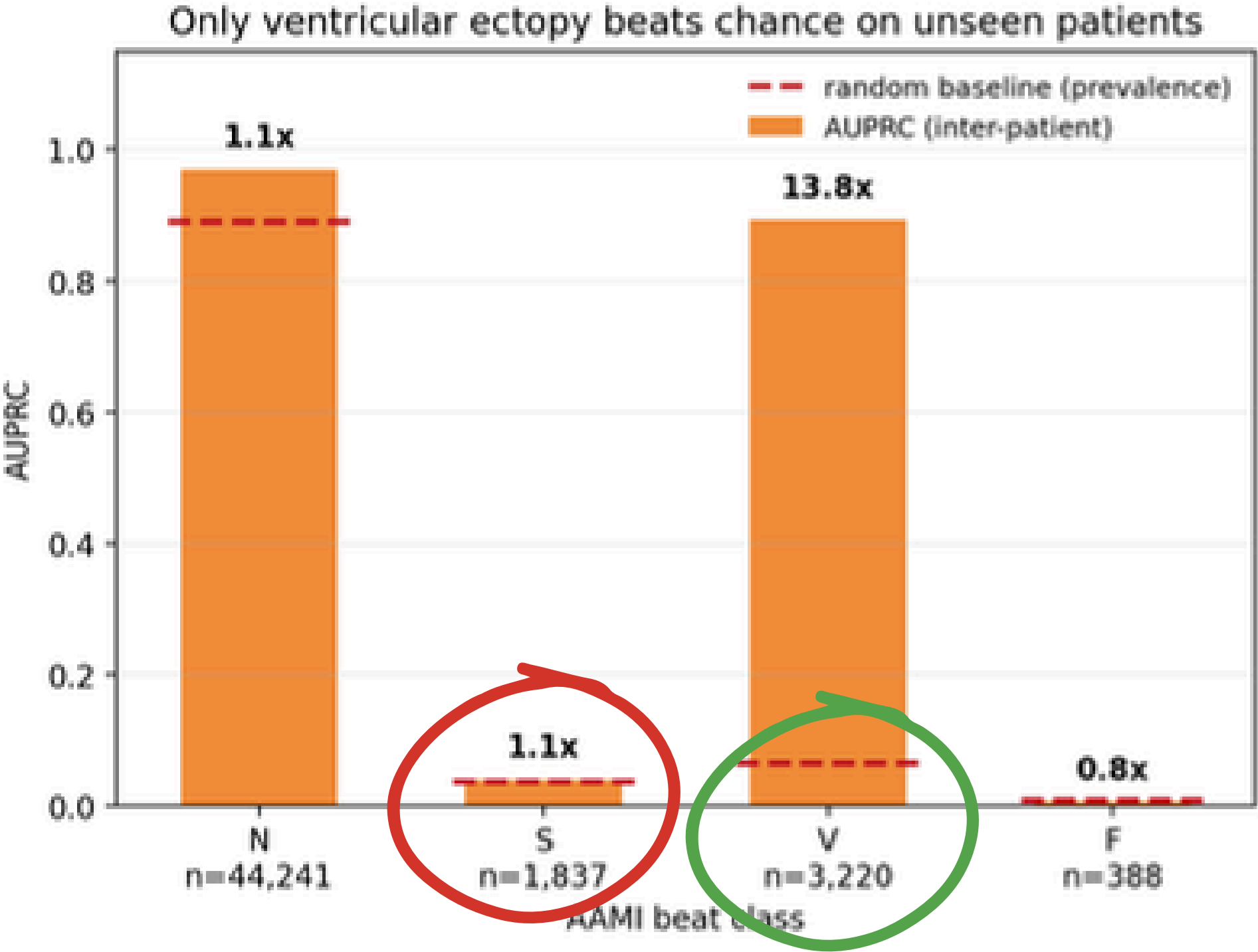
Shuffle beats randomly so a patient in both training and testing ("leaky"): **96% accurate**

Split so no patient appears in both ("de Chazal"): **77% accurate**

Beyond project scope: most highly accurate ECG accuracies published, that are above 99%, are leaky.

What we tested	How we tested it
Can it detect arrhythmia in patients it's never seen?	Public ECG database (MIT-BIH, PhysioNet)

 **The model is able to detect PVCs very well, but not PACs.**



PVC = wide, misshapen beat.
Detected ~14x better than chance.

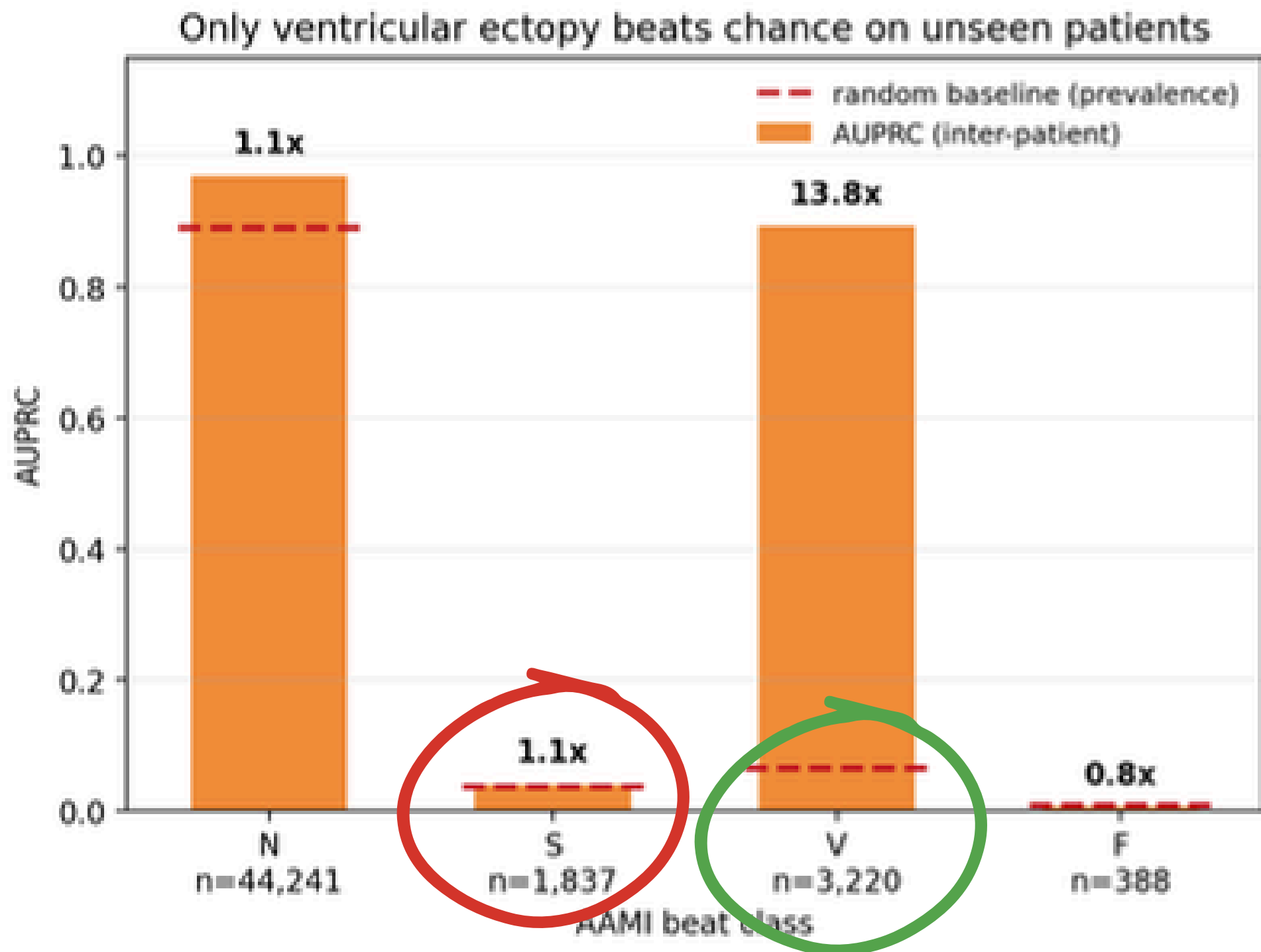
PAC = normal-looking beat that arrives early. Detected almost same as chance.



[9] <https://www.qaly.co/post/premature-ventricular-contraction-single-vs-premature-atrial-contraction-single-on-your-watch-ecg>

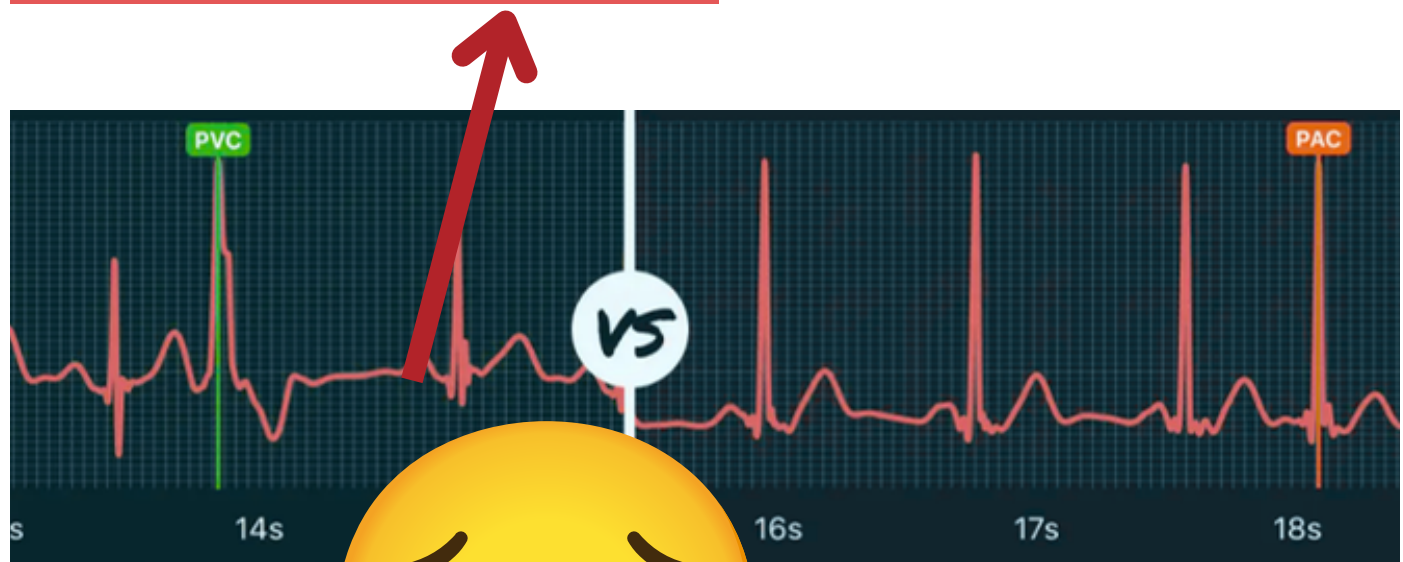


The model is good at shapes, useless at timing. Training failure or can we revise?



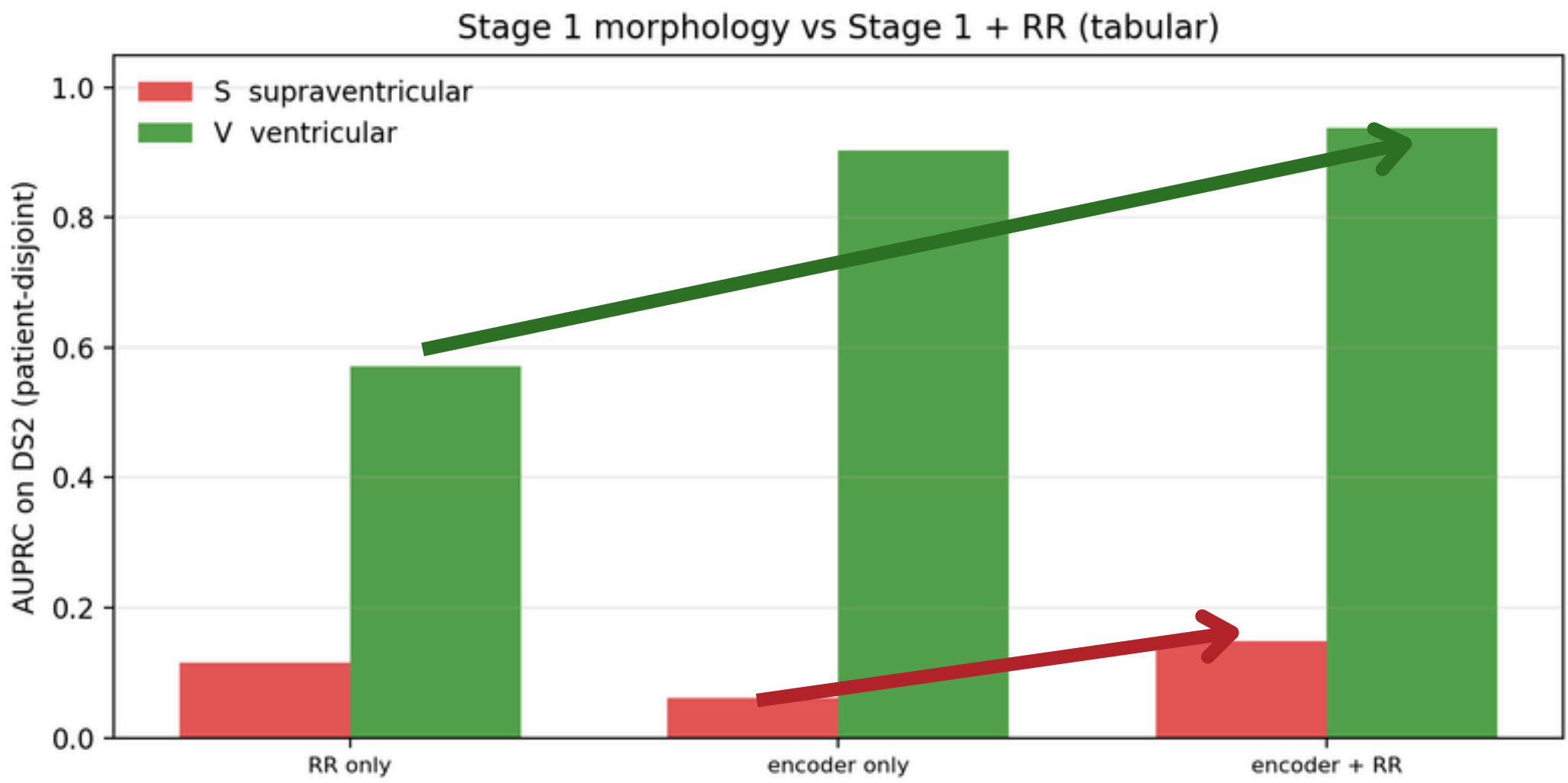
PVC = wide, misshapen beat.
Detected ~14x better than chance.

PAC = normal-looking beat that arrives early. Detected almost same as chance.



Good at shapes, useless at timing. Training failure or can we revise?

 Each half of the model does a job the other one can't, which is why we have 2.



Analysis window: 256 samples at 360 Hz = 0.711 s = one RR interval at 84.4 BPM
One RR interval: 72 BPM = 0.833s
WINDOW IS SHORTER THAN ONE BEAT-TO-BEAT INTERVAL so it carries no timing. Above the 84.4BPM equivalent, a neighbouring R peak starts entering window. Likely why encoder has weak PAC signal. Limited by physiology!

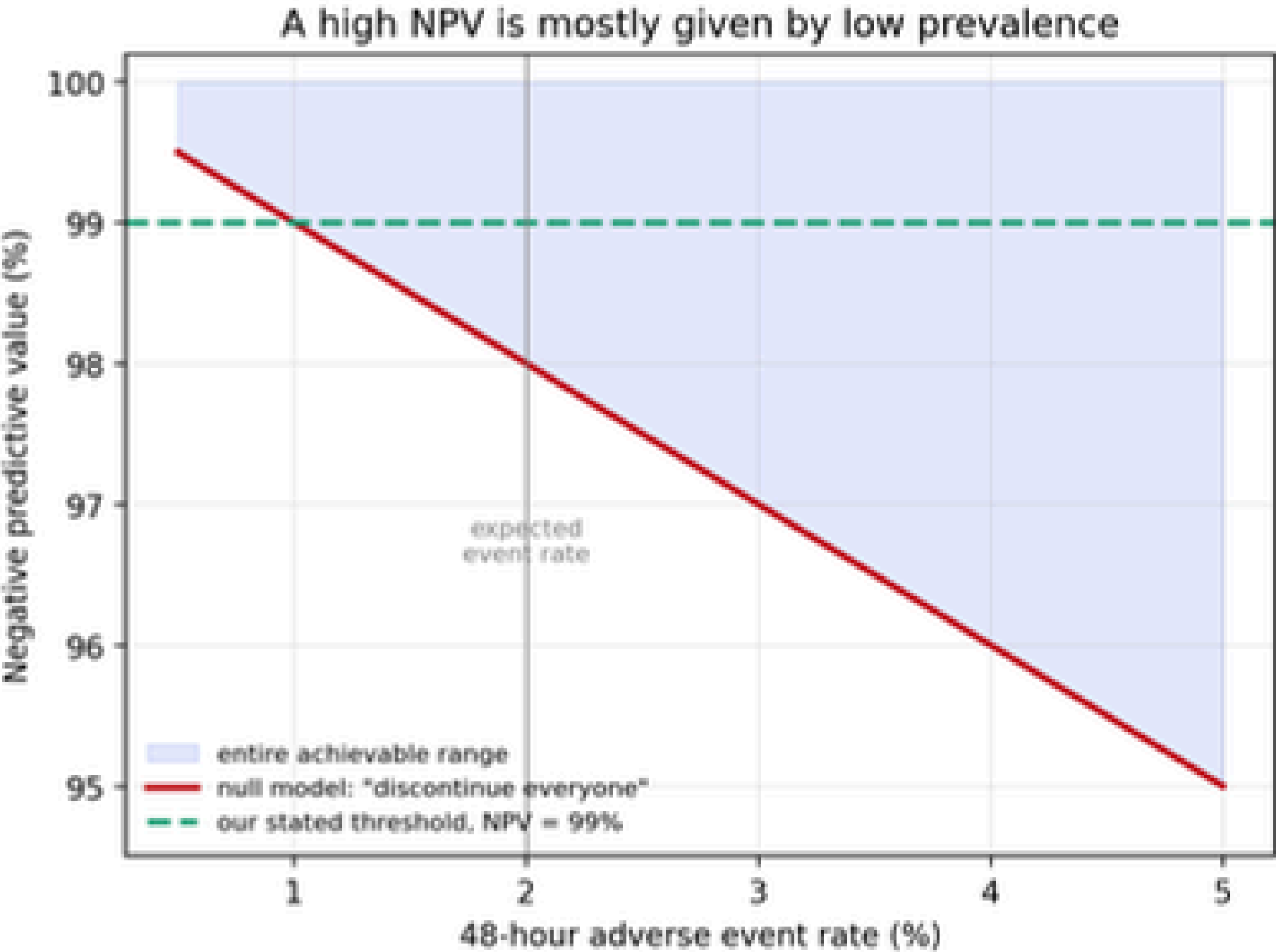
Adding timing features more than doubled PAC detection! (how long since the last beat, how long to the next).

Input	PAC	lift	PVC	lift
Timing only	0.116	3.13x	0.571	8.81x
Encoder only	0.061	1.65x	0.904	13.94x
Encoder and timing	0.149	4.04x	0.938	14.47x

NOTE!



Auditing our safety threshold failed. “99% safe” can mean missing half of events.



03



Maintain safety:
NPV ≥99% at 48 h

(O3)

Initial (default in research) NPV=99%.

- At a 2% event rate, recommending discontinuation for everyone scores 98%.

NPV=99% has 51% sensitivity.
Half of events are still missed.

What we tested	How we tested it
Is our own safety threshold meaningful	Statistical analysis



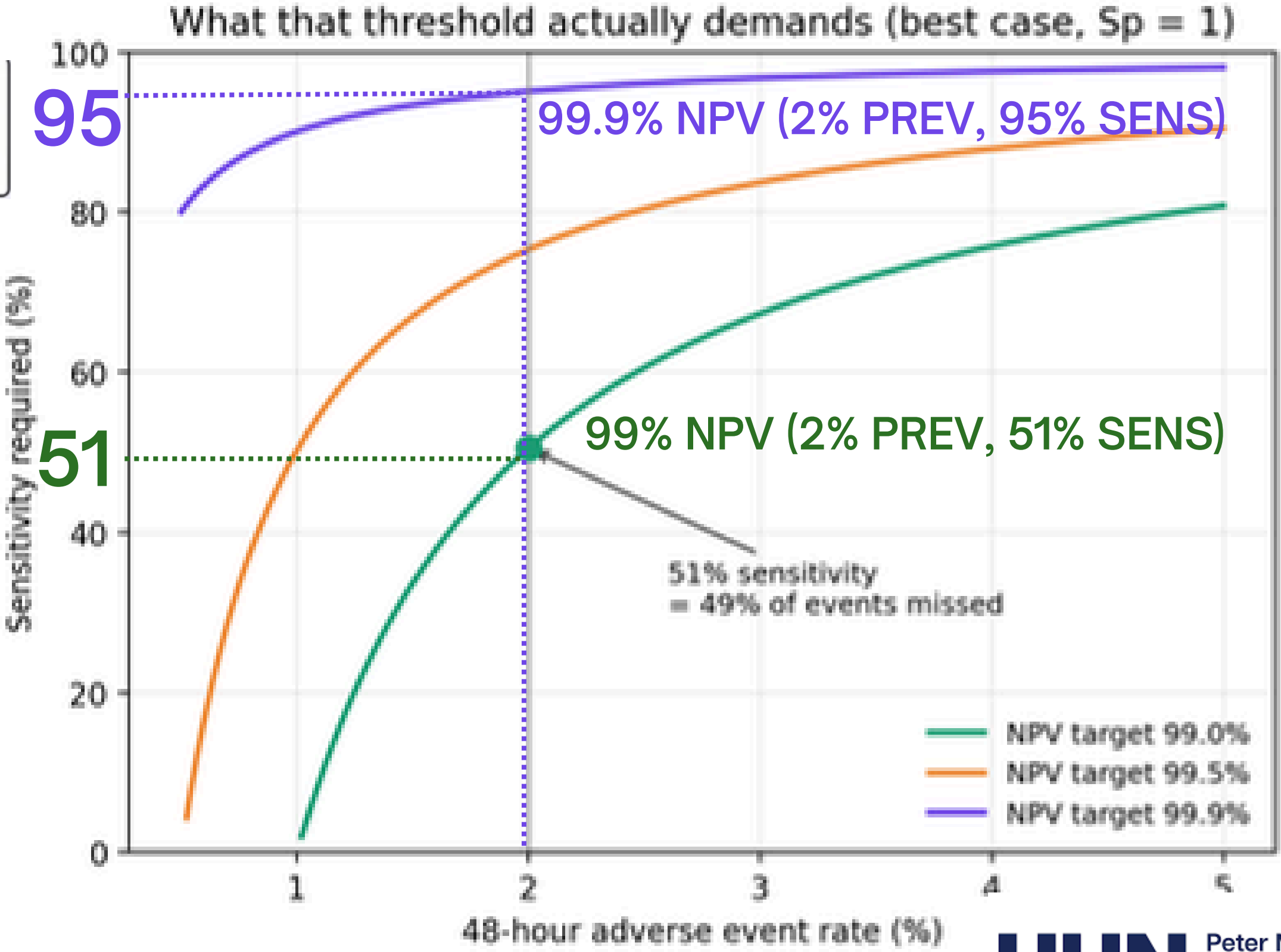
Auditing our safety threshold failed. To get sensitivity 95%, use NPV 99.9%.

03



99.9%
Maintain safety:
NPV > 99% at 48 h

(O3)





Capacity Model

We can use a LOSS system, with Markovian (Poisson) arrivals, Markovian service times, c servers, and total system capacity being c .

- Implies NO queue. If all servers busy, arrival is LOST and not held in line (similar to TGH)

Model: **Erlang-B**, that asks “how many circuits does a switchboard need so callers rarely get a busy signal?” [10]. Today, the MOST applicable model to hospital bed allocation [11].

- Matches TGH because finite pool of channels, stochastic arrivals, variable holding times

offered load: $A = \lambda \times h$

- λ = arrivals per hour
- h = mean holding time
- A = average number of channels that would be busy if we had unlimited tele. Has units “Erlangs”.

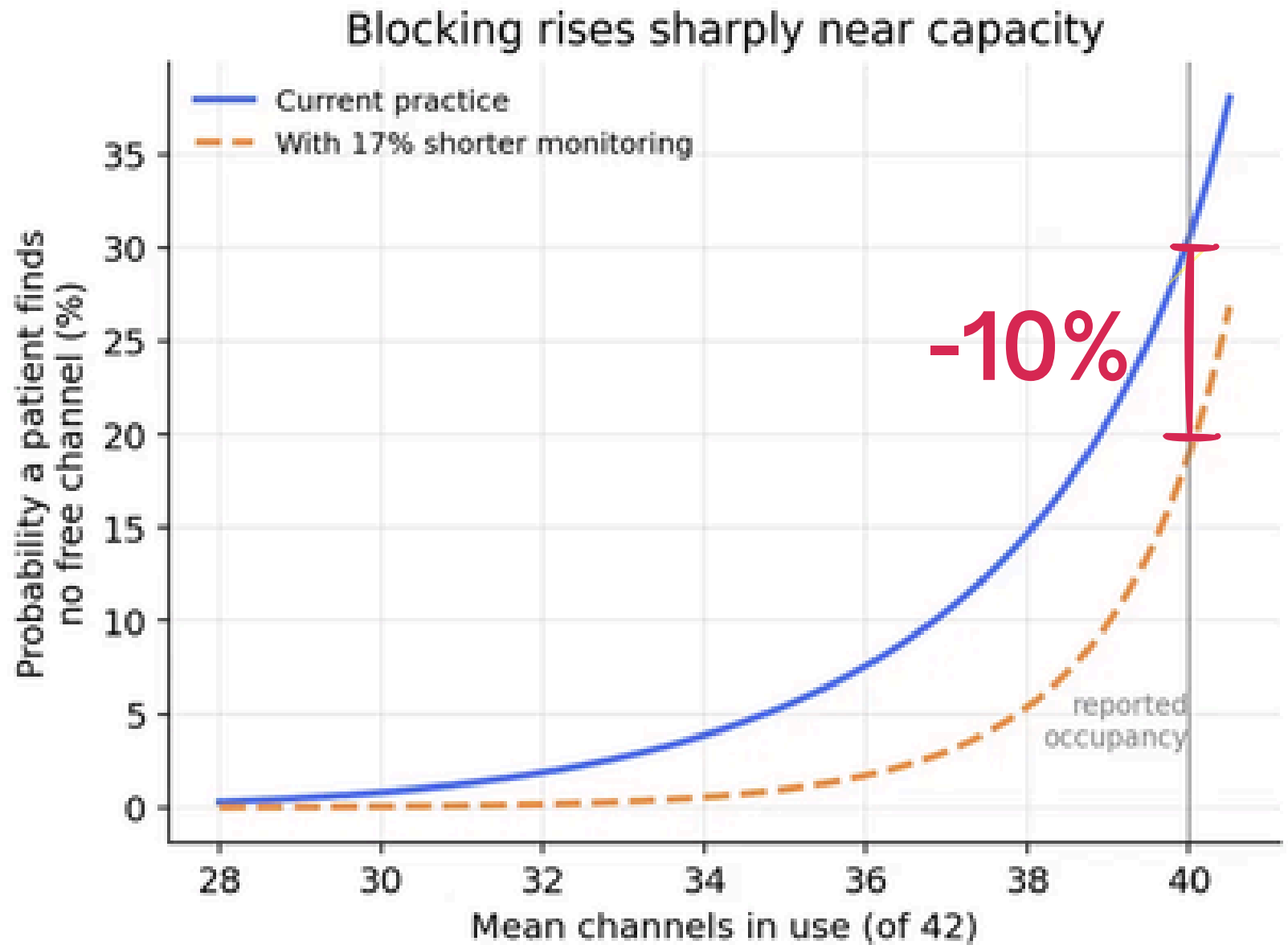

$$B(c, A) = \frac{A^c / c!}{\sum_{k=0}^c A^k / k!}$$

[10] <https://www.semanticscholar.org/paper/adf5b0d899df65c58caab65f2dd64d2f5e7ebaaa>

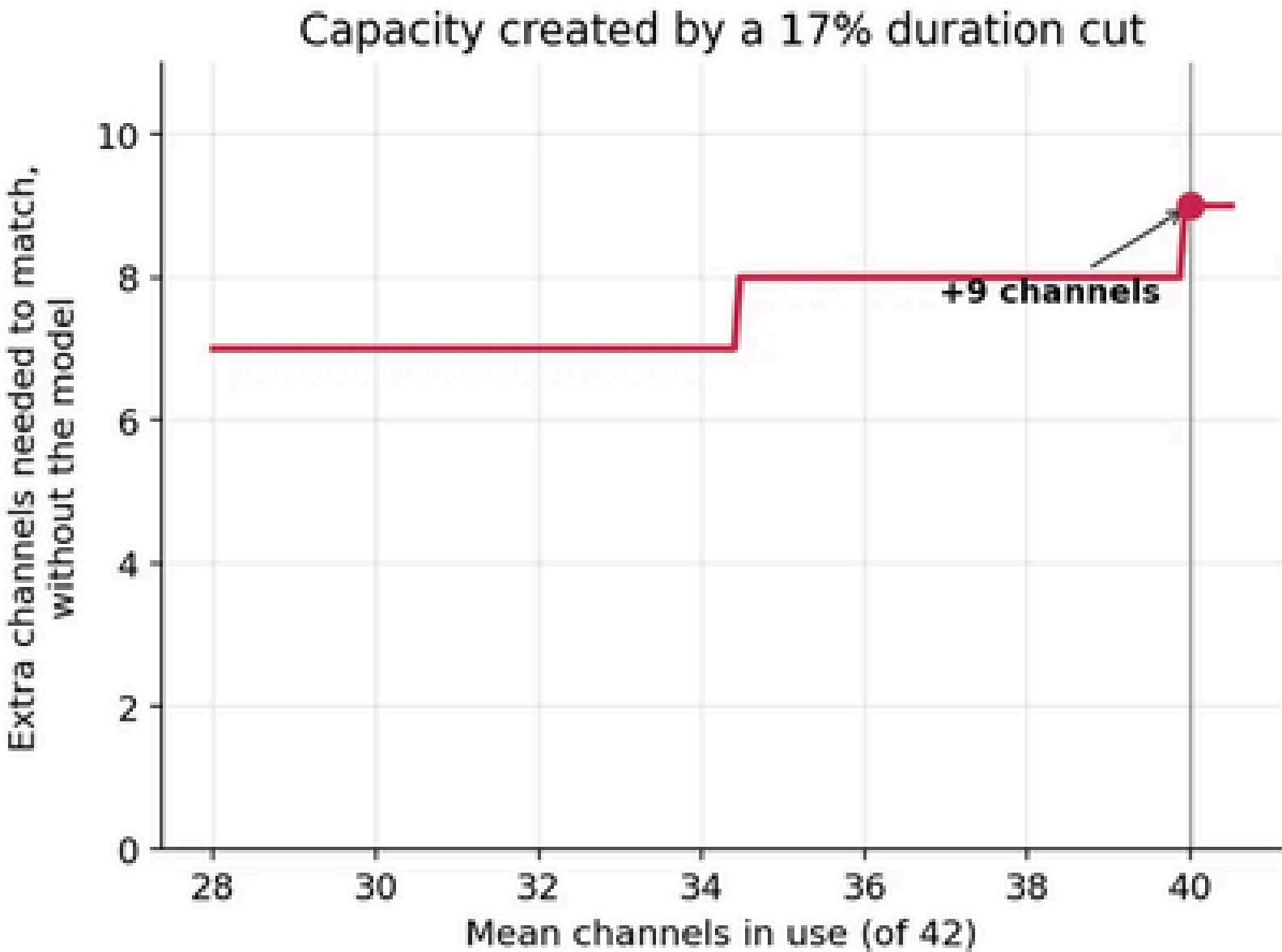
[11] [https://www.researchgate.net/publication/220461897_Dimensioning_hospital_wards_using_the_Erlang_loss_model\[2\]](https://www.researchgate.net/publication/220461897_Dimensioning_hospital_wards_using_the_Erlang_loss_model[2])



“Nine more monitors without buying nine more monitors.”



NOTE: we were told “40/40 beds full” is a daily thing, from our own research at the 5B ward in TGH.



EXTRA physical channels (at 50h duration) to get the blocking probability that the 17% reduction gives for free

~10% reduction of full beds

+9 Channels free



Evaluating Objectives

01

 Reduce telemetry duration $\geq 17\%$

02

 Reduce non-actionable alarm burden

03

 Maintain safety: NPV $\geq 99\%$ at 48 h

04

 Maximise interpretability: traceable to named clinical features

05

 Free monitored-bed capacity

Req	Test	Metric	Statistical method	Acceptance threshold
01	Simulated vs actual discontinuation time, paired	Median change in telemetry hours	Wilcoxon signed-rank, median + IQR, 95% CI	$\geq 17\%$ reduction, $p < 0.05$
02	Non-actionable alarms per patient-day	Internal rate of return (IRR)	Paired negative binomial	$IRR < 1$, CI excludes 1
03	Adverse events within 48 h of simulated discontinuation	NPV	Wilson score 95% CI	$NPV \geq 99\%$ AND sensitivity reported alongside
04	Explanation fidelity	Clinician review of n sampled SHAP rationales	Qualitative	Pre-specified agreement rate
05	Bed-days saved = sum of hours saved $\div$ 24	Bed-days/month, projected cost	Non-parametric bootstrap, 1000 resamples	95% CI excludes zero
-	Model discrimination	<u>AUPRC</u> primary, <u>AUROC</u> secondary		Pre-specify floor
-	Model calibration	Brier, calibration intercept + slope	Calibration plot	Slope within 0.8–1.2

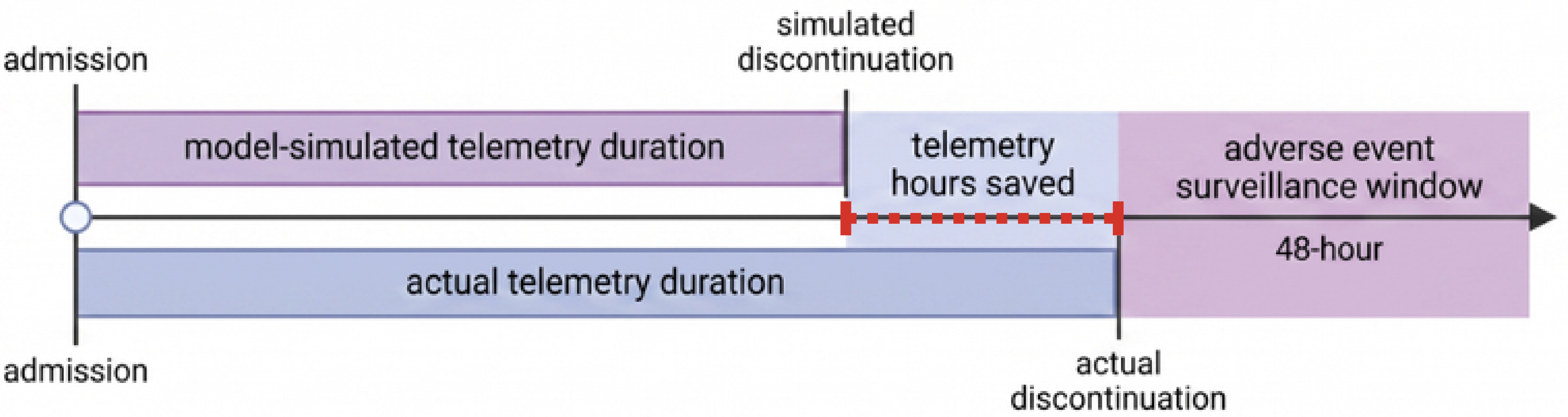
Evaluating Objectives (#1)

Req	Test	Metric	Statistical method	Acceptance threshold
01	Simulated vs actual discontinuation time, paired	Median change in telemetry hours	Wilcoxon signed-rank, median + IQR, 95% CI	≥ 17% reduction, p < 0.05

01

Reduce telemetry duration ≥17%



What could go wrong

The counterfactual problem (most significant)

If telemetry caught an arrhythmia -> triggered an intervention that prevented a code, our "no adverse event" label is contaminated by the very monitoring we propose removing.

Mitigation: Restrict primary safety events to those **detectable independent of telemetry** (e.g. Code Blue, triggered by nurses and vitals, not only by the monitor). Sensitivity analysis excluding encounters where the **intervention was demonstrably telemetry-triggered**.

Asymmetric error

Under-triage can be fatal, over-triage wastes a channel.

Mitigation: **loss weighting is deliberately asymmetric: 3.0 on adverse complications vs 1.0 on baseline**. The model abstains rather than guesses when signal quality fails the gate (F6).



Generalisability and fairness

Single-centre, single-vendor (Philips), external validation not yet performed.

Mitigation: Subgroup performance reported by sex, age band, BMI, and ethnicity per **TRIPOD+AI**. Ten-year window spans practice change; temporal split is our test for drift.

03. Future Directions

Future Directions



- Validate TELESafe using retrospective TGH data:
 - On actual tele
 - Use tele bed admission data to strengthen Erlang-B argument
 - On actual UHN tele costs
- Evaluate safety and accuracy of telemetry-discontinuation predictions
- Progress toward prospective clinical validation and implementation

Epic EHR



Phillips Telemetry



Operational Data





TELESAFE



- Earlier safe discontinuation of telemetry
- Reduce telemetry hours
- Improve patient flow, save bed-days
- Preserve limited telemetry capacity for patients who need it most
- Support -NOT replace- clinical judgment

...“Nine more monitors without buying nine more monitors.”.

References

- [1] R. Chong-Yik, A. L. Bennett, R. V. Milani, and D. P. Morin, "Cost-saving opportunities with appropriate utilization of cardiac telemetry," *Am. J. Cardiol.*, vol. 122, no. 9, pp. 1570–1573, Nov. 2018, doi: 10.1016/j.amjcard.2018.07.016.
- [2] S. Jacques and E. Williams, "Reducing the safety hazards of monitor alert and alarm fatigue," PSNet, Agency for Healthcare Research and Quality, Rockville, MD, USA, May 1, 2016. [Online]. Available: <https://psnet.ahrq.gov/perspective/reducing-safety-hazards-monitor-alert-and-alarm-fatigue>
- [3] R. Dressler, M. M. Dryer, C. Coletti, D. Mahoney, and A. J. Doorey, "Altering overuse of cardiac telemetry in non-intensive care unit settings by hardwiring the use of American Heart Association guidelines," *JAMA Intern. Med.*, vol. 174, no. 11, pp. 1852–1854, Nov. 2014, doi: 10.1001/jamainternmed.2014.4491.
- [4] S. S. Pendharkar, I. B. Barry, S. Patibandla, T. I. Leung, A. Gupta, A. N. Lin, et al., "AHA telemetry guidelines improve telemetry utilization in the inpatient setting," *Am. J. Manag. Care*, vol. 26, no. 11, pp. 476–481, Nov. 2020, doi: 10.37765/ajmc.2020.88525.
- [5] G. P. de Jesus, L. R. G. Dolenkei, V. A. L. Cerqueira, and J. V. M. K. dos Santos, "Bridging efficiency and interpretability: Explainable edge AI for clinical decision-making in medical telemetry," *Iberoamerican J. Appl. Comput.*, vol. 14, no. 1, 2026.
- [6] Unity Health Toronto, "AI tool study." Unity Health Toronto. Sept. 2024. [Online]. Available: <https://unityhealth.to/2024/09/ai-tool-study/> (accessed Aug. 22, 2026).
- [7] InfoBionic.Ai, "MoMe ARC." InfoBionic.Ai. [Online]. Available: <https://infobionic.ai/mome-arc/> (accessed Aug. 22, 2026).
- [8] Qaly, "Qaly." Qaly. [Online]. Available: <https://www.qaly.co/> (accessed Aug. 22, 2026).
- [9] Qaly, "Premature ventricular contraction (single) vs. premature atrial contraction (single) on your watch ECG." Qaly. [Online]. Available: <https://www.qaly.co/post/premature-ventricular-contraction-single-vs-premature-atrial-contraction-single-on-your-watch-ecg> (accessed Aug. 22, 2026).
- [10] T. Kite, U. T. Siam, B. Ayers, N. Houstis, and A. D. Aguirre, "Unlocking telemetry potential: Self-supervised learning for continuous clinical electrocardiogram monitoring," 2024, arXiv:2406.16915.
- [12] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH Arrhythmic Database," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, May–Jun. 2001, doi: 10.1109/51.932724.
- [13] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000, doi: 10.1161/01.CIR.101.23.e215.
- [14] P. de Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 7, pp. 1196–1206, Jul. 2004, doi: 10.1109/TBME.2004.827359.
- [15] N. Latifi, T. Johnson, A. M. Knight, L. Prichett, B. Modanloo, T. Dungarani, et al., "Optimizing decision support alerts to reduce telemetry duration: A multicenter evaluation," *Appl. Clin. Inform.*, vol. 15, no. 5, pp. 860–868, Oct. 2024.
- [16] A. K. Erlang, "Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges," *Post Office Elect. Eng. J.*, vol. 10, pp. 189–197, 1917.

THANK YOU!
Any Questions?



Hyperparameters

- Encoder:
 - 32 filters
 - $k=5$
 - $s=2$

} → 64 filters $k=3$ $s=1$,
- PReLU
- Spatial dropout 0.25–0.40
- L2
- 2-layer BiLSTM/GRU
- penultimate dense 128–256
- **Training:** batch 64–128, 50–100 epochs
 - EarlyStopping patience 10
 - ReduceLROnPlateau factor 0.1 patience 5.



The way we split the data changed our results

Held constant:

- architecture (1D ResNet stem → 3 residual blocks → 2-layer BiGRU → 128-d dense)
- hyperparameters: Adam, lr 1e-3, batch 128, 12 epochs, ReduceLROnPlateau (factor 0.1, patience 2)
- preprocessing: 256-sample window on R-peak and per-beat Z-score
- inverse-freq class weighting
- random seed (7) to NumPy, PyTorch, CUDA

WHAT CHANGED:

1. ARM A: all beats pooled, shuffled, 80/20 split. Patients on both sides
2. ARM B; de Chazal DS1 → DS2, no patient crosses boundary

COHORT:

- 44/48 MIT-BIH records, 4 excluded because of paced rhythms, per AAMI recommendation
- **DS1: 22 records, 51002 beats (train). DS2: 22 records, 49693 beats (test)**
- MLII lead selected by header name, not channel index. Record 114 stores on MLII on channel 1, and taking channel 0 blindly would feed the model a diff lead for same patient.



Adverse Event Defns

For the purpose of this study, adverse cardiac events will refer to primary outcomes reporting:

- Code Blue activation
- Rapid Response Team (RRT) activation
- Unplanned ICU transfer / ICU step-up
- Clinically significant arrhythmia
 - a. Ventricular tachycardia (VT)
 - b. Ventricular fibrillation (VF)
 - c. Torsades de pointes
 - d. Sustained supraventricular tachycardia requiring intervention
 - 3,5 M e. New-onset atrial fibrillation/flutter requiring treatment
 - f. Symptomatic bradycardia
 - g. High-grade AV block
 - h. Asystole / significant pauses



TRIPOD+AI / PROBAST alignment.

TRIPOD+AI: Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis [1]

- 22-item checklist, which aims to improve the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes.
- Aims to improve the transparency of the reporting of a prediction model study regardless of the study methods used

PROBAST: Prediction model Risk Of Bias ASsessment Tool [2]

- A tool for assessing the risk of bias (ROB) and applicability of diagnostic and prognostic prediction model studies
- 4 domains: participants, predictors, outcome, and analysis, each with a total of 20 signaling questions to occur when shortcomings in study design, conduct, or analysis lead to systematically distorted estimates of model predictive performance.

[1] <https://pubmed.ncbi.nlm.nih.gov/25560730/>

[2] <https://pubmed.ncbi.nlm.nih.gov/30596875/>



Why XGBoost and not a neural net end-to-end?

Sparse tabular data, sample size, native missingness handling, and SHAP tractability.



Isn't 95% sensitivity extremely demanding?

Yes, deliberately. The errors are asymmetric: missing a deteriorating patient can be fatal, over-monitoring wastes a channel. **Rather set a hard target and report honestly against it than set one we're guaranteed to clear.**

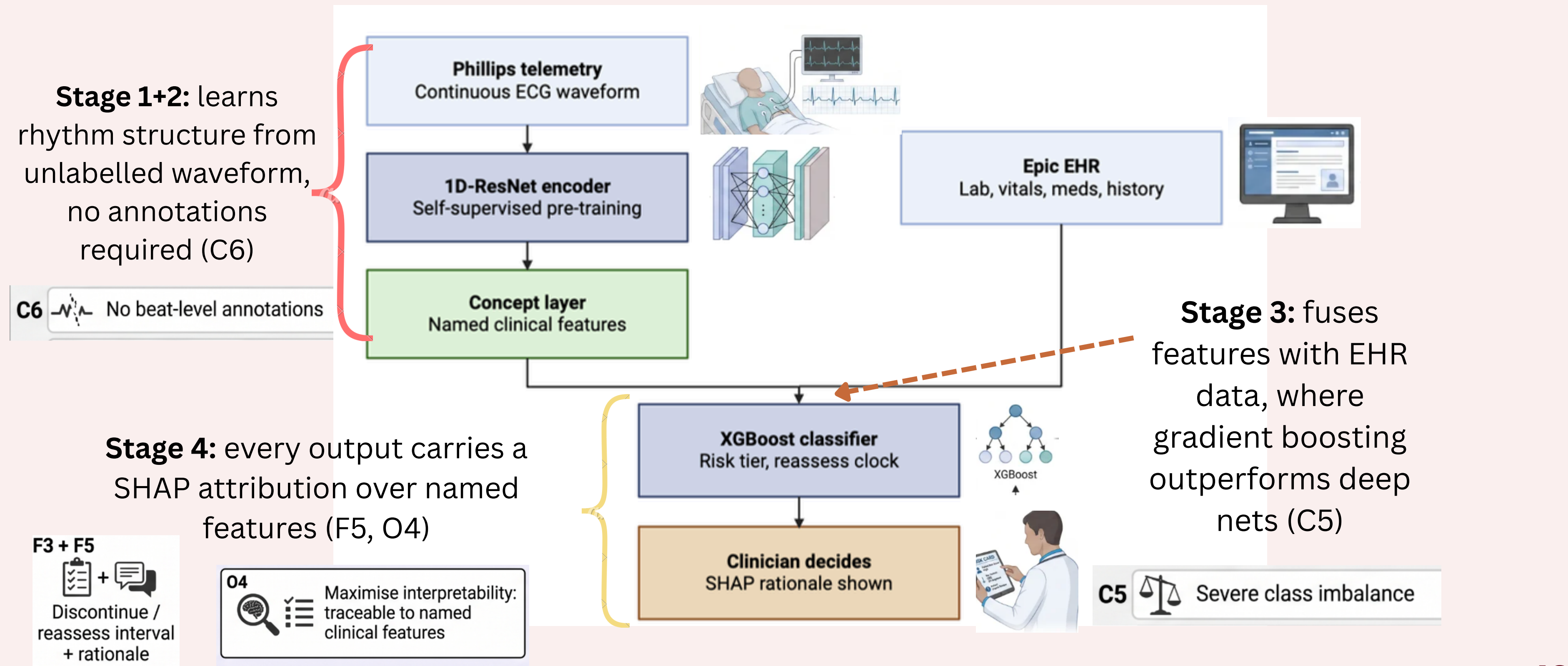


**SHAP tells you feature importance for the model.
How do you know that's the actual clinical reason?**

SHAP explains the model, not the physiology, and explanations can use model outputs while being misleading. That is why we ensured the encoder named clinical concepts rather than explaining raw embeddings.



Two-stage hybrid: waveform encoder into tabular classifier



 Why not just use one model?

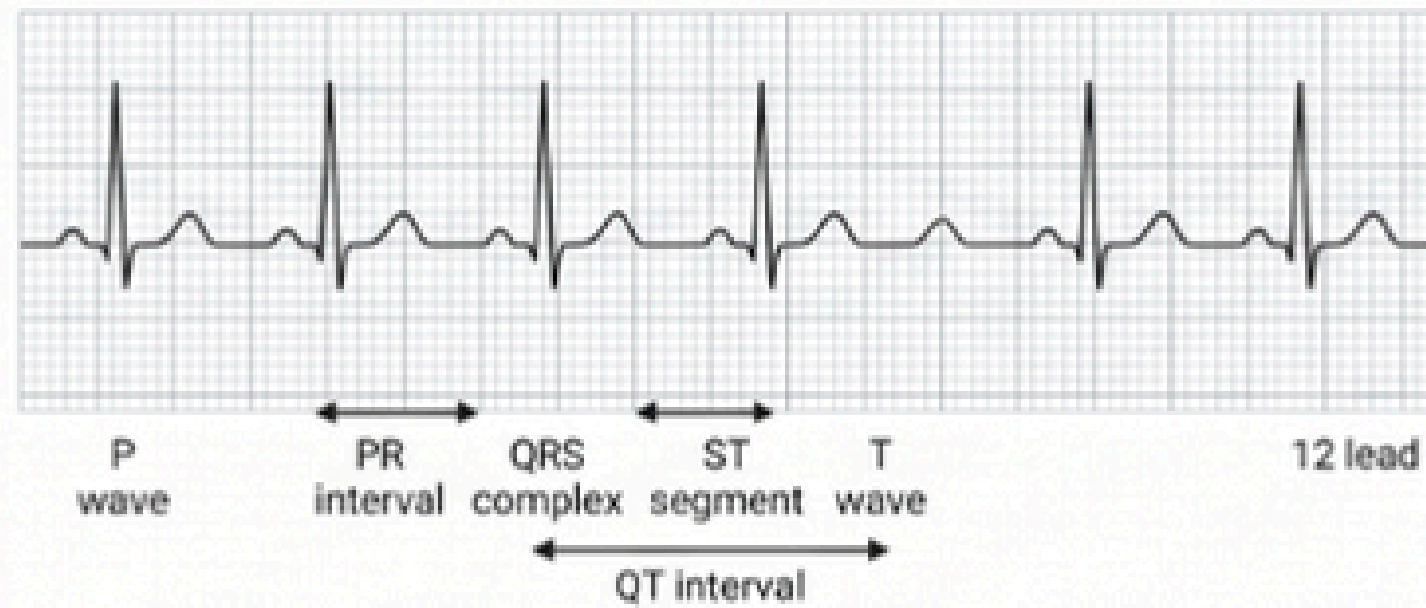
Within the encoder family we compared plain 1D-CNN (no long-range memory), SlowFast (multi-scale but heavier), and DNN-RB with a bidirectional recurrent head.

Option	Uses waveform	Handles sparse EHR	Interpretable
AHA rules-only CDS (current practice)	No	N/A	Yes
Tabular ML on EHR only	No	Yes	Yes
End-to-end deep learning	Yes	No	No
Hybrid encoder to GBT	Yes	Yes	Yes

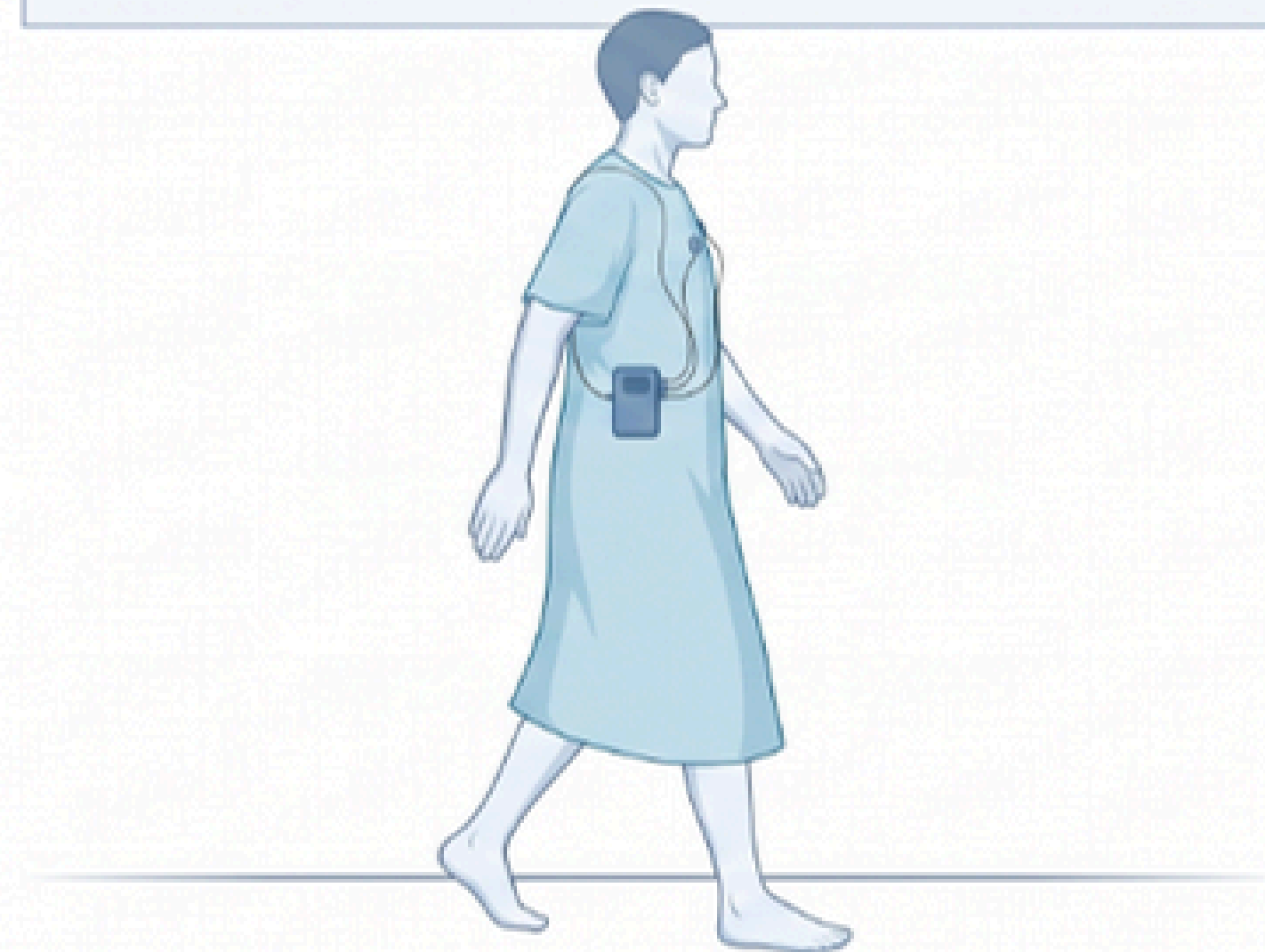
DNN-RB was selected for single-lead-II compatibility with tele and for its residual depth without gradient collapse.

Clean ECG vs. Telemetry, issue with raw data

10-second ECG — clean, annotated, 12 lead

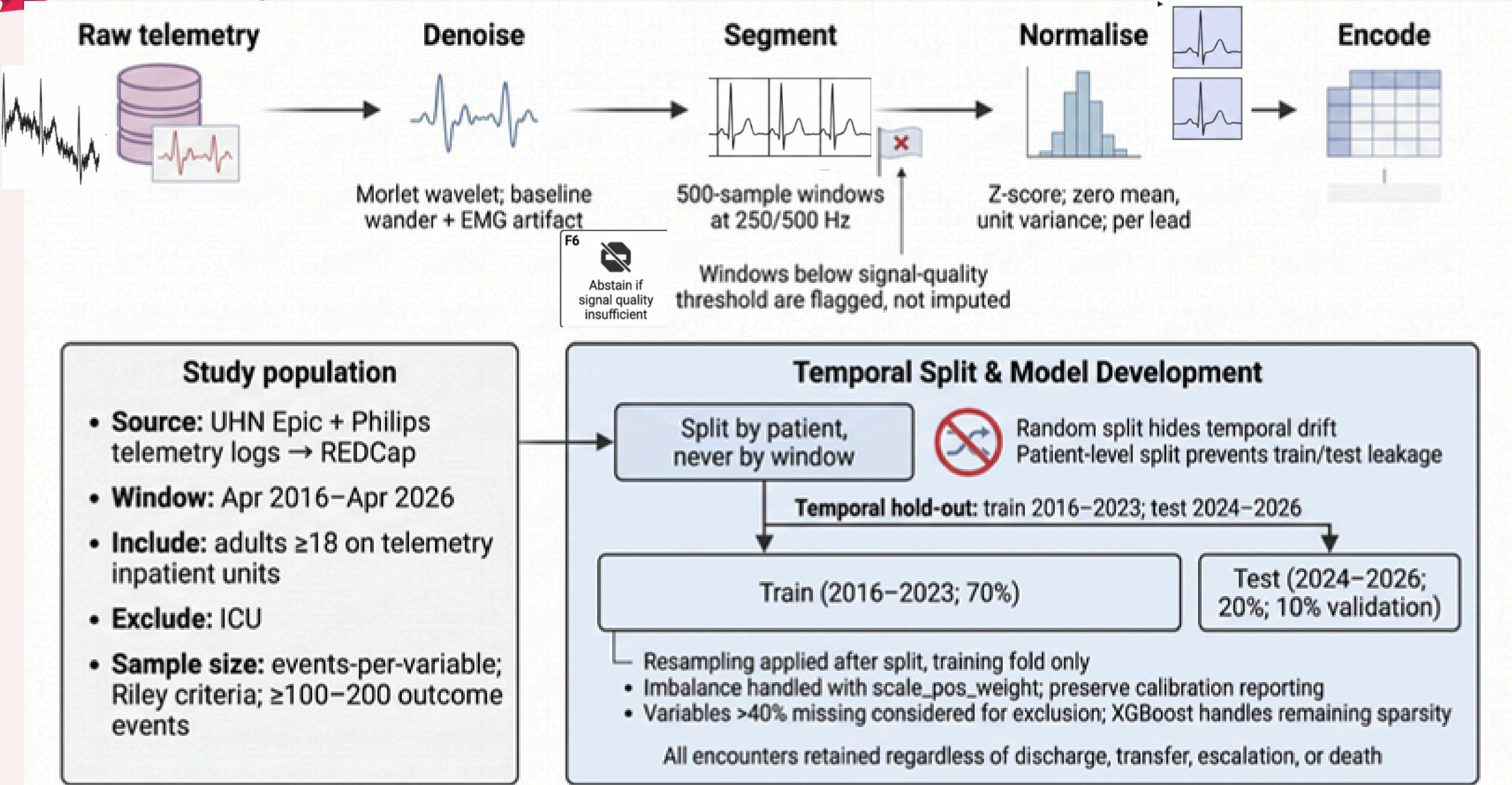


72-hour telemetry — noisy, unlabelled, single lead

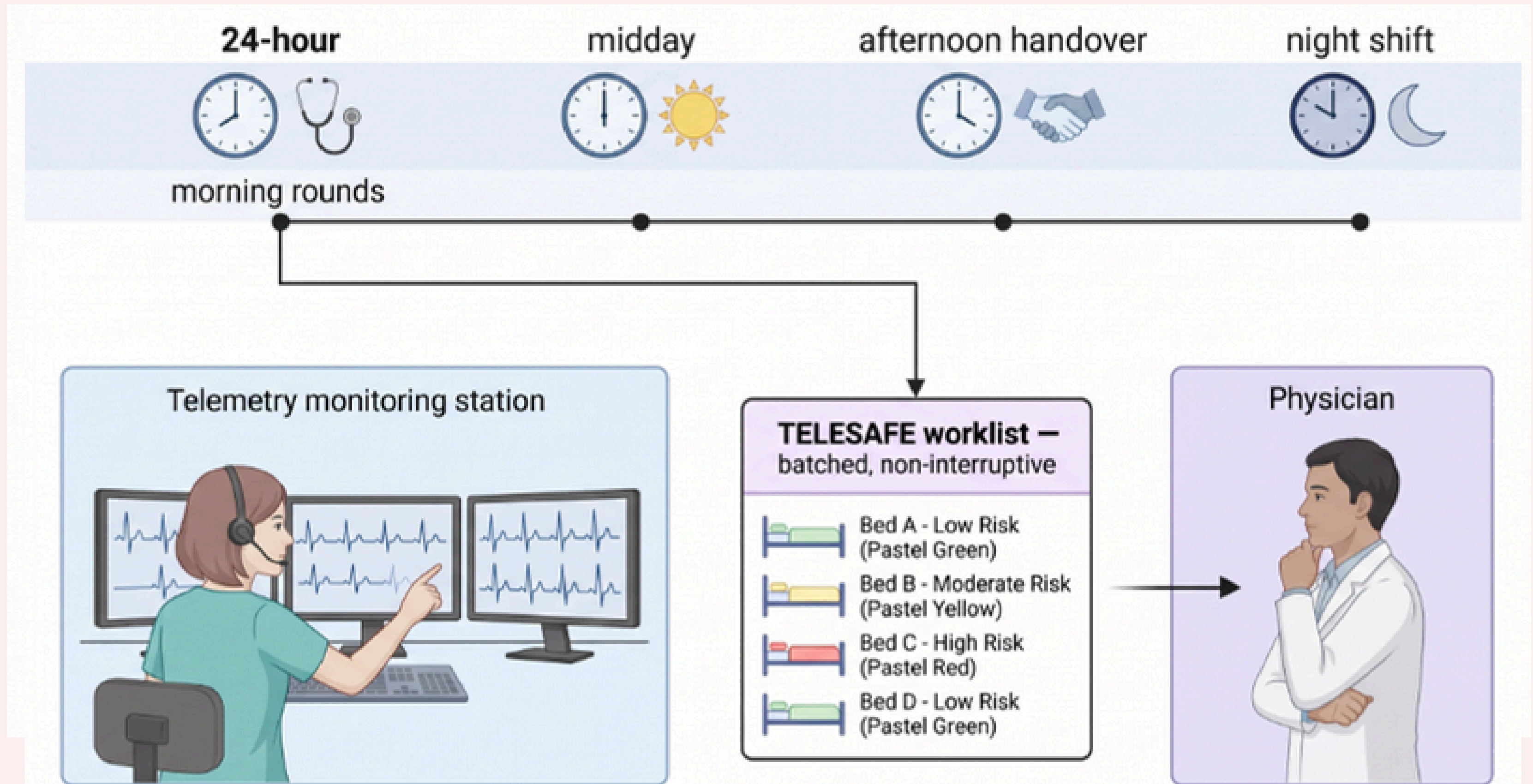




Raw signal -> model input



Where it is used in the ward day





Window Geometry and PAC ceiling. Why not widen the window?

We could, and that's the obvious next step. We used four timing features deliberately as a minimal ablation, to isolate the mechanism rather than compete on benchmark performance. Widening the window changes several things at once.