

Time-Series, Monitoring, and Real-Time Systems

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 11 of "Healthcare Analytics and AI: Building Systems That Actually Work"

The essentials: what to know, why it matters at the bedside, and how models fail.

One-sentence thesis. A monitoring system earns its place only when it respects time in two ways: the model never peeks at the future, and it fires rarely enough that a human still looks up when it does.

PART 1: The Structure of Clinical Time-Series

Irregular Sampling Across Care Settings

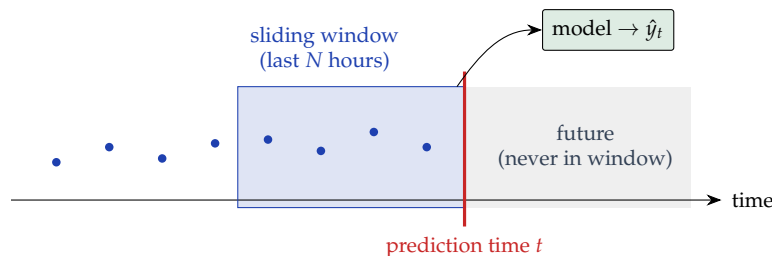
The same patient is filmed by different cameras at wildly different frame rates, and a model trained on one tape cannot read the others. In the ICU a heart rate is sampled every second (about 86,400 points a day); on the ward, every four to eight hours (about six points a day); at home on a watch, every few minutes with gaps for charging. A model trained on dense ICU streams silently breaks when applied to sparse ward vitals, and on the ward the *missingness itself is signal*: a vital gets rechecked hourly precisely because someone is worried.

Trajectory Beats the Snapshot

A lab value is almost meaningless without its trend. A creatinine of 1.8 mg/dL is an emergency if the base-

line was 0.9, an unremarkable baseline in chronic kidney disease, and *good news* if it is falling from 3.2. A speedometer reading 40 mph tells you nothing about safety until you know whether the car is accelerating toward a wall or coasting to a stop. Models that use only the most recent value systematically underperform models that ingest the whole trajectory, and because vital signs move together (a falling blood pressure drives a compensatory rise in heart rate), treating each stream as independent misses the earliest coupling patterns of deterioration. The standard fix is a sliding window: feed the model the last N hours as a sequence. **Watch out.** Choose N deliberately, and never let a window straddle the prediction time point.

Care setting	Interval	Points/day	Modeling implication
ICU	Every second	~86,400	Dense; ready for sequence models
Step-down / ward	Every 4–8 h	~6	Sparse; missingness is signal
Home wearable	5–15 min, gaps	~200	Missing-not-at-random; calibration risk



A sliding window. Only the last N hours (blue) feed the prediction at time t ; the window must end at or before t so no future information (grey) ever leaks in.

■ The Answer Key on the Desk

Leakage is any feature that would only be knowable *after* the outcome it is meant to predict, and it is the single biggest inflator of clinical model performance. It is the modeling equivalent of the answer key drifting onto the student's desk minutes before the exam ends: the score looks brilliant and means nothing. A 2025 University of Chicago study (Ramadan et al., *JAMA Network Open*) found 40.2 percent of models predicting same-admission outcomes used ICD diagnostic codes, which are assigned at discharge by coders reading the finished chart. Strip that leak and a model that showed an apparent AUC near **0.95** collapses to some-

thing like **0.72**, which is the real number.

The subtler form is *clinical action leakage*: “blood culture ordered” predicts sepsis only because the physician already suspected sepsis. The model learns to read the doctor's mind, not the disease, and falls apart the moment it must predict independently. Three disciplines defend against it. First, define an exact prediction time point and use only data available at or before it. Second, split temporally, not by random shuffle within a stay. Third, audit every feature for causal direction and drop consequences of the outcome, however predictive. **Watch out.** The clean 0.72 is the number to report; the 0.95 is a lie.

Two Different Questions

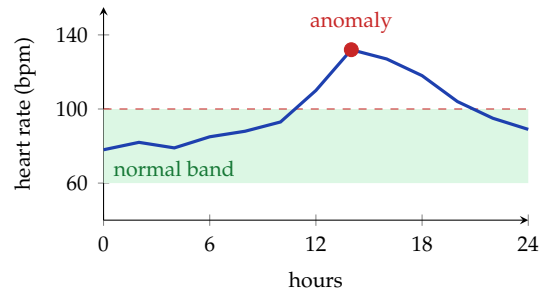
Two questions that look alike need different machinery. Forecasting asks “what will this value be next?” Anomaly detection asks “is this value unexpected given everything else about *this* patient?” A heart rate of 120 is alarming at rest and entirely expected at 39.5 degrees Celsius of fever.

ANALOGY

Forecasting is predicting tomorrow’s weather; anomaly detection is a smoke alarm that has learned what your kitchen normally smells like at dinnertime.

The best anomaly detectors learn a *patient-specific* baseline. An SpO2 of 88 percent is an emergency in most adults and the stable home baseline of a severe COPD patient, and population thresholds that ignore that are

a primary engine of alarm fatigue (Part 6). **Watch out.** A Transformer that beats an LSTM on MIMIC-Sepsis is not automatically deployable; a heavier model is a heavier maintenance and failure surface.



Anomaly detection against a patient-specific band. The reading is normal until it breaks out of the expected range; the flagged point is what an alert should fire on, not every value above a fixed population threshold.

Model	Strength	Weakness	Best clinical use
ARIMA	Interpretable, cheap	Linear only	Stable patient, short horizon
LSTM/GRU	Captures slow deterioration	Training cost	Multi-day trends
Transformer	Global attention, handles treatments	Data-hungry	Multi-step with interventions

NEWS2: Why a Lookup Table Still Wins

Before reaching for a deep model, respect what a lookup table already achieves. The National Early Warning Score (NEWS2) sums points from six vital signs into a 0-20 total, a table a nurse computes with a pen, yet it reaches AUROC 0.85-0.90 for cardiac arrest, unplanned ICU admission, or death within 24 hours. A cheap, dependable bicycle everyone can ride beats a Formula 1 car that needs a pit crew, in every hospital that lacks a pit crew, which is most of them. NEWS2 is transparent (you see which vital drives the score), robust (works on paper, with no EHR or network), and validated across hundreds of thousands of patients, so a deep model must prove better *clinical outcomes*, not merely a higher AUC.

Sepsis EWS: Epic vs. TREWS

The cautionary counterexample is Epic's proprietary sepsis model, deployed across hundreds of hospitals and found on external validation (Wong et al., 2021, *JAMA Internal Medicine*) to have only **33 percent** sensitivity while alerting on **18 percent** of *all* hospitalized patients. It was a car alarm that shrieks whenever a truck drives past: people stop checking, so it fails on the one night the car is actually stolen. The failure was not discrimination but the operating threshold, and dropping alerts into a workflow already saturated with noise, and a confidently wrong model is worse than none because it manufactures false safety. Systems clinicians actually engage with (the Johns Hopkins TREWS line of work) help only when the alert lands with an actionable next step. **Watch out.** Judge an EWS by its false-alarm burden at a fixed sensitivity and by whether the alert attaches to an action, never by AUROC alone.

NEWS2 score	Risk	Action
0-4	Low	Routine monitoring
5-6 (or any single = 3)	Medium	Urgent clinical review
7 or above	High	Emergency response; consider ICU

Sequential Decisions as an MDP

Supervised learning takes a snapshot; reinforcement learning tries to steer the film. Many treatments are sequences, not single choices, raise the vasopressor, reassess in two hours, taper if the pressure holds, and RL is the natural language for that. It frames care as a Markov Decision Process, which is really a very formalized ICU flow sheet: states (where the patient is now), actions (what you can do next, such as a dose choice), a transition function (how physiology is likely to respond), a reward (survival, organ preservation), and a discount factor γ . The goal is a *policy*, a state-to-action map, that maximizes expected discounted cumulative reward:

$$\pi^* = \arg \max_{\pi} \mathbb{E} \left[\sum_{t=0}^T \gamma^t R(s_t, a_t) \right]$$

Offline Q-Learning and Its Ceiling

In healthcare, RL is almost always *offline* (batch): you cannot ethically experiment with vasopressor doses on live patients, so the agent learns from historical records where different clinicians made different choices.

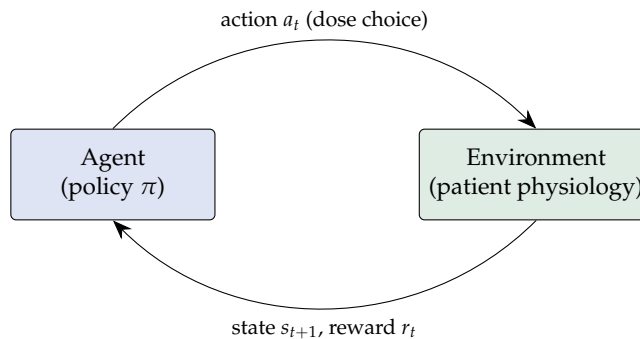
ANALOGY

Offline RL is learning to drive only by watching dashcam footage of thousands of other drivers. You can spot patterns, but you never got to try the turn yourself.

The workhorse is Q-learning, which estimates the value of taking action a in state s and then acting optimally thereafter:

$$Q(s, a) = \mathbb{E} [R(s, a) + \gamma \max_{a'} Q(s', a')]$$

Komorowski et al. (2018, “The Artificial Clinician”) modeled sepsis as 750 states and 25 actions and found clinicians tended to give too much fluid and too little vasopressor, and offline RL is now scaling from single ICU episodes to chronic care (the 2025 Dura-max lipid-management framework learned from 3.6 million patient-months). **Watch out.** Observational data hides confounding (patients who got less aggressive treatment may have been less sick), reward functions are crude (survival ignores organ damage and quality of life), and out-of-distribution actions have unknown consequences. No RL policy has been validated prospectively; treat every one as decision *support* with hard safety constraints (never dose insulin below 80 mg/dL regardless of the Q-values). Dynamic treatment regimes and SMART trials are how a suggestive policy becomes a causal finding (Chapter 13).



The reinforcement-learning loop. The agent chooses an action, the patient’s physiology returns a new state and a reward, and the goal is a policy π that maximizes discounted cumulative reward.

Alarm Arithmetic and Asymmetric Liability

A single ICU bed can generate roughly **1,578** alarms a day, and 70-99 percent are non-actionable. It is a fire alarm that shrieks every time someone makes toast: the rational human response is to pull the battery, and then the real fire is invisible. The Joint Commission documented 80 alarm-related deaths over three and a half years, likely a tenfold undercount, and better algorithms cut false alarms 30-60 percent but do not fix the root cause, because the root cause is an incentive, not a signal.

The liability is asymmetric. A missed alarm creates lawsuits and regulatory action, borne by the manufacturer and the hospital; a false alarm creates only nurse overload, borne by a clinician with no vote in the threshold. So thresholds stay conservative and the noise persists. Alarm fatigue is a liability and cultural problem wearing a signal-processing costume.

Case: ChatGPT Health and the Apple Watch

In January 2026, 40 million daily users began feeding consumer wearable data to an LLM expecting clinical interpretation. Reporters found it assigned an “F” cardiac grade, misread Apple Watch data, and produced *different* scores on repeated submissions of identical input, like a confident stranger reading your fortune

from a fitness tracker. The FDA cleared the Apple Watch for narrow tasks (AFib detection, sleep apnea, hypertension risk), not comprehensive health assessment, so wellness metrics carrying 5-10 percent error get extrapolated into clinical judgments no validation supports, in a regulatory vacuum. The device generates data; the surrounding system decides whether that data becomes safe care.

Should You Build This?

Build it when the signal is genuinely temporal and someone can act on it in time: an arrivals forecast a charge nurse staffs against, or an early warning score built without label leakage, monitored for drift, that fires rarely and is right when it fires. **Think twice when** your alert joins a stream already running over a thousand alarms per bed per day, 70 to 99 percent of them non-actionable, or when the input is a consumer sensor carrying 5 to 10 percent error that no regulator cleared for the question you are asking. **Who bears the risk:** bedside nurses, who inherit every non-actionable alarm and the blame when a silenced one was real; patients, harmed as surely by alarm noise as by a missed event; and consumers acting on authoritative-sounding interpretations no regulator has validated.

Quick Self-Test

1. A model hits AUC 0.95 in cross-validation using “blood culture ordered” as a feature. What is wrong, and what will the real-world AUC look like?
2. Why does the same vital-sign model that works in the ICU fail on a general ward, and what does missingness tell you there?
3. NEWS2 is “just a lookup table.” Give three concrete reasons it still beats deep models in real hospitals.
4. The Epic sepsis model had acceptable standalone discrimination. Why did it fail at the bedside anyway?
5. Why is clinical reinforcement learning almost always offline, and what two failure modes does that create?
6. Explain, using incentives rather than algorithms, why alarm fatigue persists after twelve years of recognition.

If You Remember One Thing

Respecting time means two separate disciplines: never letting future information leak into a feature, and never letting so many alerts through that a clinician stops looking. A model can be technically excellent and still kill, because a monitoring system is judged by what a tired human does at 3:47 a.m., not by its AUROC.

Richard J. Young, Ph.D. · richard@deepneuro.ai · Senior AI Research Scientist, UnitedHealth Group · UNLV Lee Business School · ORCID 0000-0002-1109-7552