

The Attention Economy in Healthcare

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 6 of "Healthcare Analytics and AI: Building Systems That Actually Work"

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

One-sentence thesis. Human attention is a finite channel, so every alert you add spends a clinician's cognitive budget, and a system that spends more than it returns makes care less safe no matter how accurate its model is.

PART 1: The Channel Is Overloaded

Detection Is No Longer the Bottleneck

Start with the failure the chapter opens on: a sepsis alert fires with a correct six-hour lead time and is clicked through in under three seconds. What failed there was not detection. For most of the field's history the hard problem was catching the event, and the field largely solved it: we now flag deterioration, sepsis, arrhythmias, and dangerous drug combinations earlier and more reliably than the overloaded humans we are trying to help. What we have not solved is what happens next. A health system can always buy more compute; it cannot buy more clinician attention. **In modern clinical AI, detection is no longer the bottleneck. Attention is.** Every model you deploy is therefore a claim against a fixed cognitive budget, and that alert entered a channel that had already delivered 748 competing claims on one nurse's attention that night, so it was priced accordingly. Everything below is about learning to spend that budget honestly.

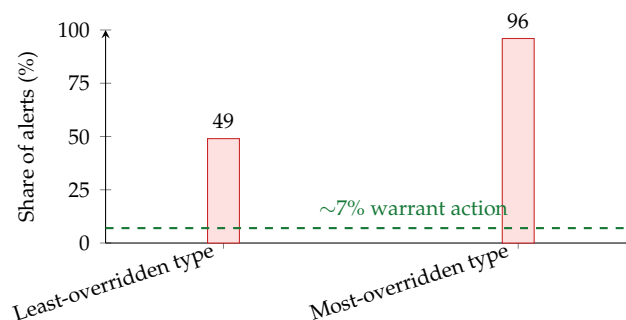
Signal-to-Noise in Clinical Alerting

Treat the clinician as a communication channel and the whole problem comes into focus. Shannon (1948)

framed information as the resolution of uncertainty through a finite-capacity channel: here the EHR is the transmitter, the clinician is the receiver, and alerts are the traffic. Signal is an alert that changes behavior; noise is everything else, and forcing every message through one blaring klaxon means the urgent page still arrives but sounds exactly like the fifty pointless ones, so no one can pick it out. Engineers measure that ratio in decibels:

$$\text{SNR}_{\text{dB}} = 10 \log_{10} \frac{P_{\text{signal}}}{P_{\text{noise}}}$$

In one study only 7.3 percent of passive medication alerts were clinically appropriate and 92.9 percent were overridden, roughly a -11 dB channel. At -11 dB a radio engineer would declare the channel useless: push more data through a channel than it can carry and signal is mathematically guaranteed to be lost in noise, which is precisely the state of modern clinical alerting.



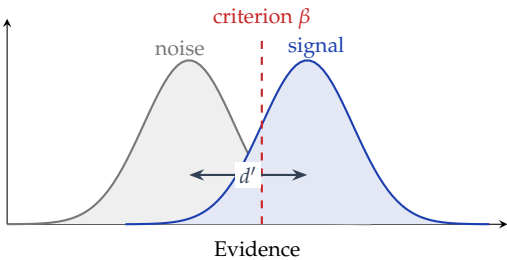
Even the least-ignored category of drug-safety alert is overridden about half the time: rates run 49 to 96 percent across alert types (van der Sijs, 2006). Fewer than 1 in 13 passive medication alerts (green line) actually warrant attention.

Signal Detection Theory (d-prime and Criterion)

Why do good clinicians ignore real alerts? Signal detection theory answers it by splitting every yes/no decision into two independent knobs. Sensitivity is how far apart the signal and noise distributions sit, a property of the system; response bias (the criterion, beta) is how much evidence the human demands before saying “this is real,” a property of the person. A smoke detector’s sensitivity is how well it tells toast from a house fire; your criterion is whether you still get up to check after it has chirped falsely all week. The system half is captured by d' :

$$d' = z(\text{hit rate}) - z(\text{false-alarm rate})$$

The four possible outcomes are hit, miss, false alarm, and correct rejection. When 92.7 percent of alerts are false, clinicians rationally raise their criterion, demanding more before they interrupt their workflow. That is a mathematically optimal adaptation to a broken system, not negligence, but as beta rises the miss rate rises with it. The 3:47 a.m. nurse who dismisses the real alert was trained to do exactly that.



Signal detection theory. d' (a property of the system) is the separation of the signal and noise distributions; the criterion β (a property of the human) is where they decide “real.” Raising β cuts false alarms but raises misses.

	Condition present	Condition absent
Alert fires	Hit (true positive)	False alarm (false positive)
No alert	Miss (false negative)	Correct rejection (true negative)

Three Types of Cognitive Load

Not all mental effort is equal, and design decides which kind a clinician pays. Sweller's Cognitive Load Theory says three loads compete for the same working memory, like a small kitchen countertop. Intrinsic load is the meal you must cook, the irreducible complexity of the task (tracking vitals, dosing, sensitivities). Extraneous load is clutter someone left on the counter, friction from bad design (six clicks to a lab,

four pop-ups before an order), and it is pure waste. Germane load is learning a better technique, the effortful reasoning that builds expertise. The design mandate is asymmetric: you cannot cut intrinsic load, you must ruthlessly cut extraneous load, and you must protect germane load. If a system adds information without removing at least as much extraneous load, it is net-negative for patient care regardless of model accuracy.

Load type	Definition	Design response
Intrinsic	Complexity inherent to the task	Cannot reduce; accept and support
Extraneous	Burden from poor design	Minimize ruthlessly
Germane	Effort building mental models	Protect; do not automate away

The Four-Slot Limit

That working memory is smaller than most designers assume. Miller's famous 1956 "seven plus or minus two" was revised down by Cowan (2001) to about **four plus or minus one** for novel, unrelated items, and in a high-acuity setting those four slots are the entire real-time decision space, four burners on a stove where lighting a fifth dish knocks something already cooking to the floor. An ICU physician computing a vasopressor drip (2 to 3 slots) who gets a pop-up for a drug interaction she already accepted (1 to 2 slots) has almost nothing left for the ventilator alarm, the new lab, and the nurse's question. That single useless alert consumed **30 to 50 percent** of her bandwidth for zero clinical value.

Interruption Cost and Resumption Lag

The real price of an interruption is not the interruption, it is the resumption lag afterward, the way

knocking over a house of cards does not just pause your build but collapses the structure you have to rebuild from scratch. Altmann and Trafton (2002) found it takes 15 to 25 minutes to return to full cognitive depth. Task-switching adds a switch cost (disengage and reorient) plus a residue cost, the fragments of the interrupting task that linger, Leroy's "attention residue." Westbrook (2010) found each interruption during medication administration raised procedural errors **12.7 percent** and clinical errors 12.1 percent, near-linearly, with nurses interrupted about once every two minutes; Chisholm found ED physicians interrupted nearly ten times an hour. The infamous three-second dismiss is too fast to read an alert, let alone judge it, the dismiss has become a motor reflex, not a decision.

PART 3: The Alert Fatigue Crisis

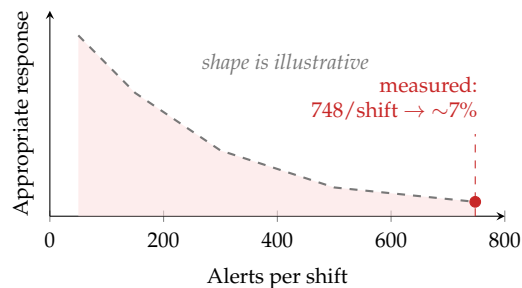
The Numbers and the Body Count

The scale is hard to believe until you count it. At 187 alerts per ICU patient per day, a nurse fields about 748 alerts per 12-hour shift, roughly one every 58 seconds; override rates run 49 to 96 percent across drug-safety alert types (van der Sijs, 2006), and fewer than 1 in 13 passive medication alerts warrant attention.

ANALOGY

The boy who cried wolf, automated and running 748 times a shift. When the real wolf appears, it looks exactly like every false alarm before it.

Alert fatigue is directly implicated in patient deaths; in a 2013 Baylor Scott and White case a pharmacist overrode a fatal drug-allergy alert after overriding hundreds of identical false ones that week. The instinct to “alert on everything so we never miss anything” is precisely backwards, because signal detection theory says raising false alarms without raising sensitivity makes the system worse. The crisis is driven by liability-driven vendor defaults, compliance incentives that rewarded turning alerts on, and additive bias with no mechanism for alert retirement.



Alert fatigue. Only the marked point is measured: at roughly 748 alerts per 12-hour shift, fewer than 1 in 13 alerts warrant attention. The dashed curve shows the direction of the relationship, not a measured dose-response, and the vertical axis deliberately carries no scale.

Case Study: Epic's Sepsis Model

The most-deployed sepsis model is a case study in cognitive cost, not just accuracy. Wong and colleagues' 2021 external validation (JAMA Internal Medicine) found the Epic Sepsis Model alerts on 18 percent of all hospitalized patients (about 90 alerts a day in a 500-bed hospital), misses **67 percent** of actual sepsis (sensitivity 33 percent), and has a positive predictive value of only about 10 to 12 percent, so roughly 9 of 10 flags are false. It is a metal detector at the beach that beeps on almost everyone and still misses two-

thirds of the actual coins, so you quickly stop digging. Internal validation looked fine due to temporal leakage, outcome-definition contamination (features triggered by the clinician already suspecting sepsis), and a threshold set so low that 18 percent of patients tripped it. A model with the worst of both worlds, a firehose of false positives and inadequate true detection, is clinically worse than no model, because it consumes the very attention that direct patient assessment (which catches the 67 percent it misses) would have used. Model accuracy divorced from cognitive-cost analysis is not clinical utility.

Metric	Reported value	Operational consequence
Alert rate	18 percent of patients	~90 alerts/day in a 500-bed hospital
Sensitivity	33 percent	Misses ~2 of 3 sepsis cases
Positive predictive value	10 to 12 percent	~9 of 10 alerts are false

The CDS Five Rights

A useful alert is not just correct, it is correctly aimed, the difference between a good text message and a shouted announcement in a crowded hall. Campbell's Five Rights (2013) say a clinical decision support intervention must deliver the right information (actionable, patient-specific, not already known), to the right person (who can act), in the right format (name, risk, action, one-click path), through the right channel (urgency-matched modality), at the right time (the moment of decision). Violating any one converts a useful alert into extraneous load. Telling an allergist a patient has a penicillin allergy they themselves recorded adds nothing; telling a covering physician who has never seen the patient adds value. Most current systems violate several Rights at once and force every alert through the same interruptive pop-up, the design equivalent of making every email an all-caps urgent flag.

Tiered Alerting and the Subtraction Principle

The fix is to stop treating alerts as binary and match the interruption to the consequence, the way a hospital PA system tiers a code blue that overrides everything against a visitor reminder that is a quiet lobby sign. Tier 1 hard stops (life-threatening, non-dismissible) should be under 1 percent of alerts; Tier 2 interruptive alerts (significant, documented override) about 5 to 10 percent; Tier 3 passive notifications (badges, in-basket); Tier 4 silent logging (regulatory only, never shown). The subtraction principle enforces discipline: every alert added requires removing or downgrading one existing alert. Most systems behave as if every alert is Tier 1, but the few institutions that run quarterly reviews, measure override rates by type, and retire alerts above a 90 percent override rate see lower fatigue and higher compliance with the alerts they keep. Interruptibility detection (task phase, click cadence, time of shift, concurrent alert volume) further defers non-critical alerts to natural task boundaries, where the same alert produces fewer errors.

Tier	Severity	Target volume	Delivery
1	Life-threatening	Under 1 percent	Non-dismissible hard stop
2	Clinically significant	5 to 10 percent	Interruptive; documented override
3	Informational	Bulk of remainder	Passive badge or in-basket
4	Logging only	Silent	Audit log; never shown

Clinical Deskilling

The subtlest cost is not to attention but to expertise: AI meant to augment reasoning can quietly erode it. Expertise is built by germane load, so when an AI hands over the differential automatically, the resident learns to check the machine instead of generating her own answer, and the pathways for independent reasoning atrophy.

ANALOGY

The aviation “Children of the Magenta”: pilots who could follow the automated flight-path line but could not fly the plane when the automation dropped out.

ECRI's 2026 hazard list ranked AI chatbot misuse the number one health technology hazard and named deskilling as a contributing factor, and the Wolters Kluwer 2026 Shadow AI report showed unauthorized AI already in routine use. The danger is recursive:

more reliance erodes skill, eroded skill increases reliance and reduces the human's ability to catch the machine's errors, so the AI ends up with less oversight than it was designed for. The design fix is to automate the extraneous tasks (documentation, order formatting, prior auth) but preserve the germane ones, for example by showing the AI's differential only after the clinician commits her own assessment.

Should You Build This?

Build it when detection is genuinely solved and you are prepared to engineer the delivery: an operating threshold chosen against the unit's existing alert burden, tiered presentation matched to urgency, and the Five Rights governing who gets interrupted, how, and when. Sometimes the highest-value move is subtraction, retiring alerts to buy back attention for the ones that matter. **Think twice when** the plan is a binary pop-up added to a channel already firing 187 alerts per patient per day, or when validation is internal

only, the way Epic's sepsis model looked strong on vendor numbers and degraded badly under external evaluation. **Who bears the risk:** the clinician holds the pager and the liability, since every false alarm

taxes her attention and every missed true alarm is charged to her judgment, and the patient in bed 3 absorbs the outcome when the one correct alert arrives priced at three seconds.

Quick Self-Test

1. In signal detection terms, which parameter belongs to the system and which to the human, and what happens to the miss rate when a clinician raises their criterion?
2. Why is a useless pop-up during a vasopressor calculation more costly than it looks, in working-memory terms?
3. The cost of an interruption is not the interruption itself. What is it, and roughly how long does it last?
4. The Epic Sepsis Model has an "acceptable" internal AUC. Give two reasons its real-world deployment still fails, using specific numbers.
5. State the subtraction principle in one sentence and explain why additive bias makes alerting worse over time.
6. Name one task that is safe to automate and one that is dangerous to automate, and explain the difference using cognitive load.

If You Remember One Thing

The central constraint in healthcare AI is not computational, it is cognitive. A model with 99 percent AUC that fires 200 alerts a day is clinically weaker than an 85 percent model that fires 10, because the second is more likely to be read, trusted, and acted on. Every system you build must answer one question: does it respect human attention or consume it?

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