

FOR AI COMMITTEES

Reading a vendor validation study in 15 minutes

What to look at first. Six questions. And three tells that a study doesn't say what its summary claims.

Written for the committee member who isn't a methodologist and has fifteen minutes.

“FDA-authorized” and “shown to work in patients like yours” are two different questions.

A 2024 review in [Nature Medicine](#) looked at 521 FDA-authorized AI-enabled devices cleared through 2022. **226 of them — 43% — had no clinical validation data behind the authorization at all.**

Of the 292 that did, 144 were retrospective only. Twenty-two had a randomized controlled trial.

WHAT THAT MEANS FOR YOU

The second question is yours to ask. Nobody upstream asked it on your behalf, and the vendor's one-page summary is not the study.

Read four things before you read the abstract

1 The table of participant characteristics

Usually Table 1. Age, sex, race, payer mix, site of care. If those patients don't look like yours, the headline number is about somebody else.

2 The flow diagram

How many patients were screened, how many excluded, and why. Large unexplained exclusions are where performance gets manufactured.

3 The funding and author affiliation statement

You're checking whether this is independent evidence or the developer's own evaluation. Both are useful and they are not the same thing.

4 The dates

When the data was collected, and which version of the model was studied.

Three minutes here will tell you whether the next twelve are worth spending.

One to three

ASK THIS

A REAL ANSWER LOOKS LIKE

1. Is this development, or evaluation?

A separate dataset: different site, different time period, ideally both. A random split of one dataset is development wearing an evaluation's clothes.

2. Who is in the denominator?

Consecutive eligible patients across a defined window. “Cases and matched controls” inflates every discrimination metric relative to your clinic.

3. Compared against what?

Current practice at the study site, measured. Not “no model,” which nobody was offering, and not a benchmark from a different population.

Four to six

ASK THIS

A REAL ANSWER LOOKS LIKE

4. What was the outcome rate, and what's ours?

The study prevalence, stated. Then ask for positive predictive value at **your** prevalence. That number, not AUC, decides whether your staff will trust the alert.

5. Does it hold across subgroups?

Performance broken out by age, sex, race, language, payer and site. A description of the training population's demographics is a true sentence answering a different question.

6. Is the studied version the sold version?

Model version, date, and what has changed since. Under a predetermined change control plan a model can move legitimately. The study does not move with it.

When the summary and the study disagree

“EXTERNAL” THAT ISN'T

Same institution, same years, random split — described in the abstract as validation.

Check the methods for a second site or a later time window. If neither is there, the word is doing work the design doesn't support.

THE DENOMINATOR MOVES

The headline figure and the results table describe different populations, or different thresholds.

Read the abstract number, then find the same number in a table. If you can't, that's the finding.

THE COMPARATOR IS ABSENCE

The study shows the model beats nothing.

You are not currently doing nothing. Ask what current practice achieved on the same patients.

Some flaws survive validation because validation reproduces them

In May 2026, [npj Digital Medicine](#) published a review of 57 reinforcement-learning studies for sepsis treatment, 2017 to 2025. **Forty-nine of them — 86% — used a data preparation step that lets the model's state be realized after the action it supposedly caused.**

The part that matters for governance: the authors show the flawed and the corrected formulations “produce nearly identical results when assessed with common off-policy evaluation metrics.” Models trained and evaluated on identically misaligned data “appear internally consistent despite being clinically invalid.”

THE GOVERNANCE CONCLUSION

If a peer-reviewed literature can carry a flaw that large invisibly, a vendor's self-reported metric certainly can. You cannot inspect your way to safety. You have to monitor after go-live.

What external validation found

WHAT WAS BELIEVED

WHAT ONE HEALTH SYSTEM MEASURED

AUC of 0.76 to 0.83,
per the developer's
internal
documentation

AUC of 0.63 (95% CI, 0.62–0.64)

A widely deployed
sepsis early-
warning model,
live at hundreds of
hospitals

Sensitivity of 33% at the recommended
threshold

Earlier detection of
sepsis

1,709 of 2,552 sepsis hospitalizations (67%)
never triggered an alert

A manageable
alert burden

Alerts on 6,971 of 38,455 hospitalizations (18%);
number needed to evaluate, 8

Wong et al., JAMA Internal Medicine, June 21, 2021. 27,697 patients, 38,455 hospitalizations, Michigan Medicine, Dec 2018 – Oct 2019. The model didn't break. One institution went and looked.

Write four lines and stop

- What population was this shown to work in
One sentence. If you can't write it, that is the review finding and you can stop here.
 - How far is that from ours
Name the specific gap: age, acuity, payer mix, language, site of care.
 - What we still don't know
The questions the study doesn't answer. This becomes the evidence request, and it should be a list, not a paragraph.
 - What we would need to see locally before we trust it
Written before go-live, while you still have leverage. This is your monitoring plan's first draft.
-

Cite the standard, not your opinion

“This study is thin” invites an argument. “This submission doesn't address items 11, 13 and 20” invites a document.

THE REFERENCE TO NAME

TRIPOD+AI — a 27-item reporting checklist for clinical prediction models using regression or machine learning. Collins, Moons, Dhiman et al., BMJ 2024;385:e078378. Public, free, and written by methodologists rather than by you, which is exactly why it works in a vendor conversation.

You are not asking the vendor to run a new study. You are asking them to report the one they ran.

You don't need to name the flaw to reject the study.

The most useful reviewers I've watched aren't statisticians. They are people who ask who was in the study, what it was compared against, and whether the version tested is the version being sold — and who are willing to sit in the silence after the question.

Troy Bannister — Founder & CEO, Onboard AI

Steal this. Change it. Tell me what I got wrong and I'll reissue it.