7 Reading an AI study critically

AUTHOR
R. Heller

R. Heller · Date: 2026-07-26 · Last changed: 2026-11-09

At a glance

Guiding question: Which findings support measurement quality, and which support better care?

For: clinicians preparing a journal club, members of committees that decide on clinical software, doctoral researchers and peer reviewers · **Prerequisites:** Chapter 5 (validation designs) and Chapter 6 (2 × 2 table, calibration, thresholds)

In the course: W2 · Language models and evidence and W3 · Diagnostics, prediction and evidence · **Data:** *published* figures from one external validation study and a *synthetic* (fictitious) abstract; no patient data · **Core reading time:** about 11 minutes

7.1 Case and decision

The committee for clinical decision support of a fictitious hospital meets in two weeks. On the agenda is a request to switch on a sepsis warning that comes with the hospital's electronic health record and, according to the supplier, is used in many hospitals. A resident who runs the internal medicine journal club is asked to prepare a one-page brief. The committee wants to know two things: how well does the score recognise sepsis, and does its use lead to better care for patients? The resident finds one independent external validation of a widely implemented proprietary sepsis model, a set of reporting guidelines with unfamiliar acronyms and a supplier brochure. The recommendation — switch on, pilot under observation, or decline — will change the work of every ward. A brief that confuses accuracy with benefit could commit the hospital to thousands of alerts without any evidence that patients gain.

7.2 Core reading

Core reading · required for W2 · about 11 minutes · ends before 'Lab'

A ladder of evidence. Studies of medical tests and models answer questions at different levels. Fryback and Thornbury described a hierarchy for diagnostic imaging that still serves well: technical quality, diagnostic accuracy, effect on the clinician's diagnostic thinking, effect on management, effect on patient outcomes, and societal costs and benefits. Evidence at a lower level is necessary but not sufficient for a claim at a higher level (Fryback and Thornbury 1991). An AI system climbs the same ladder. A *reader study* or *diagnostic accuracy study* compares the system's output with a reference standard. An *external validation* applies a prediction model to patients from another place or time and measures discrimination and calibration there (Chapter 6). An *early clinical*

evaluation places the system into a real workflow on a small scale and asks how clinicians use it, when they override it and what goes wrong. Only a *comparative study*, ideally randomised, can show whether patients are better off with the system than without it.

The higher rungs are sparsely populated. A systematic review of deep learning studies in medical imaging that compared algorithms with clinicians found only 10 records of randomised trials, two of them published; of 81 non-randomised studies only nine were prospective, 58 were at high risk of bias, and 61 stated in their abstract that the algorithm performed at least as well as clinicians (Nagendran et al. 2020). Claims in abstracts can therefore run ahead of the design.

Four tools, four tasks. Reporting guidelines list what a paper must report so that a reader can judge it; an appraisal tool guides the judgement itself. Table 7.1 assigns each of the four AI-specific tools to one concrete check. The item numbers refer to the published checklists.

Table 7.1: Four tools for four questions about clinical AI. For randomised trials of AI interventions, CONSORT-AI adds 14 items to CONSORT 2010 (Liu et al. 2020).

Study type and question	Tool	One concrete check	What the study cannot show
Diagnostic accuracy: does the output agree with the reference standard?	STARD-AI, reporting (Sounderajah et al. 2025)	How were the data partitioned into training, validation and test sets (item 15b), and does the test set represent the target condition and population (items 25, 28)?	Whether clinicians decide differently or patients do better
Prediction model development or evaluation: how well does the model discriminate and calibrate?	TRIPOD+AI, reporting (Collins et al. 2024)	Are all measures for discrimination, calibration and clinical utility specified (item 12e) and reported with confidence intervals, including key subgroups (item 23a)?	Whether acting on the prediction changes outcomes
Same studies, appraised: can the result be trusted and does it apply here?	PROBAST+AI, quality, risk of bias and applicability (Moons et al. 2025)	Model evaluation, domain 4: was data leakage avoided (4.5), were all development steps repeated in resampling (4.6), were calibration, discrimination and net benefit evaluated (4.7)?	Anything beyond the population and setting of the evaluation
Early clinical evaluation in live use: how is the system used, and is it safe?	DECIDE-AI, reporting (Vasey et al. 2022)	Is user agreement reported, with instances of and reasons for variation from the recommendation (item 12), and are errors and risks to patient safety listed (item 13)?	Comparative effectiveness

A missing item is not a verdict. It is a question to the authors, the supplier or the local team. If a paper on a prediction model reports an AUC but no calibration, the right response is “How do predicted and observed risks compare in your test set?”, not silent acceptance or silent rejection. Such questions can go to the authors through the journal, to the supplier as part of procurement, or to the local team as the first item of a plan for a silent evaluation in the hospital’s own patients. The answer may change the verdict in either direction.

Five questions for the journal club. Whatever the study type, five questions organise the reading, and the lab turns them into a card. *Recruitment:* who was included, from where and when, and who was left out? A test set collected from patients with clear-cut disease and clear-cut health tends to overstate accuracy for the mixed patients of everyday care. *Reference:* how was the target condition or outcome established, by whom, and without knowledge of the model's output? Discharge codes, for example, capture only episodes that were documented and coded. *Analysis:* were tuning, model choice and testing kept apart (Chapter 5), and are discrimination, calibration and a threshold reported with denominators (Chapter 6)? *Generalisation:* does the evaluation population resemble the patients on whom the model would be used here — case mix, measurement devices, time, care pathway? *Decision:* does the conclusion stay at the level the design reached, or does an accuracy study recommend implementation "to improve outcomes"? Each question can be answered from the paper, from a supplement, or not at all; the last answer is itself a finding.

One statement, checked against its source. The committee brief will contain numbers, and each number needs a source that can be checked. Here is one statement, traced to its origin.

Table 7.2: A checked statement: claim, original, locator, unit and verdict. Numbers are *published* (Wong et al. 2021).

Element	Content
Statement in the brief	In an independent external validation, the sepsis model discriminated poorly (AUC 0.63) and missed two thirds of sepsis cases while alerting in almost one in five hospitalisations.
Original study	Wong et al. (2021), retrospective cohort at one US academic health system, December 2018 to October 2019
Locator	Abstract, "Results": AUC 0.63 (95% CI 0.62–0.64) at hospitalisation level; the model "did not identify 1709 patients with sepsis (67%) despite generating alerts for an ESM score of 6 or higher for 6971 of all 38 455 hospitalized patients (18%)"
Unit and denominator	38,455 hospitalisations of 27,697 patients; 2,552 with sepsis (7%)
Data status	*published*; abstract read in PubMed, DOI checked against Crossref; the article carries a correction (2 August 2021) of the number needed to evaluate in Table 2 and the Results, so only the corrected version may be quoted for that number
Verdict	Supported for measurement quality at this centre. Not evidence about patient outcomes, nor about other hospitals or later versions of the model

The lab reconstructs the 2 × 2 table behind this statement: a sensitivity of about 33%, a positive predictive value of about 12%, and roughly 8 hospitalisations with an alert for each hospitalisation with sepsis that the alert caught. The study also compared the score with contemporary care: it identified 183 of the 2,552 patients with sepsis (7%) who did not receive timely antibiotics. The authors read this as low sensitivity compared with contemporary clinical practice: the score seldom flagged sepsis that clinicians had not already treated in time. That is a question one level above accuracy on the Fryback–Thornbury ladder — whether the information changes what clinicians do.

The most important assumption when reading any AI study is that a study supports only the claim its design can answer: accuracy in a test population is not evidence of better care, and better care in one hospital is not evidence for another. The supplier's statement that the model is "used in many hospitals" is evidence of dissemination, not of benefit.

Question for the workshop: The committee asks for a single sentence. Write one sentence about the sepsis warning that is fully supported by Table 7.2, and one sentence that the brochure might contain but that the study cannot support. Which study type would be needed for the second sentence?

7.3 Lab

7.3.1 Part A · Reconstruct the 2 × 2 table from the published abstract

All inputs are *published* counts from the abstract of Wong et al. (2021). Combining them into one table rests on an *assumption* that must be stated: the counts refer to the same unit (hospitalisation), the same alert threshold (score 6 or higher) and the same time window. The abstract places them in one sentence but does not say this explicitly; the full text (Table 2, corrected version) is where a reader would confirm it.

```
n_hosp <- 38455          # hospitalisations analysed (published)
n_sepsis <- 2552         # hospitalisations with sepsis (published)
missed <- 1709           # sepsis not identified by the alert (published, 67%)
alerted <- 6971          # hospitalisations with an alert, score >= 6 (published, 18%)

tp <- n_sepsis - missed                    # sepsis with alert
fp <- alerted - tp                         # alert without sepsis
fn <- missed                               # sepsis without alert
tn <- n_hosp - n_sepsis - fp               # neither
sens <- tp / n_sepsis
spec <- tn / (n_hosp - n_sepsis)
ppv <- tp / alerted
npv <- tn / (tn + fn)
alerts_per_case <- alerted / tp            # alerts per sepsis case caught

# Consistency with the percentages printed in the abstract
stopifnot(tp + fp + fn + tn == n_hosp,
          round(100 * missed / n_sepsis) == 67, round(100 * alerted / n_hosp) == 18,
          round(100 * n_sepsis / n_hosp) == 7, tp == 843)
round(c(sensitivity = sens, specificity = spec, ppv = ppv, npv = npv,
        alerts_per_case = alerts_per_case), 3)
```

```
#>     sensitivity     specificity             ppv             npv alerts_per_case
#>           0.330           0.829           0.121           0.946           8.269
```

```
two_by_two <- data.frame(
  row = c("Alert (score ≥ 6)", "No alert", "Total"),
  sepsis = c(tp, fn, n_sepsis),
  no_sepsis = c(fp, tn, n_hosp - n_sepsis),
  total = c(alerted, fn + tn, n_hosp))
knitr::kable(two_by_two, col.names = c("", "Sepsis", "No sepsis", "Total"),
             format.args = list(big.mark = ","), row.names = FALSE)
```

Table 7.3: Reconstructed 2 × 2 table for the sepsis alert (score ≥ 6) at the hospitalisation level, from the published abstract of Wong et al. (2021). Unit: hospitalisation; n = 38,455 from 27,697 patients; counts derived from published figures under the stated assumption.

	Sepsis	No sepsis	Total
Alert (score ≥ 6)	843	6,128	6,971
No alert	1,709	29,775	31,484
Total	2,552	35,903	38,455

Read the result as a committee member would. Of 1,000 hospitalisations, about 181 receive an alert; about 22 of these involve sepsis and about 159 do not; about 44 hospitalisations with sepsis receive no alert. The unit matters: 38,455 hospitalisations come from 27,697 patients, so some patients contribute more than one hospitalisation, and the hospitalisations are not fully independent observations (Chapter 19).

7.3.2 Part B · A journal-club card for a fictitious abstract

The second exercise uses a *synthetic* abstract written for this book. It describes no real study, product or hospital.

> **Fictitious abstract.** *Machine learning predicts postoperative delirium after hip fracture.* We developed a gradient-boosting model using 46 routinely collected variables from 2,140 admissions of patients aged 65 or older at two hospitals (2019–2023). Delirium was identified from discharge diagnosis codes. Admissions were randomly split 80 : 20 into training and test sets; hyperparameters were optimised on the test set. The model achieved an AUC of 0.89 (95% CI 0.86–0.92) and outperformed a clinical points rule (AUC 0.71). Feature importance identified postoperative opioid dose, age and dementia as the strongest predictors. The model reliably identifies patients at risk and should be implemented to reduce delirium.

The card has five rows — recruitment, reference (here: outcome definition), analysis, generalisation and decision. Each row names the tool item it draws on and turns a missing or doubtful item into a question.

```
card <- data.frame(
  domain = c("Recruitment", "Reference / outcome", "Analysis", "Analysis",
        "Generalisation", "Decision"),
  abstract_says = c("2,140 admissions, two hospitals, 2019-2023",
                    "delirium from discharge diagnosis codes",
                    "random 80:20 split of admissions",
                    "hyperparameters optimised on the test set; AUC only",
                    "test set from the same two hospitals and period",
                    "'should be implemented to reduce delirium'"),
  item = c("PROBAST+AI 1.3 (representative inclusions)",
           "PROBAST+AI 3.1 (outcome defined and assessed appropriately)",
           "PROBAST+AI 4.5 (data leakage avoided); 2.4 (predictors available at prediction
        time)",
           "TRIPOD+AI 12c, 12e; PROBAST+AI 4.7 (calibration, net benefit)",
           "STARD-AI 28 / TRIPOD+AI 23a (target population, subgroups)",
           "DECIDE-AI; CONSORT-AI (impact needs live use and comparison)"),
  rating = c("unclear", "doubtful", "doubtful", "not reported", "not reported", "not
        supported"))
```

```r
card$question <- ifelse(card$rating == "reported", "",
  c("How many eligible admissions were excluded, and why?",
    "How sensitive are discharge codes for delirium compared with a structured bedside
        assessment?",
    "Were repeat admissions of the same patient kept on one side of the split? Was the
        postoperative opioid dose known at the time of prediction?",
    "What was the performance on data not used for tuning, and how well calibrated are the
        predicted risks?",
    "How does the model perform in another hospital, a later period and in key subgroups?",
    "Which study showed that acting on the model reduces delirium?"))
n_questions <- sum(card$question != "")
stopifnot(n_questions == nrow(card))        # every row of this abstract triggers a
        question
knitr::kable(card, col.names = c("Domain", "Abstract says", "Tool item", "Rating",
        "Question to the authors"),
             row.names = FALSE)
```

Table 7.4: Journal-club card for the fictitious delirium abstract. Every row rated other than ‘reported’ produces a question to the authors.

Domain	Abstract says	Tool item	Rating	Question to the authors
Recruitment	2,140 admissions, two hospitals, 2019-2023	PROBAST+AI 1.3 (representative inclusions)	unclear	How many eligible admissions were excluded, and why?
Reference / outcome	delirium from discharge diagnosis codes	PROBAST+AI 3.1 (outcome defined and assessed appropriately)	doubtful	How sensitive are discharge codes for delirium compared with a structured bedside assessment?
Analysis	random 80:20 split of admissions	PROBAST+AI 4.5 (data leakage avoided); 2.4 (predictors available at prediction time)	doubtful	Were repeat admissions of the same patient kept on one side of the split? Was the postoperative opioid dose known at the time of prediction?
Analysis	hyperparameters optimised on the test set; AUC only	TRIPOD+AI 12c, 12e; PROBAST+AI 4.7 (calibration, net benefit)	not reported	What was the performance on data not used for tuning, and how well calibrated are the predicted risks?
Generalisation	test set from the same two hospitals and period	STARD-AI 28 / TRIPOD+AI 23a (target population, subgroups)	not reported	How does the model perform in another hospital, a later period and in key subgroups?
Decision	‘should be implemented to reduce delirium’	DECIDE-AI; CONSORT-AI (impact needs live use and comparison)	not supported	Which study showed that acting on the model reduces delirium?

Two findings stand out. Tuning on the test set removes the independence of the test, so the AUC of 0.89 is an optimistic estimate (Chapter 5). And postoperative opioid dose cannot be known on the first morning after admission: if the model is meant to trigger prevention before surgery, one of its strongest predictors is not available at the time of use (PROBAST+AI, signalling question 2.4). The concluding sentence makes an impact claim from an accuracy study.

7.4 Counter-check

Change the target population. Suppose the committee considers the alert only for an intensive care unit, where sepsis is far more common — say 20% of stays instead of 7% (*assumption*).

Before you compute, predict: With the published sensitivity and specificity, does the positive predictive value rise or fall? And what happens to the number of alerts per sepsis case caught?

```
ppv_at <- function(prev, se = sens, sp = spec) se * prev / (se * prev + (1 - sp) * (1 -
        prev))
prevalences <- c(published = n_sepsis / n_hosp, icu_assumed = 0.20)
cc <- data.frame(setting = names(prevalences), prevalence = prevalences,
                 ppv = ppv_at(prevalences), alerts_per_case = 1 / ppv_at(prevalences))
stopifnot(abs(cc$ppv[1] - ppv) < 1e-9, cc$ppv[2] > cc$ppv[1])
knitr::kable(cc, digits = 3, row.names = FALSE,
             col.names = c("Setting", "Prevalence", "PPV", "Alerts per case caught"),
             caption = "Counter-check: published sensitivity and specificity applied to an
        assumed higher prevalence.")
```

Counter-check: published sensitivity and specificity applied to an assumed higher prevalence.

Setting	Prevalence	PPV	Alerts per case caught
published	0.066	0.121	8.269
icu_assumed	0.200	0.326	3.067

The positive predictive value rises to about 33% and the alerts per case caught fall to about 3.1. The calculation, however, assumes that sensitivity and specificity carry over unchanged to intensive care. That is rarely true: patients there are sicker, measured more often and treated earlier, so the score's behaviour can change in either direction. The counter-check therefore yields a hypothesis for a local evaluation, not a result (Methods in R: Predictive Values and Prevalence).

7.5 Consequences for patients and operations

Patients. Two harms sit on either side of a sepsis alert. A missed case may delay treatment, although clinicians often recognise sepsis without the alert, as the comparison with timely antibiotics in Wong et al. (2021) shows. A false alert can lead to unnecessary tests, antibiotics and interruptions, and many false alerts teach staff to ignore all of them. Which harm dominates depends on how the alert is embedded in the workflow — a question for an early clinical evaluation (DECIDE-AI), not for the accuracy study.

Operations. A brief for the committee should give the expected number of alerts per ward and day, the person who responds to an alert, and the rule for stopping the pilot (Chapter 20). Local data from a silent, non-alerting phase can show discrimination and calibration in the hospital's own population before any alert reaches a clinician.

Regulation. Software intended to support the diagnosis of sepsis in individual patients is typically assessed as medical device software under the Medical Device Regulation (European Parliament and Council of the European Union 2017); the qualification guidance MDCG 2019-11 Rev. 1 is orientation, not binding law (Medical Device Coordination Group 2025). Under Article 6(1) of the AI Act (European Parliament and Council of the European Union 2024), an AI system is high-risk if it is a safety component of a product, or itself a product, covered by the legislation listed in Annex I — which includes the Medical Device Regulation — *and* that product has to undergo a third-party conformity assessment. After the amendment by Regulation (EU) 2026/1744 (European Parliament and Council of the European Union 2026), which entered into force on 27 July 2026, Chapter III, Sections 1 to 3 apply to such systems from 2 August 2028 (Article 113, as amended; read on EUR-Lex in the English text of the amendment; status: November 2026). Whether a particular warning function is a medical device, and who is its manufacturer, depends on its intended purpose and is a question for the hospital's regulatory function (*open*; Chapter 22).

Data protection. A local validation uses patient data for a purpose other than treatment. Its legal basis, minimisation and access rules must be settled before data are extracted (Chapter 23, Chapter 26).

7.6 Look up and transfer

Reviewer question: "For each claim in your abstract, which result in which table supports it, at which level of evidence — accuracy, effect on decisions, or effect on patient outcomes — and in which population?"

Self-test: A paper reports that after an AI fracture-detection tool was introduced in an emergency department, the rate of missed fractures in the following six months was lower than in the six months before. The authors conclude that the tool improves patient outcomes. Which level of evidence does the study reach, and what is the main alternative explanation?

Answer

Annotated sources:

- Fryback and Thornbury (1991) defines the hierarchy of diagnostic efficacy that structures the core reading.
- Nagendran et al. (2020) documents how few prospective and randomised studies exist for deep learning in imaging and how often abstracts claim parity with clinicians.
- Sounderajah et al. (2025) (STARD-AI; an author correction of July 2026 concerns only the consortium list), Collins et al. (2024) (TRIPOD+AI), Moons et al. (2025) (PROBAST+AI) and Vasey et al. (2022) (DECIDE-AI) are the four tools assigned to concrete checks in Table 7.1.

- Liu et al. (2020) is the reporting extension for randomised trials of AI interventions, the top rung of the ladder.
- Wong et al. (2021) is the external validation from which the checked statement and the reconstructed 2 × 2 table come.
- The AI Act (European Parliament and Council of the European Union 2024) and its 2026 amendment (European Parliament and Council of the European Union 2026) define the high-risk timetable cited in the consequences; the Medical Device Regulation (European Parliament and Council of the European Union 2017) and MDCG 2019-11 (Medical Device Coordination Group 2025) the device question.

7.7 Further reading

- Chapter 6 — the measures (discrimination, calibration, net benefit) that the checks in Table 7.1 ask for.
- Chapter 20 — how a silent phase, a pilot and monitoring turn an accuracy result into evidence about use.
- Chapter 27 — the model passport that collects these answers for one's own model.
- Methods in R: TRIPOD-AI — prediction models — a reporting template to compare a paper against.
- Methods in R: CONSORT — randomised trials — the core items that CONSORT-AI extends.
- CTTIR Courses, Course 4 · TRIPOD-AI, fairness auditing, reproducibility at scale — a lab that applies TRIPOD-AI to a model.
- CTTIR Tutorials: Diagnostic Test Accuracy — the accuracy measures behind Part A.
- Scientometrics in R: Retraction Analysis — checking corrections and retractions before a paper is cited.