

# Four questions for any medical AI tool

A one-page checklist for medical and pharmacy students · Ali Ganjizadeh, MD, MBA

An AI tool performs like a diagnostic test and lives like a drug, so judge it like both. **None of these questions is about the model.** All of them draw on clinical training.

## 1 DATA PROVENANCE

### Where did the data come from, and do those patients look like mine?

Where did the data and labels come from? Do those patients, sites and devices look like mine?

Like a drug's trial population

How were the "right answers" decided, and was it tested at other hospitals and in subgroups?

Like a gold standard, then replication

**From the talk:** purple surgical ink on benign moles cut a melanoma AI's specificity from 84% to 46% (Winkler, *JAMA Dermatol* 2019).

## 2 LABEL DEFINITION

### What exactly does it predict, and who acts on it, when?

Does it predict what matters, or a stand-in such as cost or a billing code?

Like a surrogate endpoint

Who acts on the output, for which patient, at what moment? Does it arrive before we'd know anyway?

Like a drug's indication

**From the talk:** an algorithm that predicted healthcare cost instead of need gave Black patients the same risk score as white patients who were less sick (Obermeyer, *Science* 2019).

## 3 EVALUATION

### What evidence is there beyond accuracy, at my prevalence?

What is the positive predictive value (PPV) at *my* prevalence, and how many false alarms come with each true case?

Like a screening test's PPV

How far up the ladder has it climbed: accuracy on past data, silent mode, a supervised pilot, a randomised trial of patient outcomes?

Like trial phases, preclinical to Phase III

**From the talk:** at 1% prevalence, a tool that is 95% sensitive and 95% specific has a PPV of about 16%. Five of every six alerts are false.

## 4 AFTER GO-LIVE

### Who watches it once it's in use?

Is there a label stating its intended use, population, and when *not* to use it?

Like a package insert

Who monitors it after go-live, and what happens when the scanner, workflow or patient mix changes?

Like pharmacovigilance

**From the talk:** Duke's "Model Facts" label for its sepsis model states who it is for and warns, for example, that it is not for ICU patients (Sendak, *npj Digit Med* 2020).

**Red flags:** no clear answer on where the training data came from · only accuracy or AUC reported, with no PPV at your prevalence · never tested at a hospital like yours · nobody named to monitor it after go-live.

## FURTHER READING

- Liu Y et al. How to read articles that use machine learning: Users' Guides to the Medical Literature. *JAMA* 2019;322:1806–16. doi.org/10.1001/jama.2019.16489
- Kelly CJ et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17:195 (open access). doi.org/10.1186/s12916-019-1426-2
- Obermeyer Z et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366:447–53. doi.org/10.1126/science.aax2342
- Finlayson SG et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med* 2021;385:283–6. doi.org/10.1056/NEJMc2104626
- Sendak MP et al. Presenting machine learning model information to clinical end users with model facts labels. *npj Digit Med* 2020;3:41 (open access). doi.org/10.1038/s41746-020-0253-3
- US FDA. Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products (draft guidance, January 2025). fda.gov (search the title)