

# Clinical AI evidence review checklist

Use this worksheet to identify what has actually been proved for the exact system, version and workflow under review. Evidence at a higher rung does not repair a weakness below it.

## Review record

|                   |  |                    |  |
|-------------------|--|--------------------|--|
| System or product |  | Supplier           |  |
| Intended task     |  | Primary user       |  |
| Evaluated version |  | Local version      |  |
| Reviewer and date |  | Proposed authority |  |

## Six-level evidence ladder

| Rung | Evidence level                                                                           | Question                                                    | Proof located / gap | Status                                                                                               |
|------|------------------------------------------------------------------------------------------|-------------------------------------------------------------|---------------------|------------------------------------------------------------------------------------------------------|
| 1    | Technical validity<br>Version-bound test report,<br>representative set, regression suite | Can this exact configuration perform its defined task?      |                     | <input type="checkbox"/> Met<br><input type="checkbox"/> Partial<br><input type="checkbox"/> Missing |
| 2    | Task and safety validity<br>Failure-mode tests, limits, escalation and recovery evidence | Is it safe enough for this bounded task, including failure? |                     | <input type="checkbox"/> Met<br><input type="checkbox"/> Partial<br><input type="checkbox"/> Missing |
| 3    | External validation<br>Independent validation, subgroup results, exclusions              | Does performance travel beyond development data?            |                     | <input type="checkbox"/> Met<br><input type="checkbox"/> Partial<br><input type="checkbox"/> Missing |
| 4    | Local prospective validation<br>Silent-mode results, burden, overrides and stop criteria | Does it work safely and usefully here?                      |                     | <input type="checkbox"/> Met<br><input type="checkbox"/> Partial<br><input type="checkbox"/> Missing |
| 5    | Outcome evidence<br>Comparative evaluation and transparent benefit-harm analysis         | Does use improve something decision-relevant?               |                     | <input type="checkbox"/> Met<br><input type="checkbox"/> Partial<br><input type="checkbox"/> Missing |
| 6    | Monitored deployment<br>Version traceability, drift, incidents and tested rollback       | Can we detect when the evidence no longer holds?            |                     | <input type="checkbox"/> Met<br><input type="checkbox"/> Partial<br><input type="checkbox"/> Missing |

# Twelve buyer questions

Record where the answer is evidenced, not merely asserted. If a material detail is unknown, treat it as a gap in the current decision.

| #  | Question                                                                 | Evidence location / notes | Status                                                         |
|----|--------------------------------------------------------------------------|---------------------------|----------------------------------------------------------------|
| 1  | What exact task and user is the system evaluated for?                    |                           | <input type="checkbox"/> Clear<br><input type="checkbox"/> Gap |
| 2  | Which deployed configuration does the evidence cover?                    |                           | <input type="checkbox"/> Clear<br><input type="checkbox"/> Gap |
| 3  | What is the benchmark population, and how does it differ from ours?      |                           | <input type="checkbox"/> Clear<br><input type="checkbox"/> Gap |
| 4  | How does the system behave when information is missing or contradictory? |                           | <input type="checkbox"/> Clear<br><input type="checkbox"/> Gap |
| 5  | When must it abstain or escalate to a human?                             |                           | <input type="checkbox"/> Clear<br><input type="checkbox"/> Gap |
| 6  | What independent and subgroup validation has been completed?             |                           | <input type="checkbox"/> Clear<br><input type="checkbox"/> Gap |
| 7  | What local prospective evaluation is expected before authority expands?  |                           | <input type="checkbox"/> Clear<br><input type="checkbox"/> Gap |
| 8  | Which workflow, patient or operational outcomes will be measured?        |                           | <input type="checkbox"/> Clear<br><input type="checkbox"/> Gap |
| 9  | How are false alarms, overrides and displaced work assessed?             |                           | <input type="checkbox"/> Clear<br><input type="checkbox"/> Gap |
| 10 | Which changes trigger regression testing or revalidation?                |                           | <input type="checkbox"/> Clear<br><input type="checkbox"/> Gap |
| 11 | How are incidents, drift and subgroup performance monitored?             |                           | <input type="checkbox"/> Clear<br><input type="checkbox"/> Gap |
| 12 | Who can pause, roll back or reduce the system's authority?               |                           | <input type="checkbox"/> Clear<br><input type="checkbox"/> Gap |

## Decision and conditions

|                             |                                                                                                                                                                                         |
|-----------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Decision                    | <input type="checkbox"/> Do not proceed <input type="checkbox"/> Evidence gap to close <input type="checkbox"/> Bounded local evaluation <input type="checkbox"/> Proceed with controls |
| Authority ceiling           |                                                                                                                                                                                         |
| Required conditions         |                                                                                                                                                                                         |
| Revalidation triggers       |                                                                                                                                                                                         |
| Named owner and review date |                                                                                                                                                                                         |

Use with the full guide: Clinical AI's Proof Problem: The Evidence Ladder from Benchmark to Bedside. This checklist is educational and does not replace clinical, legal, regulatory or information-security review. Evidence expectations should be adapted to the use case, jurisdiction, risk and authority of the deployed system.