



PART 3

Read the text below and answer questions 27–40.

ARTIFICIAL INTELLIGENCE IN MEDICAL DIAGNOSIS

A

The arrival of artificial intelligence in diagnosis is often presented as a simple technological upgrade: faster pattern recognition applied to scans, slides, and photographs. Yet the deeper change is epistemological. Traditional diagnosis rests on a chain of interpretive acts—history-taking, examination, and the clinician’s judgement about what counts as relevant evidence. Diagnostic AI, by contrast, converts clinical materials into a computational problem: learn statistical regularities in data and output a probability. This shift raises an awkward question about “ground truth”. In medicine, the gold standard is rarely pure fact. A biopsy may confirm malignancy, but many conditions lack a single definitive test, and even “confirmed” labels are mediated by sampling, reporting conventions, and human interpretation. AI therefore does not merely automate diagnosis; it forces a reconsideration of what diagnosis is, and whether correlation-rich models can be trusted when the underlying truth they learn from is itself partially constructed.

B

The “ground truth” problem becomes evident when models learn from labelled datasets. Labels frequently derive from clinician judgement, radiology reports, or pathology reads, and these are subject to inter-observer variability. Two experts can view the same image and disagree, especially when findings are subtle or borderline. Records can also contain errors produced by time pressure, incomplete follow-up, or the tendency to document uncertainty in free text rather than structured fields. When such imperfect labels are treated as if they were objective, models may internalise local habits—how one hospital defines pneumonia on X-ray, how aggressively another classifies suspicious lesions—rather than learning a universal clinical reality. The result is epistemic opacity: high headline accuracy can coexist with hidden fragility, because the model is optimised to replicate an imperfect consensus rather than to discover an independent truth.

C

Even if labels were flawless, diagnostic AI would still confront generalisation. A model that performs impressively on a development dataset can deteriorate when moved to a new hospital, a new scanner, or a new population. This is not simply “bad luck”; it is domain shift. Differences in disease prevalence, demographic composition, imaging protocols, and even how clinicians order tests can change the statistical structure of the data. Models may also overfit—capturing quirks of a particular dataset rather than stable clinical



signals. A tool trained largely on high-income populations, for example, may underperform in settings with different comorbidities or presentation patterns. These failures matter because medicine is often decided at the margins: rare diseases, atypical cases, and under-represented groups are precisely where clinicians need help most, and precisely where data are thinnest.

D

In response, developers often invoke interpretability—heatmaps, saliency maps, and feature attributions that appear to show “why” a model reached its conclusion. Yet interpretability can become a performative reassurance. Saliency maps may highlight regions correlated with the output, but correlation is not causal reasoning, and visual explanations can be manipulated or can remain stable even when a model’s decision is driven by confounders. Moreover, the trade-off between accuracy and explainability is not easily resolved: some highly performant models are complex ensembles whose internal logic resists human-friendly narratives. Hospitals and regulators therefore increasingly emphasise transparency of development, external validation, and monitoring over simplistic “explanations” that risk giving clinicians a false sense of understanding. The dilemma is practical as well as philosophical: clinicians need reasons they can act on, but the model may offer only probabilities wrapped in persuasive imagery.

E

The impact of AI in diagnosis is ultimately mediated by human behaviour. When algorithms are embedded into workflow as triage, second reads, or prioritisation tools, they shape attention. This can improve speed, but it can also create automation bias, the tendency to over-trust machine outputs even when they conflict with other evidence. The opposite failure also occurs: clinicians may ignore AI entirely if it generates frequent false positives, disrupting routine and producing alarm fatigue of a different kind. A longer-term concern is deskilling. If clinicians repeatedly defer to algorithmic judgement, the skills needed to detect subtle patterns may erode, leaving practitioners less able to challenge the system when it is wrong. Effective deployment therefore requires training, explicit protocols for disagreement, and a culture in which AI is treated as an assistant whose outputs must be interpreted in context, rather than as an oracle.

F

Ethics and liability sharpen these tensions. Many deployments rely on the “human in the loop” principle: a clinician remains responsible and can override the model. In practice, this can be a legal fiction. If institutional policy encourages reliance on an approved tool, if time pressure makes independent review difficult, or if the tool is marketed as outperforming humans, then meaningful override may be socially and operationally



constrained. When an AI-assisted diagnosis is wrong, responsibility can fall into a liability vacuum: vendors may claim the clinician made the final decision; clinicians may argue they were following standard-of-care tools; hospitals may point to regulatory clearance. Consent adds another layer. Some argue patients should be told when algorithmic systems contribute to decisions; others worry that disclosure becomes symbolic without genuine choice, particularly when alternatives are limited.

G

The future of diagnostic AI depends on whether it reduces inequity or codifies historical bias. Optimists argue that in under-resourced regions, AI could deliver specialist-level support where clinicians are scarce, shortening queues and preventing missed diagnoses. Pessimists note that models trained on skewed datasets can misdiagnose populations that differ from the training context, thereby reinforcing disparities. Algorithmic fairness therefore requires diversity in data, evaluation across settings, and continuous surveillance for performance drift as scanners, workflows, and disease patterns change. In this sense, AI tools should be treated like medical devices with a lifecycle: validated externally, audited after deployment, and recalibrated when conditions shift. The key question is not whether an algorithm can match expert performance in a controlled trial, but whether institutions can build governance—technical, clinical, and ethical—that keeps the system safe, useful, and accountable over time.

Questions 27–31

Select the correct answer.

27 What is the fundamental problem with “ground truth” labels described in the passage?

- A** They are unnecessary because algorithms can diagnose without any training data
- B** They often reflect imperfect human judgement and inconsistent records rather than objective fact
- C** They can be corrected fully by using larger datasets from a single hospital
- D** They are reliable only when created by junior clinicians rather than specialists

28 Why do diagnostic models often fail when transferred to a new hospital?

- A** Because external validation is prohibited by most regulators
- B** Because models cannot be used on digital images outside the training site
- C** Because clinicians refuse to adopt tools that have been tested elsewhere
- D** Because domain shift can alter patient populations, equipment, and prevalence patterns

29 What does the passage suggest about visual explanation tools like heatmaps?



- A** They demonstrate that a model has identified the true cause of disease
- B** They guarantee that clinicians will calibrate trust appropriately
- C** They may give a persuasive impression of understanding without revealing real reasoning
- D** They remove the need for external validation and monitoring

30 What legal complication arises from the “human in the loop” concept?

- A** It can create a liability vacuum if real-world override is constrained and responsibility is disputed
- B** It ensures vendors are always legally responsible for incorrect diagnoses
- C** It makes consent unnecessary because clinicians remain in control
- D** It prevents any ethical debate by assigning blame clearly to regulators

31 According to the passage, how can AI exacerbate health inequalities?

- A** By improving access in low-resource settings too quickly
- B** By misdiagnosing groups under-represented in training data and reinforcing historical bias
- C** By preventing hospitals from collecting any new clinical data
- D** By requiring all clinicians to specialise in computer science

Questions 32–36

Complete each sentence with the correct ending, **A–H**, below.

- 32** The reliability of “ground truth” data is compromised
- 33** Diagnostic performance often degrades in new settings
- 34** Visual explanations such as saliency maps may be deceptive
- 35** Automation bias can cause clinicians to rely too heavily on AI outputs
- 36** Determining legal liability remains legally ambiguous

List of Endings

- A** because statistical correlations do not equate to causal reasoning.
- B** because human experts can show inter-observer variability.
- C** because the “human in the loop” framework can obscure who is truly accountable.
- D** because domain shift can involve differences in scanners, populations, and disease prevalence.
- E** because clinicians may trust algorithmic outputs over their own judgement.



F because external validation always eliminates all diagnostic errors.

G because heatmaps are legally required in every jurisdiction.

H because clinical data never changes once a model is deployed.

Questions 37–40

Choose **YES** if the statement agrees with the claims of the writer, choose **NO** if the statement contradicts the claims of the writer, or choose **NOT GIVEN** if it is impossible to say what the writer thinks about this.

37 The writer claims that medical diagnosis always rests on a single definitive test.

38 The passage suggests that full explainability of “black box” models may be unachievable.

39 The passage states that most patients prefer AI diagnosis over human assessment.

40 The writer claims that matching expert performance in a controlled trial is sufficient to ensure diagnostic AI remains safe over time.

Answer Key

27. B

28. D

29. C

30. A

31. B

32. B

33. D

34. A

35. E

36. C

37. NO

38. YES

39. NOT GIVEN



40. NO