RESTRICTED ACCESS

CASE-09 · HEALTHCARE & NEW TECHNOLOGIES

# Investing €25 million in diagnostic AI across the whole hospital network

Adopting a more accurate system (in the vendor's tests) to improve efficiency and quality of care.

THE JUDGMENT

whole network → validated pilot

CALIBRATED CONFIDENCE

47%

66

Δ-CSI

SECTOR

Healthcare

EXECUTIVE VERDICT

# On your patients accuracy often drops — and blind trust breeds new errors.

THE CALL

whole network → validated pilot

CONFIDENCE

47%

IN ONE LINE

Avoided **€25M** of unused technology and a clinical-legal risk from over-trusting the machine.

## § THE PROTAGONIST & THE CONTEXT

The chief medical officer of the hospital network, with a digital innovation mandate and a transformation budget approved by the board for the current three-year plan.

A network of eight medium-sized public hospitals serving a heterogeneous population: high proportion of elderly patients, high prevalence of chronic conditions, significant ethnic variability across sites. The current diagnostic imaging systems are fragmented across three different platforms. The vendor of the new AI system validated its solution on an international multicentre study of 120,000 scans.

## § BACKGROUND

Diagnostic AI systems have reached and in some benchmarks surpassed the average accuracy of human radiologists. Yet the real-world implementation literature tells a more complex story: independent studies have documented performance drops of 10–25% when a model trained on international data is applied to local populations with different demographic profiles. The phenomenon of automation bias — the tendency of clinicians to accept machine judgments without critical verification — is well documented and intensifies when the system displays a high numerical confidence score. Without explicit human-review protocols and clear escalation criteria, the risk is not that the system fails to work: it is that it works well enough to appear trustworthy, and then fails systematically on a specific patient subgroup.

## § COUNTER-INTUITIVE SCENARIO

On real patients accuracy often **drops**: your population differs from the test one (age, comorbidities, origins). And there's an opposite risk — doctors who **over-trust** the machine and stop double-checking, introducing new errors. The value isn't the software itself, but how you integrate it into daily work and how you govern it. Without that, it's €25 million of technology left on the shelf — and risky, at that.

### IMPLICIT ASSUMPTIONS

- The test accuracy will carry over to real patients.
- Doctors will adopt it and trust the results.
- The return comes from fewer errors and shorter times.

### FALSIFICATION TESTS

- Test it first on YOUR patients, not on the vendor's tests.
- In a pilot: how often do doctors correct it? How much do they over-trust it?
- Does it work the same way across all patient groups?

### QUESTIONS THAT RAISE THE BAR

- What is the accuracy on your real patients, not on the published ones?
- Who is legally accountable when the machine is wrong?

### CALIBRATED CONFIDENCE & PROVENANCE

that the test accuracy holds across the whole network

Provenance: vendor validation study · local pilot data · literature on automation over-trust · red-team base.

**Declared accuracy vs local pilot**

ILLUSTRATIVE DATA


Diagnostic accuracy of the AI system in the vendor's validation study vs in the pilot run on a single network site with the local population (illustrative data).

**Uncritical acceptance rate in pilot**

ILLUSTRATIVE DATA

**64%**

OF AI REPORTS ACCEPTED WITHOUT CRITICAL REVIEW IN THE PILOT

Share of AI reports accepted by physicians without modification or critical review in the local pilot — a signal of automation bias already at the experimental stage (illustrative data).

OUTCOME

Rollout made conditional on a **local field validation** and on clear governance and medical-accountability rules.

VALUE

Avoided **€25M** of unused technology and a clinical-legal risk from over-trusting the machine.

METHODOLOGICAL NOTE — READ FIRST

Composite cases, in the method of the Harvard Business Review: reconstructions based on real, recurring situations in each sector, merged and anonymized to protect confidentiality. The decision dynamics are authentic; names, figures and details are altered and not traceable to any single client or case. The «provenance» notes describe the type of evidence the engine cites with traceability in production. The Δ-CSI values illustrate the intensity of the pressure the contradiction put on the assumptions.