Ecaveo

AI ADOPTION
WHITEPAPER SERIES

PAPER ONE · POPULATION HEALTH RESEARCH

Responsible AI Adoption in Population-Scale Health Research

Why evaluation capability, and not model capability, decides how far adoption can go

Paul Forrest
Founder, Ecaveo
July 2026

About this paper

This is the first of a series of papers on the responsible deployment of artificial intelligence in regulated and evidence-led organisations. Each paper stands alone. Together they set out a single argument, which is that the durable advantage from artificial intelligence comes from better decisions, and that better decisions require systems designed around the person who is accountable for them.

Population health research opens the series because it is the setting where the consequences of a bad deployment are least reversible. Its outputs enter the scientific record, its inputs are special category data, and its licence to operate rests on consent that participants can withdraw.

Editorial balance. Roughly eight parts in ten of this document are evidence and sector analysis, and two parts in ten are original framework. Readers who want the argument without the operating model can stop at Chapter 11.

HOW TO READ THE EVIDENCE TAGS

Every substantive claim in this paper carries a tag stating what kind of evidence supports it. The tags are used consistently across the series.

PEER-REVIEWED A primary study in a peer-reviewed journal.

REVIEW A systematic review, scoping review or meta-analysis.

PREPRINT Posted but not yet through peer review, and flagged wherever it is used.

LEGISLATION An Act, a statutory instrument or the text of a regulation.

REGULATOR Guidance, a consultation or a published position from a regulator or national body.

PRACTICE Practitioner opinion, including the author's own.

ECAVEO The author's own framework, proposed and not yet proven.

A claim resting on two kinds of evidence carries both tags. Where the evidence is thin the paper says so, and in this field it is thin more often than the technology market implies.

Legal and regulatory notice. This paper describes data protection legislation, medical device regulation and clinical safety standards, in particular in Chapters 9, 10 and 11. Nothing in this document constitutes legal advice. The provision of legal advice sits outside the terms of any engagement with the author or with Ecaveo, and any question of a particular organisation's compliance is a matter for its own qualified advisers.

Statements about what a source says are attributed to that source, and where the author draws a practical implication the source does not itself state, the text says so. The position described is the one the author was able to verify in July 2026.

ILLUSTRATIVE MATERIAL

Every organisation, programme and scenario in this paper is invented, and none describes a real client or engagement. Cases are marked **illustrative scenario** wherever they appear. Published research is cited and numbered, and invented material is never presented as evidence.

CITATION

Forrest, P. (2026). *Responsible AI Adoption in Population-Scale Health Research: Why Evaluation Capability, and Not Model Capability, Decides How Far Adoption Can Go.* Ecaveo AI Adoption Whitepaper Series, Paper One. Version 1.0, July 2026.

Comments and corrections are welcome and will be acknowledged in later versions.

ALSO IN THIS SERIES

In preparation. *The Governed Agent.* Bounded autonomy in regulated operations.

In preparation. *Renting the Model, Owning the Learning.* Where advantage actually accumulates.

In preparation. *The Assurance Record.* What a board should be able to ask for, and receive.

Contents

FRONT MATTER

	A note from the author	4
	Executive summary	6

PART ONE · WHAT THE EVIDENCE SUPPORTS

1	The evaluation gap	8
2	Why this sector is different	11
3	Knowledge work and software engineering	14
4	Clinical and research tasks	17
5	Evidence synthesis, where the evidence is worst	19
6	Predictive genomics and the utility boundary	21
7	Where the positive evidence is strongest	23

PART TWO · OVERSIGHT AND THE PERIMETER

8	The oversight problem	26
9	Data protection after the 2025 Act	29
10	Medical devices and the clinical boundary	31
11	Research governance, security and standards	33

PART THREE · BUILDING THE CAPABILITY

12	An operating model	36
13	Evaluation and assurance	39
14	A portfolio built to the evidence	42
15	Delivery over six to nine months	45
16	The intelligent adoption programme	47

END MATTER

	Board and executive checklist	49
	Glossary	50
	Method and evidence grading	51
	About the author	53
	References	54

A NOTE FROM THE AUTHOR

Four thousand studies and nineteen trials

I came to this paper expecting to write about model capability, because that is where the sector's attention sits and where the procurement conversations start. The evidence sent me somewhere else.

The largest recent survey of large language models in clinical medicine screened nearly 13,000 records and included 4,609 studies. Nineteen of them were prospective randomised trials. A quarter of the included studies had fewer than 30 participants. And the advantage these systems show over human performance shrinks as the evaluation gets closer to real clinical work, which is the opposite of the pattern a buyer is invited to assume.[4] REVIEW

Set beside that, the developer productivity literature contains a 26 per cent gain in one study and a 19 per cent slowdown in another, both randomised, both competently run.[1,2] PEER-REVIEWED The spread is not a puzzle to be resolved by averaging. It is the finding.

What changed my mind about the useful contribution of software was not any of that, though. It was the oversight evidence. Human review is the control that nearly every AI governance framework leans on hardest, and it has the weakest empirical support of anything in common use. Reviewers given erroneous model output lost 14 percentage points of diagnostic accuracy despite 20 hours of prior AI literacy training.[10] PEER-REVIEWED Endoscopists exposed to AI assistance detected fewer adenomas in their unassisted procedures afterwards than the same centres had recorded before.[11] PEER-REVIEWED

A control that names a person and omits a procedure is an assertion about intent. This paper is largely about the difference between the two, and about the unglamorous capability that has to exist before either can be tested.

TWO PAGES THAT CIRCULATE ON THEIR OWN

Executive summary

The constraint on responsible AI adoption in population-scale health research is not the capability of the models. It is the capability of the institution to establish what a workflow costs today, to compare it against a version assisted by a model, to detect a quality regression before it reaches a publication or a participant, and to repeat that exercise when the model changes underneath it.

This paper argues that the portfolio should follow that capability. Where it leads instead, the organisation is buying something it cannot yet assess. An organisation that can run an instrumented baseline may pilot a reversible internal workflow, and one that can run a pre-registered controlled comparison may consider a use case whose output leaves the building. It argues equally firmly against several deployments that are currently easy to buy.

1

Published effect sizes do not transfer. Three pooled randomised field experiments across 4,867 developers report a 26.08 per cent increase in completed tasks. A randomised study of 16 experienced developers in repositories they knew well reports a 19 per cent increase in completion time.

2

Measured advantage falls as realism rises. Across 1,046 head-to-head comparisons, models outperformed humans in 38.4 per cent of exam-style studies and 25.9 per cent of studies using real clinical data.

3

The evidence base is thinner than the market. Of 4,609 included studies of large language models in clinical medicine, 19 were prospective randomised trials and at least a quarter had samples below 30.

4

Literature searching is the worst-evidenced candidate. In comparative studies, models missed between 68 and 96 per cent of relevant studies, median 91 per cent, and the failure leaves no trace in the output the researcher receives.

5

Human review fails in three distinct ways. Through indifference, where a model neither helped nor was helped by clinicians. Through credulity, where AI literacy training did not protect against erroneous output. And through atrophy, where unassisted detection rates fell after AI was introduced.

6

The regulatory question has moved. Three regulators now ask what the organisation claims a product does, who holds clinical risk when it is deployed, and what recourse a person has. None is answered by a model evaluation.

What follows for a board, a funder or a programme

- **Build the measurement first.** Before selecting a use case, establish what one named workflow costs today in time, rework and escalation, using data. Recollection flatters the workflow by about the margin the pilot is meant to measure. An organisation that cannot do that for work it has already changed should not commit to work whose output leaves the building.
- **Tier by recall cost.** The expense and incompleteness of withdrawing an output matters more than the classification of the input. A participant newsletter drafted from no personal data carries more exposure than an analysis inside an audited environment.

- **Specify the review, or stop calling it a control.** Name what the reviewer sees, what they compare it against, how long they have, what authority they hold to reject, and how anyone would know afterwards whether the review was performed with care.
- **Refuse four things.** No literature search presented as complete. No model treated as an authorisation layer. No participant-facing health interpretation outside clinical and regulatory governance. And no agent with production write access before bounded autonomy has been tested and can be stopped.

Chapter 16 proposes an AI adoption design workshop for a research programme, a health data infrastructure or a funder. It begins with what to measure and why, which is unglamorous, cheap, and the part most programmes skip.

PART ONE

What the evidence supports

Seven chapters on how good the evidence for artificial intelligence in health research actually is, and on why the sector keeps buying ahead of it.

CHAPTER ONE

The evaluation gap

Every large research programme reaches the same three questions in the same wrong order, and all three presuppose an answer to a fourth that nobody asks.

Someone asks whether generative AI should be allowed, someone else asks which tool to buy, and a third person asks what the policy should say. The order is wrong. All three questions presuppose an answer to a fourth that nobody asks, and that fourth question decides whether any of the others can be answered honestly. How would this organisation know whether a deployment worked?

The contention of this paper is that the binding constraint on responsible AI adoption in population-scale health research is institutional evaluation capability rather than model capability. Candidate use cases and vendor offers are abundant in this sector. What is scarce is the ability to establish what a workflow costs today, to compare it against a version assisted by a model, to detect a quality regression before it reaches a publication or a participant, and to repeat that exercise when the model changes underneath them. Where that capability is absent, governance documents describe controls that nobody can test, pilots produce adoption statistics that nobody can interpret, and the organisation acquires exposure it cannot measure in exchange for benefits it cannot demonstrate.

THE EVALUATION CAPABILITY LADDER

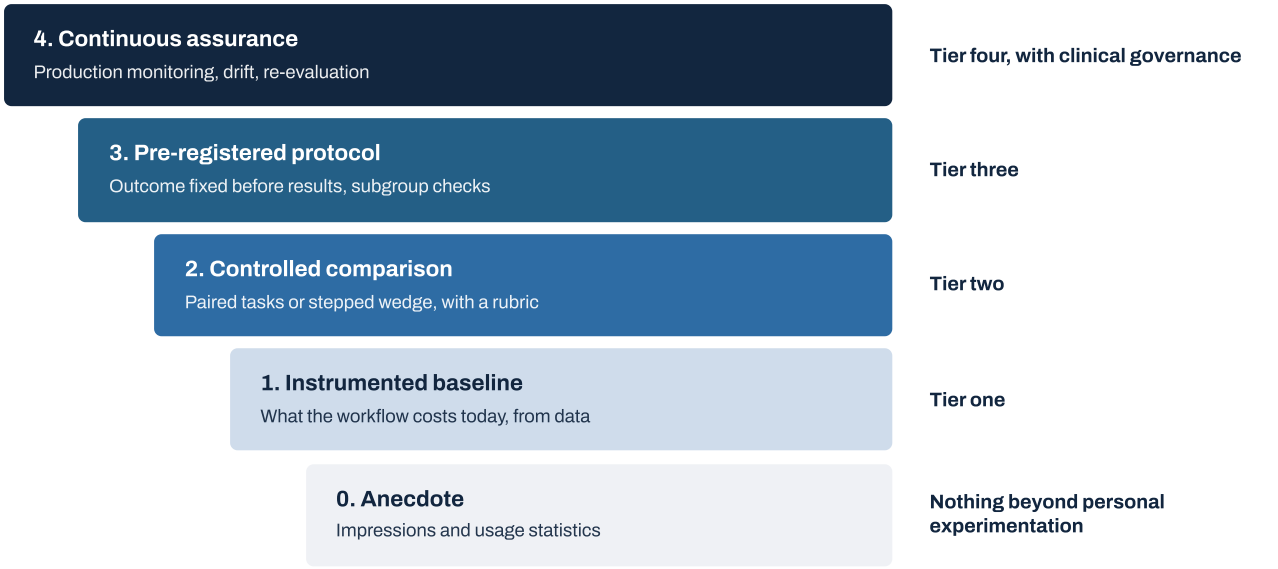


FIGURE 1 Five levels of evidence an organisation can produce about work it has already changed, and the highest risk tier each level supports. Most organisations approaching their first pilots operate at level one and believe they operate at level three.

Ecaveo framework. ECAVEO

1.1 Three lines of evidence

The first concerns transferability. Published effect sizes for generative AI in skilled work range from a pooled 26.08 per cent increase in completed tasks across three randomised field experiments to a 19 per cent increase in completion time in a randomised study of experienced developers working in repositories they knew well, a result its authors revisited a year later with returning and new participants.[1,2,3] PEER-REVIEWED That spread is the finding. Averaging it away produces a number that describes no real workplace.

The second concerns realism. A systematic review published in 2026, covering 4,609 included studies of large language models in clinical medicine, reports that the measured advantage of these systems over human performance falls as evaluation conditions become more realistic. Models outperformed humans in 38.4 per cent of head-to-head comparisons in exam-style studies and in 25.9 per cent of comparisons using real clinical data. Nineteen of the 4,609 were prospective randomised trials.[4] REVIEW

The third concerns oversight. Human review is the control that almost every AI governance framework leans on hardest, and it has the weakest empirical support of any control in common use. In a randomised trial, physicians using a large language model scored two percentage points above physicians using conventional resources, a difference that did not reach significance, while the model working alone scored an adjusted 16

points above the unaided physicians.[9] PEER-REVIEWED Reviewers given deliberately erroneous model output lost 14 percentage points of diagnostic accuracy despite 20 hours of prior AI literacy training.[10] PEER-REVIEWED Endoscopists at four centres detected fewer adenomas in their unassisted procedures after AI was introduced than the same centres had recorded before it arrived.[11] PEER-REVIEWED

THE ARGUMENT IN ONE PARAGRAPH

Health research has evidence abundance and evaluation scarcity. The useful contribution of artificial intelligence governance is therefore measurement before deployment. Establish a baseline, fix the outcome before the results are seen, tier by how expensive an error is to withdraw, and specify the human review that everything else rests on. Everything in this paper is either evidence about how good the systems are, or an argument about what a programme should refuse to do with them.

1.2 What follows from it

Build the evaluation capability first, at a level the organisation can honestly sustain, then let the portfolio follow it. An organisation that can run an instrumented baseline may pilot a reversible internal workflow. An organisation that can run a pre-registered controlled comparison with subgroup analysis may consider a use case whose output leaves the building. No organisation should deploy above the level of evidence it is equipped to produce, and the cheapest way to discover that level is to try to measure something small before committing to something large. ECAVEO

This is a paper about a sector. Population cohorts, health data infrastructures, genomics initiatives and research platforms differ in scale and remit, and they share the four features that make AI adoption in this setting distinctive. They process special category data at scale. They depend on public consent that can be withdrawn. Their outputs enter the scientific record, where errors propagate and corrections travel slowly. And they operate national digital services alongside their research mission, which means an AI decision taken inside one of them is operational and scientific at the same moment, and regulatory and reputational a little later.

CHAPTER TWO

Why this sector is different

A research programme recovers from a bad quarter more easily than from a breach of trust, and that asymmetry should shape every adoption decision it makes.

A population health research programme sits awkwardly between two organisational types, and that awkwardness causes most of its AI governance difficulty. Measured by its technology estate it resembles a national digital service, running clinics, identity systems, data pipelines, security operations and a researcher-facing product at considerable scale. Measured by its licence to operate it is a research institute, dependent on consent, ethical approval, scientific quality and a continuing public willingness to take part. Those two identities fail differently. Software companies recover from a bad quarter, whereas a programme built on voluntary participation recovers from a breach of trust much more slowly, if at all.

Three features of the setting deserve separate treatment, because each defeats a control that works adequately elsewhere.

Errors propagate through the scientific record. A wrong answer in an internal memorandum embarrasses its author and is corrected in the next meeting. A wrong cohort definition, an omitted confounder or a fabricated citation enters an analysis, then a manuscript, then the literature, where it acquires citations of its own. Retraction is slow and partial, and it rarely reaches everyone who read the original. The relevant property of an AI-assisted research error is therefore its persistence.

Consent is specific and revocable. Participants agree to a described use of their information, and the description was written before generative AI became an operational question in most programmes. The Data (Use and Access) Act 2025 has widened what broad consent to an area of scientific research can cover, which Chapter 9 sets out. Widening a lawful basis does not widen a participant's expectations.[15] LEGISLATION My own view is that a programme whose participants would be surprised by a deployment has a trust problem whether or not it has a compliance problem, and that the two should be assessed separately. PRACTICE

The same tool changes character with its intended purpose. A retrieval assistant over internal policy is an efficiency measure. The same architecture pointed at participant health information and asked to summarise it for a participant may acquire a medical intended purpose, at which point an entirely different regulatory regime applies. The Medicines and Healthcare products Regulatory Agency provides that a product's intended purpose is not specifically determined by the inclusion of any particular technology, and that use in a medical environment or context alone does not qualify a product as a medical

device.[21] REGULATOR The corollary, which matters more in practice, is that the claim an organisation makes about a tool can create obligations the tool's design never anticipated.

2.1 Recall cost, and why sensitivity is the wrong primary variable

Most AI risk frameworks tier use cases by data sensitivity. That ordering is intuitive and it is, in my view, the wrong way round for this sector. ECAVEO

Consider two deployments. The first summarises pseudonymised clinical measurements inside a trusted research environment for an analyst who will check every figure against the source before using it. The second drafts an item of general health information for a participant newsletter using no personal data whatsoever. A sensitivity-led framework ranks the first as the higher risk, because it touches health data. A recall-led framework ranks the second higher, because a published error reaches tens of thousands of people who trusted the sender, and no subsequent correction reaches all of them.

THE SAME ERROR, TWICE

An error inside an audited environment can be found and corrected. The same error in a participant newsletter reaches people who will never see the correction.

RECALL COST AND DATA SENSITIVITY ARE INDEPENDENT

FIGURE 2 A framework that tiers by sensitivity alone ranks the upper-left quadrant as low risk, which is where communications work sits. The two axes are related and they are not the same.

Ecaveo framework. ECAVEO

Recall cost is the expense and the incompleteness of withdrawing an output after it has left the control of the person who produced it. It combines four things, namely how far the output travels, how many people act on it, how long it remains in use, and how completely a correction can reach the people it misled. Sensitivity of the input remains important and belongs in the assessment as a modifier, since the

two variables are related but not the same. An output derived from special category data that never leaves an audited environment is recoverable. A public statement derived from nothing personal at all may not be.

The practical consequence is a tiering scheme whose spine is the reversibility of the output, which Chapter 12 sets out. The classification of the input does less work here than it appears to. My own view, and it is only a view, is that this reordering explains a pattern I have seen repeatedly, where organisations apply heavy scrutiny to well-bounded internal analytical uses and comparatively light scrutiny to communications work that carries far greater exposure. PRACTICE

CHAPTER THREE

Knowledge work and software engineering

Two competently run randomised studies disagree about the sign of the effect. That disagreement is more useful than either number on its own.

A white paper that reports the productivity literature as a single number is easier to write and worse than useless, because the reader takes that number into a business case and defends it in a board meeting. What follows sets out the strongest available studies with their design, their limitations, their sample sizes and their disagreements intact.

Cui and colleagues pooled three randomised controlled field experiments at Microsoft, Accenture and an anonymous Fortune 100 electronics manufacturer, covering 4,867 developers given access to a commercial coding assistant, and reported a 26.08 per cent increase in completed tasks with a standard error of 10.3 per cent. The work is peer-reviewed, having been published in February 2026. Less experienced developers gained most, with junior developers increasing output by 21 to 40 per cent against 7 to 16 per cent for senior developers.[1] PEER-REVIEWED

Four features of that study matter to anyone building a business case on it. The headline figure is a precision-weighted average of three separate estimates, and only the Microsoft estimate reached conventional significance on its own. Adoption was weak, with the authors reporting that around 30 to 40 per cent of developers in the treatment groups never tried the product. On quality the authors conclude that they find no negative effect on build success rates, although the Accenture estimate taken alone shows a fall of 17.40 percentage points with a standard error of 7.12, which is a reminder that a pooled reassurance can conceal a firm-level problem. And the outcome measure counts pull requests, which is a proxy for output and says nothing about the value delivered.

TWO MEASURES, TWO DIRECTIONS, NO TRANSFERABLE NUMBER

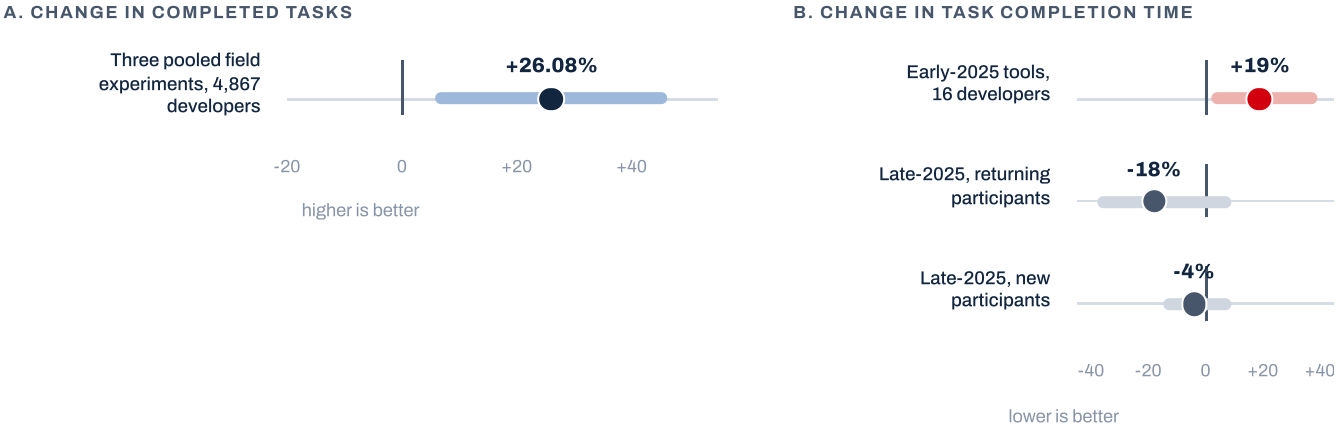


FIGURE 3 Panel a reports the change in completed tasks across three pooled randomised field experiments, with an interval derived from the reported standard error. Panel b reports the change in completion time across three randomised comparisons by the same research group.

References 1, 2 and 3. PEER-REVIEWED

3.1 The counterweight, and its own sequel

A randomised study of 16 experienced open-source developers completing 246 tasks in repositories averaging more than a million lines of code found that access to early-2025 AI tools increased completion time by 19 per cent. The perception gap is the part worth carrying forward. Participants forecast a 24 per cent reduction before starting and estimated a 20 per cent reduction afterwards, having in fact been slowed down, while external experts predicted improvements of 38 to 39 per cent. The interval around the 19 per cent figure runs from 2 to 39 per cent, which appears numerically only in a later publication, and it sits close enough to zero to deserve caution.[2] PREPRINT

The same group then ran a larger follow-up between August 2025 and February 2026, covering 57 developers and more than 800 tasks across 143 repositories. That study found point estimates in the opposite direction, with confidence intervals crossing zero in both participant groups, and its authors describe their own data as only very weak evidence, citing severe selection effects, since between 30 and 50 per cent of developers told them they were holding back tasks they did not want to attempt without AI.[3] PREPRINT A 2026 paper citing the 19 per cent slowdown without its sequel is quoting a result the authors have themselves qualified.

Read together, these studies support a narrower claim than either supports alone. Generative AI assistance changes the time cost of skilled work by an amount that depends on task type, workforce experience, codebase familiarity and tooling, and the dependence is strong enough that the sign of the effect reverses across settings. My own view is that the spread across these studies, running from a 26 per cent gain in output

to a 19 per cent loss of speed, is itself the central finding, and that an organisation quoting either endpoint in a business case has misread both papers. PEER-REVIEWED / PRACTICE

WHAT THIS RULES OUT

No published productivity percentage imported into a business case without a local baseline. Elapsed time on its own will not do as a pilot measure, since it is the number most likely to improve and least likely to capture what degraded underneath it. Nor will an inference from a developer study to a research or clinical workflow, which differs in task, workforce, verification and consequence. The pattern in all three is the same, which is a number borrowed from a setting that was never described.

CHAPTER FOUR

Clinical and research tasks

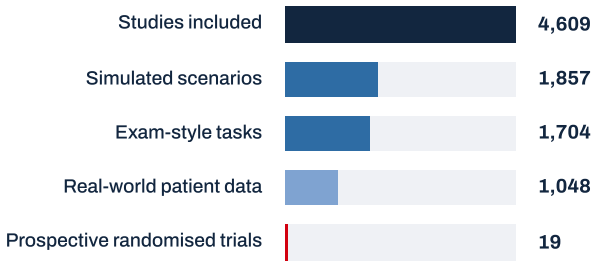
The literature is large and fast-growing, and thin where it matters, and the apparent advantage shrinks every time the evaluation gets more realistic.

The most useful recent survey of the clinical evidence is a systematic review published in Nature Medicine in March 2026, which screened 12,894 records and included 4,609 peer-reviewed studies of large language models in clinical medicine published between January 2022 and September 2025, sorting them into tiers by evidential strength.[4] REVIEW Its findings describe a field publishing faster than it is accumulating usable evidence.

Publication ran at roughly 3.2 papers a day. Of the 4,609 studies, 1,048 used real-world patient data, and 19 of those were prospective randomised trials, while most of the remainder addressed simulated scenarios, at 1,857 studies, or exam-style tasks, at 1,704. At least a quarter had sample sizes below 30. Closed-source models accounted for 87.7 per cent of evaluations, with models from a single vendor making up 65.7 per cent.

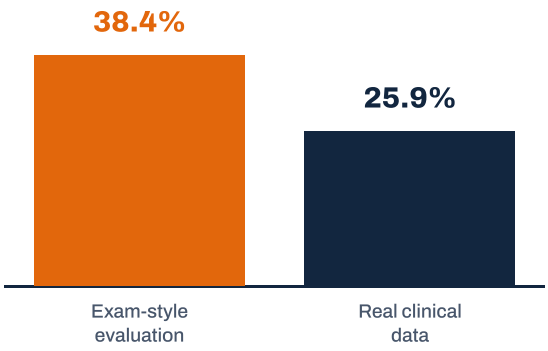
THE SHAPE OF THE CLINICAL EVIDENCE

A. WHAT THE 4,609 INCLUDED STUDIES CONTAIN



Categories overlap. The trials are a subset of the real-world group.

B. COMPARISONS WON BY THE MODEL



Share of 1,046 head-to-head comparisons. The advantage falls as the evaluation becomes more realistic.

FIGURE 4 Panel a shows the composition of the 4,609 studies included from 12,894 screened to September 2025, where the categories overlap because the trials are a subset of the real-world group. Panel b shows the share of 1,046 head-to-head comparisons won by the model, split by how realistic the evaluation was.

Reference 4. REVIEW

The gradient inside that literature is the part which should change how a research programme reads vendor material. Across 1,046 head-to-head comparisons, models outperformed humans in 33 per cent of cases overall, and the rate differed by the realism of the evaluation, running at

38.4 per cent in exam-style studies against 25.9 per cent in studies using real clinical data. The review concludes that performance on contrived knowledge-based examinations does not correlate well with real-world clinical practice, and that rigorous patient-centred evidence remains scarce. The review was itself conducted with model assistance, which the authors disclose, and screening was performed on titles and abstracts only.

4.1 A taxonomy of failure, without a rate

An earlier systematic review covered 89 studies across 29 medical specialties and 124 distinct models. Its method was qualitative thematic coding, so what it produces is a taxonomy of failure. There is no rate in it. Output limitations were coded under nine second-order and 32 third-order categories including incorrectness, non-comprehensiveness, non-reproducibility, unsafety and bias.[5] REVIEW Because its search window closed at the end of 2023, it predates the reasoning models now in common use, and it should be cited for the structure of the failure modes. Its frequencies are already out of date.

The absence of a transferable error rate is itself the operational finding. A programme cannot buy an accuracy figure and inherit it. It has to measure the error rate on its own task, with its own corpus, against its own definition of a critical error, which is the argument Chapter 13 makes concrete.

A USEFUL ABSENCE

Neither review reports a hallucination rate that transfers between settings. Any vendor quoting one is quoting a benchmark, not a deployment.

CHAPTER FIVE

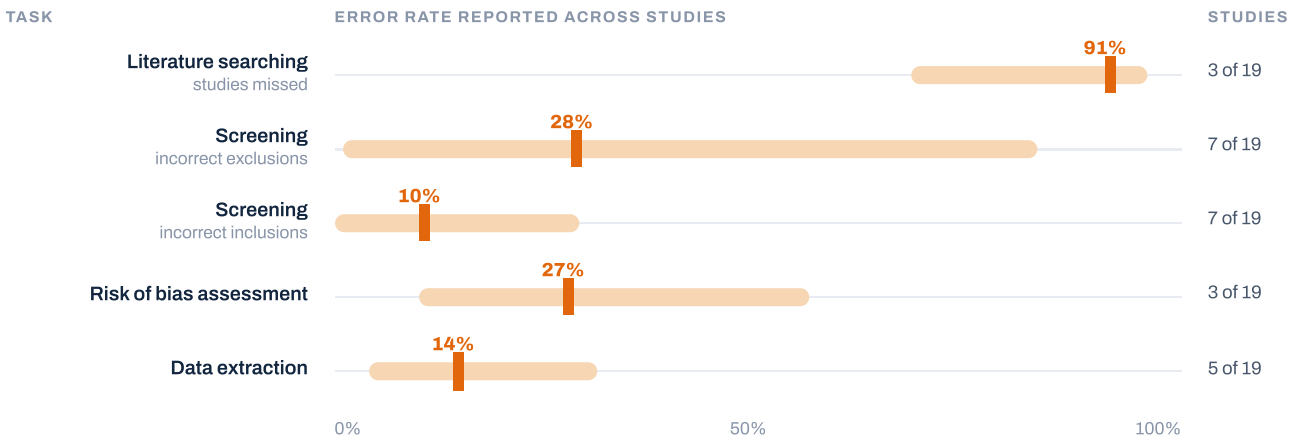
Evidence synthesis, where the evidence is worst

Literature searching sits near the top of almost every candidate list and near the bottom of the evidence, and its failure mode is silent.

Literature searching and screening appear near the top of almost every research organisation's list of candidate AI use cases. On the available evidence they belong near the bottom, and the gap between where they sit and where they belong is the most actionable finding in this paper.

Clark and colleagues systematically reviewed 19 comparative studies that benchmarked generative AI against human experts across evidence synthesis tasks. For literature searching, models missed between 68 and 96 per cent of relevant studies, with a median of 91 per cent. For title and abstract screening, incorrect inclusions ran from 0 to 29 per cent with a median of 10 per cent, and incorrect exclusions ranged from 1 to 83 per cent with a median of 28 per cent. Data extraction produced incorrect extractions in 4 to 31 per cent of cases, median 14 per cent. Risk of bias assessment produced errors in 10 to 56 per cent of cases, median 27 per cent, with better performance on randomised trials than on observational studies. The authors conclude that the current evidence does not support generative AI use in evidence synthesis without human involvement or oversight, and they recommend against its use for searching specifically.[6] REVIEW The practical difficulty underneath that recommendation is duller than the statistics and probably more decisive, which is that checking one of these searches properly takes longer than running the search by hand, so the control that would catch the problem is the first thing to go when the protocol is due. PRACTICE

HOW THE TASKS COMPARE, AND ON HOW MANY STUDIES



Bars show the reported range, and the marker shows the median. Each median rests on a handful of studies.

FIGURE 5 Median error across four evidence synthesis tasks, with the range reported in the review and the number of included studies behind each figure. Searching rests on three studies, which is the caveat that belongs beside the headline.

Reference 6. REVIEW

Two qualifications belong beside those numbers. The searching result rests on three of the 19 included studies, screening on seven, extraction on five and risk of bias on three, so each median summarises a handful of observations. The same review also reports a very large time saving in searching, at five to 57 minutes against 644 for the human comparison, which is the trade-off a business case would be built on.

Even with those qualifications, the direction of the finding should be read slowly by anyone about to commission a literature assistant. A tool that misses most of the relevant studies is a systematically biased sampling instrument whose output looks like a search result, and the failure is silent, because a missing study leaves no trace in the artefact the researcher receives. A large apparent saving beside an invisible loss is the most dangerous shape a use case can take.

ILLUSTRATIVE SCENARIO

The review that could not be reviewed

An invented research team uses an AI assistant to scope the literature for a grant application. The assistant returns 40 studies in four minutes, correctly formatted and plausibly relevant, against the eight hours the team would have spent. Nine months later a reviewer asks why a well-known cohort study is absent from the background. Nobody can answer, because the search that produced the 40 studies was never recorded, the assistant has since been updated, and the artefact contains no statement of what it did not find.

The failure is not that the tool was wrong. It is that the tool produced an object indistinguishable from a complete search, and the team had no way to tell the difference at the moment it mattered.

CHAPTER SIX

Predictive genomics and the utility boundary

A more diverse cohort is scientifically valuable and it is not, by itself, evidence that a model derived from it will perform equitably.

Programmes holding genotype data at population scale face a version of this question that is older than generative AI and is now being asked with new urgency, because the arrival of capable general-purpose models tends to raise the ambition of everything around them.

A scoping review of 37 genomic prediction methods published between 2013 and 2023 found that 12 were suitable for cross-ancestry prediction, and that transferability across ancestries varies with trait heritability and the diversity of the training data, while admixed populations remain difficult. The same review identifies the absence of standard error measurements for individual polygenic scores as a critical gap, which matters because a score reported to a person without an interval invites a certainty the method cannot support. It reports that absolute accuracies often fall short of clinical utility.[13] REVIEW

The consortium literature frames the equity problem directly. A multi-site consortium description records how the historical oversampling of populations with European ancestry in genome-wide association studies causes polygenic scores to perform less well in understudied populations, leading to concerns that clinical use in current forms could widen healthcare disparities.[14] REVIEW That paper describes a programme of work. No result has been reported yet, and it is cited here for the framing it supplies.

A more diverse cohort is scientifically valuable and it is not, by itself, evidence that a model derived from it will perform equitably. Representation in the training data is a necessary condition for equitable performance. It is not a sufficient one, and demonstrating sufficiency requires subgroup-specific validation with reported uncertainty. My own view is that the distinction between research utility and clinical utility is the single most important line for a programme of this kind to hold in writing, because it is the line that internal enthusiasm erodes first and the line a regulator will examine first. PRACTICE

WHAT THIS RULES OUT

No individual-level risk score communicated without an interval, and where the method cannot produce an interval, no score. Cohort diversity does not demonstrate equitable model performance, and only subgroup validation with reported uncertainty does. A research instrument re-described as clinical decision support needs the regulatory and clinical safety route that Chapter 10 sets out, and re-describing it is the part organisations do without noticing.

One question sits underneath all three, and this paper does not answer it. A polygenic score that is well calibrated across a cohort and close to uninformative for the individual in front of you is a legitimate research output and a poor basis for anything a participant reads, and the distance between those two statements is not a technical quantity that a validation study can settle. Where the line falls is a judgement about the person receiving the result, made by people who will not be in the room when it is read. I have not yet seen a programme that has written that judgement down in a form anyone could audit. PRACTICE

CHAPTER SEVEN

Where the positive evidence is strongest

The successes share a mechanism rather than a category, and the mechanism is a better guide to a first pilot than any taxonomy of use cases.

The clearest positive findings in health settings concern ambient documentation. A pre-post survey study across six United States health systems, covering 263 ambulatory practitioners over 30 days, reported burnout falling from 51.9 to 38.8 per cent, an odds ratio of 0.26 with a 95 per cent confidence interval from 0.13 to 0.54, alongside improvements in note-related cognitive task load and a reduction of 0.90 hours in after-hours documentation.[7] PEER-REVIEWED The design has no control arm and the outcomes are self-reported, so the authors state that AI may help reduce administrative burden, and stop there.

A larger observational study at a single health system compared 698 adopters against 867 non-adopters across approximately 1.2 million ambulatory encounters, reporting 1.81 additional work relative value units a week, a 5.8 per cent increase, and 0.80 additional encounters a week.[8] PEER-REVIEWED Adoption was voluntary, so adopters and non-adopters differ systematically, and the authors state that they cannot separate a genuine productivity gain from improved coding practice.

THE THREE CONDITIONS THE POSITIVE EVIDENCE SHARES

1

Verification is immediate

The output is checked at the moment it is produced, while the reviewer still holds the context that would reveal an error.

2

The verifier is accountable

The person checking the output is the person who owns the consequence, rather than a reviewer downstream of the decision.

3

The reviewer knows independently

Verification draws on knowledge the reviewer holds separately from the tool, so agreement carries information.

Where all three hold, the evidence is encouraging

Ambient clinical documentation satisfies all three, and its evidence is the strongest in the health literature. Evidence synthesis satisfies none of them, and its evidence is the weakest. The mechanism is a better guide to a first pilot than the category.

Proposed as a reading across studies that were not designed to test it.

FIGURE 6 Where all three hold, the evidence is encouraging. Where any one fails, it weakens. Evidence synthesis fails all three, which is consistent with Chapter 5.

Ecaveo framework, read against references 6, 7 and 8. ECAVEO

My own view is that these successes share a mechanism, and that the mechanism is more useful than the category. In each case the output is

checked by the person who is accountable for it, at the moment it is produced, against something that person already knows. A clinician reads a draft note about a consultation they have just conducted. Verification is immediate, it is performed by the accountable party, and it draws on knowledge the reviewer holds independently of the tool. Where all three conditions hold, the evidence is encouraging. Where any one fails, and evidence synthesis fails all three, the evidence weakens considerably. ECAVEO

7.1 Retrieval, and the limits of grounding

Retrieval-augmented generation is the standard proposed remedy for fabrication, and the strongest empirical test of that remedy comes from law. Magesh and colleagues put 202 pre-registered queries to three commercial legal research tools and to a general-purpose model, defining a hallucination as a response containing incorrect information or a false assertion that a source supports a proposition. The best performing tool answered 65 per cent of queries accurately and hallucinated on 17 per cent. A second tool was accurate on 42 per cent and hallucinated on 33 per cent. A third was accurate on 20 per cent, failing mostly by declining or half-answering, which is a different problem and a less dangerous one. The authors conclude that vendor claims of hallucination-free operation are overstated.[12] PEER-REVIEWED

Two qualifications apply. The study examined legal research tools at a point in time, and those products have since been revised. The domain is not medicine, so the finding transfers by argument, and the argument is that both domains share the properties the authors identify as defeating retrieval, namely questions without a single definitive answer, relevance that depends on non-textual factors, and synthesis that requires reasoning retrieval does not supply. I could find no peer-reviewed study of comparable rigour testing retrieval-augmented generation against hallucination in a medical or research setting, which is itself informative about the state of the evidence. PRACTICE

SPECIFICATION

What a system should display alongside a retrieved answer

- Which sources were searched, and the date of the corpus, because a superseded document answers confidently.
- Which retrieved passages support each claim, at the level of the individual claim.
- What the search did not cover, stated on the face of the output.
- A route to the source that a reviewer will actually follow, which in practice means one click and no login.

PART TWO

Oversight and the perimeter

Four chapters on the control every framework relies on and the evidence that it fails, and on where the regulatory question has moved since 2024.

CHAPTER EIGHT

The oversight problem

Human review is the control almost every framework relies on, and it has the weakest empirical support of anything in common use.

Almost every AI governance framework in circulation, including the one proposed in this paper, relies on human review as its principal safeguard. The empirical record for that safeguard is poor, and a sector whose entire assurance model rests on qualified people checking outputs should look at the record directly. Three studies show the control failing in three different ways.

Through indifference. Goh and colleagues randomised 50 physicians in a single-blind trial, giving one arm a large language model alongside conventional resources and the other conventional resources alone, with 60 minutes to work through up to six clinical vignettes. Median diagnostic reasoning scores were 76 per cent in the model arm and 74 per cent in the control arm, a difference of two percentage points with a 95 per cent confidence interval from minus four to plus eight, which did not reach significance. Time per case did not differ significantly either. The model working alone reached a median of 92 per cent, an adjusted 16 percentage points above the unaided physicians with a confidence interval from two to 30.[9] PEER-REVIEWED Read carefully, the trial reports two results, because access to the model left the clinicians roughly where it found them, and the pairing performed well below what the model achieved on its own. Participants received no training in prompting, the vignettes were curated summaries, the encounters were not live, and the sample is small, so the result should be treated as indicative.

NEITHER PARTY IMPROVED THE OTHER

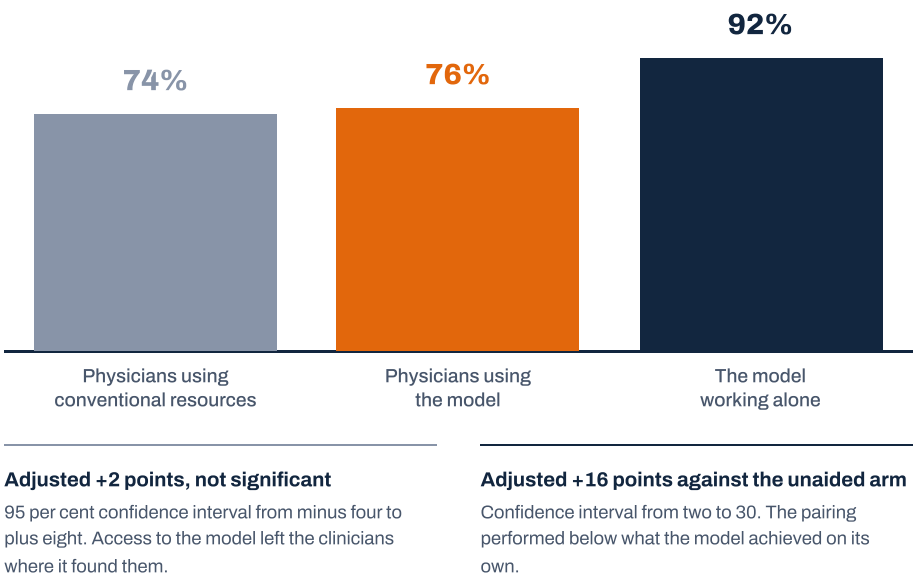


FIGURE 7 Median diagnostic reasoning scores from a single-blind randomised trial of 50 physicians working through up to six clinical vignettes. The marked differences are the adjusted between-group estimates reported by the authors. They are not the arithmetic gaps between the medians shown.

Reference 9. PEER-REVIEWED

Through credulity. A randomised trial of 44 physicians, all of whom had completed 20 hours of AI literacy training, gave one arm error-free model output and the other deliberately erroneous output in half the vignettes. Accuracy fell from 84.9 to 73.3 per cent, an adjusted difference of 14.0 percentage points with a confidence interval from 8.3 to 19.7, and the effect on the top-choice diagnosis was larger at 18.3 points.[10] PEER-REVIEWED The finding that matters for governance design is that the training did not inoculate. The errors were injected, the vignettes are not live encounters, and the trial ran at a single site with a small sample, so it should be read as a warning.

Through atrophy. A multicentre observational study across four Polish endoscopy centres compared adenoma detection rates in standard colonoscopy during the three months before and the three months after AI assistance was introduced, covering 1,443 procedures performed without AI. Detection fell from 28.4 to 22.4 per cent, an absolute difference of 6.0 percentage points with a confidence interval from 1.6 to 10.5 and a p value of 0.0089, and exposure to AI was independently associated with lower detection at an odds ratio of 0.69.[11] PEER-REVIEWED The authors state that continuous exposure to AI might reduce the detection rate of standard colonoscopy, suggesting a negative effect on endoscopist behaviour. The design is retrospective and confined to one country, so secular trends cannot be excluded. I went looking for the opposite result here, on the assumption that experienced people checking machine output must do better than this, and the search did not

produce one. PRACTICE

8.1 What the failures have in common

The three studies describe different mechanisms and share one property. In every case the review step existed, because a physician read the output, a clinician performed the procedure, and nobody skipped a stage. What failed was the design of the task the human had been given.

My own view, which goes beyond what any of these studies claims, is that human in the loop as written in most AI policies specifies a person and omits a procedure. It names who reviews. It leaves unstated what they see, what they compare it against, how long they have, what authority they hold to reject, and how anyone would know afterwards whether the review was performed with care. A control specified at that level of abstraction is an assertion about intent. It produces no evidence, and the three failures above are what an unspecified control looks like in operation. ECAVEO

SPECIFICATION

Four properties that turn a nominal review into an auditable control

- The reviewer sees the evidence and not only the conclusion, because a summary that omits its own uncertainty leaves nothing to disagree with.
- Disagreement is cheap. A workflow in which rejecting the output means redoing the work by hand will produce agreement regardless of quality, and that is a design fault.
- Review quality is itself observable, which means sampling accepted outputs for independent checking and reporting the results to the forum that approved the deployment.
- The workflow preserves the unaided competence it depends on, whether through rotation, periodic unassisted work or deliberate calibration, since the colonoscopy finding suggests competence does not preserve itself.

Each of those four is a design decision with a cost. An organisation unwilling to pay any of them has decided, in effect, to deploy without the control it has written down.

CHAPTER NINE

Data protection after the 2025 Act

The reform is widely misdescribed. It is liberalising for ordinary personal data and it keeps a tight gate where special category data is involved.

> Nothing in this chapter or the two that follow constitutes legal advice. The material supports discussion and further review by qualified advisers, and statements about how any particular organisation stands under these instruments are inference, and are offered as such.

The United Kingdom has no cross-sector artificial intelligence statute. The House of Lords Library recorded in May 2026 that the 2026 King's Speech contained no general AI bill, and the policy direction stated by ministers has been to address specific areas one at a time.[36] REGULATOR What has changed is less visible and more consequential than a statute would have been, because three regulators have converged on the same question from different directions.

The Data (Use and Access) Act 2025 received Royal Assent on 19 June 2025, and sections 67, 68, 80 and 86 came into force on 5 February 2026 under the sixth commencement regulations.[15,16] LEGISLATION Section 80 replaces Article 22 of the UK General Data Protection Regulation with new Articles 22A to 22D.

The direction of that reform is widely misdescribed, and the misdescription matters for this sector. For ordinary personal data the change is liberalising, since the previous general prohibition on solely automated significant decision-making no longer applies and a much wider set of lawful bases becomes available. For special category data the gate remains tight. Article 22B provides that a significant decision based entirely or partly on special category data may not be taken solely automatically unless the data subject has given explicit consent, or the decision is either necessary for entering into or performing a contract or required or authorised by law, and in that second case Article 9(2)(g) must also apply. Article 22B also prohibits a solely automated significant decision where the processing relies wholly or partly on the new recognised legitimate interests basis, so the widening described above stops short of the full set of lawful bases. Article 22C retains four safeguards, requiring the controller to provide information about such decisions, to enable the data subject to make representations, to enable human intervention on the part of the controller, and to enable the data subject to contest the decision. The Secretary of State's regulation-making power under Article 22D cannot amend Article 22C.[15] LEGISLATION

THE OPERATIVE DATE

5 February 2026. The automated decision reform and the research provisions are in force, and guidance written against the previous regime is still being applied to them.

9.1 The research provisions, which matter more

Sections 67 and 68 matter more to research operations than the automated decision provisions do. Section 67 defines processing for scientific research purposes to include any research that can reasonably be described as scientific, whether publicly or privately funded and whether commercial or non-commercial. Section 68 permits consent to an area of scientific research where the purpose cannot be fully specified at the time, provided that seeking consent in that form is consistent with generally recognised ethical standards and the data subject can consent to part only.[15,20] LEGISLATION

Two cautions attach to reading this reform as room to move. The Information Commissioner's Office guidance on AI and data protection was last updated on 15 March 2023 and now carries a notice that it is under review and may be subject to change because of the Act, so guidance written against the previous automated decision regime is being applied to a regime that has changed.[18] REGULATOR Separately, the Commissioner acquired a statutory duty in 2026 to prepare a code of practice on developing and using artificial intelligence and on automated decision-making, under regulations made on 16 April 2026 and in force from 12 May 2026. That code is not yet published.[17] LEGISLATION

My own view is that an organisation designing its AI intake process this year should design it so that the code, when it arrives, can be met by an update, which in practice means recording intended purpose, data flow, human involvement and evaluation evidence in a form that would survive a change of vocabulary. PRACTICE

On impact assessments the existing position is unchanged. The Commissioner's guidance cites Article 35(3) and identifies large-scale processing of special categories of personal data, and systematic and extensive evaluation of personal aspects based on automated processing on which decisions producing legal or similarly significant effects are made, among the circumstances requiring an assessment.[18] REGULATOR For most programmes in this sector the first of those is satisfied continuously, which is an argument for a screening step at intake. Treating every use case as requiring a full assessment wastes the assessment. Procurement should test vendor claims against contracts and technical configuration, and the Commissioner's audit framework toolkit on contracts and third parties is the practical instrument for that.[19] REGULATOR

CHAPTER TEN

Medical devices and the clinical boundary

The marketing copy for an internal tool can create a regulatory obligation the engineering never contemplated, and a disclaimer will not undo it.

The Medicines and Healthcare products Regulatory Agency published guidance on ambient voice technology enabled products on 29 July 2026, and although its subject is documentation tools it is the clearest recent statement of how the regulator approaches this class of software generally. It sits alongside the Agency's standing guidance on software and artificial intelligence as a medical device, which it does not displace.[21,22] REGULATOR

Three propositions in that guidance carry beyond ambient scribes. The guidance provides that use in a medical environment or context alone does not qualify a product as a medical device, so deployment inside a clinic is not the test. It provides that a product's intended purpose is not specifically determined by the inclusion of any particular technology, so being built on a large language model neither creates nor avoids device status. And it provides that intended purpose is derived from claims in a product's instructions for use, labelling and the manufacturer's promotional materials, adding that general disclaimers, for example that a product is not for diagnosis, are not acceptable where medical claims appear elsewhere. On generative models specifically, the guidance states that hallucination is a well-known behaviour with potentially broad impact in such products, to be mitigated by design.

WHAT MOVES A TOOL ACROSS THE BOUNDARY

OUTSIDE THE DEVICE PERIMETER		INSIDE IT, ON THE SAME CODE
Helps researchers find datasets and fields.	→ A CHANGE OF DESCRIPTION, NOT OF CODE	Supports a clinical decision.
Drafts internal documentation for staff review.		Produces a participant-specific health interpretation.
Summarises a policy for an employee.		Is described in any document as doing either.

> The regulator provides that intended purpose is derived from claims in instructions for use, labelling and promotional materials, and that general disclaimers are not acceptable where medical claims appear elsewhere. Use inside a clinic is not the test, and neither is the technology.

FIGURE 8 The regulator locates intended purpose in what the organisation claims, not in the technology it used or the room it sits in. The move is a change of description. The engineering is untouched.

Ecaveo reading of reference 21. Inference on the author's part, and not a regulatory statement. ECAVEO

The practical reading, and this is inference on my part, is that the

marketing copy for an internal tool can create a regulatory obligation the engineering never contemplated. A research support assistant described internally as helping researchers find datasets sits comfortably outside the device perimeter. The same tool described in a funding paper as supporting clinical decisions may not. Of the two descriptions, the second is the one a regulator would read. PRACTICE

The regulator's sandbox programme has continued through this period. The AI Airlock pilot report was published in October 2025, phase two completed in May 2026 with seven innovators across clinical note-taking, cancer diagnostics, eye disease detection and obesity treatment support, and the phase two programme report followed on 27 July 2026, with phase three in design.[23] REGULATOR On post-market obligations, the Medical Devices (Post-market Surveillance Requirements) (Amendment) (Great Britain) Regulations 2024 came into force on 16 June 2025.[24] LEGISLATION The pre-market statutory instrument remains in draft, and no commencement date should be assumed.

10.1 Clinical risk management in deployment

Two NHS standards govern clinical risk management for health software, namely DCB0129 for manufacturers and DCB0160 for deploying care organisations. Both remain mandatory in their current versions, and both are under national review.

NHS England published supporting information for that review on 29 June 2026, opening a consultation that closes on 11 September 2026. The review states that artificial intelligence, machine learning, ambient voice technologies and complex algorithmic systems are increasingly deployed in clinical settings, and that these technologies present novel risk profiles requiring specific governance frameworks that current standards do not adequately address.[25] REGULATOR That sentence is the most candid official acknowledgement available that the deployed-safety framework has not caught up with the technology, and an organisation near the clinical boundary should read it as a reason to document its own reasoning now.

Separately, NHS England guidance on AI-enabled ambient scribing products, whose second iteration was published on 29 July 2026, requires suppliers to provide DCB0129 evidence at procurement and deploying organisations to complete DCB0160 documentation and a data protection impact assessment before deployment. The guidance states that it does not determine which products meet the definition of a medical device, deferring that question to the Agency.[26] REGULATOR

WHAT THIS RULES OUT

No participant-facing health interpretation produced by a general-purpose model outside the regulatory and clinical safety route. A disclaimer will not undo a medical claim made elsewhere in the same organisation's materials, and the regulator says so directly. And a research label protects nothing once the described purpose has moved.

CHAPTER ELEVEN

Research governance, security and standards

The guidance for this field is being written now, and the bodies writing it are open to organisations that engage early.

In May 2026 the Health Research Authority published a two-year plan to enable safe AI-powered innovation in health and social care research, working across the four nations. Its three priorities are to be clear about where AI development, evaluation and implementation activities count as research, to clarify when health information may be accessed using AI-enabled and data-driven approaches to identify and contact people about relevant research, and to make the review of such research better informed, more rigorous and more consistent.[27] REGULATOR

The Authority also supports the National Commission on the Regulation of AI in Healthcare, launched in September 2025 under the chairmanship of Professor Alastair Denniston, which is advising on the regulatory framework.[28] REGULATOR Both initiatives carry the same significance for a programme planning AI-assisted research operations. The guidance is being written now, and the bodies writing it are open to organisations that engage early. Underneath both, the UK Policy Framework for Health and Social Care Research, last updated on 10 January 2025, continues to supply the governing principles for research conduct and is displaced by none of the above.[29] REGULATOR

11.1 Security and technical standards

The National Cyber Security Centre published Guidelines for Secure AI System Development in November 2023, structured around secure design, development, deployment, and operation and maintenance.[30] REGULATOR It remains the most useful starting point for a programme building its own internal standard, and its four-stage structure maps onto the assurance record set out in Chapter 13.

The United States National Institute of Standards and Technology published its Generative AI Profile, AI 600-1, on 26 July 2024, and it remains the most complete public taxonomy of generative AI risk suitable for adapting into an internal framework.[31] REGULATOR The parent AI Risk Management Framework is under revision and no revised version has been published. The OWASP Top 10 for large language model applications, 2025 edition, lists prompt injection, sensitive information disclosure, supply chain, data and model poisoning, improper output handling, excessive agency, system prompt leakage, vector and embedding weaknesses, misinformation and unbounded consumption.[32] REGULATOR Excessive agency is the entry that most directly concerns any programme considering tools that act on the world, and it is the reason

Chapter 14 defers broad autonomy.

Two further instruments are worth holding. The government's AI Playbook, published in February 2025, sets out ten principles for public sector use and is the document a partner organisation is most likely to cite back.[33] REGULATOR The AI Opportunities Action Plan of January 2025 explains the policy direction that funders are working within.[34] REGULATOR For assurance vocabulary specifically, the Department for Science, Innovation and Technology's introduction to AI assurance remains the common reference.[35] REGULATOR Organisations with European operations should also track the staged application of the EU AI Act and the deferrals agreed in 2026, which change the dates and leave the obligations where they were.[37] LEGISLATION

11.2 What has actually changed

Read together, the instruments above describe a shift in where the regulatory question sits. In 2024 the common question was whether a model was safe enough to use. In 2026 the question the Agency asks is what the organisation claims the product does, the question NHS England asks is who holds clinical risk when it is deployed, and the question the Commissioner asks is what decision is being made about a person and what recourse that person has. None of those questions is answered by a model evaluation, and all three are answered by documentation the organisation controls. ECAVEO

PART THREE

Building the capability

Five chapters on what to measure, what to tier, what to pilot and what to refuse, and on the sequence that makes each of those decisions answerable.

CHAPTER TWELVE

An operating model

Ordinary controls, specified before deployment rather than assembled after an incident, and connected to a measurement that can test them.

The controls in this chapter are ordinary. Their value lies in being specified before deployment, and in being connected to the evaluation capability described in Chapter 13, without which each of them is a statement of intention.

12.1 A tier for every use case

The tiers below are ordered by recall cost, as argued in Chapter 2, with data sensitivity operating as a modifier that can raise a use case by one tier but never lower it. ECAVEO

RISK TIERS ORDERED BY RECALL COST, WITH DATA SENSITIVITY AS A MODIFIER

TIER	RECALL POSITION	MINIMUM CONTROLS	APPROVAL
One	Output stays with its author and can be discarded without trace. Internal drafting, retrieval and exploration on non-personal content.	Approved tool, user training, source checking, no external use without review.	Team owner
Two	Output informs an internal decision or reaches colleagues, and a correction reaches everyone who saw it.	Registered use case, security and privacy screening, evaluation set, logging, named reviewer.	Function or product owner, with security or data protection as relevant
Three	Output leaves the organisation, enters the scientific record or triggers a material operational action, so a correction reaches some of its audience at best.	DPIA screen and the assessment where indicated, threat model, documented data flow, representative testing, incident and rollback plan, formal sign-off.	Data protection officer, security, and the accountable scientific or operational owner
Four	Output shapes a decision about a person's health, or forms part of care, where the consequence cannot be recalled at all.	All tier three controls, plus a clinical safety case, a regulatory classification decision, prospective evidence, subgroup validation and post-deployment surveillance.	Executive sponsor with data protection, security, clinical safety and regulatory governance

12.2 Extend the Five Safes, and add a sixth

Programmes in this sector already operate the Five Safes, which govern people, projects, data, settings and outputs. The framework is a good fit for AI precisely because it was designed to control access and disclosure in a shared environment, and a language the organisation already speaks is worth far more than a parallel governance structure nobody reads.

THE FIVE SAFES EXTENDED FOR AI, WITH A SIXTH PROPOSED

SAFE	EXTENSION FOR AI	THE CONTROL QUESTION
Safe people	Authorised and trained users with named accountability for outputs and actions.	Who may prompt, approve, override and investigate, and what competence does each of those require?
Safe projects	A registered intended use with a measurable benefit and a list of prohibited uses.	What problem is being solved, is a model necessary to solve it, and which decisions must remain human?
Safe data	Minimised inputs, lawful processing and protection against reuse of content for model training.	What leaves the boundary, is it retained, and are prompts, attachments, telemetry and embeddings all covered?
Safe settings	Approved models, tools, connectors, network paths and identity controls.	Where does inference happen, and which services, sub-processors and regions are involved?
Safe outputs	Evidence-backed, reviewed and monitored outputs and actions.	How is factual accuracy checked, what can an agent do unaided, and could an output leak or re-identify?
Safe evidence proposed	Whether the tool works, measured before and after deployment.	Was the benefit defined and the baseline captured before access was granted, was the evaluation designed before the results were seen, and is it repeated when the system changes?

The five extensions above cover access and disclosure, which is what the framework was built for. They do not cover whether the tool works. My own view is that AI adoption requires a sixth Safe, which might be called safe evidence, covering whether the intended benefit was defined before deployment, whether a baseline exists, whether the evaluation was designed before the results were seen, whether quality and equity were measured alongside speed, and whether the evaluation is repeated when the model, prompt, retrieval corpus or workflow changes. The existing five will stop an organisation disclosing something it should not. None of them will stop it deploying something that does not work. ECAVEO

12.3 Technical controls that are not negotiable

The following are stated as absolutes because each one, if traded away, removes the evidential basis for a claim made elsewhere in an assurance record.

- No participant, genetic, health record or research data enters a model or connector until the exact data flow and processor arrangement have been approved and recorded.
- Enterprise agreements disable provider training on organisational content where that option exists, and they define retention, region, sub-processors, deletion and incident duties in the contract itself.
- Agents receive short-lived, least-privilege credentials and a bounded set of tools, run with default-deny networking, require explicit approval for high-impact writes, log every action in the same way as user activity, and can be stopped quickly.

- Retrieval systems enforce source permissions before retrieval, because a model is not an authorisation layer and cannot be made into one. Filtering after generation is too late.
- Prompt injection, malicious documents, sensitive information disclosure, improper output handling and excessive agency are covered in threat modelling and in red-team testing, before deployment and after material change.
- Model, prompt, retrieval corpus, tool and policy versions are recorded, so that an evaluation can be reproduced and an incident can be investigated.
- Outputs used in code, in science or in external communication pass the same tests, reviews and publication controls as work produced without assistance.

These align with the Centre's secure development guidance, the Commissioner's expectations on AI supply chains, the Institute's generative AI profile and the OWASP taxonomy.[30,19,31,32] REGULATOR None of them is novel. The discipline lies in applying them as rigorously to the last use case of the year as to the first. Whether any organisation sustains that into its second year is a question I cannot answer from the engagements I have seen.

CHAPTER THIRTEEN

Evaluation and assurance

An organisation should not deploy above the level of evidence it can produce, which requires knowing which level it is at.

The central recommendation of this paper is that a programme should not deploy above the level of evidence it can produce. That requires knowing which level it is at, and the five levels in Figure 1 are offered as a working scale, and no more than that.

Most organisations approaching their first pilots operate at level one and believe they operate at level three. The diagnostic is simple and uncomfortable, since an organisation at level two can state, for a workflow it has already changed, what the baseline was, how the comparison was structured and what the result was. An organisation that cannot answer those three questions about work already done should not commit to a use case whose output leaves the building. ECAVEO

THE EVALUATION CAPABILITY LADDER, AND THE DEPLOYMENT EACH LEVEL SUPPORTS

LEVEL	WHAT THE ORGANISATION CAN DO	HIGHEST TIER IT SHOULD DEPLOY
0. Anecdote	Collect user impressions and usage statistics.	Nothing beyond personal experimentation
1. Instrumented baseline	State what a named workflow costs today in time, rework and escalation, using data rather than recollection.	Tier one
2. Controlled comparison	Run a paired-task or stepped-wedge comparison with a quality rubric and a defined sample.	Tier two
3. Pre-registered protocol	Fix the primary outcome and analysis before results are seen, include subgroup and equity checks, and use blinded review where feasible.	Tier three
4. Continuous assurance	Monitor quality and harm in production, detect drift, and re-evaluate automatically when the model, prompt, corpus or workflow changes.	Tier four, alongside clinical and regulatory governance

13.1 Designing an evaluation before deployment

Each pilot should open with a one-page protocol written before anyone is given access, recording the current workflow, the target outcome, the sample, the comparison method, the quality rubric, the plausible harms, the subgroup checks and the conditions under which the pilot stops. Adoption figures are evidence of interest rather than evidence of value, and a pilot that reports only usage has answered a question nobody needed answering.

Pre-registering the primary outcome is the cheapest protection available against self-deception, and the developer productivity literature shows why it is needed. Both of the major studies discussed in Chapter 3

produced noisy estimates whose intervals came close to zero, and in each case the direction of the finding was contested afterwards by serious people. Those were well-resourced research teams with no commercial stake in the answer. An organisation running a single pilot, with a smaller sample and a business case already written, is far more exposed to post-hoc framing than either of them.

SPECIFICATION

Five design choices that cover most situations

- Paired tasks, where the same class of work can be completed with and without assistance without contamination.
- A stepped-wedge rollout, where teams need access at different times and withholding a tool indefinitely is impractical.
- Blinded expert review, for factual, scientific or code quality comparisons.
- Adversarial testing covering prompt injection, misleading evidence, ambiguous requests, permission boundaries and unsafe tool use.
- Repetition after material change to the model, the system prompt, the retrieval corpus, the connectors or the workflow, because a deployment evaluated once was evaluated against a system that no longer exists.

13.2 What to measure, and what would stop a rollout

WHAT TO MEASURE, AND THE GATE EACH MEASURE CONTROLS

DIMENSION	EXAMPLE MEASURES	GATE BEFORE SCALING
Outcome	Cycle time, queue time, tasks completed, time to a validated answer.	A predefined improvement over baseline, reported with an interval rather than a point estimate.
Quality	Defect rate, factual precision, citation support, reviewer corrections, rework.	No material quality regression, and a critical-error rate below the agreed threshold.
Safety and privacy	Policy breaches, sensitive data exposure, unauthorised actions, red-team findings.	No unresolved critical findings, and every high finding either mitigated or formally accepted by a named owner.
Scientific validity	Reproducibility, test pass rates, concordance with independently verified analysis, subgroup calibration.	The statistical and scientific acceptance criteria set out in the protocol.
Human factors	Review time, rate of agreement with model output, user comprehension of limits, escalation quality.	Reviewers can identify limits, disagree at low cost, and recover from a failure.
Equity and access	Performance by relevant subgroup, language and access need.	No unexplained material disparity, with mitigation and monitoring agreed where one is found.
Economics	Licences, compute, integration, assurance and support cost per useful outcome.	A sustainable total cost set against the verified benefit rather than the projected one.
Operational fit	Reliability, latency, maintainability, exposure to vendor change.	A named service owner, a support route, a rollback path and monitoring in place.

13.3 The minimum assurance record

A use case that cannot produce the following record has not been governed, whatever policy it complied with.

THE MINIMUM ASSURANCE RECORD FOR A DEPLOYED USE CASE

RECORD FIELD	REQUIRED CONTENT
Purpose and owner	The problem, the intended users, the benefit claimed, the accountable owner and the review date.
Boundaries	Permitted and prohibited uses, data classes, environments and actions.
System description	Model and version, vendor, prompts, retrieval sources, tools, interfaces and dependencies.
Risk evidence	The impact assessment decision, the threat model, the equality and ethics assessment, and the regulatory and clinical safety determinations.
Evaluation	The dataset or task sample, the method, the results with their limitations, and the approval decision.
Operations	Monitoring, incidents, user feedback, change control, rollback and retirement.

CHAPTER FOURTEEN

A portfolio built to the evidence

Literature searching moves down the list, documentation assistance moves up, and anything reaching a participant carries a communications and clinical safety gate.

The portfolio below revises the conventional opportunity list in light of Part One, and the revisions are deliberate. Literature searching moves down. Documentation assistance moves up. Anything whose output reaches a participant carries a communications and clinical safety gate, whatever the technology behind it. ECAVEO

A CANDIDATE PORTFOLIO, ORDERED BY THE STRENGTH OF THE EVIDENCE BEHIND IT

USE CASE	VALUE HYPOTHESIS	TIER	POSITION
Meeting and documentation synthesis	Time returned from note-taking, with the author verifying immediately.	One to two	The positive evidence in health settings fits this shape, since the author verifies immediately, though it comes from clinical scribing and from uncontrolled designs.
Internal policy and engineering knowledge assistant	Less time searching, and more consistent answers linked to approved sources.	Two	Start here. Access-controlled, citation-required retrieval over non-participant content, with superseded-document failures tested explicitly.
Engineering assistance in controlled repositories	Faster tests, documentation, refactoring and investigation.	Two	Pilot by task class, with branch protection, tests, review and secret scanning, and with defect and review-burden measures compulsory.
AI feature assessment for corporate software	Faster data protection, legal and security assessment with clearer data flow records.	Two	Build a standard questionnaire and evidence pack. People make the decisions, and the tool prepares the evidence.
Researcher support case classification	Shorter response times and better routing.	Two to three	Classification and drafting only, with staff approving every external reply.
Metadata and documentation assistant for the research environment	Faster discovery of datasets, fields, caveats and approved analytic patterns.	Three	Keep retrieval and inference inside the environment boundary and cite authoritative metadata rather than generating descriptions.
Data quality anomaly investigation	Quicker root-cause hypotheses for pipelines and releases.	Three	Investigation support only. Every finding is confirmed by a deterministic check before it informs a decision.
Literature search and evidence synthesis	Quicker evidence mapping and first drafts.	Three	Defer. The comparative evidence is thin, resting on three studies for searching, and what it reports is a median of 91 per cent of relevant studies missed, with the failure leaving no trace in the output.
Participant communications assistance	Faster drafting and more accessible variants.	Three	Approved-source retrieval and human publication control only. High recall cost despite using no personal data.
Participant risk or eligibility decisions	Potential support for targeted prevention or trial recruitment.	Four	Outside an adoption programme. Route to formal research, regulatory and clinical safety governance from the outset.
Autonomous agents with production write access	Fewer manual operational steps.	Four	Defer broad autonomy. Permit bounded actions with least privilege, default-deny networking, approvals, logging and tested rollback.

14.1 Three pilots worth starting with

The first pilot should make internal policy, security and engineering guidance easier to retrieve. It exercises retrieval, access control, citation enforcement, evaluation and user training without touching participant data, and it leaves behind an evaluation set the organisation can reuse on everything that follows. Success means fewer minutes spent searching and fewer escalations, with no unsupported answer accepted as policy. Test explicitly for the confident answer drawn from a superseded

document, because that is the form in which retrieval systems usually fail.

The second pilot should test AI-assisted engineering on named task classes. It exercises sandboxing, repository boundaries, secure coding, test discipline and objective measurement. Given that one of the three field experiments discussed in Chapter 3 recorded a fall of 17.40 percentage points in build success rate, defect and review burden measures are compulsory, and a pilot that reports only elapsed time has measured the thing most likely to improve while ignoring the thing most likely to degrade.

The third pilot should improve how AI features in corporate software are assessed. It exercises cross-functional work between technology, procurement, legal, data protection and security, and its output is a reusable supplier questionnaire, a data flow record and a decision log. This is the least glamorous of the three and probably the most valuable, because without a standard assessment process, AI arrives in the estate anyway and nobody assesses it.

ILLUSTRATIVE SCENARIO

Three uses of one architecture

The Ashdown Cohort does not exist. It is a composite assembled from features common to programmes of this type, and it appears here to make the tiering argument concrete. No real organisation is being described.

Ashdown holds questionnaire, clinic and genotype data for rather more than a million consented participants. In a single quarter its technology team proposes three uses of the same retrieval architecture. The first indexes internal engineering runbooks for the platform team. The second indexes dataset documentation inside the research environment so that approved researchers can find fields, caveats and analytic patterns. The third drafts items for a quarterly participant newsletter from an approved library of published health information, touching no personal data at all.

Ranked by data sensitivity, the second proposal is the one that attracts scrutiny and the third is waved through. Ranked by recall cost, the order changes. The runbook assistant produces outputs that stay with their author. The documentation assistant shapes analyses that reach publications, so its errors persist, which places it in tier three. The newsletter assistant reaches a million people who trust the sender, and a correction issued three months later reaches a fraction of them, which places it in tier three as well despite using no participant information whatsoever.

In the composite, as in every version of this story I have encountered, the newsletter tool is the one that reached production without an evaluation.

CHAPTER FIFTEEN

Delivery over six to nine months

Five phases, three pilots and one evidence template, arranged so that the organisation owns the capability when the engagement ends.

A PHASED PROGRAMME OVER SIX TO NINE MONTHS

PHASE	WEEKS	PRIMARY OUTPUTS	EXIT CONDITION
Understand	1 to 4	Workflow observation, tool and use-case inventory, disclosure survey, risk taxonomy, baseline measures, governance map.	The executive agrees priorities, risk appetite and pilot owners, and the organisation knows its position on the capability ladder.
Prove	5 to 10	Three controlled pilots, evaluation protocols, training, office hours, supplier and data-flow reviews, first red-team tests.	At least two pilots show value without a material quality or risk regression.
Standardise	11 to 18	Intake process, approved patterns, extended Five Safes, assurance record, reusable sandbox patterns, champions network.	Teams can start low-risk work through a clear, repeatable route without consultant involvement.
Scale	19 to 28	Expanded pilots, production monitoring, role-based training, adoption reporting, researcher-facing capability, benefits tracking.	Scaled uses meet their gates and named service owners accept ongoing operation.
Transfer	29 to 36	Strategy refresh, prioritised backlog, lessons report, governance handover, capability assessment, next-year roadmap.	Ownership, documentation, metrics and support survive the end of the engagement.

15.1 The first 30 days

Meet the technology, security, data protection, legal, scientific, product, digital, people and communications teams before proposing anything, and observe real workflows, because a described workflow is a remembered one. Quantify delay, repetition, error and review burden, because those four numbers are the baseline that every later claim will be measured against. Inventory the AI already present in software, browsers, code editors, meeting tools and vendor products, and classify the data paths and controls that inventory reveals, since most organisations discover at this point that adoption began without them. Publish interim safe-use guidance and an easy route for staff to disclose experiments without blame, on the reasoning that undisclosed use is a governance failure and punishing disclosure guarantees more of it. Select three pilots with executive owners, user champions, baselines, evaluation protocols and stop conditions. And demonstrate one bounded agentic workflow in a sandbox, so that leaders see the capability and the control design in the same sitting.

15.2 Operating rhythm

A weekly delivery review tracks pilot evidence, risks, decisions and blockers. A fortnightly clinic with data protection, legal and security resolves intake questions quickly, since the main cause of shadow use is a governance process slower than the work it governs. A monthly executive review makes scale, redesign or stop decisions using one evidence template, so that the comparison between use cases is meaningful. Proposals that touch participants or the public add the relevant ethics, communications, scientific and clinical voices before work begins.

THE MONTHLY DECISION, AND THE EVIDENCE EACH OPTION REQUIRES

DECISION	EVIDENCE NEEDED
Scale	The target benefit achieved, quality maintained, residual risks formally accepted, and support and monitoring funded.
Redesign	A value signal exists, and errors, review burden, workflow fit or controls fall short of the threshold.
Stop	No verified benefit, unacceptable residual risk, poor fit with how people work, vendor dependency, or a simpler method performing better.

15.3 What success looks like at the end

WHAT SUCCESS LOOKS LIKE, AND HOW IT WOULD BE EVIDENCED

OUTCOME	EVIDENCE
Visible value	A small portfolio of scaled workflows with verified improvements in time, quality or service, each traceable to a baseline captured beforehand.
Safer adoption	One intake route, approved patterns, documented tiers, monitored tools, and fewer undisclosed uses than at the start.
Internal capability	Named owners who can design an evaluation, explain a limit and support colleagues without external help.
Stronger governance	Existing risk, assurance and Five Safes processes absorbing AI without a duplicate committee structure.
A credible next step	A backlog that separates internal productivity, research enablement and clinically significant innovation, and prices each accordingly.

CHAPTER SIXTEEN

The intelligent adoption programme

What to adopt, in what order, and what to measure. The first step costs nothing and is the one everybody skips.

The adoption sequence that follows from this paper is deliberately unimpressive, which is the point. The sector has spent three years buying capability and has not yet built the measurement that would tell it whether the capability helped.

16.1 The sequence

First, instrument one workflow, in the ordinary sense of recording what it costs today in time, rework and escalation. Nothing else in this list works without it. Second, record intended purpose for every proposed use, in one line, in the words of the person who wants it, because that line is what a regulator will read. Third, tier by recall cost rather than by data sensitivity. Fourth, specify the review, including what the reviewer sees and what it costs them to disagree. Fifth, pre-register the outcome before anyone is given access. Sixth, and only sixth, consider anything agentic, and expect to conclude that the first five have already delivered most of the value. ECAVEO

16.2 Questions a board should ask before approving any of this

The following are the questions that most reliably expose whether a programme is ready, and a leadership team that can answer them has already done most of the work this paper describes.

- Which organisational outcomes should improve, by when, and which of the relevant metrics already exist in a form somebody trusts?
- Which AI experiments are visible today, and where is undisclosed use most likely to be occurring?
- Which data classes may approved services process at present, and which boundaries are absolute?
- Who holds the decision when data protection, security, science, clinical safety, product and procurement disagree?
- Where does researcher friction currently limit data use, and is that friction a technology problem or a process problem?
- Which participant-facing capabilities are planned, and might any of them acquire a medical intended purpose through the way it is described?
- What should remain deliberately human even after automation becomes technically possible, and who has the standing to hold that line?

- What evidence would give the board and its public representatives confidence that adoption is increasing public benefit, and what would tell them it is only increasing activity?

16.3 What good looks like in three years

A programme that can say what it changed and whether it worked. An intake route fast enough that people use it instead of working around it. A set of reviewers who can explain what their tool cannot tell them, and whose disagreement costs them nothing. And a portfolio in which the uses that reach participants are the best evidenced ones, which is the reverse of where most organisations start.

That is a modest description of progress in a field that has been sold something far more exciting for three years. It is also, on the evidence assembled here, most of what is actually available.

A model evaluation answers none of the three questions the regulators are now asking. The documentation the organisation controls answers all of them.

CHAPTER ELEVEN

END MATTER

Board and executive checklist

Eighteen items. The first four cost nothing and deliver most of the value.

- [ ] **One workflow is instrumented**
  What it costs today in time, rework and escalation, taken from data.
- [ ] **Tiering is by recall cost**
  How expensive the error is to withdraw, with data sensitivity as a modifier.
- [ ] **Disagreement is cheap for the reviewer**
  Rejecting an output must not mean redoing the work by hand.
- [ ] **Unaided competence is preserved**
  Rotation, periodic unassisted work or deliberate calibration.
- [ ] **Quality is measured alongside speed**
  Defect rate and review burden, not elapsed time alone.
- [ ] **A model is never an authorisation layer**
  Retrieval enforces source permissions before retrieval, not after generation.
- [ ] **Agents are bounded and stoppable**
  Least privilege, default-deny networking, approvals, logging and tested rollback.
- [ ] **Participant-facing output has a publication gate**
  Approved-source retrieval and human publication control, whatever data it used.
- [ ] **A DPIA screen sits at intake**
  Deciding whether a full assessment is needed, before any tool is selected.
- [ ] **Every proposed use has an intended purpose**
  One line, in the words of the person who wants it. A regulator will read that line.
- [ ] **The review is specified, not assumed**
  What the reviewer sees, how long they have, and what authority they hold to reject.
- [ ] **Review quality is sampled**
  A proportion of accepted outputs is independently checked and reported.
- [ ] **The primary outcome is pre-registered**
  Fixed before anyone is given access to the tool.
- [ ] **Subgroup performance is checked**
  Cohort diversity is not evidence of equitable performance.
- [ ] **Versions are recorded**
  Model, prompt, corpus, tools and policy, so an incident can be investigated.
- [ ] **Literature search is treated as high risk**
  The comparative evidence reports a median of 91 per cent of relevant studies missed.
- [ ] **The device boundary is reviewed by description**
  What the organisation claims the tool does, across every document.
- [ ] **Evaluations repeat after material change**
  A deployment evaluated once was evaluated against a system that no longer exists.

END MATTER

Glossary

Terms used in a specific sense in this paper.

Recall cost

The expense and incompleteness of withdrawing an output after it has left the control of the person who produced it. The primary tiering variable proposed here.

Safe evidence

A proposed sixth Safe, covering whether a deployed tool works, which the existing five do not address.

Pre-registration

Fixing the primary outcome and the analysis before results are seen. The cheapest available protection against post-hoc framing.

Special category data

Under UK data protection law, data including health and genetic information, subject to additional conditions for processing.

Deskilling

Loss of unaided competence following sustained exposure to assistance. Observed in the colonoscopy evidence in Chapter 8.

Retrieval-augmented generation

Grounding a model's output in retrieved source documents. Reduces fabrication without eliminating it.

Evaluation capability

What an organisation can demonstrate about work it has already changed. Graded on a five-level ladder in Chapter 13.

Instrumented baseline

A record of what a named workflow costs today in time, rework and escalation, derived from data.

Intended purpose

In medical device regulation, what the manufacturer claims a product does, derived from instructions for use, labelling and promotional material.

Automation bias

The tendency of a reviewer to accept an automated output they would have questioned had a person produced it.

Excessive agency

In the OWASP taxonomy, the risk arising when a system is granted more autonomy, permission or functionality than its task requires.

Five Safes

A long-standing framework governing safe people, projects, data, settings and outputs in shared research environments.

END MATTER

Method and evidence grading

How the sources in this paper were selected, checked and graded, and what the reader should distrust.

Sources were gathered from the peer-reviewed literature, primary legislation, statutory instruments and published regulator guidance, with priority given to systematic reviews, randomised evidence and methodological criticism over single primary studies and over vendor material, of which none is cited. Every reference in the list that follows was taken from the publishing journal, publisher or issuing body itself, and the evidence position is the position as at 31 July 2026.

What could not be verified

Four things were searched for and not found, and the pattern of absence shapes this paper more than any single finding. No peer-reviewed study of comparable rigour testing retrieval-augmented generation against hallucination in a medical or research setting, which is why the retrieval argument in Chapter 7 rests on legal-domain evidence and transfers by argument. No rigorous study of prompt injection in a production healthcare deployment, as distinct from benchmark environments. No transferable error rate for large language models on clinical or research tasks, which is the reason Chapter 13 insists on local measurement. And no published AI-specific policy for use inside a trusted research environment that the author could locate, which is an absence of evidence from a reasonable search, and no more than that.

What to distrust in this paper

Three things. The evaluation capability ladder in Chapter 13, the recall cost framework in Chapter 2 and the sixth Safe proposed in Chapter 12 are the author's own frameworks, offered and not yet validated, and each is tagged as such wherever it appears. None has been tested outside the engagements it came from. The three-condition mechanism in Chapter 7 is an inference drawn across studies that were not designed to test it. And the reading of intended purpose in Chapter 10 is the author's summary of regulator guidance and not a ruling; any question of a particular product's status is for the Agency and for qualified advisers.

Limits of the evidence base

The literature is dominated by software engineering and clinical documentation, so its transfer to research operations rests on argument, and carries whatever weight the reader thinks that argument deserves. It also moves faster than peer review, which is why one study cited here remains a preprint and is marked as such, and the systems it evaluates are revised continuously, so a finding about a model in early 2025 may

not describe the same product name a year later. The practical response to all three is local measurement.

Conflict of interest

The author advises organisations on the deployment of artificial intelligence, including in regulated and research settings, and would be a plausible supplier of several things this paper recommends. It also argues against several product categories that would be commercially straightforward to sell, including literature search assistance and participant-facing interpretation. Readers should weigh the argument accordingly.

END MATTER

About the author

Paul Forrest is the founder of Ecaveo and works as a fractional head of AI with private equity and venture capital portfolio businesses on the deployment of artificial intelligence in regulated settings. He began in business process reengineering in the 1990s and moved into natural language and machine learning work from 2014, building small, domain-specific models in preference to general-purpose ones.

He is the author of *Tales from AI Development Hell* and *AI Made the Decision: Who Answers When Nobody Owns the Outcome?*, and holds an adjunct professorship teaching digital acumen and digital entrepreneurship. A third book on where advantage accumulates in the use of artificial intelligence is in preparation.

This is the first paper in the Ecaveo AI adoption series. Comments and corrections are welcome.

END MATTER

References

37 sources, cited as at 31 July 2026.

Evidence on capability, productivity and error

1. **Cui, K. Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S., & Salz, T.** (2026). The effects of generative AI on high-skilled work. Evidence from three field experiments with software developers. *Management Science*, published online 27 February 2026. doi:10.1287/mnsc.2025.00535
2. **Becker, J., Rush, N., Barnes, E., & Rein, D.** (2025). Measuring the impact of early-2025 AI on experienced open-source developer productivity. arXiv:2507.09089 v2, 25 July 2025. Preprint, not peer reviewed.
3. **METR** (2026). Measuring the impact of late-2025 AI on experienced open-source developer productivity. 24 February 2026. Preprint, not peer reviewed. The authors describe the data as only very weak evidence.
4. **Chen, S. F., Alyakin, A., Seas, A., Yang, E., Choi, J. J., Lee, J. V., et al.** (2026). LLM-assisted systematic review of large language models in clinical medicine. *Nature Medicine*, 32(3), 1152-1159. doi:10.1038/s41591-026-04229-5
5. **Busch, F., Hoffmann, L., Rueger, C., van Dijk, E. H. C., Kader, R., Ortiz-Prado, E., et al.** (2025). Current applications and challenges in large language models for patient care. A systematic review. *Communications Medicine*, 5, 26. doi:10.1038/s43856-024-00717-2
6. **Clark, J., Barton, B., Albarqouni, L., Byambasuren, O., Jowsey, T., Keogh, J., et al.** (2025). Generative artificial intelligence use in evidence synthesis. A systematic review. *Research Synthesis Methods*, 16(4), 601-619. doi:10.1017/rsm.2025.16
7. **Olson, K. D., Meeker, D., Troup, M., Barker, T. D., Nguyen, V. H., Manders, J. B., et al.** (2025). Use of ambient AI scribes to reduce administrative burden and professional burnout. *JAMA Network Open*, 8(10), e2534976. doi:10.1001/jamanetworkopen.2025.34976
8. **Holmgren, A. J., Fenton, C. L., Thombley, R., Soleimani, H., Croci, R., DeMasi, O., et al.** (2026). Ambient artificial intelligence scribes and physician financial productivity. *JAMA Network Open*, 9(1), e2553233. doi:10.1001/jamanetworkopen.2025.53233
9. **Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., et al.** (2024). Large language model influence on diagnostic reasoning. A randomized clinical trial. *JAMA Network Open*, 7(10), e2440969. doi:10.1001/jamanetworkopen.2024.40969
10. **Qazi, I. A., Ali, A., Khawaja, A. U., Akhtar, M. J., Sheikh, A. Z., & Alizai, M. H.** (2026). Automation bias in large language model-assisted diagnostic reasoning among physicians trained in AI literacy. A randomized clinical trial. *NEJM AI*, 3(5). doi:10.1056/AIoa2501001
11. **Budzyn, K., Romanczyk, M., Kitala, D., Kolodziej, P., Bugajski, M., Adami, H. O., et al.** (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy. A multicentre, observational study. *The Lancet Gastroenterology and Hepatology*, 10(10), 896-903. doi:10.1016/S2468-1253(25)00133-5
12. **Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E.** (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*, 22(2), 216-242. doi:10.1111/jels.12413. Cited as analogical evidence from a non-medical domain.
13. **Jayasinghe, D., Eshetie, S., Beckmann, K., Benyamin, B., & Lee, S. H.** (2024). Advancements and limitations in polygenic risk score methods for genomic prediction. A scoping review. *Human Genetics*, 143(12), 1401-1431. doi:10.1007/s00439-024-02716-8
14. **Kullo, I. J., Conomos, M. P., Nelson, S. C., Adebamowo, S. N., Choudhury, A., Conti, D., et al.** (2024). The PRIMED Consortium. Reducing disparities in polygenic risk assessment. *American Journal of Human Genetics*, 111(12), 2594-2606. doi:10.1016/j.ajhg.2024.10.010. A consortium description rather than an empirical study.

Legislation and data protection

15. **Data (Use and Access) Act 2025** (c. 18). Royal Assent 19 June 2025. Section 80 replaces Article 22 of the UK GDPR with Articles 22A to 22D. Sections 67 and 68 concern scientific research and consent.
16. **The Data (Use and Access) Act 2025 (Commencement No. 6 and Transitional and Saving Provisions) Regulations 2026**, S.I. 2026/82. Made 29 January 2026. Sections 67, 68, 80 and 86 in force 5 February 2026.
17. **The Data Protection Act 2018 (Code of Practice on Artificial Intelligence and Automated Decision-Making) Regulations 2026**, S.I. 2026/425. Made 16 April 2026, in force 12 May 2026. The code itself was unpublished at the date of this paper.
18. **Information Commissioner's Office** (2023). *Guidance on AI and data protection.* Last updated 15 March 2023. The page carries a notice that the guidance is under review following the Data (Use and Access) Act.
19. **Information Commissioner's Office** (2024). *Data protection audit framework. Artificial intelligence toolkit, contracts and third parties.* Last updated 20 November 2024.
20. **Information Commissioner's Office** (2026). *Data (Use and Access) Act 2025.* Guidance hub, consulted July 2026.

Medical devices, clinical safety and research governance

21. **Medicines and Healthcare products Regulatory Agency** (2026). *Ambient voice technology-enabled products.* Published 29 July 2026. Cited for intended purpose, disclaimers and hallucination by design.
22. **Medicines and Healthcare products Regulatory Agency** (2025). *Medical devices. Software and artificial intelligence (AI).* Published 6 April 2023, last updated 3 February 2025.
23. **Medicines and Healthcare products Regulatory Agency** (2026). *AI Airlock. The regulatory sandbox for AI as a medical device.* Phase 2 programme report, 27 July 2026. Phase 3 in design at the date of this paper.
24. **The Medical Devices (Post-market Surveillance Requirements) (Amendment) (Great Britain) Regulations 2024**, S.I. 2024/1368. Made 16 December 2024, in force 16 June 2025.
25. **NHS England** (2026). *National review of clinical risk management standards DCB0129 and DCB0160. Supporting information.* 29 June 2026. Consultation open to 11 September 2026.
26. **NHS England** (2026). *Guidance on the use of AI-enabled ambient scribing products in health and care settings.* Second iteration, last updated 29 July 2026.
27. **Health Research Authority** (2026). *Plan to enable safe AI-powered innovation in health and social care research.* Published May 2026. A two-year, four-nations plan.
28. **Health Research Authority** (2025). *National Commission on the Regulation of AI in Healthcare.* Launched 26 September 2025, chaired by Professor Alastair Denniston.
29. **Health Research Authority** (2025). *UK Policy Framework for Health and Social Care Research.* Last updated 10 January 2025.

Security, assurance and policy

30. **National Cyber Security Centre** (2023). *Guidelines for secure AI system development.* 27 November 2023. Issued jointly with international partners.
31. **National Institute of Standards and Technology** (2024). *Artificial intelligence risk management framework. Generative artificial intelligence profile*, NIST AI 600-1. 26 July 2024. doi:10.6028/NIST.AI.600-1
32. **OWASP GenAI Security Project** (2025). *Top 10 for LLM applications*, 2025 edition. Cited for the ten categories listed in Chapter 11.
33. **Government Digital Service** (2025). *Artificial intelligence playbook for the UK government.* Published 10 February 2025.
34. **Department for Science, Innovation and Technology** (2025). *AI opportunities action plan.* Published 13 January 2025.
35. **Department for Science, Innovation and Technology** (2024). *Introduction to AI assurance.* Published 12 February 2024.
36. **House of Lords Library** (2026). *AI regulation in the UK. Debate on the need for cross-sector legislation.* 20 May 2026.
37. **European Commission** (2026). *Regulatory framework for artificial intelligence.* Consulted July 2026, including the staged application timetable and the 2026 deferrals.

WHERE THIS GOES NEXT

An AI adoption design workshop

The argument in this paper is testable, and the cheapest way to test it is to instrument one workflow for one team for one quarter, then ask in three months what became answerable that was not answerable before. That costs nothing but a habit, and in my experience it changes more than any purchase.

Ecaveo runs design workshops with research programmes, health data infrastructures, funders and technology partners. The output is a measurement standard the organisation owns, an intended purpose statement for every tool in use, a risk tier for each of them and a first evaluation protocol. It is not a product demonstration.

What a workshop involves

- Two sessions with the people who actually do the work, not only with the people who buy the software.
- A measurement standard covering time, rework, escalation and error for the workflows in scope.
- An intended purpose statement for every AI feature already present in the estate, including the ones nobody procured.
- A risk tier for each, ordered by recall cost, with the absolute boundaries written down.
- One pre-registered evaluation protocol, ready to run, with its stop conditions agreed in advance.