Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Do Documented Risk Controls Hold? A Pre-Registered Method for Verifying Control Performance and Hand-Off Responsibility in Multi-Supplier Clinical AI Workflows

Florian Odi Stummer

florian.stummer@uk-halle.de

Martin Luther University Halle-Wittenberg

Thomas Frese

Martin Luther University Halle-Wittenberg

Research Article

Keywords: Clinical risk management, Artificial intelligence, Primary care, Risk controls, Human oversight, Hazard log

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11325071/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. F.O.S. holds equity in a company that develops governance services for clinical AI. The company did not fund this work and had no role in its design, analysis or reporting. No supplier or product named or implied in the worked example is real. T.F. declares no competing interests.

Do Documented Risk Controls Hold? A Pre-Registered Method for Verifying Control Performance and Hand-Off Responsibility in Multi-Supplier Clinical AI Workflows

Florian Odi Stummer[1*] and Thomas Frese[1]

[1*]Institute of General Practice and Family Medicine, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany.

*Corresponding author(s). E-mail(s): florian.stummer@uk-halle.de;

Abstract

Primary care practices increasingly chain AI tools from several suppliers, such as triage, ambient scribes, automated filing and coding, and task agents. Governance work on clinical AI has so far taken a single system, or a system together with its human reviewer, as the unit of analysis. The risk in a multi-supplier workflow, however, sits in the controls that each supplier documents and in the hand-offs that no supplier owns. The UK National Commission into the Regulation of AI in Healthcare (September 2026) asks manufacturers to specify the controls needed for safe deployment and contracts to allocate responsibility for delivering them; the government accepted these recommendations in October 2026. This paper presents a pre-registered method for checking both. Each documented control is classified against its documented effectiveness and against a severity-based floor using Wilson intervals on walkthrough counts from synthetic test cases, with exact power and the walkthrough sample size required. Parameter uncertainty is propagated through the workflow in closed form, separating escape to the record from undetected escape, and every gap is mapped to a draft hazard-log entry under the English clinical risk management standards DCB0129 and DCB0160. In an invented six-step workflow with six controls and ten fault types, three controls fell below their documented effectiveness, and no control could be shown to meet a floor of 0.80 with 40 walkthrough cases; about 100 cases per control would be needed. Three of five hand-offs had no owner, one fault type had no control, and two fault types were covered only by a retrospective audit. Full correlation between a scribe confidence flag and clinician review removed the

benefit of the second barrier. The method turns control performance and hand-off responsibility into checkable statements; all inputs of the worked example are invented.

Keywords: Clinical risk management, Artificial intelligence, Primary care, Risk controls, Human oversight, Hazard log

1 Introduction

A single primary care consultation can now pass through several AI-enabled tools from different suppliers. An online triage tool routes the request, an ambient scribe drafts the consultation note, robotic process automation files incoming documents and assigns codes, and an agent sets recalls and tasks. Each supplier documents its own risk controls, and the practice adds its own procedures. Safety then depends on two things that no single supplier owns: whether the documented controls catch the faults they are meant to catch in the workflow as it is actually lived, and whether someone is responsible for what happens at each hand-off between tools. The layered-defence view of safety describes this situation: each control is a barrier with holes, and harm occurs when holes in successive barriers line up [1]. In a multi-supplier workflow the barriers belong to different organisations, are documented in different risk management files, and are rarely tested together.

The National Commission into the Regulation of AI in Healthcare addressed this situation in its recommendations of September 2026 [2]. Recommendation 26 asks the Medicines and Healthcare products Regulatory Agency (MHRA) to require manufacturers to specify, in pre-market submissions and risk management files, the operational conditions needed for safe deployment, including the controls that must exist in the user environment. Recommendation 28 states that the contract between manufacturer and healthcare provider should explicitly allocate responsibility for delivering required risk controls, and Recommendation 24 asks that the distribution of responsibility at each stage of the product lifecycle be clear. In its response of 6 October 2026, the government accepted these recommendations: the MHRA will develop guidance on operational conditions for safe use that includes illustrative examples of best-practice risk controls, and the Department of Health and Social Care will review how responsibility can be allocated more clearly between manufacturers and providers, with further details expected by spring 2027 [3]. In England, clinical risk management for health IT follows DCB0129 for manufacturers and DCB0160 for deploying organisations [4, 5], both resting on the hazard concepts of ISO 14971 [6]. Neither standard prescribes how to verify control performance across a chain of suppliers.

Recent work in this journal has advanced the governance of individual clinical AI systems across their lifecycle. It has proposed a minimum-viable governance protocol and algorithmovigilance [7], a decision-oriented approach to post-deployment surveillance within institutions [8], safety envelopes with continuous assurance for continuously updated products [9], lifecycle assurance beyond regulatory approval [10], a

minimum safety case for record-connected consumer health assistants [11], and governance of clinical readiness claims derived from benchmark results [12]. A second line of work questions human review as a safeguard. Clinicians may remain normatively responsible for AI-assisted decisions without being cognitively able to review their basis [13], and AI generators and their human reviewers should be tested together rather than separately [14]. For ambient scribes, validation practice remains heterogeneous and relies mostly on simulated or retrospective data [15, 16], and the clinical note itself is a weak reference standard for what was said [17]. Outside this journal, the medical algorithmic audit [18], acceptance testing and quality control for AI tools [19], and risk-driven quality management of a clinical process that includes a machine learning system [20] address related parts of the problem, and a recent scoping review of reviews found that governance recommendations have advanced faster than evidence on how they are put into operation [21].

Common to this work is the unit of analysis: one AI system, one process within one organisation, or one system together with its reviewer. Three questions fall between these contributions. The first is whether the risk controls that each supplier documents perform at the level the safety case assumes when they are operated in the lived workflow. The second is who owns each hand-off between suppliers, where faults originate that no single product's risk management file covers. The third is how much walkthrough evidence a provider needs before it can state that a control holds. This paper addresses these questions with a pre-registered verification method whose output is a draft hazard log in the format a Clinical Safety Officer works with. The contribution is methodological. The method is demonstrated on a worked example whose inputs are invented, so that its behaviour, including where it cannot reach a conclusion, is visible. The example is not evidence about any product, supplier or provider.

2 Methods

2.1 Inputs

The method needs three inputs, all taken from documents and from a walkthrough in the provider's workflow. The first is the workflow itself: an ordered sequence of steps and hand-offs, in which each step names its supplier and each hand-off names its source step, its target step and, where one is assigned, its owner. The second is a list of fault types, each with a known point of origin at a step or a hand-off. The third is the set of documented controls. For each control the method records its location, its owner and its type, where a preventive control acts before information enters the patient record and a retrospective control acts afterwards. It also records the fault types the control claims to catch, its documented effectiveness as stated in the risk management file, the instructions for use or a standard operating procedure, and walkthrough counts: the number of injected synthetic test cases the control caught out of the number presented, and the number of false alarms on fault-free cases. Because each injected fault has a known origin and content, the injected case serves as the reference standard, which avoids relying on the clinical note as ground truth [17]. Rates are defined per case passing through the workflow, so for a sampling control such as a monthly audit they already include the sampling.

2.2 Primary endpoint: per-control classification

Each control receives two pre-specified classifications based on the 95% Wilson interval of its walkthrough capture rate [22, 23]. Against its documented effectiveness, a control is classified as below documented when the upper bound of the interval lies below the documented value, as above documented when the lower bound lies above it, and as consistent otherwise. Against a floor, a control meets the floor when the lower bound reaches it, falls below the floor when the upper bound lies below it, and otherwise has insufficient evidence. The floor is set per control by the Clinical Safety Officer according to the severity of the hazard the control mitigates. A control whose documented effectiveness already lies below its floor is additionally flagged as designed below the floor.

The primary hypothesis is that at least one control does not hold in the lived workflow. It is supported if at least one control is classified as below documented or below the floor. No confidence interval is computed over the set of controls, because the controls of one workflow are not a sample from a population. To make the strength of the evidence visible, the method also reports for each control the posterior probability that its capture reaches the floor, the exact binomial power to show that the floor is met at the walkthrough sample size if the control works exactly at its documented effectiveness, and the sample size needed for 80% power under the same assumption.

2.3 Escape through the workflow

For each fault type the method computes the probability that the fault escapes. A control covers a fault only if it sits at or after the fault's origin and claims that fault type. Escape to the record is the probability that no preventive control catches the fault before it enters the record; undetected escape is the probability that no control catches it at all, retrospective controls included. Each control's capture probability has a Beta(k+1, n-k+1) posterior, where k of n walkthrough cases were caught. The parameter of each control is drawn once per draw, using 4,000 draws in the confirmatory run, and shared across all fault types and scenarios. Within a draw, escape is computed in closed form as the product of (1 - capture) over the covering controls. Results are reported as the posterior mean with a 95% interval from the 2.5th and 97.5th percentiles across draws, and escape is also computed from documented effectiveness as a comparison scenario.

2.4 Structural checks and secondary endpoints

Some gaps can be read from the workflow map without any walkthrough. The method lists hand-offs without an owner, hand-offs with a named owner but no operating control, fault types without any control, and fault types without a preventive control. It also compares escape for faults that originate at hand-offs with escape for faults that originate inside a tool, with and without the fault types that have no control; because each group contains few fault types and controls are placed by design, this comparison is descriptive. False-alarm burden is reported as alarm events per 1,000 fault-free cases and as fault-free cases with at least one alarm per 1,000.

2.5 Sensitivity scenarios

The base case assumes that controls act independently. Automation bias, in which a human reviewer defers to an automated output [24], and alert fatigue, in which frequent false alarms reduce attention to all alarms [25], can break this assumption, and they are the mechanisms behind the call to test generators and reviewers together [14]. The method therefore runs three families of scenarios, each varying a single parameter. In the first, capture is multiplied by 0.9 or 0.75. In the second, a share ρ of cases (0.5 or 1.0) is decided by one shared draw for all controls on the same fault, so that escape in those cases equals one minus the highest capture. In the third, each control's capture is multiplied by one minus a penalty times its own false-alarm rate, with penalties of 0.5 and 1.0. These scenarios rest on assumed parameters; they show how far the results depend on the independence assumption and do not estimate either mechanism.

2.6 Output for clinical risk management

Every gap becomes a draft hazard-log entry with fields for cause, hazard, hazardous situation, harm, severity, likelihood proxy, existing controls, residual risk, log, proposed owner and recommendation alignment. Gaps on vendor-owned controls are directed to the manufacturer's DCB0129 log, with the reliance recorded in the deploying organisation's DCB0160 log; practice-owned controls and unowned hand-offs go to the DCB0160 log. The hazard statement follows from the classification above and distinguishes performance below documented effectiveness, documented effectiveness below the floor, performance below the floor, and insufficient evidence. Hazardous situation, harm, severity and residual risk are left empty, because a registered Clinical Safety Officer must rate and sign off each entry. Recommendation numbers are recorded as alignment with the National Commission report [2], not as obligations. For hand-offs between two suppliers the entry notes that the contractual allocation in Recommendation 28 is bilateral, so such a hand-off needs an interface agreement.

2.7 Pre-registration, quality control and use of language models

The protocol, configuration and analysis script were frozen by a SHA-256 hash lock before the confirmatory run. A run refuses to start if any locked file has changed unless it is labelled as a pilot, and pilot runs use a separate seed. Random numbers come from a separate seeded stream for each control and purpose, so that changing one input does not shift unrelated results. Every text artefact was checked by an external rule-based gate before and after each edit. The first locked version of the analysis was reviewed with three independent methods: analytic recomputation, an audit of code and pre-registration, and a review of construct validity. The review confirmed the arithmetic but found that the original primary endpoint was largely determined by the walkthrough sample size, that per-case parameter draws reduced the reported intervals to Monte Carlo noise, that the hazard text was not derived from the data, and that one hypothesis had no comparison group. The method reported here addresses these findings; the changes were logged in a dated deviation record before re-locking, and the first confirmatory run is retained.

A large language model (Claude Opus 5.5, Anthropic; accessed through the Claude Code command-line agent, 6 to 7 October 2026, default settings) was used to write and revise the analysis code, to carry out the three-method review as three separately instructed agents, to retrieve and check reference metadata, and to draft the manuscript. Prompts were free-text task instructions; they are available from the corresponding author on request. F.O.S. specified the study, checked the code and outputs, verified every reported number against the locked results file, and checked every reference against PubMed or Crossref. Both authors reviewed the manuscript and take responsibility for its content.

2.8 Worked example and governance

The worked example uses invented inputs, shown in Fig. 1 and Table 1. Six steps from four suppliers and the practice are joined by five hand-offs, ten fault types originate at these steps and hand-offs, and six controls each have 40 walkthrough cases and 40 fault-free cases. The floor is set to 0.80 for every control as a placeholder for floors that a Clinical Safety Officer would set. The example assumes that the clinician reviewing the scribe note can detect lost triage context, which requires the triage output to be displayed next to the note. The example involves no patients, no personal data and no human participants. Before walkthroughs take place in a provider's workflow, that stage should be classified with the Health Research Authority decision tool [26] and registered locally as a service evaluation or audit. The method is not a medical device and does not give clinical advice.

3 Results

All numbers come from the confirmatory run of the locked analysis with 4,000 parameter draws and describe invented inputs.

3.1 Per-control classification

The primary hypothesis was supported. Against documented effectiveness, three controls fell below (controls 1, 4 and 6), three were consistent (controls 2, 3 and 5) and none was above. Against the floor of 0.80, no control met it, four had insufficient evidence and two were below it (controls 2 and 4); for both of these the documented effectiveness itself lay below the floor (Table 2).

The power column shows the main limitation of a 40-case walkthrough. A control working exactly at a documented effectiveness of 0.90 would be shown to meet the floor of 0.80 with probability 0.423, and about 100 cases would be needed for 80% power. Control 5, whose walkthrough capture of 0.825 exceeded its documented effectiveness of 0.80, still could not be shown to meet the floor, because its documented value equals the floor.

3.2 Escape through the workflow

Fig. 2 and Table 3 show escape per fault type. The intervals reflect the uncertainty from 40 walkthrough cases per control and are wide. The three fault types covered

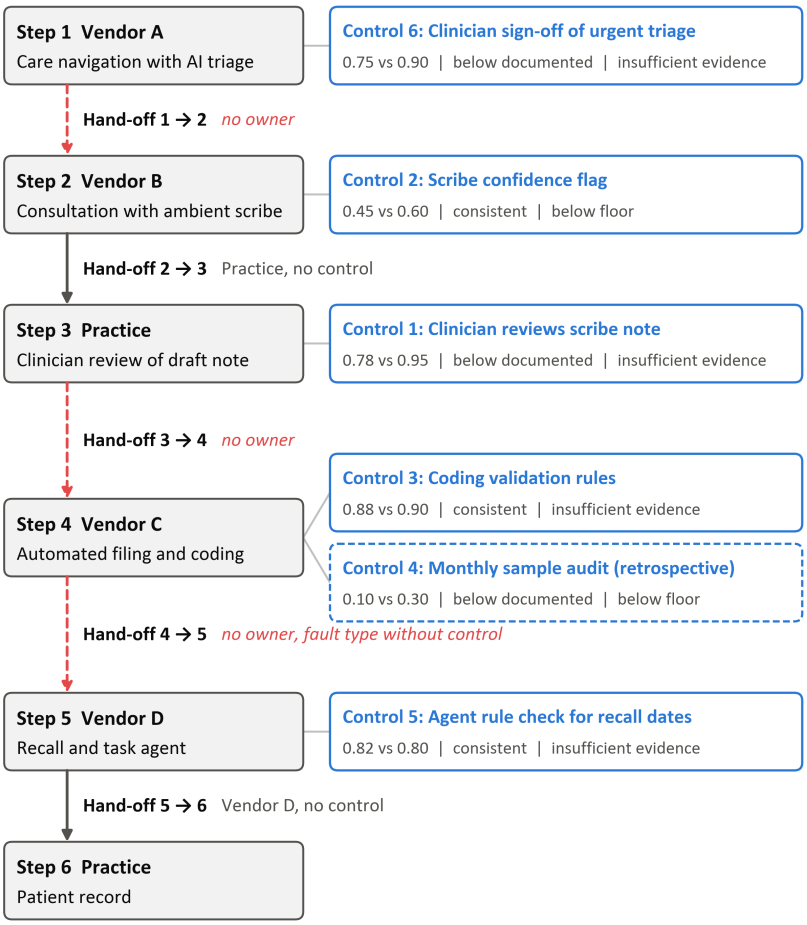

Fig. 1 Worked example workflow (all inputs invented). Grey boxes are workflow steps with their supplier; arrows are hand-offs, dashed and red where no owner is assigned. Blue boxes are the documented controls at their location, with walkthrough capture against documented effectiveness and the two classifications of each control. The dashed blue box marks the retrospective control.

only by control 1 share one parameter and therefore have identical estimates. For most fault types the walkthrough-based estimate lies well above the value implied by documented effectiveness; for a recall that is not set, the two coincide because the agent rule check performed as documented. Two fault types, a structured field lost at filing and an action in a letter filed without a task, were covered only by the retrospective monthly audit and therefore always reached the record. A third, a task duplicated or lost between the automation and the agent, had no control at all.

Table 1 Worked example: workflow, controls and walkthrough counts (all invented)

Step or hand-off	Supplier	Owner of hand-off
Step 1: Care navigation with AI triage	Vendor A	
Hand-off 1 → 2		none
Step 2: Consultation with ambient scribe	Vendor B	
Hand-off 2 → 3		Practice
Step 3: Clinician review of draft note	Practice	
Hand-off 3 → 4		none
Step 4: Automated filing and coding	Vendor C	
Hand-off 4 → 5		none
Step 5: Recall and task agent	Vendor D	
Hand-off 5 → 6		Vendor D
Step 6: Patient record	Practice	

Control	Location	Owner	Type	Documented	Caught / presented	False alarms / clean
Control 1: Clinician reviews scribe note before saving	Step 3	Practice	preventive	0.95	31 / 40	2 / 40
Control 2: Scribe confidence flag on uncertain content	Step 2	Vendor B	preventive	0.60	18 / 40	6 / 40
Control 3: Coding validation rules in automation	Step 4	Vendor C	preventive	0.90	35 / 40	1 / 40
Control 4: Monthly sample audit of filed documents	Step 4	Practice	retrospective	0.30	4 / 40	0 / 40
Control 5: Agent rule check for recall dates	Step 5	Vendor D	preventive	0.80	33 / 40	2 / 40
Control 6: Clinician sign-off of urgent triage outcomes	Step 1	Practice	preventive	0.90	30 / 40	3 / 40

3.3 Hand-offs, false alarms and sensitivity

Three of the five hand-offs had no owner: from triage to scribe (hand-off 1 → 2), from clinician review to automated filing (hand-off 3 → 4), and from automated filing to the agent (hand-off 4 → 5), the last of which lies between two suppliers. Two further hand-offs had a named owner but no operating control (hand-offs 2 → 3 and 5 → 6). Because all five hand-offs crossed supplier boundaries, ownership could not be compared between hand-offs and steps within one supplier. On average, undetected escape was 0.411 higher for faults originating at hand-offs than for faults originating inside a tool (95% interval 0.376 to 0.442); without the fault type that had no control, the difference was 0.264 (0.215 to 0.307). For escape to the record the difference was 0.431 (0.399 to 0.460), and it fell to 0.037 (−0.037 to 0.121) once fault types without a

Table 2 Per-control classification (walkthrough capture with 95% Wilson interval)

Control	Documented	Walkthrough capture (95% CI)	Against documented	Against floor 0.80	Probability capture ≥ floor	Power at documented, n = 40	n for 80% power
Control 1	0.95	0.775 (0.625 to 0.877)	below	insufficient evidence	0.303	0.862	40
Control 2	0.60	0.450 (0.307 to 0.602)	consistent	below	0.000	0.000	not reachable
Control 3	0.90	0.875 (0.739 to 0.945)	consistent	insufficient evidence	0.853	0.423	100
Control 4	0.30	0.100 (0.040 to 0.231)	below	below	0.000	0.000	not reachable
Control 5	0.80	0.825 (0.680 to 0.913)	consistent	insufficient evidence	0.603	0.028	not reachable
Control 6	0.90	0.750 (0.598 to 0.858)	below	insufficient evidence	0.181	0.423	100

"Not reachable": documented effectiveness at or below the floor, so the floor cannot be shown to be met even with unlimited walkthrough cases.

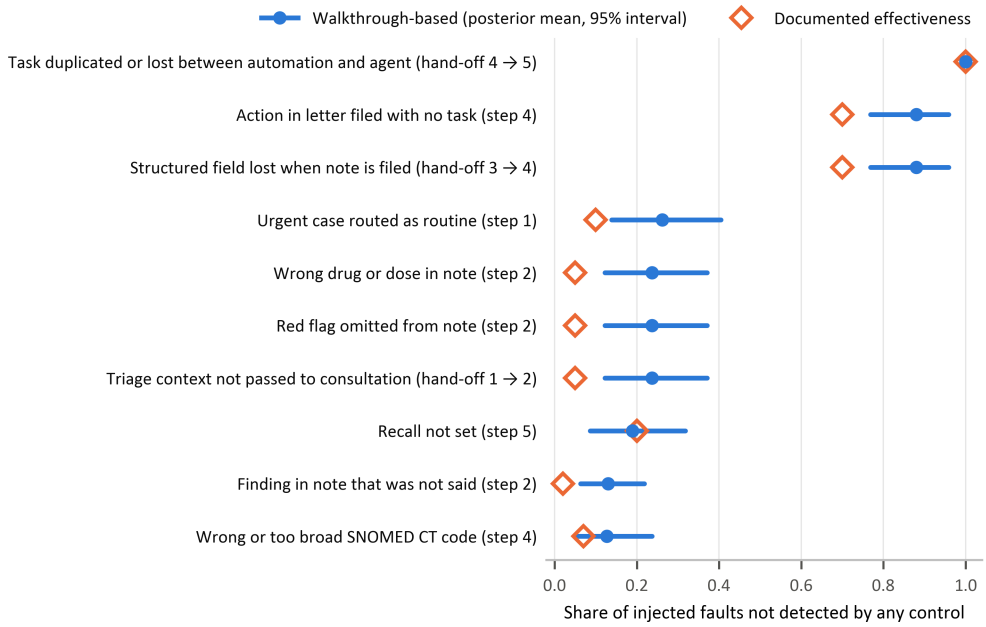

Fig. 2 Share of injected faults not detected by any control, per fault type and its origin (all inputs invented). Blue: estimate from walkthrough counts (posterior mean with 95% interval from parameter uncertainty). Orange: value implied by documented effectiveness.

Table 3 Escape per fault type (lived: posterior mean with 95% interval; documented: from documented effectiveness)

Fault type (origin)	Controls	Undetected, documented	Undetected, lived	To record, documented	To record, lived
Urgent case routed as routine (step 1)	Control 6	0.100	0.262 (0.139 to 0.405)	0.100	0.262 (0.139 to 0.405)
Triage context not passed to consultation (hand-off 1 → 2)	Control 1	0.050	0.237 (0.123 to 0.371)	0.050	0.237 (0.123 to 0.371)
Red flag omitted from note (step 2)	Control 1	0.050	0.237 (0.123 to 0.371)	0.050	0.237 (0.123 to 0.371)
Wrong drug or dose in note (step 2)	Control 1	0.050	0.237 (0.123 to 0.371)	0.050	0.237 (0.123 to 0.371)
Finding in note that was not said (step 2)	Controls 2 and 1	0.020	0.130 (0.063 to 0.218)	0.020	0.130 (0.063 to 0.218)
Structured field lost when note is filed (hand-off 3 → 4)	Control 4	0.700	0.880 (0.769 to 0.959)	1.000	1.000
Wrong or too broad SNOMED CT code (step 4)	Controls 3 and 4	0.070	0.127 (0.049 to 0.237)	0.100	0.144 (0.056 to 0.268)
Action in letter filed with no task (step 4)	Control 4	0.700	0.880 (0.769 to 0.959)	1.000	1.000
Task duplicated or lost between automation and agent (hand-off 4 → 5)	none	1.000	1.000	1.000	1.000
Recall not set (step 5)	Control 5	0.200	0.189 (0.087 to 0.318)	0.200	0.189 (0.087 to 0.318)

preventive control were set aside, so in this example the difference is largely explained by where the controls were placed.

Across all six controls, fault-free cases produced 477.0 alarm events per 1,000 (95% interval 302.0 to 687.3), and 396.3 per 1,000 fault-free cases (268.7 to 530.0) raised at least one alarm. Scaling capture by 0.75 raised undetected escape for an urgent case routed as routine from 0.262 to 0.446 (0.354 to 0.554). Full correlation between the scribe flag and clinician review ($\rho = 1.0$) raised escape for a finding that was not said from 0.130 to 0.237 (0.123 to 0.371), which equals the escape under clinician review alone, so that the second barrier added nothing. An alert-fatigue penalty of 1.0 raised escape for a red flag omitted from the note from 0.237 to 0.291 (0.170 to 0.432).

The run produced 14 draft hazard-log entries. Five concern hand-offs, three of them without an owner and two with an owner but no control; six concern the controls; one concerns the fault type without a control; and two concern the fault types covered only retrospectively. Each entry names its log (DCB0129 or DCB0160) and a proposed owner, while hazardous situation, harm, severity and residual risk are left for the Clinical Safety Officer.

4 Discussion

The governance protocols, surveillance models and assurance frameworks published in this journal [7–12] take a single system as the object that is governed over time. The method presented here takes the multi-supplier workflow as the object and the

documented risk controls as the thing that is verified, at the point where a provider deploys several tools together. It complements post-deployment surveillance [8, 9] by establishing, before and during deployment, whether the controls that the safety case relies on work at the documented level, and it produces the hazard-log entries that a safety case [11] and the DCB0129 and DCB0160 processes [4, 5] require.

The method also gives the debate on human oversight a measurable form. The distinction between normative and cognitive oversight [13] maps onto two separate checks: ownership of a step or hand-off records who is normatively responsible, while walkthrough capture measures whether the responsible reviewer actually catches the injected fault. The call to test generators and reviewers together [14] corresponds to the correlation scenario. When the scribe's confidence flag and the clinician's review fail on the same cases, the second barrier adds nothing, and an independence assumption in the safety case would overstate the protection by a factor of almost two in this example. For ambient scribes in particular, synthetic test cases with injected faults provide a reference standard that does not depend on the clinical note [15–17].

The worked example makes four further mechanisms visible, all of them properties of the method under the stated inputs rather than findings about practice. Insufficient evidence is a result in its own right: with 40 walkthrough cases, a control documented at 0.90 has less than an even chance of being shown to meet a floor of 0.80, so a provider needs about 100 cases per control or must accept that "not shown" is the likely outcome. A documented control can be designed below the level its hazard requires, and the method separates this from underperformance. Retrospective controls do not prevent, because an audit finds a fault only after it has entered the record. Unowned hand-offs and uncovered fault types are structural gaps that need only the workflow map to find. Recommendation 28 addresses the contract between one manufacturer and one provider [2], so a hand-off between two suppliers, such as the one from automated filing to the task agent, needs matching clauses in both contracts or an interface agreement among the provider and both suppliers. The review of responsibility allocation that the government has committed to [3] offers an occasion to address such supplier-to-supplier hand-offs explicitly.

Several limitations apply. All inputs in the worked example are invented, so its numbers describe how the method behaves and not any product, supplier or provider. In the base case the controls act independently, and correlated failure, automation bias and alert fatigue enter only as scenarios with assumed parameters. Detection is treated as correction, without latency. Fault types are not weighted by severity, so escape values are not risk estimates. The workflow is linear, without loops, parallel paths, workarounds or re-introduced faults. The comparison between hand-off and in-tool faults rests on few fault types and on the placement of controls, and the ownership comparison had no comparison group in the example. The mapping of controls to fault types reflects a single rater; the method supports a second independent mapping with Cohen's kappa, which was not run here. Whether automation steps such as filing, coding and task agents are medical devices, and so fall within the scope of the recommendations cited, is not established. A first application needs a provider's own documents and walkthrough counts, floors set for each control by a Clinical Safety

Officer, walkthrough samples sized from the power calculation, a second independent control mapping, and local registration of the walkthrough stage.

5 Conclusions

Documented controls in multi-supplier clinical AI workflows can be verified against their own documentation, against severity-based floors, and for responsibility at every hand-off, using a pre-registered and reproducible procedure. The method moves the unit of governance from the single system to the supplier chain and reports whether a control holds, whether the evidence suffices to say so, and where hand-offs and fault types have no owner or no control.

Declarations

Author Contributions. F.O.S. conceived the study, designed the method, supervised the software, analysed the results, prepared Figures 1 and 2, and drafted the manuscript. T.F. contributed the primary-care governance framing and the interpretation of the results, and critically revised the manuscript. Both authors reviewed and approved the final manuscript and are accountable for the work.

Funding. The authors received no external funding for this work.

Data Availability. All inputs are invented and fully reported in Table 1. The pre-registered protocol, the lock record with file hashes and the deviation record are provided as Supplementary Information. The configuration, analysis script and results of the confirmatory run are available from the corresponding author on reasonable request.

Ethical Approval. Not applicable. The study used invented inputs only and involved no human participants, patients or personal data.

Consent to Participate. Not applicable.

Consent to Publish. Not applicable.

Competing interests. F.O.S. holds equity in a company that develops governance services for clinical AI; the company did not fund this work. T.F. declares no competing interests. No supplier or product named or implied in the worked example is real.

Clinical trial number: not applicable.

References

[1] Reason J (2000) Human error: models and management. BMJ 320:768-770. doi:10.1136/bmj.320.7237.768

[2] Medicines and Healthcare products Regulatory Agency (2026) National Commission into the Regulation of AI in Healthcare: recommendations for a future regulatory framework. GOV.UK, London. https://www.gov.uk/government/pu

blications/national-commission-into-the-regulation-of-ai-in-healthcare-recommendations-for-a-future-regulatory-framework. Accessed 6 October 2026

[3] Medicines and Healthcare products Regulatory Agency (2026) Government response to the National Commission's recommendations on the regulation of AI in healthcare. GOV.UK, London. https://www.gov.uk/government/publications/government-response-to-the-national-commissions-recommendations-on-the-regulation-of-ai-in-healthcare. Accessed 7 October 2026

[4] NHS England. DCB0129: Clinical risk management: its application in the manufacture of health IT systems. https://data.standards.nhs.uk/current-standards/nhsd-clinical-risk-management-its-application-in-the-manufacture-of-health-it-systems. Accessed 6 October 2026

[5] NHS England. DCB0160: Clinical risk management: its application in the deployment and use of health IT systems. https://standards.nhs.uk/published-standards/clinical-risk-management-its-application-in-the-deployment-and-use-of-health-it-systems. Accessed 6 October 2026

[6] International Organization for Standardization (2019) ISO 14971:2019 Medical devices: application of risk management to medical devices. ISO, Geneva

[7] Sirago G, Vinci A, Visci P, Solarino B, Dell'Erba A, Ferorelli D (2026) From black-box clinical AI to algorithmovigilance: a minimum-viable governance protocol for safety, accountability, and prevention of algorithmic harm. J Med Syst 50:52. doi:10.1007/s10916-026-02382-5

[8] Lampignano JAP, Kale A, Kumar S, Shah D, Reddy B, Lowe DJ (2026) Beyond validation: operationalising post-deployment surveillance of AI medical devices in clinical practice. J Med Syst 50:98. doi:10.1007/s10916-026-02426-w

[9] Nandagopal S (2026) Safety envelopes and continuous assurance for updates and postmarket surveillance in digital mental health technologies. J Med Syst 50:58. doi:10.1007/s10916-026-02375-4

[10] Shujaat S (2026) Beyond regulatory approval: lifecycle assurance for clinical artificial intelligence. J Med Syst 50:138. doi:10.1007/s10916-026-02466-2

[11] Bair H (2026) A minimum safety case for record-connected consumer health assistants. J Med Syst 50:70. doi:10.1007/s10916-026-02401-5

[12] Dong Y, Cheng J, Ding C, Lu R (2026) Governing clinical readiness claims derived from medical AI benchmark results. J Med Syst 50:119. doi:10.1007/s10916-026-02445-7

[13] Ver Berne J, Jacobs R (2026) When useful clinical AI exceeds meaningful oversight. J Med Syst 50:121. doi:10.1007/s10916-026-02446-6

[14] Sorin V, Klang E (2026) Clinical AI generators and reviewers must be tested together. J Med Syst 50:125. doi:10.1007/s10916-026-02454-6

[15] Kerimoğlu E, Notermans FV, Silkens MEWM, de Mul M, Ahaus KCTB, van der Boon RMA (2026) Validating digital scribes: a scoping review of evaluation practices and clinical use. J Med Syst 50:62. doi:10.1007/s10916-026-02392-3

[16] Nair R, Hashmi MM, Kassim SS, Singer A (2026) The current state of digital scribes in primary care: a scoping review. J Med Syst 50:2. doi:10.1007/s10916-025-02319-4

[17] Sezgin E (2026) Hospital phone encounters as upstream evidence for AI scribe evaluation. J Med Syst 50:39. doi:10.1007/s10916-026-02366-5

[18] Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L (2022) The medical algorithmic audit. Lancet Digit Health 4:e384-e397. doi:10.1016/S2589-7500(22)00003-6

[19] Mahmood U, Shukla-Dave A, Chan HP, Drukker K, Samala RK, Chen Q, et al. (2024) Artificial intelligence in medicine: mitigating risks and maximizing benefits via quality assurance, quality control, and acceptance testing. BJR Artif Intell 1:ubae003. doi:10.1093/bjrai/ubae003

[20] Mercolli L, Rominger A, Shi K (2024) Towards quality management of artificial intelligence systems for medical applications. Z Med Phys 34:343-352. doi:10.1016/j.zemedi.2024.02.001

[21] El Arab RA, Mustafa MH, Almagharbeh WT, Saleem NH, Al Abdulmohsen S, Boathab R, et al. (2026) Beyond model development in healthcare AI: post-development robustness, post-deployment monitoring, and lifecycle governance. A scoping review of reviews. Healthcare (Basel) 14:1459. doi:10.3390/healthcare14111459

[22] Wilson EB (1927) Probable inference, the law of succession, and statistical inference. J Am Stat Assoc 22:209-212. doi:10.1080/01621459.1927.10502953

[23] Brown LD, Cai TT, DasGupta A (2001) Interval estimation for a binomial proportion. Stat Sci 16:101-133. doi:10.1214/ss/1009213286

[24] Goddard K, Roudsari A, Wyatt JC (2012) Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 19:121-127. doi:10.1136/amiajnl-2011-000089

[25] Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R, et al. (2017) Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak 17:36. doi:10.1186/s12911-017-0430-8

[26] Health Research Authority. Is my study research? Decision tool. https://www.hra-decisiontools.org.uk/research/. Accessed 6 October 2026

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformationStummerFreseJMS.pdf