Supplementary 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, Thomas Frese
Institute of General Practice and Family Medicine
Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

1. Purpose of this file

This file documents the pre-registration of the analysis reported in the manuscript: the record of both hash locks, the protocol in its second locked version, the pilot runs, and the deviation that led from the first to the second locked version. It is provided so that reviewers can check that the protocol was frozen before each confirmatory run and that every change between the two versions was recorded before the second lock. All inputs of the worked example are invented. Labels for steps, hand-offs and controls follow the manuscript (Table 1 and Fig. 1).

2. Lock record

Before each confirmatory run, the protocol, the configuration file and the analysis script were frozen by computing a SHA-256 hash of each file and a combined hash over all three. The analysis refuses to run when any locked file differs from its recorded hash, unless the run is explicitly labelled as a pilot. Table S1 lists both locks. The confirmatory results reported in the manuscript come from the run under the second lock. The run under the first lock is retained unchanged.

Table S1 Lock record

Version	Locked (UTC)	Combined SHA-256 and confirmatory run
First locked version	6 Oct 2026, 16:28:18	`a5d89193b14c6a4e5c05a3425937da63d2a61644bbc8b69da77f0ecdde94cc43`; confirmatory run started 16:28:19
Second locked version	6 Oct 2026, 16:47:36	`db12f5a254a1f108f376f67619b6daaf5a5191530d4a9b8265f2469cad9c5bbc`; confirmatory run started 16:47:37 (reported in the manuscript)

Table S2 File hashes of the second locked version

File	SHA-256
Protocol	`01efb7ffe793899eabb62f38689827aff0c39992587f58b36431e851768dbbda`
Configuration	`b0987d4e6d494dfd61a243a9fef48762dbd4e665a114f176110c48f9224f7412`
Analysis script	`8cc2dc49b6d9b977f2092d240bc6789265ba918bfb8f527de92a08afd92376d8`

The three file hashes were re-checked against the locked files on 7 October 2026 and still match.

3. Protocol (second locked version)

Question

Do the operational controls specified by manufacturers (risk management file, instructions for use) and by the provider (standard operating procedures) catch the faults they are meant to catch in the lived workflow, and is responsibility assigned at every hand-off between tools?

Hypotheses

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 its documented effectiveness or below its floor. The second hypothesis, treated as descriptive, is that faults originating at hand-offs between tools escape more often than faults originating inside a tool; it is reported with and without fault types that have no control, because the groups contain few fault types and controls are placed by design. The third hypothesis, also descriptive, is that hand-offs between suppliers lack an assigned owner; in the worked example all hand-offs cross suppliers, so no comparison group exists and the result is reported as a count.

Design

The method has two stages applied to one workflow at a time. In the document analysis, every documented control receives a location, an owner, a type (preventive, acting before the record, or retrospective, acting after it), the fault types it claims to catch, its documented effectiveness, and walkthrough counts: caught cases out of injected synthetic test cases, and false alarms on fault-free cases. Walkthrough rates are defined per case passing through the workflow, so that for a sampling control they already include the sampling. In the parameter-uncertainty analysis, each control's capture probability is drawn from $\text{Beta}(k+1,\, n-k+1)$ and its false-alarm probability from the corresponding Beta posterior, once per draw and shared across all fault types and scenarios. For each draw, escape per fault type is computed in closed form over the controls located at or after the fault's origin. Documented effectiveness is run as a deterministic comparison scenario. A control covers a fault only if it sits at or after the fault's origin; unknown locations or origins stop the run.

Endpoints

The primary endpoint is a per-control classification based on the 95% Wilson interval of walkthrough capture. Against documented effectiveness, a control is below documented when the upper bound lies below the documented value, above documented when the lower bound lies above it, and consistent otherwise. Against the floor, which the Clinical Safety Officer sets per control according to the severity of the hazard the control mitigates (example default 0.80), a control meets the floor when the lower bound reaches it, is below the floor when the upper bound lies below it, and otherwise has insufficient evidence; a control whose documented effectiveness lies below its floor is flagged as designed below the floor. For each control the method also reports the posterior probability that capture reaches the floor, the exact power to show that the floor is met at the walkthrough sample size if true capture equals documented effectiveness, and the sample size needed for 80% power. Secondary endpoints are escape per fault type, both to the record (preventive controls only) and undetected (all controls), each as posterior mean with 95% parameter-uncertainty interval and as the documented-effectiveness value; the hand-off comparison described above; hand-offs without owner and hand-offs with a named owner but no operating control; fault types without any control and without a preventive control; and the false-alarm burden as alarm events and as fault-free cases with at least one alarm, each per 1,000 fault-free cases and with intervals.

Controls of the analysis and sensitivity scenarios

Fault-free cases provide the false-alarm burden. An optional second, independent mapping of controls to fault types allows Cohen's kappa; it was not run for the worked example. Sensitivity scenarios vary one assumption at a time: capture scaled by 0.9 and 0.75; correlation between controls on the same fault, in which a share ρ of cases (0.5 and 1.0) is decided by one shared draw so that escape equals one minus the highest capture; and an alert-fatigue penalty that multiplies capture by one minus the penalty (0.5 and 1.0) times the control's own false-alarm rate. Each control and purpose uses its own seeded random stream. Pilot runs use a separate pilot seed and fewer draws; the confirmatory run uses the confirmatory seed and 4,000 draws.

Outputs and governance

The outputs are a control table, an escape table, a scenario table and a draft hazard-log table. Each gap carries cause, hazard, hazardous situation, harm, severity, likelihood proxy, existing controls, residual risk, log (DCB0129 for vendor-owned controls, DCB0160 for the deploying organisation), proposed owner and alignment with the recommendations of the National Commission into the Regulation of AI in Healthcare;

hazardous situation, harm, severity and residual risk are left for the Clinical Safety Officer. No patients and no personal data are involved. The current stage is synthetic modelling without human participants; the provider stage is to be classified with the Health Research Authority decision tool and registered locally before walkthroughs. A registered Clinical Safety Officer rates and signs off the hazard-log entries before they enter a DCB0129 or DCB0160 safety case.

Limits stated in the protocol

All inputs of the worked example are assumptions for demonstrating the method; real conclusions require a provider's own documents and walkthrough counts. The workflow is linear, detection is treated as correction without latency, fault types are not severity-weighted, and controls are independent in the base case, with correlation and alert fatigue entering only as scenarios. Whether automation steps such as filing, coding and task agents are medical devices is not established.

4. Pilot runs

The first version was piloted on 6 October 2026 at 16:27 UTC without a lock. The pipeline ran as specified. One wording fix was made before the first lock: the gap text for a control that failed the hold criterion was changed to name the threshold and the lower bound, because a control can fail the threshold while its walkthrough capture exceeds its documented effectiveness. The pilot also showed that with 40 walkthrough cases a control could meet a threshold of 0.80 only with at least 37 of 40 cases caught; this was recorded without changing the protocol. The first pilot used the same seed as the first confirmatory run; this was one of the issues addressed in the second version.

The second version was piloted at 16:47 UTC with the separate pilot seed and 1,000 draws. All endpoints were computed as specified. One presentation fix was made before the second lock: intervals are written as "lower to upper" so that negative bounds read correctly, and per-1,000 false-alarm figures use one decimal. Hypotheses, endpoints and inputs were not changed by either pilot.

5. Deviation record

Deviation 1 (6 October 2026): revision of the first locked version

The first locked version and its confirmatory run were reviewed internally with three independent methods: analytic recomputation, an audit of the code against the protocol and the lock, and a review of construct validity (Methods, manuscript). The review confirmed the arithmetic of the first run and found that its inference design and reporting did not answer the study question. The deviation was recorded before the second lock. The inputs of the worked example were not changed.

The following changes were made. First, uncertainty: the first version drew a new capture probability for every simulated case, which reduced the reported intervals to Monte Carlo noise; the second version draws each control's parameter once per draw, shares it across fault types and scenarios, and computes escape in closed form. Second, the primary endpoint: the first version reported the share of controls whose lower Wilson bound reached 0.80, with an interval over the six controls; the second version classifies each control against its documented effectiveness and against a severity-based floor, reports power and the required sample size, defines a decision rule for the primary hypothesis, and drops the interval over the set of controls. Third, the hazard text is derived from this classification instead of one sentence for every control that missed the threshold. Fourth, the hand-off comparison is reported descriptively with a parameter-uncertainty interval, with and without fault types that have no control, instead of a bootstrap over fault types. Fifth, the ownership hypothesis is restated as a descriptive count, and hand-offs with a named owner but no operating control are listed separately. Sixth, controls are typed as preventive or retrospective (control 4, the monthly sample audit, is retrospective), and escape is reported both to the record and as undetected. Seventh, coverage became position-aware, and unknown control locations or fault origins stop the run. Eighth, each control and purpose received its own seeded random stream, the confirmatory run received a new seed, and pilots use a separate pilot seed, so that the confirmatory run no longer replays the pilot. Ninth, the false-alarm burden is reported as events and as cases with at least one alarm, with intervals, and the rate actually used (the posterior mean) is shown next to the raw rate. Tenth, correlation and alert-fatigue scenarios were added and labelled as assumptions. Eleventh, the hazard-log table received the structure of a clinical risk management log, with the fields that require clinical judgement left for the Clinical Safety Officer. Twelfth, recommendation numbers are stated

as alignment with the National Commission report, not as obligations, and recommendations outside the declared scope were removed from the mapping; vendor-to-vendor hand-offs carry a note that the contractual allocation is bilateral and needs an interface agreement. Thirteenth, the report states on its first page that all inputs are invented, names the assumptions that drive the result and lists its limitations. Fourteenth, the governance wording distinguishes the current synthetic stage from the provider stage.

The expected effect, stated before the second lock, was that the primary result would change from "no control holds" to a per-control classification in which most controls move to insufficient evidence or design below floor, that escape intervals would widen by about an order of magnitude, that the hand-off difference would shrink once the uncovered fault type is shown separately, and that the number of unowned hand-offs and the uncovered fault type would not change. The second confirmatory run matched these expectations; its results are those reported in the manuscript.