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.

Planning Human Review for Large-Scale Medical Device Safety Surveillance: An Operating Map Across 18.2 Million MAUDE Narratives

You-Ra CHO
Chung-Ang University
Sunghee Kim
sung1024@cau.ac.kr

Chung-Ang University

Research Article

Keywords: MAUDE, medical device surveillance, text mining, human-in-the-loop systems, review prioritisation, operating characteristics

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11029425/v1

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

Additional Declarations: No competing interests reported.

Abstract

Post-market medical-device surveillance generates millions of free-text narratives, but limited staffing precludes exhaustive human review. In this retrospective observational study of 18,228,617 narratives from the U.S. Food and Drug Administration Manufacturer and User Facility Device Experience (MAUDE) system (2016–2025), we evaluated transparent, rule-defined queues for first-stage prioritisation. Component lexical rules were frozen before a procedurally blinded audit of 300 reports by nurse clinicians; a separate 30-report calibration set was excluded from performance estimates. The candidate queue family was finalised after audit summaries and evaluated descriptively. Nine queues were compared using exact corpus review burden, design-weighted clinician-reference yield, coverage, and reports requiring review per clinician-positive narrative, with 10,000 shared within-stratum bootstrap resamples. For a fixed corpus and rule definition, queue size is known exactly, whereas clinician-reference yield requires estimation. The triple-domain and software-plus-harm queues required review of 3.2% and 4.1% of the corpus and yielded 78.5% (95% confidence interval [CI], 63.0–92.2) and 76.0% (63.3–87.1), respectively, compared with a 36.5% random-review reference. Their coverage was 6.9% and 8.5%, leaving more than 90% of clinician-positive narratives outside each queue. Direct workflow screening required review of 31.3% of reports, yielded 52.6% (37.0–69.3), and covered 45.2%. An exposure-free restriction removed 23 of 107 nominal workflow-positive reports, including 16 clinician-positive reports, without improving the within-audit positive predictive value. These queues represent capacity choices rather than ranked solutions. They support auditable review-volume planning for human assessment, not autonomous classification or validated safety-signal detection. Prospective evaluation should measure review time, downstream actionability, and transportability.

1 Introduction

Medical-device adverse-event reporting captures suspected device-associated deaths, serious injuries, and malfunctions [1]. Post-market surveillance is therefore essential for identifying recurring problems and supporting corrective action after devices enter routine use. Its value, however, is shaped by how safety information is captured, reviewed, and acted on within a sociotechnical work system [2–5]. Free-text narratives can preserve contextual detail about device behaviour, care interruption, troubleshooting, and recovery that is not fully represented in structured fields. However, incident reports are selective and incomplete accounts rather than direct measures of incidence, causality, or patient risk [6–7].

The Manufacturer and User Facility Device Experience (MAUDE) database combines structured regulatory fields with linked free-text narratives in a complex multi-file reporting system [8]. Manufacturer submissions may be delayed [9], reporting volume has changed substantially over time, and the database lacks device-use denominators [10]. Narrative review can nevertheless identify safety-relevant information that is not represented by structured event categories [11]. Across millions of reports, the practical problem is therefore not simply whether relevant text can be identified, but which narratives should be reviewed first when expert time is limited.

Related informatics approaches address different aspects of this problem. Large language models (LLMs), expert-labelled transformer classifiers, and hybrid topic-modelling/LLM pipelines have been applied to MAUDE narratives [12–14], whereas incident-report natural language processing studies have largely evaluated classification tasks across heterogeneous outcomes and methods [15]. Governance frameworks emphasise traceability, monitoring, and accountable use of computational systems [16], while change-point analysis can identify temporal shifts in spontaneous reporting [17]. The present study addresses a different operational question: how text-derived rules translate into review queues with known review volumes.

For resource-constrained surveillance, a proposed queue must be judged on three linked dimensions: the proportion of the corpus requiring review, the concentration of clinician-relevant documentation within that queue, and the proportion of all clinician-positive reports retained. A narrow queue may be efficient yet leave most relevant narratives outside review, whereas a broad queue may improve coverage but be infeasible. Deterministic rules are useful for examining this trade-off because their versions and triggers can be inspected, reproduced, and governed.

We therefore conducted an exploratory evaluation of rule-defined review queues across 10 years of MAUDE data. We asked: (1) how exact review burden, clinician-reference yield, and coverage varied across single-domain, multidomain, direct-workflow, and structured-outcome queues; (2) how automated workflow labels compared with procedurally blinded nurse-clinician contextual judgements; and (3) whether lexical-separation and repeated-text analyses materially changed the observed operating profiles.

The intended contribution is an operational evaluation rather than a new predictive model. We combine exact population queue sizes, known without sampling error, with design-weighted estimates from a clinician-reference audit. No confirmatory hypothesis, preferred capacity threshold, or optimal queue was prespecified. The intended use is transparent first-stage prioritisation for human assessment; autonomous case assignment, clinical diagnosis, incidence estimation, causal inference, professional attribution, and safety-signal confirmation are outside the study scope.

2 Methods

2.1 Study design and data source

This retrospective observational study used publicly available MAUDE files and was reported in accordance with applicable items from the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) and REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statements [18–19]; completed reporting checklists accompany the submission. MAUDE master, device, and narrative files were analysed for receipt years 2016–2025; report year was derived from the DATE_RECEIVED field in the MAUDE master file. The source files were downloaded from the United States Food and Drug Administration (FDA) on 14 May 2026 [1, 20]. The FDA cautions that

MAUDE reports cannot be used alone to establish incidence, prevalence, or causality [1]. The unit of analysis was the report-level narrative identified by MDR_REPORT_KEY. Reports were not assumed to represent unique clinical events because related content may be submitted by different sources or through follow-up processes.

2.2 Corpus construction

The initial linkage database contained 18,575,099 reports. After excluding 21,027 reports without linked narrative text, 18,554,072 report-level narratives remained. Narrative segments were whitespace-normalised, trimmed, deduplicated within each report, and concatenated in recorded order. Operational data-quality thresholds excluded narratives shorter than 30 characters (n = 315,872; 1.702% of linked narratives) or longer than 20,000 characters (n = 9,583; 0.052%), leaving 18,228,617 eligible reports (Fig. 1; Table 1). Together, these thresholds removed 1.754% of reports with linked narrative text.

Structured event outcome was coded as malfunction, injury, death, or unknown. Report source, narrative-length category, and retained post-deduplication segment count characterised reporting composition. Device records were linked for product-code and manufacturer analyses. Missing or multiple linked values were retained as explicit categories rather than removed.

2.3 Rule development, frozen representations, and intended use

Deterministic dictionaries were developed iteratively. Broad software and workflow labels were overinclusive because of generic screen, display, connection, setting, reset, and restart language, prompting a stricter software definition. STRICT_SW_SIGNAL was not preregistered but was frozen before adjusted modelling and the blinded audit; reviewer ratings did not alter the rules (Supplementary Tables S1–S2).

Three exposure domains were retained: STRICT_SW_SIGNAL (SW; strict software signal), MECH_SIGNAL (MECH; mechanical signal), and HAZ_SIGNAL (HAZ; broad harm-related lexical signal). These defined eight mutually exclusive strata: NONE (no-signal), STRICT_SW_ONLY, MECH_ONLY, HAZ_ONLY, STRICT_SW_MECH, STRICT_SW_HAZ, MECH_HAZ, and STRICT_SW_MECH_HAZ (triple-domain). The post hoc operating-map family included the individual domains, their intersections, direct workflow-response screening (WF_ANY), and structured injury/death. It was formulated after the original audit summaries were available and is treated as exploratory (Supplementary Table S2).

WF_ANY identified documented delay, care-process interruption, repeated or corrective activity, technical escalation, workaround, or documentation gaps; WF_ANY_NO_REWORK excluded rework-only hits. Lexical outputs did not establish actor, clinical recovery, causality, or true workflow disruption. Because the manual framework used aligned categories, the audit is a contextual measurement layer, not an independent external construct.

The system was evaluated for human-review prioritisation, not autonomous classification. Queues were compared across capacity levels without designating a preferred option. Intended use, prohibited interpretations, governance, and failure modes are documented in Supplementary Table S3.

2.4 Blinded clinician reference audit

A 30-report reviewer-calibration set preceded the formal audit and was excluded from all performance estimates. The formal audit contained 300 reports sampled from the frozen labelled corpus. No formal power calculation was performed for the clinician audit; the 300-report sample was designed to provide representation across the nine sampling strata (Supplementary Table S4) while maintaining the feasibility of duplicate clinician review. The subsequent operating-map analysis was exploratory and was specified after the original audit summaries were available. To prevent reuse, 1,050 previously reviewed report keys were excluded from the formal audit sampling frame. Within the triple-domain group, 25 reports meeting the nested high-specificity adverse-condition criterion (HAZ_STRICT_ACTUAL: HAZ_SIGNAL plus an explicit, non-hypothetical adverse-condition assertion) and 20 remaining reports were sampled using an auxiliary stratification rule fixed before the 17 July 2026 final sample (Supplementary Table S2). Splitting the triple-domain stratum into these two internal subsets yielded nine sampling strata from the eight mutually exclusive exposure strata.

Three non-author registered nurses served as reviewer 1 (R1), reviewer 2 (R2), and reviewer 3 (R3); no manuscript author served in these roles. R1 had more than 10 years of emergency nursing experience, R2 more than 10 years of intensive-care nursing experience, and R3 was an emergency-care nurse manager. After the manual codebook was locked, R1 and R2 independently completed separate masked reviewer files; neither reviewer could access the other reviewer's ratings during independent review. Automated labels, sampling strata, report identifiers, and model results were withheld. R3 adjudicated disagreements without automated information. Reviewers interpreted negation, temporality, hypothetical events, clinical context, and actor. The resulting reference is a procedurally blinded nurse-clinician contextual reference, not engineering adjudication or an external gold standard (Supplementary Table S5).

The audit was standardised to its 18,227,567-report post-exclusion frame using nine internal sampling strata. Each report received design weight N_h/n_h [21], where N_h and n_h denote the audit-frame population size and audited sample size, respectively, for stratum h. Confidence intervals (CIs) were derived from 10,000 nonparametric bootstrap resamples within each stratum while preserving its formal-audit sample size [22]. All queue metrics were recalculated within each shared resample, preserving shared-denominator uncertainty and between-queue covariance. Unweighted Cohen's kappa confidence intervals used 10,000 paired nonparametric bootstrap resamples of the 300 R1–R2 rating pairs (seed 20260826). Standardisation to complete-corpus stratum counts produced nearly identical point estimates.

2.5 Exploratory operating-map estimands and statistical analysis

Let Q denote a review queue, N the full corpus, y_i the clinician-reference indicator, and w_i the audit design weight. Review burden was $B(Q) = |Q|/N$. Clinician-reference yield was $Y(Q) = \Sigma(i \in Q) w_i y_i / |Q|$, using the known full-corpus queue size; for WF_ANY and structured injury/death, this replaced audit-based denominator estimation. Accordingly, the operational yield combines an audit-frame-standardised clinician-positive numerator with the known full-corpus queue denominator; complete-corpus standardisation produced nearly identical point estimates (Supplementary Table S4). The 1,050 reports excluded to prevent reuse represented < 0.006% of the corpus.

Coverage was $C(Q) = \Sigma(i \in A \cap Q)\, w_i y_i / \Sigma(i \in A)\, w_i y_i$, with its denominator re-estimated in every bootstrap resample; by construction, this equals sensitivity against the clinician reference. Number needed to review (NNR) was $1/Y(Q)$. For random review, expected yield equalled the overall design-weighted clinician-positive proportion and expected coverage equalled the review fraction. No superiority ranking was prespecified.

Paired descriptive contrasts used the shared bootstrap resamples. The triple-domain queue (SW∩MECH∩HAZ) was compared with SW∩HAZ, where ∩ denotes intersection of rule-defined domains. HAZ restrictions within SW and MECH were assessed as rule-narrowing rather than causal effects. No multiplicity adjustment was applied to these exploratory comparisons. WF_ANY and structured injury/death were checked against known corpus margins as weight-mapping quality control.

Secondary aggregated-cell binomial models characterised the developed lexical system by standardising year, event outcome, source, narrative length, and segment count using predictive margins [23]. These contrasts are developmental because STRICT_SW_SIGNAL was refined after exploratory summaries. Product-code clustering used the CR1 cluster-robust covariance estimator [24]; full specifications are in the Supplement.

All data processing and statistical procedures were executed in Python using archived scripts. Frozen scripts, random seeds, file checksums, audit records, and aggregate provenance materials are retained in the reproducibility archive; analytic specifications are summarised in Supplementary Table S6.

No predictive model determined queue membership, so train/validation/test partitioning and calibration were not applicable. Random review and structured injury/death were operational comparators. A learned model trained and evaluated on the same 300 clinician labels would have compounded reuse and addressed a different question. Governance records appear in Supplementary Tables S2, S3, and S6.

2.6 Robustness and failure-mode analyses

Frozen scripts regenerated all exposure and workflow labels for the 300 audited reports before the exposure-free analysis. A post hoc stress test then required a workflow-trigger sentence without an exposure-domain trigger; it reused the clinician reference and enumerated removed positives directly.

Repeated-text dependence was assessed using the report-level bootstrap, an exact-normalised-text cluster bootstrap, and a one-report-per-text sensitivity. Other robustness analyses are reported in the

Supplement as failure-mode characterisation.

2.7 Ethics, reproducibility, and software tools

This study used publicly available regulatory files, involved no participant contact or institutional clinical records, and made no re-identification attempt. The Chung-Ang University Institutional Review Board determined this retrospective secondary analysis of publicly available FDA MAUDE reports to be exempt from ethics review (No. 1041078-20260523-HR-220; 6 June 2026).

The reproducibility archive contains frozen scripts and rules, audit manifests, random seeds, file checksums, provenance records, narrative-stripped ratings, and aggregate outputs (Supplementary Table S6).

Randomised procedures used prespecified random seeds recorded in the reproducibility archive. Figure 2 and Supplementary Figs. S1–S2 were generated in Python using Matplotlib. Figure 1 and Supplementary Fig. S3 were created using draw.io (diagrams.net).

3 Results

3.1 Corpus and exact review fractions

The analytic corpus comprised 18,228,617 report-level narratives: 11,514,305 malfunctions (63.2%), 6,596,224 injuries (36.2%), 108,093 deaths (0.6%), and 9,995 unknown outcomes (0.1%). The median narrative length was 739 characters (interquartile range 381–1,519), and the median retained segment count was 2 (Table 1).

Table 1
Corpus characteristics

Characteristic	n	% or summary
Initial linkage database	18,575,099	—
No linked narrative	21,027	0.1% of initial database
Narratives < 30 characters excluded	315,872	1.7% of linked narratives
Narratives > 20,000 characters excluded	9,583	0.1% of linked narratives
Final analytic corpus	18,228,617	100.0
Structured event outcome: malfunction	11,514,305	63.2
Structured event outcome: injury	6,596,224	36.2
Structured event outcome: death	108,093	0.6
Structured event outcome: unknown	9,995	0.1
Narrative length, characters, median (IQR)	—	739 (381–1,519)
Retained segments after deduplication, median (IQR)	—	2 (2–3)

Note. IQR = interquartile range; MAUDE = Manufacturer and User Facility Device Experience. Length-exclusion percentages use the 18,554,072 reports with linked narrative text as denominator; event-outcome percentages use the final analytic corpus. Percentages may not sum to 100.0 because of rounding.

Review burden varied by more than fifteen-fold across candidate queues. It ranged from 3.21% for the triple-domain queue to 49.24% for MECH. SW∩HAZ was similarly narrow (4.08%), whereas direct WF_ANY and structured injury/death required review of 31.31% and 36.78% of the corpus, respectively. Exact counts and fractions for all candidate queues are shown in Table 2 and Supplementary Tables S7 and S9; these are full-corpus quantities, not audit estimates.

3.2 Clinician-reference operating map

Candidate queues showed different combinations of review burden, yield, and coverage. Narrower conjunctions generally concentrated higher-yield subsets, whereas broader queues captured a larger share of clinician-positive reports, although the relationship was not monotonic. The overall design-weighted clinician workflow-positive proportion in the audit frame was 36.5% (95% CI 29.4–43.7), defining the expected yield of random review. The triple-domain queue yielded 78.5% (63.0–92.2) with 6.9% coverage, and SW∩HAZ yielded 76.0% (63.3–87.1) with 8.5% coverage. Their yield difference was 2.5 percentage points (95% CI − 2.3 to 7.4); these adjacent low-burden queues could not be reliably ranked.

Broader queues traded yield for coverage. Direct WF_ANY screening required review of 31.31% of the corpus, with 52.6% (37.0–69.3) clinician-reference yield and 45.2% (33.5–57.5) coverage. MECH required review of 49.24% and reached 58.8% coverage with 43.6% yield. The HAZ queue and the structured injury/death queue were near the random-review reference; structured injury/death yielded 35.6% (23.7–48.4) (Table 2; Fig. 2). Based on point estimates, WF_ANY required less review than structured injury/death while providing higher yield and coverage; this descriptive comparison was not a superiority test.

Table 2
Exploratory review-burden, clinician-reference yield, and coverage operating map

Queue	Review burden (% of corpus)	Clinician-reference yield (95% CI)	Coverage (95% CI)	NNR (95% CI)
Triple (SW∩MECH∩HAZ)	3.21%	78.5% (63.0–92.2)	6.9% (5.3–9.0)	1.27 (1.08–1.59)
SW∩HAZ	4.08%	76.0% (63.3–87.1)	8.5% (6.7–10.9)	1.32 (1.15–1.58)
SW∩MECH	9.51%	64.4% (52.7–76.1)	16.8% (13.1–21.5)	1.55 (1.31–1.90)
MECH∩HAZ	12.75%	57.2% (44.9–69.3)	20.0% (15.5–25.7)	1.75 (1.44–2.22)
SW	14.54%	57.5% (48.6–66.5)	22.9% (18.3–29.0)	1.74 (1.50–2.06)
HAZ	24.57%	38.0% (29.8–46.6)	25.6% (19.8–32.7)	2.63 (2.15–3.36)
WF_ANY direct	31.31%	52.6% (37.0–69.3)	45.2% (33.5–57.5)	1.90 (1.44–2.70)
Structured injury/death	36.78%	35.6% (23.7–48.4)	35.9% (25.2–47.4)	2.81 (2.07–4.21)
MECH	49.24%	43.6% (33.9–53.5)	58.8% (48.8–69.3)	2.29 (1.87–2.95)
Random review reference[a]	f	36.5% (29.4–43.7)	= f	2.74 (2.29–3.40)

Note. SW = strict software signal; MECH = mechanical signal; HAZ = broad harm-related lexical signal; ∩ = conjunction. CI = confidence interval; NNR = estimated reports requiring review per clinician-positive narrative identified (1/yield). Review burdens are exact proportions of the 18,228,617-report corpus and carry no sampling error; yield uses the known queue size as denominator, whereas coverage is design-weighted to the 18,227,567-report audit frame (Section 2.5). Intervals are from 10,000 shared within-stratum bootstrap resamples. The deliberately stratified formal audit included 45 reports in the automated triple-domain stratum (25 high-specificity and 20 remaining). Manual three-domain concurrence is a separate measurement layer and is reported in Section 3.3 and Supplementary Table S8. Estimates are exploratory; no queue is designated optimal.
a Analytical reference, not an observed queue; expected coverage equals the chosen review fraction f.

3.3 Audit validity and measurement-layer comparison

Reviewer agreement was high across the main manual fields. For workflow-response documentation, R1–R2 agreement was 94.7% (kappa = 0.893; 95% CI 0.840–0.940); agreement was 92.0% (kappa = 0.840; 95% CI 0.776–0.899) for broad harm-related content and 95.7% (kappa = 0.815; 95% CI 0.708–0.903) for the high-specificity adverse-condition reference. Software, mechanical, and component workflow fields showed 100% agreement. In all three contextual fields with disagreements, R2 assigned more positive ratings than R1 (176 vs 152, 47 vs 34, and 149 vs 133, respectively); R3 adjudicated all discordant items. Kappa was interpreted alongside percentage agreement because it depends on marginal prevalence. Automated-versus-manual eight-group agreement was 70/300 (23.3%; 95% CI 18.9–28.4) unweighted and 30.8% (24.0–38.0) after design weighting; six reports met the manual three-domain reference. This low cross-layer concordance may reflect both differences in measurement granularity and imperfect lexical capture; the audit could not separate these mechanisms. No final reference rating remained uncertain (Supplementary Tables S4, S5, and S8).

Restriction effects were asymmetric across domains. Within SW, requiring HAZ increased clinician-reference yield by 18.5 percentage points (95% CI 6.2–30.1) and decreased coverage by 14.4 points (− 18.9 to − 10.8). Within MECH, requiring HAZ increased yield by 13.6 points (0.3–26.4) and decreased coverage by 38.9 points (− 48.6 to − 29.0). These are rule-narrowing contrasts, not causal contributions of HAZ. SW and MECH∩HAZ occupied nearby operating regions: their yield difference was 0.3 points (95% CI − 14.2 to 14.4) and their coverage difference was 2.9 points (− 2.4 to 8.4).

The archived automated labels were reproduced for all 300 audited reports before the exposure-free analysis. In observed unweighted counts, the exposure-free restriction reduced nominal WF_ANY positives from 107 to 84; 16 of the 23 removed reports were clinician workflow-positive. For WF_ANY_NO_REWORK, 18 of 83 nominal positives were removed, 15 of which were clinician-positive. Using design-weighted audit-ratio classifier metrics, calculated as within-audit ratios and therefore distinct from the fixed-known-denominator operational queue yield in Table 2, sensitivity decreased from 45.2% to 37.1%, specificity increased from 74.5% to 78.6%, and positive predictive value (PPV) was essentially unchanged (50.4% vs 49.8%). WF_ANY coverage therefore equals sensitivity (45.2%), whereas its 52.6% clinician-reference yield uses the known full-corpus queue size as denominator.

Repeated-text dependence also had little effect on the principal map. Triple-domain yield remained 78.5%, with an exact-text cluster-bootstrap interval of 63.0–91.8% and a one-report-per-text yield of 78.5% (Table 3; Supplementary Tables S9–S11).

Table 3
Same-sample restriction, lexical-separation, and duplicate-sensitivity findings

Analysis	Estimate	95% CI / supporting interval	Interpretation
Triple vs SW∩HAZ yield	+ 2.5 pp	−2.3 to + 7.4 pp	Adjacent low-burden queues not reliably ranked
HAZ restriction within SW	Yield + 18.5 pp; coverage − 14.4 pp	Yield + 6.2 to + 30.1; coverage − 18.9 to − 10.8 pp	Narrower queue: higher yield, lower coverage
HAZ restriction within MECH	Yield + 13.6 pp; coverage − 38.9 pp	Yield + 0.3 to + 26.4; coverage − 48.6 to − 29.0 pp	Narrower queue: higher yield, large coverage loss
Nominal WF_ANY classifier metrics (audit-ratio)	Sensitivity 45.2%; specificity 74.5%; PPV 50.4%	See Supplementary Table S8, Panel E2	Same-sample lexical layer; not Table 2 fixed-denominator queue yield
Exposure-free WF_ANY classifier metrics (audit-ratio)	Sensitivity 37.1%; specificity 78.6%; PPV 49.8%	Point estimates only; see Supplementary Table S10	Same-sample stress test; not Table 2 fixed-denominator queue yield
Exposure-free removal	23/107 nominal positives removed; 16/23 clinician-positive	—	Removed reports were not selectively false positives
Triple duplicate-text sensitivity	Yield 78.5%; coverage 6.9%	Text-cluster yield CI 63.0–91.8%	No material duplicate-text effect

Note. pp = percentage points. Paired restriction differences use shared within-stratum bootstrap resamples. HAZ comparisons describe rule narrowing, not causal contribution. Exposure-free metrics use the same 300 clinician references after successful 300/300 nominal-label regeneration. PPV values in the nominal and exposure-free rows are same-sample design-weighted audit-ratio classifier metrics used for within-audit comparison; they are distinct from the fixed-known-denominator operational queue yield in Table 2. Unweighted sample-conditional equivalents for the nominal automated labels appear in Supplementary Table S8, Panels B–D; paired restriction results are detailed in Supplementary Table S10. Exact-text analyses are dependence sensitivities; the primary sampling design remains report-level stratified sampling. Percentage-point differences were calculated from unrounded estimates.

3.4 Population-scale automated characterisation

Detailed secondary full-corpus lexical analyses are reported as development-stage characterisation in the Supplement. In the final rule system, adjusted WF_ANY probability was 47.1% in triple-domain reports and 27.0% in NONE reports (difference 20.1 percentage points, 95% CI 20.0–20.2); the exposure-free restriction reduced this difference to 11.4 points. These narrow intervals quantify model-based sampling uncertainty conditional on the automated labels; they do not capture uncertainty in construct definition, iterative rule development, or temporal transportability. Product-code and manufacturer adjustment retained positive point estimates for contrasts with the NONE group, but product-code

cluster-robust intervals widened substantially, and intervals for contrasts between the triple-domain and MECH_ONLY groups included the null (Supplementary Tables S12–S14; Supplementary Figs. S1–S2).

4 Discussion

4.1 Principal findings

The low-burden triple-domain and SW∩HAZ queues concentrated clinician-positive narratives, with yields of 78.5% and 76.0%, but covered only 6.9% and 8.5% of the estimated clinician workflow-positive reports. Consequently, an estimated 93.1% and 91.5% of true positive reports remained outside these highly targeted queues. Direct WF_ANY screening increased coverage to 45.2% but required review of 31.3% of all reports. These findings show that a high-yield queue can still leave most clinician-positive narratives unreviewed.

The operating map places the triple-domain findings within a broader capacity trade-off rather than identifying a single best rule. SW∩HAZ occupied an adjacent low-burden region and could not be reliably distinguished from the triple-domain queue, whereas WF_ANY and MECH exchanged lower yield for greater coverage. Choice among these operating points would require explicit information on available staff time, review duration, acceptable missed-case burden, and the consequences of unnecessary review. None of those deployment outcomes was measured here.

This operational question differs from both safety-signal detection and predictive text classification. Change-point analysis identifies temporal changes in spontaneous-report patterns [17], and targeted text mining can retrieve additional health-record cases after a candidate signal has been identified [25]. The present study instead evaluates first-pass allocation of human review across the eligible regulatory narrative corpus. Clinician workflow-response documentation is an operational review construct; it is not a validated safety signal, severity endpoint, or measure of patient benefit.

4.2 Review-queue breadth and review-volume trade-offs

Adding domain requirements changed selectivity without uniformly improving the operating profile. HAZ alone yielded 38.0%, close to the 36.5% random-review reference. Requiring HAZ within SW reduced burden from 14.5% to 4.1%, increased yield by 18.5 percentage points, and reduced coverage by 14.4 points; within MECH, burden fell from 49.2% to 12.8%, yield increased by 13.6 points, and coverage fell by 38.9 points. These contrasts quantify rule narrowing and should not be interpreted as a causal role for HAZ.

At corpus scale, these differences translate into materially different review volumes. The triple-domain queue contained 585,547 reports, direct WF_ANY contained 5,706,888, and exhaustive review contained 18,228,617. Direct WF_ANY therefore contained approximately 9.7 times as many reports to review as the triple-domain queue, whereas exhaustive review contained approximately 31.1 times as many. Unlike clinician-reference yield, these queue sizes are exact corpus quantities, so review-volume implications

can be assessed before local performance validation. Averaged over the 10-year corpus and assuming 250 working days per year, the triple-domain and direct WF_ANY queues correspond to approximately 234 and 2,283 reports per working day, respectively. These historical averages are illustrative and do not establish current queue volumes or staffing requirements. The appropriate operating point will depend on local capacity and the value assigned to coverage.

The comparison with random review is also operationally informative. Structured injury/death fields did not concentrate clinician workflow-positive narratives beyond the random-review reference, while direct WF_ANY showed greater enrichment and coverage on point estimates. This finding supports examining narrative context when allocating review, but it does not establish that WF_ANY improves regulatory decisions or patient safety.

4.3 Measurement layers, lexical dependence, and human verification

The full-corpus lexical analysis and clinician audit assessed different aspects of the same narratives. Because STRICT_SW_SIGNAL was refined after exploratory summaries and automated exposure and workflow labels came from the same text, the 20.1-percentage-point contrast characterises the developed rule system rather than an independent clinical association. Clinicians additionally interpreted negation, temporality, hypothetical events, and contextual referents.

The exposure-free analysis illustrates this distinction. Sixteen of 23 removed WF_ANY-positive reports were clinician workflow-positive, while audit-ratio PPV was essentially unchanged (50.4% vs 49.8%). Sentence-level separation therefore removed both clinician-positive and clinician-negative reports. Given the small number of removed reports and the conceptually aligned reference, this should be interpreted as a lexical sensitivity analysis rather than clinical validation.

4.4 Transparency, governance, and intended use

One advantage of deterministic rules in this setting is that their operating behaviour can be inspected directly: versions can be rerun, triggers traced, and lexical overlap, repeated templates, and device-mix dependence examined directly. Supplementary Table S3 documents intended use, prohibited interpretations, governance requirements, monitoring, and failure modes; such documentation supports accountable use [16] but does not replace local validation.

The intended architecture is human-in-the-loop: rules prioritise narratives and reviewers determine context, relevance, actor, and actionability (Supplementary Fig. S3). Because the same 300-report audit informed candidate characterisation and performance estimation, the operating map is descriptive and exploratory. A mature holdout dataset for 2026 was unavailable due to inherent reporting delays and temporal accumulation [9–10]; future evaluation should lock rules, capacity bands, endpoints, sample size, sampling allocation, and decision criteria prospectively.

Practical implementation would additionally require local validation, monitoring, and prespecified drift-response criteria; transfer to other systems, languages, or device portfolios would require revalidation.

4.5 Strengths and limitations

Strengths include the 10-year MAUDE corpus, exact full-corpus review-volume calculations, explicit reporting of the post hoc candidate family, duplicate nurse-clinician review under procedural blinding, design-based weighting, shared-resample contrasts, duplicate-text analyses, and archived provenance. Together, these features make the assumptions and review-volume implications of each queue inspectable.

Limitations include iterative refinement of the strict software rule and same-source coupling because exposure and workflow labels were derived from the same narrative. The procedurally blinded clinician reference was conceptually aligned with the automated taxonomy and lacked independent engineering or regulatory adjudication. The narrative-length exclusions (< 30 and > 20,000 characters) were operational data-quality thresholds, and sensitivity to alternative cutoffs was not evaluated.

The 300-report audit provided imprecise estimates for several queues; NONE and MECH_ONLY each contained only 40 audited reports, and only six reports met the manual three-domain reference. Coverage estimates therefore depend partly on weighted clinician-positive counts from sparsely sampled strata; the stratified bootstrap reflects sampling uncertainty but not uncertainty in the clinician-reference construct. The same audit informed all candidate operating points, and the exposure-free removal count was a small-sample descriptive result. The reported intervals are conditional on the final post hoc candidate family and do not account for uncertainty introduced by its formulation.

MAUDE lacks a device-use denominator and is subject to selective and delayed reporting [1, 9–10]; reports need not represent unique clinical events. One exact-normalised narrative appeared in both calibration and audit samples, and exact-text sensitivity addresses only one dependence mechanism. Product-code CR1 estimates remain vulnerable to extreme cluster imbalance. The retained archive supports reconstruction of rule, file, and analysis provenance, although an exact package-version lockfile for the May analyses was not retained. No independent temporal cohort or prospective queue was evaluated, and review time, actionability, regulatory response, missed-case consequences, downstream safety action, and patient benefit were not measured.

5 Conclusions

Rule-defined MAUDE queues showed distinct review-volume, clinician-reference yield, and coverage profiles. Narrow multidomain queues concentrated clinician-positive narratives at a 3–4% review burden but left more than 90% of the estimated clinician-positive population outside review; broader workflow screening increased coverage at substantially greater review volume. The triple-domain and SW∩HAZ queues could not be reliably ranked. This exploratory operating map offers an auditable approach to prioritising narratives for human assessment using exact queue sizes and weighted reference data. For

a fixed corpus and rule definition, queue size is determined exactly, whereas clinician-reference yield requires estimation; this asymmetry allows capacity planning to be separated analytically from classification-performance validation before implementation. The analysis does not establish autonomous classification performance, safety-signal validity, or clinical utility. Prospective, independently specified evaluation is needed before deployment.

Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests

The authors declare no competing interests.

Ethics Approval

The study used publicly available regulatory files and was determined by the Chung-Ang University Institutional Review Board to be exempt from ethics review (No. 1041078-20260523-HR-220; 6 June 2026). The exemption was granted for retrospective secondary analysis of publicly available FDA MAUDE reports; no participants were recruited, and individual informed consent was not applicable.

Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

Data Availability

Source MAUDE files are publicly available from the United States Food and Drug Administration [1, 20]. A versioned, narrative-stripped reproducibility package will be provided to editors and reviewers through the submission system during peer review and deposited in a public repository with a persistent identifier no later than acceptance. The package will include reviewer-calibration and audit documentation, reviewer ratings without redistributed narrative text, adjudicated references, rule and provenance materials, operating-map analysis materials, design-weighted analysis code, aggregate

outputs, random seeds, file checksums, duplicate-text assessments, and full-corpus aggregate queue counts. Raw regulatory narratives will not be redistributed by the authors.

Materials and Code Availability

Frozen rule specifications, preprocessing and modelling scripts, provenance records, audit manifests, design-weighted operating-map code, random seeds, checksums, and aggregate outputs will be included in the reproducibility package. The package will be provided to editors and reviewers through the submission system during peer review and deposited in a public repository with a persistent identifier no later than acceptance.

Author Contributions

Author contributions are provided on the title page and in the submission interface.

Clinical trial number

Not applicable.

References

1. U.S. Food and Drug Administration (2025) About Manufacturer and User Facility Device Experience (MAUDE) database. https://www.fda.gov/medical-devices/mandatory-reporting-requirements-manufacturers-importers-and-device-user-facilities/about-manufacturer-and-user-facility-device-experience-maude-database. Accessed 2 July 2026
2. Sittig DF, Singh H (2010) A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care 19(Suppl 3):i68–i74. https://doi.org/10.1136/qshc.2010.042085
3. Carayon P, Schoofs Hundt A, Karsh BT, Gurses AP, Alvarado CJ, Smith M, Brennan PF (2006) Work system design for patient safety: the SEIPS model. Qual Saf Health Care 15(Suppl 1):i50–i58. https://doi.org/10.1136/qshc.2005.015842
4. Holden RJ, Carayon P, Gurses AP, Hoonakker P, Schoofs Hundt A, Ozok AA, Rivera-Rodriguez AJ (2013) SEIPS 2.0: a human factors framework for studying and improving the work of healthcare professionals and patients. Ergonomics 56:1669–1686. https://doi.org/10.1080/00140139.2013.838643
5. Gong Y, Kang H, Wu X, Hua L (2017) Enhancing patient safety event reporting: a systematic review of system design features. Appl Clin Inform 8:893–909. https://doi.org/10.4338/ACI-2016-02-R-0023
6. Macrae C (2016) The problem with incident reporting. BMJ Qual Saf 25:71–75. https://doi.org/10.1136/bmjqs-2015-004732

7. World Health Organization (2020) Patient safety incident reporting and learning systems: technical report and guidance. World Health Organization, Geneva. https://www.who.int/publications/i/item/9789240010338. Accessed 2 July 2026
8. Ensign LG, Cohen KB (2017) A primer to the structure, content and linkage of the FDA's Manufacturer and User Facility Device Experience (MAUDE) files. eGEMs 5:12. https://doi.org/10.5334/egems.221
9. Everhart AO, Karaca-Mandic P, Redberg RF, Ross JS, Dhruva SS (2025) Late adverse event reporting from medical device manufacturers to the US Food and Drug Administration: cross sectional study. BMJ 388:e081518. https://doi.org/10.1136/bmj-2024-081518
10. Mishali M, Sheffer N, Mishali O, Negev M (2025) Evaluation of reporting trends in the MAUDE database: 1991 to 2022. Digit Health 11:20552076251314094. https://doi.org/10.1177/20552076251314094
11. Lalani C, Kunwar EM, Kinard M, Dhruva SS, Redberg RF (2021) Reporting of death in US Food and Drug Administration medical device adverse event reports in categories other than death. JAMA Intern Med 181:1217–1223. https://doi.org/10.1001/jamainternmed.2021.3942
12. Shi Y, Gong Y (2025) Leveraging LLMs to understand narratives in MAUDE reports. Stud Health Technol Inform 323:194–198. https://doi.org/10.3233/SHTI250076
13. Luschi A, Nesi P, Iadanza E (2023) Evidence-based clinical engineering: health information technology adverse events identification and classification with natural language processing. Heliyon 9:e21723. https://doi.org/10.1016/j.heliyon.2023.e21723
14. Kesharwani R, Cvijic L, Denecke K (2026) Automating safety surveillance for software-based medical devices: insights from FDA MAUDE data. Stud Health Technol Inform 336:615–619. https://doi.org/10.3233/SHTI260244
15. Young IJB, Luz S, Lone N (2019) A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. Int J Med Inform 132:103971. https://doi.org/10.1016/j.ijmedinf.2019.103971
16. Crigger E, Reinbold K, Hanson C et al (2022) Trustworthy augmented intelligence in health care. J Med Syst 46:12. https://doi.org/10.1007/s10916-021-01790-z
17. Takahashi K, Yokoi H, Mikami F et al (2026) Signal detection using change point analysis to identify safety signals from spontaneous reports in aortic stent grafts in Japan. J Med Syst 50:61. https://doi.org/10.1007/s10916-026-02388-z
18. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP (2007) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 147:573–577. https://doi.org/10.7326/0003-4819-147-8-200710160-00010
19. Benchimol EI, Smeeth L, Guttmann A et al (2015) The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med 12:e1001885. https://doi.org/10.1371/journal.pmed.1001885

20. U.S. Food and Drug Administration (2026) MDR data files. https://www.fda.gov/medical-devices/medical-device-reporting-mdr-how-report-medical-device-problems/mdr-data-files. Accessed 2 July 2026
21. Lohr SL (2022) Sampling: Design and Analysis, 3rd ed. Chapman & Hall/CRC, Boca Raton. https://doi.org/10.1201/9780429298899
22. Davison AC, Hinkley DV (1997) Bootstrap Methods and Their Application. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511802843
23. Williams R (2012) Using the margins command to estimate and interpret adjusted predictions and marginal effects. Stata J 12:308–331. https://doi.org/10.1177/1536867X1201200209
24. Cameron AC, Miller DL (2015) A practitioner's guide to cluster-robust inference. J Hum Resour 50:317–372. https://doi.org/10.3368/jhr.50.2.317
25. Abedian Kalkhoran H, Zwaveling J, van Hunsel F, Kant A (2024) An innovative method to strengthen evidence for potential drug safety signals using Electronic Health Records. J Med Syst 48:51. https://doi.org/10.1007/s10916-024-02070-2

Figures

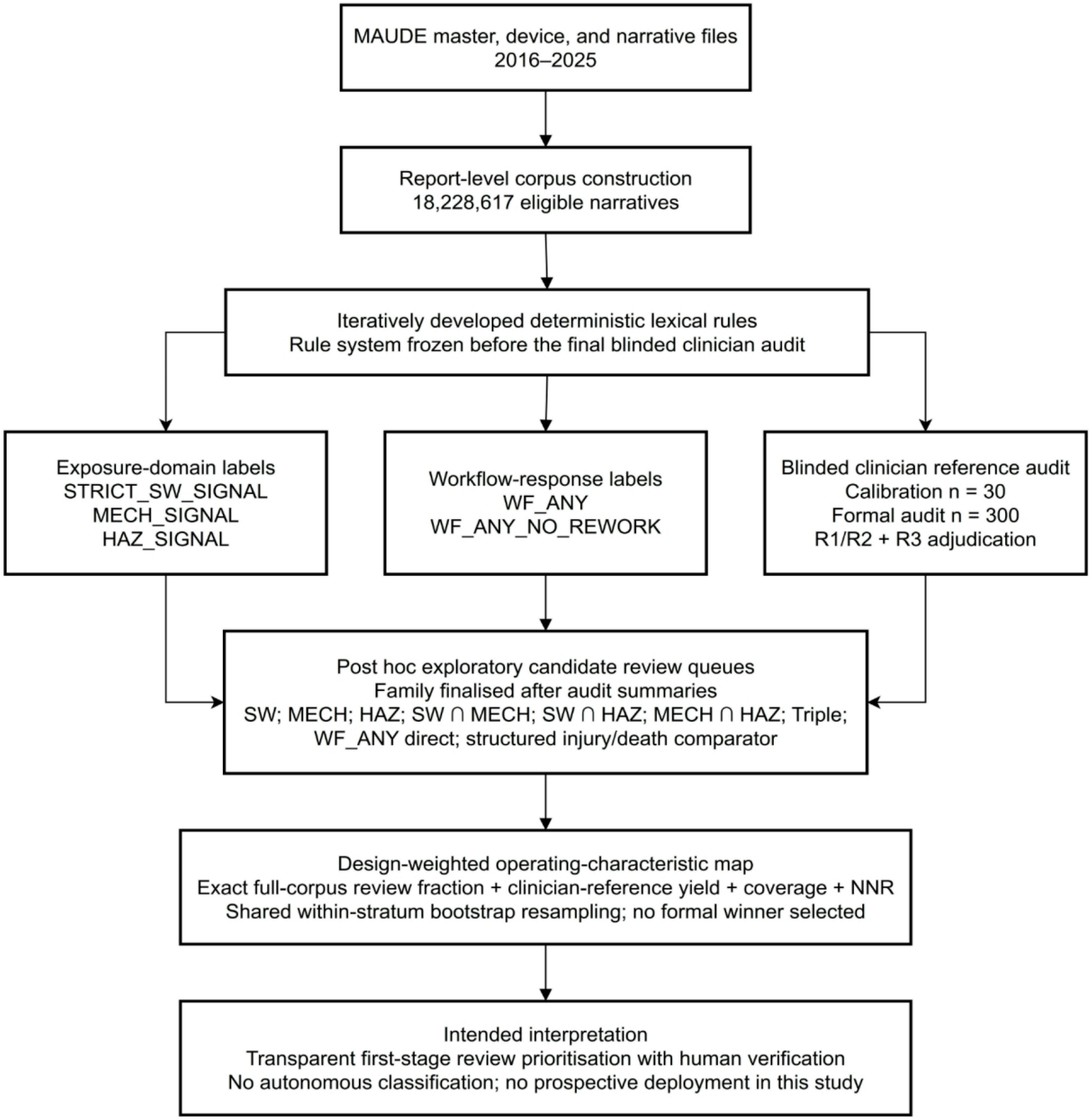


Figure 1

Study architecture and evidence hierarchy. Component lexical rules were frozen before the formal procedurally blinded clinician audit. The candidate family was defined post hoc after the original audit summaries and is interpreted as exploratory. Shared within-stratum bootstrap resamples were used for uncertainty estimation

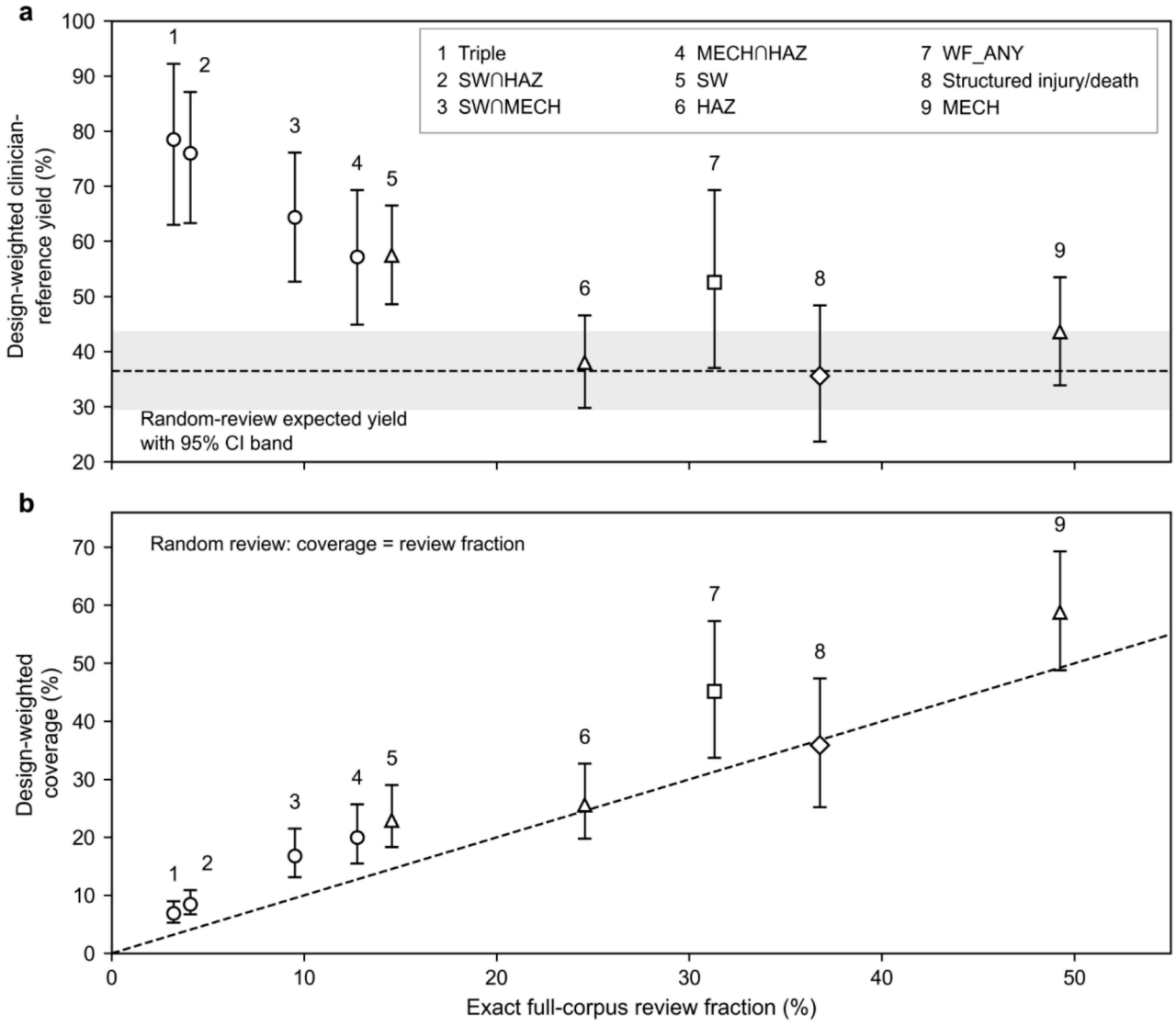


Figure 2

Exploratory operating-characteristic map of candidate review queues. a Exact full-corpus review fraction versus design-weighted clinician-reference yield for workflow-response documentation. The dashed horizontal line indicates the expected random-review yield (36.5%), and the shaded band its 95% confidence interval (29.4%–43.7%). b Exact full-corpus review fraction versus design-weighted coverage. The dashed diagonal represents random review, for which expected coverage equals the review fraction. Error bars show 95% confidence intervals from 10,000 within-stratum bootstrap resamples. Queue numbers correspond to the key shown in panel a. SW = strict software signal; MECH = mechanical signal; HAZ = broad harm-related lexical signal; ∩ = conjunction

Supplementary Files

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

- 03JOMSSupplementaryFINAL.docx