The Eastasouth Journal of Information System and Computer Science
Vol. 2, No. 02, December 2024, pp. 386-397
ISSN: 3025-566X, DOI: 10.58812/esiscs.v2i02

Routing Unreliable Risk Scores in Capacity Limited Decision Systems

Saliheen Ahmaad Rayeef
Gulf University for Science and Technology, Mubarak Al-Abdullah, Kuwait

Article Info

Article history:

Received Dec, 2024
Revised Dec, 2024
Accepted Dec, 2024

Keywords:

Data Quality;
Decision Analytics;
Monte Carlo Simulation;
Review Capacity;
Selective Routing

ABSTRACT

Organizations often have more flagged records than staff can review. A data-quality flag can trigger backup scoring or remove a record from the queue, but these responses have different consequences for event capture. This study compared four policies using 200 paired simulated cohorts of 4,000 records and a review budget of 600 per cohort. The primary score was more precise than the backup on clean records. A subset of primary scores was inflated, and an imperfect quality flag identified potentially corrupted scores. At 15% corruption, routing flagged records to backup scoring captured 217.9 events per cohort versus 171.5 with primary scoring alone. The mean paired difference was 46.3 events (95% Monte Carlo bootstrap interval [44.9, 47.8]). The clean scenario quantified the opportunity cost of using a less precise backup. These results characterize a queue allocation mechanism under specified assumptions and provide a reproducible design for testing the mechanism on operational data.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Name: Saliheen Ahmaad Rayeef
Institution: Gulf University for Science and Technology, Mubarak Al-Abdullah, Kuwait
Email: marium.taj20@gmail.com

1. INTRODUCTION

A review queue converts a statistical score into a decision about who receives attention first. A modest upward error can matter when it moves a record just above the last available slot. A large upward error can dominate a whole queue. When the same data pipeline supports thousands of decisions, a measurement problem can therefore change the population that receives review even if the underlying event process stays stable.

The operational choice is what to do with a record whose score has become suspect. Keeping it in the primary ranking preserves access but risks crowding out reliable candidates. Removing it protects the ranking from suspect values but may exclude records with genuine events. Routing it to a second score preserves access at the cost of accepting a noisier signal. The best choice depends on the reliability of the flag, the backup and the available capacity.

We ask how event capture and queue composition change as primary score corruption increases. The experiment holds latent event risk and review capacity fixed, so comparisons isolate the consequences of ranking and routing. The unit of analysis is the simulated cohort, and the estimand is the paired difference in events captured by alternative policies under the same conditions.

A review slot has a boundary that most predictive metrics do not describe. The candidate ranked six hundredth receives

attention in this study, while the candidate ranked six hundred and first waits. Neither record needs to have a substantially different estimated risk from the other. A change in the information used to produce the ranking can nonetheless determine who receives attention. The distinction matters whenever the operational objective is to find as many events as possible with a staff budget that cannot expand on demand. An analyst can improve the overall ranking and still make little difference near the relevant cutoff. Conversely, a failure in a small subset of records can dominate the cutoff if the failure moves those records far enough upward.

The problem arises even in a setting where the event definition and the population remain stable. Consider a queue assembled from a recent measurement and a secondary source that is usually less informative. If the recent measurement is unreliable for some records, a score computed from it can imply more urgency than the available evidence warrants. A quality check might identify some of those records before allocation. It will also miss some failures and may occasionally mark sound records. The organization then has to decide whether a flag is a reason to use a different source, a reason to wait, or simply information for the reviewer. The study isolates this decision by holding the underlying risks fixed across policy comparisons.

A queue policy can make use of a warning without changing the underlying prediction model. In this study, routing acts after both scores have been generated and before any review slots are assigned. The rule is deliberately simple: use the better score when it is presumed available and switch to a noisier backup when the quality check raises a concern. This is a useful benchmark because its behavior is visible in the record-level decisions. It also gives a clear comparison with the operational practice of holding flagged records aside. Under deferral, the shortage is transferred from a scoring problem to an eligibility problem. The review staff may see a cleaner queue while flagged records with genuine events wait.

The four policies answer distinct questions. The primary-only policy estimates what happens when the queue ignores the quality flag. The backup-only policy asks how much could be gained by giving up the stronger but vulnerable signal across the entire cohort. Routing asks whether the flag contains enough information to preserve the stronger signal for most records while correcting some errors. Deferral asks whether restricting eligibility is sufficient to protect review performance. These comparisons make the cost of imperfect flagging interpretable. A useful signal can still trigger the wrong action if the response is too costly for the records it marks.

The study makes two contributions. It specifies a small, reproducible experiment in which the same records can be evaluated under all four policies. It then reports both event capture and the provenance of selected records, which helps explain changes in ranking that a single performance measure would obscure. The design gives a controlled view of one failure mode. The simulation also establishes a vocabulary for subsequent empirical tests, where the flag error rates, backup accuracy and distribution of failures would need to be measured rather than stipulated.

2. RELATED WORK AND DESIGN RATIONALE

Comparisons of prediction, feature reliance and ranking stability in diabetes progression motivate evaluating the stability of the review order [1]. Survey-weighted cardiovascular analytics illustrates the distinction between estimating population burden and ranking individual records for follow-up [2].

Chronic disease risk stratification motivates measuring who reaches the review stage when scores determine access [3]. Multi-omics research illustrates why missing or unreliable modalities can matter for downstream candidate prioritization [4]. Cancer analytics links predictive ranking to the separate question of the cost of follow-up [5].

Bioinformatics for anticancer discovery illustrates another kind of prioritization, where candidate ranking precedes expensive follow-up [6]. AI-assisted drug discovery illustrates the distinction between candidate prioritization and downstream confirmation [7]. Real-time epidemiological surveillance illustrates the importance of recording data freshness before a score enters an operational queue [8]. Machine learning for epidemic prediction provides a related setting where surveillance models depend on the condition of incoming records [9]. Multisource geospatial analytics provides a parallel example of decision support using inputs with different provenance [10].

These examples describe a family of decision problems with a common structure. A score draws on evidence collected for some purpose, and a person or team must decide where to spend scarce attention. In health analytics, the distinction between estimating disease burden and prioritizing individual follow-up is especially consequential. The former can be useful even if no individual record is placed in a queue. The latter requires the ranking to remain meaningful for the specific cases near the review threshold. Both depend on how the input was measured, but they address different questions. The present study concentrates on individual allocation while using a generic binary outcome to avoid assuming a particular intervention or clinical pathway.

Applications that combine several sources also raise an issue of provenance. A secondary signal may come from an earlier observation, a different instrument, or a less detailed record. Its usefulness depends on whether it tracks the same outcome in the same population and on whether the original measurement problem also affects it. The simulation uses a separate error term for the backup to represent an alternative source that remains available during the specified failure. In practice, two systems may draw from the same upstream record even when their interfaces and model architectures differ. A backup that inherits the same error offers less protection than the independent error in the present experiment suggests.

The relationship between data governance and analytics success motivates assigning ownership for the quality flag and its review rules [11]. Work on explainable decision support motivates exposing which scoring path produced a recommendation [12]. Cloud-based project analytics motivates separating the scoring infrastructure from the rule governing allocation of reviewer time [13]. Agile project management provides an implementation context for testing queue policies through staged releases [14]. AI-driven project management provides an example of routing analytic outputs through existing management systems [15].

For decision support, an auditable workflow needs to identify the evidence behind a ranking at the point when a reviewer encounters it. A record that enters on a backup score should carry enough information to explain the switch and distinguish it from a routine update. Otherwise a quality event may look like a change in underlying risk. The same distinction matters in retrospective evaluations. A team that retains only the final ranking cannot reconstruct which outcomes would have been captured by primary-only scoring or how many flagged records were held aside. Preserving decision inputs allows the team to compare candidate policies against the same set of eligible records, subject to the limits of its outcome data.

A rollout also creates a question of ownership. Someone must define what a flag means, decide when the condition expires and determine who handles a record that cannot receive a reliable score. A flag can represent a temporary delay in data delivery or a more persistent measurement defect. The appropriate handling of those cases may differ even when both receive the same indicator in a simplified simulation. An operational evaluation therefore needs to record the reasons for flags and the time until those reasons are resolved. Those fields support analysis of queue performance and allow an organization to review whether the

policy is creating a persistent backlog for some sources.

Predictive supply-chain analytics gives a nonclinical example of triage when disruption review capacity is limited [16]. Comparative supermarket forecasting illustrates why model choice should be evaluated in its operational context [17]. Smart manufacturing analytics offers an example of a workflow with multiple signals and finite inspection capacity [18]. Customer experience analytics provides an example of a setting in which a score may affect who receives follow-up [19].

Employee retention analytics supplies a separate example in which scores can inform the order of follow-up under constrained resources [20]. Research on microcredit provides socioeconomic context for considering who may lose access when uncertain records are deferred [21]. Energy-market analytics illustrates that the cost of a backup service belongs in an eventual resource analysis [22]. Work on data-center efficiency highlights that maintaining a backup scoring path has operational costs even when rarely invoked [23].

The business applications in the preceding literature also illustrate why the performance objective should be specified at the outset. A retention team may care about the cases that can reasonably benefit from outreach. An inspection team may instead care about how many faults are found per shift. The score can serve either objective only if the endpoint, horizon and capacity are aligned with the actual decision. Counting events under a fixed budget creates a common experimental outcome across contexts, but it leaves those application-specific choices to a subsequent study. It also assumes every captured event contributes equally, an assumption that would need examination wherever consequences differ substantially across cases.

The burden of obtaining a backup score can change the preferred design. A secondary source that already exists may add little marginal cost. A source that requires a new measurement, a paid request or more staff time competes with the review budget itself. The current comparison assumes that obtaining either score uses no review slots. This assumption isolates the ranking mechanism and makes backup-only scoring a meaningful comparator. In an application where the backup is expensive, a useful evaluation would record the number of backup requests as a separate resource outcome and compare policies using both staff time and scoring costs.

Cloud data protection provides a setting in which a quality signal can travel with a record as metadata [24]. A review of cloud-integrated information systems places data governance and privacy alongside defensive analytics [25]. AI and management information systems provide a context for joining scoring services to human review workflows [26]. Research on cybersecurity threat detection places the queue problem in the context of deciding which alerts receive analyst attention [27]. Strategic work on emerging cyber threats provides organizational context for the design of defensive review queues [28]. Work integrating threat intelligence and project management motivates making queue review a defined organizational responsibility [29]. Cybersecurity training research points to the role of operator behavior when flags are shown in a review interface [30]. Federated threat intelligence provides a distributed setting where score provenance and interoperability are important design considerations [31].

The cybersecurity literature contributes a parallel operational concern. Alert triage often takes place under uneven workloads, and a surge in questionable measurements can consume attention that would otherwise go to routine investigations. The simulation represents that problem as an error in a score before ranking. It neither specifies a cause for the error nor models how an incident unfolds. This abstraction keeps the comparison centered on defensive allocation and separates two issues that can be difficult to disentangle in operational logs: whether the score reflects the underlying event and whether an analyst has enough time to act on the cases selected for review.

Across the cited areas, the decision problem is more specific than the broad question of whether predictive analytics is useful. It concerns what to do when the preferred source becomes unreliable for some records and a secondary source remains available. The bibliography motivates the choice to measure ranking stability, provenance and operational capacity together. The numerical setup comes from the declared simulation assumptions. Consequently, the results can be interpreted on their own terms and do not depend on treating findings from one application domain as estimates for another.

3. METHODS

The simulation used NumPy's default random number generator with seed 7361. Each of 200 independent cohorts contained 4,000 records. For each record, a latent signal Z was drawn from a standard normal distribution. A binary event was generated from a Bernoulli distribution with probability logistic (−1.9 + Z). Primary and backup scores shared Z and added independent zero-mean Gaussian errors with standard deviations 0.35 and 0.85, respectively. This construction makes backup scoring less precise while preserving the scale on which the routing rule compares scores.

The event model makes outcomes more likely as the latent signal increases. It leaves substantial variation among records with similar signals because each outcome is realized independently from its probability conditional on that signal. Even a policy that knew the latent signal would therefore capture fewer than 600 events in 600 reviews. The simulation does not use the realized outcomes to build a score or set a threshold. It uses them after selection to measure how well the policy concentrated events in the review queue. This temporal separation is built into the experiment and is the equivalent of evaluating a frozen rule against outcomes observed later.

The two scores share their underlying signal so that, on average, both contain information about the same outcome. The primary score has the smaller measurement error on clean records. Its advantage follows from the construction, which gives a clear interpretation to the clean comparison. The backup retains the same relationship to latent risk at each corruption prevalence because only the primary score receives the extra shift. Backup performance is consequently constant across prevalence scenarios within each cohort. This constancy offers a check on the implementation and gives the results a stable comparison point.

Corruption added 2.5 score units to the primary value. Corruption prevalence ranged from 0% to 45%. Uniform random draws generated corruption and flag status independently of latent risk conditional on corruption status. The flag marked 80% of corrupted records and 5% of clean records in expectation. The same latent outcomes, base scores and uniform draws were used across policies and scenarios within a cohort. The design models a measurement defect affecting the primary score while the backup remains available.

Using the same uniform draw to assign corruption across the five scenarios creates nested sets of affected records. Every record corrupted in the 5% scenario is also corrupted in the 15% scenario for that cohort. The same is true across the higher prevalences. Nested assignments let the comparison follow a larger affected population while keeping the original affected records in view. Reusing the flag draw preserves comparability as well. A clean record receives a flag when its draw falls below the false-positive rate, and a corrupted record receives a flag when the same draw falls below the higher sensitivity. A record can therefore become flagged as the prevalence scenario changes and its primary score becomes corrupted.

The flag's practical meaning changes with prevalence even though its sensitivity and false-positive rate stay fixed. At 5% corruption, for example, the expected fraction of all records that are corrupted and flagged is 4.0%. The expected fraction that are clean and flagged is 4.75%. At 15% corruption, the corresponding proportions are 12.0% and

4.25%. These are expectations implied by the generator, not measured properties of a deployed quality check. They help explain why the same rule switches many clean records at low prevalence and sends a greater proportion of truly affected records to the backup as prevalence rises.

The design uses a positive shift because the principal mechanism of interest is crowding at the top of a queue. Subtracting from a score can also harm event capture, but it would do so by hiding affected records below the threshold. A mix of upward and downward errors would combine both mechanisms and could make the direction of the results harder to interpret. Holding the direction and magnitude fixed gives the prevalence comparison a specific meaning. As prevalence grows, an increasing number of records receive the same increment regardless of their original risk.

Table 1. Queue policies

Policy	Priority rule
Primary only	Rank everyone by primary score.
Backup only	Rank everyone by backup score.
Route flagged cases	Use backup when flagged and primary otherwise.
Defer flagged cases	Remove flagged records and rank remaining records by primary score.

Each policy selected at most the 600 highest-scored eligible records. Stable sorting resolved ties in input order. The primary outcome was the number of true events among selected records. Secondary outcomes were the fraction of all cohort events selected and the percentage of reviewed records with corrupted primary scores. Corruption in the review queue describes record provenance even when the ranking policy uses the backup.

Routing is applied before ranking. Each record receives exactly one score for the purpose of sorting. A flagged record uses the backup and an unflagged record uses the primary score. It does not receive an extra review slot or a reserved position in the queue. Deferral removes flagged records from consideration in the current allocation. The resulting queue may still include corrupted primary scores because a flag can miss a corruption event. No policy changes the review capacity to accommodate a larger number of flags. This puts the policies on the same nominal budget and keeps the comparison focused on their ordering and eligibility rules.

Event capture is an absolute count with an immediate operational interpretation. At a budget of 600, ten additional events mean the staff would encounter ten more events on average among the same number of reviewed records, provided the modeled outcome were observable at review. The proportion of cohort events captured adds context when the number of events varies from cohort to cohort. The corruption share describes the quality of the original measurement among selected records and remains meaningful even under backup scoring. It does not indicate how many selected records were ranked using their corrupted scores, since some entered through a backup path. Those interpretations are kept separate throughout the analysis.

We calculated policy means across cohorts and estimated the routing-minus-primary contrast within each cohort. A percentile bootstrap resampled the 200 cohort-level contrasts 3,000 times using seed 103. The reported intervals characterize Monte Carlo uncertainty in mean contrasts under this generator. They do not quantify uncertainty about the assumptions or about performance in any real organization.

The paired design reduces noise from comparisons between cohorts with different outcomes. A draw with many high-risk records can yield more events under every policy than a draw with fewer. Taking the difference within that cohort removes much of this common variation. For the bootstrap, the cohort is the independent sampling unit. Resampling records inside one selected queue would ignore the ranking constraint and would treat a slot as independent of the cases competing for it. Resampling entire cohorts

retains the policy comparison already computed under the same ranking budget and synthetic conditions.

The experiment reports pointwise intervals for routing versus primary-only scoring. Each interval concerns one specified corruption prevalence. The five scenarios share records within a cohort, so reading the intervals as independent replications would overstate the evidence. Likewise, the design does not estimate a continuous prevalence threshold at which one policy overtakes another. An observed ordering between two tested values can motivate closer investigation, but it leaves the shape of the transition unresolved. The tables show the complete set of evaluated scenarios so that the conclusions can remain tied to the tested grid.

4. RESULTS

Table 2. Mean events captured per 600 reviews

Corruption	Primary	Backup	Routing	Deferral	Routing minus primary [95% CI]
0%	237.9	205.9	235.9	234.8	-2.0 [-2.4, -1.6]
5%	214.3	205.9	230.2	227.5	15.9 [15.0, 16.8]
15%	171.5	205.9	217.9	213.6	46.3 [44.9, 47.8]
30%	166.8	205.9	201.4	193.9	34.7 [33.1, 36.3]
45%	187.3	205.9	187.8	173.1	0.4 [-1.0, 1.9]

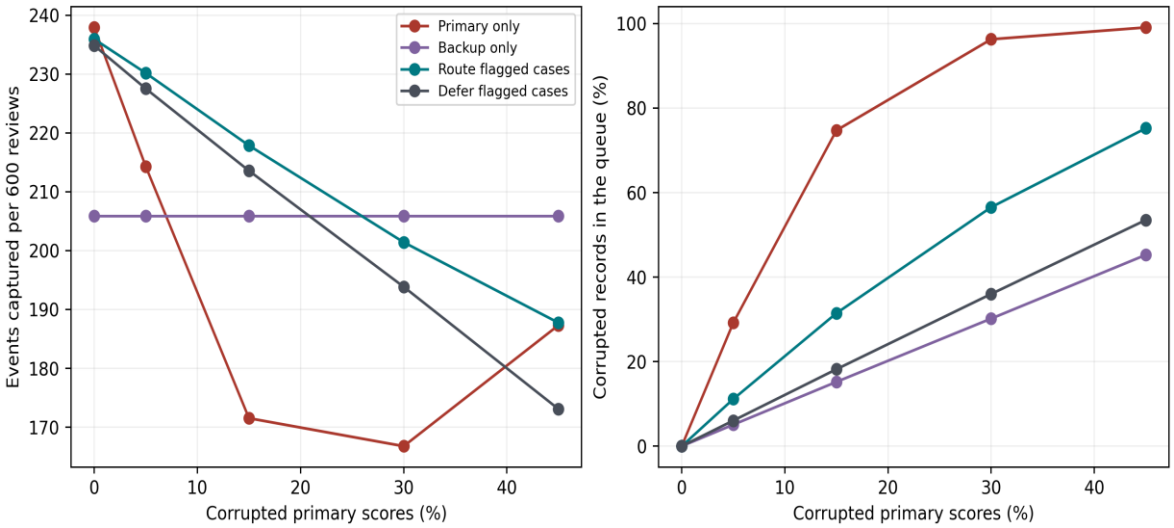


Figure 1. Event capture and queue composition by policy and corruption prevalence

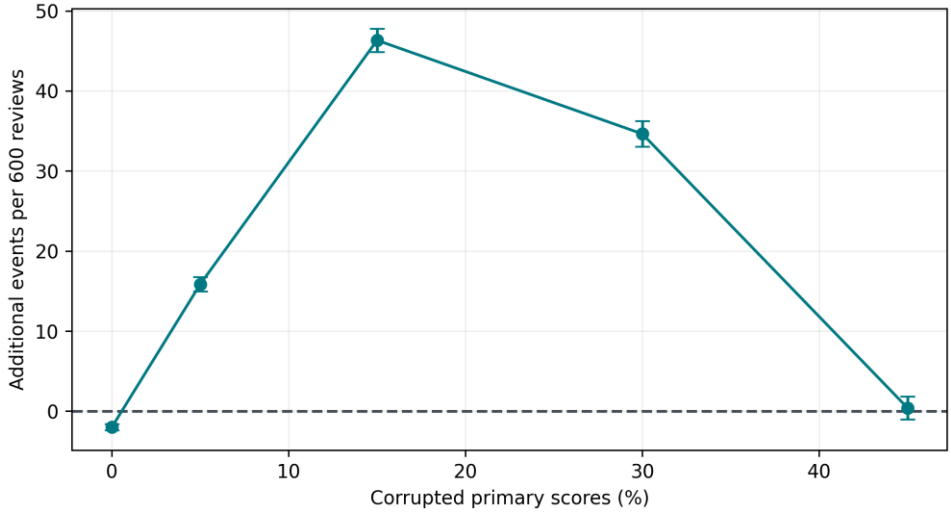


Figure 2. Paired effect of routing versus primary-only scoring. Bars show 95% bootstrap intervals

With clean inputs, primary scoring captured 237.9 events on average. Routing captured 235.9, reflecting the false flags that substituted a noisier score. At 30% corruption, primary-only scoring captured 166.8 events, compared with 201.4 under routing. The corresponding fractions of cohort events captured were 25.0% and 30.2%.

At 30% corruption, 96.3% of records reviewed under primary scoring had corrupted primary scores. The routing queue contained 56.5% such records. The composition measure must be interpreted against prevalence. A policy can capture more events without driving the corruption share to zero because genuinely high-risk records can also have corrupted measurements.

The 5% scenario shows the cost of concentrating errors near the top of a ranking. Only a small share of primary measurements is affected in the cohort, yet primary-only capture falls from 237.9 to 214.3 events on average. Routing recovers much of that loss, reaching 230.2. Backup-only scoring remains at 205.9, so it would sacrifice useful primary information for too many clean records in this particular scenario. The comparison is relevant when an organization suspects a limited defect that can be located with a reasonably informative quality check. It does not establish that routing will work equally well for smaller shifts or different error rates.

At 15% corruption, the routing contrast is largest among the evaluated scenarios. Its improvement over primary-only scoring is 46.3 events per cohort, based on the unrounded means. On a 600-slot budget, that corresponds to about 7.7 additional events per 100 slots. Deferral also improves capture relative to the primary-only ranking, but it yields 213.6 events compared with 217.9 under routing. The point estimates suggest that restoring eligibility with a backup can help when flagged records still include people or cases worth reviewing. The reported interval in Table 2 is for routing versus primary-only scoring, so it does not establish a corresponding interval for the difference between routing and deferral.

At the highest prevalence, the estimated routing gain falls to 0.4 events and its interval spans zero. The same scenario gives backup-only scoring an advantage in the descriptive comparison. This pattern helps distinguish two questions. A quality flag can improve the use of the primary score relative to ignoring the flag, while a globally stable backup can still yield a better queue once unresolved corruption becomes widespread. The best comparator therefore depends on the practical decision. For a team considering whether to retain the primary signal at all, the primary-only baseline gives only part of the answer.

5. DISCUSSION

Policy rankings changed with corruption prevalence. Routing captured more events than backup-only scoring at 15% corruption (217.9 versus 205.9), but backup-only scoring captured more at 30% (205.9 versus 201.4). At 45%, the two paths still differed, while routing and primary-only scoring had nearly the same mean. The rise in primary-only capture from 30% to 45% reflects the changing composition of the inflated pool. As it grows, competition among inflated records increasingly reflects their underlying score differences. It would therefore be misleading to extrapolate the routing gain monotonically from the lower corruption scenarios.

An inflated score creates a displacement effect. A record moves up the queue because its measured priority increased, even though its latent probability of an event stayed fixed. Routing reverses this inflation for records the flag catches. False negatives retain the inflated value and can still dominate the top of the ranking. False positives use a less informative score, which explains why routing can have a cost even when there is little or no primary corruption.

Deferral has a different consequence. It makes flagged records ineligible for the current review cycle and can therefore discard genuine events. Backup-only scoring, by contrast, preserves eligibility for everyone but gives up the extra resolution of the primary score on clean records. These mechanisms explain why event capture and corruption

share should be reported together. A low corruption share may reflect exclusion, while a high share may reflect exposure to unresolved score inflation.

The comparison assumes score commensurability. A production backup might produce scores on a different scale or with a different prediction horizon. Mixing unaligned scores could shift queue positions for reasons unrelated to quality. Before a prospective evaluation, both paths would need a common target and calibration procedure fitted on earlier data. Score source, timestamps, flag reason and policy version should travel with each queue entry so reviewers can reconstruct how priority was assigned.

Event capture measures the number of true events available for review, and the study does not estimate the effect of review on those events. Additional events detected might bring different benefits depending on case severity, intervention effectiveness and follow-up costs. A deployment evaluation would measure completion time and outcomes after review as well as capture. It would also compare the same policies within predeclared groups because flag errors could concentrate in a particular site or population.

For policy design, the highest capture in an individual scenario should be read together with the assumptions that make that scenario possible. A team rarely knows corruption prevalence at the time it must allocate reviews. The fraction flagged can be observed more easily, but prevalence cannot be inferred from that fraction without credible estimates of sensitivity and false positives. A change in the flag rate might reflect a real increase in measurement problems, a change in upstream reporting or a revision to the detection rule. Before using the results to define an operational switch between routing and backup-only scoring, an organization would need to validate the meaning of its flag under those conditions.

A useful pilot would freeze the four policy definitions in advance and score incoming batches without changing which cases staff actually review. The four queues could then be reconstructed on the same eligibility set. Reliable outcome ascertainment is essential, because cases outside the current queue may never receive confirmation. If only reviewed cases have labels, a simple retrospective comparison rewards policies that resemble the incumbent. A pilot could use an independently observed outcome where one is available. Alternatively, an approved audit sample drawn from records outside the current queue could help estimate missed events, provided the sampling design and any resulting weights were recorded.

Another issue is the time spent waiting. The model studies a single allocation window and counts what happens within it. In a repeated queue, deferred cases can return after a data correction, remain pending or leave eligibility altogether. Any evaluation of a continuing workflow should distinguish those paths. A policy that looks acceptable on one-day capture could create a backlog that raises waiting times for the same records every week. Conversely, a brief delay may be tolerable when reliable evidence is expected shortly and review is reversible. The study's fixed-window results identify the immediate trade-off and leave the longer-term consequences open.

Reporting should also distinguish a policy's mean performance from the distribution of its effects. If a fallback source performs well for one site but poorly for another, pooled capture can disguise a persistent loss for the site relying on weaker data. A practical study could report flag rates, review selection rates, capture and time to review within prespecified strata. Those comparisons would need sample sizes sufficient to characterize uncertainty. They would also benefit from a written account of how eligibility is determined, since records missing both scores can disappear before the ranking is even evaluated.

6. LIMITATIONS

The independent upward shift is only one type of corruption. Downward shifts can suppress high-risk records without making the queue look crowded by corrupted records. Drift in the backup, correlated failures,

changing review capacity, delayed labels and outcome-dependent corruption can alter the policy ordering. The experiment also keeps every review equally costly. Unequal case complexity would require a time budget or a constrained allocation model in place of a fixed record count.

The supplied bibliography spans several domains. Its application papers provide context for the design rather than supplying the synthetic parameter values. The results cannot establish empirical performance in healthcare, cybersecurity, energy or business settings. Publication metadata follow the supplied bibliography, and full methods and findings were not available for every reference.

The correlation structure between primary and backup errors is a particularly important boundary on interpretation. Both simulated scores depend on the same underlying risk, but their added measurement errors are independent. Real systems may share a data feed, preprocessing step or common identifier. In that case, the flagged record might carry the same defect into both scores. A useful sensitivity study would vary the extent of shared failure and compare results under different combinations of backup noise and flag accuracy. The present experiment holds those parameters fixed to isolate the prevalence comparison, so it cannot estimate how much independence is required for routing to retain its advantage.

The simulation also assumes that review can be allocated solely by score. Actual workflows may reserve slots for urgent referrals, geographic coverage or tasks requiring specialized staff. Those constraints change which records compete directly for a slot. They can also make the marginal cost of a mistaken priority depend on which reviewer handles the case. Extensions could evaluate multiple queues with separate budgets or use simulated review times to model differences in workload. Such analyses would answer a different operational question and should retain the simpler single-budget benchmark as a reference.

Finally, the simulated event is observed for all 4,000 records regardless of review selection. This avoids selective labeling in the benchmark but makes the evaluation easier than it would be in many applications. The narrow bootstrap intervals reflect repeated draws from a specified model with known outcomes. They do not reduce uncertainty about the underlying data-generating assumptions. Empirical performance would need independent evidence about label availability, the frequency and direction of errors, the staff budget and the response after detection. The magnitude of the simulated effects should be read as a property of the experimental setting.

7. CONCLUSION

When a primary risk score is susceptible to upward measurement corruption, a backup pathway can preserve access to review while reducing the displacement caused by suspect scores. Its value depends on the quality of the flag and the independent reliability of the backup. The paired simulation makes those assumptions explicit and offers an evaluation design that can be repeated with operational data.

The clearest practical lesson is that a quality flag and the policy attached to it should be evaluated separately. The flag identifies a record for which the preferred signal may be unreliable. The routing policy determines whether that information changes a score, eligibility, or both. A strong flag can still be paired with a poor response if the backup is weak or the cost of deferral is high. Conversely, a moderately informative flag can help preserve review capacity when the backup remains reliable and the primary score continues to add information for most of the cohort. This distinction gives an operational team two separate questions to investigate when a policy underperforms.

A replication with operational records should begin by specifying what counts as a review, when outcomes become observable and which records are eligible at the point of allocation. It should retain both scores and the flag for each record so that the four policies can be evaluated on comparable cohorts. The published benchmark supplies a

transparent starting point for that comparison. Its central result is conditional on a particular upward error and a stable backup. Across the tested scenarios, the preferred policy changed as more primary scores were affected, which makes local measurement of both failure rates and backup quality central to deciding how a flagged record should be handled.

REFERENCES

[1] M. Adnan *et al.*, "Interpretable Data Analytics for One-Year Diabetes Progression: Comparative Prediction, Feature Reliance, and Ranking Stability," *Eur. J. Med. Heal. Res.*, vol. 2, no. 6, pp. 257–269, 2024, doi: https://doi.org/10.59324/ejmhr.2024.2(6).34.

[2] E. Hossain *et al.*, "Cardiovascular Disease Burden and Clinical Associations Among US Adults Using Survey-Weighted Data Analytics," *Eur. J. Med. Heal. Res.*, vol. 2, no. 3, pp. 269–281, Jun. 2024, doi: https://doi.org/10.59324/ejmhr.2024.2(3).31.

[3] M. J. Hossain *et al.*, "Data-Driven Predictive Modeling for Early Diagnosis and Risk Stratification of Chronic Diseases," *Eastasouth J. Inf. Syst. Comput. Sci.*, vol. 2, no. 01, pp. 162–172, Aug. 2024, doi: https://doi.org/10.58812/esiscs.v2i01.1220.

[4] M. M. T. G. Manik, "Multi-Omics Integration with Machine Learning for Early Detection of Ischemic Stroke Through Biomarkers Discovery," *J. Ecohumanism*, vol. 2, no. 2, pp. 175 – 187, Jun. 2023, doi: https://doi.org/10.62754/joe.v2i2.6800.

[5] M. H. Rahman *et al.*, "Harnessing big data and predictive analytics for early detection and cost optimization in cancer care," *J. Comput. Sci. Technol. Stud.*, vol. 6, no. 5, pp. 278–293, 2024, doi: https://doi.org/10.32996/jcsts.2024.6.5.22.

[6] M. K. Ahmed *et al.*, "Big Data in Plant Biotechnology: Leveraging Bioinformatics to Discover Novel Anticancer Agents from Flora," *J. Med. Heal. Stud.*, vol. 4, no. 6, pp. 126–133, 2023, doi: https://doi.org/10.32996/jmhs.2023.4.6.15.

[7] S. I. Nilima *et al.*, "Advancement of Drug Discovery Using Artificial Intelligence and Machine Learning," in *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)*, 2024, pp. 1–7. doi: https://doi.org/10.1109/COMPAS60761.2024.10796748.

[8] M. I. Faruk *et al.*, "Global health intelligence platforms data science infrastructure for real-time epidemiological surveillance and pandemic forecasting," *Spectr. Eng. Sci.*, vol. 2, no. 5, pp. 749–761, Dec. 2024, [Online]. Available: https://thesesjournal.com/index.php/1/article/view/3425

[9] S. Hossain *et al.*, "Big Data Analysis and prediction of COVID-2019 Epidemic Using Machine Learning Models in Healthcare Sector," *J. Ecohumanism*, vol. 3, no. 8, pp. 14468 – 14477, Nov. 2024, doi: https://doi.org/10.62754/joe.v3i8.6775.

[10] M. H. Rahman *et al.*, "Scalable AI Models for Climate Change Mitigation Using Multisource Geospatial Big Data," *J. Comput. Anal. Appl.*, vol. 33, no. 08, pp. 5836–5856, Dec. 2024, [Online]. Available: https://eudoxuspress.com/index.php/pub/article/view/3431

[11] M. A. R. Chy *et al.*, "Understanding the relationship between data governance and business analytics success: A case study of global corporations," *J. Inf. Syst. Eng. Manag.*, vol. 9, no. 4s, pp. 4120–4136, 2024, doi: https://doi.org/10.52783/jisem.v9i4s.14807.

[12] P. Chakraborty *et al.*, "Toward autonomous decision intelligence: Integrating explainable AI and scalable DSS architectures in modern management information systems," *J. Inf. Syst. Eng. Manag.*, vol. 9, no. 4s, pp. 3893–3917, 2024, doi: https://doi.org/10.52783/jisem.v9i4s.14591.

[13] F. Mahmud *et al.*, "Big data and cloud computing in IT project management: A framework for enhancing performance and decision-making," *Fuel Cells Bull.*, vol. 2023, no. 9, 2023, doi: https://doi.org/10.52710/fcb.166.

[14] N. Das *et al.*, "Leveraging management information systems for agile project management in information technology: A comparative analysis of organizational success factors," *J. Bus. Manag. Stud.*, vol. 5, no. 3, pp. 161–168, 2023, doi: https://doi.org/10.32996/jbms.2023.5.3.17.

[15] K. B. Siddiqa *et al.*, "AI-Driven Project Management Systems: Enhancing IT Project Efficiency Through MIS Integration," in *2024 International Conference on Progressive Innovations in Intelligent Systems and Data Science (ICPIDS)*, 2024, pp. 114–119. doi: https://doi.org/10.1109/ICPIDS65698.2024.00027.

[16] M. A. Goffer *et al.*, "Leveraging predictive analytics in management information systems to enhance supply chain resilience and mitigate economic disruptions," *Educ. Adm. Theory Pr.*, vol. 30, no. 4, pp. 11134–11144, 2024, doi: https://doi.org/10.53555/kuey.v30i4.9641.

[17] S. Sultana *et al.*, "A Comparative Review of Machine Learning Algorithms in Supermarket Sales Forecasting with Big Data," *J. Ecohumanism*, vol. 3, no. 8, pp. 14457 – 14467, Nov. 2024, doi: https://doi.org/10.62754/joe.v3i8.6762.

[18] M. E. Hossin *et al.*, "Harnessing business analytics in management information systems to foster sustainable economic growth through smart manufacturing and Industry 4.0," *Educ. Adm. Theory Pr.*, vol. 30, no. 10, pp. 730–739, 2024, doi: https://doi.org/10.53555/kuey.v30i10.9643.

[19] U. Haldar *et al.*, "The Role of Artificial Intelligence in Enhancing Customer Experience in Business," *J. Comput. Anal. Appl.*, vol. 33, no. 08, pp. 7138–7155, Dec. 2024, [Online]. Available: https://eudoxuspress.com/index.php/pub/article/view/4068

[20] M. Hossain, M. M. T. G. Manik, A. Tiwari, J. Ferdousmou, N. Vanu, and A. Debnath, "Data Analytics for Improving Employee Retention in the U.S. Technology Sector," in *2024 International Conference on Intelligent Cybernetics*

Technology & Applications (ICICyTA), 2024, pp. 344–349. doi: https://doi.org/10.1109/ICICYTA64807.2024.10913216.

[21] F. Karim *et al.*, "Empowering Women through Microcredit: An Analysis of the Grameen Bank Model's Impact on Socio-Economic Outcomes," *Int. J. Comput. Exp. Sci. Eng*, vol. 10, no. 4, 2024, doi: https://doi.org/10.22399/ijcesen.4074.

[22] F. B. Khair *et al.*, "Sustainable Economic Growth Through Data Analytics: The Impact of Business Analytics on U.S. Energy Markets and Green Initiatives," in *2024 International Conference on Progressive Innovations in Intelligent Systems and Data Science (ICPIDS)*, 2024, pp. 108–113. doi: https://doi.org/10.1109/ICPIDS65698.2024.00026.

[23] M. D. Hossain, M. S. Uddin, M. S. Sikder, T. Hossen, B. Uddin, and R. M. Ahsan, "Green and secure data centers: Balancing energy efficiency with advanced cybersecurity measures," *J. Comput. Sci. Technol. Stud.*, vol. 6, no. 5, pp. 300–315, 2024, doi: https://doi.org/10.32996/jcsts.2024.6.5.24.

[24] P. Chakraborty *et al.*, "An Intelligent Cybersecurity Framework for Data Protection in Cloud Computing Environments," *Spectr. Eng. Sci.*, vol. 2, no. 3, pp. 657–677, Oct. 2024, doi: https://doi.org/10.5281/zenodo.21538159.

[25] K. S. Mudaliyar and S. S. Kumar, "Intelligent Cybersecurity Frameworks for Data Protection in Cloud-Integrated Management Information Systems," *Eastasouth J. Inf. Syst. Comput. Sci.*, vol. 2, no. 02, pp. 305–323, Dec. 2024, doi: https://doi.org/10.58812/esiscs.v2i02.1165.

[26] M. D. Hossain, M. S. Sikder, M. S. Uddin, R. M. Ahsan, B. Uddin, and T. Hossen, "Cognitive Cyber Defense: AI–MIS Integration through Big Data and Cloud Frameworks for Next-Generation Digital Resilience," *Eastasouth J. Inf. Syst. Comput. Sci.*, vol. 1, no. 02, pp. 140–152, Dec. 2023, doi: https://doi.org/10.58812/esiscs.v1i02.764.

[27] S. N. Hasan *et al.*, "Enhancing cybersecurity threat detection and response through big data analytics in management information systems," *Fuel Cells Bull*, vol. 2023, no. 12, 2023, doi: https://doi.org/10.52710/fcb.137.

[28] J. Kaur *et al.*, "Advanced cyber threats and cybersecurity innovation—Strategic approaches and emerging solutions," *J. Comput. Sci. Technol. Stud.*, vol. 5, no. 3, pp. 112–121, 2024, doi: https://doi.org/10.32996/jcsts.2023.5.3.9.

[29] S. M. Orthi *et al.*, "AI-Driven Cyber Threat Intelligence as a Management Information System: Integrating Cybersecurity Governance and IT Project Management for Organizational Resilience," *Eastasouth J. Inf. Syst. Comput. Sci.*, vol. 1, no. 02, pp. 194–213, Dec. 2023, doi: https://doi.org/10.58812/esiscs.v1i02.873.

[30] A. Shan-A-Alahi, M. Mustafizur, K. M. R. Hossan, A. Al Zaiem, and M. M. Rahman, "Cybersecurity training and its influence on employee behavior in business environments," *Comput. Fraud Secur.*, vol. 2024, no. 12, pp. 506–515, 2024, doi: https://doi.org/10.52710/cfs.689.

[31] M. K. Tuhin *et al.*, "Federated Machine Learning for Privacy-Preserving Cyber Threat Intelligence in Global Cloud-Integrated MIS Platforms," *Eastasouth J. Inf. Syst. Comput. Sci.*, vol. 2, no. 02, pp. 284–293, Dec. 2024, doi: https://doi.org/10.58812/esiscs.v2i02.1123.