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.

Review Capacity and Alert Prioritization in Pharmaceutical Field Management

TEVFIK TOLGA KAVUN

tevfik.tolga.kavun@ieee.org

Independent Researcher https://orcid.org/0009-0000-2073-0201

Research Article

Keywords: managerial attention, alert prioritization, simulation study, pharmaceutical field management, decision support, review capacity

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11848612/v1

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

Additional Declarations: The authors declare no competing interests.

Abstract

Background

A field-management dashboard can generate more possible exceptions than a manager can review. Ranking accuracy alone does not determine how many relevant cases will be examined when review capacity is limited.

Methods

A Monte Carlo experiment compared five selection policies across 12 prespecified design cells combining two regional prevalence profiles, two signal-separation levels and three review capacities. Each cell used 200 independent replications of 12 batches containing 160 synthetic tasks, with the same cases presented to every policy within a replication.

Results

With heterogeneous prevalence, moderate signal separation and 32 reviews per batch, score-only ranking selected 13.12 true exceptions on average, compared with 15.44 using correctly specified contextual probabilities and 8.04 using deliberately misassigned probabilities. The paired gain for correct context was 2.32 true exceptions per batch, with a Monte Carlo 95% interval of 2.23–2.40. Correct contextual ranking also concentrated coverage across regions: its mean regional recall range was 0.712 versus 0.144 for score-only ranking.

Conclusions

Under this constructed model, contextual information improved total detection when correct but reduced it when misassigned. Review allocation should therefore be evaluated jointly with capacity, contextual validity and regional coverage. These are computational findings from assumed distributions, not evidence of improved pharmaceutical sales, employee performance or clinical outcomes.

1 Introduction

An operational dashboard may identify many records that deserve attention while a manager has time to examine only a subset. The decision is then a capacity-constrained selection problem: which cases should enter the review set? In pharmaceutical commercial field management, possible applications include incomplete account plans, unresolved follow-up tasks or operating exceptions. These examples define an intended use context; the present study does not use records from a pharmaceutical company.

The distinction matters because a useful ranking can still leave many relevant cases unreviewed. In addition, incorporating regional context can change which areas receive attention even if every region has the same number of candidate cases. A policy that maximizes the number of true exceptions selected is not automatically a policy that provides similar recall across regions. Both objectives need explicit measurement.

Organizational decision-making research describes different ways of combining human and AI decisions [1]. A recent meta-analysis also finds that combining people and AI does not automatically outperform the better component alone [2]. The current experiment addresses a smaller question: how do a fixed review budget and correctly or incorrectly specified regional prevalence interact in a stylized selection process? It does not simulate human judgment, measure human–AI synergy or propose a new machine-learning architecture.

The contribution is an openly reproducible benchmark that contrasts detection yield, precision, recall and regional coverage under the same synthetic cases. Its design deliberately includes a favorable, correctly specified contextual policy and an unfavorable contextual policy. This makes the assumption behind any apparent benefit visible rather than treating context as inherently helpful.

2 Methods

The experiment is reported using the aims, data-generating mechanisms, estimands, methods and performance measures structure recommended for simulation studies [3]. The design, code and assumptions were fixed before inspecting the experiment's outputs. It was not preregistered. A supplementary archive contains the executable code, all replication-level metrics, aggregate results and the run manifest.

2.1 Synthetic tasks and design cells

Each batch contains 160 independent binary review tasks divided equally among four anonymous regions. A true exception is a task that warrants review under the simulated ground truth. It is not a poor-performing employee, a compliance violation or a patient outcome. Twelve independently generated batches form one replication. Cases do not persist across batches, and unreviewed tasks do not enter a backlog.

Two prevalence profiles are used. The homogeneous profile assigns probability 0.175 to every region. The heterogeneous profile assigns probabilities 0.05, 0.10, 0.20 and 0.35. Both profiles therefore contain 28 expected true exceptions per batch. These probabilities are analyst-selected stress-test assumptions, not estimates of regional business conditions. Equal regional case counts are likewise a simplifying assumption.

For each task i in region g, the exception indicator Y_i follows a Bernoulli distribution with probability p_g. Conditional on Y_i, the score Z_i follows a normal distribution with mean delta × Y_i and variance 1. The

separation parameter delta is either 1 or 2. The corresponding population score AUC values are approximately 0.760 and 0.921, obtained as Phi(delta / sqrt(2)); these values describe the assumed generator, not fitted model performance. No model training or parameter estimation is performed.

Review capacity k is set to 16, 32 or 64 tasks per batch, representing 10%, 20% or 40% of candidates. Combining the two prevalence profiles, two separation values and three capacities gives 12 design cells. Each cell contains 200 independent replications. The experiment consequently generates 4,608,000 synthetic tasks, each evaluated by the same five policies within its batch. Policy evaluations are paired rather than independent samples.

2.2 Selection policies

The random policy selects k tasks without replacement and provides a capacity-matched baseline. The score-only policy selects the k highest raw scores. The correctly specified context policy ranks tasks by their posterior log odds under the known generator: $L_i = \text{delta} \times Z_i - \text{delta}^2 / 2 + \log[p_g / (1 - p_g)]$. Because the transformation from log odds to probability is monotone, selecting the highest L_i is equivalent to selecting the highest posterior probabilities.

The equal-quota policy uses the same contextual ranking but selects k/4 tasks from each region. Within a region the prevalence term is constant, so this ordering is identical to raw-score ordering within that region. The misassigned-context policy uses the posterior expression with the regional prevalence vector rotated by two positions. In the heterogeneous setting it therefore uses 0.20, 0.35, 0.05 and 0.10 in place of the true regional probabilities. This is a deliberate misspecification stress test, not a measured rate or pattern of contextual error.

All policies receive identical labels and scores for a given batch and the same capacity. The context policy is an idealized correctly specified reference, not a realistically estimated model. All selected true exceptions are assumed to be recognized during review. Reviews require equal time and have no downstream effect on the next batch. These assumptions isolate selection from the quality of subsequent managerial action.

2.3 Outcomes and Monte Carlo uncertainty

The primary outcome is true exceptions selected per batch, averaged across the 12 batches of a replication. Precision is selected true exceptions divided by all selected tasks. Recall is selected true exceptions divided by all true exceptions. False reviews and missed exceptions are also reported per batch. Precision and recall are reported together because the prevalence of the target event affects the practical meaning of classification performance [4].

Regional recall is calculated after pooling each region's 12 batches within a replication. The regional recall range is the largest minus the smallest regional recall. It is a descriptive coverage measure, not a legal or ethical fairness verdict. Regions are anonymous operating groups rather than protected

demographic categories. The code records a missing regional recall if a denominator is zero; this is handled explicitly rather than assigning zero recall.

Summary values are means across independent replications. Monte Carlo standard errors are the replication standard deviations divided by sqrt(200). Paired policy differences subtract the score-only outcome within each replication; a descriptive 95% Monte Carlo interval is the mean paired difference plus or minus 1.96 times its Monte Carlo standard error. These intervals quantify simulation sampling error only. They do not express uncertainty about actual companies or the validity of the assumed probabilities.

2.4 Implementation and verification

The primary random seed is 20261123. NumPy SeedSequence creates independent streams indexed by design cell and replication. The code uses a fixed policy order, stable score sorting and shared cases within a batch. Verification checks confirm review counts, unique selected indices, the known posterior expression at the likelihood-ratio midpoint and identical rankings for score-only and correctly specified context when all regional prevalences are equal. The supplement states the software versions and includes instructions for reproducing the results.

Selective prediction research formalizes a trade-off between coverage and error when some decisions are withheld [5]. The present study uses a related resource constraint but does not implement SelectiveNet or abstaining prediction: tasks are ranked for review, and unselected tasks remain unreviewed. This distinction prevents an evaluation of review allocation from being interpreted as validation of a selective classifier.

3 Results

Every policy used its complete assigned capacity in every simulated batch. In the homogeneous prevalence cells, score-only and correctly specified contextual ranking selected exactly the same tasks. Their outcomes therefore matched, as expected from the monotone score transformation. This provides an implementation check rather than an empirical discovery.

Table 1 reports the heterogeneous-prevalence setting with capacity fixed at 32 reviews. Under moderate separation, correct context selected 15.44 true exceptions per batch compared with 13.12 for score-only ranking. The paired difference was 2.32, with Monte Carlo standard error 0.044 and a 95% Monte Carlo interval of 2.23–2.40. Misassigned context selected 8.04, a paired difference of − 5.08 relative to score-only ranking (interval − 5.17 to − 4.99).

Table 1
Policy outcomes with heterogeneous prevalence and 32 reviews per batch

δ	Policy	True selected	Precision	Recall	Regional recall range
1	**Random**	**5.58**	**0.174**	**0.200**	**0.113**
1	Score only	13.12	0.410	0.470	0.144
1	Correct context	15.44	0.482	0.553	0.712
1	Equal regional quota	11.97	0.374	0.429	0.190
1	Misassigned context	8.04	0.251	0.288	0.731
2	Random	5.60	0.175	0.199	0.109
2	Score only	20.56	0.643	0.733	0.128
2	Correct context	21.54	0.673	0.768	0.395
2	Equal regional quota	17.53	0.548	0.625	0.344
2	Misassigned context	18.10	0.566	0.645	0.390

Means across 200 independent replications. True selected is per batch. Regional recall range is calculated within each replication and then averaged.

With stronger score separation, score-only ranking selected 20.56 true exceptions per batch and correct context selected 21.54. The paired gain was 0.98 (Monte Carlo standard error 0.030; interval 0.92–1.04). Misassigned context selected 18.10, while equal quotas selected 17.53. The smaller gain from correct context in this setting is consistent with the raw signal already providing more separation between exception and non-exception tasks.

Correct context increased total recall but also increased the regional recall range. At moderate separation and capacity 32, the mean range was 0.712 with correct context and 0.144 with score-only ranking. Equal quotas produced a range of 0.190, showing that equal numbers of reviews are not the same as equal regional recall. At stronger separation the corresponding ranges were 0.395, 0.128 and 0.344. These differences arise in the constructed setting and do not identify which allocation objective an organization should adopt.

Table 2
Capacity sensitivity for correctly specified context

Delta	Reviews	True selected	Precision	Recall
1	16	9.65	0.603	0.344
1	32	15.44	0.482	0.553
1	64	22.39	0.350	0.798
2	16	13.83	0.865	0.493
2	32	21.54	0.673	0.768
2	64	26.35	0.412	0.943

Heterogeneous prevalence; 200 independent replications per cell. All values are simulation outputs.

Table 2 shows the capacity sensitivity for the correctly specified contextual policy. With moderate separation, increasing capacity from 16 to 64 raised mean recall from 0.344 to 0.798 while reducing precision from 0.603 to 0.350. With stronger separation, recall increased from 0.493 to 0.943 and precision fell from 0.865 to 0.412. Additional review capacity therefore found more true exceptions while admitting a less concentrated set of cases. Complete results for every design cell and policy are supplied rather than only the favorable comparisons.

4 Discussion

The experiment illustrates three separate decisions: how much review capacity to provide, how to rank tasks and how to distribute attention across regions. Under the known generator, ranking by correct posterior probabilities is expected to improve total detection relative to ignoring informative regional prevalence. The numerical results illustrate that familiar consequence at finite capacity. They should not be presented as discovery of a generally superior pharmaceutical management algorithm.

The misspecification comparison is equally important. A regional label adds value only through the information it represents. In this stress test, associating the wrong prevalence with a region displaced useful tasks from the review set. An organization would need evidence that contextual variables are measured correctly, appropriate to the decision and stable enough to use. The simulation does not establish how those checks should be performed or how often contextual estimates will be wrong in practice.

The coverage results also qualify the aggregate benefit. Correct contextual ranking concentrated review toward cases with larger modeled probabilities, producing a wider regional recall range. Whether that trade-off is acceptable depends on the purpose of the review process and consequences of missed tasks. Equal quotas were not a universal remedy because equal allocation of reviews did not produce

equal recall under unequal prevalence. Coverage constraints require a specified objective rather than a generic claim of fairness.

A prospective field evaluation would require validated exception labels, realistic review times, an explicit comparator and separate assessment of what happens after review. The NIST AI Risk Management Framework identifies validity, transparency and ongoing risk management as lifecycle concerns [6]. In this application, those concerns imply documenting the source of contextual estimates and measuring task selection separately from business outcomes. The simulated ground truth must not be replaced by unverified assumptions about employees.

The main limitations are deliberate. The score distributions are Gaussian with equal variance; true probabilities are fixed within a design cell; case counts are balanced; labels are independent; review is perfectly accurate; and false reviews have the same time cost as true reviews. There is no model-estimation error, concept drift, backlog, strategic behavior or employee response. The misassigned-context policy is a strong constructed error, not a calibrated model of organizational mistakes. The regional recall range is noisy at low prevalence and is not a complete account of distributional consequences.

The study is framed around pharmaceutical commercial management because that is the intended operational context. Its data-generating process is generic and contains no pharmaceutical-specific estimates. External validity to that industry, or to any actual company, remains untested. Further work should evaluate unequal review costs, persistent cases, estimated rather than known prevalence and alternative regional coverage constraints before drawing operational recommendations.

5 Conclusion

In a reproducible synthetic experiment, review capacity and regional information jointly affected the number and distribution of true exceptions selected. Correct contextual probabilities improved total detection within the assumed model; deliberately wrong probabilities reduced it. Higher capacity increased recall while lowering precision. These findings support evaluating review allocation with both aggregate and regional measures, while preserving a clear distinction between a computational benchmark and demonstrated field effectiveness.

Declarations

Ethics approval and consent to participate: Not applicable. The study used exclusively synthetic decision tasks and involved no human participants, patient records or company data.

Consent for publication: Not applicable; no identifiable individual data are included.

Availability of data and materials: The supplementary reproducibility archive contains the simulation code, replication-level outcomes, summary tables, paired comparisons and run manifest. The code

regenerates all synthetic tasks from the stated seed; no external dataset is required.

Competing interests: The author declares no competing interests.

Funding: No specific external funding was received for this work.

AI assistance: OpenAI ChatGPT assisted with study specification, code drafting, computational execution, reference checking and manuscript drafting. The reported numerical results were generated by the accompanying executable simulation, not supplied as invented observations. Final review, authorship and submission responsibility remain with the author.

Preprint status: This manuscript has not undergone peer review. No clinical, sales or employee-performance benefit has been established.

References

1. Shrestha YR, Ben-Menahem SM, von Krogh G (2019) Organizational Decision-Making Structures in the Age of Artificial Intelligence. Calif Manag Rev 61(4):66–83. https://doi.org/10.1177/0008125619862257
2. Vaccaro M, Almaatouq A, Malone T (2024) When combinations of humans and AI are useful: A systematic review and meta-analysis. Nat Hum Behav 8(12):2293–2303. https://doi.org/10.1038/s41562-024-02024-1
3. Morris TP, White IR, Crowther MJ (2019) Using simulation studies to evaluate statistical methods. Stat Med 38(11):2074–2102. https://doi.org/10.1002/sim.8086
4. Saito T, Rehmsmeier M (2015) The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS ONE 10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432
5. Geifman Y, El-Yaniv R (2019) SelectiveNet: A Deep Neural Network with an Integrated Reject Option. Proceedings of Machine Learning Research, 97, 2151–2159. https://proceedings.mlr.press/v97/geifman19a.html
6. Tabassi E (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Supplementary Files

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

- KavunResearchSquareSupplement.zip