AI-Assisted Test Prioritization:

A Simulation Study of Diagnostic Uncertainty Reduction Across Point-of-Care and Centralized Settings

Anonymous Author(s)
Bhome Gen Co., Ltd.
Seoul, Republic of Korea
lana@bhomegen.com

Abstract

Diagnostic decision-making often begins with symptom presentations that overlap across multiple etiologies and proceeds by selecting tests to reduce diagnostic uncertainty under practical constraints. While AI-driven triage and diagnostic decision-support systems are increasingly proposed to improve efficiency and access, their evaluation typically assumes a homogeneous testing environment, most often reflecting centralized laboratory settings. Here, we present a simulation framework that explicitly formalizes diagnostic decision-making across distinct testing regimes, including centralized testing and point-of-care testing (POCT), by incorporating regime-specific constraints on test availability and capacity. Using synthetic but biologically constrained scenarios derived from curated symptom–disease and test–disease mappings, we evaluate six diagnostic test-selection strategies spanning baseline, clinically motivated, and algorithmic-driven approaches. Diagnostic performance is quantified as the reduction in the number of diseases remaining in the differential diagnosis (ΔD), and regime-dependent effects are assessed using paired nonparametric analyses. We show that strategy performance differs systematically between regimes, that regime dependence is strategy-specific and capacity-dependent, and that identical decision policies can yield qualitatively different outcomes when deployed under constrained testing environments. These findings demonstrate that testing regime is a significant determinant of diagnostic decision-making and highlight the necessity of regime-aware evaluation for the development and deployment of AI-driven diagnostic support systems, particularly as healthcare delivery increasingly shifts toward decentralized and resource-limited settings.

1 Introduction

Diagnostic decision-making typically begins with a clinical presentation whose signs and symptoms are shared across multiple possible disease etiologies. This lack of specificity creates diagnostic uncertainty that clinicians seek to reduce through the judicious selection of tests. In practice, test selection is shaped not only by clinical reasoning but also by the operational constraints of the care setting. Centralized testing environments—encompassing off-site reference laboratories and higher-level hospitals with comprehensive in-house diagnostic infrastructures—support a broad array of modalities such as high-throughput molecular panels, culture assays, serology, sequential reflex testing, and advanced imaging (e.g., X-ray, CT, MRI). These infrastructures allow layered and expansive diagnostic strategies that progressively refine the differential diagnosis.

By contrast, Point-of-Care Testing (POCT) refers to diagnostic testing performed at or near the site of patient care and is characterized by rapid turnaround and simplified workflows designed to inform immediate clinical decisions without needing extensive laboratory infrastructure. POCT devices range

Submitted to 1st 2026AI Co-Scientist Challenge Korea. Do not distribute.

from portable immunoassay readers to handheld molecular platforms capable of same-encounter results, enabling clinicians to adjust treatments in real time rather than waiting hours or days for centralized laboratory results. POCT has been recognized for its potential to improve clinical or economic outcomes by facilitating earlier decision-making and appropriate treatment implementation, particularly where delays in centralized processing impede care delivery (Larkins et al., 2025).

Operationally, POCT settings impose constraints that are distinct from centralized settings: (i) narrower breadth of condition coverage per available test, (ii) strict limits on the number of tests permissible within an encounter, and (iii) exclusion of certain modalities such as traditional imaging and extended reagent-dependent assays (Larkins et al., 2025). These differences correspond to fundamentally different regimes of diagnostic feasibility that must be formalized when evaluating diagnostic processes and decision-support tools.

Clinicians naturally adapt their reasoning strategies to these structural realities. Common patterns include syndrome-anchored algorithms (e.g., the respiratory workup), conservative adherence to tested clinical pathways that prioritize high-yield investigations, and urgency-driven reasoning that expedites evaluation for life-threatening conditions (Naumann et al., 2023). These heuristics, while effective in many cases, are vulnerable to variation in execution, cognitive load, and inconsistency across clinicians and settings.

Software-based decision support—particularly systems built on clinical decision support systems (CDSS)—offers a structured approach to applying explicit mapping knowledge (linking symptom sets to differential diagnoses and to tests that cover them) under defined constraints. A high-impact overview of CDSS highlights these tools' aims to enhance healthcare delivery by matching patient characteristics to computerized clinical knowledge and presenting evidence-based recommendations at the point of care (Sutton et al., 2020). Moreover, AI-driven extensions of CDSS are increasingly designed to leverage machine learning and predictive analytics to improve diagnostic accuracy and workflow efficiency across settings.

In parallel, AI-based triage tools and symptom checkers have been proposed to support both patient-facing and clinician-facing decision workflows. Systematic evaluations illustrate that digital triage tools vary in diagnostic and urgency-classification accuracy, reflecting the challenges of automated reasoning under symptom ambiguity (Wallace et al., 2022). While generative AI models and LLM-based tools show promise in generating differential diagnoses, meta-analyses indicate that their overall diagnostic performance remains variable and often does not yet surpass expert human clinicians (Takita et al., 2025).

Despite extensive discussion of the potential of AI in diagnostic triage and clinical decision support, existing work often under-specifies the testing context—in particular, the operational constraints that fundamentally differentiate centralized and decentralized settings. Rigorous evaluation therefore requires explicit regime definitions that formalize (i) what tests and modalities are feasible in each setting and (ii) decision policies that correspond either to recognizable clinical workflows or to computational rules deployable in software.

Here, we present a simulation framework that formalizes (i) symptom-driven differential construction, (ii) constraint-aware test eligibility under different regimes, and (iii) quantitative metrics of diagnostic uncertainty reduction computed over disease states. Using synthetic yet biologically constrained scenarios derived from a curated symptom–disease matrix, we systematically compare diagnostic strategies across centralized and POCT regimes, quantifying differences via paired, nonparametric significance testing. This structured approach provides a regime-aware basis for evaluating decision-support strategies, addressing a key translational gap in understanding how computational tools perform across distinct diagnostic environments.

2 Methods

2.1 Symptom-disease and test-disease mappings

A symptom–disease matrix was constructed to encode clinically plausible associations between presenting symptoms and candidate diseases. Each disease was associated with a defined set of symptoms, allowing initial differentials to be generated by identifying all diseases consistent with a given symptom presentation. Separately, a test–disease mapping was defined to specify which diseases are covered by each diagnostic test. Test coverage was treated as binary and deterministic, reflecting whether a test is capable of ruling out or confirming a disease under idealized conditions.

Where relevant, construction of these mappings was informed by established clinical and regulatory guidance, including diagnostic recommendations and testing considerations published by the World

Health Organization and the U.S. Food and Drug Administration (World Health Organization, 2023). These sources were used to ensure that symptom groupings, disease inclusion, and test coverage assumptions reflected widely accepted diagnostic practice, while remaining abstracted and non-probabilistic for simulation purposes. All mappings were shared across strategies and regimes, ensuring that observed performance differences arose solely from decision logic and operational constraints rather than differences in underlying diagnostic knowledge.

2.2 Scenario generation

Diagnostic scenarios were generated synthetically by sampling symptom sets consistent with one or more diseases in the symptom–disease matrix. For each scenario, an initial differential diagnosis was constructed as the set of all diseases compatible with the sampled symptoms. Scenarios were designed to reflect realistic levels of symptom overlap across etiologies, resulting in nontrivial differentials that required testing to resolve. All strategies were evaluated on identical scenario sets, enabling paired comparisons across strategies and regimes.

2.3 Testing regimes and operational constraints

Two testing regimes were formalized: centralized testing and point-of-care testing (POCT). Centralized testing represents environments with access to broad diagnostic menus and the capacity to perform multiple sequential tests, including advanced laboratory assays and imaging-based diagnostics, as typically available in centralized laboratories or higher-level hospitals. POCT represents constrained environments in which only a limited number of tests can be performed and where individual tests typically cover a narrower subset of diseases.

Operational constraints were encoded by a single parameter, K, representing the maximum number of tests that could be selected for a given scenario. This parameter reflects testing capacity rather than monetary cost. Centralized regimes were evaluated at $K = 1,2,3$, while POCT regimes were evaluated at $K = 1,2$, consistent with realistic differences in testing infrastructure. Regime-specific constraints were implemented by restricting the set of eligible tests available to strategies at each decision step.

2.4 Diagnostic test-selection strategies

Six diagnostic test-selection strategies were evaluated. A random baseline selected eligible tests uniformly at random without using diagnostic information. A single-best strategy selected the test with the highest disease coverage at each decision step. An urgency(disease) strategy prioritized tests associated with clinically urgent diseases, whereas an urgency(test) strategy prioritized tests based on diagnostic efficiency under urgency constraints. A syndrome-best strategy first restricted eligible tests to those consistent with an inferred syndrome and then optimized test selection within that subset, while a syndrome-lock strategy restricted all test selections to the initially inferred syndrome throughout the decision process. Strategies differed only in how eligible tests were prioritized; all operated under identical regime constraints and shared symptom–disease and test–disease mappings.

2.5 Diagnostic uncertainty metric

Diagnostic performance was quantified using $\boldsymbol{\Delta D}$, defined as the reduction in the number of diseases remaining in the differential diagnosis after test selection. For each scenario, the initial differential size was computed from the symptom–disease mapping, and the final differential size was computed after applying the selected tests. The difference between these values constitutes $\boldsymbol{\Delta D}$, with higher values indicating greater reduction in diagnostic uncertainty. This metric was chosen because it is agnostic to disease prevalence and test ordering and is well suited for comparing strategies under formal operational constraints.

2.6 Statistical analysis

To assess regime-dependent effects, strategies were evaluated on identical scenarios across centralized and POCT regimes at matched values of $\boldsymbol{K}$. Paired differences in $\boldsymbol{\Delta D}$ were computed on a per-scenario basis. Absolute performance distributions were visualized using boxplots. Regime dependence was summarized by computing the median paired difference $\boldsymbol{\Delta D}_{\text{POCT}} - \boldsymbol{\Delta D}_{\text{Centralized}}$ for

each strategy and test capacity. Statistical significance of paired differences was assessed using two-sided Wilcoxon signed-rank tests. Cases in which all paired differences were zero were treated as exact regime invariance and reported separately. No correction for multiple testing was applied, as comparisons were pre-specified and limited in number.

2.7 Use of AI

AI-assisted tools were used extensively to support both the computational and manuscript-preparation aspects of this study. In addition to language editing and stylistic refinement, AI agents were employed to assist in writing and iteratively refining Python scripts used to collect and organize relevant external sources, construct the simulation and analysis pipeline, process scenario-level results, perform statistical analyses, and generate figures and tables. These tools enabled rapid prototyping, debugging, and scaling of analyses that would have been difficult to accomplish within practical time constraints using manual coding alone.

All methodological decisions, modeling assumptions, data analyses, and scientific interpretations were developed and validated by the authors. All code and outputs generated with AI assistance were reviewed, tested, and verified by the authors prior to use. Human oversight was maintained at all stages to ensure correctness, reproducibility, and alignment with the scientific objectives of the study. Without the use of AI-assisted tools, the scope, scale, and iterative evaluation of regime-dependent diagnostic strategies presented here would not have been feasible.

2.8 Implementation and reproducibility

All simulations and analyses were implemented in Python using standard scientific computing libraries. Experiments were performed using CPU-based execution on standard desktop computing resources, without reliance on specialized hardware or GPU acceleration, and typical runs completed within minutes. Scenario-level results were stored in structured comma-separated value files and used consistently across analyses, enabling straightforward replication of all reported results. The code used in this study is available from the corresponding author upon reasonable request.

3 Results

3.1 Formalizing diagnostic decision-making across testing regimes

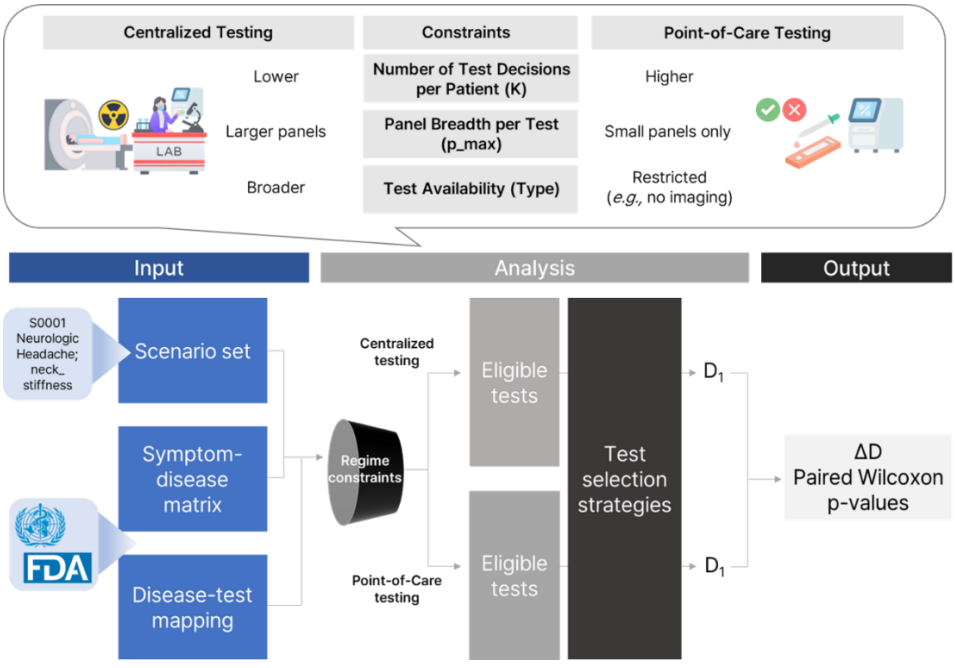

Figure 1: Schematic overview of the simulation framework and regime definitions

We first formalized diagnostic decision-making under two operationally distinct testing regimes: centralized testing and point-of-care testing (POCT) (Figure 1). Centralized testing encompasses environments with access to broad test menus and the capacity to perform multiple sequential tests, including advanced laboratory assays and imaging-based diagnostics. POCT represents constrained settings in which both the number of tests that can be performed and the breadth of disease coverage per test are limited.

Within each regime, diagnostic decision-making was modeled as an iterative process in which an initial symptom-driven differential diagnosis is refined through sequential test selection. Operational constraints were encoded by a single parameter, K, denoting the maximum number of tests that can be selected. Importantly, Kreflects testing capacity rather than monetary cost. Centralized regimes were evaluated at $K = 1,2,3$, while POCT regimes were evaluated at $K = 1,2$, consistent with realistic differences in testing infrastructure.

Diagnostic performance was quantified as ΔD, defined as the reduction in the number of diseases remaining in the differential diagnosis following test selection.

3.2 Evaluation of diagnostic strategies

Table 1: Overview of diagnostic strategies

Category	Strategy	Decision Principle
Baseline	Random	Randomly selects eligible tests without using diagnostic information
Clinical flow	Urgency(disease)	Prioritizes tests associated with clinically urgent diseases
	Syndrome-best	Selects tests consistent with an inferred syndrome, then optimizes within the set
	Syndrome-lock	Restricts all test selections to the initially inferred syndrome
Algorithmic selection	Single-best	Selects the test with the highest disease coverage
	Urgency(test)	Prioritizes tests based on diagnostic efficiency under urgency constraints

Six diagnostic test-selection strategies were evaluated within this framework and are summarized in Table 1. The strategies span baseline approaches, clinically motivated workflows, and optimization-driven policies, while operating on a shared underlying symptom–disease and test–disease mapping. Differences between strategies arise solely from how eligible tests are prioritized under regime-specific constraints.

This table provides a conceptual reference for the analyses that follow and clarifies the decision logic of each strategy independently of regime or performance outcomes.

3.3 Absolute uncertainty reduction across regimes and test capacities

We next examined the absolute reduction in diagnostic uncertainty achieved by each strategy across regimes and test capacities (Figure 2). For each value of K, distributions of ΔD are shown across simulated scenarios, allowing assessment of both central tendency and variability.

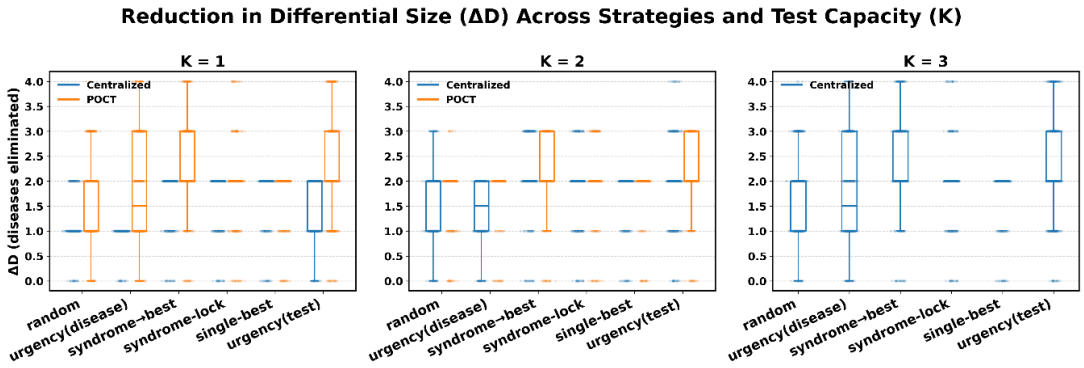

Figure 2: Strategy-dependent reduction in diagnostic uncertainty across test capacity

Across all strategies and regimes, increasing K led to monotonic improvements in uncertainty reduction. However, substantial differences were observed in both median performance and dispersion across strategies. Optimization-driven strategies generally achieved higher median ΔD, whereas clinically motivated strategies exhibited broader distributions, reflecting sensitivity to scenario composition and syndrome structure.

For matched values of K, POCT and centralized regimes frequently exhibited distinct ΔD distributions for the same strategy, indicating that regime constraints influence not only overall performance but also its variability. While these distributional differences are evident in Figure 2, their directionality and magnitude are not easily inferred from distributions alone.

3.4 Regime-dependent effects on diagnostic strategies

To directly quantify the impact of testing regime on diagnostic decision-making, we next analyzed paired differences in uncertainty reduction between POCT and centralized settings at matched values of K(Figure 3). For each strategy, the heatmap summarizes the median paired difference $\Delta D_{\text{POCT}} - \Delta D_{\text{Centralized}}$, with statistical significance assessed using paired Wilcoxon signed-rank tests.

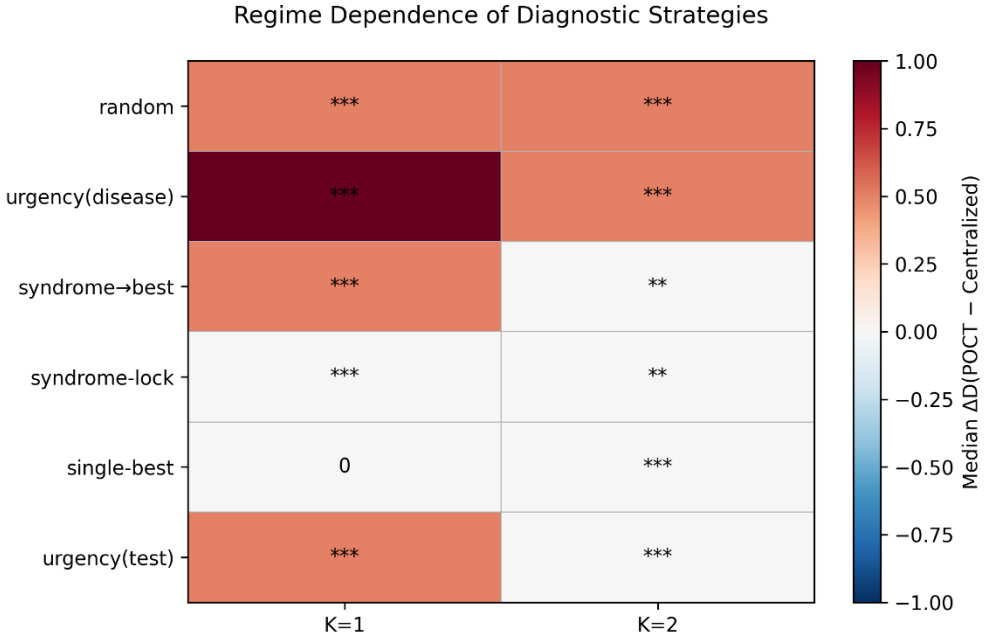

Figure 3: Regime-dependent differences in diagnostic uncertainty reduction across strategies

This analysis reveals pronounced heterogeneity in regime dependence across strategies. Several strategies exhibited statistically significant regime-dependent effects, particularly at low test capacity ($K = 1$), indicating that identical decision policies can yield systematically different diagnostic outcomes depending on the testing environment. In contrast, other strategies showed near-zero median differences, consistent with relative robustness to regime constraints.

Notably, regime dependence was strategy-specific and not strictly monotonic with respect to K. For some strategies, regime effects diminished as test capacity increased, whereas others remained sensitive even at higher capacity. These findings demonstrate that testing regimes do not merely scale diagnostic performance but can qualitatively alter the relative effectiveness of different decision strategies.

4 Discussion

This study demonstrates that the effectiveness of diagnostic decision strategies is not an intrinsic property of the strategy itself, but depends critically on the testing regime in which it is deployed. Even when strategies operate on identical symptom–disease and test–disease mappings, their ability to reduce diagnostic uncertainty differs between centralized and point-of-care testing (POCT) environments. This regime dependence is strategy-specific and varies with test capacity, indicating that operational constraints shape diagnostic decision-making and should be treated as significant components of evaluation frameworks.

Much of the existing literature on AI-driven triage and diagnostic optimization implicitly assumes a homogeneous testing environment, typically corresponding to centralized laboratory settings. Under this assumption, decision policies are evaluated without explicit consideration of how constraints on test availability, test breadth, or sequencing may alter their behavior. Our results challenge this view by showing that strategies optimized or assessed under centralized conditions do not necessarily generalize to POCT contexts. In constrained environments, differences in test menus and permissible decision pathways alter the structure of the decision space itself, leading to qualitative changes in strategy performance rather than simple rescaling of outcomes.

These findings have direct implications for the design and deployment of AI-based diagnostic decision-support systems. AI-driven triage tools are often framed as broadly universal solutions that can be applied across clinical settings. Our results suggest that such portability should not be assumed. Strategies that perform well in centralized regimes may exhibit reduced effectiveness, altered rankings, or increased variability when applied to POCT settings, particularly at low test capacity. Conversely, some strategies demonstrate relative robustness to regime constraints, while others benefit disproportionately from the structure imposed by POCT test menus. This heterogeneity underscores the need for regime-aware evaluation and, potentially, regime-specific optimization of decision policies.

A notable feature of our analysis is that regime dependence is not monotonic with respect to test capacity. For several strategies, regime effects are most pronounced when only a single test can be selected and diminish as additional tests become available. This pattern suggests that regime constraints interact with strategy logic in a nonlinear manner, especially in low-capacity settings where early test selection decisions dominate downstream outcomes. Importantly, these effects are not driven solely by reduced capacity in POCT environments. Rather, they reflect differences in the structure of the eligible test set, including the breadth of disease coverage and the degree of redundancy across tests. As a result, POCT regimes can, under certain strategies, achieve uncertainty reduction comparable to or exceeding that of centralized regimes at matched capacity, highlighting the importance of constraint-aware design over unconstrained expansion of test menus.

Several of the evaluated strategies were motivated by patterns commonly observed in clinical practice, such as syndrome-based workups and urgency-driven prioritization. The observed regime-dependent behavior of these strategies indicates that clinical reasoning heuristics, when formalized algorithmically, may interact with testing infrastructure in non-obvious ways. In particular, strategies that rigidly adhere to an initial syndrome tend to exhibit relative robustness to regime constraints, whereas strategies that rely on flexible prioritization are more sensitive to differences in test availability. These findings reinforce the importance of evaluating diagnostic decision-support tools in settings that reflect their intended clinical use. As POCT continues to expand across emergency departments, outpatient clinics, and resource-limited regions, evaluation frameworks developed for centralized laboratories may fail to capture critical performance characteristics in

these environments.

This study has several limitations. The analysis relies on synthetic scenarios generated from curated symptom–disease and test–disease mappings. While these mappings are biologically and clinically constrained, they do not capture all sources of uncertainty present in real-world clinical settings, such as imperfect test sensitivity, variable symptom reporting, or clinician-specific decision biases. In addition, uncertainty reduction was quantified solely in terms of the size of the remaining differential diagnosis. Although this metric is well suited to comparing strategies under formal constraints, future work could incorporate outcome-oriented measures such as time to correct diagnosis, downstream testing burden, or clinical risk mitigation. Extending the framework to include probabilistic test outcomes, longitudinal decision-making, and patient-level outcomes would further enhance its applicability.

As healthcare systems continue to shift toward decentralized diagnostic models, including POCT and near-patient testing, the conditions under which diagnostic decisions are made increasingly resemble those formalized in this study: symptom presentations with substantial etiological overlap, coupled with strict constraints on what tests can be performed and in what sequence. Our results indicate that these constraints do not merely limit diagnostic capacity but actively reshape the decision space in which diagnostic reasoning operates. Strategies that appear effective when evaluated under centralized assumptions may behave differently—or even qualitatively change their relative performance—when applied to constrained environments. By explicitly formalizing testing regimes and quantifying regime-dependent effects on diagnostic uncertainty reduction, this work provides a framework for evaluating diagnostic decision-support policies in a manner that is aligned with real-world deployment contexts. Such regime-aware evaluation is essential for ensuring that AI-driven triage and diagnostic tools are practically effective in the settings where diagnostic uncertainty is greatest and resources are most limited.

References

[1] Larkins, M. C., Zubair, M., & Thombare, A. (2025). Point-of-Care Testing. In *StatPearls*. StatPearls Publishing..

[2] Naumann, M., Scharfenberg, S. R., Seleznova, Y., Wein, B., Bruder, O., Stock, S., Simic, D., Scheckel, B., & Müller, D. (2023). Factors influencing adherence to clinical practice guidelines in patients with suspected chronic coronary syndrome: a qualitative interview study in the ambulatory care sector in Germany. *BMC health services research*, *23*(1), 655. https://doi.org/10.1186/s12913-023-09587-1

[3] Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). *An overview of clinical decision support systems: Benefits, risks, and strategies for success*. npj Digital Medicine, 3, Article 17. https://doi.org/10.1038/s41746-020-0221-y.

[4] Wallace, W., Chan, C., Chidambaram, S., Hanna, L., Pannu, D., Bickerdike, L., … Fraser, H. (2022). *The diagnostic and triage accuracy of digital and online symptom checker tools: A systematic review*. npj Digital Medicine, 5, Article 118. https://doi.org/10.1038/s41746-022-00667-w

[5] Takita, H., Kabata, D., Walston, S. L., Ito, S., Yamashita, R., Kido, S., … Komatsu, M. (2025). *A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians*. npj Digital Medicine, 8, Article 175. https://doi.org/10.1038/s41746-025-01543-z

[6] World Health Organization. (2023). *The selection and use of essential in vitro diagnostics: Report of the fourth meeting of the WHO Strategic Advisory Group of Experts on In Vitro Diagnostics, 2022 (including the fourth WHO model list of essential in vitro diagnostics)* (WHO Technical Report Series No. 1053). World Health Organization. https://www.who.int/publications/i/item/9789240074553

AI Co-Scientist Challenge Korea Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: Yes

   Justification: The abstract and introduction accurately reflect the scope and contributions of the study by focusing on regime-dependent evaluation of diagnostic decision strategies under centralized and POCT constraints, without making claims beyond simulation-based analysis. The claims are directly supported by the simulation framework and results presented in the *Methods* section and *Results* (Figures 1–3), with implications contextualized in the *Discussion*.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: Yes

   Justification: The limitations of the study are explicitly discussed in the *Discussion*, including the use of synthetic scenarios, idealized test coverage assumptions, and the focus on uncertainty reduction rather than clinical outcomes, with directions for future extensions clearly outlined.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: Yes

   Justification: All theoretical results are derived from explicitly stated modeling assumptions, including the construction of symptom-driven differentials, regime-specific test eligibility, and the definition of the uncertainty reduction metric (ΔD), as detailed in the *Methods*. The correctness and completeness of these results are supported through formal definitions and systematic simulation-based evaluation described in the *Methods* and demonstrated in the *Results* (Figures 2–3).

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: YES

   Justification: The paper fully specifies the simulation framework, regime definitions, diagnostic strategies, evaluation metric (ΔD), and statistical analyses in the *Methods*, providing all information necessary to reproduce the main experimental results that support the paper's claims. Additionally, data generation procedures and implementation details are described in the *Methods.*

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While AI Co-Scientist Challenge Korea does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: No

   Justification: The entire code will be available from the corresponding author upon reasonable request.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: Yes

   Justification: Although the study does not involve model training or learning-based optimization, all details necessary to understand the results—including scenario generation, strategy definitions, regime constraints, evaluation metrics, and statistical testing procedures—are fully

specified in the *Methods*, ensuring transparency comparable to training and testing disclosures in learning-based studies.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: Yes

   Justification: The paper reports uncertainty and variability using distributional summaries (boxplots) and quantifies statistical significance of paired differences using two-sided Wilcoxon signed-rank tests, as described in the *Methods* and presented in the *Results* (Figures 2–3), with clear definitions of the underlying metrics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: Yes

   Justification: The computational requirements for reproducing the experiments are modest and are described in the *Methods* under *Implementation and reproducibility*, including the use of standard Python-based analysis on commodity hardware, without reliance on specialized accelerators or large-scale computing resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://nips.cc/public/EthicsGuidelines?

   Answer: Yes

   Justification: The research conforms to the NeurIPS Code of Ethics by relying exclusively on synthetic, non-identifiable data and by clearly disclosing assumptions, limitations, and appropriate use cases, thereby avoiding risks related to privacy, misuse, or harm. Ethical considerations and responsible scope are addressed in the *Methods* (Use of AI-assisted tools) and reinforced in the *Discussion*, where limitations and deployment implications are explicitly stated.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: Yes

    Justification: The paper discusses potential positive societal impacts, such as improved access to diagnostics and more appropriate deployment of decision-support tools in constrained settings, as well as potential negative impacts and risks of misapplication, including overgeneralization across testing regimes. These considerations are addressed in the *Introduction* and expanded in the *Discussion*, particularly in the sections on regime dependence, limitations, and deployment implications.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: NA

    Justification: [TODO]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: NA

    Justification: [TODO]

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: NA

    Justification: [TODO]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their

submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: NA

    Justification: [TODO]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: NA

    Justification: [TODO]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.