International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

EVIDENCE QUALITY IN URBAN SEARCH AND RESCUE READINESS ASSESSMENT: REGISTRY AMBIGUITY, OPERATIONAL INFLUENCE, AND COMPLIANCE-RECORD ERROR

Mohammad Syafii[1]; Ma'mun Murod[2]; Andriansyah, Andriansyah[3]
[1] Doctoral Program in Public Administration, Faculty of Social and Political Sciences, Universitas Muhammadiyah Jakarta, Jalan K.H. Ahmad Dahlan, Cireundeu, East Ciputat District, South Tangerang City, Banten 15419, Indonesia.
[2] Faculty of Social and Political Sciences, Universitas Muhammadiyah Jakarta, Jalan K.H. Ahmad Dahlan, Cireundeu, East Ciputat District, South Tangerang City, Banten 15419, Indonesia.
[3] Faculty of Social and Political Sciences, Universitas Muhammadiyah Jakarta, Jalan K.H. Ahmad Dahlan, Cireundeu, East Ciputat District, South Tangerang City, Banten 15419, Indonesia.
Email: mohammad.syafii@umj.ac.id[1]

Abstract: Readiness judgments can be misleading when classification labels, operational endpoints, and requirement records are treated as interchangeable evidence. We assess these boundaries using 61 first INSARAG classifications in 48 countries, 43 BASARNAS offices, descriptive national indicators, and controlled simulations. Exact Brier-loss envelopes and exhaustive refitting cover four disputed labels. Under equal team weights, ridge improvement over prevalence ranges from −0.02841 to +0.00348; under equal country evaluation weights, the same predictions remain favorable across all 16 assignments. Robustness is therefore estimand-dependent. The office training coefficient is imprecise and reverses sign after removing one office. Source checks confirm that its influential annual means agree with published totals and that the endpoint measures readiness to depart, not incident arrival. A descriptive category decomposition localizes the recorded increase without identifying its cause. In 1600 paired synthetic experiments, exact gate features reduce hard-gate ridge RMSE at n = 27 from 0.2223 to 0.0688. An independent 800-experiment recording-error sweep yields RMSE 0.0511 with exact gates and 0.1526 with 10% per-gate error at n = 240. The contribution is a reproducible evidence audit with explicit interpretation boundaries, not a validated certification model or a causal estimate of training effectiveness.

Keywords (English): *Urban search and rescue; evidence quality; emergency preparedness; INSARAG; administrative data; outcome-label uncertainty*

INTRODUCTION

Urban search and rescue (USAR) readiness is a safety-relevant judgment about what a deployable team can do, under which conditions, and with what supporting evidence. INSARAG external classification evaluates a specified configuration of management, search, rescue, medical, and logistics capacity. Medium and Heavy classifications represent different operational capacities; they are not interchangeable with a universal ranking of team quality, survival benefit, or effectiveness in every incident (Switzerland, 2020). A predictive model of a team's first classification therefore answers a narrower question than either a study of progression from Medium to Heavy or an evaluation of rescue outcomes. (Bortolin, 2024).

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

Urban search and rescue (USAR) readiness is a safety-relevant judgment concerning the extent to which a deployable team is capable of performing assigned functions, the operational conditions under which those functions can be carried out, and the quality of evidence available to demonstrate that the required capacities have been established. Readiness therefore should not be interpreted simply as a general indication that a team is "good" or "bad," but rather as an assessment of whether a particular organizational and operational configuration is prepared to undertake specified search-and-rescue responsibilities. Within this context, the INSARAG external classification process evaluates a defined configuration of management, search, rescue, medical, and logistics capacities, each of which contributes to the overall ability of a team to operate within the scope of the applicable classification framework.

The distinction between Medium and Heavy classifications is particularly important because these categories represent different levels and configurations of operational capability rather than a universal hierarchy of organizational quality. Consequently, a Heavy classification should not automatically be interpreted as proof that a team will always achieve better outcomes than a Medium-classified team in every emergency setting, since classification status does not by itself provide a universal ranking of team quality, survival benefit, deployment effectiveness (Bortolin, 2024), or operational success across all possible incidents

This distinction is also relevant when classification data are used for statistical or predictive analysis. A predictive model designed to estimate a team's first INSARAG classification is addressing a relatively specific analytical question: whether available characteristics or contextual indicators are associated with the classification initially recorded for participating teams. A study of progression from Medium to Heavy would instead need to examine changes in capacity and organizational development over time, whereas an evaluation of rescue outcomes would require evidence that more directly reflects operational performance in specific incidents (Alabede, 2023). Maintaining these analytical boundaries is therefore essential to prevent a predictive classification model from being interpreted more broadly than the evidence it was designed to address

LITERATURE REVIEW

Research on classification has already documented organizational preparation and operational constraints. The Japanese experience examined capacity development through classification and reclassification, while the BASARNAS case described preparation, mentoring, and institutional coordination (Okita, Sugita, Katsube, & Minato, 2018). Evidence from the 2015 Nepal earthquake connected classification to international deployment and acceptance decisions rather than treating accreditation as sufficient proof of timely arrival (Okita & Shaw, Effects of the classification system in international SAR: The case of the 2015 Nepal earthquake, 2019). The Beirut case described adaptation of USAR roles under pandemic-related constraints. Together, these studies motivate attention to the evidence used in readiness judgments; they do not provide a validated numerical readiness function. Table 1 specifies how the present analysis differs without claiming that classification, coordination, or capacity development are previously unstudied subjects (Okita, Glassey, & Shaw, COVID-19 and the expanding role of international urban search and rescue (USAR) teams: The case of the 2020 Beirut explosions, 2022).

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

Table 1 Relationship to selected empirical USAR studies; this is a focused comparison, not a systematic review.

Study and setting	Evidence or focus	Question addressed here
Okita et al. (Okita, Sugita, Katsube, & Minato, 2018); Japan	Classification/reclassification and team capacity development.	How sensitive is a predictive comparison to unresolved outcome records?
Okita and Shaw (Okita & Shaw, Search and rescue team classification: BASARNAS of Indonesia, 2020); BASARNAS	Organizational preparation, mentoring, and coordination.	Can coarse administrative proxies support team-level readiness claims?
Okita and Shaw (Okita & Shaw, Effects of the classification system in international SAR: The case of the 2015 Nepal earthquake, 2019); Nepal	Classification in international deployment and acceptance.	How should classification prediction be separated from operational performance?
Okita et al. (Okita, Glassey, & Shaw, COVID-19 and the expanding role of international urban search and rescue (USAR) teams: The case of the 2020 Beirut explosions, 2022); Beirut	Operational adaptation and expanded roles in a constrained deployment.	Which conclusions remain conditional on mission and measurement definitions?

Source: Compiled by the author based on field observations, 2026.

Three distinctions guide the study. Construct validity concerns whether indicators measure an intended capability; predictive performance concerns loss on held-out observations; causal identification concerns the consequences of an intervention (Cronbach & Meehl, 1955). None follows automatically from either of the others. A fourth distinction is especially important in safety assessment: evidence that a model predicts an administrative label is not evidence that it can authorize deployment or verify a mandatory requirement. Misclassifying this boundary could encourage unjustified assurance even when a statistical analysis is internally consistent (Shmueli, 2010).

The practical gap addressed here is a traceable assessment of how different evidence failures alter conclusions drawn from the same readiness-related records. Source disagreement can change both training outcomes and scoring labels. Aggregation can disconnect a training exposure from the personnel and tasks of interest. An influential office can dominate a squared-error summary. In a simulation, a learner given the generator's exact threshold functions can appear superior because of its representation, not because its learning algorithm is generally better. These problems require different diagnostics rather than a single composite uncertainty score.

Grouped validation, proper scoring rules, and specification sensitivity provide established tools for this purpose (Roberts et al., 2017). We combine them with exhaustive outcome-label refitting (Gneiting & Raftery, 2007), an explicit comparison of team- and country-weighted loss,

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

a source-level examination of an influential office, and paired synthetic experiments. The applied contribution is the connection between a diagnosed evidence limitation and a narrower, defensible interpretation. The exact loss identity and the descriptive decomposition below are transparent algebraic tools, not claims of a new universal statistical method (Simonsohn, Simmons, & Nelson, 2020).

The study asks four questions. RQ1: Is predictive improvement over a simple baseline invariant to the documented label disagreements, and does the answer depend on the unit given equal weight? RQ2: Does a recorded office-level training count add stable information about next-year administrative mobilization-response time, and what do the source totals reveal about the most influential office? RQ3: How much of the simulated hard-gate advantage is associated with supplying exact gate features rather than changing the regression algorithm? RQ4: How do recording error and training sample size jointly affect the selected gate-augmented models? The empirical and simulated answers are kept separate throughout.

RESEARCH METHODS

Evidence architecture and the boundary of a readiness claim

The organizing framework retains five proposed capability domains: Leadership and Organizational Learning (L), Human Resource Capacity (H), Coordination and Interoperability (C), Institutional Commitment and Stability (I), and Facilities and Logistics Readiness (F). Collaborative-governance research motivates attention to leadership, institutional arrangements, and capacity for joint action (Ansell & Gash, 2008), but does not establish this particular five-domain measurement model. No coefficient for a direct L–H–C–I–F readiness score is estimated from the country or office records. Appendix A retains the conceptual maturation function, its mathematical interpretation, and a proposed direct-measurement matrix (Emerson, Nabatchi, & Balogh, 2012).

The analysis uses four evidence layers (Table 2). Their units, outcomes, and exposures differ, so they are neither pooled nor treated as independent validations of one latent readiness score. The first-IEC analysis concerns participating teams and national context; the office analysis concerns administrative mobilization across the full SAR portfolio; national indicators provide organizational description; simulation concerns explicitly constructed continuous responses. A favorable result in one layer cannot repair an invalid interpretation in another.

Table 2 Evidence layers, estimands, and limits on interpretation.

Layer	Unit and sample	Estimand or diagnostic	Interpretation boundary
Registry prediction	61 first IECs; 48 countries; 2005–2024	Paired held-out Brier loss under disputed labels; team/country weighting.	First classification among participants, not progression or general-population readiness.
Office operations	43 matched offices; 2023–2024	Training coefficient, deletion sensitivity, and out-of-office prediction.	All-SAR administrative mobilization, not a causal training effect or collapse-rescue effectiveness.

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

Layer	Unit and sample	Estimand or diagnostic	Interpretation boundary
National indicators	Five series; five years each	Original-unit descriptions and aggregation checks.	No trend inference, latent-index validation, or causal triangulation.
Controlled simulation	Four generators; paired models and feature sets	Independent-test RMSE and paired contrasts with Monte Carlo error.	Behavior under stated generators, not an empirical effect in rescue teams.

Source: Compiled by the author based on field observations, 2026.

Source provenance, targeted verification, and analysis history

The processed archive was constructed from two official INSARAG registry pages (INSARAG, 2005), the Worldwide Governance Indicators (WGI) 2025 Revision and its documentation (World Bank, 2026), and World Development Indicators (WDI). We distinguish byte integrity, computational reproducibility, and verification against original sources. These are separate checks: a correct hash does not establish a correct transcription, and reproducing an analysis does not establish that its estimand is appropriate (Olympia, 2025).

Fifty recovered archive files matched their original SHA-256 manifest. The present execution reconstructed 180 archived cross-validation Brier scores from per-team predictions, reran the exhaustive label and matched-feature analyses, and recalculated office regression and prediction diagnostics. Original BASARNAS source PDFs and World Bank workbooks were not recovered inside that archive. Accordingly, the supplement contains processed inputs and source locators, not a claim that the entire original acquisition pipeline has been independently reconstructed.

Targeted current-source checks on 16 September 2026 examined the four registry disagreements, the BASARNAS definition of response time, the 2023 and 2024 Manado rows and their category totals, and the internally conflicting 2024 certified-HR indicator. The source locators specify both PDF and printed page numbers to avoid ambiguity. These checks are narrower than re-extracting all office outcomes, training rows, and historical covariates. A historical WGI API-scale problem remains an archived acquisition warning; the former API behavior was not re-enacted in this revision.

All analyses are exploratory and were not preregistered. The initial archive used seed 20260718. Exhaustive refitting and matched-feature experiments used seed 20260909 and were rerun here with the same settings. The current weighting summaries, additional absolute-error reporting, and targeted Manado decomposition are retrospective diagnostics. The recording-error sweep reported in Section 3.7 is a newly executed, independent simulation run using seed 20260916; it is not presented as a recovery of the earlier seed-20260910 run. This chronology prevents retrospective refinements from being described as confirmatory tests.

Requirement interpretation must also preserve the applicable standard. The current INSARAG guidance page specifies that the newly issued guidelines take effect on 1 January 2027 and retains the 2020 edition for reference (INSARAG, 2027). That announcement does not make a later checklist retroactively applicable to every 2005–2024 classification. The empirical outcome remains the historical recorded classification, not a reassessment against a contemporary checklist.

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

First-IEC sample, covariates, and validation protocols

The sample includes each team's first recorded Medium or Heavy IEC during 2005–2024. Later reclassification events and Light classifications are excluded by definition. Detailed-page coding yields 36 Heavy and 25 Medium outcomes; updated-page coding yields 38 Heavy and 23 Medium outcomes. Four first-IEC levels disagree between the pages: GER-02 (2007), UAE-01 (2009), ALG-01 (2017), and CHI-01 (2017). Detailed coding was the archived reference because of its event detail, not because a registry owner had adjudicated its correctness. Excluding the four records gives a separate 57-team sample and changes the evaluated population.

All covariate reference years are no later than IEC year minus one. The context logistic model uses Government Effectiveness, Control of Corruption, and Rule of Law estimates, with median imputation, standardization, and a single principal component fitted exclusively on the training fold. Principal component analysis is dimension reduction, not construct validation (Jolliffe & Cadima, 2016). The full ridge logistic model instead uses eight predictors directly: the three governance estimates, Political Stability, log GDP per capita at purchasing-power parity, urban population share, log total population, and IEC year. Its pipeline uses training-fold imputation and standardization, without replacing the governance variables by the context component.

No values are missing in those processed predictor columns, although the imputation steps are retained in the executable pipelines. The context model uses logistic regularization parameter C = 10; full ridge uses C = 1. Both use L-BFGS and a 5000-iteration limit. Hyperparameters are retained from the archive and were not optimized on the evaluation observations. The baseline assigns every held-out team the Heavy prevalence in its training fold. It must be recomputed when training labels change.

Temporal lagging addresses reference-year order but not historical release availability. The WGI revision recalculates earlier years, and WDI values can be revised. Moreover, grouped random folds allow later classification years to occur in the training set of an earlier test observation. The analysis therefore estimates retrospective transfer to held-out countries within the assembled sample. It is neither a historical deployment backtest nor a temporal forecast. A prospective evaluation would require contemporaneous releases, a specified assessment cohort, and an appropriate time split.

The archived benchmark used outcome-stratified, country-grouped five-fold cross-validation repeated ten times. Its selected results are retained in Appendix C for traceability. Exhaustive refitting uses a different design: sorted country identifiers are randomly permuted with seed 20260909 plus the repetition index and partitioned into five groups. The resulting country folds do not depend on outcome labels and are reused across all 16 assignments and both fitted models. There are no country overlaps between training and test partitions, and every training partition contains both outcome classes. The procedure fits 1600 logistic models, in addition to training-prevalence baselines.

Loss estimands and the ambiguity set

Brier loss is the primary predictive measure (Brier, 1950). Let y_i be a binary label, p_{ir} a candidate prediction, and b_{ir} a prevalence-baseline prediction for team i in repetition r. The paired loss difference is negative when the candidate is better (Gneiting & Raftery, 2007). With non-negative weights summing to one, the repetition-averaged comparison is:

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

$$d_{ir}(y_i) = (p_{ir} - y_i)^2 - (b_{ir} - y_i)^2 = p_{ir}^2 - b_{ir}^2 - 2y_i(p_{ir} - b_{ir}). \quad \textbf{(1)}$$

$$\Delta_w(y) = \sum_{i=1}^{n} w_i \bar{d}_i(y_i), \qquad \bar{d}_i(y_i) = \frac{1}{R}\sum_{r=1}^{R} d_{ir}(y_i). \quad (2)$$

The primary estimand assigns equal weight to every team, $w_i = 1/61$. The country sensitivity first averages team losses within each country and then averages the 48 country means, giving $w_i = 1/(48n_c)$ for a team in country c with n_c sampled teams. Predictions, fitted models, folds, and the team-prevalence baseline are unchanged in this sensitivity. Thus, it changes evaluation weights, not training weights or the baseline definition. It should not be interpreted as a comparison against a separately optimized, country-weighted baseline.

Let U contain the four disputed records. The ambiguity set contains all $2^4 = 16$ assignments of their binary outcomes while holding other labels fixed. We assume the two values for each disputed field are independently admissible. Only the all-detailed and all-updated assignments are directly observed page-wide codings; mixed assignments are sensitivity scenarios, not additional official registries. If an adjudicator establishes cross-record constraints, the admissible set should be restricted accordingly. No probability weights are assigned to the 16 scenarios.

Exact envelope for fixed predictions

For fixed predictions, define C as the weighted contribution of undisputed labels, and let L_w and U_w denote the lower and upper endpoints. Because each disputed label appears in a separate additive term, the minimum and maximum loss differences are obtained by selecting each record's smaller and larger contribution:

$$L_w = C + \sum_{i \in U} w_i \min\{\bar{d}_i(0), \bar{d}_i(1)\}, \qquad C = \sum_{i \notin U} w_i \bar{d}_i(y_i). \quad \textbf{(3)}$$

$$U_w = C + \sum_{i \in U} w_i \max\{\bar{d}_i(0), \bar{d}_i(1)\}. \quad (4)$$

The proof is the separability of the finite optimization: choosing the label for one disputed team does not constrain another under the stated ambiguity set. Contributions are averaged across repetitions before optimization, so a team receives one consistent hypothetical label in every repetition. This calculation is linear in the number of stored predictions. It is not a sampling confidence interval, and a negative upper endpoint establishes dominance only for those fixed predictions and that ambiguity set.

A useful descriptive consequence is the contribution of each disputed record to the total envelope width:

$$W_i = w_i\left|\bar{d}_i(1) - \bar{d}_i(0)\right| = 2w_i\left|\frac{1}{R}\sum_{r=1}^{R}(p_{ir} - b_{ir})\right|, \qquad U_w - L_w = \sum_{i \in U} W_i. \quad \textbf{(5)}$$

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

This identity identifies records responsible for sensitivity in a frozen comparison. It is not an optimal source-adjudication policy: verification costs are unspecified, no prior over correct labels is supplied, and refitting may change predictions elsewhere. A record with a large fixed-prediction contribution is a transparent candidate for source checking, not a proven highest-value intervention.

Exhaustive refitting and interpretation of scenario counts

The fixed-prediction envelope does not bound a refitted model. Each label assignment is therefore used to rebuild training labels, refit both models and the baseline on the shared folds, and score the resulting held-out predictions. Team- and country-weighted means are computed from the same prediction records. The minimum and maximum across assignments characterize this finite exercise. A count such as 15 of 16 favorable scenarios is not a 93.75% probability of superiority. Repetition variability describes partition sensitivity rather than independent population replications; it is not used to construct an inferential p-value.

Office-level administrative mobilization analysis

The 43 offices are matched by office name between the 2023 and 2024 reports. The exposure is the number of 2023 Potensi SAR training events attributed to each office: 33 offices have at least one recorded event, and the extracted exposure totals 43 office-attributed events. A zero means no event is recorded in the selected table, not verified absence of all training. The exposure does not establish whether deployable INASAR members attended, whether competencies improved, or whether the trained personnel participated in subsequent incidents (Milani, Roveri, Falla, Cappello, & Strapazzon, 2023).

The outcome requires a precise definition. The 2023 report defines response time as the interval from a received report being declared valid until the search and rescue unit is ready to depart; it does not measure arrival at the incident or completion of rescue (Badan Nasional Pencarian dan Pertolongan, Laporan Kinerja Badan Nasional Pencarian dan Pertolongan Tahun 2023; BASARNAS: Jakarta, Indonesia, 2024, 2023). We therefore use the term administrative mobilization-response time for the office means. The reported measure covers multiple SAR incident categories, not structural collapse alone. This restriction is substantive: a training association with this endpoint would not directly estimate improved search proficiency, reduced travel time (Alabede, 2023), or patient survival (Ouyang, Zhang, Argon, & Ziya, 2024).

For office i, the primary specification adjusts 2024 mean response time for 2023 response time, log(1 + 2023 operation count), the 2023 training-event count, and four indicators representing five analytic regions (Huang & Wang, 2024). One region is the reference:

$$R_{i,2024} = \alpha + \beta R_{i,2023} + \theta \log\left(1 + N_{i,2023}\right) + \delta T_{i,2023} + \sum_{r=2}^{5} \lambda_r I_{ir} + \varepsilon_i. \quad \textbf{(6)}$$

The coefficient δ is an adjusted association in minutes per recorded training event. HC3 heteroskedasticity-consistent covariance is used with normal-reference intervals, matching the archived implementation (White, 1980). We rerun the full model, delete each office once for influence diagnosis, and retain Manado in the primary analysis. Original no-region, binary-exposure, change-score, Huber, and median-regression sensitivity estimates are preserved separately as archived results (Huber, 1964). Cook's distance and studentized residuals in the original archive belong to its no-region model, not to every specification (Koenker & Bassett, 1978).

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

Leave-one-office-out cross-validation compares persistence (the office's 2023 mean), the training-set mean, baseline adjustment, training augmentation, and training-plus-region adjustment. Every regression is refitted without the held-out office. Performance is summarized by RMSE, mean absolute error (MAE), median absolute error, and the largest absolute error. Reporting absolute-error summaries alongside RMSE distinguishes broad prediction quality from concentration in a large residual. Repeating the procedure after excluding Manado changes the evaluation population; it does not estimate the deployment benefit of deleting difficult cases.

The archive also includes a 5000-draw region-stratified bootstrap and 10,000 within-region Freedman–Lane residual permutations (Efron & Tibshirani, 1993). These are retained as historical sensitivity analyses rather than rerun claims (Freedman & Lane, 1983). In particular, blocking by region does not guarantee exchangeable residuals under heteroskedasticity or sparse strata. None of these adjustments identifies a causal training effect without assumptions about assignment, baseline differences, personnel exposure, and incident composition (Hernán & Robins, 2020).

Source-level diagnosis of the influential office and national aggregation

To investigate whether the Manado residual reflects a simple mismatch with the published tables, we checked its incident-category operation counts and total minutes against the official page images. The 2023 row appears on PDF page 56 (printed page 45); the 2024 row appears on PDF page 66 (printed page 56) (Badan Nasional Pencarian dan Pertolongan, Laporan Kinerja Badan Nasional Pencarian dan Pertolongan Tahun 2023; BASARNAS: Jakarta, Indonesia, 2024, 2023). Four categories have operations in both years; aircraft operations are zero in both. For each non-empty category c, let n_{ct} be the operation count, a_{ct} the total recorded minutes, $r_{ct} = a_{ct}/n_{ct}$ the category mean, and $w_{ct} = n_{ct}/\Sigma_c\, n_{ct}$ its operation share (Badan Nasional Pencarian dan Pertolongan, Laporan Kinerja Badan Nasional Pencarian dan Pertolongan Tahun 2024; BASARNAS: Jakarta, Indonesia, 2025, 2026).

We describe the change in the aggregate mean using a symmetric algebraic decomposition into changes in category means and changes in category shares:

$$\bar{r}_{2024} - \bar{r}_{2023} = \sum_c \bar{w}_c \left(r_{c,2024} - r_{c,2023}\right) + \sum_c \bar{r}_c \left(w_{c,2024} - w_{c,2023}\right). \quad \textbf{(7)}$$

Here, $\bar{w}_c$ and $\bar{r}_c$ are the arithmetic averages of the two yearly weights and category means. Expanding the right-hand side cancels the cross terms and recovers the aggregate difference exactly. Categories empty in both years contribute nothing; no undefined zero-count rate is imputed. The identity describes reported aggregates, not the causal effects of incident mix or training. Changes in category means may still reflect changing conditions within a broad category, timing practices, or source errors.

For the national response indicator, we distinguish an unweighted mean across offices from a mean weighted by operation counts and from total reported minutes divided by total operations. The five national 2020–2024 performance series remain descriptive (Appendix C). We do not fit trend models, correlate five-point series for hypothesis testing, or standardize unlike indicators into an overall readiness score.

Paired synthetic experiments

The simulation follows the aims, data-generating mechanisms, estimands, methods, and performance-measures structure used to report statistical simulation studies (Morris, White, &

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

Crowther, 2019). Its objective is to diagnose representation and recording-error effects under controlled functions. Each observation draws five independent Uniform(0,1) dimensions; the outcome is a continuous score clipped to after Gaussian noise. Four generators represent a linear bounded process, a smooth sigmoid, a smooth interaction process, and a hard-gate process. Appendix B gives every equation, coefficient, threshold, and noise scale. The settings are numerical experiments, not calibrated operational standards (INSARAG, INSARAG Guidelines 2020: Volume II, Preparedness and Response, Manual C—INSARAG External Classification and Reclassification; United Nations Office for the Coordination of Humanitarian Affairs: Geneva, Switzerland, 2020., 2020).

The matched-feature experiment uses training n = 27 and 60, 200 independent replications per generator–size cell, and 1500 independent test observations per replication. Both ridge regression and penalized sigmoid regression receive the same three feature regimes: five dimensions alone; dimensions plus three exact gate indicators and their conjunction; and dimensions plus three independently corrupted gate indicators with their recomputed conjunction. The hypothetical symmetric flip probability is 0.10 per individual gate. Corrupted records are shared between algorithms within a replication; training and test corruption are independently generated. The same outcomes are used across the six fits in each training/test experiment.

Exact gates are deterministic transformations of the original dimensions. They supply the generator's threshold representation, not independent raw information. Ridge uses training-set standardization, a squared-error penalty parameter of 1, and predictions clipped to (INSARAG, INSARAG Guidelines 2020: Volume II, Preparedness and Response, Manual C—INSARAG External Classification and Reclassification; United Nations Office for the Coordination of Humanitarian Affairs: Geneva, Switzerland, 2020., 2020). The sigmoid minimizes mean squared error plus 0.05 times the squared non-intercept coefficients divided by training size. It uses the same standardized predictors, a zero initial vector, L-BFGS-B, a 2000-iteration limit, and objective tolerance 10^{-12}. Optimization failure raises an error rather than silently returning a prediction. Hyperparameters are fixed and test observations are never used for tuning (Hoerl & Kennard, 1970).

There are 1600 distinct matched training/test experiments and 9600 fitted models. The additional seed-20260916 sweep uses only the hard-gate generator, training sizes 27, 60, 120, and 240, and error probabilities 0, 0.025, 0.05, 0.10, and 0.20, together with dimensions-only baselines. Two hundred replications per size produce 800 distinct training/test experiments and another 9600 fitted models. Within each replication, training data, test data, noise, and uniform corruption draws are shared across error rates and algorithms. This pairing reduces irrelevant Monte Carlo variation when comparing conditions.

The primary performance estimand is mean test RMSE across independent replications. For a comparison of two conditions, the replicate-level RMSE difference is calculated before averaging. Its Monte Carlo standard error (MCSE) is the standard deviation of the paired differences divided by $\sqrt{200}$. For a single condition, MCSE is its replicate standard deviation divided by $\sqrt{200}$. Bars of mean ± 1.96 MCSE describe numerical precision under the generator, not uncertainty about operational performance in real teams. MAE and all replicate-level scores are also exported. Sample sizes and replication counts are fixed stress-test settings, not an empirical power analysis.

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

Computational reproducibility and AI assistance

The current execution used Python 3.13.5, NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0, scikit-learn 1.8.0, and statsmodels 0.14.6 (Harris et al., 2020). Matplotlib generates all analytical graphics from stored observations (Virtanen et al., 2020), predictions, or simulation outputs. The supplement separates recovered outputs from current outputs, records environments and seeds, and provides code, figure provenance, and file hashes. Re-execution of the recovered core is a reproducibility check, not an independent implementation of every estimator (Hunter, 2007).

OpenAI ChatGPT assisted with methodological review, source searching, code development, manuscript revision, and figure preparation. Numerical calculations were executed in conventional scientific software, and no image-generation model produced the scientific figures. Simulated observations are explicitly generated numerical experiments, not fabricated field observations. Author verification of the source interpretations, code, conclusions, and declarations remains a prerequisite for submission; no inference of human approval is made from successful code execution.

RESULTS AND DISCUSSION

Evidence integrity, source agreement, and sample description

All 50 recovered files tested against the embedded manifest matched their recorded hashes. Recomputed Brier scores for all 180 archived score records agreed with the exported values to a maximum absolute difference of 1.67×10^{-16}. The shared-fold rerun reproduced all 29,280 saved prediction records without a numerical difference, and the matched-feature rerun reproduced all 9600 saved RMSE and MAE records. These checks support execution consistency of those outputs. They do not independently verify every original extraction or establish construct validity.

The targeted source checks confirmed that the four first-IEC disagreements remain present between the two official pages. They also established a narrower meaning for the office response endpoint and agreement of the influential Manado office means with the displayed total-minute and operation-count rows. Table 3 distinguishes those current checks from unresolved archival warnings. This distinction avoids treating every anomaly as either a confirmed source error or a corrected record.

Table 3 Material source findings and their verification status.

Issue	Evidence and status	Analytic handling
Four first-IEC levels disagree	Current comparison of the two official registry pages; no adjudication received.	Retain all teams; enumerate the 16 independently admissible assignments.
Response-time meaning	2023 report, PDF p. 55 / printed p. 44: valid report to unit ready to depart.	Interpret as administrative mobilization, not arrival or completion of rescue.
Manado annual change	Category and total rows checked in both reports; 803/43 and 2795/64 minutes per operation.	Keep the office in the primary model; investigate reported categories without asserting an error.

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

Issue	Evidence and status	Analytic handling
2024 certified-HR indicator	Current tables/chart show 96.29; five-year comparison table shows 96.26.	Retain discrepancy; display 96.29 descriptively; neither value enters a fitted model.
Training-table totals and historical WGI scale warning	Preserved from original audit; complete event rows and former API behavior were not independently reconstructed here.	No claim of current API failure; 2024 training counts are not the exposure in the office model.

Source: Compiled by the author based on field observations, 2026.

The global descriptive governance component explains 97.05% of standardized variance in its three input indicators. This indicates redundancy among macro-context variables in the selected sample, not validity of five team-level capabilities. The unweighted office mean increases from 16.424 to 17.207 minutes. Recorded-training offices have a mean annual change of +0.309 minutes, compared with +2.342 among offices without a recorded event. These are unadjusted comparisons of a coarse exposure and do not establish a training effect.

Registry robustness depends on both label assignment and weighting

For the archived detailed-registry predictions held fixed, the team-weighted exact envelope is [−0.00903, +0.00070] for context logit and [−0.02189, +0.00532] for full ridge. Both cross zero. Exhaustive evaluation of all assignments recovers the analytic endpoints. Neither frozen comparison therefore has an advantage invariant to the specified unresolved labels.

Table 4 Shared-fold refitting across all 16 assignments: the same predictions and baseline, with different evaluation weights.

Model	Loss weighting	Minimum Δ	Maximum Δ	Δ < 0
Context logit	Country-equal	-0.015655	-0.000044	16/16
Context logit	Team-equal	-0.013898	+0.000038	15/16
Ridge logit	Country-equal	-0.041266	-0.003820	16/16
Ridge logit	Team-equal	-0.028413	+0.003483	15/16

Source: Compiled by the author based on field observations, 2026.

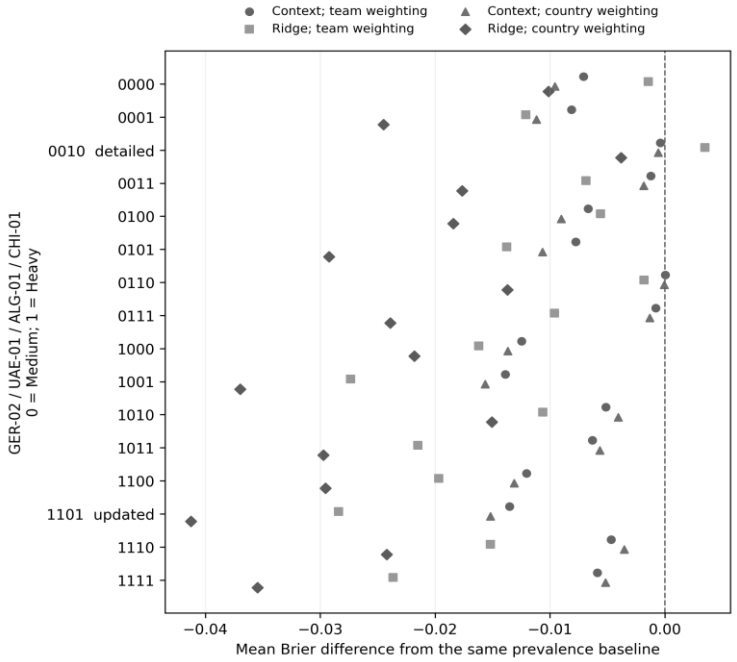
Δ is candidate minus training-fold prevalence Brier loss, averaged over ten shared country-fold repetitions. Ranges are sensitivity summaries, not confidence intervals. Country weighting does not alter model fitting or the prevalence baseline.

Under shared-fold refitting with equal team weights, full ridge ranges from −0.028413 to +0.003483 relative to prevalence and is favorable in 15 of 16 assignments. Context logit ranges from −0.013898 to +0.000038 and is also favorable in 15 assignments. For ridge, the detailed-page assignment 0010 produces the unfavorable endpoint and the updated-page assignment 1101 produces the most favorable endpoint. The identities of the two near-zero context scenarios should not be interpreted through rounded signs alone; the supplement retains full precision.

The country-equal sensitivity changes this conclusion: ridge differences range from −0.041266 to −0.003820, and context differences from −0.015655 to −0.000044. Both comparisons are negative in every enumerated assignment (Figure 1). The context margin is only about 4.4×10^{-5} at its least favorable assignment. Thus, a blanket statement that improvement is never label-invariant would be incorrect. The defensible result is that label-invariance fails for

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

the primary team-equal comparison but holds, within this finite experiment, when countries receive equal evaluation weight. Neither result is a population guarantee or a comparison with a separately country-weighted fitted.



Picture 1 Mean paired Brier.

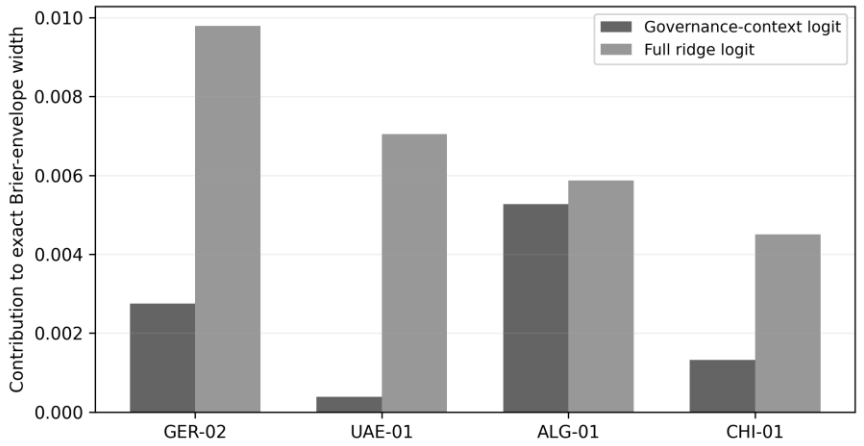
This distinction is substantive rather than cosmetic. Equal team weighting gives countries with more sampled teams more total influence. Equal country weighting gives each represented country the same total influence, irrespective of its number of participating teams. Neither is automatically the right target for an intended decision. The primary weighting is retained rather than replaced after inspecting the more favorable sensitivity. A deployment target would have to specify its team and country distribution before evaluating performance.

The original stratified-fold benchmark remains compatible with this account but is not numerically identical to the shared-fold experiment. Under detailed coding, ridge Brier loss was 0.2479 versus 0.2425 for prevalence; under updated coding, it was 0.2140 versus 0.2353. Since the fold construction also differs, these absolute values cannot isolate the label effect by comparison with the shared-fold values. Appendix C preserves the original protocol-specific summaries.

Frozen-loss sensitivity identifies model-specific source priorities

Equation (5) decomposes the exact frozen-loss envelope into four record contributions (Figure 2). GER-02 has the largest contribution for full ridge, approximately 0.00979 Brier units, followed by UAE-01 (0.00705), ALG-01 (0.00587), and CHI-01 (0.00451). Context logit is most sensitive to ALG-01 (0.00527); the UAE-01 contribution is much smaller (0.00038). The record ordering is therefore model-dependent even before refitting. It should be described as sensitivity localization, not a model-independent ranking of registry quality.

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal



Picture 2 Contributions to the exact team-weighted Brier-envelope width with archived predictions

In practical terms, checking GER-02 would remove more of ridge's present frozen-score ambiguity than checking CHI-01, provided verification is definitive and other labels are unchanged. This statement does not determine which source is correct and does not predict how all coefficients will change after retraining. The distinction matters when using an audit to set a data-cleaning agenda: uncertainty can be localized without overstating the information that an audit supplies.

Office training association and out-of-office prediction

The primary region-adjusted estimate is −1.327 minutes per recorded training event, with a 95% HC3 interval of [−5.004, +2.351] and p = 0.480. Excluding Manado yields +0.402 minutes, with interval [−0.520, +1.325] and p = 0.393 (Table 5). These broad intervals include effects in both directions; they are not equivalence results. Across the 43 single-office deletion fits, the coefficient ranges from −1.656 to +0.402, with the positive endpoint obtained by removing Manado. Figures 3 and 4 retain the full set of office observations and deletion diagnostics.

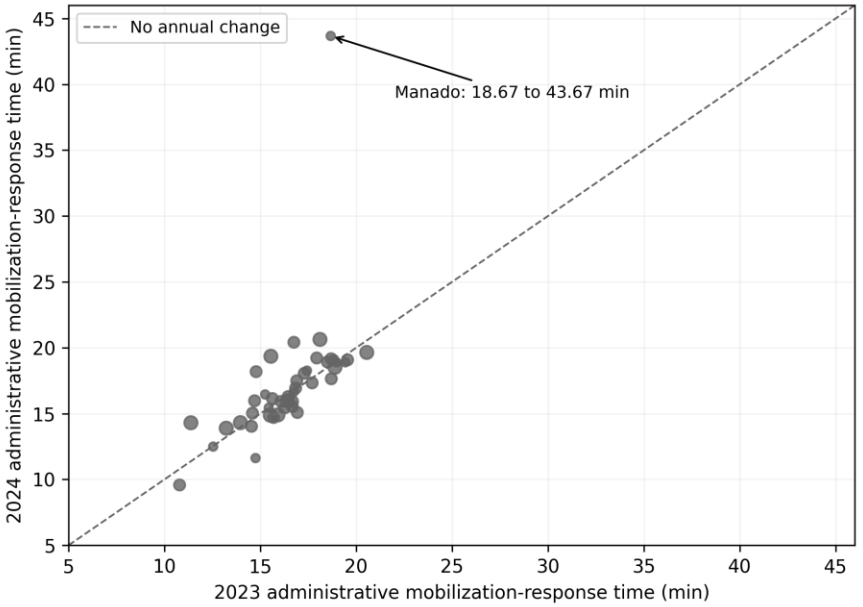
Table 5 Training association under current reruns and retained archival sensitivities.

Specification	Coefficient (min)	95% interval	Status
Region-adjusted primary	−1.327	[−5.004, +2.351]	Rerun; HC3; n = 43
Region-adjusted, excluding Manado	+0.402	[−0.520, +1.325]	Rerun; HC3; n = 42
No-region adjustment	−0.808	[−3.333, +1.716]	Archived HC3; n = 43
Any recorded training	−1.951	[−7.075, +3.172]	Archived binary contrast; n = 43
Change-score sensitivity	−1.384	[−5.219, +2.451]	Archived; n = 43
Huber regression + region	+0.014	[−0.563, +0.591]	Archived robust fit; n = 43
Median regression + region	+0.219	[−0.628, +1.066]	Archived quantile fit; n = 43

Source: Compiled by the author based on field observations, 2026.

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

All coefficients are per recorded event except the binary any-training contrast. Intervals belong to their respective estimators; they are not interchangeable. Archival rows are retained for transparency and were not all rerun in this revision.



Picture 3 Administrative mobilization-response time in 43 matched offices.

Persistence is the strongest of the five currently rerun prediction baselines on the full office set: RMSE 4.064 minutes, MAE 1.565, and median absolute error 0.670 (Table 6). Adding the training count to baseline adjustment increases RMSE from 4.196 to 4.341 and MAE from 1.975 to 2.156. Adding region increases them further to 4.796 and 2.652. Thus, the lack of a predictive increment is not confined to a single squared-error summary. After excluding Manado, persistence remains best among these models, with RMSE 1.423 and MAE 1.007; this is a different 42-office target population.

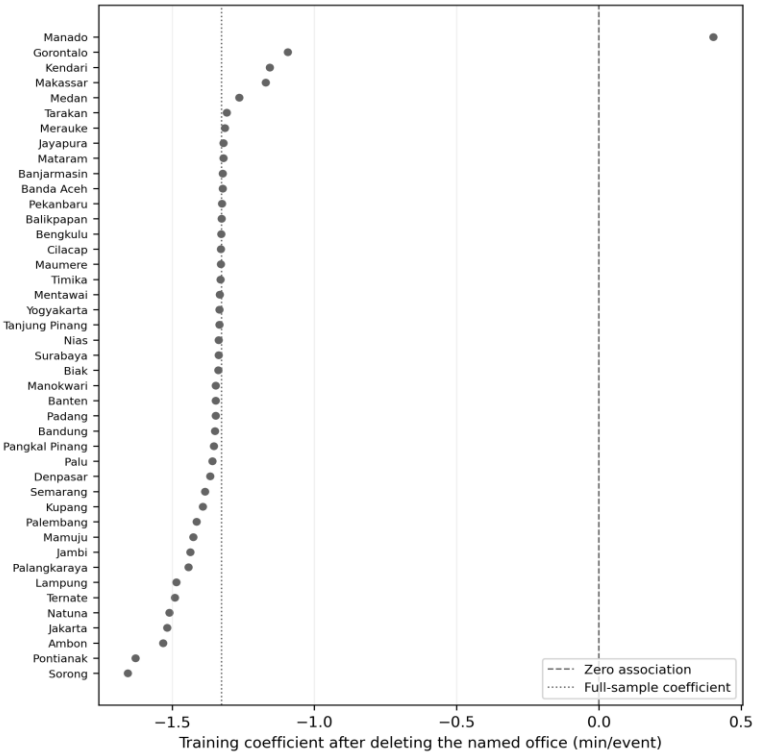
Table 6 Recomputed leave-one-office-out prediction errors, in minutes.

Model	RMSE 43	MAE 43	Median AE 43	RMSE 42	MAE 42
Persistence	4.064	1.565	0.670	1.423	1.007
Training mean	4.826	2.644	1.829	2.441	1.992
Baseline ANCOVA	4.196	1.975	1.348	1.500	1.124
ANCOVA + training	4.341	2.156	1.384	1.515	1.168
ANCOVA + training + region	4.796	2.652	1.447	1.690	1.296

The 42-office set excludes Manado; all models are refitted within each leave-one-out split. The supplement also provides maximum absolute errors and every held-out prediction.

Manado contributes 88.0% of persistence squared error. Its 25.00-minute difference between the two rounded office means is also the largest absolute persistence error. The change in RMSE after exclusion therefore mainly describes error concentration, not evidence that the remainder of the country has become easier to serve. The original no-region influence values Cook’s distance 0.862 and externally studentized residual 16.862—are consistent with this diagnosis but do not establish a transcription mistake.

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal



Picture 4 Region-adjusted training coefficient after deleting each named office once. What the Manado source rows establish—and do not establish

The checked source totals support the office values used in the analysis: 803 recorded minutes across 43 operations in 2023 and 2795 minutes across 64 operations in 2024. Division yields 18.6744 and 43.6719 minutes, respectively, which round to the published 18.67 and 43.67. The anomaly is therefore not resolved by replacing a copied office mean with the displayed total/count ratio. Table 7 reports the four non-empty categories used in the source-level arithmetic check.

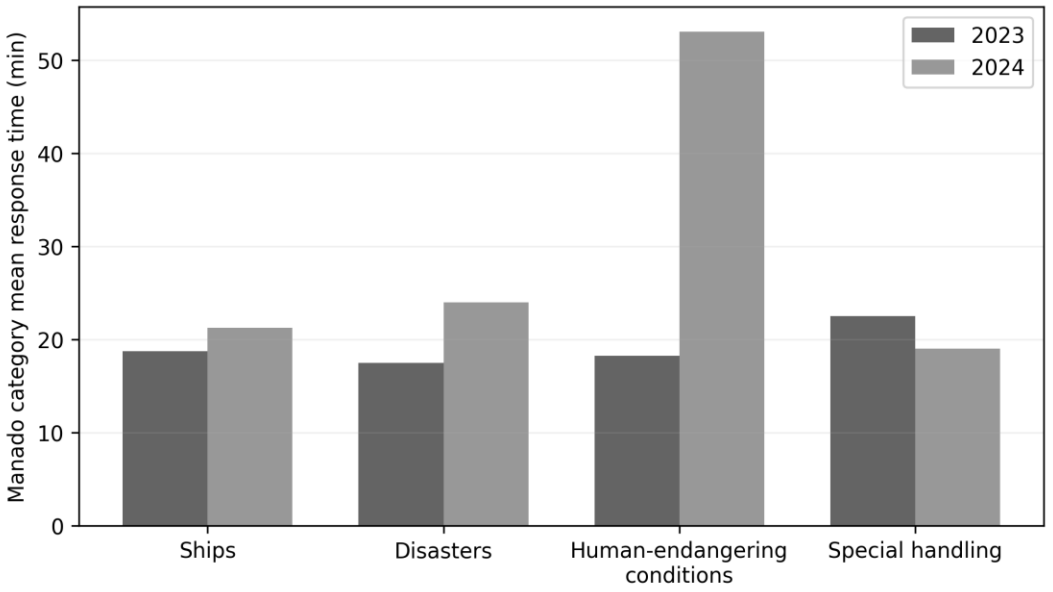
Table 7 Manado source totals and derived category means.

Incident category	Ops 23	Min 23	Mean 23	Ops 24	Min 24	Mean 24
Ships	21	394	18.762	16	340	21.250
Disasters	2	35	17.500	2	48	24.000
Human-endangering conditions	18	329	18.278	45	2388	53.067
Special-handling incidents	2	45	22.500	1	19	19.000

Ops = operation count; Min = total recorded minutes; Mean = Min/Ops. Source: BASARNAS 2023 PDF p. 56 / printed p. 45 and 2024 PDF p. 66 / printed p. 56 (Badan Nasional Pencarian dan Pertolongan, Laporan Kinerja Badan Nasional Pencarian dan Pertolongan Tahun 2023; BASARNAS: Jakarta, Indonesia, 2024, 2023). Category means are calculated from the totals, not copied from rounded display cells. Aircraft operations are zero in both years (Badan Nasional Pencarian dan Pertolongan, Laporan Kinerja Badan Nasional Pencarian dan Pertolongan Tahun 2024; BASARNAS: Jakarta, Indonesia, 2025, 2026).

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

The human-endangering-conditions category changes from 18 operations and 329 minutes to 45 operations and 2388 minutes. Its derived mean rises from 18.278 to 53.067 minutes (Figure 5). The symmetric decomposition of the unrounded aggregate increase, 24.9975 minutes, is 20.5744 minutes from the category-mean component and 4.4230 minutes from the category-share component. Their sum reproduces the aggregate difference to floating-point precision. This identifies where the increase is recorded, not why it occurred. In particular, "within-category" does not mean constant incident severity, personnel, recording practice, or operational conditions.



Picture 5 Manado mean administrative mobilization-response time by reported incident category

The source check changes the appropriate next step: request clarification or more granular incident records for the relevant category rather than delete the office automatically. The increase could reflect a genuine operational pattern, a changed mixture of incidents within the category, or an error already present in the official source. The present aggregate evidence cannot distinguish those explanations. No allegation about individual staff or office performance follows from the residual.

Aggregation also explains a smaller apparent discrepancy. The unweighted office means are 16.4244 and 17.2065 minutes, whereas total recorded minutes divided by total operations yields 16.3981 and 17.1793. These round to the national 16.40 and 17.18 values. The office-equal and operation-weighted summaries target different quantities; their difference is not, by itself, an inconsistency. The remaining national series are retained in original units in Appendix C without inferential trend claims.

Matched features, rather than algorithm labels, explain the large hard-gate gain

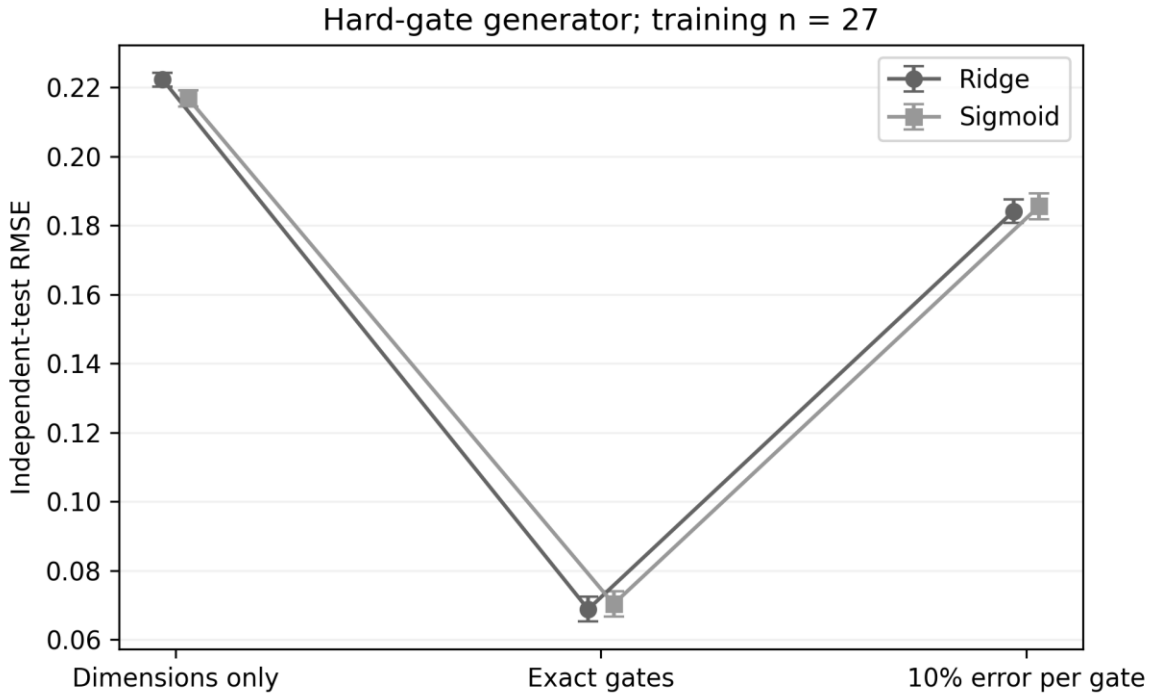
In the hard-gate process at n = 27, ridge RMSE falls from 0.22228 with dimensions alone to 0.06879 with exact gate features. The matched sigmoid values are 0.21681 and 0.07032 (Table 8). Paired improvements are 0.15348 for ridge (MCSE 0.00191) and 0.14649 for sigmoid (MCSE 0.00201). With the exact representation, the sigmoid-minus-ridge RMSE difference is only 0.00152 (MCSE 0.00018). Under these fixed settings, the representation effect is much larger than the difference between the two algorithms.

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

Table 8 Matched-feature hard-gate experiment, seed 20260909.

Training n	Shared feature regime	Ridge RMSE ± MCSE	Sigmoid RMSE ± MCSE
27	Dimensions	0.2223 ± 0.0010	0.2168 ± 0.0012
27	Exact gates	0.0688 ± 0.0018	0.0703 ± 0.0019
27	10% error per gate	0.1841 ± 0.0017	0.1855 ± 0.0019
60	Dimensions	0.2093 ± 0.0005	0.2004 ± 0.0007
60	Exact gates	0.0547 ± 0.0002	0.0551 ± 0.0002
60	10% error per gate	0.1639 ± 0.0008	0.1639 ± 0.0009

Two hundred independent replications per cell; 1500 independent test observations per replication. The complete four-generator results and paired contrasts are in Supplement S1.



Picture 6 Matched-feature comparison under the hard-gate generator at n = 27.

At n = 27, corrupting each gate with probability 0.10 increases RMSE to 0.18409 for ridge and 0.18554 for sigmoid. At n = 60, the corresponding noisy-gate values are 0.16388 and 0.16392, compared with exact-gate values near 0.055. These losses are not equivalent to a 10% error probability for the entire gate vector or its conjunction. Multiple gate records can be corrupted, and a flip need not change the conjunction in every record.

The other generators prevent the hard-gate result from being mistaken for a universal ranking. For the interaction generator with dimensions alone, sigmoid RMSE is lower than ridge by 0.01957 at n = 27 and 0.02089 at n = 60. The smooth-sigmoid differences are much smaller, and the linear process slightly favors ridge. All 48 generator–size–feature–model summaries are retained. All 4800 sigmoid optimizations in the matched experiment satisfied the implementation's convergence check. These are comparisons of selected fitted function classes and penalties, not validation of the five proposed governance constructs.

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

Independent recording-error replication across training sizes

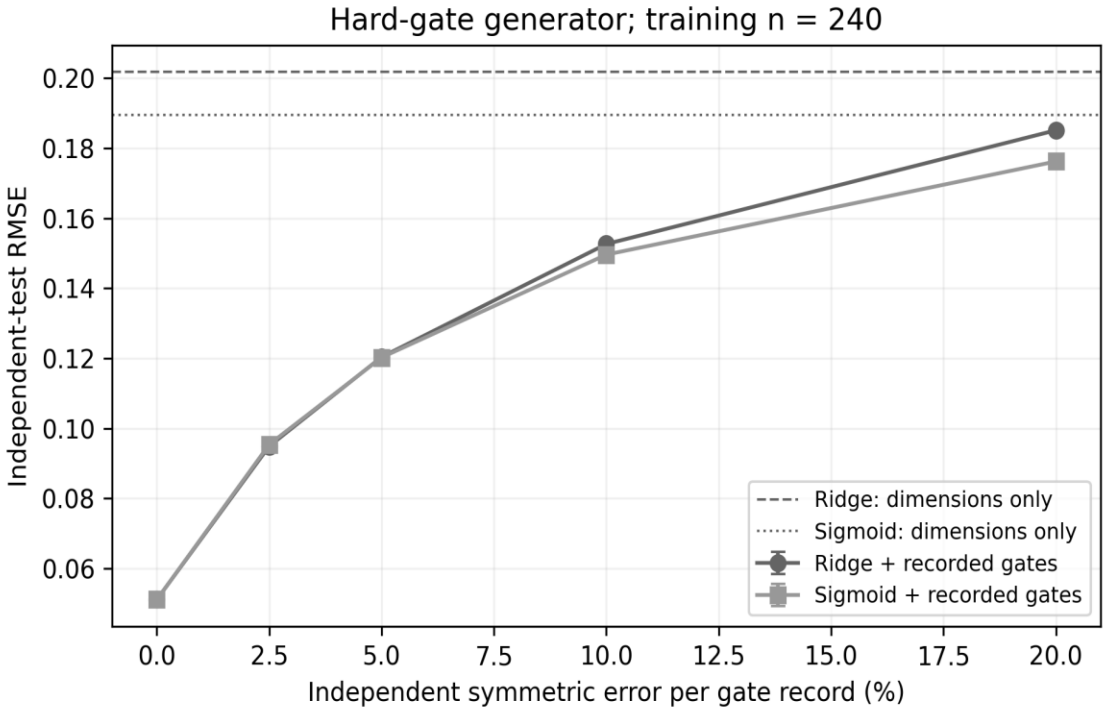
The new seed-20260916 experiment reproduces the qualitative error sensitivity without reusing the former gate-error draws. At n = 27, ridge RMSE is 0.22194 with dimensions alone and 0.06528 with exact gates. It rises to 0.18314 at q = 0.10 and 0.22351 at q = 0.20. Thus, at the largest hypothetical recording-error rate, gate augmentation provides no improvement over the dimensions-only ridge mean in this small-sample experiment. The paired differences and MCSEs are supplied rather than inferred from overlapping marginal error bars.

Table 9 Newly executed gate-error sweep at the smallest and largest training sizes, seed 20260916.

n	Model	Dimensions	Exact gates	10% gate error	20% gate error
27	Ridge	0.2219 (0.0011)	0.0653 (0.0013)	0.1831 (0.0017)	0.2235 (0.0015)
27	Sigmoid	0.2180 (0.0015)	0.0667 (0.0013)	0.1837 (0.0018)	0.2233 (0.0018)
240	Ridge	0.2018 (0.0003)	0.0511 (0.0001)	0.1526 (0.0004)	0.1851 (0.0004)
240	Sigmoid	0.1896 (0.0003)	0.0511 (0.0001)	0.1496 (0.0004)	0.1762 (0.0004)

Entries are mean RMSE (MCSE), with 200 replications and 1500 test cases per replication. This is an independent run, not the former seed-20260910 output. Complete results include n = 60 and 120 and q = 0.025 and 0.05.

At n = 240, ridge RMSE is 0.05109 with exact gates, 0.15263 at q = 0.10, and 0.18511 at q = 0.20. Sigmoid RMSE is 0.05112, 0.14960, and 0.17621, respectively (Figure 7). More training observations improve the selected fits but do not recover their exact-gate performance while test-time gate records remain corrupted. All 4800 sigmoid fits in this independent sweep passed the convergence check. MCSEs quantify repeatability under this generator; the hypothetical error rates were not estimated from INSARAG or BASARNAS data.



Picture 7 Newly executed recording-error experiment at n = 240, seed 20260916.

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

The unchanged dimensions contain enough information to compute the exact gates when the thresholds are known. A different method could exploit that fact or learn a more flexible boundary. Consequently, the curves do not establish an irreducible error bound or a general claim that additional data cannot overcome measurement problems. They show the cost of relying on corrupted engineered features within these fixed learners and regularization settings. Correlated errors, unknown thresholds, missing records, asymmetry, and distribution drift remain outside the experiment.

Discussion

What has been established for readiness assessment

The study establishes three bounded findings with direct relevance to how readiness evidence is interpreted. First, predictive robustness is jointly conditional on outcome adjudication and the population weighting encoded in the loss. The team-equal and country-equal results are not contradictory; they answer different evaluation questions. Second, a coarse office training count has no stable predictive increment under the examined specifications, while a large office residual remains supported by the published aggregate rows. Third, exact threshold representations explain most of the large hard-gate gain for two selected learners, and recording error degrades that advantage in an independent replication.

The contribution is an applied evidence audit, not a new readiness certification rule. The algebraic bounds make a finite source-disagreement problem transparent; the weighting comparison prevents a sensitivity result from being generalized beyond its estimand; the Manado check moves from statistical influence to a traceable source question; and the paired simulation separates representation from algorithm identity. Keeping these operations distinct is essential. Combining their numerical outputs into a score would obscure rather than resolve their different units and assumptions.

The safety relevance lies in avoiding unwarranted assurance. The study does not estimate injuries prevented, rescue survival, or an implemented safety-control effect. Its conclusions concern the defensibility of evidence used before such decisions. Classification, capability records, and operational endpoints can inform management, but a retrospective model with unresolved labels and indirect predictors should not certify a team or override an applicable external assessment.

Interpreting the office result without blaming training or deleting difficult cases

The office exposure is a record of events attributed to an organization. It does not identify attendance, instructional quality, competency gain, recency, or deployment of the trained personnel. Likewise, the outcome ends at readiness to depart and aggregates several types of SAR operation. A weak association can arise even when a particular training program improves a particular task. Conversely, a favorable association would remain vulnerable to selection into training, staffing differences, and changes in incident conditions. The current results cannot discriminate among these mechanisms.

Manado demonstrates why influence diagnosis and data correction should not be conflated. The two annual means can be reproduced from the official totals, and the recorded increase concentrates in a broad incident category. Neither finding proves that all underlying timings are correct. The defensible response is a targeted request for the timing definition in practice,

International Journal of Interdisciplinary Cultural Studies
ISSN: 2327-008X (Print), ISSN: 2327-2554 (Online)
Volume 21, Issue 2, 2026
https://cgscopus.com/index.php/journal

incident-level records where permissible, and an explanation of category composition. Removing the office because it changes a regression sign would conceal an unresolved evidence question.

from requirement verification

A proposed evidence-management design follows from the limitations, but has not been implemented or evaluated in this study. It keeps capability indicators, requirement verification, and operational outcomes in separate linked records. A capability indicator may support training or maintenance planning; a requirement record documents whether a specified condition has been verified for a particular assessment; an outcome record documents what happened in an exercise or incident. One should not substitute for another merely because all can be displayed on a dashboard (Selvakumaran, et al., 2026)

.

Table 10 Proposed evidence record and review actions; these are design requirements, not validated controls.

Record element	Minimum content	Reason for inclusion
Identity and applicability	Team/module, intended classification, checklist edition, requirement identifier.	Avoid transferring a label or condition across an inapplicable unit or standard.
Verification state	Verified pass, verified fail, or unknown/unverified; expiry where relevant.	Avoid turning missing evidence into an assumed pass.
Provenance	Source locator, assessment date, verifier, extraction rule, version/hash where appropriate.	Support correction and independent reconstruction without treating a hash as truth.
Capability measurement	Role-specific competency, exercise conditions, scoring rubric, recency and assessor agreement.	Distinguish a direct capability measure from an administrative proxy.
Outcome definition	Start/stop timestamps, incident category, denominator, exclusions, safety-critical errors.	Preserve the operational meaning and permit comparable evaluation.
Decision and audit trail	Human decision owner, unresolved conflicts, permitted model use, reasons for override.	Keep the scope of a model claim visible and contestable.

An unverified mandatory condition is not equivalent to a verified failure: both can prevent a positive verification claim, but they call for different actions. A documented failure requires corrective action or reassessment; an unknown state requires obtaining evidence. Whether either state legally or operationally precludes deployment depends on the applicable authority and context, not on the numerical rule in this paper. The simulation thresholds 0.22, 1.15, and 2.8 must never be adopted as official readiness requirements.

The error experiments motivate careful record management rather than the indiscriminate collection of more predictors. Their oracle regime assumes precisely known functions of existing dimensions. Real compliance evidence may instead involve inspection, documentary authorization, or independent competency checks. The experiment does not establish that an additional real-world field will improve prediction, and it does not quantify the cost-

effectiveness of a proposed verification workflow. Those questions require a field study with specified error sources and decision costs.

CONCLUSION

This study demonstrates that reliable urban search and rescue readiness assessment depends not only on the availability of data, but also on the alignment between the unit of analysis, operational endpoint, evidence source, and scope of the conclusion that can legitimately be drawn. The findings indicate that predictive performance is sensitive to unresolved classification records and to the weighting framework used to evaluate model performance. In the first-IEC analysis, the four disputed registry records produced different conclusions under equal team and equal country weighting. Although the ridge and contextual logistic models were not uniformly superior to the prevalence baseline when teams received equal weight, both models remained favorable across all sixteen admissible label assignments when countries received equal evaluation weight. This finding suggests that predictive robustness should be interpreted as estimand-dependent rather than presented as an unconditional indicator of model validity. Such a distinction is particularly important in safety-related contexts, where statistical improvement should not automatically be translated into operational assurance.

The BASARNAS analysis further shows that administrative training records provide limited evidence for assessing operational readiness. The recorded number of training events demonstrated an unstable adjusted relationship with subsequent administrative mobilization-response time and did not provide a consistent predictive improvement over the evaluated baseline models. Importantly, the outcome represents the time required for a unit to become ready to depart after a valid report, rather than arrival at the incident location or successful rescue completion. The Manado observation illustrates the importance of distinguishing statistical influence from data error: the reported annual means were consistent with the published totals and operation counts, while the increase was concentrated within a particular incident category. However, the available aggregate evidence does not establish the underlying cause. The controlled simulations similarly demonstrate that representation quality can be more consequential than algorithm selection, while recording errors substantially reduce the benefit of exact gate information. Overall, the study contributes an evidence-audit perspective that clarifies when readiness-related conclusions are defensible and when additional verification, measurement, and prospective evidence remain necessary.

Recommendations

Based on these findings, future readiness assessment should prioritize evidence quality, measurement specificity, and transparent interpretation rather than relying primarily on increasingly complex predictive models. First, disputed classification records should undergo systematic source adjudication before being used for high-stakes predictive or comparative analysis. Each record should identify the applicable team, classification, assessment edition, source location, verification status, and responsible reviewer. Where conflicting records cannot

immediately be resolved, the uncertainty should remain explicit and be incorporated through sensitivity analysis rather than being silently converted into a single assumed value. This approach would strengthen reproducibility while reducing the risk that an unresolved registry discrepancy is interpreted as a substantive organizational difference.

Second, BASARNAS and related emergency-management institutions should distinguish administrative training exposure from direct measures of competency and operational capability. Training-event counts should not be treated as evidence that personnel acquired, retained, or subsequently applied the required competencies. Future studies should collect team-level information on participant identity, competency assessment, training recency, exercise performance, role assignment, and subsequent deployment. Operational outcomes should also be defined with greater precision, including timestamps, incident characteristics, denominators, exclusions, and safety-relevant performance measures. For influential observations such as Manado, researchers should seek incident-level clarification and examine changes in category composition before considering exclusion or correction.

Third, readiness information should be organized into separate but linked evidence domains covering capability measurement, requirement verification, and operational outcomes. A verified requirement should not be inferred solely from a statistical prediction, while an unknown or missing record should not automatically be treated as a failed requirement. Provenance, assessment date, verification status, source locator, assessor, and audit history should be retained to facilitate independent review. Finally, predictive models should be evaluated prospectively against explicitly defined baselines and decision targets. Simulation evidence should guide methodological caution rather than establish operational thresholds. The numerical thresholds used in the simulations should not be adopted as certification standards. Future research should therefore examine prospective field data, measurement error, temporal validation, personnel exposure, and decision consequences before any predictive system is considered for operational readiness decisions.

ACKNOWLEDGEMENTS

The author would like to express sincere gratitude to the Rector of Universitas Muhammadiyah Jakarta, Prof. Dr. Ma'mun Murod; the Dean of the Faculty of Social and Political Sciences, Universitas Muhammadiyah Jakarta, Prof. Dr. Evi Satispi, S.Pd., M.Si.; and the Head of the Doctoral Program in Public Administration, Prof. Dr. Andriansyah, M.Si., for their valuable support, guidance, and encouragement throughout the research process. The author greatly appreciates the opportunities, direction, and continuous guidance provided from the initial stage to the completion of this research.

REFERENCES

Alabede, L. (2023). Enhancing emergency response readiness through autonomous drones for rapid search, rescue, and situational awareness. *4(2)*, 49–58. doi:10.22271/27078221.2023.v4.i2a.91

Ansell, C., & Gash, A. (2008). Collaborative governance in theory and practice. J. *Public Adm. Res. Theory, 18*, 543–571. doi:10.1093/jopart/mum032

Badan Nasional Pencarian dan Pertolongan. (2023). Laporan Kinerja Badan Nasional Pencarian dan Pertolongan Tahun 2023; BASARNAS: Jakarta, Indonesia, 2024. *Available online: https://basarnas.go.id/storage/files/9/LAPKIN%202023/LAKIP_*. Retrieved from https://basarnas.go.id/storage/files/9/LAPKIN%202023/LAKIP_2023_Basarnas.pdf

Badan Nasional Pencarian dan Pertolongan. (2026). Laporan Kinerja Badan Nasional Pencarian dan Pertolongan Tahun 2024; BASARNAS: Jakarta, Indonesia, 2025. *Available online: https://basarnas.go.id/storage/files/9/LAKIP%202024/Laporan%20Kinerja%20Basarnas%202024.pdf (accessed on 16 September*. Retrieved from https://basarnas.go.id/storage/files/9/LAKIP%202024/Laporan%20Kinerja%20Basarnas%202024.pdf

Bortolin, M. (2024). Urban Search and Rescue. 359–363. doi:10.1016/b978-0-323-80932-0.00055-0

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Mon. Weather Rev, 78*, 1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull, 56*, 81–105. doi:10.1037/h0046016

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychol. Bull, 52*, 281–302. doi:10.1037/h0040957

Efron, B., & Tibshirani, R. (1993). An Introduction to the Bootstrap; Chapman & Hall/CRC: New York, NY, USA, 1993.

Emerson, K., Nabatchi, T., & Balogh, S. (2012). An integrative framework for collaborative governance. J. *Public Adm. Res. Theory, 22*, 1–29. doi:10.1093/jopart/mur011

Freedman, D., & Lane, D. (1983). A nonstochastic interpretation of reported significance levels. J. *Bus. Econ. Stat, 1*, 292–298. doi:10.1080/07350015.1983.10509354

Gneiting, T., & Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. J. *Am. Stat. Assoc, 102*, 359–378. doi:10.1198/016214506000001437

Harris et al. (2020). Array programming with NumPy. *Nature, 585*, 357–362. doi:10.1038/s41586-020-2649-2

Hernán, M., & Robins, J. (2020). Causal Inference: What If; Chapman & Hall/CRC: Boca Raton, FL, USA, 2020.

Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 55–67. doi:10.1080/00401706.1970.10488634

Huang, L., & Wang, W. (2024). Search and rescue - optimizing emergency response through the allocation of critical services. 239–258. doi:10.4337/9781839102981.00026

Huber, P. (1964). Robust estimation of a location parameter. *Ann. Math. Stat, 35*, 73–101. doi:10.1214/aoms/1177703732

Hunter, J. (2007). Matplotlib: A 2D graphics environment. *Comput. Sci. Eng*, *9*, 90–95. doi:10.1109/MCSE.2007.55

INSARAG. (2005). IEC/Rs since 2005. *Available online: https://insarag.org/uncategorized/iec-rs-since-*. Retrieved from https://insarag.org/uncategorized/iec-rs-since-2005/

INSARAG. (2020). INSARAG Guidelines 2020: Volume II, Preparedness and Response, Manual C—INSARAG External Classification and Reclassification; United Nations Office for the Coordination of Humanitarian Affairs: Geneva, Switzerland, 2020.

INSARAG. (2027). INSARAG Guidelines. Available online: https://insarag.org/methodology/insarag-guidelines/ (accessed on 16 September 2026). *The guidance page identifies 1 January*. Retrieved from https://insarag.org/methodology/insarag-guidelines/

INSARAG. (Updated) IEC/Rs since 2005. (2005). Available online: https://insarag.org/iec/iec-rs/ (accessed on 16 September 2026). Retrieved from https://insarag.org/iec/iec-rs/

Jolliffe, I., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philos. Trans. R. Soc. A, 374*, 20150202. doi:10.1098/rsta.2015.0202

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46*, 33–50. doi:10.2307/1913643

MacKinnon, J., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. J. *Econom, 29*, 305–325. doi:10.1016/0304-4076(85)90158-7

Magid, E. (2023). Simulation Tools for Urban Search and Rescue Robotics. *28*, 4–11. doi:10.5954/icarob.2023.ps2

McKinney, W. (2010). Data structures for statistical computing in Python. doi:10.25080/Majora-92bf1922-00a

Milani, M., Roveri, G., Falla, M., Cappello, T., & Strapazzon, G. (2023). Occupational Accidents Among Search and Rescue Providers During Mountain Rescue Operations and Training Events. *81(6)*, 699–705. doi:10.1016/j.annemergmed.2022.12.015

Morris, T., White, I., & Crowther, M. (2019). Using simulation studies to evaluate statistical methods. *Stat. Med, 38*, 2074–2102. doi:10.1002/sim.8086

Okita, Y., & Shaw, R. (2019). Effects of the classification system in international SAR: The case of the 2015 Nepal earthquake. *Disaster Prev. Manag, 28*, 359–370. doi:10.1108/DPM-08-2018-0240

Okita, Y., & Shaw, R. (2020). Search and rescue team classification: BASARNAS of Indonesia. *Aust. J. Emerg. Manag, 35*, 73–79. Retrieved from https://knowledge.aidr.org.au/resources/ajem-april-2020-search-and-rescue-team-classification-basarnas-of-indonesia/

Okita, Y., Glassey, S., & Shaw, R. (2022). COVID-19 and the expanding role of international urban search and rescue (USAR) teams: The case of the 2020 Beirut explosions. *Int. J. Humanitarian Action, 7*, 8. doi:10.1186/s41018-022-00116-z

Okita, Y., Sugita, M., Katsube, T., & Minato, Y. (2018). Capacity building of international search and rescue teams through the classification system: Example from Japan Disaster Relief team and INSARAG External Classification and Reclassification. J. *Jpn. Assoc. Earthq. Eng, 18*, 6_23–6_40. doi:10.5610/jaee.18.6_23

Olympia, R. (2025). Emergency Preparedness. 25–32. doi:10.1007/978-3-031-85765-2_4

Ouyang, H., Zhang, Z., Argon, N., & Ziya, S. (2024). Dynamic Resource Allocation in Urban Search and Rescue Operations. doi:10.2139/ssrn.5063953

Pedregosa et al. (2011). Scikit-learn: Machine learning in Python. J. *Mach. Learn. Res*, *12*, 2825–2830.

Roberts et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography, 40*, 913–929. doi:10.1111/ecog.02881

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. doi:10.25080/Majora-92bf1922-011

Selvakumaran, S., Ma, W., Cobo, M., Thomas, D., He, N., & Marinoni, A. (2026). Multimodal, uncertainty-aware structural damage assessment for post-disaster Urban Search and Rescue (USAR) decision-making. doi:10.5194/egusphere-egu26-21653

Shmueli, G. (2010). To explain or to predict? Stat. *Sci, 25*, 289–310. doi:10.1214/10-STS330

Simonsohn, U., Simmons, J., & Nelson, L. (2020). Specification curve analysis. *Nat. Hum. Behav, 4*, 1208–1214. doi:10.1038/s41562-020-0912-z

Virtanen et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods, 17*, 261–272. doi:10.1038/s41592-019-0686-2

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*, 817–838. doi:10.2307/1912934

World Bank. (2026). The Worldwide Governance Indicators: Revised Methodology for Measuring Governance Using Perception Data; World Bank: Washington, DC, USA, 2025. *Available online: https://www.worldbank.org/en/publication/worldwide-governance-indicators/documentation (accessed on 16 September*. Retrieved from https://www.worldbank.org/en/publication/worldwide-governance-indicators/documentation

World Bank. (2026). World Development Indicators [dataset]. *Available online: https://databank.worldbank.org/source/world-development-indicators (original extraction reported as 18 July*. Retrieved from https://databank.worldbank.org/source/world-development-indicators

World Bank. (2026). Worldwide Governance Indicators, 2025 Revision [dataset]. *Available online: https://www.worldbank.org/en/publication/worldwide-governance-indicators (original extraction reported as 18 July*. Retrieved from https://www.worldbank.org/en/publication/worldwide-governance-indicators.