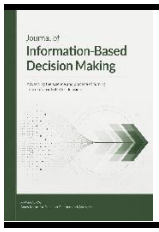
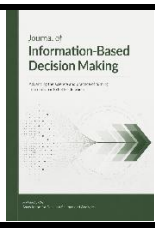


Journal of Information-Based Decision Making

Journal homepage: www.jibdm.journal-publishing.org
ISSN: 2838-1750



How Much Does the Aggregation Method Matter? Rank Reversal and Method Sensitivity in Multi-Criteria FMEA for Rural Hospital Sustainability

Sheng-Wei Lin[1,*]

[1] Department of Financial Management, National Defense University, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 7 June 2026
Received in revised form 29 July 2026
Accepted 12 September 2026
Available online 14 September 2026

Keywords:
Failure mode and effects analysis; multi-criteria decision making; rank reversal; method sensitivity; rural healthcare; Kendall rank correlation.

ABSTRACT

Failure mode and effects analysis (FMEA) is routinely coupled with multi-criteria decision-making (MCDM) methods, yet the sensitivity of the resulting risk priorities to the choice of aggregation method is seldom quantified on a fixed expert dataset. This study holds the data constant and varies the analysis. Using a four-factor FMEA questionnaire completed by twenty experts on twelve failure modes threatening sustainable healthcare in rural hospitals, risk rankings are computed under twenty-five configurations that cross eight ranking methods (TOPSIS, VIKOR, MARCOS, CoCoSo, MAIRCA, EDAS, WASPAS, and ARAS) with three weighting schemes (equal, questionnaire-derived, and entropy), plus a four-factor risk priority number baseline, with the classical three-factor index as an external reference. Agreement is high overall (mean Kendall tau-b of 0.931), but disagreement is systematic rather than random: it concentrates in the detection criterion, which correlates negatively with the other three factors, and in the compensation assumption, with regret-based VIKOR demoting the consensus top risk from first to fifth under equal weights. Removing any single expert perturbs rankings less (mean tau-b of 0.972) than switching methods does, so the analyst is a larger source of ranking variance than any individual panelist. The findings are distilled into a reporting protocol that treats method sensitivity as a declarable component of FMEA uncertainty.

1. Introduction

Rural hospitals occupy a structurally disadvantaged position in most health systems. They serve populations with heavier chronic disease burdens using thinner staffing, older equipment, and more fragile finances than their urban counterparts, and these disadvantages compound one another [1, 2]. When such institutions commit to sustainability programs that span environmental measures, telemedicine, community engagement, and workforce retention, the number of ways the commitment can fail multiplies, while the resources available for mitigation do not. Systematic risk prioritization is therefore not an administrative refinement for these organizations but a condition of survival: the ordering of failure modes determines where scarce mitigation capacity is spent first.

* *Corresponding author.*
E-mail address: shengwei.lin1985@gmail.com

Failure mode and effects analysis (FMEA) is the dominant instrument for this purpose in healthcare, endorsed by accreditation bodies and adapted for hospital use as healthcare FMEA [3, 4]. Its classical output, the risk priority number (RPN), multiplies severity, occurrence, and detection scores into a single index. The deficiencies of that index are among the best documented findings in the risk analysis literature: the implicit equal weighting of factors, the sensitivity of products to small rating changes, and the identical scores produced by very different risk profiles have been cataloged repeatedly [5–7]. The standard remedy has been to replace the product with a multi-criteria decision-making (MCDM) aggregation, and the literature now offers dozens of such replacements, from early multi-attribute formulations [8] through fuzzy hybrid designs [9–11] to recent granular and linguistic models [12].

This proliferation creates a question that the proliferation itself rarely answers. If a risk analyst can choose among many defensible aggregation methods, how much does the choice matter for the decision that FMEA is supposed to inform? The question is empirical and distinct from the question of which method is best. Simulation studies have compared MCDM methods on synthetic matrices [13–15], and the rank reversal literature has established that several methods can reorder alternatives when the set of alternatives changes [16–19]. What is scarce is a controlled accounting, on a single real expert panel dataset with a concrete decision at stake, of how much the priority ordering shifts when only the analysis changes, and how that shift compares with the movement induced by the composition of the expert panel itself.

Information-based decision frameworks of this kind, in which structured expert or operational data are converted into rankings that commit real resources, now span supply chains, finance, energy, and sustainability management [20–24], and risk-informed variants increasingly support reliability and policy decisions under uncertainty [25, 26]. In all of these settings, the analyst chooses an aggregation rule, and in almost none of them is the sensitivity of the output to that choice reported. FMEA is a natural setting for examining the omission because its inputs are standardized, its output is an ordering, and its use in resource-constrained hospitals makes the ordering have immediate consequences.

1.1 MCDM Extensions of FMEA

The movement of FMEA from a product index toward multi-criteria aggregation began with the recognition that severity, occurrence, and detection are neither equally important nor commensurable through multiplication [5, 8]. The subsequent literature can be read as a sequence of substitutions at two decision points: how factor weights are obtained, and how weighted factor scores are aggregated into an ordering. Kutlu and Ekmekcioglu [9] combined fuzzy AHP weighting with fuzzy TOPSIS ordering; Wang *et al.,* [10] moved the ratings into interval-valued intuitionistic fuzzy sets; Tian *et al.,* [11] integrated fuzzy best-worst weighting with entropy adjustment and VIKOR ordering; Lo *et al.,* [27] assembled a hybrid model for manufacturing failure modes; and Huang *et al.,* [12] introduced probabilistic linguistic term sets with TODIM ordering for reliability applications. Comprehensive reviews document well over a hundred such combinations [6, 7], and the uncertainty-representation frontier continues to advance through Fermatean, spherical, and other extended fuzzy environments in adjacent risk assessment domains [28].

Healthcare has been a prominent application area throughout. Structured healthcare FMEA emerged from patient safety programs [3], was consolidated in dedicated methodological treatments [4], and continues to be applied to service redesign and digital care, including simulation-supported failure reduction in clinical practice [29] and the FMEA-based implementation of televisiting programs [30]. The present dataset belongs to this tradition, with one extension that matters for resource-constrained settings: expected mitigation cost enters as a fourth factor alongside severity,

occurrence, and detection, since a rural hospital that cannot afford a remedy needs the cost of remedies inside the prioritization rather than outside it. The cost extension also changes what disagreement between aggregation rules means in practice. When prioritization determines which mitigations are funded within a fixed budget cycle, a two-position move in the middle of the ranking can carry a failure mode across the funding line, so the dependence of the ordering on the analysis is not a methodological curiosity but a budgeting risk in its own right, and it is in this sense that the present study, although it proposes no new aggregation, is addressed to practitioners as much as to methodologists.

1.2 Rank Reversal and the Consequences of Method Choice

A separate literature examines what happens to MCDM outputs when elements of the analysis change. Its founding observation is that AHP can reorder existing alternatives when a new alternative is added [16], and subsequent work established that rank reversal in various forms affects additive variants, ELECTRE methods, and other families [17, 18, 31]. Aires and Ferreira [19] systematize the phenomenon across methods and definitions. A complementary strand compares methods head to head: Zanakis *et al.,* [13] ran large simulation comparisons of additive and distance-based methods; Ceballos *et al.,* [14] compared fuzzy variants; Salabun *et al.,* [15] asked directly whether MCDA methods can be benchmarked and showed that recommendations diverge across reasonable choices; and selection frameworks now exist that treat the method itself as an object of choice [32, 33].

Two gaps remain at the intersection of this strand with FMEA. First, the comparative evidence is dominated by synthetic decision matrices, whose correlation structure is controlled by the simulator rather than produced by real experts assessing real risks; the structure of a genuine FMEA matrix, in which detection often opposes the other factors, is exactly the kind of feature simulations tend to average away. Second, comparisons across methods are rarely benchmarked against a source of variance that decision makers understand well, namely the composition of the expert panel. A ranking difference between two methods is hard to interpret in isolation; the same difference expressed as a multiple of the perturbation caused by the loss of one panelist is immediately interpretable.

1.3 Positioning, Research Gap, and Contributions

Within the taxonomy above, this paper is a method-sensitivity study on a fixed empirical FMEA dataset rather than a proposal of a new aggregation. The study takes no position on which configuration is correct; it measures how much the answer depends on the configuration, and prices that dependence against panel composition. The design philosophy parallels comparative weighting-and-ranking audits in other domains, where the robustness of a decision recommendation is assessed before it is acted upon [34–36].

The empirical vehicle is a four-factor FMEA dataset in which twenty experts rated twelve failure modes for rural hospital sustainability on severity, occurrence, detection, and expected cost. The analysis varies over a grid of twenty-five configurations: eight widely used ranking methods (TOPSIS, VIKOR, MARCOS, CoCoSo, MAIRCA, EDAS, WASPAS, and ARAS) crossed with three weighting schemes (equal, questionnaire-derived subjective, and entropy-based objective), plus a four-factor RPN baseline, with the classical three-factor RPN as an additional reference. Agreement is measured with Kendall rank correlation; the failure modes whose positions are unstable are located; each method is tested for rank reversal under single-alternative removal; and all of this is benchmarked against a leave-one-expert-out perturbation of the panel.

The gap, in short, is that the FMEA literature multiplies aggregation options without measuring how much the options matter on real expert data, and the significance of closing it is direct: for a

resource-constrained hospital, the difference between two defensible analyses can be the difference between funding one mitigation and funding another. The objective of this study is therefore threefold: to quantify the structure of agreement and disagreement across a broad grid of defensible configurations on one real dataset; to attribute the disagreement that exists to identifiable causes, namely the dissenting detection criterion and the compensation assumption; and to benchmark method-induced ranking variance against panel-induced ranking variance, converting an abstract methodological concern into a quantity decision makers can interpret. Section 2 details the data, the comparative design, and the methods. Section 3 reports the results, Section 4 discusses their implications, and Section 5 concludes.

2. Methodology

2.1 Setting, Instrument, and Expert Panel

The empirical setting is the sustainability program of rural hospitals in Taiwan, where sparse transport infrastructure, workforce attrition, and constrained budgets jointly threaten the continuity of care. A structured questionnaire was administered in October 2024 to twenty experts with professional knowledge of rural healthcare operations and management [EXPERT PANEL DETAILS: positions, affiliations profile, and experience to be added]. The instrument contained two parts. The first elicited the relative importance of four risk evaluation factors: severity of consequences (S), occurrence probability (O), detectability before occurrence (D), and expected cost of prevention or remediation (E). Each expert ordered the four factors from most to least important and stated, on a nine-point ratio scale, the importance of each factor relative to the next factor in that ordering. The second part asked each expert to rate twelve failure modes across the four factors using ten-point anchored scales, with higher scores indicating greater severity, higher occurrence probability, poorer detectability, and higher expected cost. Responses were de-identified prior to analysis, and no patient data were involved [ETHICS WORDING TO BE CONFIRMED BY AUTHOR]. The four-factor design extends the conventional severity, occurrence, and detection triple with expected cost because the decision the instrument serves is explicitly budgetary: a rural hospital prioritizing mitigations must weigh what a failure would do against what preventing it would consume, and a prioritization that ignores remediation cost systematically favors expensive fixes for visible risks over affordable fixes for quiet ones.

The twelve failure modes were developed from the rural healthcare sustainability literature and program documents and are listed with working definitions in Table 1. They span resource and finance (FM1, FM2, FM3), access and technology (FM4, FM5, FM6), workforce and quality (FM7, FM8), and organization and community (FM9 through FM12).

Table 1
Failure modes for rural hospital sustainable healthcare

ID	Failure mode	Working definition
FM1	Insufficient resource allocation	Limited funding and staffing prevent the hospital from sustaining service continuity and quality.
FM2	Policy change risk	Shifts in government support alter the funding and resource commitments that underpin sustainability programs.
FM3	Difficult implementation of eco-friendly measures	Financial and technical constraints impede energy-saving equipment and green-facility initiatives.
FM4	Transportation inaccessibility	Weak transport infrastructure restricts outreach services and patient access, particularly in adverse conditions.
FM5	Unstable telemedicine technology	Connectivity issues and platform instability disrupt remote consultations and timely diagnosis.

Table 1
Continued

ID	Failure mode	Working definition
FM6	Aging medical equipment	Constrained maintenance budgets leave long-serving equipment prone to failure and degraded performance.
FM7	Medical workforce shortage	Difficult living conditions drive staff attrition, undermining team stability and professional capacity.
FM8	Difficulty controlling service quality	Expanding service scope outpaces monitoring capacity, raising the risk of adverse events and dissatisfaction.
FM9	Weak interdepartmental coordination	Insufficient communication across health, government, community, and business actors impairs execution.
FM10	Insufficient community participation	Low resident awareness and engagement limit program uptake and effective use of resources.
FM11	Unmet mental health needs	Scarce professional mental health services leave local psychological needs unaddressed.
FM12	Difficult health-education outreach	Low health literacy hampers preventive education and the early detection of health problems.

Note: Factor scales: S, O, D, and E, rated from 1 to 10 with anchored descriptors; higher values indicate higher risk on each factor, with D coded so that 10 denotes the poorest detectability.

2.2 Data Screening

Screening of the 960 ratings (12 failure modes by 20 experts across 4 factors) identified exactly one out-of-range entry: expert 15 recorded 23 on the detection factor for FM5, outside the 1 to 10 scale. The entry was treated as a transcription error and replaced by the cross-expert median of the FM5 detection ratings, which equals 4. Because any treatment of a single cell is somewhat arbitrary, Section 3.7 repeats the entire analysis with the value imputed as 2, 3, and 10; no rank in any configuration moves by more than one position under any of these treatments. All remaining ratings fell within scale, and every expert completed both parts of the instrument, so no other cleaning was required.

2.3 Comparative Design

The design principle is to hold the data fixed and vary only the analysis. From the screened ratings, one group decision matrix is constructed (Section 2.5) and submitted to a grid of ranking configurations formed by crossing eight MCDM ranking methods with three weighting schemes, yielding twenty-four configurations, to which the four-factor product RPN4 is added as a twenty-fifth. The classical three-factor RPN3 is computed as an additional reference outside the grid, since it ignores the cost factor that the instrument deliberately collected. The eight methods were selected to span the main aggregation families in the FMEA literature: distance to ideal solutions (TOPSIS), compromise with a regret component (VIKOR), ratio-based compromise (MARCOS, CoCoSo), gap analysis (MAIRCA), distance from average (EDAS), and weighted aggregation of sum and product forms (WASPAS, ARAS). The three weighting schemes span the standard positions: agnostic equal weights, the panel's own stated importance, and dispersion-based objective weights. Every configuration receives the same matrix, so any difference among the twenty-five orderings is attributable to the analysis alone. Method parameters follow their conventional midpoints throughout, with the VIKOR trade-off and the CoCoSo and WASPAS mixing parameters at 0.5, so that no configuration is tuned toward or away from agreement; the grid compares the methods as an informed practitioner would encounter them, at their textbook settings.

Four evaluation layers are then applied. Pairwise Kendall rank correlations quantify overall agreement among configurations. Rank bandwidth locates the failure modes whose positions are configuration-dependent. A leave-one-alternative-out protocol tests each method for rank reversal

when the alternative set changes. A leave-one-expert-out protocol reruns the entire grid twenty times with one panelist removed, which prices method-induced variance against panel-induced variance on a common scale.

2.4 Data Statement

Aggregated matrices and complete computational results are provided in the supplementary workbook, and de-identified raw responses are available from the author on reasonable request.

2.5 Notation and Group Aggregation

Let x(k) with elements x(k) sub ij denote the rating of failure mode i on factor j by expert k, for i = 1, ..., m failure modes, j = 1, ..., n factors, and k = 1, ..., K experts, with m = 12, n = 4, and K = 20. Individual ratings are aggregated by the arithmetic mean into the group decision matrix in Eq. (1), the standard aggregation for cardinal group judgments [37]. All four factors are benefit-oriented toward risk, so larger entries indicate greater risk and rank 1 denotes the highest priority failure mode throughout.

$$\bar{x}_{ij} = \frac{1}{K}\sum_{k=1}^{K} x_{ij}^{(k)} \tag{1}$$

2.6 Weighting Schemes

Three weight vectors are used. The equal scheme (EQ) sets every weight to 1/n and represents the agnostic position implicit in the classical RPN. The questionnaire scheme (SUB) converts each expert's chained importance ratios into a weight vector and aggregates across experts. Writing the expert's ordering of the factors as a permutation and the stated ratio of the p-th to the (p+1)-th factor as r sub p, weights are recovered recursively as in Eq. (2), normalized to sum to one, and combined across experts by the normalized geometric mean in Eq. (3), the aggregation that preserves ratio judgments [37]. The construction is the chained-ratio counterpart of structured subjective weighting families such as the best-worst method [38], and its behavior in group settings mirrors that of Bayesian extensions of such families [35].

$$w_{\sigma(n)} = 1, \quad w_{\sigma(p)} = r_p\, w_{\sigma(p+1)}, \quad p = n-1, \ldots, 1 \tag{2}$$

$$w_j^{\mathrm{SUB}} = \frac{\left(\prod_{k=1}^{K} w_{jk}\right)^{1/K}}{\sum_{l=1}^{n}\left(\prod_{k=1}^{K} w_{lk}\right)^{1/K}} \tag{3}$$

The entropy scheme (ENT) derives objective weights from the dispersion of the group matrix [39]. With column-normalized shares p sub ij, the entropy of factor j and the resulting weight are given in Eq. (4) and Eq. (5); factors along which failure modes differ more receive larger weights.

$$e_j = -\frac{1}{\ln m}\sum_{i=1}^{m} p_{ij}\ln p_{ij}, \quad p_{ij} = \frac{\bar{x}_{ij}}{\sum_{i=1}^{m}\bar{x}_{ij}} \tag{4}$$

$$w_j^{\mathrm{ENT}} = \frac{1-e_j}{\sum_{l=1}^{n}(1-e_l)} \tag{5}$$

2.7 Ranking Methods

Each method receives the group matrix and one weight vector and returns a complete ordering of the twelve failure modes. The formulations follow the original sources; only the scoring functions are restated here, with method-specific normalizations noted. Throughout, w sub j denotes the active weight of factor j. The methods differ not only in their scoring functions but in the reference objects those functions consult: TOPSIS and MARCOS score against ideal profiles constructed from the data,

EDAS scores against column averages, ARAS against an appended optimal alternative, VIKOR and MAIRCA against gap structures, and WASPAS and the product baselines against fixed column maxima or against nothing outside the alternative's own row. These reference structures are catalogued here because Section 3.5 shows that they, rather than the scoring functions themselves, govern susceptibility to rank reversal when the alternative set changes.

TOPSIS ranks by relative closeness to the ideal solution [40]. With vector-normalized weighted ratings, distances d plus and d minus to the ideal and anti-ideal profiles yield the closeness coefficient in Eq. (6), ranked in descending order. Because the vector normalization divides each column by the root of its sum of squares, the effective spread of a criterion after normalization depends on the dispersion of its column, a property that matters for the detection factor in Section 3.3. Applied audits of TOPSIS-family pipelines in performance evaluation follow the same construction [34].

$$CC_i = \frac{d_i^-}{d_i^+ + d_i^-} \tag{6}$$

VIKOR ranks by a compromise between group utility and individual regret [41, 42]. With ideal and worst factor values f star and f minus, Eq. (7) defines the weighted average gap S sub i and the maximum weighted gap R sub i, and Eq. (8) blends them with trade-off parameter v, set to 0.5; smaller Q is better, so ranks ascend in Q. The regret term makes VIKOR the only method in the grid whose aggregation is deliberately non-compensatory at the margin: sufficiently poor performance on one criterion cannot be fully offset by strength on the others, which is precisely the behavior a cautious risk owner may want and a throughput-oriented one may not.

$$S_i = \sum_{j=1}^{n} w_j \frac{f_j^* - \bar{x}_{ij}}{f_j^* - f_j^-}, \quad R_i = \max_j \; w_j \frac{f_j^* - \bar{x}_{ij}}{f_j^* - f_j^-} \tag{7}$$

$$Q_i = v \, \frac{S_i - S^*}{S^- - S^*} + (1 - v) \, \frac{R_i - R^*}{R^- - R^*} \tag{8}$$

MARCOS scores each alternative against appended ideal and anti-ideal rows [43]. Utility ratios K plus and K minus against the two references combine into the utility function of Eq. (9), ranked in descending order; the method's use inside weighted evaluation pipelines follows Lo and Lin [36]. Because the ideal and anti-ideal rows are recomputed from whatever alternatives are present, the method is in principle exposed to set dependence, although Section 3.5 shows the exposure to be negligible on this matrix.

$$f(K_i) = \frac{K_i^+ + K_i^-}{1 + \frac{1 - f(K_i^+)}{f(K_i^+)} + \frac{1 - f(K_i^-)}{f(K_i^-)}} \tag{9}$$

CoCoSo combines a weighted sum S sub i and a weighted power product P sub i of min-max normalized ratings through three appraisal strategies, aggregated in Eq. (10) with balance parameter set to 0.5 [44]; larger k is better. The min-max normalization maps the worst observed value of each criterion exactly to zero, so an alternative sitting at a column minimum contributes nothing to the power-product component on that criterion, a discontinuity visible in the behavior of the workforce mode under equal weights.

$$k_i = (k_{ia} \, k_{ib} \, k_{ic})^{1/3} + \frac{k_{ia} + k_{ib} + k_{ic}}{3} \tag{10}$$

MAIRCA measures the gap between an equal a priori preference allocation and the realized weighted performance [45]. With theoretical evaluation t sub pj equal to w sub j divided by m and realized evaluation t sub rij on min-max normalized ratings, the total gap in Eq. (11) is ranked in ascending order. Since the theoretical evaluation spreads preference equally across alternatives, the

method rewards profiles that convert their share of preference into realized performance uniformly across the criteria.

$$Q_i = \sum_{j=1}^{n} \left(t_{pj} - t_{rij} \right) \tag{11}$$

EDAS evaluates alternatives by their positive and negative distances from the column averages [46]. Weighted sums of the two distance types, normalized by their maxima, average into the appraisal score of Eq. (12), ranked in descending order. The average-based reference shifts whenever the alternative set changes, making EDAS the natural candidate for mild set dependence among additive designs.

$$AS_i = \frac{1}{2} (NSP_i + NSN_i) \tag{12}$$

WASPAS blends the weighted sum and weighted product models on max-normalized ratings with mixing parameter set to 0.5 [47], as in Eq. (13), and ARAS ranks by the ratio of each alternative's additive utility to that of an appended optimal alternative on column-sum normalized ratings [48], as in Eq. (14); both rank in descending order. WASPAS consults only the column maxima and ARAS an appended optimal row, so their reference objects move only when a removed alternative happens to hold a column extreme, and the product component of WASPAS inherits the multiplicative penalty structure of the RPN in attenuated form.

$$Q_i = \lambda \sum_{j=1}^{n} w_j \; r_{ij} + (1 - \lambda) \prod_{j=1}^{n} r_{ij}^{\,w_j} \tag{13}$$

$$K_i = \frac{S_i}{S_0} \tag{14}$$

Finally, the product baselines are the classical three-factor index and its four-factor extension in Eq. (15), both ranked in descending order. RPN4 uses the same information set as the MCDM configurations and belongs to the grid; RPN3 discards the cost factor and serves as the classical reference. Each product depends only on the alternative's own row, so the baselines are structurally immune to rank reversal by construction, an immunity that Section 3.5 confirms and Section 4 places against their validity problems.

$$RPN_{3,i} = \bar{x}_{iS}\, \bar{x}_{iO}\, \bar{x}_{iD}, \quad RPN_{4,i} = \bar{x}_{iS}\, \bar{x}_{iO}\, \bar{x}_{iD}\, \bar{x}_{iE} \tag{15}$$

2.8 Comparison Metrics

Agreement between two configurations is measured by the Kendall tau-b rank correlation [49], which corrects for ties, as in Eq. (16), where C and Dc are the numbers of concordant and discordant pairs and T1 and T2 are tie corrections. Positional instability of failure mode i is summarized by the rank bandwidth of Eq. (17), the spread of its rank across the twenty-five configurations, complemented by the standard deviation of its ranks. Consistency of the action set is measured by the mean pairwise Jaccard similarity of the top-three sets across configurations.

$$\tau_b = \frac{C - D_c}{\sqrt{(C + D_c + T_1)(C + D_c + T_2)}} \tag{16}$$

$$B_i = \max_c \mathrm{rank}_{ic} - \min_c \mathrm{rank}_{ic} \tag{17}$$

Rank reversal is tested by single-alternative removal: for each configuration, each failure mode is deleted in turn, the method is recomputed on the reduced matrix with weights held fixed, and the share of pairwise orderings among the surviving alternatives that flip relative to the original configuration is recorded, as in Eq. (18), where F sub a is the set of flipped pairs after removing alternative a. Panel sensitivity is measured by a leave-one-expert-out protocol: the entire pipeline,

including the SUB and ENT weight vectors, is recomputed twenty times with one expert excluded, and each replicate ranking is compared with the full-panel ranking of the same configuration by tau-b. All computations were performed in Python 3 with a single reproducible script.

$$\varphi_c = \frac{1}{m}\sum_{a=1}^{m}\frac{|F_a|}{C(m-1,2)} \tag{18}$$

3. Results

3.1 Criterion Structure and Weights

Table 2 reports the group decision matrix with cross-expert standard deviations. Two features organize everything that follows. First, the workforce shortage mode FM7 has an extreme profile: it dominates on severity (8.30), occurrence (8.45), and expected cost (8.45) while recording the lowest detection score in the matrix (3.00), meaning the panel considers it grave, likely, expensive, and easy to see coming. Second, the detection factor opposes the other three across the whole matrix: the correlations of D with S, O, and E on the group matrix are negative and strong (r = −0.75, −0.81, and −0.84), while S, O, and E are nearly collinear with pairwise correlations above 0.94. The failure modes the panel judges most dangerous are, in this dataset, also the ones it judges easiest to detect, so any aggregation must reconcile one dissenting criterion with three agreeing ones. Cross-expert dispersion is moderate, with standard deviations between 1.1 and 2.3 rating points across the matrix, so the aggregate structure is not the artifact of a few extreme raters; the detection column shows the largest average dispersion of the four factors, indicating that visibility before occurrence is also where the individual experts disagree most.

Table 2
Group decision matrix: mean expert rating with standard deviation in parentheses

ID	Severity (S)	Occurrence (O)	Detection (D)	Expected cost (E)
FM1	7.00 (1.45)	6.95 (1.70)	4.65 (2.11)	7.30 (1.78)
FM2	6.30 (1.59)	5.95 (1.57)	5.05 (1.47)	6.55 (1.88)
FM3	4.70 (1.81)	5.75 (1.48)	4.30 (1.84)	5.70 (1.49)
FM4	5.85 (1.35)	6.45 (1.39)	4.15 (2.32)	6.50 (1.79)
FM5	5.20 (1.20)	5.00 (1.34)	4.55 (1.70)	5.70 (1.72)
FM6	6.55 (1.39)	6.65 (1.50)	3.25 (1.16)	7.70 (1.26)
FM7	8.30 (1.13)	8.45 (1.19)	3.00 (1.72)	8.45 (1.23)
FM8	6.25 (1.48)	6.10 (1.80)	4.40 (1.57)	6.15 (1.69)
FM9	4.85 (1.27)	5.00 (1.17)	4.80 (1.54)	5.00 (1.21)
FM10	4.00 (1.34)	4.95 (1.32)	5.30 (1.81)	4.30 (1.17)
FM11	4.30 (1.56)	4.85 (1.50)	5.55 (1.79)	4.60 (1.27)
FM12	4.10 (1.74)	4.75 (1.68)	5.05 (1.73)	4.25 (1.33)

Note: Means over twenty experts after the single-cell imputation described in Section 2.2; higher values denote higher risk on every factor.

Table 3 reports the three weight vectors. Twelve of the twenty experts named severity the most important factor, five named expected cost, two named detection, and one named occurrence, and the aggregated questionnaire weights reflect this concentration: severity carries 0.554, expected cost 0.213, detection 0.126, and occurrence 0.108. The entropy weights, driven by dispersion, also favor severity (0.320) and cost (0.300) over occurrence (0.200) and detection (0.180). Both informed schemes therefore assign detection the least influence, whereas the equal scheme, and by construction the classical RPN, give the dissenting criterion a full quarter of the vote. This single observation anticipates the pattern of disagreement in Section 3.3.

Table 3
Weighting schemes over the four risk factors

Scheme	S	O	D	E
Equal (EQ)	0.250	0.250	0.250	0.250
Questionnaire (SUB)	0.554	0.107	0.126	0.213
Entropy (ENT)	0.320	0.200	0.180	0.300
Chosen most important (of 20 experts)	12	1	2	5

Note: SUB aggregates individual chained-ratio weight vectors by the normalized geometric mean; the final row reports how many experts named each factor most important.

3.2 Rankings and Overall Convergence

Table 4 presents the complete ranking table for the twenty-five grid configurations, organized into three panels by weighting scheme, with the RPN3 reference alongside the equal-weight panel. Figure 1 maps the pairwise tau-b correlations among the twenty-five configurations. Agreement is high in the aggregate: the mean pairwise correlation is 0.931, the top-three action set is identical or nearly identical across most configurations with a mean pairwise Jaccard similarity of 0.713, and the extremes of the ordering are effectively locked. FM1 appears in the top three in all twenty-five configurations and FM7 in twenty-four; FM7 ranks first in twenty-two configurations and FM1 in the remaining three; FM12 and FM10 occupy the bottom two positions almost everywhere. A decision maker who acts only on the extremes of this ranking would take the same action under any configuration in the grid.

Table 4
Risk ranks of the twelve failure modes across the twenty-five grid configurations (rank 1 = highest priority)

ID	TOP	VIK	MAR	CoC	MAI	EDA	WAS	ARA	RPN3	RPN4
Panel A. Equal weights (EQ), with product baselines										
FM1	2	1	2	1	2	2	2	2	1	2
FM2	4	3	3	3	3	3	3	3	3	3
FM3	9	7	8	7	8	8	8	8	9	8
FM4	6	2	6	6	6	6	6	6	5	6
FM5	8	8	7	8	7	7	7	7	7	7
FM6	3	6	4	4	4	5	4	4	6	4
FM7	1	5	1	2	1	1	1	1	2	1
FM8	5	4	5	5	5	4	5	5	4	5
FM9	11	9	10	9	10	9	9	9	8	9
FM10	10	11	11	11	11	11	11	11	11	11
FM11	7	10	9	10	9	10	10	10	10	10
FM12	12	12	12	12	12	12	12	12	12	12
Panel B. Questionnaire weights (SUB)										
FM1	2	2	2	2	2	2	2	2	.	.
FM2	4	4	4	4	4	4	4	4	.	.
FM3	9	9	8	8	8	8	8	8	.	.
FM4	6	6	6	6	6	6	6	6	.	.
FM5	7	7	7	7	7	7	7	7	.	.
FM6	3	3	3	3	3	3	3	3	.	.
FM7	1	1	1	1	1	1	1	1	.	.
FM8	5	5	5	5	5	5	5	5	.	.
FM9	8	8	9	9	9	9	9	9	.	.
FM10	11	12	11	11	11	12	11	11	.	.
FM11	10	10	10	10	10	10	10	10	.	.
FM12	12	11	12	12	12	11	12	12	.	.

Table 4
Continued

ID	TOP	VIK	MAR	CoC	MAI	EDA	WAS	ARA	RPN3	RPN4
Panel C. Entropy weights (ENT)										
FM1	2	1	2	2	2	2	2	2	.	.
FM2	4	3	4	3	4	3	4	4	.	.
FM3	8	8	8	8	8	8	8	8	.	.
FM4	6	6	6	6	6	6	6	6	.	.
FM5	7	7	7	7	7	7	7	7	.	.
FM6	3	4	3	4	3	4	3	3	.	.
FM7	1	2	1	1	1	1	1	1	.	.
FM8	5	5	5	5	5	5	5	5	.	.
FM9	9	9	9	9	9	9	9	9	.	.
FM10	11	11	11	11	11	11	11	11	.	.
FM11	10	10	10	10	10	10	10	10	.	.
FM12	12	12	12	12	12	12	12	12	.	.

Note. TOP = TOPSIS, VIK = VIKOR, MAR = MARCOS, CoC = CoCoSo, MAI = MAIRCA, EDA = EDAS, WAS = WASPAS, ARA = ARAS. RPN3 and RPN4 are weight-free and shown once, in Panel A; RPN4 is the twenty-fifth grid configuration.

Reading Table 4 down the panels shows the texture behind these statistics. Under the questionnaire and entropy weights, the eight methods produce almost interchangeable columns: FM7 first, FM1 second, FM6 and FM2 alternating between third and fourth, FM8 fifth, FM4 sixth, and the community-facing modes closing the order, with only single-position swaps involving FM3, FM9, and the bottom pair. The equal-weight panel is where the columns diverge: TOPSIS lifts FM11, the mode with the highest detection score, three positions to seventh while reordering the flat-profile coordination mode FM9 to eleventh; VIKOR restructures the entire upper half as examined below; and the remaining six methods stay close to the informed-weight consensus. The disagreement is therefore not diffuse noise spread over the grid; it is localized in one weighting scheme and, within that scheme, in two methods whose geometries process the detection factor differently.

Beneath the aggregate agreement, Figure 1 shows visible structure: the questionnaire-weight and entropy-weight blocks are internally almost unanimous, with within-block mean correlations of 0.974 and 0.976, while the equal-weight block is markedly noisier, with a within-block mean of 0.869 and the grid minimum of 0.576 between TOPSIS and VIKOR under equal weights. Decomposing the grid the other way, holding the method fixed while changing the weighting scheme yields a mean correlation of 0.927, and holding the weights fixed while changing the method yields 0.940, so method choice and weight choice contribute variance of the same order. Table 5 collects these agreement statistics.

3.3 Where the Methods Disagree: The Dissenting Criterion

Figure 2 resolves the disagreement to individual failure modes by plotting, for each mode, its full set of twenty-five ranks. The band structure is sharply bimodal. Seven of the twelve modes have a rank bandwidth of at most one position, including the entire top of the ordering (FM1 spans ranks 1 to 2, FM2 spans ranks 3 to 4, FM8 spans ranks 4 to 5) and the entire bottom (FM12 spans ranks 11 to 12). The instability concentrates in a middle band and, in one exceptional case, at the top: FM4 spans ranks 2 to 6, FM7 spans ranks 1 to 5, and FM6, FM9, and FM11 each span three positions.

The exceptional case is the informative one. FM7, the consensus top risk, is demoted to fifth by exactly one configuration, VIKOR under equal weights, which simultaneously promotes the balanced transportation mode FM4 from the consensus sixth position to second. The mechanism is the regret term: FM7 sits at the worst observed value of the detection factor, so its individual regret R sub i

attains the maximum possible weighted gap on D, and with equal weights that single gap is large enough to overturn the advantage conferred by three dominant factors. Compensatory aggregations allow severity, occurrence, and cost to offset the detection deficit; the regret component deliberately refuses that offset. Under the questionnaire and entropy weights, both of which cut the influence of detection by roughly half, the same method returns FM7 to second and first place, and TOPSIS, whose equal-weight geometry mildly elevates the high-detection mode FM11 for the mirror-image reason, likewise falls into line. The practical reading is precise: whether workforce shortage is the first priority of the sustainability program or its fifth depends not on the experts, who were unanimous in the aggregate, but on whether the analyst's aggregation permits strong factors to compensate for a weak one, and on how much voice the dissenting criterion receives.

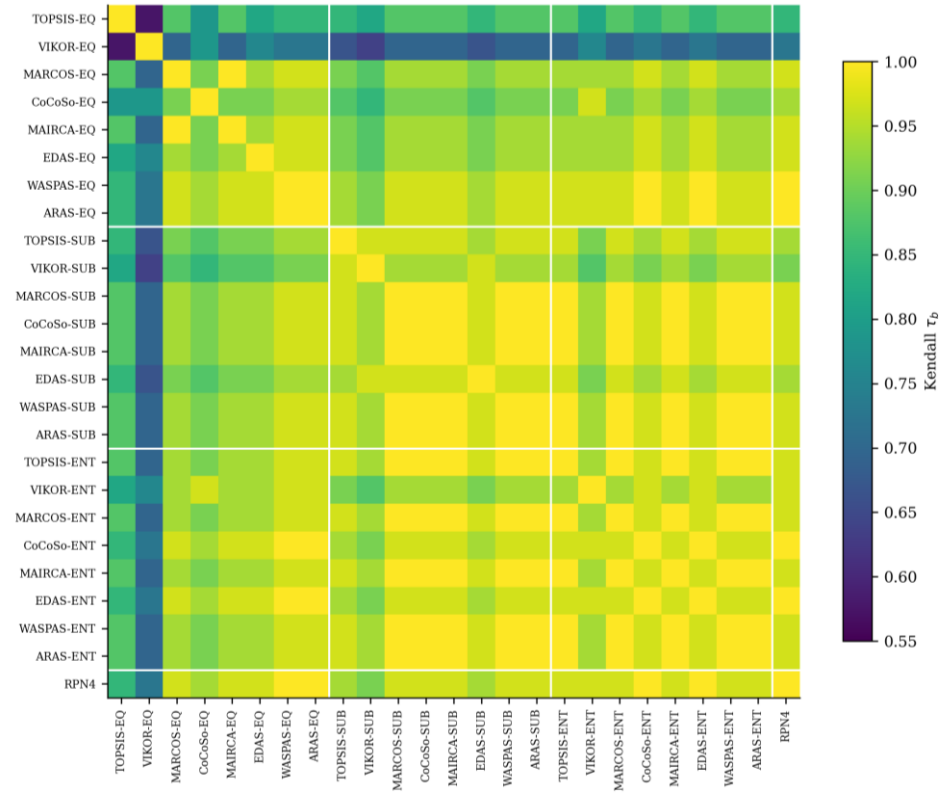


Fig. 1. Pairwise Kendall tau-b correlations among the twenty-five ranking configurations, ordered by weighting block

Table 5
Agreement structure across configurations (Kendall tau-b)

Comparison	Mean	Minimum
All 25 configurations	0.931	0.576
Within equal-weight block (8 methods)	0.869	0.576
Within questionnaire-weight block (8 methods)	0.974	0.939
Within entropy-weight block (8 methods)	0.976	0.939
Same method, different weighting schemes	0.927	.
Same weighting scheme, different methods	0.940	.
RPN3 against the 24 MCDM configurations	0.852	.
RPN4 against the 24 MCDM configurations	0.953	.
RPN3 against RPN4	0.879	.

Note. Entries marked with a period are not applicable. The grid minimum of 0.576 occurs between TOPSIS and VIKOR under equal weights.

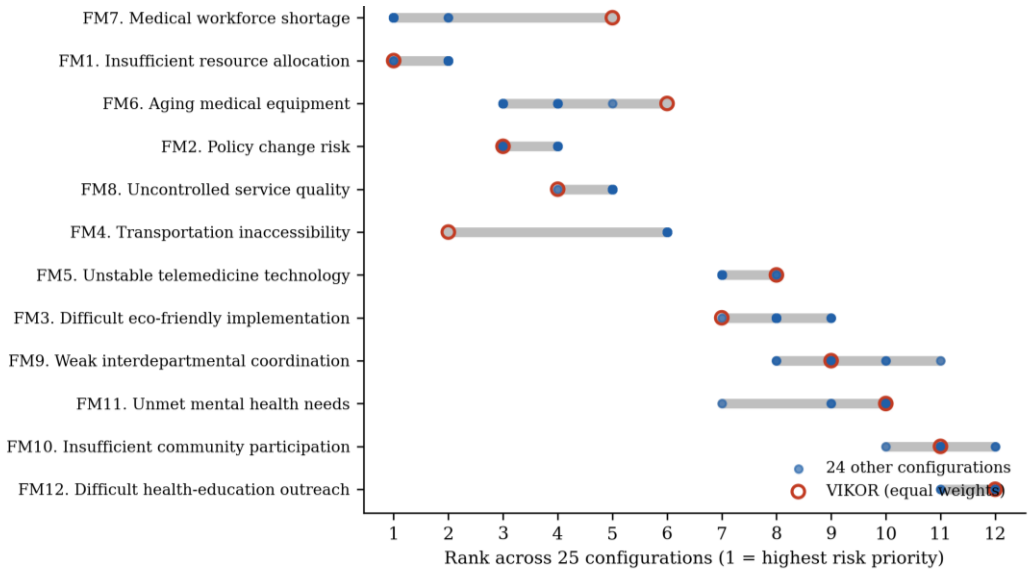


Fig. 2. Rank of each failure mode across the twenty-five configurations, sorted by median rank; open markers identify VIKOR under equal weights

The mirror image of the demotion appears at the bottom of the equal-weight panel. The community-facing modes FM10, FM11, and FM12 combine the weakest severity, occurrence, and cost scores in the matrix with the strongest detection scores, and equal weighting hands that lone strength a quarter of the total influence. TOPSIS, whose vector normalization preserves the comparatively large dispersion of the detection column, responds by lifting FM11 to seventh place; under the questionnaire and entropy weights, the lift disappears, and the community-facing modes settle into the final positions in every method. The two visible disagreements in the grid are thus a single phenomenon observed from both ends: a criterion that dissents from its peers, granted equal voice, is processed differently by different aggregation geometries, and the modes with extreme profiles on that criterion absorb the difference.

3.4 The Classical Baseline

The product baselines sharpen the point. RPN3, the classical index, is the configuration farthest from the modern consensus, with a mean correlation of 0.852 against the twenty-four MCDM configurations, and it disagrees with them on the question that matters most: it places FM1 first and FM7 second, because multiplying by the low detection score of FM7 penalizes the product directly. Adding the cost factor in RPN4 raises the mean correlation to 0.953 and restores FM7 to first place, in agreement with twenty-one of the twenty-four MCDM configurations. The panel's own weight statements compound the critique: the experts collectively assign detection an eighth to a fifth of the influence of severity, while the product form implicitly weights the factors equally. The classical index thus contradicts both the modern methods and the panel it summarizes, and Section 3.5 will add the irony that it is nonetheless the most structurally stable configuration in the grid. The distance of RPN3 from the consensus has a mechanical reading worth stating. A product treats a low score on any factor as a divisor of overall risk, so the detection scores act on the classical index exactly as the regret term acts on VIKOR, but with equal implicit weights and without the informed correction that returned VIKOR to the consensus; the classical index hard-codes the most divergence-prone configuration in the grid, equal influence for the dissenting criterion inside a non-compensating functional form.

3.5 Rank Reversal under Single-Alternative Removal

Table 6, Panel A, and Figure 3 report the leave-one-alternative-out test. Reversal is rare everywhere: the worst mean flipped-pair share is 1.97 percent, recorded by TOPSIS under equal weights, where nine of the twelve single removals flip at least one pair; EDAS shows mild set dependence through its data-driven average reference; VIKOR, CoCoSo, MAIRCA, and MARCOS flip occasional pairs; and WASPAS, ARAS, and the product baselines record no reversal at all in any of the thirty-six removal experiments. The immunity of the products is structural, since each alternative's score depends only on its own row, and the near-immunity of WASPAS and ARAS reflects reference points that this dataset leaves essentially undisturbed under removal. Method choice therefore matters far more through the compensation channel of Section 3.3 than through classical rank reversal, which on this real matrix is a second-order concern. The pattern in Panel A tracks the reference structures catalogued in Section 2.7 almost exactly: the methods that consult data-dependent references, the ideal profiles of TOPSIS and the average solution of EDAS, account for nearly all observed reversals, and the reversals concentrate under equal weights, where the dissenting detection criterion keeps those reference objects sensitive to which alternatives are present, while the methods whose references are fixed or trivially perturbed record none. Reversal on this matrix is not an exotic pathology; it is a predictable footprint of reference-point design, and it is small.

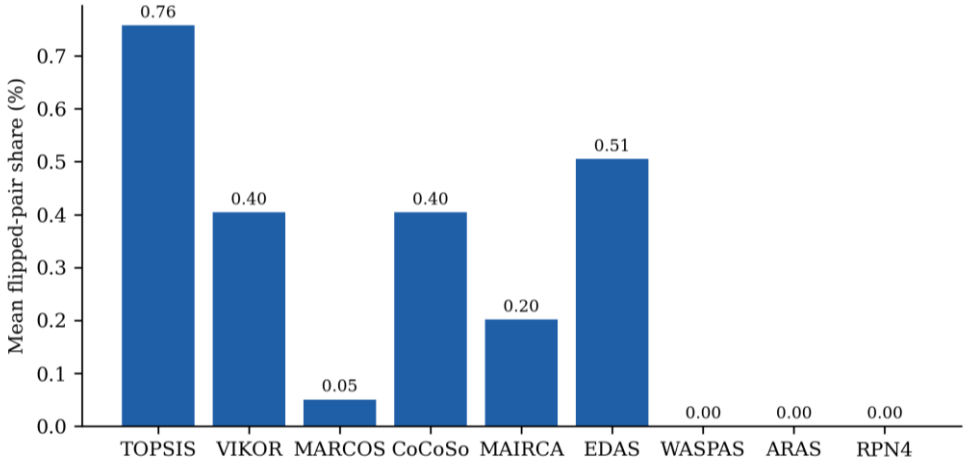


Fig. 3. Mean flipped-pair share under single-alternative removal, averaged over weighting schemes

Table 6
Structural robustness of the configurations

Panel A. Rank reversal under single-alternative removal				
Method	EQ flip (%)	SUB flip (%)	ENT flip (%)	Removals with a flip (EQ / SUB / ENT, of 12)
TOPSIS	1.97	0.00	0.30	9 / 0 / 2
VIKOR	0.45	0.00	0.76	2 / 0 / 2
MARCOS	0.15	0.00	0.00	1 / 0 / 0
CoCoSo	0.45	0.30	0.45	1 / 1 / 3
MAIRCA	0.45	0.00	0.15	1 / 0 / 1
EDAS	0.61	0.76	0.15	4 / 5 / 1
WASPAS	0.00	0.00	0.00	0 / 0 / 0
ARAS	0.00	0.00	0.00	0 / 0 / 0
RPN4	0.00	.	.	0 / . / .

Table 6
Contined

Panel B. Leave-one-expert-out stability (tau-b against full panel)				
Method	Mean	Minimum	.	.
TOPSIS	0.961	0.848	.	.
VIKOR	0.973	0.879	.	.
MARCOS	0.974	0.939	.	.
CoCoSo	0.966	0.818	.	.
MAIRCA	0.973	0.939	.	.
EDAS	0.972	0.909	.	.
WASPAS	0.978	0.909	.	.
ARAS	0.979	0.909	.	.
RPN4	0.974	0.939	.	.

Note. Panel A reports the mean share of pairwise orderings that flip when one failure mode is removed and the method is recomputed with weights fixed; RPN4 is weight-free. Panel B pools the twenty leave-one-expert replicates over the applicable weighting schemes. Entries marked with a period are not applicable.

3.6 Panel Composition against Method Choice

The leave-one-expert-out protocol supplies the benchmark that makes the preceding numbers interpretable. Across all twenty-five configurations and twenty replicates, the mean correlation between a reduced-panel ranking and its full-panel counterpart is 0.972, with per-method means between 0.961 and 0.979 and a worst single replicate of 0.818 (Table 6, Panel B, and Figure 4). No failure mode moves more than three positions under any expert removal, and the top of the ordering barely moves, with FM7 and FM1 shifting at most one position. The comparison with Section 3.2 is the study's central quantitative finding: removing an entire expert from the panel perturbs the ranking less, on average, than switching the aggregation method does, since the mean cross-method correlation of 0.931 lies clearly below the mean leave-one-expert correlation of 0.972. The analyst's choice of configuration is a larger source of ranking variance than the identity of any single panelist, and it deserves at least the same reporting discipline that panel composition already receives. The per-mode view reinforces the reading. The largest displacement any expert removal produces anywhere in the grid is three positions, recorded by the transportation mode FM4, and the modes that move at all under panel perturbation are the same middle-band modes that move under method perturbation, so the two sources of variance point at the same alternatives. For the program owner, the operational translation is direct: no single change in panel composition and no defensible change of method alters who is first, second, eleventh, or twelfth; both kinds of sensitivity live in the contested middle, and that is where deliberative attention belongs.

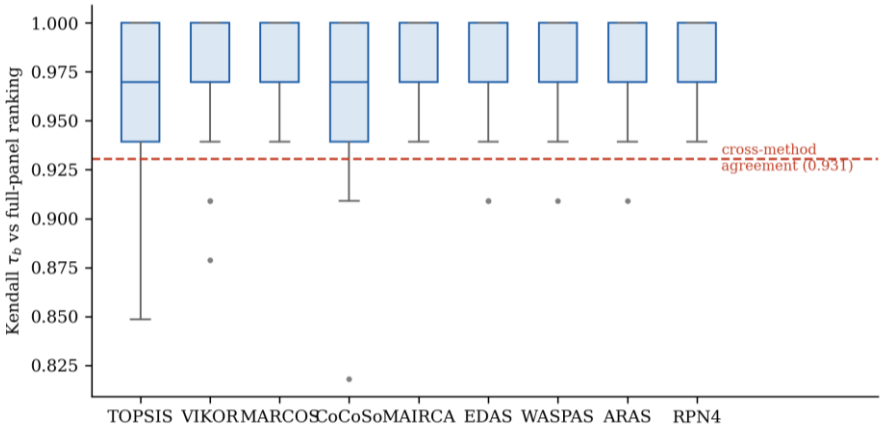


Fig. 4. Leave-one-expert-out agreement with the full-panel ranking by method; the dashed line marks the mean cross-method agreement, which every method exceeds

3.7 Sensitivity Checks

Two checks confirm that the findings do not hinge on incidental choices. First, replacing the single out-of-range detection entry in Section 2.2 with 2, 3, or 10 instead of the median 4 moves the rank in any configuration by no more than one position, and every statistic reported above remains unchanged to the stated precision. Second, the two configurations that drive the equal-weight disagreement, VIKOR and TOPSIS, converge to the consensus ordering under both informed weighting schemes, so the headline divergence is not an artifact of a particular parameterization but a property of the equal-weight treatment of the dissenting criterion; the trade-off parameter of VIKOR and the mixing parameters of CoCoSo and WASPAS were held at their conventional midpoints of 0.5 throughout.

4. Discussion

For the rural hospital program that motivated the data, the results are directly actionable. The extremes of the ordering are configuration-proof: resource allocation and workforce shortage lead under every defensible analysis, community participation and health-education outreach trail under every defensible analysis, and mitigation planning for these modes can proceed without methodological anxiety. The genuine managerial question lives in the middle band, where transportation access, equipment aging, coordination, and mental health provision trade places depending on the configuration. For decisions that hinge on ranks three through six, the analysis recommends triangulation rather than method loyalty: compute several aggregations, report the rank band, and treat a mode whose band straddles the funding cutoff as a case for deliberation rather than for mechanical selection.

For FMEA methodology, the study reframes two familiar debates. The first concerns the classical RPN. Its standard indictment, implicit equal weighting and product pathologies, is confirmed here in an unusually concrete form: the panel itself assigns detection a small fraction of the influence that the product form gives it, and the classical index is the farthest configuration from the modern consensus while disagreeing with it on the identity of the top risk. Yet the index is also completely immune to rank reversal, and its four-factor extension sits comfortably inside the consensus. The lesson is not that products are defensible but that stability and validity are separate properties, and criticisms of the RPN should say which one they are about. The second debate concerns compensation. The VIKOR demotion of the top risk is not an error; it is the designed behavior of a regret component confronting an alternative that is catastrophic on three factors and benign on one. Whether poor performance on detection should offset severity is a substantive question about risk attitude in healthcare, not a technicality, and the choice of aggregation method answers it silently. Analysts should answer it out loud instead, by stating whether the decision context calls for compensatory logic before selecting the machinery that embodies the answer.

A further implication concerns the meaning of the equal-weight default. Equal weighting is usually adopted as the neutral option, the choice an analyst makes to avoid making a choice. In this dataset, it is the least neutral configuration in the grid: the only scheme under which the methods disagree materially, precisely because it maximizes the influence of the criterion the panel itself considers least important, and because a criterion that conflicts with its peers converts influence into divergence. The panel's stated weights and the dispersion-based weights, obtained by entirely different logics, agree in muting the detection factor and thereby agree in producing near-unanimous rankings. When informed weights of either kind are available, declining to use them does not avoid a judgment; it substitutes a judgment that the data and the experts both contradict.

The comparison with panel perturbation carries a general implication for information-based decision practice. Decision reports routinely document who the experts were, how many there were,

and how their judgments were elicited, because readers understand that the panel shapes the answer. On this dataset, the analysis pipeline shapes the answer more. A reporting protocol follows naturally: alongside the point ranking, report the configuration grid or at least a small method ensemble, the rank bandwidth of each alternative, the agreement statistics among configurations, and an explicit statement of the compensation assumption. The cost of the protocol is minutes of computation; the benefit is that method sensitivity becomes visible exactly where it is decision-relevant, in the alternatives whose bands cross an action threshold. The prescription echoes robustness practices emerging in adjacent evaluation domains [32, 33] and extends them with a benchmark, panel perturbation, that practitioners already find meaningful.

The findings come with boundaries. They describe one panel of twenty experts, twelve failure modes, four factors, and one sector, and the sharp negative correlation of detection with the remaining factors, while substantively plausible for risks that announce themselves, is a feature of this matrix that other FMEA datasets may not share; in matrices where the criteria align, the equal-weight disagreement documented here should shrink. The grid, though broad, is finite, and we excluded outranking families because their pairwise structure does not deliver the complete cardinal-free ordering shown here. Group judgments entered through the arithmetic mean, and distributional treatments of expert heterogeneity lie beyond the scope of this study. The agreement statistics also weight all rank positions equally; correlation measures that emphasize the top of the ordering would sharpen the analysis for contexts where only the leading positions are funded, and the bandwidth and top-set statistics reported here are first steps in that direction.

5. Conclusions

This study examined how much the choice of aggregation method matters in MCDM-based FMEA and addressed this using a fixed, real expert dataset from rural healthcare sustainability. Across twenty-five configurations, agreement is high, and the extremes of the priority ordering are configuration-proof, but the disagreement that remains is systematic: it concentrates in the criterion that dissents from the others and in the compensation assumption; it is capable of moving the consensus top risk to fifth place, and it exceeds, on average, the perturbation caused by removing any single expert from the panel. The classical RPN emerges as simultaneously the least valid and the most structurally stable configuration, a combination that clarifies what its critics should be claiming. The constructive output is a low-cost reporting protocol that declares rank bandwidth and the compensation stance alongside the point ranking.

Three extensions follow. Applying the same audit to FMEA matrices from other sectors would establish whether the dissenting-criterion mechanism generalizes; embedding the grid inside distributional and fuzzy treatments of expert judgment would unify method sensitivity with input uncertainty in a single accounting; and eliciting decision makers' compensation preferences directly, before method selection, would convert the silent assumption identified here into a documented modeling choice. Each extension serves the same goal: to measure, report, and price the sensitivity of a risk ranking to its own analysis, rather than discover it in the decision it misleads.

Acknowledgement

This research was not funded by any grant.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Strasser, R. (2003). Rural health around the world: Challenges and solutions. Family Practice, 20(4), 457–463. https://doi.org/10.1093/fampra/cmg422

[2] Weinhold, I., & Gurtner, S. (2014). Understanding shortages of sufficient health care in rural areas. Health Policy, 118(2), 201–214. https://doi.org/10.1016/j.healthpol.2014.07.018

[3] DeRosier, J., Stalhandske, E., Bagian, J. P., & Nudell, T. (2002). Using health care failure mode and effect analysis: The VA National Center for Patient Safety's prospective risk analysis system. The Joint Commission Journal on Quality Improvement, 28(5), 248–267. https://doi.org/10.1016/s1070-3241(02)28025-6

[4] Liu, H.-C. (2019). Improved FMEA methods for proactive healthcare risk analysis. Springer. https://doi.org/10.1007/978-981-13-6366-5

[5] Bowles, J. B., & Pelaez, C. E. (1995). Fuzzy logic prioritization of failures in a system failure mode, effects and criticality analysis. Reliability Engineering & System Safety, 50(2), 203–213. https://doi.org/10.1016/0951-8320(95)00068-d

[6] Liu, H.-C., Liu, L., & Liu, N. (2013). Risk evaluation approaches in failure mode and effects analysis: A literature review. Expert Systems with Applications, 40(2), 828–838. https://doi.org/10.1016/j.eswa.2012.08.010

[7] Huang, J., You, J.-X., Liu, H.-C., & Song, M.-S. (2020). Failure mode and effect analysis improvement: A systematic literature review and future research agenda. Reliability Engineering & System Safety, 199, 106885. https://doi.org/10.1016/j.ress.2020.106885

[8] Braglia, M. (2000). MAFMA: Multi-attribute failure mode analysis. International Journal of Quality & Reliability Management, 17(9), 1017–1033. https://doi.org/10.1108/02656710010353885

[9] Kutlu, A. C., & Ekmekcioglu, M. (2012). Fuzzy failure modes and effects analysis by using fuzzy TOPSIS-based fuzzy AHP. Expert Systems with Applications, 39(1), 61–67. https://doi.org/10.1016/j.eswa.2011.06.044

[10] Wang, L.-E., Liu, H.-C., & Quan, M.-Y. (2016). Evaluating the risk of failure modes with a hybrid MCDM model under interval-valued intuitionistic fuzzy environments. Computers & Industrial Engineering, 102, 175–185. https://doi.org/10.1016/j.cie.2016.11.003

[11] Tian, Z.-P., Wang, J.-Q., & Zhang, H.-Y. (2018). An integrated approach for failure mode and effects analysis based on fuzzy best-worst, relative entropy, and VIKOR methods. Applied Soft Computing, 72, 636–646. https://doi.org/10.1016/j.asoc.2018.03.037

[12] Huang, J., Liu, H.-C., Duan, C.-Y., & Song, M.-S. (2022). An improved reliability model for FMEA using probabilistic linguistic term sets and TODIM method. Annals of Operations Research, 312(1), 235–258. https://doi.org/10.1007/s10479-019-03447-0

[13] Zanakis, S. H., Solomon, A., Wishart, N., & Dublish, S. (1998). Multi-attribute decision making: A simulation comparison of select methods. European Journal of Operational Research, 107(3), 507–529. https://doi.org/10.1016/s0377-2217(97)00147-1

[14] Ceballos, B., Lamata, M. T., & Pelta, D. A. (2016). A comparative analysis of multi-criteria decision-making methods. Progress in Artificial Intelligence, 5(4), 315–322. https://doi.org/10.1007/s13748-016-0093-1

[15] Salabun, W., Watrobski, J., & Shekhovtsov, A. (2020). Are MCDA methods benchmarkable? A comparative study of TOPSIS, VIKOR, COPRAS, and PROMETHEE II methods. Symmetry, 12(9), 1549. https://doi.org/10.3390/sym12091549

[16] Belton, V., & Gear, T. (1983). On a short-coming of Saaty's method of analytic hierarchies. Omega, 11(3), 228–230. https://doi.org/10.1016/0305-0483(83)90047-6

[17] Triantaphyllou, E. (2001). Two new cases of rank reversals when the AHP and some of its additive variants are used that do not occur with the multiplicative AHP. Journal of Multi-Criteria Decision Analysis, 10(1), 11–25. https://doi.org/10.1002/mcda.284

[18] Wang, X., & Triantaphyllou, E. (2008). Ranking irregularities when evaluating alternatives by using some ELECTRE methods. Omega, 36(1), 45–63. https://doi.org/10.1016/j.omega.2005.12.003

[19] Aires, R. F. de F., & Ferreira, L. (2018). The rank reversal problem in multi-criteria decision making: A literature review. Pesquisa Operacional, 38(2), 331–362. https://doi.org/10.1590/0101-7438.2018.038.02.0331

[20] Lin, S.-W. (2026). Strategic interactions and operational efficiency in supply chain networks: A game-theoretic network directional distance function approach. International Journal of Production Economics, 301, 110157. https://doi.org/10.1016/j.ijpe.2026.110157

[21] Lin, S.-W., & Lu, W.-M. (2025). A multi-period decision framework for mutual fund efficiency: Integrating range directional DEA with machine learning for dynamic investment optimization. INFOR: Information Systems and Operational Research, 64(3), 649–683. https://doi.org/10.1080/03155986.2025.2572139

[22] Lin, S.-W., & Lu, W.-M. (2026). An integrated data envelopment analysis–machine learning framework for evaluating pension fund management efficiency. Applied Soft Computing, 197, 115154. https://doi.org/10.1016/j.asoc.2026.115154
[23] Lin, S.-W., & Lu, W.-M. (2026). A strategic framework for supplier performance assessment and anomaly detection in high-tech supply chains: Integrating shared-resource DEA with machine learning. IEEE Transactions on Engineering Management, 73, 2470–2484. https://doi.org/10.1109/tem.2026.3675293
[24] Lin, S.-W., Lo, H.-W., & Lu, W.-M. (2026). Mapping the evolution of low-carbon supply chain research: A systematic review with non-negative matrix factorization-assisted topic modeling. Process Integration and Optimization for Sustainability. Advance online publication. https://doi.org/10.1007/s41660-026-00858-y
[25] Lin, S.-W. (2026). Evaluating renewable energy policy effectiveness: A quantitative approach for decision support and risk mitigation. Annals of Operations Research. Advance online publication. https://doi.org/10.1007/s10479-026-07296-6
[26] Lin, S.-W., & Lu, W.-M. (2026). Proactive supplier performance reliability prediction: A chance-constrained network DEA and machine learning integration. Annals of Operations Research. Advance online publication. https://doi.org/10.1007/s10479-026-07371-y
[27] Lo, H.-W., Shiue, W., Liou, J. J. H., & Tzeng, G.-H. (2020). A hybrid MCDM-based FMEA model for identification of critical failure modes in manufacturing. Soft Computing, 24(20), 15733–15745. https://doi.org/10.1007/s00500-020-04903-x
[28] Lo, H.-W., Chen, T.-H., Fang, T.-Y., & Lin, S.-W. (2025). Airport sustainability and risk assessment using interval-valued Fermatean fuzzy MCDM network analysis. Research in Transportation Business & Management, 62, 101454. https://doi.org/10.1016/j.rtbm.2025.101454
[29] Leeftink, A. G., Visser, J., de Laat, J. M., van der Meij, N. T. M., Vos, J. B. H., & Valk, G. D. (2021). Reducing failures in daily medical practice: Healthcare failure mode and effect analysis combined with computer simulation. Ergonomics, 64(10), 1322–1332. https://doi.org/10.1080/00140139.2021.1910734
[30] Tartaglia, R., Parretti, C., Candido, G., La Regina, M., Scaldaferri, F., Rumi, G., Napolitano, D., Vetrugno, G., & Barach, P. (2026). Implementing a televisiting program in a tertiary university hospital: Failure modes and effect analysis (FMEA) and solutions for improving patient safety and sustainable care. Safety Science, 196, 107096. https://doi.org/10.1016/j.ssci.2025.107096
[31] Triantaphyllou, E. (2000). Multi-criteria decision making methods: A comparative study. Springer. https://doi.org/10.1007/978-1-4757-3157-6
[32] Watrobski, J., Jankowski, J., Ziemba, P., Karczmarczyk, A., & Ziolo, M. (2019). Generalised framework for multi-criteria method selection. Omega, 86, 107–124. https://doi.org/10.1016/j.omega.2018.07.004
[33] Cinelli, M., Kadzinski, M., Miebs, G., Gonzalez, M., & Slowinski, R. (2022). Recommending multiple criteria decision analysis methods with a new taxonomy-based decision support system. European Journal of Operational Research, 302(2), 633–651. https://doi.org/10.1016/j.ejor.2022.01.011
[34] Chiu, Y.-H., Lo, H.-W., & Lin, S.-W. (2026). An integrated CRITIC–TOPSIS framework for warehouse personnel performance evaluation: Evidence from an electronics manufacturing enterprise. Journal of Intelligent Decision Making and Granular Computing, 2(1), 30–45. https://doi.org/10.31181/jidmgc21202633
[35] Lo, H.-W., & Lin, S.-W. (2025). A hybrid Bayesian BWM–machine learning framework for university digital transformation assessment: Integrating expert clustering and predictive validation. Technology in Society, 85, 103170. https://doi.org/10.1016/j.techsoc.2025.103170
[36] Lo, H.-W., & Lin, S.-W. (2026). A Shapley value-based argumentation framework integrating CRITIC and MARCOS for sustainable supplier evaluation: Evidence from a global semiconductor supply chain. Argumentation Based Systems Journal, 2, 304–323. https://doi.org/10.59543/cg894606
[37] Forman, E., & Peniwati, K. (1998). Aggregating individual judgments and priorities with the analytic hierarchy process. European Journal of Operational Research, 108(1), 165–169. https://doi.org/10.1016/s0377-2217(97)00244-0
[38] Rezaei, J. (2015). Best-worst multi-criteria decision-making method. Omega, 53, 49–57. https://doi.org/10.1016/j.omega.2014.11.009
[39] Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
[40] Hwang, C.-L., & Yoon, K. (1981). Multiple attribute decision making: Methods and applications. Springer. https://doi.org/10.1007/978-3-642-48318-9
[41] Opricovic, S., & Tzeng, G.-H. (2004). Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. European Journal of Operational Research, 156(2), 445–455. https://doi.org/10.1016/s0377-2217(03)00020-1

[42] Opricovic, S., & Tzeng, G.-H. (2007). Extended VIKOR method in comparison with outranking methods. European Journal of Operational Research, 178(2), 514–529. https://doi.org/10.1016/j.ejor.2006.01.020

[43] Stevic, Z., Pamucar, D., Puska, A., & Chatterjee, P. (2020). Sustainable supplier selection in healthcare industries using a new MCDM method: Measurement of alternatives and ranking according to compromise solution (MARCOS). Computers & Industrial Engineering, 140, 106231. https://doi.org/10.1016/j.cie.2019.106231

[44] Yazdani, M., Zarate, P., Zavadskas, E. K., & Turskis, Z. (2019). A combined compromise solution (CoCoSo) method for multi-criteria decision-making problems. Management Decision, 57(9), 2501–2519. https://doi.org/10.1108/md-05-2017-0458

[45] Pamucar, D., Vasin, L., & Lukovac, V. (2014). Selection of railway level crossings for investing in security equipment using hybrid DEMATEL-MAIRCA model. Proceedings of the XVI International Scientific-Expert Conference on Railways, 89–92.

[46] Keshavarz Ghorabaee, M., Zavadskas, E. K., Olfat, L., & Turskis, Z. (2015). Multi-criteria inventory classification using a new method of evaluation based on distance from average solution (EDAS). Informatica, 26(3), 435–451. https://doi.org/10.15388/informatica.2015.57

[47] Zavadskas, E. K., Turskis, Z., Antucheviciene, J., & Zakarevicius, A. (2012). Optimization of weighted aggregated sum product assessment. Elektronika ir Elektrotechnika, 122(6), 3–6. https://doi.org/10.5755/j01.eee.122.6.1810

[48] Zavadskas, E. K., & Turskis, Z. (2010). A new additive ratio assessment (ARAS) method in multicriteria decision-making. Technological and Economic Development of Economy, 16(2), 159–172. https://doi.org/10.3846/tede.2010.10

[49] Kendall, M. G. (1938). A new measure of rank correlation. Biometrika, 30(1–2), 81–93. https://doi.org/10.1093/biomet/30.1-2.81