Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Urban Density and Vertical Complexity in High-Rise Women's Hospitals: Development and Validation of the HRWHPI for Hot-Arid Cities

IBRAHIM AHMED ADAM AHMED

ibrahimahmedeng5@gmail.com

Omdurman Islamic University

Research Article

Keywords: High-rise hospitals, women's health, performance-based design, Delphi, inter-rater reliability, evidence governance

Posted Date: September 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10940028/v1

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

Additional Declarations: No competing interests reported.

Abstract

High-rise healthcare development can reconcile urban land constraints with access to specialized care, but verticalization intensifies dependencies among clinical adjacency, elevators, fire and smoke protection, critical utilities, infection control, commissioning, and patient experience. This study developed the High-Rise Women's Hospital Performance Index (HRWHPI), a performance- and evidence-governed framework for hot-arid cities. Evidence synthesis and lineage review produced 72 indicators across nine domains, comprising 27 non-compensable Critical Gates, 26 Mandatory, 17 Performance, and 2 Enhancement indicators. A scope-filtered modified Delphi process generated 343 eligible expert-indicator evaluations from a 17-member multidisciplinary panel. All eligible Round-1 ratings were positive at the prespecified threshold, yielding item- and scale-level content-validity indices (CVIs) of 1.00 and modified kappa of 1.00; no second round was required. Expert-confirmed importance judgments and four domain comparisons produced locked hierarchical weights. Controlled inter-rater calibration met all estimable acceptance criteria: quadratic weighted Cohen kappa = 0.817, exact agreement = 83.3%, within-one agreement = 100%, and mean absolute difference = 0.167. Critical-Gate exact agreement was 100%, with no decision-changing discordance; gate kappa was not estimable because all paired outcomes were PASS/PASS. In an evidence-constrained Mayo/Riyadh real-case application, authentic context was available for 14/72 indicators (19.4%; 20.9% of locked weight), but 0/72 indicators were final-score eligible. The complete-project HRWHPI was therefore withheld. Sensitivity testing confirmed that this decision was robust. HRWHPI provides a structured performance framework while explicitly preventing unsupported scoring when safety-critical project evidence is inaccessible. This dual function supports transparent interpretation under constrained real-world access.

Article Highlights

- The new index combines 72 hospital-design checks across nine safety and performance domains.
- Independent reviewers applied the locked scoring rules with strong agreement.
- The real-case study withheld a score when confidential technical evidence was insufficient.

1. Introduction

Urban growth, high land values, and concentration of tertiary services increasingly push complex hospitals toward vertical or high-density configurations. In such projects, stacking, zoning, and routing are not merely space-planning problems: they determine the travel paths of patients, staff, supplies, waste, emergency responders, and critical services [1]. Elevator planning becomes a clinical logistics problem when inpatient movement, emergency transfer, service segregation, and degraded-mode operation all depend on the same vertical infrastructure [2].

The consequences of verticalization are particularly important during emergencies. Hospital evacuation differs from ordinary high-rise egress because many occupants require assistance, beds or mobility devices, clinical support, and continuity of life-sustaining services. Simulation research has shown that

movement devices, bed handling, assisted horizontal transfer, and evacuation sequencing materially affect evacuation time and resource demand [3–5]. These dependencies make fire, smoke, evacuation, vertical transport, and utility resilience inseparable in high-rise healthcare assessment.

Women's hospitals add further spatial and experiential requirements. Birthing environments are associated with maternal experience, perceived control, privacy, support, and staff workflow [6–8]. Neonatal settings similarly require design approaches that support family engagement while preserving clinical efficiency and infection-control needs [9, 10]. These requirements mean that a high-rise women's hospital cannot be assessed adequately by generic tower metrics or by a single accreditation, sustainability, or patient-experience standard.

Environmental and infection-control performance also interact with the hot-arid context. Built-environment interventions can influence nosocomial infection risk, and ventilation design remains a central engineering control [11, 12]. Hospital resilience requires consideration of multiple hazards and system interdependencies [13], while facade, thermal-comfort, and energy-performance studies show that climate-responsive design can materially influence cooling demand and indoor conditions in healthcare buildings [14–16]. International health-facility guidance further treats functional relationships, parking/access, patient movement, and women's-health units as integrated planning concerns [17], while its engineering guidance addresses critical services and vertical transportation [18]. WHO/UNICEF guidance separately emphasizes WASH and hygiene infrastructure as foundational to safe and dignified healthcare, including maternity care [19].

Despite this evidence, the available literature remains fragmented by discipline. Existing work addresses optimization, evacuation, maternity environments, infection control, resilience, or environmental performance separately, while broader hospital-resilience instruments are not specifically structured around the combined urban, vertical, women/neonatal, hot-arid, and evidence-governance demands of a high-rise women's hospital [13, 20]. The research gap is therefore not the absence of individual standards, but the absence of a traceable framework that integrates them, distinguishes non-compensable safety conditions from weighted performance, and prevents missing evidence from being misclassified as failure or compliance.

This study aimed to develop and validate the High-Rise Women's Hospital Performance Index (HRWHPI) for hot-arid cities. The objectives were to: (1) synthesize standards and research into a traceable indicator architecture; (2) establish content validity and Critical-Gate status through multidisciplinary expert review; (3) derive hierarchical weights; (4) test the reproducibility of the locked scoring process through controlled inter-rater calibration; and (5) demonstrate an evidence-constrained real-case application in Riyadh without fabricating performance scores where confidential technical evidence was unavailable.

2. Materials and Methods

2.1 Study design and evidence synthesis

A sequential instrument-development design was used: evidence synthesis -> indicator derivation and lineage control -> expert content validation -> weighting -> scoring-manual lock -> controlled inter-rater reliability (IRR) calibration -> evidence-constrained real-case application -> uncertainty and sensitivity analysis. The framework was designed as a performance-assessment instrument rather than a substitute for statutory approval, accreditation, or clinical governance.

The evidence synthesis was purposive and standards-led for instrument derivation rather than a formal systematic review or meta-analysis. The evidence base combined peer-reviewed research, Saudi regulatory and engineering references available to the study, international health-facility guidance, hospital safety and resilience literature, and specialist evidence on maternity/neonatal design. Peer-reviewed literature considered in the evidence base was current through 29 July 2026, with priority given to 2020–2026 literature while retaining seminal methodological sources for Delphi and content validity. The technical evidence hierarchy prioritized applicable authority requirements, health-facility guidance, integrated international design/engineering guidance, peer-reviewed research, and then case/vendor evidence. Key references included International Health Facility Guidelines, WHO/UNICEF guidance on safe health-facility infrastructure, Saudi electrical and mechanical codes, and IBC/IFC [17–19, 21–24].

2.2 Indicator derivation, lineage control, and final architecture

The initial evidence-derived framework contained 63 indicators organized conceptually across urban/regulatory, access, vertical transport, clinical adjacency, infection control, fire/evacuation, building-services reliability, hot-arid environmental performance, and women-centered care. Indicator refinement used an explicit crosswalk rather than silent rewriting. The V1-to-V2.1 reconciliation preserved 123 lineage relationships and produced a 72-item candidate set. Lineage decisions were independently reviewed and reconciled before the Delphi release.

The final instrument was organized into nine operational domains: D1 Regulatory, Site & Access Governance; D2 Spatial, Clinical & Women/Neonatal Planning; D3 Vertical Transportation & Mobility Resilience; D4 Fire & Life Safety; D5 Critical MEP, Utilities & System Resilience; D6 Environmental, Infection Control & Healing Environment; D7 Constructability, Commissioning & Operational Readiness; D8 Operations, Facilities & Lifecycle Performance; and D9 Digital Intelligence, Evidence & Governance. Indicators were classified as Critical Gate, Mandatory, Performance, or Enhancement. Critical Gates were defined as non-compensable: a safety-critical failure cannot be offset by high scores in lower-consequence domains.

2.3 Modified Delphi and content-validity analysis

A multidisciplinary panel of 17 experts was used. Eligibility was domain-specific rather than requiring every expert to rate every indicator. For each item within an expert's assigned scope, Round 1 requested

ratings of relevance, clarity, and measurability on a 1–4 scale, together with Critical-Gate judgment and a Keep/Modify/Delete/Merge decision. 'Not within expertise' was excluded from the denominator. This scope-filtered approach yielded 343 eligible expert-indicator evaluations, with 3–9 eligible experts per item (median 5). Healthcare-focused modified Delphi applications informed panel and instrument design [25–27], while consensus reporting and content-validity methods followed established guidance [28–31].

For each item, I-CVI was calculated as the proportion of eligible experts providing a positive rating (3 or 4). S-CVI/Ave and S-CVI/UA summarized scale-level performance, and modified kappa adjusted item-level agreement for chance. A relevance I-CVI threshold of 0.78 was prespecified in line with established content-validity guidance [30, 31], with equivalent positive thresholds applied to clarity and measurability; safety-critical items were not to be deleted automatically without examining wording and panel expertise. A second Delphi round was required only for revised, low-consensus, or unresolved items.

2.4 Expert-confirmed weighting

Weighting was performed after content validation. The Stage 5 dataset comprised 17 expert-confirmed sets of indicator-importance judgments (1–9 scale; 343 scope-eligible judgments) and four complete Best-Worst Method (BWM) comparisons across the nine domains. BWM identifies the most and least important criteria, collects best-to-others and others-to-worst comparisons, and derives a consistent criterion-weight vector [32].

The four domain-weight vectors were aggregated by geometric mean and normalized. Indicator importance ratings were likewise aggregated within domains using geometric means, converted to local weights, and multiplied by the corresponding domain weight to obtain global indicator weights Wj. The final global weights sum to 1.000 (100 points). Critical-Gate classification remained separate from weight: neither a high importance rating nor a large weight could create or replace gate status.

2.5 Scoring architecture and evidence governance

The locked scoring manual used an ordinal raw score from 0 to 4 together with Evidence Status and, for Critical Gates, a separate Gate Outcome. The assessment explicitly distinguishes verified non-compliance from absence of evidence. 'Not Evidenced' therefore does not equal score 0 and does not imply compliance. Missing evidence is carried as uncertainty until sufficient project documentation supports a case-specific judgment.

For evidenced indicators, domain performance is a weighted mean and the overall index is the weighted aggregation of the domain/indicator scores. However, a complete-project result can be released only when the evidence base and Critical-Gate decisions are sufficient for the intended assessment. No universal HRWHPI pass cut-off was treated as empirically validated in this study.

2.6 Controlled inter-rater reliability calibration

Before real-case interpretation, the locked manual was tested by two independent raters with complementary healthcare expertise (healthcare architecture/planning and healthcare MEP/construction/facilities management). Both raters evaluated all 72 indicators using the same frozen evidence package and remained blinded to the other's scores until both returns were frozen.

Because authentic Mayo documents alone could not make every indicator scoreable, the Stage 7 package combined the real EVID-001/EVID-002 baseline with clearly labeled CAL-SIM B01-B14 supplements. These supplements were used only to create a controlled calibration environment and were explicitly prohibited from being treated as factual Mayo evidence. Prespecified IRR criteria were quadratic weighted Cohen kappa [33] > = 0.80, exact agreement > = 70%, within-1 agreement > = 95%, mean absolute difference (MAD) < = 0.50, Critical-Gate kappa > = 0.80 when estimable, Critical-Gate exact agreement > = 90%, and zero unresolved decision-changing gate discordance. If gate prevalence made Pe = 1, conventional Cohen kappa was reported as not estimable rather than forced to 1.00.

2.7 Evidence-constrained Mayo/Riyadh real-case application

After calibration, Mayo/Riyadh was evaluated as an Evidence-Constrained Real-Case Application. Only authentic project evidence available to the researchers was used to characterize the case; CAL-SIM materials were excluded. The principal accessible project sources were architectural/functional and project-presentation records. Detailed fire strategy, MEP design, elevator studies, commissioning records, and other restricted technical files were outside the accessible evidence boundary and were not treated as prerequisites for case inclusion.

For each of the 72 indicators, the assessment distinguished: authentic contextual evidence, final score eligibility, and Critical-Gate decision status. Contextual evidence could support case understanding without being sufficient to assign a 0–4 performance score. Evidence coverage was therefore reported independently from project performance. Source conflicts were preserved as open whenever the project documents reported inconsistent values rather than being reconciled by assumption.

2.8 Uncertainty, sensitivity, and scenario analysis

Uncertainty was treated primarily as structural evidence-access and score-identifiability uncertainty rather than sampling error. No missing performance value was imputed and no confidence interval was calculated for an unidentified project score. Evidence coverage was examined using locked global weights, equal domain weights, and unweighted indicator counts. One-at-a-time (OAT) domain-weight perturbations of +/-10% and +/-20% were applied with renormalization to test whether the evidence-coverage conclusion depended on a fragile weighting structure.

Five evidence-access scenarios were also examined: S0 current baseline; S1 verification of all currently contextual indicators; S2 S1 plus resolution of all Critical Gates; S3 S2 plus verification of all Mandatory indicators; and S4 full evidence availability. These scenarios modeled potential evidence capacity only; they did not assign hypothetical Mayo performance scores.

2.9 Use of generative AI tools

Generative AI tools (including ChatGPT, OpenAI) were used under author supervision to assist with language editing, document structuring, consistency checking, and execution of author-specified computational and document-production workflows. They were not treated as data sources, expert participants, or autonomous decision-makers. All methodological decisions, source interpretation, quantitative results, and final scientific claims were reviewed by, and remain the responsibility of, the authors.

3. Results

3.1 Final HRWHPI structure and content validity

The finalized instrument contained 72 indicators: 27 Critical Gates, 26 Mandatory indicators, 17 Performance indicators, and 2 Enhancement indicators. Domain sizes ranged from 4 indicators in D3 to 11 in D8. The largest concentration of Critical Gates occurred in D4 Fire & Life Safety (6/6) and D5 Critical MEP, Utilities & System Resilience (6/8). The final domain architecture and locked domain weights are summarized in Table 1.

Across the 343 scope-eligible Round-1 evaluations, every eligible rating was recorded as positive at the content-validity dichotomy (3/4), every gate-designated item received supporting gate votes from its eligible experts, and all 343 eligible decisions were Keep. Consequently, the minimum item-level I-CVI for relevance, clarity, and measurability was 1.00; S-CVI/Ave and S-CVI/UA were 1.00; and modified kappa was 1.00. No item required Round 2 and the 72-item wording was retained. The preserved Delphi dataset did not distinguish the exact raw value 3 from 4, so means and standard deviations of those three rating dimensions were not reconstructed. Round-1 validation outcomes are summarized in Table 2.

Table 1
Final HRWHPI domain architecture and locked domain weights

Domain	Domain name	Items	Critical Gates	Locked weight
D1	Regulatory, Site & Access Governance	9	4	11.01%
D2	Spatial, Clinical & Women/Neonatal Planning	8	3	11.52%
D3	Vertical Transportation & Mobility Resilience	4	2	9.95%
D4	Fire & Life Safety	6	6	12.18%
D5	Critical MEP, Utilities & System Resilience	8	6	16.27%
D6	Environmental, Infection Control & Healing Environment	6	2	8.75%
D7	Constructability, Commissioning & Operational Readiness	10	3	14.11%
D8	Operations, Facilities & Lifecycle Performance	11	0	9.69%
D9	Digital Intelligence, Evidence & Governance	10	1	6.52%

Table 2
Round-1 content-validation summary

Measure	Result	Interpretation
Candidate indicators	72	Frozen V2.1 set
Eligible expert-indicator evaluations	343	3–9 experts/item; median 5
Relevance I-CVI	1.00	All eligible ratings positive
Clarity I-CVI	1.00	All eligible ratings positive
Measurability I-CVI	1.00	All eligible ratings positive
S-CVI/Ave	1.00	All three dimensions
S-CVI/UA	1.00	All three dimensions
Modified kappa	1.00	Positive dichotomy
Critical Gates supported	27/27	No unresolved gate designation
Round 2	Not required	No modified/low-consensus items

3.2 Locked expert weights

The locked domain weights summed to 1.000. D5 Critical MEP, Utilities & System Resilience received the highest weight (16.27%), followed by D7 Constructability, Commissioning & Operational Readiness (14.11%) and D4 Fire & Life Safety (12.18%). D9 Digital Intelligence, Evidence & Governance received the

lowest relative domain weight (6.52%). These weights represent the expert-confirmed Stage 5 hierarchy and do not alter the non-compensable status of Critical Gates.

3.3 Controlled inter-rater reliability

The two independent raters produced 72 paired ordinal scores. Exact agreement occurred on 60/72 indicators (83.3%), while all 72 pairs were within one scale point; the 12 disagreements were therefore adjacent-category differences only. Quadratic weighted Cohen kappa was 0.8175 and MAD was 0.1667, meeting all prespecified ordinal criteria. The raters also agreed exactly on all 27 Critical-Gate outcomes, with zero decision-changing gate discordance. Critical-Gate Cohen kappa was not estimable because every paired gate outcome was PASS/PASS, yielding invariant category use (Pe = 1). Stage 7 was therefore accepted on the basis of all estimable prespecified criteria, while the gate result was interpreted cautiously as controlled calibration rather than proof of operational reliability across heterogeneous cases. The prespecified IRR metrics and decisions are summarized in Table 3.

Table 3
Controlled inter-rater reliability results

Metric	Observed	Criterion	Decision
Quadratic weighted Cohen kappa	0.8175	>=0.80	PASS
Exact agreement	83.3% (60/72)	>=70%	PASS
Within-1 agreement	100.0%	>=95%	PASS
Mean absolute difference	0.1667	<=0.50	PASS
Critical-Gate exact agreement	100.0% (27/27)	>=90%	PASS
Critical-Gate Cohen kappa	Not estimable (Pe = 1)	>=0.80 when estimable	Interpretive
Decision-changing gate discordance	0	0	PASS

3.4 Mayo/Riyadh evidence coverage and score identifiability

Authentic contextual information was available for 14/72 indicators (19.4%), representing 20.9% of the locked global indicator weight. Context was concentrated in D1-D3, where the available project documents contained regulatory/site, spatial/clinical, parking/access, and vertical-organization information. No authentic contextual evidence was available for D5, D7, D8, or D9 within the source boundary used for the case. Domain-level evidence coverage is reported in Table 4 and visualized in Fig. 1.

Importantly, contextual evidence was insufficient to satisfy the locked final-scoring evidence requirements for any indicator: 0/72 indicators were final-score eligible. Five of the 27 Critical Gates had

some contextual information, but 0/27 had sufficient evidence for a final PASS/FAIL determination; all 27 therefore remained unresolved for final case decision. The complete-project HRWHPI was withheld rather than treating absent technical records as non-compliance or assumed compliance.

Four source conflicts were retained as open: Mayo built-up/total area (227,714 vs 222,384 sqm), parking count (789 vs 1,044), footprint/local-regulation values (35% maximum vs 44% reported total footprint), and inconsistent area/FAR summary definitions across the project presentation. None was silently reconciled for scoring.

Table 4
Authentic contextual evidence coverage by HRWHPI domain in the Mayo/Riyadh case

Domain	Domain name	Context/total	Count coverage	Within-domain weighted coverage	Global weight represented
D1	Regulatory, Site & Access Governance	5/9	55.6%	51.6%	5.68%
D2	Spatial, Clinical & Women/Neonatal Planning	5/8	62.5%	65.4%	7.53%
D3	Vertical Transportation & Mobility Resilience	2/4	50.0%	45.2%	4.50%
D4	Fire & Life Safety	1/6	16.7%	17.9%	2.18%
D5	Critical MEP, Utilities & System Resilience	0/8	0.0%	0.0%	0.00%
D6	Environmental, Infection Control & Healing Environment	1/6	16.7%	12.0%	1.05%
D7	Constructability, Commissioning & Operational Readiness	0/10	0.0%	0.0%	0.00%
D8	Operations, Facilities & Lifecycle Performance	0/11	0.0%	0.0%	0.00%
D9	Digital Intelligence, Evidence & Governance	0/10	0.0%	0.0%	0.00%

3.5 Sensitivity and evidence-access scenarios

The evidence-constrained conclusion was insensitive to the tested weighting alternatives. Contextual coverage was 20.9% under locked global weights, 21.3% with equal domain weights, and 19.4% using unweighted indicator counts. OAT +/-10% perturbations produced a 20.4–21.4% range, and +/-20%

perturbations produced a 19.9–21.9% range. The complete-project HRWHPI remained withheld under every weighting basis. The domain-weight sensitivity pattern is shown in Fig. 2.

Evidence-access scenarios further separated scoreability from performance. S1 could make 14/72 indicators (20.9% of locked weight) potentially evidence-eligible; resolving all Critical Gates in S2 could increase potential eligible weight to 63.1%; adding all Mandatory indicators in S3 could increase it to 86.1%; and S4 represented 100% evidence availability. However, none of these scenarios assigned hypothetical 0–4 Mayo scores; consequently, they were not alternative HRWHPI performance results. Table 5 and Fig. 3 summarize the evidence-access scenarios.

Table 5
Evidence-access scenarios; percentages represent potential score-eligible locked weight, not simulated Mayo performance

Scenario	Potential eligible indicators	Potential eligible locked weight	Critical Gates resolved	Project HRWHPI
S0 Current baseline	0/72	0.0%	0/27	WITHHELD
S1 Verify contextual indicators	14/72	20.9%	5/27	WITHHELD
S2 + resolve all Critical Gates	36/72	63.1%	27/27	WITHHELD
S3 + verify Mandatory indicators	58/72	86.1%	27/27	WITHHELD
S4 Full evidence availability	72/72	100.0%	27/27	WITHHELD - no hypothetical performance scores

4. Discussion

The HRWHPI was developed to address a problem created by the intersection of healthcare complexity and vertical urban form. Existing research establishes that stacking and routing influence hospital configuration [1], that elevator systems govern high-rise logistics [2], and that evacuation of dependent occupants requires assisted-movement planning rather than generic egress assumptions [3–5]. Women's and neonatal environments add privacy, choice, family, and rapid-adjacency requirements [6–10]. The final nine-domain architecture therefore treats density, clinical function, vertical movement, fire safety, critical utilities, environmental/IPC performance, commissioning, operations, and evidence governance as a connected system rather than separate checklists. This design-oriented architecture also differs from broader resilience indices that focus primarily on organizational and hospital-continuity characteristics rather than integrated high-rise clinical-design performance [20].

The weighting pattern is also informative. Critical MEP and utilities received the highest domain weight, followed by constructability/commissioning and fire/life safety. This does not imply that clinical planning or women-centered experience is secondary; instead, it reflects expert recognition that a high-rise healthcare building can maintain clinical intent only if power, cooling, water, medical gases, vertical transport, and commissioning remain reliable under routine and degraded conditions. International healthcare engineering guidance similarly treats loss of critical services and vertical-transport reliability as patient-safety and operational-continuity concerns [18]. The framework deliberately prevents these relative weights from overriding Critical Gates.

The content-validity findings provide strong evidence of expert acceptance but should not be overinterpreted. All scope-eligible judgments were positive after dichotomization, producing CVI and modified-kappa values of 1.00. However, the preserved record did not retain the exact distinction between ratings 3 and 4. The appropriate interpretation is therefore unanimous positive content-validity endorsement within the eligible panel, not zero dispersion on the original 1–4 scale. This conservative interpretation is consistent with established Delphi and content-validity reporting principles [28–31] and avoids reconstructing ordinal detail that is no longer available.

Controlled IRR demonstrated that the locked manual can be applied consistently when raters receive the same complete evidence boundary. A quadratic weighted kappa of 0.8175, 83.3% exact agreement, 100% within-one agreement, and low MAD indicate that remaining differences were small ordinal judgments rather than systematic disagreement. Weighted kappa explicitly accommodates scaled disagreement between ordinal categories [33]. The Critical-Gate result is more limited: 100% exact agreement is encouraging, but invariant PASS classifications make Cohen kappa mathematically unidentifiable. Future reliability studies should intentionally include authentic cases containing failed, unresolved, and satisfied gates so that gate-category reliability can be estimated under real prevalence conditions.

The Mayo/Riyadh application demonstrates a second function of HRWHPI: evidence governance. Within this case, detailed fire, MEP, elevator, commissioning, operational, and digital records were unavailable inside the accessible evidence boundary. A conventional weighted index can produce a misleadingly precise result if missing records are silently coded as failure, ignored, or assumed compliant. HRWHPI instead maintains Not Evidenced as a distinct state and withholds the complete-project score when performance cannot be identified. This converts a common limitation of evidence-restricted healthcare case assessment into an explicit, auditable result.

The distribution of available Mayo evidence should therefore not be read as a performance profile. Higher contextual coverage in D1-D3 reflects the architectural and presentation-oriented nature of the accessible sources. Zero contextual evidence in MEP, commissioning, operations, or digital governance does not demonstrate weakness in those domains. Similarly, the four unresolved project-document conflicts show why source governance matters: choosing one parking, area, or footprint value without an authoritative revision record could bias regulatory and access interpretations.

Sensitivity testing strengthens this conclusion. Alternative weighting structures and +/-20% domain perturbations changed contextual coverage by only a small amount and never altered the withholding decision. Even complete evidence availability in the scenario analysis was not treated as high performance, because evidence capacity and observed performance are different variables. This distinction is a core methodological contribution of HRWHPI and may be useful beyond women's hospitals whenever researchers evaluate safety-critical infrastructure with incomplete access to technical records.

Several limitations remain. First, the content-validation panel was scope-filtered, so eligible N varied by item from 3 to 9. Second, exact 3-versus-4 Delphi ratings were not preserved. Third, Stage 5 weighting is best described as an expert-confirmed empirical dataset: 17 experts confirmed the recorded importance judgments and four confirmed BWM comparisons, but future studies should preserve digitally time-stamped raw returns from inception. Fourth, Stage 7 was a controlled calibration exercise using simulated supplements and therefore does not establish operational IRR under messy real-world evidence. Fifth, the Mayo case had 0/72 final-score-eligible indicators and cannot validate score distribution, discrimination, or thresholds. Finally, the study examined one evidence-constrained real case; external validity and an empirical pass-band structure require additional hospitals, preferably with owner-authorized access to complete records.

Future research should apply the locked 72-item instrument to multiple high-rise women's hospitals and comparable tertiary towers at different lifecycle stages. Priority analyses include operational IRR, gate-category variability, criterion validity against independent safety/quality outcomes, longitudinal sensitivity to commissioning and operational data, and assessment of whether observed score distributions support empirically defensible performance bands.

5. Conclusions

This study developed the HRWHPI as a nine-domain, 72-item performance framework for high-rise women's hospitals in hot-arid cities. The instrument integrates 27 non-compensable Critical Gates with weighted mandatory, performance, and enhancement indicators. Scope-filtered expert review produced unanimous positive content-validity endorsement, expert-confirmed weighting established a locked 100-point hierarchy, and controlled calibration achieved acceptable ordinal inter-rater reliability. The Mayo/Riyadh case then demonstrated the framework's evidence-governance function: despite authentic contextual information for 14 indicators, no indicator met final-score evidence requirements and all Critical Gates remained unresolved for final case judgment, so the complete-project score was withheld. Sensitivity testing showed that this decision was not an artifact of weighting assumptions. HRWHPI should therefore be interpreted as both a performance-assessment structure and a governance mechanism that defines when a healthcare performance claim is - and is not - supported by evidence.

Declarations

Author Contribution

I.A.A.A. conceived and designed the study, developed the research methodology and the HRWHPI framework, conducted the literature and standards review, developed the research instruments, coordinated the expert consultation and validation processes, performed the data analysis and interpretation, prepared the figures and tables, and wrote, reviewed, and revised the manuscript. I.A.A.A. approved the final version of the manuscript and takes full responsibility for the integrity and accuracy of the work.

Acknowledgement

The author gratefully acknowledges the engineers, project managers, academics, and other professionals who provided general guidance and constructive feedback during the development of this research. Their contributions were advisory in nature and did not meet the criteria for authorship.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request. The anonymized supplementary datasets and analytical outputs necessary to understand and reproduce the reported results are provided as supplementary information. Individual expert and rater records containing potentially identifiable information are not publicly shared to protect participant confidentiality. Project-specific documents used in the evidence-constrained case application are subject to access and confidentiality restrictions and therefore cannot be made publicly available.

References

1. Cubukcuoglu C, Nourian P, Sariyildiz IS, Tasgetiren MF. Optimal design of new hospitals: a computational workflow for stacking, zoning, and routing. Autom Constr. 2022;134:104102. https://doi.org/10.1016/j.autcon.2021.104102.
2. Srinopparatanakul P, Viriyasiri P. (2021) Design process of elevators inside high-rise hospital buildings. Sarasatr J Archit Des 4(2). https://so05.tci-thaijo.org/index.php/sarasatr/article/view/251328
3. Hunt ALE, Galea ER, Lawrence PJ, Frost IR, Gwynne SMV. Simulating movement devices used in hospital evacuation. Fire Technol. 2020;56(5):2209–40. https://doi.org/10.1007/s10694-020-00971-5.
4. Joyce MS, Lawrence PJ, Galea ER. Hospital Evacuation Planning Tool for Assistance Devices (HEPTAD). Fire Mater. 2021;45(4):564–82. https://doi.org/10.1002/fam.2856.
5. Kwak J, Lees MH, Cai W, Pourghaderi AR, Ong MEH. Estimating horizontal movement performance of patient beds and the impact on emergency evacuation time. Saf Sci. 2021;134:105038.

https://doi.org/10.1016/j.ssci.2020.105038.

6. Goldkuhl L, Gyllensten H, Begley C, Nilsson C, Wijk H, Lindahl G, Uvnas-Moberg K, Berg M. Impact of birthing room design on maternal childbirth experience: results from the Room4Birth randomized trial. HERD Health Environ Res Des J. 2023;16(1):200–18. https://doi.org/10.1177/19375867221124232.
7. Setola N, Naldi E, Cardinali P, Migliorini L. A broad study to develop maternity units design knowledge combining spatial analysis and mothers' and midwives' perception of the birth environment. HERD Health Environ Res Des J. 2022;15(4):204–32. https://doi.org/10.1177/19375867221098987.
8. Bjornson Skogstrom L, Vithal E, Wijk H, Lindahl G, Berg M. Women's experiences of physical features in a specially designed birthing room: a mixed-methods study in Sweden. HERD Health Environ Res Des J. 2022;15(3):193–205. https://doi.org/10.1177/19375867221077097.
9. Machry H, Joseph A, White RD, Allison D. Designing for family engagement in neonatal ICUs: how is the interior design of single-family rooms supporting family behaviors, from passive to active? HERD Health Environ Res Des J. 2023;16(3):238–60. https://doi.org/10.1177/19375867231168651.
10. Fay L, Real K, Haynes S, Daneshvar Z. Examining efficiency in open-bay and single-family room NICU designs. Adv Neonatal Care. 2023;23(4):355–64. https://doi.org/10.1097/ANC.0000000000001058.
11. Xiong L, Sheng G, Fan Z-M, Yang H, Hwang F-J, Zhu B-W. Environmental design strategies to decrease the risk of nosocomial infection in medical buildings using a hybrid MCDM model. J Healthc Eng. 2021;2021:5534607. https://doi.org/10.1155/2021/5534607.
12. Ren C, Wang J, Feng Z, Kim MK, Haghighat F, Cao S-J. Refined design of ventilation systems to mitigate infection risk in hospital wards: perspective from ventilation openings setting. Environ Pollut. 2023;333:122025. https://doi.org/10.1016/j.envpol.2023.122025.
13. Tang K, Chen B. Resilient hospital design: from Crimean War to COVID-19. HERD Health Environ Res Des J. 2023;16(4):36–55. https://doi.org/10.1177/19375867231174238.
14. Almhafdy A, Alsehail AM. The effect of window design factors on the cooling load in hospitals wards. Smart Sustain Built Environ. 2025;14(5):1389–406. https://doi.org/10.1108/SASBE-07-2023-0195.
15. Garces HO, Duran C, Espinosa E, Jerez A, Palominos F, Hinojosa M, Carrasco R. Monitoring of thermal comfort and air quality for sustainable energy management inside hospitals based on online analytical processing and the Internet of Things. Int J Environ Res Public Health. 2022;19(19):12207. https://doi.org/10.3390/ijerph191912207.
16. Harshalatha H, Patil S, Kini PG. Calibration of simulation model to analyze hospital building energy performance. Energy Build. 2024;313:114242. https://doi.org/10.1016/j.enbuild.2024.114242.
17. TAHPI (2017) International Health Facility Guidelines, Part B: Health Facility Briefing & Design, Version 5. International Health Facility Guidelines.
18. TAHPI (2025) International Health Facility Guidelines, Part E: Engineering. International Health Facility Guidelines.

19. World H, Organization UNICEF. Water, sanitation and hygiene in health care facilities: practical steps to achieve universal access to quality care. Geneva: World Health Organization; 2019.
20. Lirk P, Janjua H, Rogers M, Flynn S, Kuo P, Pietrobon R. Psychometric development and validation of the Hospital Resilience Index. Discover Health Syst. 2024;3:98. https://doi.org/10.1007/s44250-024-00162-x.
21. Saudi Building Code National Committee. SBC 401-CR-2024: Saudi Electrical Code. Saudi Arabia: Riyadh; 2024.
22. Saudi Building Code National Committee. SBC 501-CC-2024: Saudi Mechanical Code. Saudi Arabia: Riyadh; 2024.
23. International Code Council. International Building Code, 2021 ed. IL: ICC, Country Club Hills; 2021.
24. International Code Council. International Fire Code, 2021 ed. IL: ICC, Country Club Hills; 2021.
25. Almohanna AAS, Win KT, Meedya S, Vlahu-Gjorgievska E. Design and content validation of an instrument measuring user perception of the persuasive design principles in a breastfeeding mHealth app: a modified Delphi study. Int J Med Inf. 2022;164:104789. https://doi.org/10.1016/j.ijmedinf.2022.104789.
26. Shinners L, Aggar C, Grace S, Smith S. Exploring healthcare professionals' perceptions of artificial intelligence: validating a questionnaire using the e-Delphi method. Digit Health. 2021;7:20552076211003433. https://doi.org/10.1177/20552076211003433.
27. Nwodoh CO, Okoronkwo IL, Anarado AN, Nwaneri AC, Chinweuba AU, Ndubuisi I, Iheanacho P, Ubochi N, Ekechukwu END. A modified Delphi consensus on generic indicators for a low- and middle-income country's quality nursing care measurement. Nurs Open. 2022;9(5):2397–408. https://doi.org/10.1002/nop2.1255.
28. Junger S, Payne SA, Brine J, Radbruch L, Brearley SG. Guidance on conducting and reporting Delphi studies (CREDES) in palliative care. Palliat Med. 2017;31(8):684–706. https://doi.org/10.1177/0269216317690685.
29. Diamond IR, Grant RC, Feldman BM, Pencharz PB, Ling SC, Moore AM, Wales PW. Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. J Clin Epidemiol. 2014;67(4):401–9. https://doi.org/10.1016/j.jclinepi.2013.12.002.
30. Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Res Nurs Health. 2007;30(4):459–67. https://doi.org/10.1002/nur.20199.
31. Lynn MR. Determination and quantification of content validity. Nurs Res. 1986;35(6):382–5. https://doi.org/10.1097/00006199-198611000-00017.
32. Rezaei J. Best-worst multi-criteria decision-making method. Omega. 2015;53:49–57. https://doi.org/10.1016/j.omega.2014.11.009.
33. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213–20. https://doi.org/10.1037/h0026256.

Figures

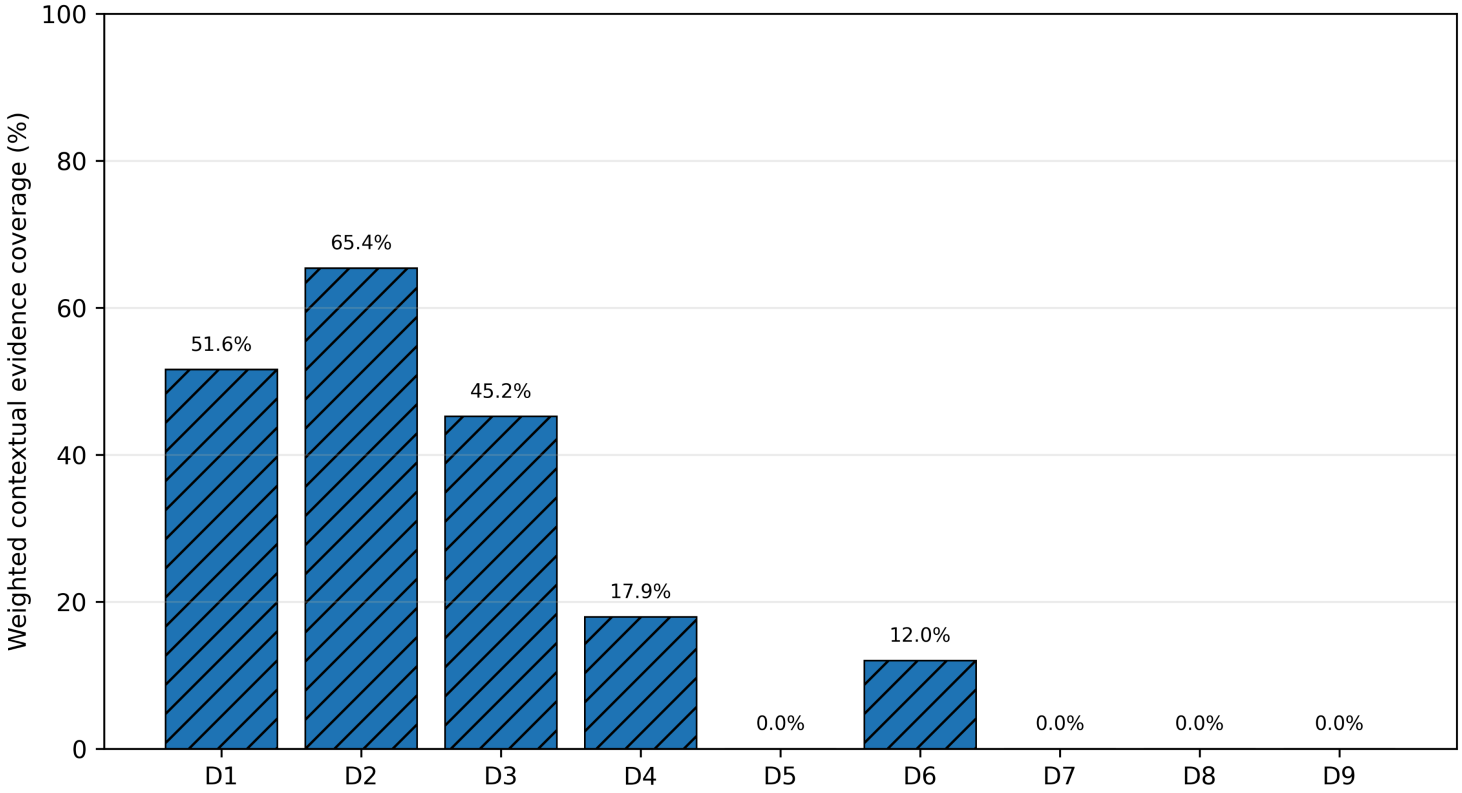


Figure 1

Distribution of authentic contextual evidence across the nine HRWHPI domains; coverage reflects evidence availability, not project performance

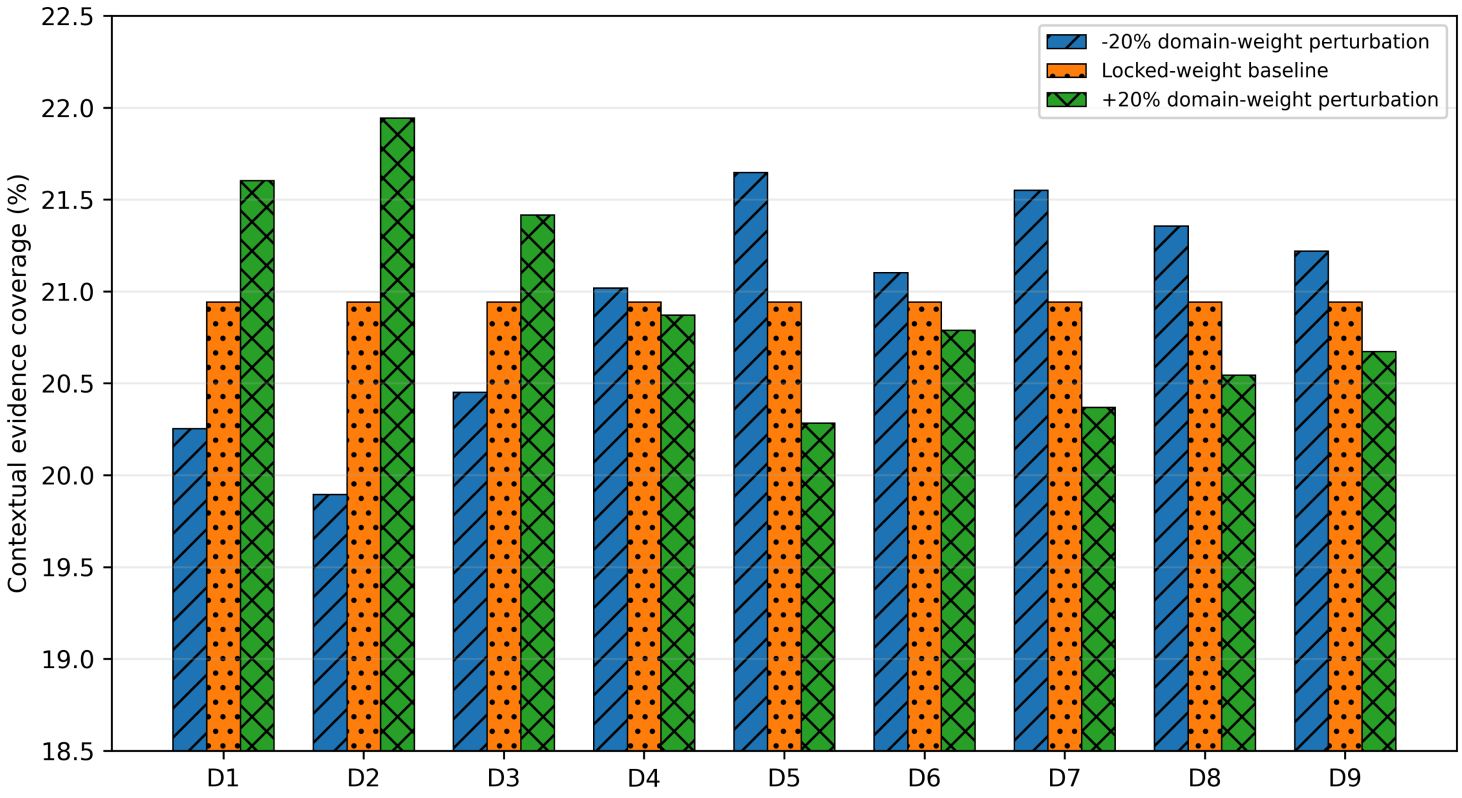


Figure 2

Deterministic sensitivity of contextual evidence coverage to domain-weight perturbation; official locked weights remain unchanged

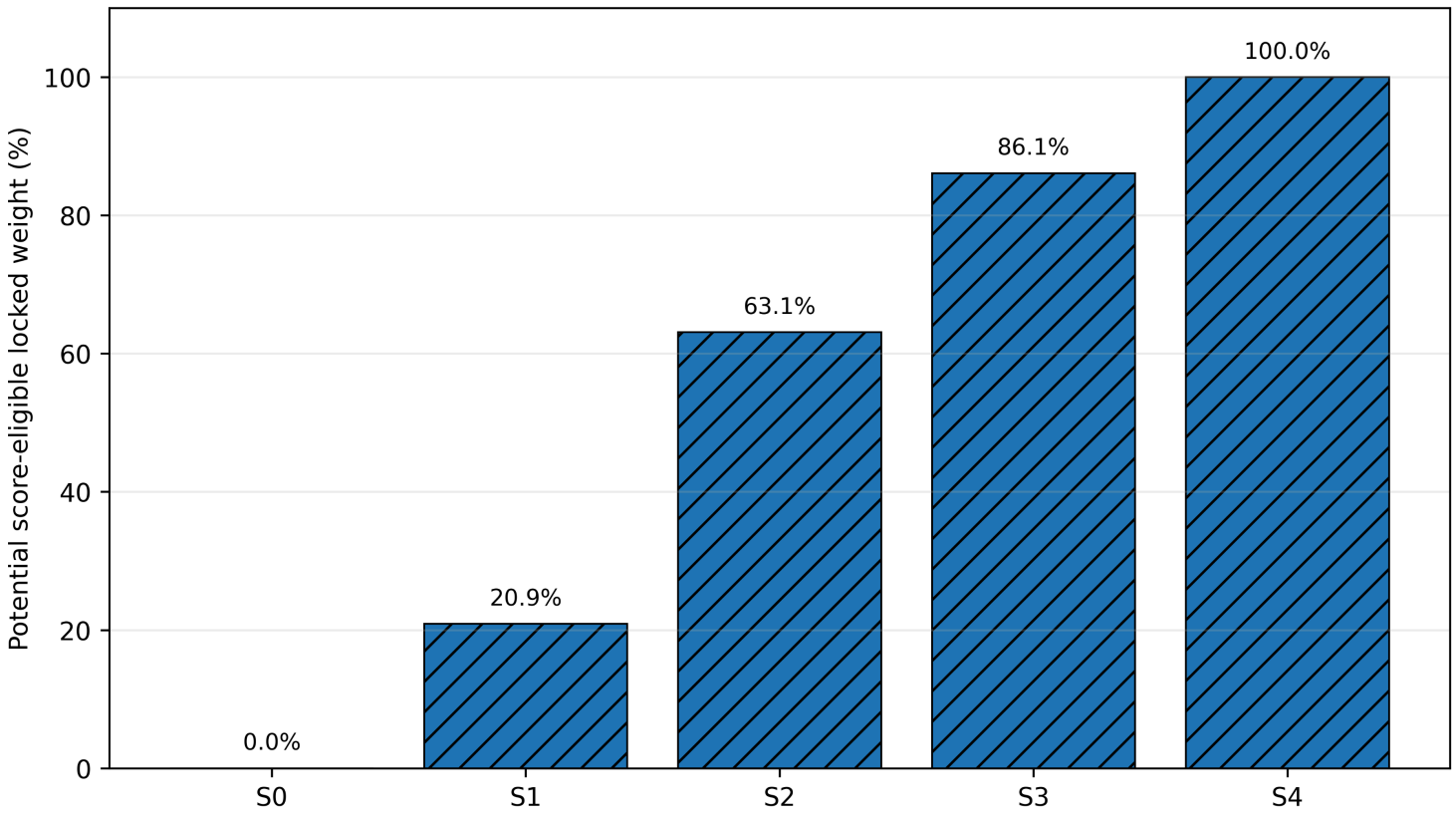


Figure 3

Evidence-access scenarios for the Mayo/Riyadh case; scenario values represent evidence capacity, not alternative HRWHPI performance scores

Supplementary Files

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

- HRWHPISUPPLEMENTARYINFORMATIONCONSOLIDATEDBLINDEDV10.xlsx