



unesco

Centre
Under the auspices
of UNESCO



ICAIRE

International
Center for
AI Research & Ethics

المركز الدولي للبحوث وأخلاقيات الذكاء الاصطناعي

Appendix A. Worked Example — One Assurance Claim End to End

Supplementary material to *Risk-Based AI Assurance in the Saudi Public Sector: Aligning International Assurance Practices with the National AI Risk Management Framework*

ICAIRE FRAMEWORK PAPER • SUPPLEMENTARY APPENDIX A

Nawaf Alsahmah, Ph.D. (Independent Researcher)

October 2026

PARENT PAPER <https://doi.org/10.68480/CEYS4608>

This appendix carries one assurance claim through every stage of the proposed framework. It is illustrative and does not describe any actual system, entity, or risk record. Its purpose is to give an adopting entity a pattern to follow and to demonstrate that the process specified in Section 7 is executable as written.

Setting. A public-sector entity operates an AI-supported screening system that ranks applications for a benefit programme and flags those requiring closer manual review. Officials make all final eligibility determinations; the system does not decide. The system is in production and has been operating for one full reporting period.

Stage 0 — Inputs drawn through the Risk-Assurance Interface. From the existing risk-management record for this system: system context and intended use; the risk register entries relevant to the system; implemented controls and their indicators; residual-risk assessments; monitoring results for the reporting period; and the most recent reassessment findings. Nothing in this stage is recalculated.

Stage 1 — Assurance scope, requirements, and claims (Section 7.2).

Scope. The screening and ranking component as configured in the current production release, for applications received in the defined reporting period, covering all applicant groups within the programme's eligibility rules. Excluded from scope: the downstream case-management system, and officials' final determination process.

Applicable substantive requirements identified Requirements arising from the program's governing rules on equal treatment of applicants; applicable data-protection requirements; the entity's own policy on human oversight of AI-supported decisions; notice, explanation, and recourse obligations owed to applicants under applicable law or policy, where applicable; and requirements arising from adopted standards where the entity has committed to them.

One claim selected for this example.

Claim C-04. *For the screening component as configured in release [x], applications processed during [reporting period] were flagged for manual review at rates, and with false-flag rates, that do not differ across the applicant groups defined in [the programme's equal-treatment rule] beyond the tolerance stated in that rule.*

The claim is clear (it admits of evidence), context-specific (release, period, population), traceable (to the named equal-treatment requirement), evaluable (from flag-rate and outcome data), and bounded in time (to the stated reporting period) — the five criteria of Section 7.2.

Stage 2 — Risk-to-assurance calibration (Section 7.3).

Risk-record linkage. Claim C-04 is linked to the existing risk

register entry describing the risk of unequal treatment of applicant groups in screening. That entry is assessed at the entity's highest risk level. A second entry, concerning data quality in an input field, is also relevant; the two are considered separately and not combined into a new score.

Calibration inputs. Risk level: highest in use. Risk nature: a risk to individual entitlement and equal treatment, for which outcome-distribution evidence is directly relevant. Context of use: a system supporting decisions about individuals' access to a benefit, with human determination downstream — which reduces but does not remove the effect on individuals, since the flag influences the review an application receives.

Presumptive floor applied. Table 9, first condition: a system whose failure could affect the eligibility or entitlements of individuals, at the highest risk level is **R3 / I3**.

Departure recorded? None. The floor is accepted.

Configuration output. Rigour R3 (enhanced verification, including sample-based verification of underlying evidence); techniques: disaggregated outcome analysis on the reporting-period data, review of monitoring records, and verification of the analysis against source records; evaluator independence I3 (internal audit or equivalent function outside the delivery chain). No binding instrument prescribes a higher minimum for this claim.

Stage 3 — Assurance execution and evidence evaluation (Section 7.4).

Evidence assembled, with source:

1. Disaggregated flag-rate and false-flag-rate analysis for the reporting period, by applicant group — *newly produced for this claim; existing monitoring reported aggregate rates only.*
2. Monitoring records for the reporting period, including exception reports — *existing risk-management record.*
3. Control indicator evidence for the control requiring periodic equal-treatment review — *existing risk-management record.*
4. Data-quality assessment for the input field identified in the second risk entry — *existing record, produced for a different purpose.*
5. Impact-assessment output covering the screening component — *existing record.*

Evaluation against the five adequacy criteria (Table 11):

Criterion	Finding
Relevance	Met. Items 1–3 concern the release and period in the claim's scope. Item 5 predates the current release and is relevant to context but not to the claim's period.
Validity	Met for items 1–3. The disaggregated analysis uses outcome data appropriate to the inference; the monitoring records corroborate it. Item 4 supports a data-quality inference, not an equal-treatment inference, and is not relied on for this claim.
Currency	Met. Items 1–3 fall within the stated reporting period and post-date the current release.
Coverage.	Not met. One applicant group defined in the equal-treatment rule is represented by too few applications in the reporting period for a stable rate estimate. Because the claim as stated covers all defined applicant groups, the evidence does not provide full coverage of the claim.
Sufficiency	Partially met. Taken together, and at the calibrated R3 rigour level, the body of evidence supports the claim for the applicant groups with sufficient representation but does not support the claim for the under-represented group. The overall evidence base is therefore only partially sufficient for the claim as stated.

Verification performed at R3. The internal audit function verified the disaggregated analysis against source application records on a sample basis, and confirmed the group-size figure that produced the coverage finding.

Interface return triggered. The coverage gap may materially affect the linked risk entry because one applicant group cannot be evaluated reliably within the current reporting period. The issue is therefore routed back through the Risk-Assurance Interface for risk-management consideration of whether the risk entry, controls, or monitoring arrangements require amendment. Assurance of C-O4 continues in parallel, while any revised risk-management outputs may subsequently re-enter the assurance process through the interface.

Stage 4 — Assurance conclusion and reassurance (Section 7.5).

Claim-level conclusion. Partially supported. The claim is supported for applicant groups [list] and is not supported by the available evidence for group [x], where the number of applications in the reporting period is insufficient to evaluate the claim reliably. This is a statement about the adequacy of the available evidence and is not a finding that the equal-treatment requirement was breached for that group.

Governance report. The partial conclusion, the coverage limitation, the R3/I3 configuration with no departures, and the interface return are reported to the body responsible for the relevant organisational decision before that decision is taken.

Organisational decision (not part of assurance). The deciding body may, for instance, continue operation with an added condition requiring cumulative monitoring across periods for

the under-represented group. That decision remains the responsibility of the entity; the assurance conclusion neither authorises nor prohibits continued operation, and the assurance record should preserve that distinction.

External communication, where applicable. Where publication or other external communication is required by applicable law or policy, or is otherwise appropriate to the system's context, a summary may state what was assured, the applicable rigour and independence levels, the assurance conclusion by category, and the material coverage limitation. Any such communication should avoid disclosing personal data, security-sensitive information, confidential information, or details that could enable circumvention of the system.

Reassurance triggers recorded. Re-evaluate C-O4 on expiry of the stated reporting period; any material change to the screening release; any change to the equal-treatment requirement; or receipt of revised risk-management outputs arising from the returned coverage finding.

What the example demonstrates, and what it does not. It demonstrates that the framework produces a specific, defensible, and reviewable output for a realistic claim, including where the available evidence supports the claim only partially. It also exercises the two-way Risk-Assurance Interface, the presumptive floors, the evidence-adequacy criteria, the conclusion categories, and the arrangements for recording and communicating the assurance conclusion where applicable. It does not demonstrate that the framework is affordable, that the presumptive floors are set at the appropriate levels, or that the effort involved is proportionate; these remain questions to be tested through the pilots specified in Table 13.

Copyright and publication

© 2026 Nawaf Alsahmah. Published by the International Center for Artificial Intelligence Research and Ethics (ICAIRE), Riyadh, Saudi Arabia. All rights reserved.

Review

This paper was reviewed by the ICAIRE Scientific Committee and approved for publication by the Committee Chair, subject to revisions, which were incorporated by ICAIRE in accordance with the ICAIRE Research Policy.

Disclaimer

The views expressed are those of the author and do not necessarily reflect the positions of ICAIRE or UNESCO.

Suggested citation

Alsahmah, N. (2026). Risk-based AI assurance in the Saudi public sector: Aligning international assurance practices with the National AI Risk Management Framework (ICAIRE Framework Paper). International Center for Artificial Intelligence Research and Ethics. <https://doi.org/10.68480/CEYS4608>