

Artificial Intelligence Review Committee (AIRC)

Comprehensive Grading Guide for All Rubrics

Version: 2.0 (December 2025)

Purpose: To provide detailed, consistent guidance on scoring all AIRC rubrics (Human Subjects, Animal Research, and Analytic/Non-Human Subjects; Streamlined and Enhanced versions) to ensure calibration and fairness across reviewers and protocols.

Table of Contents

1. Introduction & Overview
2. General Scoring Principles
3. Using the 1-4 Scale
4. Applying N/A and Insufficient Documentation
5. Domain-Specific Grading Guidance
6. Critical Deficiency Rules
7. Examples & Case Scenarios
8. Calibration & Consistency Tips

1. Introduction & Overview

Purpose of This Guide

AIRC reviewers come from diverse disciplinary backgrounds and may have varying interpretations of rubric criteria. This guide ensures that:

- A score of "3 - Proficient" means the same thing regardless of rubric type or reviewer
- Scoring decisions are defensible and transparent
- Institutional AIRC programs can calibrate reviewers using consistent examples
- Quality and consistency improve over time

Scope

This guide applies to:

- Human Subjects Research Rubric (Streamlined & Enhanced)

- Animal Research Rubric (Streamlined & Enhanced)
- Analytic/Non-Human Subjects Rubric (Streamlined & Enhanced)

When to Use This Guide

- Before your first review: Familiarize yourself with general principles and domain-specific guidance
- During a review: Reference specific domains where you're uncertain about scoring
- After a review: Use examples to calibrate your decisions with other reviewers
- During training sessions: Use to align team understanding of scoring criteria

2. General Scoring Principles

Core Principle: "Evidence-Based Scoring"

Score based on what is documented in the protocol, not what you assume or hope is true.

Principle	What This Means	Example
Score only documented information	If the protocol does not describe a data validation process, you cannot give it a "4" for data quality, even if the data source is well-known	Reviewer reads protocol and finds no data quality documentation → Score 2 or lower, note the gap
Use scoring anchors, not opinion	Base your decision on the rubric's specific 4-3-2-1 descriptions, not your personal standards	If rubric says "3" requires "adequate documentation" and the protocol has "adequate" documentation, score 3 (not 4 just because you like the protocol)
Differentiate between domains	A protocol might score 4 on Data Quality but 2 on Fairness. Score each independently	Don't let a strong score in one domain inflate scores in others

Principle	What This Means	Example
Comment generously	Your notes are as important as your score. They help the AIRC chair, the committee, and the PI	Always explain the reason behind scores of 1 or 2

Scoring Philosophy by Level

Score	Philosophy	Review Mindset
4 - Exemplary	Exceeds institutional standards; proactive, comprehensive approach	"This is a model for other protocols; could be published as a case study"
3 - Proficient	Meets institutional standards; appropriate, adequate approach	"This is acceptable; the committee can approve with confidence."
2 - Basic	Below institutional standards; minimal compliance; gaps present	"This is borderline; the committee should require modifications before approval."
1 - Deficient	Does not meet standards; major gaps or unacceptable risks	"This cannot be approved in its current form; major revision required."

3. Using the 1-4 Scale

The 1-4 Scale Explained

Score	Label	Meaning	Next Step
4	Exemplary	Comprehensive, proactive, exceeds standards	Include in final rubric; highlight in summary
3	Proficient	Clear documentation; meets standards; appropriate	Include in final rubric; note as acceptable
2	Basic	Documented but minimal; gaps present; borderline	Flag in "Modifications Required" list
1	Deficient	Major gaps; inadequate; unacceptable risk	Critical deficiency rule: "Not Acceptable"

What "Comprehensive" Means at Each Level

Exemplary (4):

- Multiple layers of evidence/safeguards (not just one approach)
- Specific details, examples, or metrics provided
- Proactive planning beyond minimum requirements
- Evidence of stakeholder engagement or external validation
- Example: "Data quality assessment includes statistical validation across demographic subgroups"

Proficient (3):

- One clear, well-documented approach meeting the standard
- Adequate specificity and clarity

- Appropriate for the risk level of the protocol
- Aligned with institutional policy
- Example: "Data quality procedures documented; quality control process in place"

Basic (2):

- Mentioned but limited detail or documentation
- One approach present but gaps remain
- Meets some but not all standard elements
- May need clarification or strengthening
- Example: "Data cleaning procedures mentioned but details not provided"

Deficient (1):

- Missing or inadequate
- Does not meet minimum standard
- Raises unacceptable risk
- Requires major revision
- Example: "No data quality assessment or procedures described"

4. Applying N/A and Insufficient Documentation

When to Use "N/A" (Not Applicable)

Use N/A when a checklist item genuinely does not apply to the protocol's context.

Examples of Appropriate N/A:

- A de-identified data analysis project has no human consent requirements → Human Subjects Privacy domain, item about "informed consent" = N/A
- A Streamlined animal research rubric doesn't include a "Group Harms" domain → Doesn't apply = N/A
- An AI tool is used for visualization only and makes no clinical or research decisions → Human oversight/decision authority = N/A

CRITICAL: Always provide a brief justification when selecting N/A:

Rubric	Item	Response	Justification
Human Subjects - Streamlined	Privacy & Consent	N/A	[Justification: "This project uses de-identified data; no consent required under IRB exemption"]

Red Flag: If you're using N/A frequently, double-check that you've selected the correct rubric.

When to Use "Insufficient Documentation"

Select this option when you cannot score because required information is missing, unclear, or inadequate.

Examples:

- Protocol mentions using "a machine learning algorithm" but does not specify which algorithm, version, or validation status
- Data source is not clearly described
- Human oversight roles are mentioned but decision authority is not defined
- Bias mitigation strategy is stated but no monitoring plan is provided

Process for Insufficient Documentation:

1. Select "Insufficient Documentation" for the domain
2. Document exactly what information is needed in Reviewer Notes
3. Return rubric to coordinator or PI with specific checklist (e.g., "Please provide: algorithm version, validation study results, performance metrics")
4. Resubmit and reschedule review after clarification

Do NOT score 1 or 2 just because information is incomplete. Use "Insufficient Documentation" to trigger resubmission for clarification.

5. Domain-Specific Grading Guidance

HUMAN SUBJECTS RUBRICS (Streamlined & Enhanced)

Domain 1: Data Quality & Provenance (Streamlined) / Data Quality & Provenance (Enhanced)

Purpose: Ensure data are high-quality, representative, and ethically sourced.

Streamlined Version - 4 Domains

Score	Criteria	Indicators	Red Flags
4 - Exemplary	Comprehensive documentation with proactive representativeness analysis	<ul style="list-style-type: none"> - Complete dataset lineage with collection dates, methods, and source documentation - Statistical validation of representativeness (e.g., demographics compared to target population) - Detailed preprocessing pipeline (outlier handling, missing value imputation, standardization) - Advanced de-identification methods with formal re-identification risk assessment 	<ul style="list-style-type: none"> - Vague data source (e.g., "obtained from hospital database") - No representativeness analysis - Preprocessing not documented - De-identification method not specified
3 - Proficient	Clear documentation with adequate representativeness	<ul style="list-style-type: none"> - Data sources identified with collection methods - Demographics provided; limitations acknowledged - Standard preprocessing steps described - Appropriate de-identification per institutional policy 	<ul style="list-style-type: none"> - Limited detail on how data were collected - Representativeness mentioned but not statistically validated - Basic preprocessing mentioned but not detailed - De-identification

Score	Criteria	Indicators	Red Flags
			method generic or unclear
2 - Basic	Minimal documentation with gaps	<ul style="list-style-type: none"> - Data source identified but limited lineage - Some representativeness info but incomplete - Basic preprocessing mentioned - De-identification claimed but not fully documented 	<ul style="list-style-type: none"> - No data source information - No representativeness consideration - Preprocessing gaps - Privacy protections unclear
1 - Deficient	Poor documentation or unrepresentative data	<ul style="list-style-type: none"> - Data source missing or undocumented - No representativeness analysis; clearly mismatched population - No preprocessing information - Inadequate privacy protections or high re-identification risk 	<ul style="list-style-type: none"> - Return to submitter; major revision required

How to Apply This Guidance:

When scoring Domain 1, ask yourself:

1. Can I trace where the data came from and how they were collected? (If no → score 2 or lower)
2. Do the data represent the population the study targets? (If unstated or unclear → score 2 or lower)
3. Are data cleaning and preprocessing steps documented? (If not → score 2 or lower)

4. Are privacy/de-identification methods appropriate for the sensitivity? (If inadequate → score 1-2)

Example Scoring:

Protocol A: "We obtained de-identified EHR data from [Hospital Name] for patients with diagnosis codes X, Y, Z from 2018-2023. Data quality was verified through automated validation checks. Demographics: 45% female, 38% Hispanic, mean age 58. De-identification per HIPAA Safe Harbor."

→ Score: 3 (Proficient) – Clear source, adequate data quality description, demographic representation provided, appropriate de-identification. Not a 4 because no formal representativeness analysis or advanced validation metrics provided.

Protocol B: "We will use AI to analyze patient data."

→ Score: 1 (Deficient) – Insufficient documentation. No data source, preprocessing, representativeness, or de-identification described.

Domain 2: Privacy & Informed Consent (Streamlined - Human) / Privacy & Informed Consent (Enhanced - Human)

Purpose: Ensure participant privacy is protected and consent appropriately addresses AI use.

Score	Criteria	Indicators	Red Flags
4 - Exemplary	Enterprise-grade security; detailed, participant-friendly consent explaining AI	<ul style="list-style-type: none"> - Encryption at rest and in transit; multi-factor authentication - Consent form includes: AI tool purpose, how AI is used, AI risks/benefits, data retention, participant rights - Access controls with audit trails documented - Formal data management plan with secure destruction procedures - Evidence of security review or compliance audit 	<ul style="list-style-type: none"> - Consent vague about AI use - Standard security without details - No data destruction plan - Access controls not documented

Score	Criteria	Indicators	Red Flags
3 - Proficient	Adequate security meeting standards; consent addresses AI	<ul style="list-style-type: none"> - Security meeting HIPAA or equivalent standards documented - Consent mentions AI use and appropriate for risk level - Access controls and authentication in place - Data management plan with retention schedule - Typical institutional security procedures 	<ul style="list-style-type: none"> - Consent does not mention AI - Security measures vaguely described - No retention/destruction plan - Limited access controls
2 - Basic	Minimal security; consent lacks AI detail	<ul style="list-style-type: none"> - Basic security measures mentioned without full documentation - Consent mentions AI but lacks specificity about risks - Limited access controls - Basic data management plan without details 	<ul style="list-style-type: none"> - No security measures described - Consent does not address AI - No data management plan - Unclear access to data
1 - Deficient	Inadequate privacy or consent	<ul style="list-style-type: none"> - Security insufficient or poorly documented - Consent does not address AI; inadequate for informed decision - No access controls - No data management or retention plan 	<ul style="list-style-type: none"> - Major vulnerabilities; return for revision

How to Apply This Guidance:

1. Read the consent form – Does it explain what AI will be used for and any risks? (If not → score 2 or lower)
2. Assess security measures – Are they documented and appropriate for the data sensitivity? (If vague → score 2)

3. Check data management – Is there a plan for storage, access, retention, and secure destruction? (If not → score 2 or lower)

Example Scoring:

Protocol A: Consent form states: "This study uses an AI tool to predict patient outcomes. The AI has been validated in [prior study]. Your data will be encrypted and stored on secure servers. You can request deletion of your data. [Details on retention period provided]."

Security: HIPAA-compliant infrastructure, encryption standard, access limited to study staff with role-based controls, audit logs maintained.

→ Score: 3 (Proficient) – Consent adequately addresses AI; standard security meets institutional policy. Not a 4 because consent could be more specific about AI limitations or benefits.

Domain 3: Fairness & Risk (Streamlined - Human) / Bias & Fairness (Enhanced - Human)

Purpose: Ensure AI performs equitably and risks are managed.

Score	Criteria	Indicators	Red Flags
4 - Exemplary	Comprehensive bias evaluation; proactive mitigation with monitoring	<ul style="list-style-type: none"> - Performance evaluated across multiple demographic dimensions (race, ethnicity, sex, age, SES) with quantitative metrics (e.g., sensitivity, specificity by group) - Fairness definition selected with stakeholder input - Documented mitigation strategies with specific interventions (e.g., threshold adjustment, rebalancing) - Continuous monitoring plan with defined thresholds for intervention - Group harm assessment with mitigation strategies - Community engagement documented 	<ul style="list-style-type: none"> - No bias evaluation - Fairness not defined - No mitigation plan - No monitoring planned - No stakeholder engagement

Score	Criteria	Indicators	Red Flags
3 - Proficient	Bias evaluated across major groups; mitigation strategy documented	<ul style="list-style-type: none"> - Performance evaluated across 2-3 major demographic dimensions - Fairness definition with clear rationale - Documented bias mitigation plan with specific actions - Monitoring procedures specified - Basic stakeholder consultation conducted - Group harms considered 	<ul style="list-style-type: none"> - Limited demographic analysis - Fairness mentioned but not formally assessed - Vague mitigation plan - Minimal monitoring - No stakeholder involvement
2 - Basic	Limited bias assessment; vague mitigation	<ul style="list-style-type: none"> - Bias audit mentions 1-2 demographic dimensions - Fairness acknowledged but not formally assessed - Mitigation plan mentioned but lacks specificity - Limited monitoring procedures - Minimal stakeholder engagement 	<ul style="list-style-type: none"> - No bias assessment - Fairness not considered - No mitigation plan - No monitoring planned
1 - Deficient	No bias assessment or unmitigated disparities	<ul style="list-style-type: none"> - No subgroup performance analysis - Fairness not addressed - No mitigation plan despite known disparities - No monitoring - No stakeholder involvement 	<ul style="list-style-type: none"> - Major ethical concern; return for major revision

How to Apply This Guidance:

1. Look for subgroup analysis – Has the protocol evaluated performance across demographic groups? (If no → score 2 or lower)
2. Assess fairness planning – Is there a definition of fairness and rationale for choice? (If not → score 2)
3. Check for mitigation – Are specific strategies documented to address identified bias? (If no → score 2)
4. Review monitoring – Is there a plan to detect emergent bias during the study? (If not → score 2)
5. Consider group harms – Could results harm specific populations if deployed? (If not addressed → score 2)

Example Scoring:

Protocol A: "We will evaluate model performance (sensitivity, specificity, AUC) overall and stratified by race/ethnicity, sex, and age groups. If performance disparities >5% are detected, we will adjust thresholds or retrain model. We consulted with [Community Advisory Board] on fairness definition and mitigation priorities."

→ Score: 4 (Exemplary) – Comprehensive bias evaluation across multiple dimensions, quantitative metrics, community engagement, clear mitigation strategy.

Protocol B: "We will ensure our model is fair to all patients."

→ Score: 1 (Deficient) – No specific bias evaluation plan, no metrics, no mitigation strategy.

Domain 4: Security & Governance (Streamlined - Human)

Score	Criteria	Indicators
4 - Exemplary	Enterprise-grade security; formal governance	<ul style="list-style-type: none"> - Multi-layered security (encryption, authentication, access controls, audit trails) - Formal data governance policy documented - Security review or compliance audit completed - Incident response plan - Regular security training for staff

Score	Criteria	Indicators
3 - Proficient	Adequate security meeting standards	<ul style="list-style-type: none">- Security measures meeting institutional standards- Access controls and authentication in place- Data governance policy referenced- Incident procedures defined
2 - Basic	Basic security; governance mentioned	<ul style="list-style-type: none">- Standard security practices mentioned- Limited documentation of governance- Vague incident procedures
1 - Deficient	Inadequate or undocumented security	<ul style="list-style-type: none">- Security measures insufficient or absent- No governance framework- No incident planning

ANIMAL RESEARCH RUBRICS (Streamlined & Enhanced)

Domain 1: Scientific Validity & Justification (Streamlined - Animal) / Data/Method Validity (Enhanced - Animal)

Purpose: Ensure AI methodology is scientifically sound and appropriate for the animal model.

Score	Criteria	Indicators	Red Flags
4 - Exemplary	Rigorous methodology with comprehensive validation	<ul style="list-style-type: none"> - AI methodology optimally matched to research question with detailed justification - Complete data provenance (training dataset source, size, characteristics) - Performance metrics exceed field standards (e.g., published benchmarks) - External validation or published evidence provided - Clear scientific rationale for AI over alternatives 	<ul style="list-style-type: none"> - AI approach not justified - Data sources unclear - Performance metrics missing - No evidence of superiority over alternatives
3 - Proficient	Sound methodology with appropriate validation	<ul style="list-style-type: none"> - AI approach appropriate for research question; rationale provided - Data sources documented - Performance metrics reported and acceptable - Validation or testing documented - Scientific rationale provided 	<ul style="list-style-type: none"> - Limited justification - Data sources vague - Performance metrics incomplete - Validation minimal
2 - Basic	Adequate methodology with validation gaps	<ul style="list-style-type: none"> - AI method applicable but limited rationale - Data sources identified but incomplete documentation 	<ul style="list-style-type: none"> - AI approach unclear - Data sources not

Score	Criteria	Indicators	Red Flags
		<ul style="list-style-type: none"> - Limited validation information - Weak justification versus alternatives 	<ul style="list-style-type: none"> - documented - No performance metrics - No alternatives considered
1 - Deficient	Inappropriate methodology or inadequate validation	<ul style="list-style-type: none"> - AI approach unsuitable for research question - Poor data quality or undocumented sources - Missing or inadequate performance metrics - No scientific justification 	<ul style="list-style-type: none"> - Major methodological concern; return for revision

Example Scoring:

Protocol A (Animal, Valid): "We will use a machine learning model trained on [reference dataset, size N=10,000] to predict which compounds are neurotoxic in rats before in vivo testing. The model was validated on [independent test set, N=2,000] with sensitivity 92%, specificity 88%, consistent with [published benchmark study]. This AI approach enables screening of 1,000+ compounds, reducing in vivo testing by 60%."

→ Score: 4 (Exemplary) – Clear scientific rationale, comprehensive validation, performance metrics exceed standards, replacement benefit documented.

Domain 2: Animal Welfare & 3Rs (Streamlined & Enhanced - Animal)

Purpose: Ensure AI advances replacement, reduction, and refinement of animal use.

Score	Criteria	Indicators	Red Flags
4 - Exemplary	AI demonstrably advances 3Rs principles	<ul style="list-style-type: none"> - Replacement: AI model demonstrates 30-50%+ reduction in required in vivo studies - Reduction: Animal numbers justified and minimized by AI screening/predictive modeling - Refinement: AI identifies procedures or dosages that minimize pain/distress - Enhanced welfare monitoring through AI capabilities - Comprehensive veterinary oversight plan 	<ul style="list-style-type: none"> - No 3Rs consideration - Animal numbers not justified - No refinement benefits - Poor welfare procedures - Insufficient veterinary oversight
3 - Proficient	AI contributes to one or more 3Rs	<ul style="list-style-type: none"> - AI contributes meaningfully to replacement, reduction, or refinement - Animal numbers justified - Refinement benefits documented - Standard welfare procedures in place - Veterinary oversight appropriate 	<ul style="list-style-type: none"> - Limited 3Rs impact - Animal numbers not justified - Minimal welfare benefits
2 - Basic	Minimal 3Rs impact; standard welfare	<ul style="list-style-type: none"> - Limited AI contribution to 3Rs - Animal numbers justified but no reduction - Minimal refinement benefits - Standard welfare procedures 	<ul style="list-style-type: none"> - No 3Rs consideration - Unjustified animal numbers - No welfare safeguards
1 - Deficient	No 3Rs consideration or welfare concerns	<ul style="list-style-type: none"> - AI does not advance 3Rs principles - Excessive animal numbers without justification 	<ul style="list-style-type: none"> - Critical animal welfare

Score	Criteria	Indicators	Red Flags
		<ul style="list-style-type: none"> - No refinement or potential for increased distress - Inadequate welfare procedures 	concern; return for revision

How to Apply This Guidance:

1. Replacement – Does AI replace or eliminate some in vivo work? (If no → score 2)
2. Reduction – Does AI reduce the number of animals needed while maintaining scientific validity? (If no reduction justified → score 2)
3. Refinement – Does AI refine procedures to minimize pain/distress? (If no → score 2)
4. Welfare Oversight – Are veterinary oversight and welfare monitoring adequate? (If not → score 2)

Domain 3: Harm Minimization & Safety (Streamlined & Enhanced - Animal)

Purpose: Ensure AI-driven procedures do not cause unnecessary harm to animals.

Score	Criteria	Indicators	Red Flags
4 - Exemplary	Comprehensive harm prevention with robust safety	<ul style="list-style-type: none"> - All AI-related risks identified and assessed - Multi-layered safety mechanisms for AI-controlled procedures - Strong human oversight with clear intervention protocols - Comprehensive error analysis with detailed contingency plans - Active monitoring system with defined response procedures 	<ul style="list-style-type: none"> - AI risks not identified - No safety mechanisms - No human oversight - No error contingencies - No monitoring

Score	Criteria	Indicators	Red Flags
3 - Proficient	Adequate harm prevention with appropriate safeguards	<ul style="list-style-type: none"> - Major AI risks identified and assessed - Safety mechanisms in place for critical procedures - Human oversight defined - Error scenarios considered with basic contingencies - Monitoring plan documented 	<ul style="list-style-type: none"> - Limited risk assessment - Vague safety mechanisms - Minimal oversight - No monitoring planned
2 - Basic	Minimal harm assessment; basic safeguards	<ul style="list-style-type: none"> - Some AI risks identified - Basic safety mechanisms mentioned - Vague human oversight - Limited error consideration 	<ul style="list-style-type: none"> - AI risks not identified - No safeguards - No oversight defined - No monitoring plan
1 - Deficient	Inadequate harm prevention or unacceptable risk	<ul style="list-style-type: none"> - AI risks not identified or underestimated - No safety mechanisms - No human oversight - Potential for increased animal harm 	<ul style="list-style-type: none"> - Critical safety concern; return for revision

Example Scenario:

Protocol: "AI will control dosing in continuous infusion pump. If AI-predicted dose exceeds safe range or sensor malfunction detected, pump defaults to 50% of standard dose. Veterinarian monitors animals hourly; any signs of distress trigger immediate review and manual override. Error logs reviewed daily."

→ Score: 4 (Exemplary) – Clear harm identification, multi-layered safeguards (dose limits, failsafe, monitoring, manual override), active oversight.

Domain 4: Transparency & Documentation (Streamlined & Enhanced - Animal)

Score	Criteria	Indicators
4 - Exemplary	Algorithm fully specified with version control	<ul style="list-style-type: none"> - Algorithm type, architecture, version with unique identifiers - Performance metrics thoroughly reported - Formal change management procedures - Complete methods documentation - Enterprise-grade data security
3 - Proficient	Adequate documentation and transparency	<ul style="list-style-type: none"> - Algorithm and version documented - Performance metrics reported - Version tracking in place - Methods adequately described
2 - Basic	Minimal documentation with gaps	<ul style="list-style-type: none"> - Algorithm identified with limited detail - Basic performance metrics - Minimal version control
1 - Deficient	Poor documentation or lack of transparency	<ul style="list-style-type: none"> - Algorithm unclear - Missing metrics - No version control

ANALYTIC/NON-HUMAN SUBJECTS RUBRICS (Streamlined & Enhanced)

Domain 1: Data Provenance & Quality (Streamlined & Enhanced - Analytic)

Score	Criteria	Indicators	Red Flags
4 - Exemplary	Complete provenance with quality validation	<ul style="list-style-type: none"> - Complete lineage with detailed documentation of source, collection methods, dates - Formal quality assessment with validation metrics - Robust preprocessing pipeline fully documented (outlier handling, missing values, standardization) - Clear inclusion/exclusion criteria with statistical justification - Representativeness analysis with statistical validation 	<ul style="list-style-type: none"> - Vague data source - No quality assessment - No preprocessing documentation - No inclusion/exclusion criteria - No representativeness analysis
3 - Proficient	Clear documentation with adequate quality	<ul style="list-style-type: none"> - Data sources clearly identified - Quality appropriate for analysis; acceptable documentation - Preprocessing documented - Inclusion/exclusion criteria defined - Basic representativeness assessment 	<ul style="list-style-type: none"> - Limited data source detail - Quality not validated - Preprocessing vague - Representativeness unclear
2 - Basic	Minimal documentation with gaps	<ul style="list-style-type: none"> - Data sources identified with limited detail - Quality acceptable but not validated - Basic preprocessing mentioned - Vague criteria - Limited representativeness 	<ul style="list-style-type: none"> - No data source info - Poor data quality - No preprocessing - No inclusion/exclusion

Score	Criteria	Indicators	Red Flags
1 - Deficient	Poor documentation or inadequate quality	<ul style="list-style-type: none"> - Sources unclear or undocumented - Poor data quality - No preprocessing information - No inclusion/exclusion criteria - No representativeness assessment 	- Major data quality concern; return for revision

Domain 2: Privacy/Re-identification Risk (Streamlined & Enhanced - Analytic)

Score	Criteria	Indicators	Red Flags
4 - Exemplary	Expert determination with comprehensive risk assessment	<ul style="list-style-type: none"> - Expert determination or equivalent validation of de-identification - Comprehensive re-identification risk analysis with quantitative metrics - Multiple safeguards throughout data lifecycle (secure storage, access controls, monitoring) - Enterprise-grade security (encryption, audit trails) - Detailed data management plan with compliance verification 	<ul style="list-style-type: none"> - De-identification not validated - No risk assessment - No safeguards - Poor security - No data management
3 - Proficient	Formal de-identification with risk consideration	<ul style="list-style-type: none"> - Formal de-identification process documented - Re-identification risk considered and acceptable - Appropriate safeguards in place - Security meeting institutional standards - Data management plan documented 	<ul style="list-style-type: none"> - De-identification method unclear - Limited risk assessment - Minimal safeguards

Score	Criteria	Indicators	Red Flags
			- Basic security
2 - Basic	Basic de-identification with limited risk assessment	<ul style="list-style-type: none"> - De-identification claimed but not validated - Limited risk assessment - Minimal safeguards - Basic security measures - Vague data management 	<ul style="list-style-type: none"> - De-identification inadequate - No risk assessment - No safeguards - Poor security
1 - Deficient	Inadequate de-identification or high risk	<ul style="list-style-type: none"> - De-identification questionable or inadequate - No risk assessment - No safeguards - Poor security - No data management plan 	- High re-identification risk; return for revision

How to Apply This Guidance:

- Is de-identification formally validated? (HIPAA Safe Harbor, Expert Determination, or equivalent)
 - If yes, consider 3 or 4
 - If no, score 2 or lower
- Is re-identification risk quantified or assessed?
 - Quantitative assessment with metrics → consider 4
 - Qualitative assessment → consider 3
 - No assessment → score 2
- Are safeguards documented? (Access controls, encryption, audit logging)

- Comprehensive safeguards → consider 3 or 4
- Basic safeguards → score 2
- No safeguards → score 1

Domain 3: Analytic Validity/Methodology (Streamlined & Enhanced - Analytic)

Score	Criteria	Indicators	Red Flags
4 - Exemplary	Rigorous methodology with comprehensive validation	<ul style="list-style-type: none"> - AI approach optimally suited to question with detailed justification - Performance metrics exceed field standards with external validation - Robust validation (cross-validation, sensitivity analysis, test set evaluation) - Statistical methods thoroughly documented with assumption checks - Comprehensive limitations discussion with mitigation strategies 	<ul style="list-style-type: none"> - AI approach unsuitable - Missing or inadequate metrics - No validation - Statistical methods unclear - No limitations
3 - Proficient	Sound methodology with appropriate validation	<ul style="list-style-type: none"> - AI approach appropriate and justified - Performance metrics adequate and documented - Validation procedures specified - Statistical methods clear - Key limitations acknowledged 	<ul style="list-style-type: none"> - Limited justification - Basic metrics - Minimal validation - Vague methods - No limitations
2 - Basic	Adequate methodology with validation gaps	<ul style="list-style-type: none"> - AI approach acceptable but limited justification - Basic performance metrics 	<ul style="list-style-type: none"> - Approach unclear - Missing

Score	Criteria	Indicators	Red Flags
		<ul style="list-style-type: none"> - Limited validation - Statistical methods vaguely described - Limited limitations discussion 	<ul style="list-style-type: none"> metrics - No validation - Methods inappropriate
1 - Deficient	Inappropriate methodology or inadequate validation	<ul style="list-style-type: none"> - AI approach unsuitable for question - Missing or inadequate performance metrics - No validation - Statistical methods unclear or inappropriate - No limitations discussion 	<ul style="list-style-type: none"> - Major analytical concern; return for revision

Example Scoring:

Analytic Protocol A: "We trained a gradient boosting model to predict disease onset using 50,000 patient records (training n=35,000, test n=15,000). Model performance: AUC 0.87 (95% CI 0.84-0.90). Validation: 5-fold cross-validation, sensitivity analysis with alternative feature sets. Model uncertainty quantified; predictions include 95% confidence intervals. Limitations: [detailed discussion of data gaps, potential biases, generalizability]."

→ Score: 4 (Exemplary) – Rigorous validation approach, confidence intervals, sensitivity analysis, comprehensive limitations.

Domain 4: Transparency/Reproducibility (Streamlined & Enhanced - Analytic)

Score	Criteria	Indicators	Red Flags
4 - Exemplary	Complete transparency enabling full replication	<ul style="list-style-type: none"> - Algorithm fully specified with public code repository - Comprehensive methods enabling exact replication - Code publicly available or 	<ul style="list-style-type: none"> - Algorithm unclear - No code availability - Poor methods

Score	Criteria	Indicators	Red Flags
		accessible upon request - Formal version control with change logs - All analysis decisions documented with justifications	description - No version control - Decisions undocumented
3 - Proficient	Adequate transparency supporting reproducibility	- Algorithm and version documented with rationale - Methods sufficiently detailed for replication - Code availability plan specified - Version tracking in place - Major analysis decisions documented	- Limited detail - Vague code availability - Minimal version control - Limited decision documentation
2 - Basic	Minimal transparency with reproducibility challenges	- Algorithm identified with limited detail - Methods description incomplete - Unclear code availability - Minimal version control - Limited analysis decision documentation	- Algorithm unclear - Methods inadequate - No code availability - No version control
1 - Deficient	Poor transparency preventing replication	- Algorithm unclear or unjustified - Methods inadequate for replication - No code availability or sharing plan - No version control - Analysis decisions undocumented	- Cannot be reproduced; return for major revision

Domain 5: Group Harms/Societal Risk (Streamlined & Enhanced - Analytic)

Score	Criteria	Indicators	Red Flags
4 - Exemplary	Detailed group harm assessment with mitigation	<ul style="list-style-type: none"> - Detailed analysis of potential group-level impacts across multiple dimensions (e.g., employment, healthcare access, lending) - Thorough bias and discrimination assessment - Proactive mitigation strategies with monitoring plans - Careful consideration of use/misuse scenarios with safeguards - Responsible dissemination plan addressing potential harms 	<ul style="list-style-type: none"> - No group harm consideration - No bias assessment - No mitigation - No use/misuse planning - Inappropriate dissemination
3 - Proficient	Major group harms assessed with mitigation	<ul style="list-style-type: none"> - Group-level impacts identified and assessed - Bias considerations documented - Mitigation strategies in place - Use/misuse scenarios considered - Appropriate dissemination plan 	<ul style="list-style-type: none"> - Limited harm assessment - Bias mentioned but not detailed - Vague mitigation - No use/misuse planning
2 - Basic	Limited group harm consideration	<ul style="list-style-type: none"> - Some group risks identified - Minimal bias assessment - Vague mitigation - Limited use/misuse consideration - Basic dissemination plan 	<ul style="list-style-type: none"> - No harm assessment - No bias consideration - No mitigation - No dissemination planning

Score	Criteria	Indicators	Red Flags
1 - Deficient	No consideration of group harms or high risk	<ul style="list-style-type: none">- No potential group harms assessed- No bias assessment despite clear risk- No mitigation plans- No use/misuse scenarios considered- Inappropriate or irresponsible dissemination	<ul style="list-style-type: none">- Significant group harm risk; return for major revision

Example Scoring:

Analytic Protocol A: "This algorithm predicts which applicants will be 'reliable' employees for low-wage positions. We assessed disparate impact by race, gender, and age using statistical tests. Results showed 12% higher false positive rate for Black applicants. Mitigation: [specific threshold adjustments, retrained models, regular fairness audits]. Use restrictions: [internal HR only; not used for initial screening]. Dissemination: [findings will disclose bias analysis and limitations]."

→ Score: 4 (Exemplary) – Comprehensive group harm assessment, specific bias identified, clear mitigation, responsible use restrictions.

6. Critical Deficiency Rules

Universal Critical Deficiency Rule

The single most important rule in the AIRC framework is the Critical Deficiency Rule. It is designed to be simple and absolute to ensure that protocols with major, unaddressed flaws are not approved.

Rule:

A score of 1 ("Deficient") in ANY domain automatically results in a final recommendation of "Not Acceptable."

This means that even if a protocol scores highly in four domains, a single score of 1 in the fifth domain is sufficient to prevent its approval in its current state. The PI must make major revisions to address the deficiency and resubmit the protocol for a new AIRC review.

If a Domain Scores 1...	Example Deficiency	Outcome
Data Quality/Provenance	Data source is undocumented, or quality is insufficient for the research question.	Not Acceptable
Privacy/Consent	Personally identifiable information is used without adequate security, or the consent process is inappropriate for AI use.	Not Acceptable
Fairness/Bias/Group Harms	Significant performance disparities across demographic groups are present with no mitigation plan.	Not Acceptable
Risk/Animal Welfare	Unacceptable risk to human participants or animal welfare is identified.	Not Acceptable
Transparency/Documentation	The AI model is a "black box" with no documentation, preventing any form of validation or oversight.	Not Acceptable

Rationale:

A score of 1 signifies a fundamental failure in one of the core pillars of responsible AI research. Strengths in other areas cannot outweigh such a failure. This rule ensures a minimum standard is met across all critical aspects of the research.

Implementation:

When completing a rubric, if you assign a score of 1 to any domain:

1. Provide a detailed, clear explanation in the "Reviewer Notes" for that domain, explaining why it is deficient.
2. Check the box in the "Critical Deficiency Rule" section.
3. Ensure the "Final Recommendation" is marked as "NOT ACCEPTABLE."

7. Examples & Case Scenarios

SCENARIO 1: Human Subjects - Streamlined Rubric

Protocol: "Predictive Model for Sepsis Risk in ICU Patients"

What You Receive:

- Training data: 20,000 ICU patients (2010-2015) from a single hospital
- Model: Logistic regression predicting sepsis within 48 hours
- Validation: Single internal test set
- Consent: Standard research consent (mentions "research" but not AI)
- Data storage: HIPAA-compliant server
- No fairness analysis conducted
- No privacy risk assessment

Scoring:

Domain	Score	Justification
Domain 1: Data Quality	2	Single hospital data limits generalizability; no representativeness analysis; a single validation set is minimal
Domain 2: Privacy & Consent	2	Consent does not explicitly address AI use; privacy is appropriate but not comprehensive
Domain 3: Fairness & Risk	1	DEFICIENT – No fairness analysis; no demographic subgroup evaluation; high risk in ICU setting with vulnerable population
Domain 4: Security	3	HIPAA-compliant; standard institutional practices
Recommendation	NOT ACCEPTABLE	Domain 3 critical deficiency. Return for: (1) fairness analysis across race/gender/age, (2)

Domain	Score	Justification
		consent revision addressing AI, (3) external validation

SCENARIO 2: Animal Research - Enhanced Rubric

Protocol: "AI-Optimized Dosing for Novel Cancer Drug in Mice"

What You Receive:

- AI system: Neural network trained on 500 prior mouse trials
- Use: Real-time dose adjustment during chemotherapy trials
- Safety: AI flagged to halt dosing if toxicity predicted; veterinarian reviews daily
- Animal numbers: 100 mice (reduced from 200 historical studies using same protocol)
- 3Rs: AI enables 50% reduction in animal numbers via predictive screening
- Documentation: Algorithm version not specified; no performance metrics provided

Scoring:

Domain	Score	Justification
Domain 1: Validity	2	Algorithm not specified; no performance metrics; training data limited (N=500); no external validation
Domain 2: AI Justification	3	Clear reduction benefit (50%); appropriate for context; alternatives considered
Domain 3: Animal Welfare & 3Rs	4	Exemplary – Significant reduction in animal numbers; refinement benefits clear; replacement opportunity identified

Domain	Score	Justification
Domain 4: Harm Minimization	3	Safety mechanisms in place; human oversight defined; error contingencies basic
Domain 5: Transparency	2	Algorithm version not specified; performance metrics missing; version control undocumented
Recommendation	MODIFICATIONS REQUIRED	Require: (1) algorithm specification with version number, (2) performance metrics from training/validation, (3) updated version control plan. Strengths: Strong 3Rs alignment and safety oversight.

SCENARIO 3: Analytic - Enhanced Rubric

Protocol: "Predictive Model for College Graduation by Race/Ethnicity"

What You Receive:

- Data: 100,000 undergraduate students (2010-2020)
- Model: Gradient boosting predicting graduation likelihood
- Prediction accuracy: 78% overall
- Fairness analysis: Performance by race shows AUC 0.82 (White), 0.71 (Black), 0.75 (Hispanic)
- Mitigation: No mitigation proposed for disparities
- Use: "To inform admissions and financial aid decisions"
- Publication: Plan to publish in peer-reviewed journal

Scoring:

Domain	Score	Justification
Domain 1: Data Provenance	3	Clear data source; N=100,000 adequate; preprocessing documented; representativeness adequate
Domain 2: Privacy	3	De-identification appropriate; risk assessment conducted; security adequate
Domain 3: Analytic Validity	3	Appropriate methodology; performance metrics reported; cross-validation mentioned
Domain 4: Transparency	2	Code availability not mentioned; version control unclear; analysis decisions partially documented
Domain 5: Group Harms	1	DEFICIENT – Significant disparities identified (AUC disparity 7-11 percentage points) but NO mitigation proposed. Use in high-stakes admissions/financial aid decisions without addressing bias is unacceptable.
Recommendation	NOT ACCEPTABLE	Domain 5 critical deficiency – Group harm risk. Return for: (1) Documented bias mitigation strategy (threshold adjustment, retraining, or decision not to deploy), (2) Fairness monitoring plan, (3) Restricted use policy, (4) Community impact assessment, (5) Reconsider publication without mitigation.

8. Calibration & Consistency Tips

Tip 1: Score Within Your Disciplinary Expertise

If you're unsure about technical details, mark "Insufficient Documentation" and request clarification rather than guessing at a score.

Example: "I cannot assess the validity of the deep learning architecture without more technical documentation. Please provide: [list specific items]."

Tip 2: Use Concrete Examples During Calibration Sessions

When multiple reviewers meet, calibrate on specific protocols:

- "How would we all score this fairness domain?"
- Compare scores and discuss reasoning
- Document consensus on decision boundaries

Tip 3: Document Your Reasoning, Not Just Your Score

Your comments are critical for:

- AIRC chair review (quality check)
- Committee understanding (informed decision-making)
- PI feedback (actionable guidance)
- Institutional learning (calibration)

Weak comment: "Data quality is good. Score 3."

Strong comment: "Data quality is proficient (3). Strengths: Complete documentation of data sources, standard preprocessing. Gaps: No external validation; single-site data limits generalizability. Recommendation: Consider external validation using [specific dataset] or multicenter data."

Tip 4: Separate "Risk" from "Grade"

A "3 - Proficient" score means the protocol meets institutional standards. It doesn't mean there's zero risk.

Examples:

- High-risk protocol (vulnerable population) with "3" on fairness = meets minimum standards, but you might recommend committee apply extra scrutiny
- Low-risk protocol (validated commercial tool) with "2" = below standards for context; request modifications

Tip 5: Know When to Use "N/A" vs. "Insufficient Documentation"

N/A Example:

- Protocol uses de-identified, analyzed data with no human consent needed
- Human Subjects rubric, Privacy domain, item about "informed consent" = N/A ✓

Insufficient Documentation Example:

- Protocol is human subjects research with identifiable data, but consent form not provided
- Cannot score privacy/consent = "Insufficient Documentation" ✓

Tip 6: Use Domain Comparisons to Catch Inconsistencies

If you give high scores (3-4) across all domains for a high-risk novel protocol, ask yourself:

- "Does this seem right for such a complex, novel AI use?"
- Consider whether you should have scored more conservatively in some domains

Conversely, if all scores are 2 or below:

- "Are there genuine strengths I'm overlooking?"
- Consider whether you should score 3 on well-documented domains

9. Quick Reference: Scoring Decision Trees

Quick Decision Tree: Data Quality (Human/Analytic)

Is the data source clearly documented (origin, methods, dates)?

- No → Score 1 or 2
- Yes → Continue

Is representativeness or applicability to target population described?

- No → Score 2
- Yes, qualitatively → Score 3
- Yes, with statistical validation → Score 3-4

Are preprocessing steps documented?

- No → Score 1-2
- Basic documentation → Score 3
- Comprehensive documentation → Score 3-4

Final Score: Lowest of the sub-scores above

Quick Decision Tree: Fairness/Bias (Human) or Group Harms (Analytic)

Has performance been evaluated across demographic subgroups?

- No → Score 1 or 2
- 1-2 groups → Score 2
- 3+ groups with quantitative metrics → Score 3-4

If disparities found, is there a mitigation plan?

- No → Score 1
- Yes, vague → Score 2
- Yes, specific → Score 3-4

Is there monitoring for emergent bias?

- No → Lower score by 1
- Yes → Keep score

Final Score: Based on analysis + mitigation + monitoring

Quick Decision Tree: Transparency/Reproducibility (All Rubrics)

Is the algorithm/method fully specified?

- No (e.g., "machine learning model") → Score 1-2
- Partially (type, not version) → Score 2
- Fully (type, version, rationale) → Score 3

Are performance metrics reported?

- No → Score 1
- Basic metrics → Score 2-3
- Comprehensive metrics with validation → Score 3-4

Is code/methods documented for reproducibility?

- No → Lower score by 1
- Yes, internal only → Score 2-3
- Yes, publicly available → Score 3-4

Final Score: Based on specificity + metrics + reproducibility

Final Summary: Key Takeaways for Reviewers

1. Score based on what is documented, not your assumptions
2. Use the scoring anchors provided in the rubric; don't create your own standards
3. Score each domain independently; don't let one strong domain inflate others
4. Comment generously, especially for scores of 1-2 or 4
5. Use N/A sparingly and with justification
6. Use "Insufficient Documentation" when information is truly missing; don't score low just because details aren't provided
7. Apply critical deficiency rules consistently
8. Calibrate regularly with other reviewers to maintain consistency
9. Remember: A score of "3 - Proficient" is good; it means the protocol meets institutional standards
10. Your role is to support ethical, rigorous, fair AI research—not to be a barrier, but to ensure safeguards are in place