Medevac Interrater Reliability Study

Part 1: Statistical Analysis Plan

Medevac Interrater Study Team

2025-11-10

Table of contents

1 Analysis Plan

1.1 Overview

This document outlines the analytic framework for evaluating **interrater reliability and decision variability** among 20 physicians reviewing 20 standardized medevac vignettes. Each physician selected one of three management options:

1. **Medevac** – Immediate air evacuation
2. **Commercial** – Next available commercial flight
3. **Remain** – Stay in village for continued observation/treatment

Each physician also provided a confidence rating (1–10 scale). Because each physician rated multiple vignettes, data are structured as **repeated measures**, with responses nested within physicians.

1.2 Vignette Classes

Class	Description	N Vignettes	Example Focus
A. Expert Consensus	Clear affirmative cases with a "suggested best" expert answer	8	Clear Medevac or Clear Remain
B. Expert Boundary	Cases where one option is *discouraged* but two are acceptable	6	Clear Not Medevac / Clear Not Remain
C. Ambiguous	All options clinically reasonable	3	Any Option / Equivocal
D. Situational	Emergencies constrained by logistics (bimodal expected)	3	Conflict Between Physiology and Logistics

1.3 Vignette Mapping Table

Q#	Question Type	Class
1	Clear Medevac	A
2	Clear Not Medevac	B
3	Any Option	C
4	Clear Medevac	A
5	Clear Not Medevac	B
6	Clear Not Medevac	B
7	Clear Not Remain	B
8	Clear Remain	A
9	Clear Commercial	A
10	Clear Not Remain	B
11	Clear Remain	A
12	Any Option	C
13	Clear Medevac	A
14	Clear Not Medevac	B
15	Clear Commercial	A
16	Conflict Between Physiology/Logistics	D
17	Conflict Between Physiology/Logistics	D
18	Conflict Between Physiology/Logistics	D
19	Clear Medevac	A

Q#	Question Type	Class
20	Any Option	C

1.4 Use of Gwet's AC1 for Interrater Reliability

Traditional kappa coefficients, such as Cohen's or Fleiss', are highly sensitive to **prevalence bias**—situations where one response option dominates. Because many vignettes in this study were designed to elicit a clear "correct" or expert-preferred decision, the distribution of responses is intentionally unbalanced. Under these conditions, often yields artificially low or even negative values despite substantial agreement, a phenomenon known as the *kappa paradox*.

To overcome this limitation, we use **Gwet's AC1** as the primary measure of interrater reliability for Classes A (Expert Consensus) and B (Expert Boundary). AC1 adjusts the expected probability of chance agreement to be independent of marginal category prevalence, resulting in a more stable and interpretable estimate of agreement when response distributions are skewed.

AC1 values range from 0 to 1, with interpretation thresholds parallel to those of (0.0–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect). This approach provides a more accurate reflection of clinician alignment with expert judgment and avoids misleading reductions in reliability due to expected prevalence.

1.5 Application of Generalized Linear Mixed Models (GLMM)

Although each physician evaluates each vignette only once, the dataset contains multiple responses per physician across different scenarios. These repeated observations are **not independent**, as each physician's individual reasoning style, experience, and risk tolerance influence all of their decisions. Similarly, each vignette varies in intrinsic difficulty or clarity.

To properly account for these nested dependencies, all inferential analyses use **Generalized Linear Mixed Models (GLMMs)** with random intercepts for both *physician* and *vignette*. This approach partitions the variance into within- and between-clinician components, controlling for clustering and reducing inflated Type I error that would arise from treating responses as independent.

The random effects capture:

- **Physician-level heterogeneity** — differences in baseline propensity to choose medevac versus conservative management.
- **Vignette-level heterogeneity** — differences in clarity, context, or complexity of the clinical scenario.

GLMMs are therefore well suited to this repeated-measures design, providing unbiased estimates of fixed effects (e.g., class-level trends, alignment with expert guidance) while preserving the natural correlation structure of the data.

1.6 Class-Specific Statistical Approach

1.6.1 Class A – Expert Consensus (Suggested Best Answer)

Goal: Measure alignment with the expert-recommended option and consistency among physicians.

Descriptive Stats:

- Proportion choosing the expert answer
- Shannon entropy (decision diversity)
- Mean ± SD confidence

Analytic Statistic: Gwet's AC1 (for overall agreement)

Visualization: Binary heatmap (physicians × vignettes; green = expert aligned), bar plots of agreement proportion.

1.6.2 Class B – Expert Boundary (One Discouraged Answer)

Goal: Quantify consistency in avoiding discouraged responses.

Descriptive Stats:

- % choosing discouraged option
- Entropy and confidence scores

Analytic Statistic: Gwet's AC1 (binary acceptable vs discouraged)

Visualization: Stacked bar charts showing acceptable vs discouraged, and 3-color heatmap across vignettes.

1.6.3 Class C – Ambiguous (All Plausible Options)

Goal: Describe distributional diversity under uncertainty.

Descriptive Stats:

- Proportion per option
- Entropy (higher = more diverse)

Analytic Statistic: ^{2} goodness-of-fit test vs equal (33.3%) distribution.

Visualization: 3-color heatmap or ternary plot of decision mix.

1.6.4 Class D – Situational (Operationally Constrained Emergencies)

Goal: Describe variability (bimodal split Medevac vs Remain).

Descriptive Stats:

- Proportion choosing each option
- Entropy and confidence summary

Analytic Statistic: Exact binomial test vs 50/50 split.

Visualization: Diverging bar plot (% Medevac vs Remain) and 2-color heatmap of clinician patterns.

1.7 Sample Size and Feasibility

- 400 total responses (20×20) provide adequate power for AC1 and GLMMs.
- Class-level AC1 and entropy estimates are stable; per-vignette was dropped.
- Small cell counts (<3 raters per category) are handled via exact tests.

1.8 Standardized Analytic Framework

Step	Metric	Applies To	Purpose
1	Proportions & Entropy	All classes	Describe decision spread
2	Gwet's AC1	A,B	Stable agreement vs expert or discouraged
3	^{2} / Binomial tests	C,D	Assess distribution shape
4	GLMM (binary/multinomial)	A,B	Account for repeated measures
5	Visualizations	All	Compare across classes

Document Generated: `r Sys.Date()`