

Medevac Interrater Reliability Study

Part 2: Descriptive Statistics

Medevac Interrater Study Team

2025-11-10

Table of contents

1	Introduction	1
1.1	Vignette Classes Overview	2
2	Table 1: Vignette Class Descriptions and Summary Statistics	2
2.1	Summary by Vignette	4
2.2	Summary Statistics	6
2.2.1	Overall Summary	6
2.2.2	Summary by Vignette Class (A → D)	6
2.2.3	Visualization: Comparison Across Classes	7
3	Table 2: Interrater Reliability Metrics (Gwet's AC1)	7
3.1	Replacement of Kappa with Gwet's AC1	7
3.2	AC1 Statistics by Class	8
3.2.1	Visualization: AC1 by Vignette	9
3.2.2	Statistical Comparison of Subclasses	9

1 Introduction

This report presents the analysis of physician medevac decision-making across 20 standardized clinical vignettes following the statistical analysis plan. Each of 20 physicians evaluated all vignettes and selected one of three management options:

- **Medevac:** Immediate medical evacuation
- **Commercial:** Next available commercial flight
- **Remain:** Remain in village for observation/treatment

Each physician also provided a confidence rating (1-10 scale) for their decision. Because each physician responds to multiple vignettes, the dataset involves **repeated measures (responses clustered within physician)**.

1.1 Vignette Classes Overview

The 20 clinical vignettes are organized into four distinct classes designed to test different aspects of clinical decision-making in remote settings. See Table 1 for detailed descriptions and summary statistics for each class and subclass.

2 Table 1: Vignette Class Descriptions and Summary Statistics

This table provides a comprehensive overview of the 20 clinical vignettes organized by class and subclass, including clear descriptions of each vignette type and summary statistics across all 20 physicians.

Table 1: Vignette Class Descriptions and Summary Statistics

Class	Type	Description	N.Vignettes	Mean.Agree.	Mean.Entropy	Mean.Conf
Right Answer						
Right Answer	Overall	Clear cases where one management option is definitively correct/preferred	8	86.2	0.341	8.2
Right Answer	Medevac	Cases where immediate medevac is the right choice (e.g., unstable patient)	4	98.8	0.050	9.1
Right Answer	Commercial	Cases where commercial flight is the right choice (e.g., stable patient)	2	70.0	0.677	7.5
Right Answer	Remain	Cases where remaining in village is the right choice (e.g., minimal symptoms)	2	77.5	0.588	7.2
Wrong Answer						
Wrong Answer	Overall	Cases where one management option should definitively be avoided	6	68.3	0.665	7.8
Wrong Answer	No Medevac	Cases where medevac is inappropriate/wrong (e.g., stable, transport would worsen condition)	4	61.2	0.795	7.3
Wrong Answer	No Remain	Cases where remaining in village is inappropriate/wrong (e.g., needs urgent care)	2	82.5	0.405	8.7
Ambiguous						
Ambiguous	Overall	Cases where all three options are clinically reasonable - physician judgment critical	3	63.3	0.789	6.7
Situational						
Situational	Overall	True emergencies with logistical constraints - decisions shaped by operational factors	3	66.7	0.606	7.5

2.1 Summary by Vignette

This section provides detailed statistics for each of the 20 individual vignettes, showing their performance across all physicians before aggregating to class-level summaries.

::: .cell ::: .cell-output-display

Table 2: Summary Statistics by Individual Vignette

Vignette	Q#	Subclass	Medevac %	Commercial %	Remain %	Modal Decision	Agreement %	Entropy	Mean Conf.	
C7	A1	1	Medevac	100	0	0	Medevac	100	0.000	9.3
	A2	4	Medevac	95	0	5	Medevac	95	0.199	8.8
	A3	8	Remain	5	15	80	Remain	80	0.613	7.3
	A4	9	Commercial	0	75	25	Commercial	75	0.562	7.3
	A5	11	Remain	0	25	75	Remain	75	0.562	7.2
	A6	13	Medevac	100	0	0	Medevac	100	0.000	9.3
	A7	15	Commercial	65	30	5	Medevac	65	0.791	7.7
	A8	19	Medevac	100	0	0	Medevac	100	0.000	9.0
	B1	2	No Medevac	20	45	35	Commercial	45	1.049	7.2
	B2	5	No Medevac	0	75	25	Commercial	75	0.562	7.7
	B3	6	No Medevac	25	45	30	Commercial	45	1.067	7.0
	B4	7	No Remain	95	5	0	Medevac	95	0.199	9.4
	B5	10	No Remain	70	30	0	Medevac	70	0.611	8.0
	B6	14	No Medevac	0	20	80	Remain	80	0.500	7.4
	C1	3	C	20	0	80	Remain	80	0.500	6.9
	C2	12	C	45	40	15	Medevac	45	1.010	6.2
	C3	20	C	65	25	10	Medevac	65	0.857	7.1
	D1	16	D	75	0	25	Medevac	75	0.562	7.2
	D2	17	D	50	0	50	Medevac	50	0.693	7.0
	D3	18	D	25	0	75	Remain	75	0.562	8.3

::: :::

2.2 Summary Statistics

2.2.1 Overall Summary

Table 3: Overall Study Metrics

Metric	Value
Total Vignettes	20
Total Physicians	20
Total Responses	400
Mean Agreement %	74.5%
Median Agreement %	75%
Agreement Range	45% - 100%
Mean Confidence	7.8
Mean Entropy	0.545
Vignettes with <U+2265>75% Agreement	13
Vignettes with <50% Agreement	3

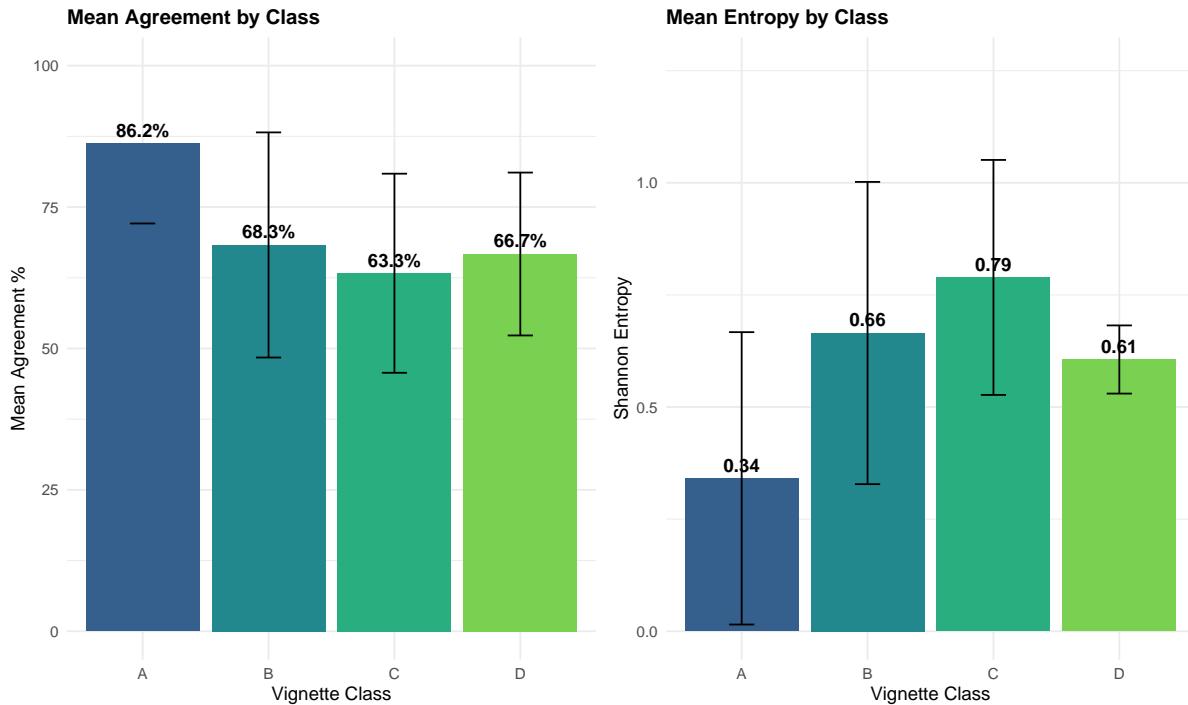
2.2.2 Summary by Vignette Class (**A → D**)

Table 4: Summary Statistics by Vignette Class and Subclass

Group	Type	N Vignettes	Mean Agree%	SD Agree	Mean Entropy	SD Entropy	Mean Conf
Class A (Right Answer)							
	A Class	8	86.2	14.1	0.341	0.326	8.2
A Subclasses							
	B Class	6	68.3	19.9	0.665	0.337	7.8
	C Class	3	63.3	17.6	0.789	0.262	6.7
	D Class	3	66.7	14.4	0.606	0.076	7.5
Class B (Wrong Answer)							
	A Commercial	2	70.0	7.1	0.677	0.162	7.5
B Subclasses							
	A Medevac	4	98.8	2.5	0.050	0.099	9.1
	B No Medevac	4	61.2	18.9	0.795	0.305	7.3
Class C (Ambiguous)							
	B No Remain	2	82.5	17.7	0.405	0.292	8.7

Class D (Situational)							
A	Remain	2	77.5	3.5	0.588	0.036	7.2

2.2.3 Visualization: Comparison Across Classes



3 Table 2: Interrater Reliability Metrics (Gwet's AC1)

3.1 Replacement of Kappa with Gwet's AC1

Because many vignettes were intentionally designed with one dominant “correct” response (e.g., clear medevac or clear remain), class prevalence is highly unbalanced. In such cases, Cohen’s and Fleiss’ produce spuriously low or even negative values despite high raw agreement—a phenomenon known as the kappa paradox.

To address this, interrater reliability was recalculated using **Gwet’s AC1**, which provides a more stable and interpretable measure under unbalanced category distributions. AC1 corrects the expected-by-chance term using an adjusted probability of agreement that is insensitive to marginal prevalence.

For this study: - **Class A (Right Answer)** and **Class B (Wrong Answer)** are analyzed using Gwet's AC1 instead of - **Classes C and D** retain their descriptive and distributional analyses (entropy, binomial/ ² tests) since agreement per se is not the central construct

The resulting AC1 values better reflect true rater alignment with expert expectations and inter-clinician consistency.

Gwet's AC1 measures agreement among multiple raters while accounting for chance agreement in unbalanced distributions.

Interpretation: - < **0.20**: Slight agreement - **0.21-0.40**: Fair agreement - **0.41-0.60**: Moderate agreement - **0.61-0.80**: Substantial agreement - **0.81-1.00**: Almost perfect agreement

Table 5: Table 2: Interrater Reliability Metrics - Gwet's AC1 (Classes A and B Only)

Vignette	Q#	Subclass	Agreement %	Entropy	Gwet's AC1	p-value	Interpretation
A1	1	Medevac	100	0.000	NA	NA	Perfect agreement
A2	4	Medevac	95	0.199	0.889	NA	Almost perfect
A3	8	Remain	80	0.613	0.576	NA	Moderate
A4	9	Commercial	75	0.562	0.368	NA	Fair
A5	11	Remain	75	0.562	0.368	NA	Fair
A6	13	Medevac	100	0.000	NA	NA	Perfect agreement
A7	15	Commercial	65	0.791	0.326	NA	Fair
A8	19	Medevac	100	0.000	NA	NA	Perfect agreement
B1	2	No Medevac	45	1.049	0.021	NA	Slight
B2	5	No Medevac	75	0.562	0.368	NA	Fair
B3	6	No Medevac	45	1.067	-0.002	NA	Slight
B4	7	No Remain	95	0.199	0.889	NA	Almost perfect
B5	10	No Remain	70	0.611	0.238	NA	Fair
B6	14	No Medevac	80	0.500	0.505	NA	Moderate

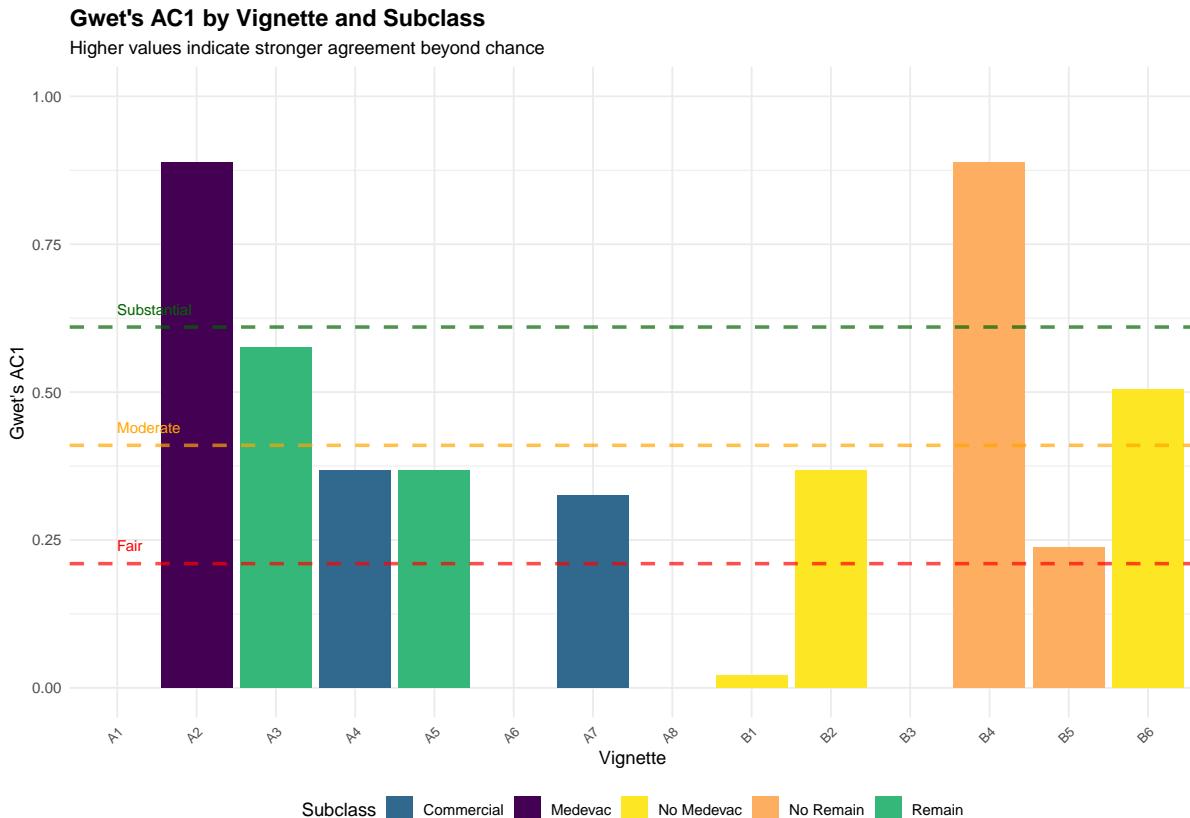
3.2 AC1 Statistics by Class

Table 6: Gwet's AC1 Statistics by Vignette Class and Subclass

Group	Type	N Vignettes	Mean AC1	Median AC1	SD AC1	Min AC1	Max AC1
Class A (Right Answer)							
A	Class	8	0.505	0.368	0.236	0.326	0.889
A Subclasses							
B	Class	6	0.336	0.303	0.334	-0.002	0.889
A	Commercial	2	0.347	0.347	0.030	0.326	0.368

A	Medevac	4	0.889	0.889	NA	0.889	0.889
Class B (Wrong Answer)							
B	No Medevac	4	0.223	0.194	0.253	-0.002	0.505
B Subclasses							
B	No Remain	2	0.564	0.564	0.460	0.238	0.889
A	Remain	2	0.472	0.472	0.147	0.368	0.576

3.2.1 Visualization: AC1 by Vignette



3.2.2 Statistical Comparison of Subclasses

Table 7: Effect Size Comparisons Between A Subclasses

Comparison	Cohen's d	Effect Size
A-Medevac vs A-Commercial	NaN	NA
A-Medevac vs A-Remain	NaN	NA

A-Commercial vs A-Remain

NaN NA
