Medevac Interrater Reliability Analysis

Right Answer vs Wrong Answer Vignette Classes

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

Q#	Question Type	Class
17	Conflict Between Physiology/Logistics	D
18	Conflict Between Physiology/Logistics	D
19	Clear Medevac	A
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

1.9 Reporting Language

> "Interrater reliability was computed using **Gwet's AC1**, which is robust to unbalanced category prevalence and provides a more accurate measure of clinician agreement for Classes A and B.
>
> For Classes C and D, decision variability was described using entropy, ^{2} and binomial tests, emphasizing the distribution of choices rather than raw agreement."

---

2 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)**.

2.1 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.

---

3 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 4: 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

3.1 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 5: Summary Statistics by Individual Vignette

Vignette	Q#	Subclass	Medevac %	Commercial %	Remain %	Modal Decision	Agreement %	Entropy	Mean Conf.
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

::: :::

3.2 Summary Statistics

3.2.1 Overall Summary

Table 6: 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**

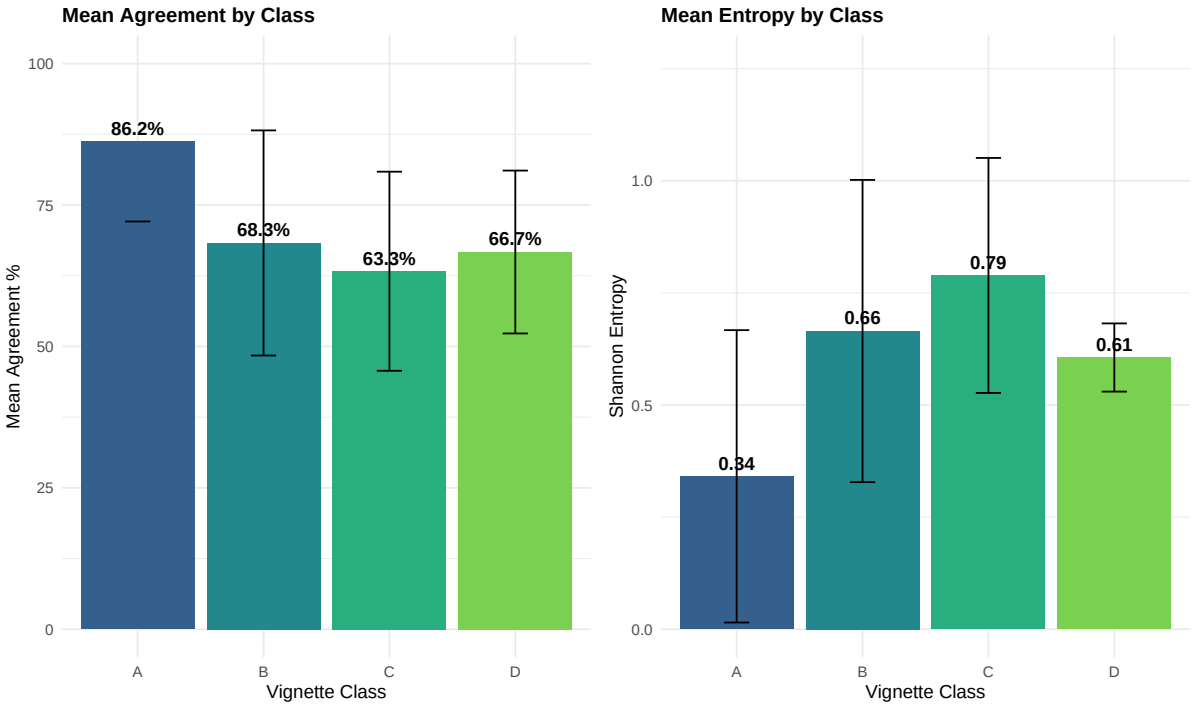
3.2.2 Summary by Vignette Class (A → D)

Table 7: 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

3.2.3 Visualization: Comparison Across Classes



---

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

4.1 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 8: 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

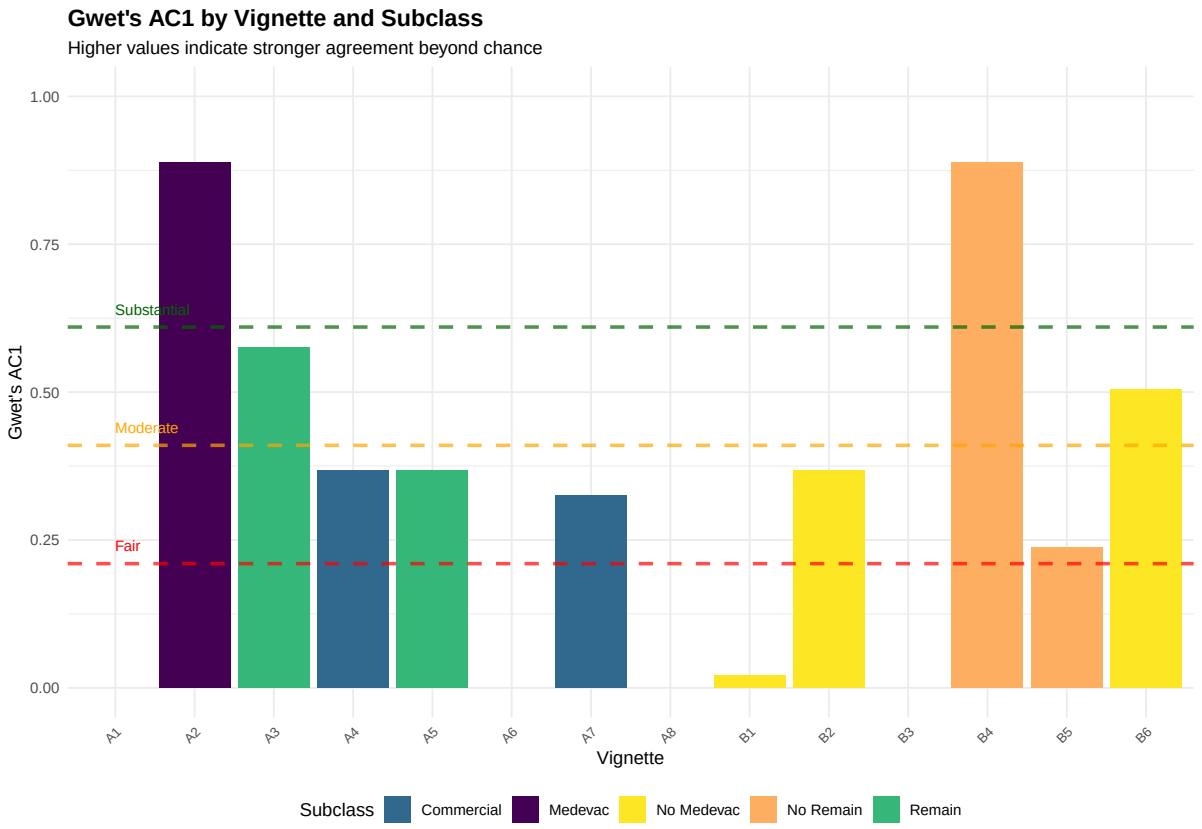
4.2 AC1 Statistics by Class

Table 9: 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

4.2.1 Visualization: AC1 by Vignette



4.2.2 Statistical Comparison of Subclasses

Table 10: 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

---

5 Vignette Class Analyses

The following sections analyze each vignette class according to its unique conceptual framework and statistical approach as outlined in the analysis plan.

---

6 Class A: Right Answer (Clear Correct Choice)

Definition: Vignettes where expert consensus suggests one preferred management option, but reasonable alternatives exist.

Objective: Measure alignment with expert guidance and degree of agreement among physicians.

Note: This analysis requires expert reference answers for each Class A vignette to be added to the dataset.

```
**Class A Summary**

- Number of vignettes: 8

- Total responses: 160

- Mean agreement: 86.2% (SD = 14.1)

- Mean entropy: 0.341 (SD = 0.326)

- Mean Gwet's AC1: 0.505
```

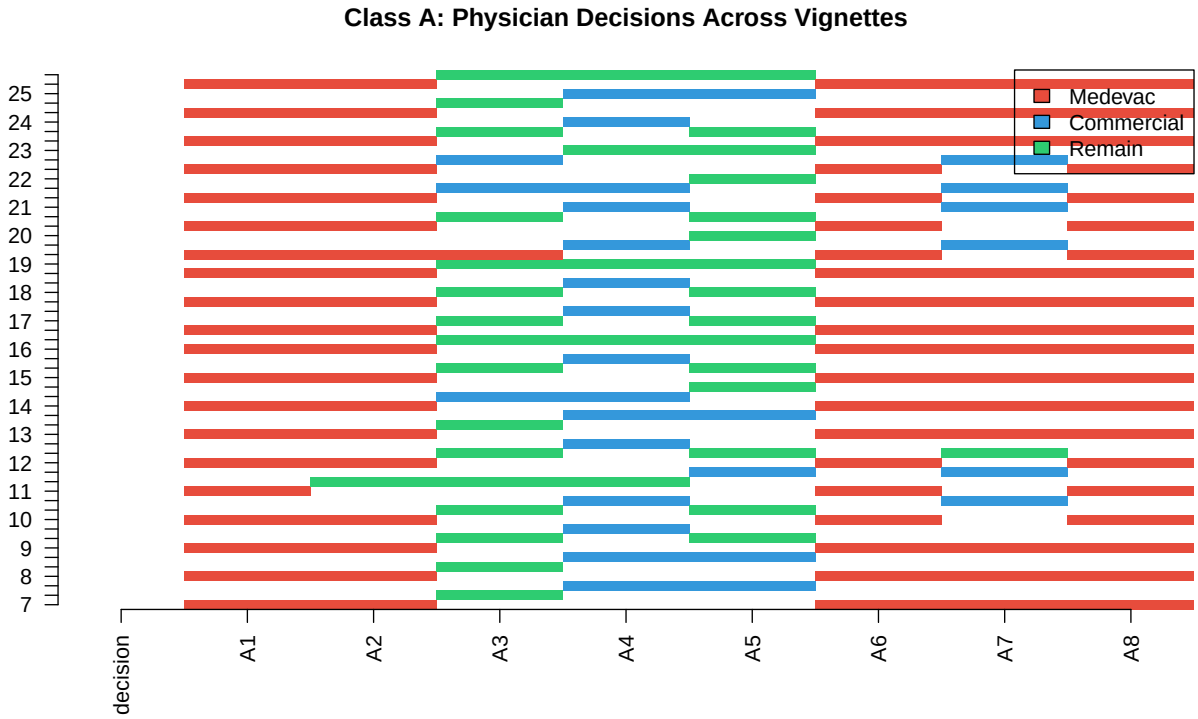
6.1 Descriptive Statistics

Table 11: Class A: Descriptive Statistics

Vignette	Q#	Question Type	Modal Decision	Agreement %	Entropy	Mean Conf.
A1	1	Clear Medevac	Medevac	100	0.000	9.3
A2	4	Clear Medevac	Medevac	95	0.199	8.8
A3	8	Clear Remain	Remain	80	0.613	7.3
A4	9	Clear Commercial	Commercial	75	0.562	7.3
A5	11	Clear Remain	Remain	75	0.562	7.2
A6	13	Clear Medevac	Medevac	100	0.000	9.3
A7	15	Clear Commercial	Medevac	65	0.791	7.7
A8	19	Clear Medevac	Medevac	100	0.000	9.0

6.2 Visualization: Decision Heatmap

```
Note: Some missing values in data, clustering disabled.
```



6.3 Mixed-Effects Logistic Regression (GLMM)

Note: To implement the full GLMM analysis as specified in the analysis plan, expert consensus decisions must be added to the dataset. The model will then estimate:

[logit(P(Consensus)) = _0 + (1|Physician) + (1|Vignette)]

---

7 Class B: Wrong Answer (Clear Incorrect Choice)

Definition: Vignettes with one clearly discouraged or clinically inappropriate option; two responses considered acceptable.

Objective: Identify how consistently clinicians avoid the discouraged response.

Note: Requires expert classification of which option is "discouraged" for each Class B vignette.

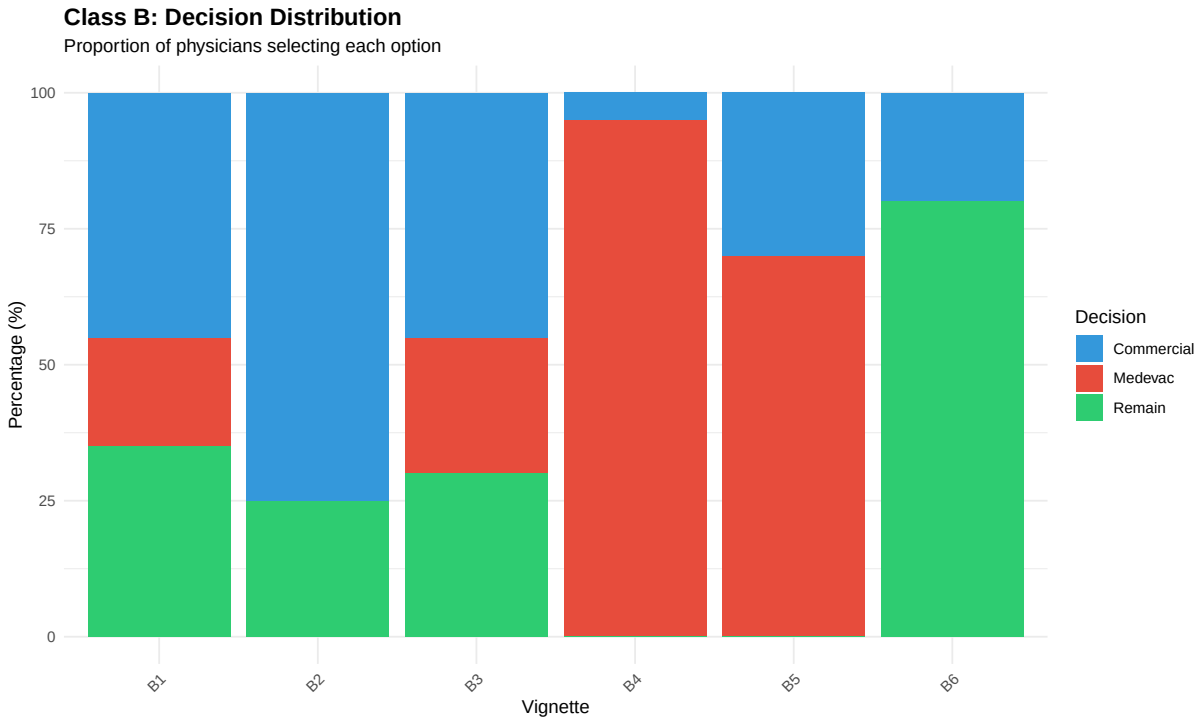
```
**Class B Summary**

- Number of vignettes: 6

- Total responses: 120

- Mean agreement: 68.3% (SD = 19.9)

- Mean entropy: 0.665 (SD = 0.337)

- Mean Gwet's AC1: 0.337
```

7.1 Descriptive Statistics

Table 12: Class B: Descriptive Statistics

Vignette	Q#	Question Type	Medevac %	Commercial %	Remain %	Entropy	Mean Conf.
B1	2	Clear Not Medevac	20	45	35	1.049	7.2
B2	5	Clear Not Medevac	0	75	25	0.562	7.7
B3	6	Clear Not Medevac	25	45	30	1.067	7.0
B4	7	Clear Not Remain	95	5	0	0.199	9.4
B5	10	Clear Not Remain	70	30	0	0.611	8.0
B6	14	Clear Not Medevac	0	20	80	0.500	7.4

7.2 Visualization: Stacked Bar Chart



7.3 Binomial Test

Testing whether the frequency of each decision differs from chance (33.3%).

Table 13: Class B: Binomial Test Results (Overall)

Decision	N Responses	Total Possible	Observed %	Expected %	p-value
Commercial	44	120	36.7	33.3	**0.4397**
Medevac	42	120	35.0	33.3	**0.6992**
Remain	34	120	28.3	33.3	**0.2867**

---

8 Class C: Ambiguous (All Plausible Options)

Definition: Vignettes intentionally designed with high uncertainty where all three options are clinically reasonable.

Objective: Characterize decision diversity and patterns under ambiguity.

```
**Class C Summary**

- Number of vignettes: 3

- Total responses: 60

- Mean agreement: 63.3% (SD = 17.6)

- Mean entropy: 0.789 (SD = 0.262) - **Higher = Greater Diversity**

- AC1 not computed (ambiguous cases - agreement not central construct)
```

8.1 Descriptive Statistics

Table 14: Class C: Descriptive Statistics

Vignette	Q#	Medevac %	Commercial %	Remain %	Entropy	Agreement %	Mean Conf.
C1	3	20	0	80	**0.500**	80	6.9
C2	12	45	40	15	**1.010**	45	6.2
C3	20	65	25	10	**0.857**	65	7.1

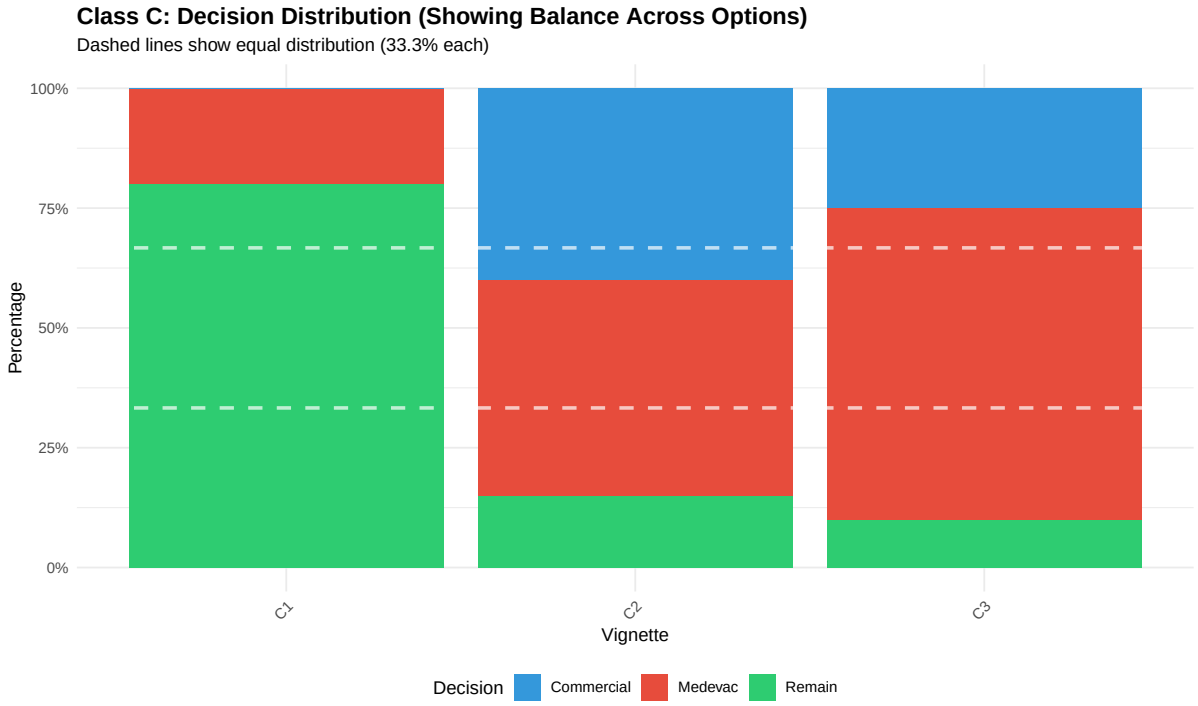
8.2 Chi-Square Goodness-of-Fit Test

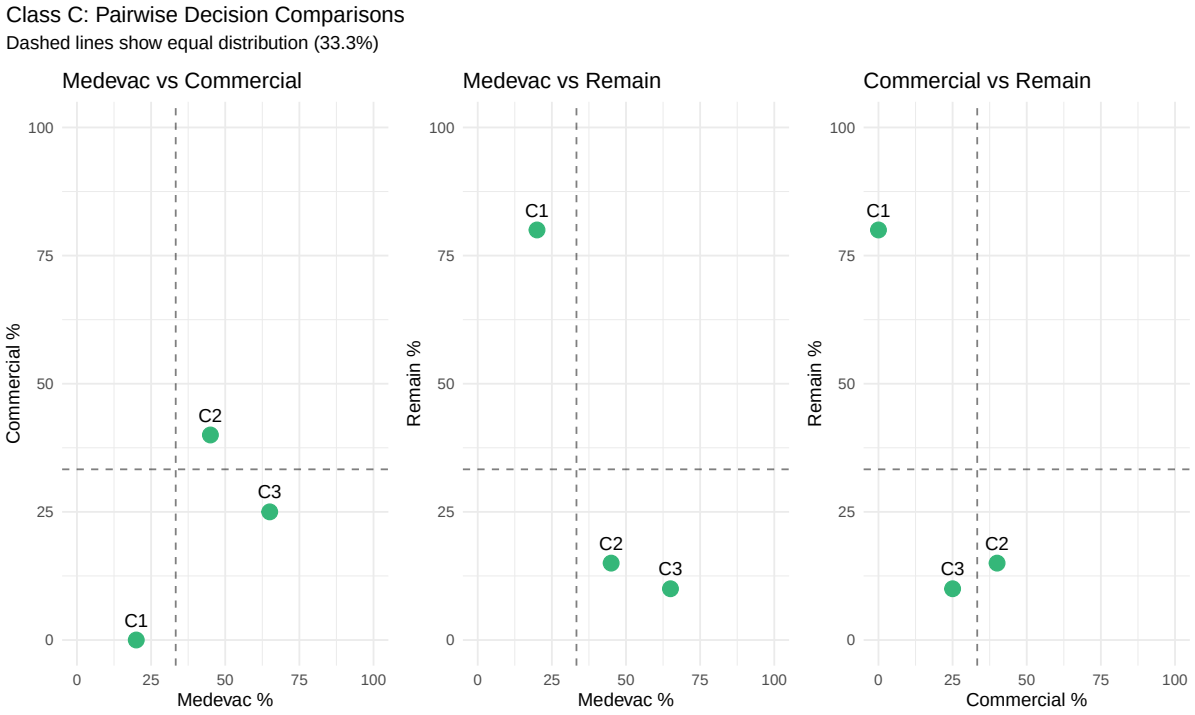
Testing for deviation from equal distribution (33.3% each option).

Table 15: Class C: Chi-Square Goodness-of-Fit Tests

Vignette	^{2} Statistic	p-value
C1	20.8	**0.0000**
C2	3.1	**0.2122**
C3	9.7	**0.0078**

8.3 Visualization: Decision Balance





---

9 Class D: Situational (Operationally Constrained Emergencies)

Definition: True emergencies where decisions are shaped by logistical constraints. Typically shows bimodal split (Medevac vs Remain).

Objective: Characterize variability under operational constraints - variation itself is meaningful.

```
**Class D Summary**

- Number of vignettes: 3

- Total responses: 60

- Mean agreement: 66.7% (SD = 14.4)

- Mean entropy: 0.606 (SD = 0.076)
```

- Mean % Medevac: 50.0%

- Mean % Remain: 50.0%

- Mean % Commercial: 0.0% (typically low)

9.1 Descriptive Statistics

Table 16: Class D: Descriptive Statistics

Vignette	Q#	Medevac %	Commercial %	Remain %	Entropy	Mean Conf.
D1	16	75	0	25	0.562	7.2
D2	17	50	0	50	0.693	7.0
D3	18	25	0	75	0.562	8.3

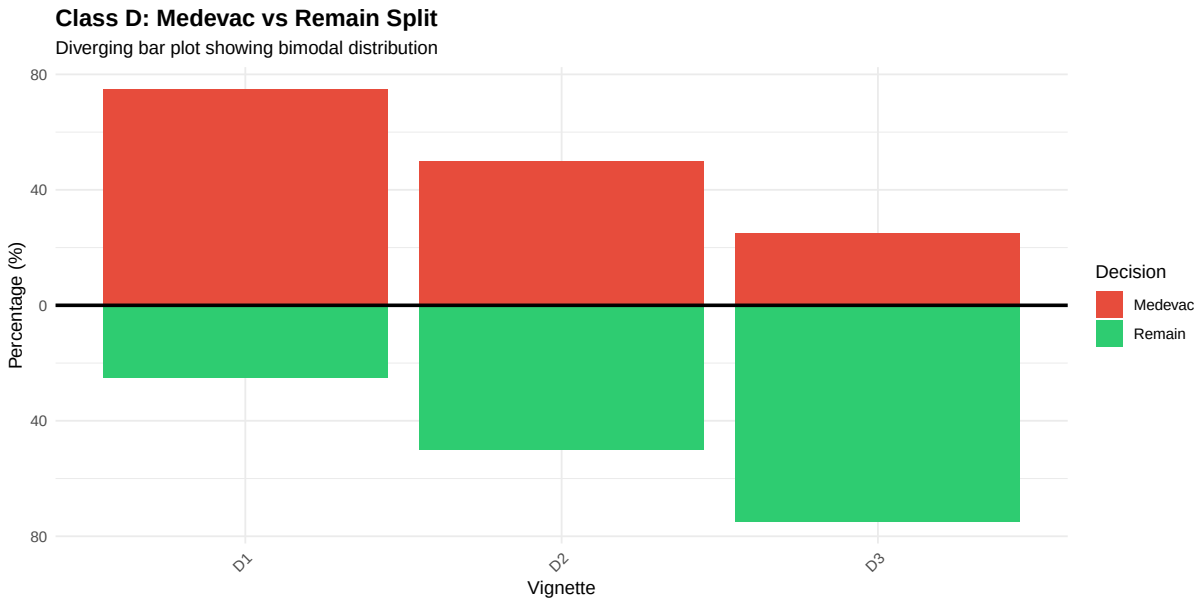
9.2 Binomial Test: Medevac vs Remain

Testing whether the Medevac/Remain split differs from 50/50.

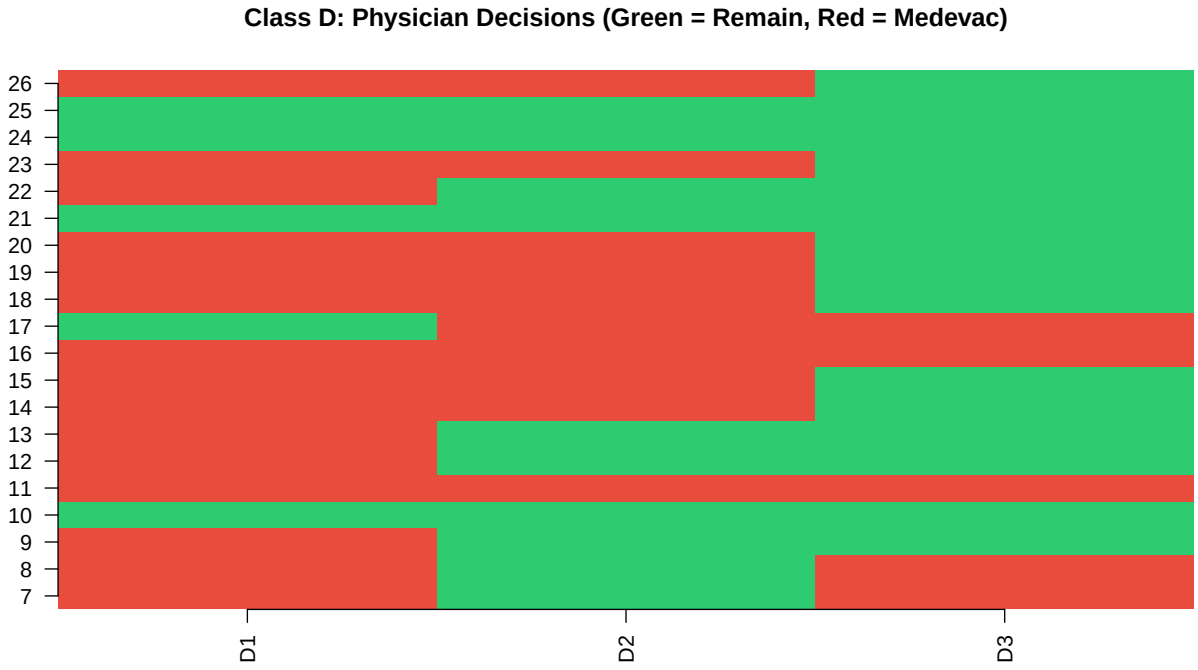
Table 17: Class D: Binomial Test Results (Medevac vs Remain)

Vignette	N Medevac	N Total	% Medevac	p-value
D1	15	20	75	**0.0414**
D2	10	20	50	**1.0000**
D3	5	20	25	**0.0414**

9.3 Visualization: Diverging Bar Plot



9.4 Visualization: Decision Heatmap



10 Summary and Conclusions

10.1 Key Findings

1. **Overall Agreement**: Mean agreement across all vignettes was 74.5%
2. **Gwet's AC1**: Mean AC1 = 0.413 (Classes A & B only)
3. **Entropy**: Mean entropy = 0.545 (higher = more diversity)

10.2 Key Findings by Subclass

- **A-Medevac** (n=0): AC1 = NaN - NA agreement
- **A-Commercial** (n=0): AC1 = NaN - NA agreement
- **A-Remain** (n=0): AC1 = NaN - NA agreement
- **B-No Medevac** (n=0): AC1 = NaN - NA agreement
- **B-No Remain** (n=0): AC1 = NaN - NA agreement

10.3 By Class

Table 18: Summary Statistics by Class

Class	N	Mean Agreement %	Mean Entropy	Mean AC1	Mean Confidence
A	8	86.2	0.341	0.505	8.2
B	6	68.3	0.665	0.336	7.8
C	3	63.3	0.789	NA	6.7
D	3	66.7	0.606	NA	7.5

10.4 Interpretation

- **Class A**: High agreement on clear affirmative cases (AC1 = 0.51)
- **Class B**: Moderate agreement on boundary cases (AC1 = 0.34)
- **Class C**: High entropy (0.79) confirms intentional ambiguity - all options plausible
- **Class D**: Lower entropy than Class C, showing bimodal split pattern

10.5 Next Steps

1. **Add expert reference answers** for Classes A and B to enable:
    - Expert consensus alignment analysis (Class A)
    - Discouraged option avoidance analysis (Class B)

- Cohen's comparisons

2. **Implement full GLMM analyses** with random intercepts for physician and vignette to account for repeated measures structure
3. **Additional visualizations** as needed for publication

Report Generated: 2025-11-10