Reliability of prehospital ultrasound in helicopter emergency transfers: cross-sectional study

DOCUMENT: SAR-2023-026-HK-v01

From: Felipe Figueiredo **To:** Hani Kuttab

2023-08-26

TABLE OF CONTENTS

Reliability of prehospital ultrasound in helicopter emergency transfers: cross-sectional study

Document version

Version	Alterations
01	Initial version

1 ABBREVIATIONS

- CI: confidence interval

2 CONTEXT

2.1 Objectives

To assess the inter-rater reliability of ultrasound measures of a mobile device used in helicopter emergency transfers.

3 METHODS

The data procedures, design and analysis methods used in this report are fully described in the annex document **SAP-2023-026-HK-v01**.

This analysis was performed using statistical software R version 4.3.0.

4 RESULTS

4.1 Prehospital ultrasound evaluation metrics

A total of 242 evaluations of ultrasound measurements were included in the analysis, divided into three metrics and evaluated independently by two experts (Table 1).

QA Score metrics were more prevalent in the 3 and 4 scores, and both examiners observed those in similar proportions. Scores 2, 3 and 4 appear to be observed in similar proportions by both raters. Score 5 HK appears to have a higher proportion in rater HK compared to rater but that score is severely under-represented in the sample and might

skew the results due to small sample effects. No observations of score 1 were available in the study sample.

Table 1 *Distribution of prehospital ultrasound evaluation metrics.*

Characteristic	**HK**, N = 242	**MV**, N = 242
QA Score, n (%)		
1	0 (0%)	0 (0%)
2	9 (3.7%)	11 (4.5%)
3	118 (49%)	138 (57%)
4	113 (47%)	92 (38%)
5	2 (0.8%)	1 (0.4%)
Interpretation, n (%)		
FN	1 (0.4%)	5 (2.1%)
FP	2 (0.8%)	3 (1.2%)
TLS	9 (3.7%)	12 (5.0%)
TN	106 (44%)	101 (42%)
TP	124 (51%)	121 (50%)
Acceptability, n (%)	234 (97%)	230 (95%)

Interpretation metric was also not sampled homogeneously, where FN. FP and TLS are under-represented in the sample and could be skewed due to small sample effects. Evaluations resulting in both TN and TP appear in similar proportions by both raters.

Acceptability metric was observed in similar proportions by both raters. It is notable that most observations were from acceptable ultrasounds, where not-acceptable had very low prevalence in the study sample.

4.2 Inter-rater reliability of prehospital ultrasound evaluation metrics

Table 2 shows how the metrics for both raters compare. All metrics had high agreement proportion, with QA Score at almost 80%, and Interpretation and Acceptability approximating 90% or above.

QA Score had moderate inter-rater reliability with kappa = 0.596 (CI: [0.50, 0.69]). This could be explained by the low representation of the extremes of the score range, where the both highest and lowest quality scores were not observed as much as the mid-range scores. the large proportion of those mid-range scores seem to offset that in the uncertainty of the estimate, given the CI is narrow, with a range of around 0.2.

Table 2 *Reliability of evaluation metrics.*

Metric	Agree (%)	Kappa	CI
QA Score	78.10	0.60	[0.50, 0.69]
Interpretation	89.26	0.81	[0.74, 0.88]
Acceptability	95.04	0.38	[0.10, 0.65]

Interpretation shows high inter-rater reliability with kappa = 0.808 (CI: [0.74, 0.88]). As see in the QA Score, the reliability of the more represented scores appear to offset the less prevalent scores.

Acceptability had the lowest reliability of the metrics under evaluation with kappa = 0.375 (CI: [0.10, 0.65]). To understand why, one needs to examine the contingency table of the cross-tabulation of scores between raters (Table 3). The mismatches between raters tend to occur in the direction of HK rating more acceptable ultrasounds when compared to rater MV. As seen in Table 1, the not-acceptable ultrasounds are under-represented in the data for a good estimate. Assuming these proportions are representative of the true prevalence of acceptable ultrasounds, the McNemar test does not indicate that raters disagree in a significant way. These two results can be jointly interpreted as there being an expectation of some disagreement in non-acceptable ultrasounds, but not large enough to disregard this metric for practical use.

Table 3 *Cross-tabulation of Acceptability between raters.*

	MV			
	0	1	**Total**	**p-value**[1]
HK, n				0.248
0	4	4	8	
1	8	226	234	
Total, n	12	230	242	

[1]McNemar's Chi-squared test

5 OBSERVATIONS AND LIMITATIONS

Recommended reporting guideline

The adoption of the EQUATOR network (http://www.equator-network.org/) reporting guidelines have seen increasing adoption by scientific journals. All observational studies are recommended to be reported following the STROBE guideline (von Elm et al, 2014).

In particular when a retrospective study is conducted using hospital records, it is recommended that the RECORD extension of the STROBE guideline is considered (Benchimol et al, 2015).

6 CONCLUSIONS

QA Score had moderate reliability between raters.

Interpretation had high reliability between raters.

Although the Acceptability shows low reliability, its measurements do not differ between raters.

7 REFERENCES

- **SAP-2023-026-HK-v01** – Analytical Plan for Reliability of prehospital ultrasound in helicopter emergency transfers: cross-sectional study

- von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Int J Surg. 2014 Dec;12(12):1495-9 (https://doi.org/10.1016/j.ijsu.2014.07.013).
- Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, Sørensen HT, von Elm E, Langan SM; RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med. 2015 Oct 6;12(10):e1001885 (https://doi.org/10.1371/journal.pmed.1001885).

8 APPENDIX

8.1 Exploratory data analysis

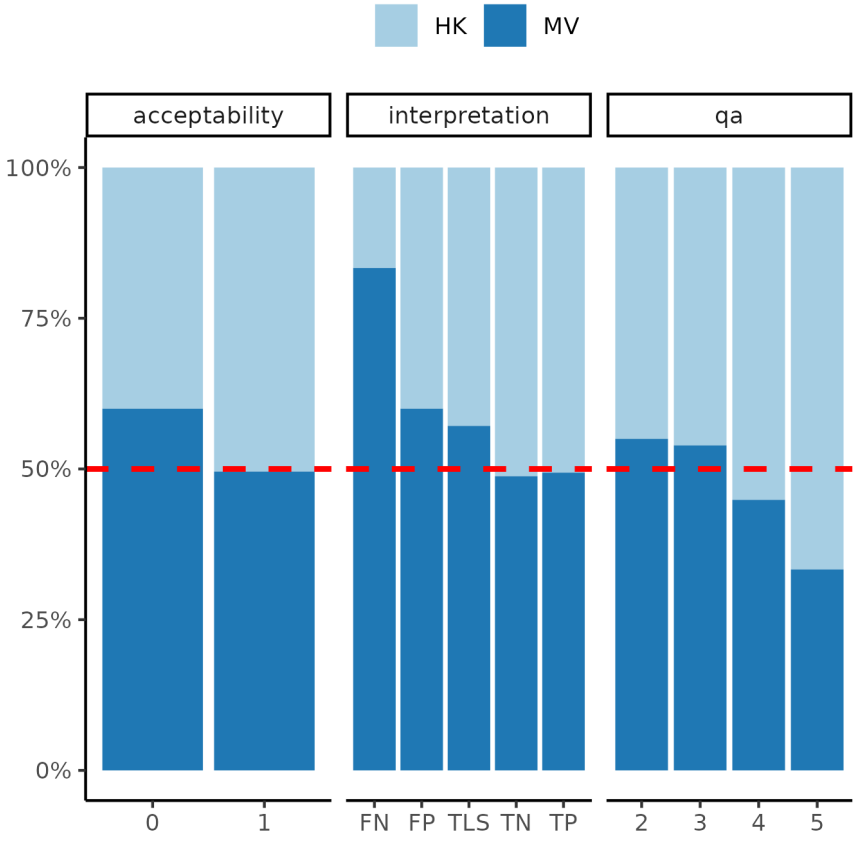

Figure A1 *Proportion of measurements per rater. Equality threshold (50%) marked with dashed line.*

8.2 Associated analyses

This analysis is part of a larger project and is supported by other analyses, linked below.

Effect of prehospital ultrasound on the time of helicopter emergency transfers: cross-sectional study

https://philsf-biostat.github.io/SAR-2023-027-HK/

8.3 Availability

All documents from this consultation were included in the consultant's Portfolio.

The portfolio is available at:

https://philsf-biostat.github.io/SAR-2023-026-HK/

8.4 Analytical dataset

Table A1 shows the structure of the analytical dataset.

Table A1 *Analytical dataset structure*

id	qa_hk	qa_mv	interpretation_hk	interpretation_mv	acceptability_hk	acceptability_mv
1						
2						
3						
…						
N						

Due to confidentiality the data-set used in this analysis cannot be shared online in the public version of this report.