Inter- and Intra-Rater Level of Agreement in Ultrasonic Video Grading of Venous Gas Emboli

Antonis Elia; Rickard Ånell; Ola Eiken; Mikael Grönkvist; Mikael Gennser

INTRODUCTION: This study aimed to evaluate whether a short familiarization session is sufficient for individuals with no prior experience of sonography to both reliably and consistently evaluate the prevalence of venous gas emboli (VGE) from precordial ultrasonic videos.

METHODS: A total of 10 adults with no prior experience of sonography were introduced to the Eftedal-Brubakk 6-grade scale and were shown 6 video sequences, each of a maximum of 10 heartbeats, representing each grading level. Thereafter, they independently evaluated the prevalence of VGE in 70 ultrasonic videos before and after a 14-d interval (test-retest; intra-rater), with these being compared to an experienced sonographer's grading (inter-rater).

RESULTS: A significant inter-rater level of agreement was found between the naïve and experienced sonographers' bubble grading both during the first (W = 0.945) and second (W = 0.952) round of bubble evaluation. The naïve observers' evaluations were on average 79% (range: 61–95%) and 75% (range: 48–95%) in complete agreement with the experienced sonographer's gradings, while the level of agreement was 99% and 98% within 1 grade unit. There was a significant intra-rater level of agreement (κ = 0.845) during the test-retest series, with a mean percentage level of agreement of 87% (range: 72–93%).

CONCLUSION: This study demonstrates that a short familiarization session enables individuals with no prior sonography experience to consistently evaluate VGE prevalence from precordial ultrasonic videos.

KEYWORDS: decompression sickness screening, aviation, diving, ultrasound operators, venous gas emboli.

Elia A, Ånell R, Eiken O, Grönkvist M, Gennser M. *Inter- and intra-rater level of agreement in ultrasonic video grading of venous gas emboli.* Aerosp Med Hum Perform. 2022; 93(1):54–57.

Pressure reductions occurring during a hypobaric (high-altitude aviation) session or following a hyperbaric exposure (diving) commonly give rise to the formation of gas emboli and, as such, incur the risk of sustaining decompression sickness.[5,7,10] Since the magnitude of gas emboli formation serves as a marker of decompression stress,[8] over the years a number of ultrasonic methods and grading systems have been developed to evaluate their prevalence.[3,4,9,10]

To date, two methods are primarily being used for assessing gas emboli in decompression research; namely, precordial imaging (e.g., ultrasonic images/videos of the cardiac four-chamber view) and ultrasound audio-Doppler [e.g., audio recordings of bubble signals (ultrasound reflections from moving blood)].[3,4] However, the latter methods' reliability has repeatedly been brought under scrutiny.[11,12] In contrast to precordial imaging where the sonographer has visual cues (e.g., full view of the four-cardiac chambers), Doppler analysis is directly dependent on the quality of the audible signal received (signal-to-noise ratio), as well as the operator's ability to differentiate between circulating gas emboli and background noise (e.g., turbulent flow, vessel walls, heart valves, etc.). Therefore, this method of assessment is by nature rather subjective and, thus, is often prone to misinterpretations.[11,12] It is, perhaps, not surprising that extensive and continuous training is required (> ~1 yr) for an individual

From the Division of Environmental Physiology, Swedish Aerospace Physiology Centre, School of Chemistry, Biotechnology, and Health, KTH, Stockholm, Sweden.

This manuscript was received for review in June 2021. It was accepted for publication in November 2021.

Address correspondence to: Antonis Elia, Ph.D., Environmental Physiology, Royal Institute of Technology, School of Engineering Sciences in Chemistry, Biotechnology and Health, Kungliga Tekniska Hogskolan, Kolan for kemi bioteknologi och halsa, Berzelius väg 13, Stockholm 1717 65, Sweden; antonise@kth.se.

DOI: https://doi.org/10.3357/AMHP.5956.2022

to be capable of independently and reliably grading audio recordings.[11,12]

Interestingly, Eftedal et al.[6] demonstrated that, following a brief introduction, inexperienced observers were able to accurately grade bubbles in ultrasonic videos. Specifically, using a 6-grade scale (0–5; see **Table I**), ~70% of their bubble grading were in complete agreement with those of experienced sonographers, while ~95% agreed within one grade unit. Yet this study did not assess the intra-rater variability, whereas Sawatzky and Nishi,[12] similarly to Eftedal et al.,[6] only examined the inter-rater variability between naïve and experienced operators using Doppler ultrasound. Thus, it is currently unclear whether inexperienced observers are able to consistently score bubbles from four-chamber heart ultrasound videos with the same grade during different occasions. Identifying whether, and to what extent, a short introductory session suffices in enabling inexperienced observers to accurately (i.e., compared with experienced sonographers; inter-rater) and consistently grade bubbles (i.e., test-retest; intra-rater) in ultrasonic videos would be of particular interest to both researchers and medical personnel involved in decompression-related fields. Duly, this study aimed to examine the inter-rater level of agreement between inexperienced and experienced sonographers as well as investigating the intra-rater variability in ultrasonic videos graded by naïve observers.

METHODS

Subjects

A total of 10 adults [7 men, 3 women; mean age of 36 (range 18–54 yr)] with no previous experience of ultrasonography nor bubble grading volunteered to take part in the experiment. The subjects gave their written, informed consent prior to enrolling and were aware that they were free to withdraw from the study at any time.

Table I. Comparing Evaluations Performed by Untrained Observers and True Values During the First (Top) and Second (Bottom) Round of Bubble Scoring.

EVALUATIONS	TRUE VALUE					
First round	**0**	**1**	**2**	**3**	**4**	**5**
0	92	14	4	1	0	0
1	4	111	50	0	0	0
2	3	17	114	5	0	0
3	0	1	19	139	6	0
4	0	0	0	20	81	4
5	0	0	0	0	12	73
Total	99	143	187	165	99	77
Level of Agreement (%)	93	78	61	84	82	95
Second round	**0**	**1**	**2**	**3**	**4**	**5**
0	94	15	8	1	0	0
1	5	97	78	3	0	0
2	0	9	95	9	0	0
3	0	0	17	148	7	0
4	0	0	0	26	68	5
5	0	0	0	0	13	72
Total	99	121	198	187	88	77
Level of Agreement (%)	95	80	48	79	77	94

Equipment and Materials

Collected were 70 4-chamber ultrasound videos (Philip CX50 Diagnostic Ultrasound System) from 2 previously conducted hypobaric experiments[1,2] which had been approved by the regional ethics review committee in Stockholm, Sweden. Ultrasonic videos were graded by an experienced sonographer (i.e., ~15 yr of ultrasonographic experience) using a 6-grade scale (**Fig. 1**),[9] with the corresponding scores serving as 'true' values.

Procedures

Prior to the evaluation of the ultrasound videos, all subjects completed a familiarization session (lasting approximately ~5-10 mins) with an experienced sonographer. During this time, subjects were introduced to the Eftedal-Brubakk (EB) grading scale and were shown 6 video sequences, each of a maximum of 10 heartbeats, together representing the 6 grading levels (Fig. 1). During this demonstration, subjects were shown which area in the image to watch (right ventricle) and what to look for [bright spots (venous gas emboli; VGE)]. Thereafter, in a randomized order and single-blinded fashion, subjects were shown 70 different video sequences and were asked to score each video independently based on the EB scale (i.e., inter-rater level of agreement). During each experiment, the EB scale was provided on paper next to the computer screen. Finally, this procedure was repeated once more after at least 14 d (range 14–20) to assess the intra-rater level of agreement.

Statistical Analyses

All data were statistically analyzed using the IBM SPSS statistics software version 21. The Kendal's coefficient of concordance W test was used to assess the level of agreement between the ultrasound operator's judgement on bubble scoring (inter-rater) and the Cohen's Kappa test was used to examine the intra-rater level of agreement. Unless otherwise stated, data are reported as means ± SD. Significance was accepted at $P < 0.05$.

RESULTS

A statistically significant level of agreement was observed among the 10 naïve observers both during the first (W = 0.945, $P < 0.001$) and second (W = 0.952, $P < 0.005$) round of bubble scoring (Table I). During both rounds of scoring, the lowest absolute level of agreement was denoted in grade 2 (Table I), with a greater inclination toward underscoring (29% first

Grade	Definition
0	No bubbles
1	Occasional bubbles
2	At least one bubble every fourth heartbeat
3	At least one bubble every heartbeat
4	At least one bubble/cm^2
5	"Whiteout", single bubbles can´t be discriminated

Fig. 1. Definitions of the image grading scale.

Table II. Comparing the Intra-rater Variability During the First and Second Round of Scoring Among the Naïve Observers.

	FIRST							
SECOND	0	1	2	3	4	5	TOTAL	LEVEL OF AGREEMENT (%)
0	93	7	1	0	0	0	101	92
1	5	131	10	0	0	0	146	90
2	0	28	94	9	0	0	131	72
3	0	1	4	143	5	0	153	93
4	0	0	0	9	76	4	89	85
5	0	0	0	0	6	73	79	92

round; 43%, second round) rather than overscoring (10% first round; 9%, second round). During the test-retest bubble scoring series, ultrasound operators presented with a statistically significant level of agreement ($\kappa = 0.845$, $P < 0.001$) (**Table II**).

DISCUSSION

The aim of the present study was to evaluate whether a short familiarization session was sufficient to enable individuals with no prior sonography experience to accurately (inter-rater) and consistently (intra-rater) evaluate VGE prevalence from precordial ultrasound videos. Our findings suggest that a short introductory session is ample to familiarize untrained observers to the EB grading scale and effectively evaluate VGE prevalence from ultrasonic videos.

The inter-rater, absolute level of agreement ranged between 48–95% across the 6-grade scale, with a statistically significant, strong level of agreement being denoted both during the first ($W = 0.945$, $P < 0.001$) and second ($W = 0.952$, $P < 0.005$) round of scoring (Table I). More specifically, on average, the complete inter-rater level of agreement was 79% and 75% during the first and second round of bubble scoring, respectively, while the naïve observers' scores agreed by 99% and 98% within 1 grade (−1 or +1) unit with the experienced sonographer's evaluations (**Fig. 2**). Our observations concur with those of Eftedal and Brubakk,[6] which likewise denoted a good, inter-rater level of agreement between naïve and experienced sonographers. Notwithstanding, in the present study we also assessed, for the first time, the intra-rater level of agreement and revealed a strong and statistically significant level of agreement ($\kappa = 0.845$, $P < 0.001$) which ranged between 72–93% (Table II). Taken together, present findings support the notion that a short introductory session is effective in: 1) introducing untrained observers to the EB grading scale, and 2) consistently evaluating the VGE prevalence from precordial ultrasound videos.

During both rounds of scoring, the least absolute level of agreement was documented in grade 2 (48% and 61%), with a tendency for the naïve observers to underscore (29% and 43%) rather than overscore (10% and 9%) (Table I). This may be explained by the nonlinear nature of the EB grading scale (Fig. 1). Specifically, the margin between grades 1–2 is rather narrow and, as such, the distinction is primarily left to the subjectivity of the observer. Considering that a low (<3) as opposed to a high bubble grade (>3) renders a lower risk of decompression sickness,[8] from a medical standpoint this margin of error is negligible. However, from a scientific perspective it bears keeping this in mind, and one should endeavor to tailor one's scientific trials such that at least one arm of the experiment will have an a priori higher bubble grade than EB 2. Therefore, in deciding whether to introduce inexperienced sonographers in the evaluation of precordial ultrasound videos, the aforementioned factor should be taken into consideration in the study design. To conclude, this study demonstrates that following a short familiarization session, untrained observers are able to consistently and effectively evaluate VGE prevalence from precordial ultrasound videos.

While precordial ultrasound imaging is less time-consuming, labor-intensive, and, as opposed to ultrasound audio-Doppler, it does not require months of extensive

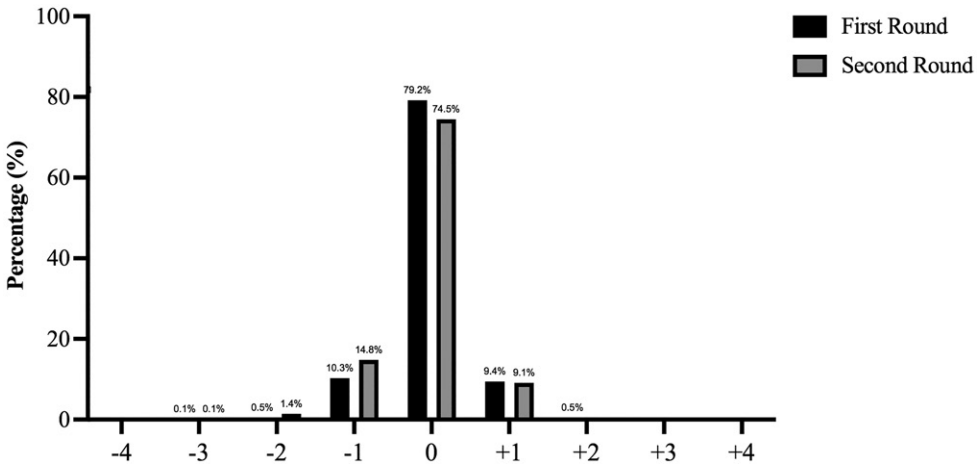

Fig. 2. Mean percentage deviation from the 'true' value in the evaluations performed by the naïve observers during the first (black filled bars) and second (gray filled bars) round of bubble scoring.

training,[11,12] there remains a certain degree of subjectivity (Fig. 1). An automated system for detection and quantification of gas bubbles in precordial imaging would, at least in theory, overcome the flaws of manual scoring and would give an objective, linear measure of detected bubbles. Accordingly, the development of such automated system(s) will be an important tool in addressing the presently reported intra- and inter-rater variability.

ACKNOWLEDGMENTS

Financial Disclosure Statement: The project was financially supported by a grant from the Swedish Armed Forces (No: 922: 0905). The authors have no competing interests to declare.

Authors and Affiliation: Antonis Elia, Ph.D., M.Sc., Ola Eiken, Prof., Ph.D., M.D., Rickard Ånell, Ph.D., M.D., Mikael Grönkvist, Ph.D., and Mikael Gennser, Ph.D., M.D., Division of Environmental Physiology, Swedish Aerospace Physiology Centre, School of Chemistry, Biotechnology and Health, Royal Institute of Technology, KTH, Stockholm, Sweden.

REFERENCES

1. Ånell R, Grönkvist M, Eiken O, Gennser M. Nitrogen washout and venous gas emboli during sustained vs. discontinuous high-altitude exposures. Aerosp Med Hum Perform. 2019; 90(6):524–530.
2. Ånell R, Grönkvist M, Gennser M, Eiken O. Evolution and preservation of venous gas emboli at alternating high and moderate altitude exposures. Aerosp Med Hum Perform. 2020; 91(1):11–17.
3. Blogg SL, Gennser M, Møllerløkken A, Brubakk AO. Ultrasound detection of vascular decompression bubbles: the influence of new technology and considerations on bubble load. Diving Hyperb Med. 2014; 44(1):35–44.
4. Brubakk AO, Eftedal O. Comparison of three different ultrasonic methods for quantification of intravascular gas bubbles. Undersea Hyperb Med. 2001; 28(3):131–136.
5. Cooper JS, Hanson KC. Decompression Sickness. StatPearls. Treasure Island (FL): StatPearls Publishing LLC; 2021.
6. Eftedal O, Brubakk AO. Agreement between trained and untrained observers in grading intravascular bubble signals in ultrasonic images. Undersea Hyperb Med. 1997; 24(4):293–299.
7. Elia A, Eiken O, Ånell R, Gronkvist M, Gennser M. Whole-body vibration preconditioning reduces the formation and delays the manifestation of high-altitude-induced venous gas emboli. Exp Physiol. 2021; 106(8): 1743–1751.
8. Gardette B. Correlation between decompression sickness and circulating bubbles in 232 divers. Undersea Biomed Res. 1979; 6(1):99–107.
9. Nishi RY, Brubakk AO, Eftedal O. Bubble detection. In: Brubakk A, Neuman T, editors. Bennett and Elliott's physiology and medicine of diving, 5th ed. Edinburgh: Saunders; 2003:501–529.
10. Pollock NW, Nishi RY. Ultrasonic detection of decompression-induced bubbles. Diving Hyperb Med. 2014; 44(1):2–3.
11. Sawatzky K. The relationship between intravascular Doppler detected gas bubbles and decompression sickness after bounce diving in humans. Toronto: York University; 1991.
12. Sawatzky KD, Nishi RY. Assessment of inter-rater agreement on the grading of intravascular bubble signals. Undersea Biomed Res. 1991; 18(5–6):373–396.