Sample size requirement for comparison of decompression outcomes using ultrasonically detected venous gas emboli (VGE): power calculations using Monte Carlo resampling from real data

David J Doolette, Keith A Gault and Christian R Gutvik

Abstract

(Doolette DJ, Gault KA, Gutvik CR. Sample size requirement for comparison of decompression outcomes using ultrasonically detected venous gas emboli (VGE): power calculations using Monte Carlo resampling from real data. *Diving and Hyperbaric Medicine*. 2014 March;44(1):14-19.)

Introduction: In studies of decompression procedures, ultrasonically detected venous gas emboli (VGE) are commonly used as a surrogate outcome if decompression sickness (DCS) is unlikely to be observed. There is substantial variability in observed VGE grades, and studies should be designed with sufficient power to detect an important effect.

Methods: Data for estimating sample size requirements for studies using VGE as an outcome is provided by a comparison of two decompression schedules that found corresponding differences in DCS incidence (3/192 [DCS/dives] vs. 10/198) and median maximum VGE grade (2 vs. 3, $P < 0.0001$, Wilcoxon test). Sixty-two subjects dived each schedule at least once, accounting for 183 and 180 man-dives on each schedule. From these data, the frequency with which 10,000 randomly resampled, paired samples of maximum VGE grade were significantly different (paired Wilcoxon test, one-sided $P \leq 0.05$ or 0.025) in the same direction as the VGE grades of the full data set were counted (estimated power). Resampling was also used to estimate power of a Bayesian method that ranks two samples based on DCS risks estimated from the VGE grades.

Results: Paired sample sizes of 50 subjects yielded about 80% power, but the power dropped to less than 50% with fewer than 30 subjects.

Conclusions: Comparisons of VGE grades that fail to find a difference between paired sample sizes of 30 or fewer must be interpreted cautiously. Studies can be considered well powered if the sample size is 50 even if only a one-grade difference in median VGE grade is of interest.

Key words

Decompression, diving, echocardiography, venous gas emboli, decompression sickness, statistics, research

Introduction

Decompression sickness (DCS) is thought to be caused by intracorporeal bubble formation. Venous bubbles (venous gas emboli, VGE) are sometimes used as an outcome in studies of decompression procedures because they can be easily detected by ultrasonic methods and graded, and because VGE grades have a general correlation with the incidence of DCS in large compilations of data.[1,2] This correlation may arise in part because VGE can cause some manifestations of DCS, but an increase in detectable VGE is also presumed to be correlated with an increase risk of bubble formation at other DCS sites. VGE grades are used to augment DCS incidence data or as a surrogate outcome if DCS is unlikely to be observed, for instance in anesthetized animals, or in studies of low-risk human procedures.

VGE occur commonly without DCS (which is rare); therefore, VGE data are potentially more information-rich than low-incidence DCS data. This additional information is counterbalanced by the facts that, owing to poor specificity, VGE grades have poor diagnostic value for DCS, and there is substantial inter- and intra-individual variability in VGE grades observed following identical exposures.[3–6] These latter facts impose a lower limit on sample size for studies of low-risk human procedures that use VGE as a surrogate outcome measure.

A common design of such studies is for two different procedures to be performed on separate occasions by the same subjects, and to test for a difference in VGE outcome using a paired statistical test such as the Wilcoxon signed-rank test. The power of a statistical test to detect a particular effect size at a particular statistical significance criterion (α) depends on the sample size, so power calculations may be used when designing an experiment to select an appropriate sample size. This study provides estimates of power for various sample sizes for human studies that use paired comparisons of VGE grades following decompression.

Methods

Monte Carlo experiments analyze outcomes in multiple computer-generated random samples. For instance, the probability of an outcome is estimated by the proportion of samples in which the outcome occurs. Monte Carlo experiments can be used to examine the properties of statistical hypothesis tests, for instance, the probability of rejecting a false null hypothesis (power) for a test procedure which produces a *P*-value and then rejects the null hypothesis if the *P*-value is less than or equal to a particular α-level. Monte Carlo estimation of the power involves computing the proportion of rejections in many random samples. Typically the random samples would be simulations generated from

parametric distributions and, in the case of a two-sample test, hypothetical effect sizes. However, in this report, samples were generated by resampling subsets of real data.

DATA

A recently published, large-scale comparison of two air decompression schedules provides unique data for estimating sample size requirements, finding corresponding statistically significant differences in DCS incidence and median peak VGE grade.[7] Eighty-one US Navy divers participated in a total of 390 man-dives, performing work during 30 minutes' bottom time at 622 kPa absolute (170 feet of sea water gauge, fsw). They were at rest and cold during either of two decompression schedules that differed only in the distribution of 174 minutes' total decompression stop time among stop depths: a shallow stop (A1) schedule and a deep stop (A2) schedule. The study reached an early stopping criterion at midpoint analysis, which found a lower incidence of DCS on the A1 than the A2 schedule at one-sided $\alpha = 0.05$ (an early 'opposite tail' finding relative to a final result that would have motivated changing US Navy procedures). DCS was diagnosed by the duty diving medical officer and full descriptions are given in the original report. During re-evaluation of the cases according to the criteria described in Temple et al,[8] one case with symptom onset 27 hours after surfacing from the A2 schedule was re-classified as not DCS. This resulted in 3/192 (DCS/dives) and 10/198 ($P = 0.0489$, one-sided Fisher's exact test), on the A1 and A2 schedules, respectively.

As a secondary outcome measure, subjects were monitored for VGE with trans-thoracic cardiac 2-D echo imaging at 30 minutes and two hours post dive. While the subjects reclined with left side down, the four heart chambers were imaged with the subject at rest and then, in turn, while they flexed each elbow and knee. VGE were graded according to the Table 1 scale, adapted from Eftedal and Brubbak.[9] The same ultrasound technician conducted all the examinations and all observed VGE grades are documented elsewhere.[7] However, in this report, only the maximum VGE grades observed at any time (rest or limb flexion, any examination) after each dive were used and will be referred to as 'VGE grade' without qualification. The median VGE grades were 2 and 3 (two-sided $P < 0.0001$, Wilcoxon rank sum test), on the A1 and A2 schedules, respectively. VGE data were missing for three man-dives: two subjects were recompressed to treat DCS before VGE examination, and results for a subject without symptoms were inadvertently not recorded. In each case, the same subject undertook the same schedule (for which data was missing) and had VGE recorded, on at least one other occasion.

The original study was not designed as a paired comparison, but of the 81 subjects who participated in the original trial, 62 dived each schedule at least once. The VGE outcome of all dives undertaken by these 62 subjects was designated the paired data set and was used to generate random samples of paired data (VGE grade after A1 and A2 schedules in the same subject). The paired data set contained 363 records, each representing one man-dive, and each comprised of a subject identifier, a schedule identifier, and the VGE grade. The distribution of VGE grades in the paired data set is given in Table 2. Median VGE grade was 2 (interquartile range [IQR] 1–3) following the A1 schedule and 3 (IQR 2–4) follow the A2 schedule. These VGE grades were significantly different (Wilcoxon rank sum test, two-sided $P < 0.0001$), and A1 less than A2 will be considered as the true outcome for power estimation. Many subjects dived the A1 and A2 schedules more than once. The mean number of dives per subject on the A1 schedule was 3 (range 1–9) accounting for a total of 183 man dives. The mean number of dives per subject on the A2 schedule was 3 (range 1–8) accounting for a total of 180 man-dives. There was no requirement in the original study for subjects to dive A1 and A2 schedules an equal number of times; however, the differences between the number of A1 and A2 schedules undertaken by each subject were relatively symmetrically distributed around zero with the absolute value of the difference/number of subjects: 0/25; 1/20; 2/12; 3/3; 4/2. Subjects refrained from any hyperbaric or hypobaric exposure for three days prior to any of the dives in the paired data set and the most common interval between these dives was seven days.

Table 1
Venous gas embolism grading (modified from reference 9)

Grade	Description
0	No bubble seen
1	Rare (< 1/s) bubble seen
2	Several discrete bubbles visible per image
3	Multiple bubbles visible per image but not obscuring image
4	Bubbles dominate image, may blur chamber outlines

RESAMPLING

For each of a range of paired sample sizes (n = 10 to 60 subjects), Monte Carlo resampling and testing of paired VGE grades was performed in the following manner. First, a subset of n subjects was randomly selected without replacement from a vector containing the 62 subject identifiers. Second, for each subject in this subset, one VGE grade was randomly selected from among the A1 schedules and one from among the A2 schedules that subject had completed. The resulting subset contained an A1-A2 pair of VGE grades for n different subjects. VGE grades from different subjects were considered independent and the resampling scheme took advantage of subjects who dived a schedule more than once by allowing different A1-A2 pairs for that subject in different subsets (there are more than 10^{41} possible such combinations in the paired data set for each value of n). Finally, for each

Table 2
VGE grades, paired data set (taken from reference 7)

	0	1	2	3	4
A1	27	36	36	53	31
A2	9	25	29	45	72

subset, the *P*-value of a paired Wilcoxon signed-rank test, with alternative hypothesis A1 less than A2 (in accord with the true outcome) was recorded. This three-step procedure was repeated 10,000 times for each value of *n*. The frequency with which *P*-values from the 10,000 subsets were less than or equal to a particular α-level provides an estimate of the probability of an α-level test on sample size of *n* subjects detecting the true one-grade difference in VGE in the paired data set (power). Power estimates are given for one sided $\alpha = 0.05$ because this level was an early stopping criterion for difference in DCS incidence in the original study that generated the data set, and for one-sided $\alpha = 0.025$ because this level is equivalent to two-sided $\alpha = 0.05$ that would commonly be used for comparisons where there is no justification for a one-sided test.

Within-subject variability in VGE grade for the same schedule was considered to be random since dives were sufficiently spaced so as not to influence each other either in terms of residual nitrogen or acclimatization. This assumption was not a requirement of the nonparametric statistical analysis. Some variability may result from measurement precision and, in particular, VGE measurements in the original study were infrequent (30 and 120 min post dive) and may not have consistently captured the peak VGE grade that occurred after each dive. To examine the consequence of possible frequent failure to record the peak VGE grade, a modified data set was drawn from the paired data set. The modified data set comprised only the maximum VGE grade observed among each repetition of the A1 schedule and each repetition of the A2 schedule for each of the 62 subjects (no intra-individual variability). The modified data set had median VGE grades of 3 (IQR 2.25–4) following the A1 schedule and 4 (IQR 3–4) following the A2 schedule (paired Wilcoxon signed-rank test, two-sided $P = 0.0056$). For each of a range of paired sample sizes ($n = 10$ to 50), a subset of *n* A1-A2 pairs of VGE grades was randomly selected without replacement from the 62 in the modified data set and tested with a paired Wilcoxon signed-rank test, with alternative hypothesis A1 less than A2. This resampling procedure was repeated 10,000 times and the power estimated as described for the paired data set. There are more than 10^{12} combinations of 50 from 62 subjects, but only 1,891 combinations of 60 from 62 subjects, so estimating power for $n = 60$ subjects by resampling from the modified data set was not considered meaningful.

Recently, a Bayesian method has been proposed to estimate the probability of DCS of a decompression procedure from maximum observed VGE grades and test for a difference in risk between two procedures.[10] We estimated the power of this latter test for comparison with the Wilcoxon test. Briefly, the method constructs posterior distributions of the probability of DCS given VGE grade (for instance based on the data given by Sawatzky[1]) and the probability of VGE grade given the test procedure, and then the total probability of DCS of a procedure is estimated by Monte Carlo simulation from these posteriors. Two procedures are tested for a difference in DCS risk by counting the frequency with which one procedure is estimated as riskier than the other (estimated confidence of the difference) in parallel Monte Carlo simulations. Using the same prior distributions as originally described[10] to produce posterior distributions from the present paired data set resulted in an estimated 99.98% confidence that the A2 schedule was riskier than the A1 schedule. Again using the same prior distributions, posterior distributions were produced from resampled subsets of the present paired data set. For each resampled subset, the confidence that the A2 schedule was riskier than the A1 schedule (in accord with the true outcome of both the Bayesian and Wilcoxon tests) was estimated. The frequency with which this confidence was greater than 95% in resampled subsets is comparable (but not identical) to the power estimate for the Wilcoxon rank sum test at one-sided $\alpha = 0.05$. Only sample sizes $n = 20$ and $n = 50$ were examined, and resampled 500 times, because the Bayesian method itself requires Monte Carlo simulations and is highly computing intensive.

Table 3
Power estimated from frequency of observed *P*-values of Wilcoxon test, paired data set

	Number of subjects					
	10	**20**	**30**	**40**	**50**	**60**
Power						
one-sided $P \leq 0.05$	0.27	0.48	0.65	0.78	0.88	0.94
one-sided $P \leq 0.025$	0.15	0.34	0.50	0.66	0.78	0.87

Data analysis was performed using R version 2.14.2 (Vienna, Austria: R Development Core Team; 2012) and MATLAB version 7.8.0.347 (R2009a) (Natwick, MA: The MathWorks Inc; 2009).

Results

Table 3 shows the power for various sample sizes for the Wilcoxon rank sum test, estimated by resampling from the paired data. These values are the probabilities of a significant test ($P \leq 0.05$ and $P \leq 0.025$) in accord with the true outcome. The fraction of results not in accord with the true outcome were usually failure to find a difference between A1 and A2 VGE grades (type II error) – the opposite tail finding of higher VGE grades on A1 than A2 was extremely rare, the highest frequency of this result was 0.0016 for $n = 10$ and $P \leq 0.05$, and otherwise zero. The choice of power depends on the consequences of making a type II error, but

Table 4
Power estimated from frequency of observed *P*-values of Wilcoxon test, modified data set

	Number of subjects				
	10	**20**	**30**	**40**	**50**
Power					
one-sided $P \leq 0.05$	0.22	0.39	0.56	0.75	0.95
one-sided $P \leq 0.025$	0.11	0.24	0.37	0.55	0.78

one convention is to design experiments with two-sided $\alpha = 0.05$ and 80% power. From the one-sided $P \leq 0.025$ row (equivalent to two-sided $\alpha = 0.05$) in Table 3, it can be seen that VGE grades from a paired sample size of about $n = 50$ subjects would have 80% power to detect a difference of one VGE grade. Power dropped quickly with sample size so that at $n = 30$ subjects ($P \leq 0.025$) there was equal probability of a true answer and a type II error.

Table 4 shows the power for various sample sizes for the Wilcoxon rank sum test, estimated by resampling from the modified data comprising only the highest VGE scores from repeated dives on the same schedules. Although there are some differences from the results of the paired data set, a sample size of about $n = 50$ is required for 80% power at two-sided $\alpha = 0.05$.

Power estimates for the Bayesian test were similar to those of the Wilcoxon rank sum test. The frequency of predicting the A2 schedule to be riskier than A1 schedule with 95% confidence was 0.40 for $n = 20$ resampled subsets and 0.80 for $n = 50$ resampled subsets. These power estimates are comparable to the values for these sample sizes in the $P \leq 0.05$ row of Table 3. The opposite tail finding (A1 riskier than A2 with 95% confidence) never occurred.

Discussion

Statistical power (or sensitivity) is the probability of rejecting a false null hypothesis (not making a type II error). In the current context, this is the probability of finding a difference (rejecting the null hypothesis of no difference) between paired samples of VGE grades for each schedule given that the VGE grades are different for each schedule in the population. The power of a statistical test depends on the magnitude of the effect to be detected, the α-value of the test, and the sample size. Power calculations are used to select appropriate sample sizes when designing experiments and Table 3 provides guidelines for designing paired comparisons using VGE as an outcome. For instance, a paired sample size of about 50 subjects is required for 80% power to detect a one-grade difference in median VGE at one-sided $\alpha = 0.025$ (equivalent to two-sided $\alpha = 0.05$) in this relatively homogenous group of subjects diving under rigidly controlled conditions.

The present results are only relevant to a one-grade difference in VGE. For instance, analysis of a simulated data set with a two-grade difference in median VGE (not shown) found a paired sample size of about 20 was required for 80% power to detect the difference at two-sided $\alpha = 0.05$. Nevertheless, the present guidelines are broadly applicable for two reasons: one VGE grade is the precision that is common across the most frequently used grading systems and many published studies report one-grade or less difference in VGE. With respect to grading precision, the present VGE grading system was a modification of the Eftedal-Brubakk system for grading VGE in 2D echocardiographic images, and the Eftedal-Brubakk grading system is broadly similar to the Spenser and Kisman-Masurel systems for aural grading of VGE detected by ultrasonic Doppler shift, in that they all grade human VGE data on an approximately equivalent zero to four ordinal scale (although the Kisman-Masurel system reports "+" and "–" intergrades and the Eftedal-Brubakk system has a grade 5 which has not been reported in humans).[2,9,11] Sample size guidelines based on the minimum measurable difference in peak VGE grade (e.g., Table 3) are useful if there is no reason to expect or require a greater difference.

The estimated power to detect a one-grade difference in median VGE is relevant to many published studies. A Medline search for the 10 years up to 2012 identified 23 publications that were paired comparisons of VGE following diving (68% of all publications found concerning VGE and diving in humans in this period). Of these, 16 reported the individual or summary statistics of the observed VGE grades (Eftedal-Brubakk, Spencer or Kisman-Masurel systems).[12–27] Only three of these 16 papers reported more than a one-grade difference in median VGE.[17,24,27] Sample sizes in these studies ranged from 6 to 28 subjects and only four of these papers reported a significant difference in VGE grades. Four papers reported no significant difference in VGE grades, and eight reported significant difference in transformations of the data. The most common transformations were to bubble count·cm^{-2} and to the Kisman-Masurel integrated severity score.[2,5] Bubble count·cm^{-2}, if a transformation from peak VGE grades (i.e., not measured directly), is subject to the same power constraints as the underlying VGE grades. The current power calculations are not applicable to the Kisman-Masurel integrated severity score which includes additional time-course information. If the Kisman-Masurel integrated severity score were demonstrated to have a stronger correlation with DCS incidence than has maximum VGE grades, sample size guidelines would be useful, but the present data did not include sufficiently frequent VGE measurements to calculate a meaningful score.

Power estimates are dependent on the precision of measurement. A limitation of the present estimates is that the paired data set may have unnecessary variance because infrequent measurements of VGE may not have always captured the true peak VGE grade. Any such aliasing may

not have been severe because the two VGE examinations (at 30 and 120 minutes) span the period during which peak VGE are typically recorded following bounce dives and VGE grades were similar at these two examinations.[28] There was no difference in VGE grades between examination times following the A1 schedule; however, there was a significant difference in VGE grades between examinations following the A2 schedule (Wilcoxon rank sum test two-sided, $P = 0.0006$) but the estimated location shift was only one-half a VGE grade. Also, the modified data set, which had no intra-individual variability in VGE scores, produced similar power estimates to those extracted from the paired data set.

The concordance between VGE grades and DCS incidence in the present data is of interest since VGE grades are often used as a surrogate for DCS (although not in the original study). The dives in the present data set were relatively risky air decompression dives; for instance, in the US Navy Diving Manual, an air dive to 170 fsw for 30 minute bottom time requires the use of oxygen decompression, and the two air schedules had a measurable difference in DCS incidences.[29] The original study planned 375 man-dives on each schedule, which would have had approximately 80% power to detect the actually observed difference in DCS incidences (a difference which was larger than expected) at two-sided $\alpha = 0.05$. This is compared with a paired sample size of about $n = 50$ subjects to detect the observed one-grade difference in median VGE at the same power and significance. While this comparison is interesting in hindsight, the objective of the original comparison of decompression procedures was to discern any practical difference in the DCS incidence, not VGE grades per se.

The concordance of differences in VGE grades and differences in DCS risk (estimated from observed DCS incidence) in the present data will not necessarily hold for all experiments. In the largest compilation of VGE and DCS incidence following diving, there was no DCS associated with Kisman-Masurel grade 0 (0 DCS/819 dives) and DCS incidence was indistinguishable between grades I (3 DCS/287 dives) and II (2 DCS/183 dives) or between grades III (27 DCS/365 dives) and IV (9 DCS/72 dives), although the DCS incidence does differ between these low and high VGE grades.[1] Therefore, an experiment that demonstrates a statistically significant difference between, for instance, median VGE grades I and II using a Wilcoxon signed rank test, may not reflect a demonstrable difference in DCS risk. Misinterpretation is less likely with the Bayesian method of Eftedal and colleagues.[10] This Bayesian method compares estimates of the probability of DCS derived from information about the distribution of DCS incidence with VGE grades, in this case a prior distribution from the data compilation noted above.[1] Because the Bayesian method incorporates this prior, it is unlikely to find a difference between a sample dominated by VGE grade I and a sample dominated by VGE grade II, unless there is also substantial difference in the distribution of other VGE grades between the samples. Conversely, any analysis of VGE may fail to identify a true difference in DCS risk between two samples dominated by grade IV VGE, since this is the highest grade observable, irrespective of DCS risk. The similarity of power and sample size estimates between the Wilcoxon and Bayesian test on the present data arises because the median VGE grades on the A1 and A2 schedule were 2 and 3 (equivalent to Kisman-Masurel grades II and III), respectively, and there is a significant difference in DCS incidence between these grades in the prior distribution.

Conclusions

Comparisons of two decompression procedures using only VGE as an endpoint that fail to find a difference between paired sample sizes of 30 or fewer must be interpreted cautiously. Studies can be considered well powered if the sample size is above 50 even if only a one-grade difference in median VGE is of interest. Maximum VGE grades can provide more power than DCS incidence to distinguish between two decompression procedures; however, a difference in VGE grades does not necessarily reflect a difference in DCS risk. If the purpose of the study is to infer a difference in DCS risk from VGE grades alone, VGE data must be interpreted cautiously, and the Bayesian method incorporating appropriate prior information about the distribution of DCS incidence with VGE grades is preferred over simple statistical tests such as the Wilcoxon signed-rank test.

References

1 Sawatzky KD. *The relationship between intravascular Doppler-detected gas bubbles and decompression sickness after bounce diving in humans* [MSc Thesis]. Toronto, ON: (Canada): York University; 1991.

2 Nishi RY, Brubakk AO, Eftedal OS. Bubble detection. In: Brubakk AO, Neuman,TS, editors. *Bennett and Elliott's physiology and medicine of diving*. 5th ed. Edinburgh: Saunders; 2003. p. 501-29.

3 Kumar VK, Billica RD, Waligora JM. Utility of Doppler-detectable microbubbles in the diagnosis and treatment of decompression sickness. *Aviat Space Environ Med*. 1997;68:151-8.

4 Gerth WA, Ruterbusch VL, Long ET. *The influence of thermal exposure on diver susceptibility to decompression sickness.* Technical Report. Panama City, FL: Navy Experimental Diving Unit; 2007 Nov. Report No.: NEDU TR 06-07. Available at http://archive.rubicon-foundation.org/xmlui/handle/123456789/5063.

5 Nishi RY, Kisman KE, Eatock BC, Buckingham IP, Masurel G. Assessment of decompression profiles and divers by Doppler ultrasonic monitoring. In: Bachrach AJ, Matzen MM, editors. *Underwater physiology VII. Proceedings of the 7th Symposium on Underwater Physiology*. Bethesda, MD: Undersea Medical Society; 1981. p. 717-27.

6 Eckenhoff RG, Hughes JS. Acclimatization to decompression stress. In: Bachrach AJ, Matzen MM, editors. *Underwater physiology VIII. Proceedings of the 8th Symposium on Underwater Physiology*. Bethesda, MD: Undersea Medical Society; 1984. p. 93-100.

7 Doolette DJ, Gerth WA, Gault KA. *Redistribution of decompression stop time from shallow to deep stops increases incidence of decompression sickness in air decompression dives.* Technical Report. Panama City, FL: Navy Experimental Diving Unit; 2011 Jul. Report No.: NEDU TR 11-06. Available at http://archive.rubicon-foundation.org/xmlui/handle/123456789/10269.
8 Temple DJ, Ball R, Weathersby PK, Parker EC, Survanshi SS. *The dive profiles and manifestations of decompression sickness cases after air and nitrogen-oxygen dives*. Technical Report. Bethesda, MD: Naval Medical Research Center; 1999. Vol 1. Report No.: 99-02. Available at http://archive.rubicon-foundation.org/xmlui/handle/123456789/4975.
9 Eftedal O, Brubakk AO. Agreement between trained and untrained observers in grading intravascular bubble signals in ultrasonic images. *Undersea Hyperb Med*. 1997;24:293-9.
10 Eftedal OS, Tjelmeland H, Brubakk AO. Validation of decompression procedures based on detection of venous gas bubbles: a Bayesian approach. *Aviat Space Environ Med*. 2007;78:94-9.
11 Brubakk AO, Eftedal O. Comparison of three different ultrasonic methods for quantification of intravascular gas bubbles. *Undersea Hyperb Med*. 2001;28:131-6.
12 Dujic Z, Duplancic D, Marinovic-Terzic I, Bakovic D, Ivancev V, Valic Z, et al. Aerobic exercise before diving reduces venous gas bubble formation in humans. *J Physiol*. 2004;555:637-42.
13 Marroni A, Bennett PB, Cronje FJ, Cali-Corleo R, Germonpre P, Pieri M, et al. A deep stop during decompression from 82 fsw (25 m) significantly reduces bubbles and fast tissue gas tensions. *Undersea Hyperb Med*. 2004;31:233-43.
14 Blatteau JE, Gempp E, Galland FM, Pontier JM, Sainty JM, Robinet C. Aerobic exercise 2 hours before a dive to 30 msw decreases bubble formation after decompression. *Aviat Space Environ Med*. 2005;76:666-9.
15 Dujic Z, Palada I, Valic Z, Duplancic D, Obad A, Wisloff U, et al. Exogenous nitric oxide and bubble formation in divers. *Med Sci Sports Exerc*. 2006;38:1432-5.
16 Blatteau JE, Boussuges A, Gempp E, Pontier JM, Castagna O, Robinet C, et al. Haemodynamic changes induced by submaximal exercise before a dive and its consequences on bubble formation. *Br J Sports Med*. 2007;41:375-9.
17 Blatteau JE, Pontier JM. Effect of in-water recompression with oxygen to 6 msw versus normobaric oxygen breathing on bubble formation in divers. *Eur J Appl Physiol*. 2009;106:691-5.
18 Bosco G, Yang ZJ, Di Tano G, Camporesi EM, Faralli F, Savini F, et al. Effect of in-water oxygen prebreathing at different depths on decompression-induced bubble formation and platelet activation. *J Appl Physiol*. 2010;108:1077-83.
19 Jurd KM, Thacker JC, Seddon FM, Gennser M, Loveman GA. The effect of pre-dive exercise timing, intensity and mode on post-decompression venous gas emboli. *Diving Hyperb Med*. 2011;41:183-8.
20 Blatteau JE, Hugon J, Gempp E, Pény C, Vallée N. Oxygen breathing or recompression during decompression from nitrox dives with a rebreather: effects on intravascular bubble burden and ramifications for decompression profiles. *Eur J Appl Physiol*. 2012;112:2257-65.
21 Schellart NA, Sterk W. Venous gas embolism after an open-water air dive and identical repetitive dive. *Undersea Hyperb Med.* 2012;39:577-87.
22 Gennser M, Jurd KM, Blogg SL. Pre-dive exercise and post-dive evolution of venous gas emboli. *Aviat Space Environ Med*. 2012;83:30-4.
23 Risberg J, Englund M, Aanderud L, Eftedal O, Flook V, Thorsen E. Venous gas embolism in chamber attendants after hyperbaric exposure. *Undersea Hyperb Med*. 2004;31:417-29.
24 Marinovic J, Ljubkovic M, Breskovic T, Gunjaca G, Obad A, Modun D, et al. Effects of successive air and nitrox dives on human vascular function. *Eur J Appl Physiol*. 2012;112:2131-7.
25 Castagna O, Brisswalter J, Vallee N, Blatteau JE. Endurance exercise immediately before sea diving reduces bubble formation in scuba divers. *Eur J Appl Physiol*. 2011;111:1047-54.
26 Dujic Z, Palada I, Obad A, Duplancic D, Bakovic D, Valic Z. Exercise during a 3-min decompression stop reduces postdive venous gas bubbles. *Med Sci Sports Exerc*. 2005;37:1319-23.
27 Møllerløkken A, Breskovic T, Palada I, Valic Z, Dujic Z, Brubakk AO. Observation of increased venous gas emboli after wet dives compared to dry dives. *Diving Hyperb Med*. 2011;41:124-8.
28 Blogg SL, Gennser M. The need for optimisation of post-dive ultrasound monitoring to properly evaluate the evolution of venous gas emboli. *Diving Hyperb Med*. 2011;41:139-46.
29 Naval Sea Systems Command. *US Navy Diving Manual*. Revision 6, NAVSEA 0910-LP-106-0957/SS521-AG-PRO-010. Arlington, VA: Naval Sea Systems Command; 2008.

Conflicts of interest: None

Acknowledgements

Aspects of this work were supported by the Naval Sea Systems Command Deep Submergence Biomedical Research Program. We thank WA Gerth for lively discussions on analysis of VGE data. The Bayesian analysis was based on the original MATLAB code authored by OS Eftedal.

Submitted: 06 August 2013
Accepted: 01 January 2014

David J Doolette[1], Keith A Gault[1] and Christian R Gutvik[2]

[1] Navy Experimental Diving Unit, Panama City, Florida, USA
[2] Technology Transfer Office, Norwegian University of Science and Technology (NTNU)

Address for correspondence:
David Doolette, PhD
Navy Experimental Diving Unit
321 Bullfinch Road
Panama City, FL 32407, USA
Phone: *+1-(0)850-230-3100*
E-mail: *<david.doolette.as@navy.mil>*