A Comparison of Water and Air Caloric Responses and Their Ability to Distinguish Between Patients with Normal and Impaired Ears

David A. Zapala, Ketil F. Olsholt, and Larry B. Lundy

Objectives: **The caloric test is a mainstay of modern vestibular assessment. Yet caloric test methods have not been well standardized, and normal response values have not been universally agreed upon. The air caloric test has been particularly problematic. In this article, we present our efforts to establish a population-based description of the caloric response evoked by water and air stimuli at both cool and warm temperatures.**

Design: **Data were collected from a retrospective record review of patients who underwent caloric testing at Mayo Clinic Jacksonville between 2002 and 2006. Two subgroups were identified. One group was found to have no vestibulopathy after comprehensive medical investigation. The second group was found to have severe bilateral vestibular weakness; this diagnosis was based on medical evaluation and objective test results. Caloric response distributions and associated probability estimates were developed from each group.**

Results: **A total of 2587 medical records were found to contain caloric response data. Of these, 693 patients met the criteria to be classified as having no identifiable vestibulopathy (otologically normal patients with normal caloric responses). Sixty-eight patients met the criteria for bilateral vestibular weakness (reduced or absent rotatory chair responses). Our analysis yielded the following results: (1) there were differences between nystagmus distributions across stimuli. On average, the magnitude of cool water (30°C) maximum slow-phase velocities was smaller than those from warm water (44°C). Maximum slow-phase velocity distributions from cool (21°C) and warm (51°C) air stimuli were more similar to each other than were responses to water stimuli and fell between the water distributions. (2) Combined metrics (combined eye speed and total eye speed) were comparable for water and air stimuli. (3) Response distributions from otologically normal patients were different from those of patients with bilateral vestibular weakness. (4) Derived probability estimates allowed for quantification of caloric response normal limits, sensitivity, specificity, and error rates.**

Conclusions: **Current bithermal test methods assume an equivalence of caloric response strength from warm and cool stimuli. Our results show standard cool and warm water stimuli provoke substantially different response magnitudes, with warm stimuli provoking stronger responses. When calibrated as described herein, air stimuli perform comparably with water stimuli for bithermal caloric test purposes, with more uniform and less variable response distributions. Both air- and water-based tests were able to distinguish between normal and abnormally weak ears with sensitivity and specificity values between 0.82 and 0.84. We advocate for the calibration of all caloric stimuli based on the test's statistical performance and not arbitrary assumptions about stimulus equivalence.**

(Ear & Hearing 2008;29;585–600)

Introduction

The caloric test is a mainstay of modern vestibular assessment. Although the test has a long history of clinical service, few studies have described the statistical behavior of the caloric response, particularly in the clinical setting. In fact, the American National Standards Institute (ANSI) bioacoustics committee defined procedures for executing the caloric test without standardizing the test performance (ANSI, 1999). So, for example, no criteria for identifying hypoactive and hyperactive responses were offered, even though this test has been in clinical use for at least 50 yrs. Additionally, consensus could not be reached for a procedural air caloric standard, despite its widespread clinical use (Burkard, 2006).

In our clinic, we have multiple electronystagmography (ENG)/videonystagmography (VNG) rooms, and each room has facilities for delivering water and air caloric stimuli. Consequently, we have needed to standardize our methods and calibrate our stimuli so that water and air caloric tests perform equivalently.

The purpose of this study was to describe the statistical characteristics of the caloric response as performed in our clinical setting. We specifically wanted to (1) investigate the response equivalence of cool and warm stimuli delivered via water and air media; (2) establish normal test limits, including criteria for hyperactive and hypoactive caloric responses; and (3) present our method for character-

Department of Otolaryngology – Head and Neck Surgery/Audiology, Mayo Clinic, Jacksonville, Florida.

0196/0202/08/2904-0585/0 • Ear & Hearing • Copyright © 2008 by Lippincott Williams & Wilkins • Printed in the U.S.A.

izing the caloric response in research and clinical settings using commonly available commercial software.

BACKGROUND

The history of vestibular testing is long and at times filled with controversy. In the 1860s, Smiederkam found that nystagmus could be observed after irrigating the ear canals with water (Sokolovski, 1966). Brown-Séquard (1860) reported the effects of thermal irrigations on the vestibular system. Barany developed the bithermal caloric test (Baloh, 2002; Sokolovski, 1966; Valli, et al., 2002–2003) for which he received the Nobel Prize in Physiology or Medicine in 1914. The test became easier to interpret when Jung and Mittermaier (1939), among others, introduced electrooculography. This allowed measurement of the slow phase of vestibular-induced nystagmus. By comparing the number of nystagmus beats, the duration of the caloric response, and the peak slow-phase velocity evoked by caloric stimulation, quantification of bithermal caloric asymmetry became possible (Jongkees & Philipszoon, 1964; Jongkees, et al., 1962). A number of investigators (Aschan, et al., 1956; Lion & Powsner, 1950; Monnier & Hufschmidt, 1950; van Egmond & Tolk, 1953) elaborated on this method, and over time it developed into the ENG test as we know it today (Milojevic, 1965).

In the 1990s, VNG advanced the possibility of eye movement recordings without the artifacts associated with the electrooculography (Geisler, et al., 2000). Over the past several years, the convenience and accuracy of VNG (Imai, et al., 2005; Vitte & Semont, 1995) have made it the method of choice for most vestibular laboratories.

Initially, water was the only irrigation used to provoke the caloric response. Barany (Baloh, 2002; Sokolovski, 1966; Valli, et al., 2002–2003) proposed that the temperatures of the irrigations be set to 7°C higher and lower than the presumed core body temperature (i.e., 37°C; Baloh, 2002) on the assumption that warm and cool stimulations provoke equally strong responses. However, no empiric data are available to validate these settings.

Over time, other methods of irrigation were suggested. For example, closed-loop systems were developed that used an ear canal balloon. Water circulated through the ear canal balloon allowing the caloric test to be performed in ears with perforated eardrums (Anderson, 1995; Ono & Kanzaki, 1976; Westhofen, 1987). These systems are no longer commercially available.

In the late 1950s and 1960s, air irrigators became commercially available (Capps, et al., 1973; Coats, et al., 1976). These systems are in wide clinical use today, but the accuracy of the air-induced caloric effect remains a point of controversy. Some have argued that the air caloric test is unreliable and air stimuli are unable to produce robust enough responses (Greven, et al., 1979; Torok, 1979; Zangemeister & Bock, 1980). Others claim that weak or unreliable air caloric responses reflect poor technique (Ford & Stockwell, 1978; Moon & Munro, 1996; Munro & Higson, 1996; Norre & Dierick, 1983). Because of the lack of consensus and limited empirical data, the air caloric test was excluded from the current ANSI standard (ANSI, 1999; Burkard, 2006).

RATIONALE

Ideally, a caloric test should reliably and accurately distinguish between normal vestibular output and abnormally reduced or increased output. To be truly standardized, the test should yield the same result regardless of where the test is performed and what stimuli are used. Ever since the introduction of the caloric test, investigators have stated the need for standardization and further improvement (Frenzel, 1959; Kobrak, 1918). Despite this need, empirical evidence to establish test accuracy and reliability is surprisingly limited. Studies reporting the statistical performance of the caloric response have often had relatively few subjects. Additionally, they report wide variation in caloric response magnitude, both within and between studies (Table 1). For example, average cool water maximum slow-phase velocities (MSPVs) considered normal have ranged from a low of 9.8 degrees/sec (Westhofen, 1987) to a high of 29 degrees/sec (Henriksson, 1956). Similarly, average warm water MSPVs considered normal have ranged from a low of 10 degrees/sec (Westhofen, 1987) to a high of 35 degrees/sec (Barber & Stockwell, 1980). Similar trends can be seen for the reported air caloric MSPVs. In six of the 21 studies found, a caloric response of less than 6 degrees/sec would have been accepted as a normal occurrence. In three studies (Karlsen, et al., 1992; Westhofen, 1987; Zangemeister & Bock, 1980), reported SDs implied the 95% normal limit of the caloric-induced MSPV to include values less than zero. Using an irrigation technique that provokes weak nystagmus is more comfortable for the patient, because of the reasonably close relationship between the magnitude of nystagmus and the perception of vertigo. Nevertheless, from a statistical point of view, when the evoked caloric response is low, a reduced caloric response from an impaired ear may be indistinguishable from a similar response from a normal ear.

The variations shown in Table 1 may be caused by several factors. Technical variables, such as the tem-

TABLE 1. Survey of literature reports presenting caloric irrigation distribution data*

	Water					Air				
		Cool		Warm			Cool		Warm	
Source	No.	Mean	Range	Mean	Range	No.	Mean	Range	Mean	Range
Henriksson, 1956	25	29	8–45	29	8–65					
Hamersma, 1957; Jongkees and Philipszoon, 1964	47	22	9–46	23	8–52					
Mehra, 1964	31	21	3–39†	26	0–52†					
Custer, et al., 1973	20	14.5		15.3						
Capps, et al., 1973	10	18	8–28	16	6–26	10	17	7–27	15	3–27
Coats, et al., 1976	25	24.9	10–44	26.5	10–68	25	22.9	10–54	24.8	4–70
Suter, et al., 1977	484	25.5		26.2		191	27.0		32.5	
Reker, 1977	25	13.2	7–29							
Sills, et al., 1977	43	14.5	5–40	19.9	6–68					
Benitez, et al., 1978						30	21	5–37†	22	6–38†
Ford and Stockwell, 1978	8	17	11–23	22	16–28	8	15	10–20	20	15–25
O'Neill, 1978	60	16.2		21.0		60	15.8		20.5	
Greven, et al., 1979	28	19.0		17.3			10.5		13.8	
Press, et al., 1979	14	29.7	1–59†	23.2	3–43†					
Barber and Wright, 1980	114	28	6–50	35	11–80	24	30	10–46	37	11–85
Zangemeister and Bock, 1980	25	17.7	5–30†	21.4	7–36†	20	12.1	0–24†	11.5	−1–24†
Norre and Dierick, 1983‡	18	12.5	7–23†			18	7.9	5–13†		
Proctor and Glackin, 1985	30	14.5	9–28	18.7	10–40					
Westhofen, 1987§	40	9.8	−7–27†	10.0	−8–28†	40	11.8	−7–30†	7.6	−5–20†
Baloh and Honrubia, 1990	44	15	5–40	21	6–68					
Karlsen, et al., 1992	24	24.8	−1–51†	12.1	−1–27†	24	17.0	−4–38†	9.1	−2–20†
Jacobson, et al., 1993	100	20.3	8–49	29.6	9–73					
Anderson, 1995§	10	25.3	9–47	15.9	6–28	10	20.5	9–30	15.2	5–23
Munro and Higson, 1996	12	23	9–37†	21	5–37†	12	14	4–24†	22	4–40†
Henry, 1999§	20	16.8	5–29†	19.9	5–35†	20	20.9	−1–42†	19.3	−1–38†
Van Der Stappen, et al., 2000¶	40	18.2	3–33†	30.9	7–55†					
Molina, et al., 2006‖**	107	13.2	3–34	19.8	3–73					
Maes, et al., 2007	42	23.4	12–35	27.0	6–48	42	13.8	4–24	11.9	2–21

** Values are reported as degrees per second unless indicated otherwise.*
† Range has been calculated using mean ± 2 SD.
‡ Used Freon instead of air.
§ Compared regular water with closed-loop water testing.
¶ Side-specific numbers in article.
‖ Sex-specific numbers in article.
*** Videonystagmography technique.*

perature and duration of the caloric stimulus, accuracy of stimulus delivery, recording method (ENG/VNG), sample rate, and filter settings are under the investigators' control. Procedural factors, such as the positioning of the patient, timing of fixation interval, room luminescence, and ambient temperature, are also under the investigators' control. Other factors, particularly patient-specific variables such as mastoid variation, may not be possible to control entirely. Unfortunately, a lot of investigators have their "own way" of doing caloric tests (Blakley, 2000) and may not always appreciate or report the technical and procedural variables unique to their setting. As a consequence, it is difficult to compare normal values from different laboratories (Bock & Zangemeister, 1978; Proctor & Glackin, 1985). In short, the caloric test is not well-standardized.

To use water and air stimuli interchangeably, the ability of each test to distinguish between normal, hypoactive, and hyperactive responses must be comparable. Thus, the magnitude of the caloric response evoked from each irrigation type in normal ears falls between the distribution of responses from hypoactive and hyperactive ears. It is also necessary to establish that each irrigation type provokes a caloric response of similar magnitude; that is, caloric stimuli provoke reliable, valid, and interchangeable measurements of vestibular output. We sought to test whether the caloric response MSPVs in our clinic were similar for both media (air and water) and both temperatures (warm and cool), with distributions that allowed us to distinguish between normal and abnormal ears with an acceptable level of statistical precision.

MATERIALS AND METHODS

The study was performed at Mayo Clinic Jacksonville, with oversight and approval of the Mayo Clinic Institutional Review Board. Mayo Clinic is an integrated multispecialty clinic that uses an electronic medical record for each patient. Although patients may be seen directly by each subspecialty, the optimal model is for the patient to be evaluated by a primary care physician, who then organizes and oversees any subsequent specialty evaluations. This integrated multidisciplinary assessment, buttressed by the electronic medical record, promotes clear communication among all providers who can follow the patient's diagnostic progress and treatment outcomes over time.

Patients are commonly referred for vestibular study as part of their evaluation for complex or obscure complaints of dizziness, ataxia, imbalance, nausea, and the like. There is an effort to be comprehensive in planning evaluations, because many patients come to Mayo Clinic after management of their conditions elsewhere has failed. As a consequence, some patients who ultimately do not have vestibulopathy are seen for vestibular study. These patients' test results were the primary focus of this study.

During the period when data were collected for this study, the vestibular laboratory at Mayo Clinic Jacksonville housed two dedicated ENG/VNG rooms. Each room had a custom examination table that allowed the patient's head and upper torso to be elevated 30 degrees during the caloric test. Each room had an air caloric irrigator (ICS model NCA-200; GN Otometrics, Taastrup, Denmark) and a water caloric irrigator (ICS model NCA-480; GN Otometrics). The water caloric irrigator was calibrated to meet ANSI s3.45–1999 standards for temperature (ANSI, 1999). The duration of the caloric irrigation was 40 sec rather than the ANSI standard of 30 sec. The reason for this deviation from the standard will be discussed later in this article. The air irrigator was calibrated so that induced caloric responses approximate the responses provoked by the water caloric irrigator (Zapala & Shaughnessy, 2005). Table 2 summarizes the caloric test stimuli characteristics.

Eye movements were recorded using the ICS medical video goggle system (CHARTR VNG, version 4.1; GN Otometrics). The system was calibrated before caloric testing. Recordings were made for at least 2 min after the onset of caloric stimulation. During the recording interval, effort was made to keep the induced eye movements close to midline to avoid parallax errors. All patients were given alerting tasks that required mental imagery (naming fruit commonly seen in the market, naming cities in the state, etc.) or counting (subtracting serial 7s from 100, etc.). The MSPV was established using the computer's internal algorithm. The typical interval between caloric irrigations was at least 5 min.

To accomplish this study, we used information in the electronic medical record to retrospectively identify two groups of patients who underwent vestibular testing at Mayo Clinic Jacksonville between 2002 and 2006. The first group of patients had a vestibular study, but these patients subsequently were found not to have vestibulopathy or other clinical conditions that would likely influence caloric test results. Specifically, patients in the first group had no indication of vestibulopathy on a comprehensive vestibular evaluation that included computerized dynamic postureography, rotary chair measurements, and vestibular-evoked myogenic potentials, as well as a standard VNG study. Additionally, the treating physicians agreed that vestibulopathy was not present in these patients. Typically, this was based on otologic or neurologic opinion.

The second group of patients was found to have bilateral vestibular weakness on rotary chair slow-harmonic acceleration testing and did not have notable caloric asymmetry. To be considered to have bilateral weakness, vestibular ocular reflex gains had to fall below the 2.5 SD normal limit at frequencies at and below 0.16 Hz on slow-harmonic acceleration. More severe forms of bilateral weakness, extending to higher-frequency ranges, were included in the patients with bilateral weakness.

Ideally, we should have had a third group consisting of patients with bilateral hyperactive responses. However, we could not find a way to identify a hyperactive vestibular response independent of the caloric test. Thus, to avoid a circular argument, we refrained from establishing such a group, and instead the 95% limit served as definition of hyperactive response.

TABLE 2. Caloric irrigation characteristics for water and air stimuli*

	Water caloric stimulator		Air caloric stimulator	
	Cool	Warm	Cool	Warm
Temperature (°C)	30	44	21	51
Irrigation duration (sec)	40	40	70	60
Flow rate	7.5 mL/sec (0.45 L/min)†		133.3 mL/sec (8 L/min)‡	

** Water caloric stimulus was calibrated using a thermometer positioned 2 cm from the tip of the delivery tube after the temperature was stabilized. Air caloric stimulus temperature was measured using a thermistor positioned approximately 5.5 cm upstream from the tip of the air delivery tube, which was hard wired into the irrigator system by the manufacturer. The air irrigation parameters were established by an internal study completed in 2001. Briefly, air-induced caloric responses were obtained from four normal subjects and compared with their analogous water-induced caloric responses. Additionally, air and water caloric distributions from patients subsequently found not to have vestibulopathy were compared using manufacturer-recommended air irrigation settings. Over the course of several months, air caloric tests were repeated on the four normal subjects, with the air settings adaptively modified to induced caloric response magnitudes approximating the average normal monothermal, CES, and TES for water. The cool air duration was increased beyond 60 sec to account for the inability to equate the air and water responses using the available temperature range of the air caloric system. There were no adverse effects noted after these calibration changes. Specifically, there was no appreciable increase in the number of studies that could not be performed because of patient nausea or emesis.*

† Calculated by dividing measured delivered volume by irrigation duration time.

‡ Dial reading.

TABLE 3. Demographic characteristics of subject groups used in this analysis

	No vestibulopathy identified		Bilateral vestibular weakness	
Caloric stimulus	Water	Air	Water	Air
No. cases	171*	522*	14	54
Sex (F:M)	1.41:1	1.52:1	1.33:1	0.69:1
Mean age (range), yr	64.35 (16.8–100.1)	63.11 (17.7–89.9)	73.44 (39.5–93.2)	72.30 (44.3–94.3)

** Five patients had both air and water testing.*

The MSPV data for patients with no identified vestibulopathy were modeled (using Microsoft Excel 2003 analysis tool pack and self-generated routines) and compared with literature references for normal caloric-evoked vestibular responses. Modeling consisted of establishing the frequency of observed MSPVs for each caloric test (water and air, cool and warm stimuli). Observed cumulative distributions were then compared with best fit normal, logarithmic, and power distributions using a least squares error criterion. The best fit distributions were then used to establish probability estimates for the observed data. Once the best fit distribution was established, the magnitude of the average response was calculated in z score units to estimate the probability that an MSPV of zero would be observed in the normal patient group.

To assess whether patient age had any effect on caloric response magnitude, total eye speeds were assessed as a function of age.

Measured MSPV curve fit data were then contrasted with analogous MSPV curve fit data from patients with bilateral vestibular weakness in an effort to evaluate how well the caloric test could differentiate between normal and hypoactive vestibular ocular reflex responses.

RESULTS

Between 2002 and 2006, a total of 2587 electronic medical records were found to contain caloric response data. Of these, 693 met the criteria to be classified as not having identifiable vestibulopathy. During the same interval, 68 records met the criteria for bilateral vestibular weakness. Bilateral, bithermal caloric response data, consisting of MSPVs, caloric stimulus type, and patient age and sex were collected from the medical record and stored for further analysis. Table 3 summarizes the demographic characteristics of each group.

Caloric-induced MSPVs in normal individuals are not normally distributed but tend to be logarithmically distributed (Jacobson, et al., 1993; Nijhuis & Huygen, 1980; Sills, et al., 1977). Figure 1 shows the distribution of caloric responses from patients with no identified vestibulopathy for water and air stimuli, with both cool and warm temperatures. The logarithmic curve fit (least squares error solution) to the distribution is also displayed. Based on the fitted curves, Table 4 presents the mean and the 90%, 95%, and 99% confidence limits for each irrigation method.

Several observations can be made from these data. First, the peak (central tendency or geometric mean) for each distribution is fairly similar. Cool water-induced responses were, on average, weaker than both cool and warm air-induced responses, which in turn were weaker than warm water-induced caloric responses. However, this is only half the story. The variability around the geometric mean was not identical across the different stimulus types and temperatures. The warm stimuli tended to provoke stronger MSPVs, and their distributions showed a stronger positive skew. This means that some individuals produced strong warm caloric-induced MSPVs relative to their cool caloric-induced MSPVs. What might be considered a hyperactive caloric response thus varies depending on the type and temperature of the stimulus. Similarly, what constitutes a weak caloric-induced response likely also varies across stimulus type and temperature, although to a much smaller extent.

On the bottom row of Table 4, the geometric means, expressed as z scores, reflect how many SDs the geometric mean is away from zero (no response). Thus, z scores greater than 3 indicate that the average caloric-induced MSPV is at least 3 SDs away from no response. With a z score of 3, a patient with normal vestibular function would be unlikely (1:450) to have a caloric-induced MSPV of zero by chance. In Table 4, calculated z scores were higher than 5.

Differences between the caloric stimuli can be further appreciated in Figure 2. The two air distributions are nearly identical, particularly from the 50% point downward (both with geometric mean of 26 degrees/sec). Above the 50% point, there was a tendency for warm air-induced caloric responses to show higher MSPVs than those induced by cool air. Even so, the two distributions seldom differ by more than 10 degrees/sec.

The water stimuli seemed to perform quite differently. On average, the cool water stimulus provoked

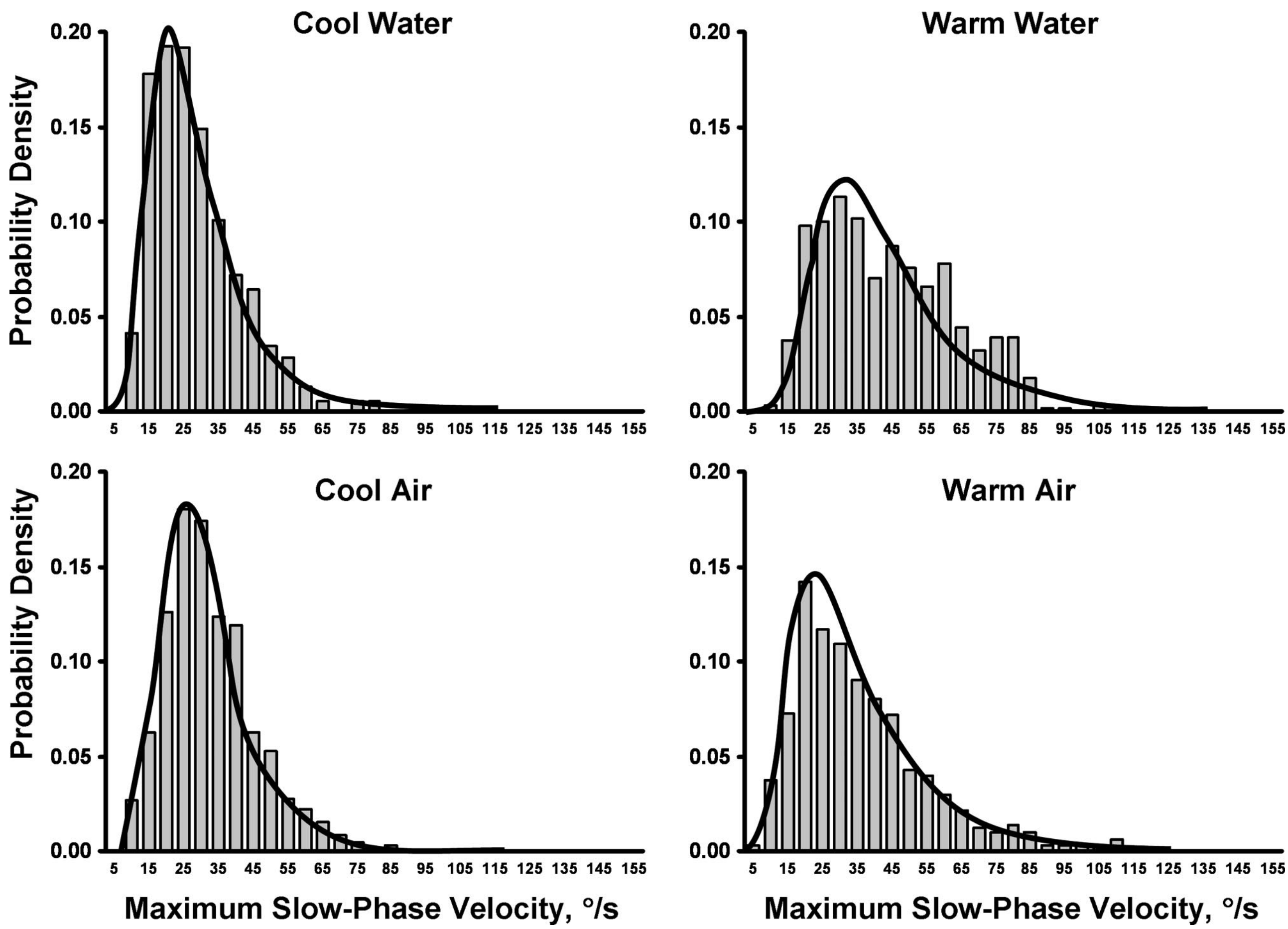

Fig. 1. Probability density distributions of maximum slow-phase velocities (MSPVs) for cool and warm water and air caloric stimuli. Bars reflect the normalized proportion of observations in each 5 degrees/sec bin, centered on labeled MSPV. For example, the bin labeled 5 degrees/sec includes observations between 2.5 and 7.5 degrees/sec. The lowest cell, containing observations between 0 and 2.5 degrees/sec is not shown and had no observation points. Curve fits are based on logarithmic transformed data using least squares solutions.

weaker responses than the warm water stimulus (geometric means of 20 and 35 degrees/sec, respectively). Further, the distributions separated by an even greater amount when the provoked MSPVs increased in magnitude. The curve fit to the warm water data was not as good as the other logarithmic solutions, although the logarithmic distribution still fit the data better than the normal or power curve fit solutions. The poorer curve fit reflected the irregular shape and larger variability of the warm water caloric response.

Figure 3 shows the modeled cumulative probability distributions for combined eye speed (CES) data. The CES is the monaural, bithermal eye speed, calculated by adding the cool and warm MSPVs, using either water or air, from one ear. As can be

TABLE 4. Maximum slow-phase velocity geometric mean values, *z* scores, and estimated population confidence limits for water and air stimuli

	Maximum slow-phase velocity			
	Water irrigation		Air irrigation	
Parameter	Cool	Warm	Cool	Warm
Geometric mean	20.60	35.17	25.51	26.19
90% CL	8.31,51.10	14.59,81.75	11.66,55.82	9.48,72.35
95% CL	6.98,60.81	12.33,100.30	10.03,64.85	7.81,87.90
99% CL	4.97,85.44	8.87,139.43	7.48,86.94	5.34,128.58
z score	5.48	6.66	6.80	5.29

CL, confidence limit.

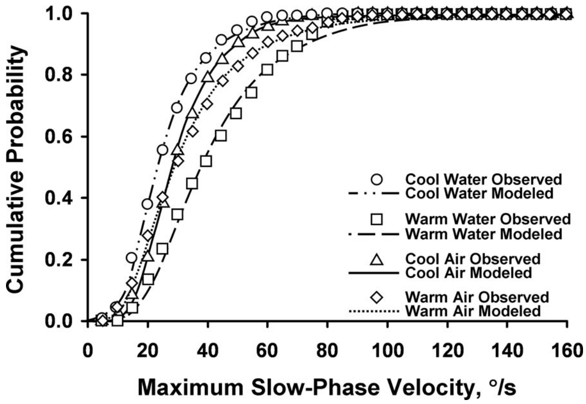

Fig. 2. Cumulative probability density of maximum slow-phase velocity for each caloric stimulus type. Symbols represent observed data. Lines represent modeled curve fits. The cool water curve demonstrated a weaker response than did warm water. Air stimuli (cool and warm) fell between the water curves.

seen, the distributions for water and air CES were similar.

Table 5 presents the geometric mean and confidence limit data derived from the modeled probability distributions. Both distributions have similar geometric means. The water caloric distribution was slightly broader, more so when CES values were greater. The upper confidence limits for water stimuli tended to be about 24% greater than analogous air values. This reflected the variability in the underlying warm water distribution.

The total eye speed (TES) is defined as the sum of the MSPVs from both cool and warm irrigation on both sides. TES cumulative probability distributions for water and air stimuli are presented in Figure 4. As with the CES distributions, the geometric means compared favorably. Again, the upper confidence limits for water tended to exceed those of air by about 24%

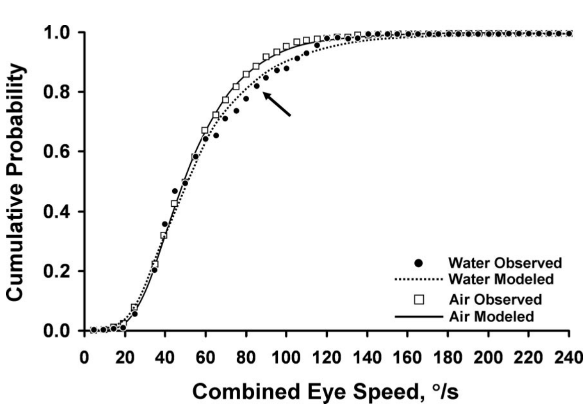

Fig. 3. Combined eye speed cumulative density probability. Symbols reflect the observed cumulative distribution. Lines represent the modeled cumulative distribution. The arrow shows where the water curve diverges from the air curve above the geometric mean.

because of the larger variability in the underlying warm water distribution. The confidence limits for the TES distributions are summarized also in Table 5.

To determine whether age affected these data, TESs were plotted as a function of age for water and air stimuli. These are shown in Figure 5. Neither air nor water results showed an appreciable trend as a function of age.

Figure 6 plots the cumulative probability distribution for CES from normal subjects and the reverse probability distribution from patients with bilateral vestibulopathy. By studying these distributions, the probability that any observed CES reflects a normal or an abnormal outcome can be established. The point where the two distributions cross each other reflects the best discriminatory performance for that test. The test discriminates better when the crossover point is associated with a lower cumulative probability. In fact, the cumulative probability associated with the crossover point reflects the misclassification rate (i.e., the false-positive and false-negative rate) at best test performance. From these data, the sensitivity and specificity of the test can be determined.

Table 6 presents the crossover point, associated cumulative probability, and sensitivity and specificity calculations for water and air CES and TES. The CES data are helpful in determining whether the output of a specific ear is normal, hypoactive, or hyperactive. The TES plots are helpful in determining whether overall vestibular output is normal or abnormally reduced or hyperactive bilaterally. As a starting point, the 95% limits may be taken as the defining limit of normal.

At the crossover point (best test performance) for CES, the sensitivity and specificity are equal. Water had a sensitivity of 0.84 and a specificity of 0.84, and air had a sensitivity of 0.82 and a specificity of 0.82, reflecting good, but not ideal performance. Moreover, if we were to accept the best test performance criterion as a clinical standard, for water and air, respectively, we would misclassify 16% and 18% of normal responses as abnormal (type I error), and we would misclassify 16% and 18% of abnormal responses as normal (type II error).

A similar analysis can be performed on TES data, only in this case, the errors reflect patient misclassifications. So, for water and air, respectively, at best test performance we would misclassify 11% and 16% of normal responses as bilaterally reduced (type I error = false positive = α), and we would misclassify 11% and 16% of bilaterally reduced responses as normal (type II error = false negative = β).

For clinical purposes, using a criterion of best test performance may not always be desirable. For example, one might choose to minimize the misclassi-

TABLE 5. Combined eye speed and total eye speed geometric mean values and estimated population confidence limits for water and air stimuli

	Combined eye speed		Total eye speed	
Parameter	Water	Air	Water	Air
Geometric mean	48.22	46.84	100.61	97.23
90% CL	19.76,117.65	21.69,101.18	43.45,233.00	46.98,201.22
95% CL	16.66,139.58	18.71,117.27	36.99,273.66	40.87,231.30
99% CL	11.93,194.93	14.02,156.46	27.01,374.76	31.13,303.70

CL, confidence limit.

fication of normal patients as abnormal (type I error rate). If the 90% (two tail) confidence limit was used as a criterion, the type I error rate would be 5%. However, the misclassification of abnormal responses as normal (type II error rate) would be much greater. Figure 7 visualizes the situation for water and air TES. Assuming a 5% type I error rate, the type II error rate would be 23% and 31% for water and air, respectively.

The curves in Figures 6 and 7 can provide estimates of the likelihood of types I and II errors for any criterion CES or TES value. The calculated variables for all the derived plots are included in the Appendix.

DISCUSSION

The purpose of this study was to evaluate the statistical performance of the caloric response as evoked in a clinical setting. We were specifically interested in establishing normal limits for cool and warm temperature stimuli, delivered via water and air media. Ideally, we would like to be able to use the same clinical norms for either medium, using air or water stimuli as clinical needs dictate. Consequently, we hoped that each caloric response distribution would be statistically similar. On this count, the results were arguably mixed. The distributions for cool and warm water stimuli differed from each other, with warm water stimuli provoking stronger caloric responses over a wider range of MSPVs than cool water stimuli. Cool and warm air stimuli provoked caloric responses with distributions that were almost identical, with average performance falling between that of warm and cool water stimuli. However, as with the water stimuli, warm air tended to provoke MSPVs over a somewhat broader range than did cool air.

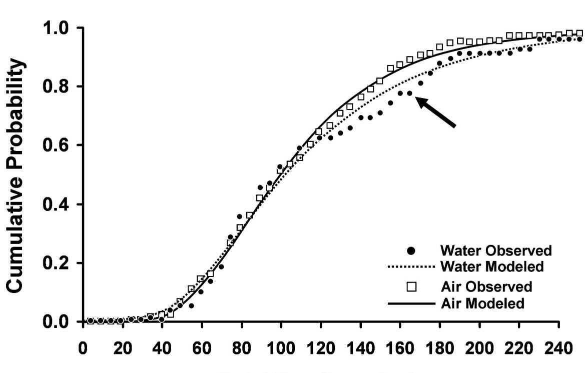

Fig. 4. Total eye speed cumulative density probability. Symbols reflect the observed cumulative distribution. Lines represent the modeled cumulative distribution. The arrow shows where the water curve diverges from the air curve above the geometric mean. The water total eye speed data varied from the model curve fit above the geometric mean, reflecting the broader and flatter warm water distribution. More observation points may have improved agreement between modeled and observed data.

Central Tendency and Variability

Before delving into the specifics of the results, it might be helpful first to review what we hoped to see in the ideal caloric test. The caloric test is based on the assumption that, in most cases, impaired ears provoke weaker caloric responses than normal ears. Because the weakest observable caloric response is zero (no evoked MSPV, i.e., no response), the distri-

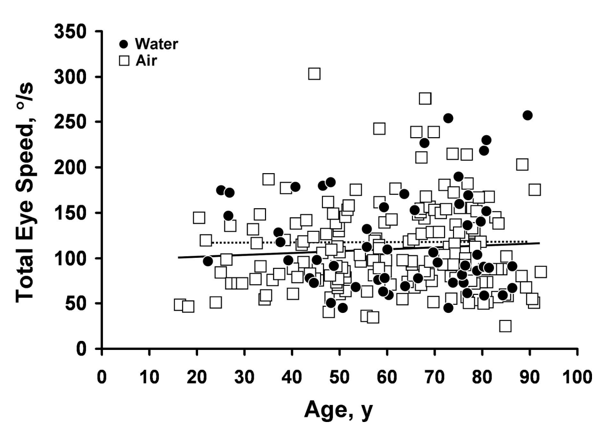

Fig. 5. Total eye speed (TES) as a function of age. Symbols represent observed age with corresponding TES. Lines represent overall trends (water, dashed line; air, solid line). The regression for the air TES was 97 ± 0.2 (age in yrs) (*p* = 0.30). The regression for the water TES was 115 ± 0.03 (age in yrs) (*p* = 0.94).

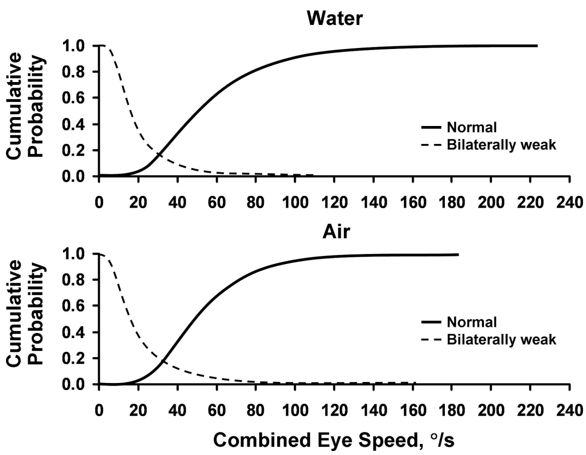

Fig. 6. Cumulative probability distribution for combined eye speed (CES) data from subjects who were otologically normal and those with bilateral vestibular weakness. The cumulative probability distributions for otologically normal subjects are shown as solid lines. The reversed cumulative probability distribution (1-cumulative probability) for subjects with bilateral weakness are shown as dashed lines. Crossover points and associated probabilities are shown in Table 6. For any CES, the cumulative probability of the value coming from the group with normal responses is the y axis value corresponding to the CES point on the solid line. The cumulative probability that a given CES came from the group with bilateral weakness is 1-cumulative probability (the y axis value) associated with the corresponding CES value on the dashed curve.

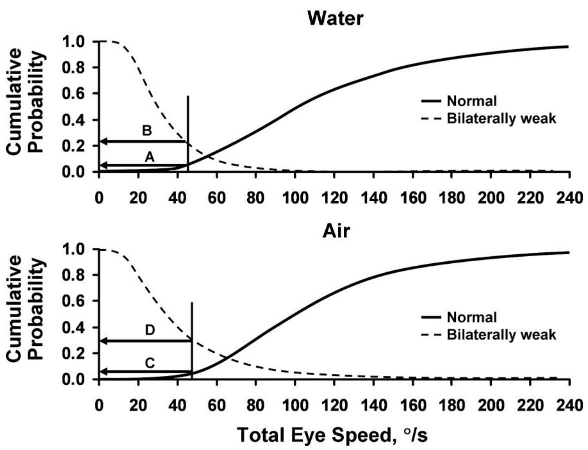

Fig. 7. Cumulative probability density distributions for total eye speed (TES). Layout is the same as in Figure 6. Vertical lines reflect TES value at the fifth percentile limit for the normal subject distribution. The horizontal arrow A marks the probability. The horizontal arrow B marks the corresponding probability for the group with bilateral weakness. In the case of the water caloric test (top panel), the fifth percentile TES value is 43.5 degrees/sec. The corresponding cumulative probability on the bilaterally weak curve is 23%. This means that 23% of subjects with bilateral weakness would have been misclassified as normal if this criterion was used. Similarly, on the lower panel, the fifth percentile (arrow C) for the air caloric test is at 47 degrees/sec. Here 31% (arrow D) of those with bilateral weakness would have been misclassified as normal.

bution of normal caloric responses should have a magnitude and variability such that MSPVs of zero would hardly ever occur by chance.

Prior studies describing the statistical characteristics of the caloric response are summarized in Table 1. With a few exceptions, caloric MSPV distributions were not carefully described in these reports. In those studies that did characterize these distributions, they were logarithmic (Jacobson, et al., 1993; Sills, et al., 1977). We also found a logarithmic distribution for the caloric response, as can be seen in Figure 1. This observation has two important implications. First, as others have noted, the use of the mean as a measure of magnitude or central tendency in a logarithmic distribution is not appropriate. Because of the positive skew in these distributions, the median response or, more desirably, the geometric mean is a more accurate measure of central tendency. Second, again because of the positive skew of these logarithmic distributions, the expected range of MSPVs in the normal population is not centered on the population geometric mean. Consequently, each distribution needs to be transformed to establish normal limits with statistical precision. When variability is analyzed using transformed data, the lower limits of the distribution appear closer to the geometric mean, and the upper limits of the distribution appear further from the geometric mean. Jacobson et al. (1993) noted this problem and used a nonparametric method of describing caloric variability. We could not find a study that characterized the caloric response using the parametric approach we used. Importantly, several articles listed in Table 1 reported untransformed mean and SD values. This means that the reported variability of the caloric response was likely inaccurate.

TABLE 6. Combined eye speed and total eye speed crossover points and associated sensitivity and specificity calculations for distinguishing between normal cases and cases with bilateral weakness

Parameter	Combined eye speed		Total eye speed	
	Water	Air	Water	Air
Crossover points (degrees/sec)	27.9	30.4	54.0	62.8
Cumulative probability at crossover point (%)	16	18	11	16
Sensitivity at crossover point	0.84	0.82	0.89	0.84
Specificity at crossover point	0.84	0.82	0.89	0.84

We believe that it is important to use a parametric approach to describe the caloric response performance. Knowing the distribution of MSPVs allows us to model caloric variability with known statistical precision. That is, we can establish the probability that any observed MSPV was evoked from a normal or an abnormal ear, obviously important for clinical decision making. It also allows us to compare more fully the performance of the caloric response evoked by different types of stimuli, such as the temperature and medium affects addressed in this article. Finally, carefully describing MSPV distributions across studies is the first step in understanding and controlling the factors that contribute to test variability across laboratories and research reports. This should aid in establishing universal standards for the caloric test (Burkard, 2006).

Our first concern in this study was to determine whether the caloric response evoked in our laboratory had sufficient magnitude to discriminate between low-normal responses and abnormal-reduced response outcomes. This can easily be determined by expressing the geometric mean in terms of a z score. The bottom row in Table 4 shows the z score values for the geometric mean associated with each caloric stimulus. All of our stimuli had z scores higher than 5, indicating that the chance of obtaining an absent response (MSPV of zero) from a normal subject was much less than 0.1%.

Prior studies have not always demonstrated sufficiently strong caloric response magnitudes. For example, the analyses by Zangemeister and Bock (1980), Westhofen (1987), and Karlsen et al. (1992) all suggest possible normal responses for MSPVs that ranged below zero. Having normal responses less than zero suggests that either the overall magnitude of the caloric response was too low or the variability of the caloric response was not calculated correctly. In either case, the ability of the caloric test to discriminate between a normal response and an abnormally hypoactive response is uncertain. This uncertainty highlights the importance of calibrating caloric stimuli to achieve a desired discriminative performance and describing test performance accurately.

Figure 2 shows the cumulative distributions for the water and air stimuli. Considering the water stimuli, it is clear that the warm stimulus provoked a stronger magnitude response than the cool stimulus (geometric means of 35.2 and 20.6 degrees/sec for warm and cool, respectively). Similar trends were reported by others (Baloh & Honrubia, 1990; Barber & Stockwell, 1980; Coats, et al., 1976; Ford & Stockwell, 1978; Hamersma, 1957; Jacobson, et al., 1993; Proctor & Glackin, 1985; Sills, et al., 1977; Zangemeister & Bock, 1980). Jacobson et al. (1993) reviewed some of the possibilities that explain this trend. First, Barany's theory of caloric-induced convection currents within the horizontal semicircular canal may be incomplete. Barany reasoned that warm and cool caloric stimuli, equidistant from core body temperature, would provoke an equal but opposite stimulus to the horizontal semicircular canal. Jacobson et al., citing the work of others, suggested that, beyond the thermal convection theory, warm stimuli directly excite sensory or neural elements within the vestibular end organ, adding to the convection effect. Similarly, Wit et al. (1990) suggested that a volume expansion or contraction effect on the fluids may occur in the horizontal semicircular canal in response to warm and cool stimulation, respectively. This expansion or contraction would modify the mechanical movement of the cupula in response to caloric stimulation. Finally, Ewald's second law may predict a higher ceiling effect from warm caloric stimuli since warm stimuli excite eighth nerve afferents from the horizontal canal.

Not all the studies in Table 1 reported stronger warm water caloric responses. For example, Capps et al. (1973) and Karlsen et al. (1992) reported stronger cool water MSPVs. It is clear from the data in Table 1 that the caloric response, although studied for the past 50 yrs, is still not standardized in a way that provides reliable, consistent results across studies. Part of the inspiration for this report was to work toward a careful inspection of issues that might contribute to this variability.

The temperatures used for the water caloric test have been standardized on the basis of Barany's original convection theory. That is, caloric stimuli of 30 and 44°C are used uniformly for open water irrigation systems (ANSI, 1999). However, other stimulus and recording parameters of the water-evoked caloric response have not been uniformly applied across studies. In our laboratory, we irrigate for 40 sec rather than 30 sec. Initially, this was chosen because it followed manufacturer recommendations for using our particular irrigation equipment. We continued to use this approach after establishment of the ANSI standards simply to ensure that the caloric responses evoked in our laboratory were of sufficient magnitude to distinguish between normal and abnormal responses. Other studies (Anderson, 1995; Karlsen, et al., 1992; Mehra, 1964; Sills, et al., 1977; Suter, et al., 1977) also used 40 sec for the irrigation duration. When comparing the performance of the water caloric test across studies, these variables are obviously important.

Air has less caloric capacity than water and is thus less efficient as a caloric stimulus (Bock & Zangemeister, 1978). We established our air irrigation methods with the aim of provoking caloric responses that are essentially equivalent to responses provoked by water stimuli (Table 2). In Figure 2, the cumulative distribution of air-evoked MSPVs fell midway between the distributions for cool and warm water stimuli. Further, the geometric means for warm and cool air stimuli were similar

(25.51 and 26.19 degrees/sec, respectively). This was a gratifying result. However, to achieve essentially equivalent warm and cool response distributions, we used stimuli that, at face value, were not equivalent. Rather, we calibrated the stimuli to produce the desired statistical performance of the test. Specifically, air temperatures were not equidistant from average core body temperature, and the irrigation durations were 60 and 70 sec for warm and cool stimuli. Having done this, the performances of warm and cool air conductive stimuli were still not perfectly identical. Warm air-induced caloric responses demonstrated a slight positive skew, meaning that the observed distribution of MSPVs included higher values from warm stimuli than for cool stimuli.

The warm water stimulus also demonstrated a distribution with a strong positive skew relative to the cool water distribution and the warm air distribution. As water is a better thermal conductor than air (Bock & Zangemeister, 1978), it is tempting to postulate that the nonconvection effects summarized by Jacobson et al. (1993) are more notable when water stimuli are used. Regardless of the cause, from a statistical point of view, the distribution of warm water MSPVs was quite broad and variable. It remains to be seen whether information is carried in this variability or if the warm water response is simply unstable. We will investigate this question in future studies. For the present purposes, it is important to remember that the greater spread of MSPVs associated with warm stimuli influences the statistical properties of all subsequent calculations that incorporate these variables.

MSPV Limits

It was our aim to establish the normal clinical limits for air and water stimuli. Such limits could then be used to describe hyperactive and hypoactive caloric responses with a known degree of statistical certainty. Our results are shown in Table 3. Even though the magnitudes of the caloric responses were relatively strong compared with the results of the other studies detailed in Table 1, the variability of the test was still problematic. For example, from Table 1, approximately one in 10 cool water caloric irrigations from healthy ears had an MSPV of less than 9 degrees/sec. Approximately one in 100 had an MSPV of less than 5 degrees/sec. These values are low enough to challenge the caloric test's ability to distinguish between normal and hypoactive responses, at least in the case of solitary irrigations. To place this into perspective, positional nystagmus of 9 degrees/sec or less can be encountered in the normal population without any caloric stimulation (Barber & Wright, 1973).

On the high side, hyperactive caloric responses varied tremendously, depending on the type of irrigation. Cool water irrigations were hyperactive when the MSPV was greater than 61 degrees/sec at the 95% confidence limit. A hyperactive warm water response at the same statistical level would have to be greater than 100 degrees/sec. The trend for having a higher cutoff for warm caloric stimuli was not unexpected. Barber and Stockwell (1980) set limits in their laboratory of 50 and 80 degrees/sec for cool and warm water MSPVs, respectively.

The upper limits for the air distributions varied less across temperature than was observed for water. The 95% limit was 65 and 88 degrees/sec for cool and warm temperature stimuli, respectively. One potential implication from these data is that it might be easier to identify a hyperactive response from an air stimulus because the normal variability of the response is less than that observed with water.

CES and TES

When using the Jongkees et al. (1962) formula to calculate a unilateral weakness percentage, the cool and warm responses from each individual ear are combined. This means that the CES from both ears are compared. The cumulative distribution for CES evoked from water and air stimuli is shown in Figure 3. Overall, the distributions compared favorably. The geometric means were similar [48.22 and 46.84 degrees/sec for water and air, respectively (Table 5)]. The water CES distribution reflected its underlying monothermal distributions. The tendency for warm water stimuli to provoke stronger caloric responses was countered by the tendency for cool water stimuli to provoke weaker caloric responses. As a result, water and air CES distributions were nearly identical, even though the underpinning monothermal distributions differed.

It is important to keep in mind that the similarity of the CES distributions occurred by design. We calibrated our air stimuli so that air and water distributions were similar. The similarity was enough to expect the Jongkees formula to behave similarly for both media.

There were subtle differences in the variability of water and air CESs, however. Overall, the air distribution showed less variability than the water distribution. Part of the variability in the water distribution came from the spread of warm water-induced MSPVs, certainly on the high side.

The same trend was seen for TESs (Fig. 4). The average TES values for air and water were similar (101 and 97 degrees/sec for water and air, respectively). Yet the probability that we would see an air caloric test TES value of 304 degrees/sec in normal patients would be approximately 1:100. The probability that we would see a water caloric test TES value of 304 degrees/sec would only be approximately 1:33. Again because of the warm water caloric response distribution, the water caloric CES and TES 99% confidence limits were about 24% higher than the comparable air confidence limits. As was the case with the monothermal data, the upper limits of normal differed greatly between air and water stimuli. Thus, the ability to define a hyperactive MSPV depends on the type of caloric stimulus selected. Moreover, the tighter distribution for air stimuli means that it may be easier to distinguish between normal, hyperactive, and hypoactive responses.

It is difficult to compare our results with those earlier reports because of the different statistical methods used in each study. However, in our laboratory, fewer than one in 100 otherwise normal subjects demonstrates a TES less than 27 degrees/sec when water stimuli were used and less than 31 degrees/sec when air stimuli were used. TES of less than 25 degrees/sec were virtually never seen in the normal group, confirming the observations of Jacobson et al. (1993) and Hain (2006).

We also investigated the effect of age on TES for both air and water stimuli. These data are shown in Figure 5. Our data show no observable trend for a change in average caloric response magnitude. In contrast, Mulch and Petermann (1979) and Aust (1991) identified peak cool water MSPVs between the ages of 30 and 40 yrs and peak warm water MSPVs between the ages of 40 and 50 yrs. Differences in caloric test method and statistical analysis of the data may explain this difference. Hajioff et al. (2000) and Davidson et al. (1988) described increasing variability on several ENG tests with age. One study even suggested that ENG was of little value in discriminating between healthy and dizzy elderly patients because of response variability (Hajioff, et al., 2002). Our material included too few data points for patients younger than 50 yrs to completely evaluate this hypothesis. Overall, however, our data did not indicate that caloric test performance changes appreciably with age.

Comparison of Normal and Abnormal Responses

From the above description, we might argue that the caloric response evoked in our laboratory has certain desirable statistical properties. First, the distributions for CES and TES were similar for air- and water-based stimuli. This similarity lays the groundwork for arguing that the performance of both tests is similar and perhaps interchangeable in the clinical setting.

Second, the magnitude of the caloric response from each stimulus type was great enough and the variability was small enough to make it unlikely that the absence of a response would be encountered in an otherwise normal subject. This is a good starting point to show that the caloric response can discriminate between normal and hypoactive ears. However, the ability to discriminate between normal and hypoactive ears depends not only on establishing the probability that a particular observation came from a normal ear, but also on the probability that a particular observation reflects an abnormal ear. To challenge the ability of the caloric test to discriminate between normal versus hypoactive caloric responses, we established a distribution for caloric responses that were measured from patients with documented bilateral vestibulopathy.

As reviewed in the Results section, the crossover points (where the normal and abnormal population distributions meet) reflected best test performance. However, they did not correspond with the confidence limits established in Table 5. Even though the abnormal population had clear evidence for bilateral vestibular weakness, there remained substantial overlap between the normal and abnormal CES and TES distributions. Naturally, in the clinic, we can change the test criterion to minimize the more costly error. To illustrate, in the TES plots in Figure 7, the crossover point occurred at a TES of 54 degrees/sec for water and 63 degrees/sec for air. These were associated with a cumulative probability of 0.11 for water and 0.16 for air (Table 6). Thus, in patients with normal ears, the probability of seeing a water caloric TES value less than 54 degrees/sec was 11%. Similarly, in patients with normal ears, the probability of seeing an air caloric TES value less than 63 degrees/sec was 16%. Thus, the probability of seeing a TES greater than these crossover points is 89% and 84%, respectively.

Taking the alternative perspective, in patients with hypoactive ears, the probability of seeing water caloric TES greater than 54 degrees/sec was 11% and the probability of seeing water caloric TES less than 54 degrees/sec was 89%. Similarly, in patients with hypoactive ears, the probability of seeing air caloric TES greater than 63 degrees/sec was 16%, and the probability of seeing air caloric TES less than 63 degrees/sec was 84%. Because we used a parametric approach to describe these distributions, these probability values reflect the sensitivity and specificity values as well.

If, rather than choosing best test performance, we chose to avoid classifying no more than 5% of normal patients as having abnormally weak vestibular responses (type I error), we would choose a TES limit of 47 degrees/sec for the air caloric test. In doing so, we would misclassify 5% of normal subjects as having bilateral weakness. We would also misclassify slightly more than 24% of patients with bilateral weakness as normal (type II error). Similarly, with water stimuli, we would choose a TES limit of 43 degrees/sec. In doing so, our type II error rate would be slightly more than 30%.

The point is not to establish rigid criteria for normal or abnormal CES or TES. Our data from patients with clear vestibular weakness demonstrated considerable overlap with the normal distributions. One may anticipate that the amount of overlap may be larger when less extreme forms of vestibulopathy are studied. The amount of overlap may vary across laboratories or test techniques. However, in as much as the distributions from normal and abnormal ears did overlap, errors in classification occurred, regardless of the criterion chosen. In certain settings, misclassifying patients with abnormal responses as normal may be more acceptable than misclassifying patients with normal responses as abnormal. In other settings, the opposite approach may be preferable. The point is that setting or changing the criteria for the purposes of clinical decision making can only be meaningfully contemplated when both the normal and abnormal distributions are known. This must be the goal we work toward if the caloric test is to remain useful in the era of evidence-based practice.

What we can clearly state is that, even though the magnitude of our caloric responses were on the stronger side relative to other studies shown in Table 1, normal and abnormally weak caloric response distributions still overlap substantially. With this in mind, it is difficult to see how studies that report average normal caloric response values substantially smaller than ours could be effective in distinguishing between normal and abnormally weak test results. With a weaker average normal response, the normal and abnormal distributions would overlap even more than we show here. Only through careful measurement of test variability, as we have presented, can the performance of the caloric test be assessed with precision. If we are to improve the performance of the caloric test as a clinical tool, we must complete this type of analysis across laboratories to ensure the validity of local test protocols and agree on a more uniform standard of best practice for this test.

Limitations of the Present Work

This study has several weaknesses that need to be recognized. First, we used test results from patients referred to our practice for vestibular testing who subsequently had normal otologic and vestibular findings at the time of testing. This determination was based, in large part, on test results as well as otologic opinion. However, because these patients were referred for vestibular study, they had an increased possibility of having undetected vestibular disease. Consequently, we cannot be sure that these patients represent an unimpaired population, although we believe it is unlikely. Our data compare favorably with those of other reports in the literature. However, these reports may be flawed as well for reasons we have reviewed.

A related limitation is that some of the patients used in this study had anxiety or depression-related disorders. These are known to increase caloric variability—either increasing or decreasing caloric response magnitudes. In particular, because anxiety can increase caloric response magnitudes, the higher response magnitudes reported herein may reflect a subgroup of patients with chronic or test-induced anxiety effects. A definitive study of the caloric response in a large group of healthy individuals would be helpful, particularly in view of the availability of videorecording methods, which may change the performance of the caloric test in ways not currently appreciated.

Procedurally, because the material in this study came from patient records, there were likely several biases in the data. Caloric asymmetry ratios (unilateral weakness and directional preponderance calculations) become unstable when caloric responses are weak (Shepard & Telian, 1996). To avoid this instability, in our practice, caloric responses less than 15 degrees/sec may be repeated or retested with a different caloric stimulus. Similarly, if one caloric response was abnormally strong, resulting in an apparently aberrant unilateral caloric weakness, the test may also have been repeated at the discretion of the examining audiologist. In these cases, only the final reported MSPV was available for consideration in the study.

The study has other limitations as well. The overall sample size in the study, although larger than that in other reports, may still not have been adequate. The curve fit solutions for the warm water caloric responses could have been better. Because this distribution had a greater skew and more variability, more observations may have helped better define its shape.

In discriminating between normal and abnormally reduced caloric responses, several other factors need to be considered. First, this analysis implied that water stimuli allowed greater distinction between normal and weak responses than did air stimuli. However, this should not be taken to mean that water is inherently a better stimulus. First, the effect is largely a reflection of the warm water caloric response. Warm water tends to evoke a stronger but more variable response. This could easily be changed by changing the irrigation method (changing temperature, duration of irrigation, etc.) of the weaker stimuli of either medium (water and air). Moreover, the ability to detect abnormally hyperactive responses may be poorer when warm water is used because of the same warm water response characteristics. We could not assess this in the current study because we could not find an independent method to determine vestibular hyperactivity.

A second reason that water stimuli may seem to perform better than air stimuli was the fact that the group with abnormal responses to water had only 14 subjects. The small number of observations makes this distribution inherently unstable. This is certainly a limitation.

It is important to keep in mind that subjects with bilaterally reduced responses were a heterogeneous group, selected on the basis of rotational tests and clinical impressions only. While many had documented ototoxic-induced bilateral vestibulopathy, diverse etiologies were represented in the group. No effort was made to characterize them or represent them as part of a larger population. They were used for illustrative purposes only. More subtle forms of vestibulopathy may be more difficult to detect with the caloric test. Regardless, we are left with the impression that, in even unambiguous cases of vestibulopathy, misclassifications do occur. Understanding the probability with which misclassifications occur, even in unambiguous cases of bilateral vestibulopathy, will help improve clinical decision making. Future efforts should focus on quantifying errors made when the magnitude of vestibular deficit is less severe.

Finally, we cannot know how representative our data compared with data from other clinics. Currently, the performance of the caloric response is not well-standardized across laboratories. As a result, each laboratory is required to establish its own normal values.

CONCLUSIONS

In this article, we report caloric test distributions from more than 600 bithermal caloric studies, some using air and some water stimuli. To our knowledge, this is the largest data set available in the world literature to date. Moreover, it is the first time air and water distributions have been described using parametric statistical methods. We have shown, in a large group of subjects, that warm and cool water MSPV distributions differ substantially from each other. As a consequence, normal limits for the caloric response are stimulus specific. Because of the way our air stimuli were calibrated, the distributions of warm and cool air MSPVs were similar and had tighter distributions than analogous water distributions. This begs a question: why do we continue to base our water caloric stimuli on the assumptions of Barany's convection theory? From a statistical point of view, our data show that, despite an elegant theory, the actual statistical performance of warm and cool water caloric response is different. It might be wise to rethink how we calibrate the water caloric test and focus on evoking responses that have the desired statistical characteristics rather than sticking to an expectation that is not supported by data. Until then, normal limits for each type of caloric stimuli are necessary.

Using the methods described in this article, the probability of any observed MSPV could be expressed as a probability that the response came from an otologically normal group. Normal limits derived from these models were presented at 90%, 95%, and 99% confidence limits for warm and cool, air and water MSPVs, as well as CES and TES values, respectively. These formed the basis for establishing normal limits for hypoactive and hyperactive responses. The parameters for all of the curves displayed in this article are provided in the Appendix.

Although our results showed differences between air and water stimuli, when combined, CES and TES distributions were similar. Thus, calculations such as directional preponderance and unilateral weakness may also be similar, regardless of the medium of caloric irrigation. From these data, we would expect that calculations of unilateral weakness would be similar across media.

Finally, with the commonplace use of desktop computers and spreadsheet software, the ability to store and analyze caloric responses is within the reach of most clinics. In completing this study, we developed several Microsoft Excel spreadsheets that automate the process of curve fitting and establishing confidence limits for the caloric test. These spreadsheets are available from the authors to interested parties on request.

ACKNOWLEDGMENTS

Editing, proofreading, and reference verification were provided by the Section of Scientific Publications, Mayo Clinic.

Dr Olsholt is the 2006–2007 recipient of the James Russell Crawford and Martha Crawford Endowed Fellowship in Otolaryngology at Mayo Clinic Jacksonville.

Address for correspondence: David A. Zapala, Department of Otolaryngology–Head and Neck Surgery/Audiology, Mayo Clinic, 4500 San Pablo Road, Jacksonville, FL 32224. E-mail: zapala.david@mayo.edu.

Received September 4, 2007; accepted December 20, 2007.

APPENDIX. Calculated variables for the derived plots*

	Normal				Bilateral weak			
	Water		Air		Water		Air	
	Central tendency	SD	Central tendency	SD	Central tendency	SD	Central tendency	SD
MSPV								
Cool	3.08606964	0.534537772	3.265538734	0.617680176	1.147928363	1.249611247	1.588102208	0.989765399
Warm	3.5600546	0.534774125	3.239014813	0.476050474	1.917838358	0.788306122	1.806962766	1.103083803
CES	3.875689965	0.542325001	3.846783754	0.468203444	2.599536494	0.720133998	2.621576227	0.855313492
TES	4.611293153	0.510517459	4.577085163	0.442166652	3.326480862	0.543804598	3.4660505	0.68217746

** To calculate the normal cumulative probability distribution from the above data, use the Microsoft Excel function LOGNORMDIST (a, b, c), where a = observed MSPV, b = central tendency value, and c = SD value. This function returns the normal distribution value from a logarithmic transformed distribution.*
CES, combined eye speed; MSPV, maximum slow-phase velocity; TES, total eye speed.

REFERENCES

Anderson, S. (1995). Caloric irrigators: air, open-loop water and closed-loop water. *Br J Audiol*, *29*, 117–128.

ANSI 3.45–1999. (1999). *American National Standard Procedures for Testing Basic Vestibular Function*. Melville, NY: Acoustical Society of America.

Aschan, G., Bergstedt, M., & Stahle, J. (1956). Nystagmography: recording of nystagmus in clinical neuro-otological examinations. *Acta Otolaryngol Suppl*, *129*, 1–103.

Aust, G. (1991). The effect of age on vestibulo-ocular reactions [German]. *Laryngorhinootologie*, *70*, 132–137.

Baloh, R. W. (2002). Robert Barany and the controversy surrounding his discovery of the caloric reaction. *Neurology*, *58*, 1094–1099.

Baloh, R. W. & Honrubia, V. (1990). *Clinical neurophysiology of the vestibular system*. (2nd ed.) Philadelphia: FA Davis.

Barber, H. O. & Stockwell, C. W. (1980). *Manual of nystagmography* (2nd ed). St. Louis: Mosby.

Barber, H. O., & Wright, G. (1973). Positional nystagmus in normals. *Adv Otorhinolaryngol*, *19*, 276–283.

Benitez, J. T., Bouchard, K. R., & Choe, Y. K. (1978). Air calorics: a technique and results. *Ann Otol Rhinol Laryngol*, *87*, 216–223.

Blakley, B. (2000). Toward a standard for electronystagmography. *J Otolaryngol*, *29*, 335–336.

Bock, O. & Zangemeister, W. H. (1978). A mathematical model of air and water caloric nystagmus. *Biol Cybern*, *31*, 91–95.

Brown-Séquard, C. -E. (1860). *Course of lectures on the physiology and pathology of the central nervous system*. Philadelphia: Collins.

Burkard, R. (2006). When standards and clinical practice collide. *Acoust Today, 1*, 50–51.

Capps, M. J., Preciado, M. C., Paparella, M. M., et al. (1973). Evaluation of the air caloric test as a routine examination procedure. *Laryngoscope*, *83*, 1013–1021.

Coats, A. C., Hebert, F., & Atwood, G. R. (1976). The air caloric test: a parametric study. *Arch Otolaryngol*, *102*, 343–354.

Custer, D. D., Black, F. O., & Hemenway W. C. (1973). The sequential bithermal binaural caloric test. I. A statistical analysis of normal subject responses. *Ann Otol Rhinol Laryngol, 82* (suppl), 3–8.

Davidson, J., Wright, G., McIlmoyl, L., et al. (1988). The reproducibility of caloric tests of vestibular function in young and old subjects. *Acta Otolaryngol*, *106*, 264–268.

Ford, C. R. & Stockwell, C. W. (1978). Reliabilities of air and water caloric responses. *Arch Otolaryngol*, *104*, 380–382.

Frenzel, H. (1959). On the problem of a routine technic of thermal vestibular stimulation [German]. *HNO*, *7*, 353–358.

Geisler, C., Bergenius, J., & Brantberg, K. (2000). Nystagmus findings in healthy subjects examined with infrared videonystagmoscopy. *ORL J Otorhinolaryngol Relat Spec*, *62*, 266–269.

Greven, A. J., Oosterveld, W. J., Rademakers, W. J., et al. (1979). Caloric vestibular test with the use of air. *Ann Otol Rhinol Laryngol*, *88*, 31–35.

Hain, T. C. (2006). *Central hearing loss [monograph on the Internet]*. Chicago: Chicago Dizziness and Hearing. Available at: http://www.dizziness-and-balance.com/disorders/hearing/cent_hearing.html. Accessed February 10, 2007.

Hajioff, D., Barr-Hamilton, R. M., Colledge, N. R., et al. (2000). Re-evaluation of normative electronystagmography data in healthy ageing. *Clin Otolaryngol Allied Sci*, *25*, 249–252.

Hajioff, D., Barr-Hamilton, R. M., Colledge, N. R., et al. (2002). Is electronystagmography of diagnostic value in the elderly? *Clin Otolaryngol Allied Sci*, *27*, 27–31.

Hamersma, H. (1957). The caloric test: a study of caloric tests in normal subjects. Amsterdam: University of Amsterdam.

Henriksson, N. G. (1956). Speed of slow component and duration in caloric nystagmus. *Acta Otolaryngol Suppl*, *125*, 1–29.

Henry, D. F. (1999). Test–retest reliability of open-loop bithermal caloric irrigation responses from healthy young adults. *Am J Otol*, *20*, 220–222.

Imai, T., Sekine, K., Hattori, K., et al. (2005). Comparing the accuracy of video-oculography and the scleral search coil system in human eye movement analysis. *Auris Nasus Larynx, 32*, 3–9.

Jacobson, G. P., Newman, C. W., & Peterson, E. L. (1993). Interpretation and usefulness of caloric testing. In: G. P. Jacobson, C. W. Newman, J. M. Kartush (Eds.), *Handbook of balance function testing* (pp. 193–233). St Louis: Mosby Year Book.

Jongkees, L. B., Maas, J. P., & Philipszoon, A. J. (1962). Clinical nystagmography: a detailed study of electro-nystagmography in 341 patients with vertigo. *Pract Otorhinolaryngol (Basel)*, *24*, 65–93.

Jongkees, L. B. & Philipszoon, A. J. (1964). Electronystagmography. *Acta Otolaryngol Suppl*, *189* (Suppl), 1.

Jung, R. & Mittermaier, R. (1939). Zur objektiven registrierung und analyse verschiedener nystagmusformen: vestibulärer, optokinetischer und spontaner nystagmus in ihren wechselbeziehungen. *Eur Arch Otorhinolaryngol*, *146*, 410–439.

Karlsen, E. A., Mikhail, H. H., Norris, C. W., et al. (1992). Comparison of responses to air, water, and closed-loop caloric irrigators. *J Speech Hear Res*, *35*, 186–191.

Kobrak, F. (1918). Beiträge zum Experimentellen Nystagmus. *Beiter Anat Physiol Path Therap D Ohres*, *10*, 214–220.

Lion, K. S. & Powsner, E. R. (1950). A new method for ergometric measurements. *Arch Phys Med Rehabil*, *31*, 508–510.

Maes, L., Dhooge, I., de Vel, E., et al. (2007). Water irrigation versus air insufflation: a comparison of two caloric test protocols. *Int J Audiol*, *46*, 263–269.

Mehra, Y. N. (1964). Electronystagmography: a study of caloric tests in normal subjects. *J Laryngol Otol*, *78*, 520–529.

Milojevic, B. (1965). Electronystagmography. *Laryngoscope, 75*, 243–258.

Molina, M. I., Zapata, C., Palma, M. J., et al. (2006). Reference values for the vestibulo-ocular reflex response to the head shaking and the bithermal caloric tests [Spanish]. *Acta Otorrinolaringol Esp*, *57*, 34–40.

Monnier, M. & Hufschmidt, H. J. (1950). Electrooculography and electronystagmography in man [French]. *Helv Physiol Pharmacol Acta*, *8*, C30–C32.

Moon, J. & Munro, K. J. (1996). A comparison of test-retest variability of caloric induced nystagmus in a normal population using an air stimulus presented via a standard and modified irrigating probe. *Br J Audiol*, *30*, 221–226.

Mulch, G. & Petermann, W. (1979). Influence of age on results of vestibular function tests: review of literature and presentation of caloric test results. *Ann Otol Rhinol Laryngol Suppl, 88* (suppl), 1–17.

Munro, K. J. & Higson, J. M. (1996). The test–retest variability of the caloric test: a comparison of a modified air irrigation with the conventional water technique. *Br J Audiol*, *30*, 303–306.

Nijhuis, B. G. & Huygen, P. L. (1980). Single-response variability of air and water caloric reactions. *ORL J Otorhinolaryngol Relat Spec*, *42*, 196–205.

Norre, M. E. & Dierick, A. (1983). Practical value of a simplified air caloric stimulation with freon. *Acta Otorhinolaryngol Belg*, *36*, 370–376.

O'Neill, G. O. (1978). Air as the caloric stimulus. *Br J Audiol*, *12*, 54–58.

Ono, H. & Kanzaki, J. (1976). A new caloric tester using ear canal balloon. *Rev Laryngol Otol Rhinol (Bord)*, *97*, 223–230.

Press, S. E., Goetzinger, C. P., Karlsen, E. A., et al. (1979). A study of five parameters of calorically-induced nystagmus in the clinical situation. *J Aud Res*, *19*, 127–135.

Proctor, L. & Glackin, R. (1985). Factors contributing to variability of caloric test scores. *Acta Otolaryngol*, *100*, 161–171.

Reker, U. (1977). Caloric diagnosis maximum stimulus and suppression of habituation effects. *Arch Otorhinolaryngol*, *214*, 247–256.

Shepard, N. T. & Telian, S. A. (1996). *Practical management of the balance disorder patient*. San Diego: Singular Publishing Group.

Sills, A. W., Baloh, R. W., & Honrubia, V. (1977). Caloric testing 2: results in normal subjects. *Ann Otol Rhinol Laryngol Suppl, 86* (suppl), 7–23.

Sokolovski, A. (1966). The influence of mental activity and visual fixation upon caloric-induced nystagmus in normal subjects. *Acta Otolaryngol*, *61*, 209–220.

Suter, C. M., Blanchard, C. L., & Cook-Manokey, B. E. (1977). Nystagmus responses to water and air caloric stimulation in clinical populations. *Laryngoscope*, *87*, 1074–1078.

Torok, N. (1979). Pitfalls in detecting vestibular decruitment with air calorics. *ORL J Otorhinolaryngol Relat Spec*, *41*, 143–146.

Valli, P., Buizza, A., Botta, L., et al. (2002–2003). Convection, buoyancy or endolymph expansion: what is the actual mechanism responsible for the caloric response of semicircular canals? *J Vestib Res, 12*, 155–165.

van der Stappen, A., Wuyts, F. L., & van de Heyning, P. H. (2000). Computerized electronystagmography: normative data revisited. *Acta Otolaryngol*, *120*, 724–730.

van Egmond, A. A & Tolk, J. (1953). The otometer. *Acta Otolaryngol, 43*, 208–211.

Vitte, E. & Semont, A. (1995). Assessment of vestibular function by videonystagmoscopy. *J Vestib Res*, *5*, 377–383.

Westhofen, M. (1987). Balloon method and water irrigation in thermal vestibular assessment: electronystagmographic comparison of both methods [German]. *Laryngol Rhinol Otol (Stuttg)*, *66*, 424–427.

Wit, H. P., Spoelstra, H. A., & Segenhout, J. M. (1990). Barany's theory is right, but incomplete: an experimental study in pigeons. *Acta Otolaryngol*, *110*, 1–6.

Zangemeister, W. H., & Bock, O. (1980). Air versus water caloric test. *Clin Otolaryngol Allied Sci*, *5*, 379–387.

Zapala, D. A. & Shaughnessy, K. (2005). The ENG caloric controversy: which is the better stimulus: water or air? *Audiol Today*, *17*, 20–21.