Interobserver Agreement by Auscultation in the Presence of a Third Heart Sound in Patients with Congestive Heart Failure*

Alnashir A. Ishmail M.D.; Simon Wing M.D.; John Ferguson M.D.; Tom A. Hutchinson M.D.; Sheldon Magder M.D.; and Kenneth M. Flegel M.D.

Although the third heart sound ($S_3$) is well recognized as an important sign in the evaluation of patients with congestive heart failure, the interobserver variability with its observation needs to be known before general applicability can be determined. Therefore, we determined the agreement among four trained observers on the presence of $S_3$ in 81 hospitalized patients. Agreement between pairs of observers varied between 48 and 73 percent. The kappa statistic, which adjusts for agreement by chance alone, showed that agreement between various observer pairs was moderate ($kappa = 0.40-0.50$) at best and slight ($kappa = 0.10-0.30$) at worst. The rate of agreement did not appear to be affected by the time interval between measurements, by the sex of the patient or by a training effect over the time of the study. In conclusion, although $S_3$ may be important as a clinical sign, clinicians cannot agree reliably about whether or not it is present.

The third heart sound ($S_3$), first described by Potain in 1856,1 has long been recognized as a sign of ventricular disease.1−8 Harvey and Stapleton,3 for example, noted that the presence of an $S_3$, even in the absence of other signs of cardiac decompensation, identified patients who are prone to develop pulmonary edema or congestive decompensation during or following a subsequent surgical procedure. Furthermore, preoperative salt restriction and the administration of diuretics improved the ability of these patients to withstand surgery. This was studied more systematically by Goldman et al8 who calculated a multifactorial index of cardiac risk during noncardiac surgery and showed that the presence of an $S_3$ is one of the principal signs associated with a greater risk of cardiac complications. The $S_3$ has also been shown to be the best predictor of a response to digoxin in patients with congestive heart failure.7 Despite the importance of $S_3$, it was our impression that there is often considerable disagreement among physicians about the actual presence of an $S_3$ in individual patients. It is therefore crucial to know the interobserver variability of an $S_3$ before its general applicability can be determined. This is particularly true in situations where patients are cared for by a team of physicians. There are no data present on this point, but from the report by Raftery and Holland8 of poor agreement among physicians on the auscultation of the first and second heart sounds, we predicted poor agreement for the more difficult $S_3$. We therefore quantitated the interobserver variability of the $S_3$.

METHODS

Selection of Patients

Consecutive patients admitted to the Royal Victoria Hospital for complaints related predominantly to the cardiovascular system were selected for this study. Each patient was examined by the observers on the same day. A total of 98 patients were enrolled, of whom 81 were seen by at least two observers and form the data base. Their diagnoses are summarized in Table 1. There were 49 men and 32 women. Forty percent were diagnosed as having congestive heart failure (CHF), 47 percent as having other cardiac disease and the rest, noncardiac diseases.

Observers and Study Protocol

There were four observers: a board-certified internist, a trainee internist, a trainee cardiologist, and a board-certified cardiologist. They were not aware of the patients’ past medical histories or the reasons for admission to hospital. They were asked to examine the patients independently for the presence of $S_3$ and to decide whether

<table>
<thead>
<tr>
<th>Ward Diagnosis</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Congestive heart failure</td>
<td>37</td>
</tr>
<tr>
<td>Myocardial infarction</td>
<td>11</td>
</tr>
<tr>
<td>Valvular heart disease</td>
<td>5</td>
</tr>
<tr>
<td>Idiopathic cardiomyopathy</td>
<td>3</td>
</tr>
<tr>
<td>Pulmonary embolus</td>
<td>5</td>
</tr>
<tr>
<td>Angina pectoris</td>
<td>13</td>
</tr>
<tr>
<td>Other cardiac disease</td>
<td>7</td>
</tr>
<tr>
<td>Noncardiac diseases</td>
<td>12</td>
</tr>
</tbody>
</table>

*Patients may have more than one ward diagnosis.
S₃ was definitely present, probably present, or definitely absent (in most cases, the observers opted only to auscultate the heart). It was not possible for each observer to examine all 91 patients, but each patient was examined by at least two observers. The maximum number of patients seen by any observer was 65 (board-certified internist) and the minimum number was 38 (trainee cardiologist).

The examinations were conducted in the hospital wards which include one- to four-bed rooms. Excessive noise was reduced such as closing windows, turning off airconditioners, or asking other people in the room to stop talking, but a completely silent room was not obtained.

Observers all used the standard approach to auscultating S₃. This included identifying the apical impulse and auscultating over that area, turning the patients on their left side if S₃ was not heard in the supine position, and using the bell of the stethoscope pressed gently on the chest. If a patient could not be properly positioned, or the apex could not be felt because of a dressing (eg, postthoracotomy), the patient was not included. The brand of stethoscope was not standardized, but all observers used good quality units. There was no systematic use of provocative measures such as handgrip, leg raising or effort, but these were often used by observers in ambiguous cases.

Data Analysis

The data were organized in pairs so that the agreement between any pair of observers could be analyzed. Each pair was organized in the form of a three by three table. The data in each of the six pairs were analyzed by three specific statistical indexes as described by Kramer and Feinstein. The first was crude percentage agreement. Next, we calculated the weighted percentage agreement as the sum of 1 = (x – N) / N(1 – N) where x represents the number in each cell of the agreement matrix, w represents the corresponding assigned weights, g represents the number of categories contained in the scale, and N equals the number of subjects compared. The values for w were 3 for maximum disagreement (ie, S₃ definitely present vs S₃ definitely absent), 2 for poor partial agreement (definitely absent vs probably present), 2 for good partial agreement (definitely present vs definitely absent), 1 for perfect agreement (absent), and 0 for perfect agreement (definitely present vs definitely absent, etc.). Finally, to correct for agreement by chance alone, we calculated the weighted kappa statistic using the same assigned weight as described above.

Results

Table 2 summarizes the results of the various tests of concordance as described under Methods. The cumulative percentage of agreement showed that any two observers agreed 57 percent of the time on the presence or absence of S₃ in a given patient. This value was improved to 68 percent when partial agreement was considered as well. There was considerable variability between pairs, and the poorest concordance was between the two board-certified observers (observer 1 vs 4). However, the above two tests do not take into account any agreement expected by chance alone. The weighted kappa statistic, as described in the Methods section, corrects for chance agreement. According to the criteria of Landis and Koch, the strength of agreement is "poor" when the value of kappa is less than zero, "slight" from 0 to 0.20, "fair" from 0.21 to 0.40, "moderate" from 0.41 to 0.60, "substantial" from 0.61 to 0.80, and "almost perfect" from 0.81 to 1.00. If these criteria are applied to Table 2, the strength of agreement between a given pair was "moderate" at best (three pairs), and "slight" at worst. The cumulative strength of agreement according to these criteria was "fair".

The results of the same tests of concordance calculated for patients with a clinical diagnosis of congestive heart failure are shown in Table 3 and for the remaining patients in Table 4. Although the weighted percentage agreement was greater than 50 percent for all six observer pairs in both groups of patients, the strength of agreement as determined by the weighted kappa statistic was "poor" to "fair" in four out of six observer pairs with congestive heart failure patients and in five out of six observer pairs in patients without congestive heart failure.

We also analyzed the factors which may have affected agreement. First, it is possible that it is more difficult to hear S₃ in women compared to men. We looked for
this possibility by using the data from the board-certified internist as he examined the largest number of patients. Out of 29 women examined, 15 had an audible S3 (50 percent), whereas 21 of 36 male patients had an audible S3 (56 percent) (p>0.1). In the same patients, 14 women (50 percent) and 23 male patients (64 percent) were noted to have an audible S3 according to any one of the other three observers (p>0.1). A similar result was obtained when the observations of the trainee internist (63 patients) were analyzed. The agreement between observers about the presence of S3 in male or female patients was comparable to the agreement observed when the sex of the patient was not considered. If anything, the agreement was marginally better with female patients.

A second factor for poor concordance might be that there was a “training effect” over the course of the study so that physicians changed their likelihood of reporting an S3 over the time of the study. This did not appear to be the case, for in 22 of the first 33 patients, the board-certified internist reported an S3, whereas he reported an S3 in 13 of the next 32 patients. In both halves of the study, the number of patients with a ward diagnosis of CHF was similar (13 vs 14). This difference in detection rate is not statistically significant (p>0.1). A similar result was obtained when the 63 observations of the trainee internist were examined. Furthermore, comparison of agreement between these two observers during the two periods of the study yielded virtually identical values for the three tests of concordance described earlier.

A third reason for poor agreement may be the fact that the audibility of an S3 varies throughout the day, and therefore, the time interval between observations is important. The average time interval between examinations, where any two observers agreed that S3 was definitely present, was 204±168 minutes (mean ± SD), which was not significantly different from 211±148 minutes when they agreed it was definitely absent. When there was maximum disagreement between two observers (ie, definitely present vs definitely absent), the time interval was shorter, 206±133 minutes. Thus, it is unlikely that the length of time between measurements was a significant factor in the low agreement for detection of an S3.

**DISCUSSION**

Phonocardiographic studies detect low frequency subaudible vibrations following S3 in almost all patients. To be clinically useful, though, these vibrations must be loud enough to be appreciated by the human ear. Shah and Jackson have stated that S3 recorded at greater than 50 cps and 24 db roll are almost always heard, but this has not been rigorously tested. These workers have also shown that an S3 in this range on phonocardiography detects patients with elevated left ventricular end diastolic pressures and decreased cardiac output. The S3 should thus be a useful clinical sign. However, the poor interobserver agreement we observed of the presence of S3 may limit its usefulness. Specifically, three out of the six observer pairs had a fair agreement or worse when overall agreement was assessed. To put the results in a more easily interpretable perspective than is provided by overall agreement or by the kappa statistic, we found that if one observer heard S3, the probability that a second observer would was between 34 and 38 percent. In contrast, physicians had an easier time agreeing that S3 was absent. When one physician said that S3 was definitely absent, the chance that a second observer would agree was between 69 and 79 percent. It should be noted that our study does not address the question of the ability of subjects to actually identify the presence of an S3, but rather the interobserver variability for we felt that this was the more important clinical question. Therefore, no attempt was made to compare the clinical detection of S3 to what could be recorded phonocardiographically.

Our failure to demonstrate a more substantial agreement between observers, particularly about the presence of an S3, could be due to several factors including the presence of breast tissue, time interval between the assessments and experience of the observer. Female breast tissue can make auscultation of an already soft S3 more difficult. However, observers did not record S3 more frequently in male patients as compared to female patients. The time interval between observations also did not appear to bias the findings of this study for we only accepted comparisons which were done within a few hours of each other. The level of training did not appear to make a difference either, for the worst agreement was between the board certified cardiologist and board certified internist. Nor, did we observe a change in frequency of S3 identification over the course of the study which indicates that a training effect did not occur.

We must conclude that the major reason for the poor agreement was simply the difficulty in discriminating S3 from the background noise because the “signal to noise ratio” was very small. This finding is not surprising because poor interobserver agreement on a number of clinical signs has been well documented and agreement on the presence of signs in cardiovascular examination is poor even with those signs that clinicians ought to be most expert, namely: the first and second heart sounds.

Factors which can affect the ability to detect S3 can be classified as “intrinsic” and “extrinsic” to the examiner. Intrinsic factors include the skill and experience of the examiner, the concentration of the examiner, and the emotional and psychological state of the examiner—all of which will effect an examiner's
ability to detect a stimulus. Since the purpose of our study was to compare our abilities to detect $S_3$, we would predict that these intrinsic factors were probably more in favor of detecting $S_3$ than during routine examinations.

Extrinsic factors include the noise in the room, the ability to manipulate the patient, and the type of stethoscope. By conducting our study on the wards, we hoped to keep these conditions as close as possible to do conditions the average physician faces daily.

Despite the difficulty in auscultating $S_3$, its association with increased left ventricular filling pressures and decreased cardiac output have been clearly established. The $S_3$ has also been shown to be useful for making important clinical decisions. For example, Lee et al. noted that an $S_3$ was the best predictor of the potential for digoxin to improve the symptoms and signs of patients with congestive heart failure. It has, therefore, been recommended that digoxin be given to patients whose congestive heart failure is controlled (ie, no rales or increased shortness of breath) with diuretics alone, but in whom an $S_3$ persists. In another study, Goldman et al. developed a multifactorial index of cardiac risk during noncardiac surgery, and found that the presence of $S_3$ was the most significant predictor of risk. The $S_3$ is thus clearly an important clinical sign when it is audible to the observer.

The fact that the observer variability is poor and yet the sign is clinically useful appears to be paradoxical. That is, if clinicians cannot agree on the presence of $S_3$, how can it be a good predictor of anything? We feel that this can best be resolved by considering the impact on the examining physician of the total gestalt of a patient. When examining a patient with a large heart and a history of heart failure, as was done by Goldman et al., the physician may have a greater expectation that the patient will have an $S_3$. The likelihood of detecting an $S_3$ would thus be influenced by the clinician’s estimate (an intrinsic factor) of the severity of the case. In our study, the observers did not have this extra information and could only rely on their auscultation. It would thus be important to determine if the findings of Goldman et al. and Lee et al. are still valid when the observer is only allowed to listen for an $S_3$.

In conclusion, although $S_3$ is an important clinical sign, poor physician agreement on whether it is actually present or not reduces its general applicability.

**References**

5. Shah PM, Gramiak R, Kramer DH, Yu PN. Determinants of atrial ($S_1$) and ventricular ($S_3$) gallop sounds in primary myocardial disease. N Engl J Med 1968; 278:753-58
10. Landis RJ, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-74