Box and whisker plots for graphic presentation of audiometric results of conductive hearing loss treatment

PAUL J. GOVAERTS, MD, THOMAS SOMERS, PhD, and **F. ERWIN OFFECIERS, PhD,** Antwerp-Wilrijk, Belgium

The Committee on Hearing and Equilibrium of the American Academy of Otolaryngology–Head and Neck Surgery has published guidelines for the reporting of audiometric results of middle ear interventions. It recommends the reporting of several audiometric variables by means of two summary parameters: means and standard deviation. This article advocates the use of other summary statistics, namely the median, quartiles, and extremes, because they do not require a normal distribution of the audiometric data and they are not sensitive to variations of the extreme values. On the basis of the exploratory data analysis, we propose a graphic method to present the Committee's variables in terms of their summary statistics. This "multiple box and whisker plot" offers a detailed and accurate overview of six variables in one graph. (Otolaryngol Head Neck Surg 1998;118:892-5.)

The reporting of treatment results in middle ear surgery is not yet standardized. Many reports use different parameters to summarize the audiologic results and different statistical tests for their analysis. This renders the comparison of different papers nearly impossible.

To improve this situation, the Committee on Hearing and Equilibrium of the American Academy of Otolaryngology–Head and Neck Surgery (AAO-HNS) has recommended clear guidelines.[1] We will comment on these guidelines and suggest a graphic method to facilitate the presentation of the recommended reporting parameters.

The aim of the AAO-HNS guidelines is to introduce a protocol for reporting otologic results that may become widely applicable and that is simple to use. The guidelines consider both audiometric and disease variables. This article will address only the audiometric results.

WHICH LEVEL TO CHOOSE

The Committee has covered two levels: level 1 for the reporting of summary data and level 2 for the reporting of raw data. The aim of any report is to transfer detailed information of a specific population (or sample) to another. The most accurate way for a surgeon to do so is to provide his or her correspondents with a detailed list of all patients and their individual preoperative and postoperative hearing levels. This may be feasible when reporting on a small number of patients, but it becomes far less interesting as this number increases. Therefore summary statistics have been developed in an attempt to summarize the original population by means of just a few parameters in such a way that, if only these parameters are transferred, the correspondent is able to reconstruct the original population. In consequence, when reporting on a series of patients, the use of summary statistics is preferred. The idea of stimulating the authors to make the raw data available on request seems good.

WHICH SUMMARY PARAMETERS TO USE

The Committee recommends that the postoperative air-bone gap, the number of decibels of closure of the air-bone gap, and the change in high-tone bone-conduction level be reported in terms of mean, standard deviation, and range. The choice of these summary statistics (mean, standard deviation, range) may be controversial. As said before, the aim of the summary statistics is to enable reconstruction of the original population. The choice of mean and standard deviation suggests a normally distributed population (gaussian population) because only such a population can be reconstructed by these two parameters. However, audiometric data are often not normally distributed, but instead show a Poisson or a bimodal distribution. This is illustrated in Fig. 1. As an example, we have analyzed 15 randomly chosen patients with otosclerosis. Twenty-two variables were collected, namely the preoperative and postoperative air-conduction levels at 250, 500, 1000, 2000, 4000, and 8000 Hz and the preoperative and postoperative bone-conduction levels at 250, 500, 1000, 2000, and 4000 Hz. Shapiro-Wilk's *W* test was

From the University Department of Otolaryngology, St. Augustinus Hospital, University of Antwerp.

Reprint requests: Paul J. Govaerts, MD, MS, University Department of Otolaryngology, St. Augustinus Hospital, Oosterveldlaan 24, B-2610 Antwerp-Wilrijk, Belgium.

0194-5998/98/$5.00 + 0 **23/77/80572**

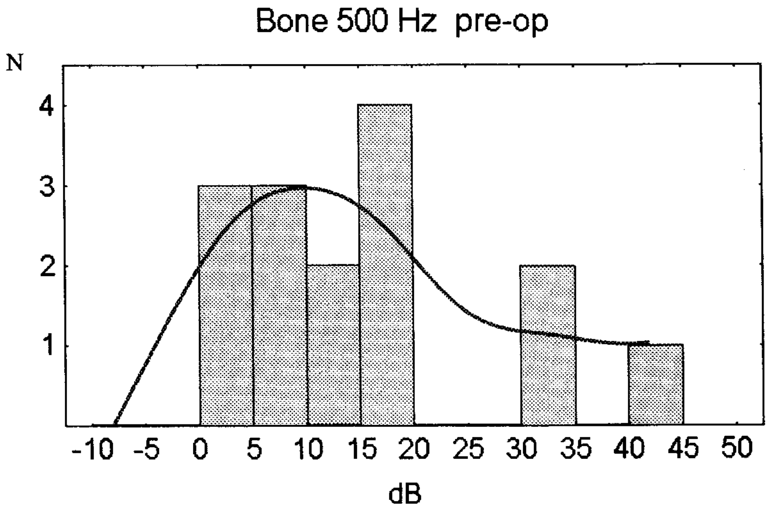

Fig. 1. Histogram of the preoperative bone-conduction levels (average 0.5, 1, and 2 kHz) of 15 randomly chosen patients with otosclerosis. Line represents best fit, which is obviously not a gaussian curve, but rather a Poisson distribution.

performed on each variable to check the normality of its distribution. Of these 22 variables 6 did not have a normal distribution. Summarizing those variables by means of mean and standard deviation would result in the erroneous reconstruction of a normal population.

Although the central limit theorem states that the distribution of a large sample ($N > 30$) from a nonnormally distributed population will be approximately normal itself,[2] we believe that it is tricky to assume a normal distribution of audiometric data without explicitly testing for it. In addition, the mean and the standard deviation are very sensitive to variations at the extreme ends of the population. For audiometric results, extreme values may be caused by several situations, ranging from extreme and rare physiologic or surgical conditions to input errors by the operator, who has to enter a vast amount of numeric data. The impact of the extremes on the mean is demonstrated in Table 1. Although the extreme values of audiometric results may be very important, we would not like to have them exert too strong an influence on our summary statistics. Otherwise, for example, a single case of deafness caused by labyrinthine invasion by a cholesteatoma would strongly deteriorate the reported global results of the surgery. In addition, there is no consensus on how to code for deafness. Quantifying it as 80 dB (as in the case of bone conduction) would have a different impact on the mean than quantifying it as 120 dB (as in the case of air conduction).

Therefore it is our conviction that other summary parameters should be used. Adequate parameters were introduced by Tukey.[3] His parameters are based on the sorting of the cases in ascending order. The rank of a case is defined by its position in the row. The following parameters are defined for a sample with N cases: lower extreme (value of the case with rank 1); upper extreme (value of the case with rank N); median (value of the case with rank [N + 1]/2); lower quartile (value of the case with rank [N + 1] × 1/4); upper quartile (value of the case with rank [N + 1] × 3/4). In a normal distribution these parameters correspond to the percentiles P0, P100, P50, P25, and P75, respectively. An example is given in Table 2.

Table 1. Impact of extreme values on the mean

	Set 1	Set 2
	10	10
	10	10
	10	10
	10	10
	60	160
Mean	20	40
Median	10	10

Two sets of five cases each are listed. Cases have the same values in both lists except for case 5, which has a value of 60 in set 1 and of 160 in set 2. Because of this difference, the mean value of set 2 is twice that of set 1. In contrast, the median is not influenced. The median therefore is more robust and less influenced by extreme values of the distribution.

In addition, outliers are values so extreme that they obviously do not belong to the main population and should be considered to be beyond the reasonable margins of the population. The reasonable margins are called "fences" and are situated at both sides of the distribution. The fences are readily defined as follows: if

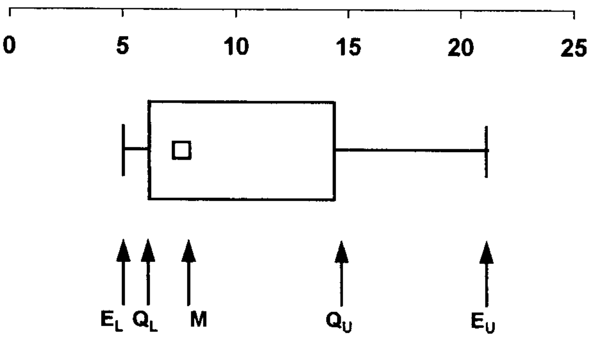

Fig. 2. Box and whisker plot of sample in Table 2. E_L, Lower extreme; Q_L, lower quartile; M, median; Q_U, upper quartile; E_U, upper extreme.

Table 2. Tukey's parameters

	Rank	Value
	1	5
	2	6
	3	7
	4	7
	5	8
	6	11
	7	16
	8	22
Summary parameters		
Median	4.5	7.5
Lower extreme	1	5
Upper extreme	8	22
Lower quartile	2.25	6.25
Upper quartile	6.75	14.75

Table shows a sample of eight cases—the values of which are sorted in ascending order—and the corresponding ranks. Ranks and values of the five summary parameters according to the exploratory data analysis are depicted.

Q_L = lower quartile and Q_U = upper quartile, then FL (lower fence) = $Q_L - (Q_U - Q_L) \times 1.5$ and F_U (upper fence) = $Q_U + (Q_U - Q_L) \times 1.5$. Any value beyond one of these fences is called an outlier.

These parameters are valid to describe any type of population, regardless of whether it is normally distributed. They describe the population in as many details as possible, and they are not sensitive to changes at the extreme sides of the population. If the distribution happens to be normal, the mean and standard deviation can be easily calculated: the mean equals the median and the standard deviation equals $(Q_U - Q_L)/1.35$.

HOW TO PRESENT THE SUMMARY STATISTICS

The population is described by five parameters: median, lower and upper extremes, and lower and upper quartiles. Tukey suggested plotting these five parameters in a box and whisker plot, in which the median is plotted as a dot and the box is delimited by the quartiles and the whiskers by the extremes (Fig. 2). If outliers exist, they are plotted as separate dots, and the whisker is modified to reach to the value closest to but still inside the fence. In this way the box and whisker plot gives a graphic view as accurate as possible of a population that does not necessarily have a normal distribution. The central tendency is read from the central dot depicting the median, the dispersion is read from the box depicting the quartile range, the margins are read from the whiskers depicting the extremes, the outliers are read from the separate dots beyond the whiskers, and the symmetry or asymmetry of the distribution is easily estimated. Most commonly used software packages for data management, such as databases, spreadsheets, and statistics, have a facility to create box and whisker plots in an easy way.

Several box and whisker plots can be mounted in one graph, providing a clear, detailed, and easy to read overview of several variables and parameters representing a tremendous amount of raw data. We propose a multiple box and whisker plot to become part of the guidelines of the Committee of the AAO-HNS. This multiple box and whisker plot should be composed of six box and whisker plots representing six different variables. The first three plots should always represent the variables' preoperative air-conduction levels, preoperative bone-conduction levels, and change in bone-conduction levels. The first two variables represent the preoperative audiometric state of the population, and the third variable is a measure for overclosure and presumed operative damage. The other three variables may be chosen freely by the author to represent any other variable, such as the preoperative air-bone gap, the postoperative air-bone gap, the postoperative air conduction, the number of decibels of closure of the air-bone gap, and so forth. An example of such a multiple box and whisker plot is given in Fig. 3.

PROPOSAL

Combining the guidelines of the Committee on Hearing and Equilibrium of the AAO-HNS with the considerations of the exploratory data analysis, we propose to add the use of multiple box and whisker plots to the recommendations. A single graph would thus present the following variables in terms of median, quartiles, extremes, outliers, and symmetry: (1) preoperative bone-conduction level = mean of the thresholds at frequencies 0.5, 1, 2, and 3 kHz; (2) preoperative air-conduction level = mean of four-tone thresholds (frequencies 0.5, 1, 2, and 3 kHz); (3) the change in high-tone bone-conduction level = the preoperative minus the postoperative high pure-tone

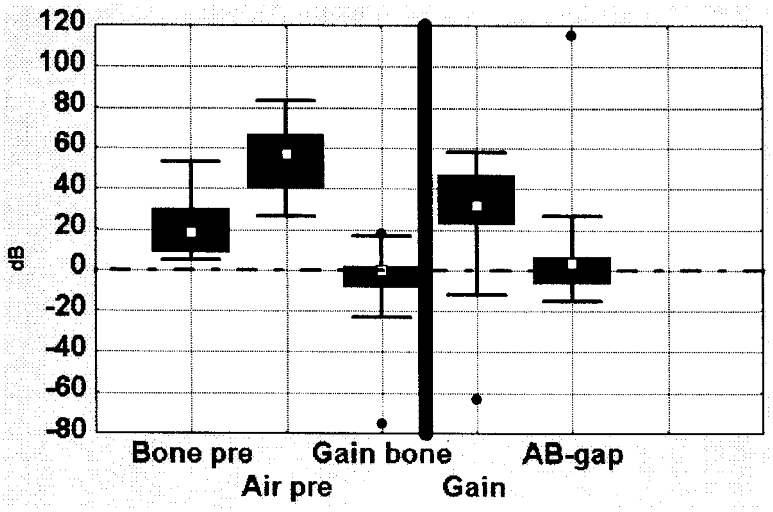

Fig. 3. Multiple box and whisker plot representing five variables from an imaginary study population: (1) preoperative bone-conduction level; (2) preoperative air-conduction level; (3) change in bone-conduction level; (4) gain of air-bone gap, which is the number of decibels of closure of the air-bone gap; and (5) postoperative air-bone gap. Median preoperative bone-conduction level and air-conduction level are 20 and 55 dB, respectively. After intervention, the median gain in bone-conduction level is 0 dB with a small dispersion. However, it is immediately clear that one case has a drop in bone-conduction level of almost 80 dB. This is an outlying value, and it is visualized as a separate dot. Advantages of the box and whisker plot are obvious: the outlier does not influence the position of the median, and it cannot be left off or hidden in the bulk of statistical data.

bone-conduction average at 1, 2, and 4 kHz; (4) postoperative air-conduction level = mean of four-tone thresholds; (5) postoperative air-bone gap = mean of four-tone thresholds for air-conduction minus the same average for bone-conduction determined at the same time; and (6) the number of decibels of closure of the air-bone gap = the preoperative minus the postoperative air-bone gap. The postoperative levels of variable 3 should be determined at 6 weeks or longer after surgery, and those of variables 5 and 6 should be determined at 1 year or longer.

If this multiple box and whisker plot were to become part of every report on the audiometric results of middle ear interventions, the results would gain in detail, accuracy, and comparability with regard to other reports and would be statistically correct.

REFERENCES

1. Committee on Hearing and Equilibrium. Committee on Hearing and Equilibrium guidelines for the evaluation of results of treatment of conductive hearing loss. Otolaryngol Head Neck Surg 113:186-7.
2. Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 5th ed. New York: John Wiley & Sons; 1987.
3. Tukey JW. Exploratory data analysis. Reading (MA): Addison-Wesley, 1977.