planned review of his recent experience with \textit{S} \textit{pneumoniae}, particularly with regard to his HIV population.

David Schlossberg, MD
Jose Miguel Torres, MD
Omar Cardenas, MD
Andres Vasquez, MD
Philadelphia, Pennsylvania

Improper Use of MDIs
Education is the Key

To the Editor:

I was very interested to read the recent article by Fink (December 1997). There is no doubt that many patients fail to use metered-dose inhalers (MDIs) properly, and therefore suffer consequently. I also share the concern raised by the author that many physicians, nurses, and respiratory care practitioners do not have the knowledge or ability to show their patients on the proper use of MDIs. Our group has recently published data on the inability of physicians to use MDIs. In agreement with previous reports, despite their confidence, most physicians in our study did not know the "gold standard" knowledge about the proper use of MDIs. In addition, upon checking, most physicians performed poorly when assessed on the six artificial stages (namely shaking of canister, full expiration, mouthpiece positioning, triggering, inspiration, and breath holding) of the inhalational process. There was also a correlation between the knowledge and actual inhalational performance scores which suggests that physicians can be educated to improve their own performance. Although the guidelines on the use of MDIs are not validated "gold standards," there is no current information on the use of MDIs otherwise. Postgraduate and undergraduate education should therefore focus on the training of physicians, nurses, and respiratory care practitioners on the proper use of MDIs.

Kenneth W. Tsang, MD, FCCP
Associate Professor
Division of Respiratory and Critical Care Medicine
University Department of Medicine
University of Hong Kong
Hong Kong

REFERENCES
1 Fink JB. Flawed paradigms drive aerosol device selection. Chest 1997; 112:1467-1449
2 Crompton GK. Problems patients have using pressurized aerosol inhalers. Eur J Respir Dis 1982; 119(suppl):101-104
8 Jones JS, Holstege CP, Rieksje R, et al. Metered-dose inhalers: do emergency health care providers know what to teach?

Ann Emerg Med 1985; 26:305-311

Using the Kappa Coefficient as a Measure of Reliability or Reproducibility

To the Editor:

Assessment of quality in medical diagnostics can be based on results of autopsies, even though the prevalence of autopsy is declining in most countries. In intensive care, evaluation of the diagnostic procedures in diagnosing of ventilator-associated pneumonia (VAP) is often based on postmortem histologic diagnosis of VAP as the gold standard. A key component in this evaluation is that the autopsy represents an accurate and reproducible technique. Recently a study of the reliability coefficient (kappa), measuring "chance corrected agreement" among four pathologists in the histologic diagnosis of pneumonia was presented in \textit{CHEST} (August 1997) by Corley and colleagues. The kappa coefficient in this study was as good as 0.92. Kappa coefficients range from $-1$ to $+1$, where $+1$ is attained in cases with perfect agreement. Kappa ($\kappa$) is defined as:

$$\kappa = \frac{P(A) - P(C)}{1 - P(C)}$$

where $P(A)$ is the probability of agreement and $P(C)$ is the probability of chance agreement.

Conducting a similar study of histologic diagnosis of VAP by six pathologists in Copenhagen ICUs, with the less impressive kappa coefficient about 0.5, we went through the statistical analysis in the study of Corley and colleagues, but were not able to retrieve the stated kappa coefficient. To us, a more likely value seemed to be about 0.70, depending on the precise definition of $P(A)$ and $P(C)$ in the study.

In our opinion, there are two uncertainties in the study by Corley and colleagues: (1) the definition of $P(A)$ is not entirely clear when more than two rates are involved; and (2) the definition of $P(C)$ is not entirely clear either. It depends on whether or not each rater is supposed to have the same marginal distribution of ratings, and is also strongly dependent on the prevalence of positive ratings.

Furthermore, two basic problems seem to be connected with the use of kappa-like measures of agreement: (1) There is really no clear interpretation of kappa, regardless of the variant chosen. In this respect, kappa of course resembles the usual coefficient of correlation. A sort of universal qualitative scaling is suggested, from poor (kappa near zero) to very good (kappa near 1). But it seems to us that any scaling must be entirely dependent on the subject matter in question. This means that levels of agreement must be defined by a panel of studies, and then used to calibrate kappa values; and (2) There is no natural hypothesis to test: testing for $kappa=0$ is usually not very interesting, since any reasonable rating panel should be able to reach levels of agreement beyond chance "agreement."

Other ways of describing levels of agreement between raters could be through statistical analysis of "patterns of disagreement" or by estimating conditional distributions of ratings, given that a certain rating has been observed by a "random" rater.

It would be of great importance to us if Corley and colleagues could comment, primarily on how the kappa coefficient was