Assessment of Interrater and Intrarater Reliability in the Evaluation of Metered Dose Inhaler Technique*

Shelly L. Gray, Pharm.D.; Alisa C. Nance, B.S.; Dennis M. Williams, Pharm.D.; and Charles C. Pullam, M.S.

Study objective: To determine if a training session using videotaped metered dose inhaler (MDI) performances can result in high Interrater and Intrarater reliability of five evaluators assessing MDI technique. Design: Five evaluators (three pharmacists, two pulmonary fellows) were trained to evaluate MDI technique during a 2-h training session. The training session consisted of verbal instruction and practical experience in evaluating MDI technique using videotaped MDI performances of six nonstudy subjects. After the training session, the evaluators independently observed the same videotaped MDI demonstrations of 14 subjects on two occasions separated by a 7- to 10-day interval. Interrater and intrarater reliability was determined for individual steps by calculating percent agreement and intraclass correlation (ICC) coefficient. Results: Interrater. The Interrater reliability for individual steps ranged from 29 to 86 percent (ICC coefficient=0.13 to 0.81). Steps in which evaluators were in agreement for less than 9 of the 14 subjects were shaking the inhaler before inhalation, exhaling, continuing to inhale slowly, and adequate breath hold. Intrarater: The overall percent agreement by step ranged from 74 to 97 percent. Exhaling to functional residual volume (76 percent) and continuing to inhale slowly and deeply (74 percent) had the lowest overall agreement between the first and second observation day. The consistency of evaluating a step between the two observation days varied considerably depending on the step and evaluator. Conclusions: High Interrater and Intrarater reliability in MDI evaluation is difficult to obtain. Clinicians and researchers involved in MDI evaluation and education should be trained to achieve consistency. A single training session using videotaped MDI demonstrations was not adequate in achieving consistency among evaluators. To improve accuracy of research results, researchers should include at least two evaluators to assess MDI technique or take other measures to show and report reliability. (Chest 1994; 105: 710-14)

I

Inhalation therapy is the primary treatment option for patients with pulmonary disease. The benefits that inhalation treatment offers over oral treatment include a rapid onset of action, a lower incidence of systemic adverse effects and direct administration to the respiratory tract. These advantages of inhalation therapy, however, can be offset if patients do not use proper inhaler technique. Inadequate inhaler technique has been reported to be as high as 62 to 69 percent.1,2 Incorrect technique substantially decreases the amount of drug reaching the functional airway and, therefore, reduces the efficacy of the inhaled agent.3,4

Several factors may account for this high rate of incorrect use. A major problem is that healthcare professionals often do not spend sufficient time educating patients regarding proper use. Verbal instruction is more effective than written instruction in teaching patients proper metered dose inhaler (MDI) technique.5,6 Several studies have shown that many physicians and pharmacists are not familiar with proper MDI technique.7,8 Thus, healthcare professionals may give inadequate and incorrect instructions to patients when they do counsel. Many patients may need repetitive instruction to master proper inhaler technique.10 Steps that pose special difficulty include the coordination of actuation and inhalation, inhaling slowly and deeply, and an adequate breath hold. Furthermore, patients may have physical or mental conditions impeding proper use of an MDI.11,12

Periodic evaluation of MDI technique is essential to maintain proper use and maximize the benefits from this type of drug delivery system.10,13 In addition to these practice related issues, the importance of optimal inhalation technique should be emphasized in clinical research. The measurement of many efficacy parameters for clinical trials in the evaluation of new agents are dependent on good inhaler technique. Thus, study subjects must have adequate inhaler technique to determine the effectiveness of the agent under investigation. A common method of assessing technique is for the clinician or investigator to observe the patient using the inhaler. Many healthcare professionals may be involved in the evaluation and education of MDI technique, so consistent and reproducible evalu-

*From the University of North Carolina, School of Pharmacy, Chapel Hill, NC.

Manuscript received February 9, 1993; revision accepted June 30.

ICC = intraclass correlation; MDI = metered dose inhaler

Evaluation of MDI Technique (Gray et al)
tion of MDI technique among evaluators over time and among multiple evaluators is important.

The evaluation of MDI technique is difficult because of the subjective nature of many of the steps. We used a training program including videotaped MDI demonstrations to assist health professionals to become effective in consistently evaluating MDI technique. By engaging the evaluators in a practice session, we attempted to show through actual MDI demonstrations what constitutes correct technique. The objective of this study was to determine if a brief training program could result in high interrater and intrarater reliability of the evaluation of MDI technique.

Methods

Fourteen patients (9 women and 5 men) who were currently using an oral MDI were included in the study. The patients were recruited from the Pulmonary Medicine Clinic at the University of North Carolina Hospitals over a 4-week period. After informed consent was obtained, the patients were asked to demonstrate their technique without prompting from the investigator. Each patient was instructed to take two inhalations during which time they were videotaped. Videotaped MDI performances were used so that the same performance could be viewed at a second period to determine intrarater reliability. All patients were given standard verbal instructions and were placed in a well-lit room at about the same angle to the video recorder. After their technique was recorded, patients were instructed on proper inhaler technique if any problems were noted.

The list of steps used to define correct inhaler technique is in Table 1. The selection of these steps was based on a literature review, product package inserts, and input from 10 pulmonary physicians.3,14,15 The National Asthma Education Program Expert Panel recognizes three acceptable methods of MDI delivery.15 Although the open mouth technique is considered optimal, the closed mouth technique is acceptable for patients having difficulty. Closed mouth technique was the standard of practice at this institution and, therefore, was considered the only acceptable method.

Three pharmacists and two pulmonary fellows were selected to evaluate MDI technique. Selection of the evaluators was based on their familiarity with and understanding of proper MDI use. The five evaluators participated in a 2-hour training session before study initiation. The training session consisted of verbal instruction and a practice session using videotaped MDI performances. Each step of inhaler technique was discussed in detail, focusing on the minimum criteria that should be met to consider that the step was performed correctly. For steps that consist of two parts, evaluators were instructed that both parts needed to be performed correctly to score the step as correct. After verbal instruction, the evaluators gained experience in evaluating technique by observing videotaped MDI demonstrations of six nonstudy patients. The videotape was stopped after each patient demonstration, and the evaluators scored the performance independently using a standardized score sheet composed of steps listed in Table 1. The scoring of each patient demonstration was then discussed as a group. If there was disagreement among evaluators for a given step, the demonstration was replayed until a consensus was reached regarding the scoring of the step. The agreement among evaluators was not formally assessed at the conclusion of the training session.

The videotape of the MDI performances of the 14 study subjects was then made available to each evaluator. The evaluators were instructed to closely observe the first inhalation of the two-inhalation sequence to become acquainted with the subject's technique. The second inhalation was evaluated and scored using a standardized score sheet. To evaluate steps 1 and 8, the evaluators were also instructed to evaluate the shake before the first inhalation and record the time between inhalations. The evaluator scored the same MDI performances on a second occasion 7 to 10 days later. The evaluators were allowed to view the videotapes only once on each study day. To prevent bias, each evaluator scored the performances independently and was blinded to the other evaluators' scores.

Statistical Analysis

Interrater reliability is the measure of association of the five evaluators' scores when evaluating a subject's MDI technique at one time period. For interrater reliability, percent agreement was calculated by counting the number of times all five evaluators were in agreement for a specific step divided by the number of subjects and multiplied by 100. Overall agreement was calculated as the number of times the evaluators were in agreement divided by the product of number of subjects and number of steps and multiplied by 100.

Intrarater reliability is the measure of variation that occurs within an evaluator when scoring the same MDI technique demonstration on more than one occasion. Similarly, for intrarater reliability, percent agreement was calculated by counting the number of times an evaluator scored a specific step the same on the first and second observation days divided by the number of subjects and multiplied by 100. Overall agreement was calculated as the number of times the evaluator was in agreement on day 1 and day 2 divided by the product of number of subjects and number of steps and then multiplied by 100.

The percent agreement is influenced by the distribution of "correct" and "incorrect" scores. If there is a high proportion of either correct or incorrect scores, then the percent agreement would be high potentially by chance agreement alone. To account for chance agreement, intra- and interrater reliability was also determined by using analysis of variance (ANOVA) techniques to calculate the intraclass correlation (ICC) coefficient (SAS Institute Inc., Cary, NC). The ICC coefficient yields identical results as the kappa coefficient.16,17 The ICC coefficient can range from 0 to 1.0. An ICC coefficient of 0 indicates the reliability is no better than chance, whereas an ICC coefficient of 1.0 indicates excellent (perfect) rater reliability.

Results

Interrater Reliability

The percentage of subjects for which the five evaluators were in total agreement by step is listed

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Shake the inhaler thoroughly</td>
</tr>
<tr>
<td>2</td>
<td>Hold the inhaler upright, neck hyperextended</td>
</tr>
<tr>
<td>3</td>
<td>Exhale normally (to functional residual volume)</td>
</tr>
<tr>
<td>4</td>
<td>Place mouthpiece in mouth, lips closed around mouthpiece</td>
</tr>
<tr>
<td>5</td>
<td>Activate canister in beginning of slow inhalation</td>
</tr>
<tr>
<td>6</td>
<td>Continue to inhale slowly and deeply (over 4 s)</td>
</tr>
<tr>
<td>7</td>
<td>Hold breath at full inspiration for at least 5 to 10 s</td>
</tr>
<tr>
<td>8</td>
<td>Shake inhaler thoroughly between inhalations</td>
</tr>
<tr>
<td>9</td>
<td>Wait at least 1 min between inhalations (this should be timed)</td>
</tr>
</tbody>
</table>

CHEST / 105 / 3 / MARCH, 1994 711
in Table 2 along with the ICC coefficients. The percent agreement varied considerably according to the step, ranging from 29 to 86 percent on the first observation day. There were differences in the percent agreement between the first and second observation days for some steps with no consistent trend for improved or reduced percent agreement on the second day. The highest percent agreement was achieved for steps 2 (86 percent-holding inhaler upright), 8 (79 percent-shaking inhaler between inhalations) and 9 (86 percent-waiting at least 1 min between inhalations). The ICC coefficient for these steps was quite variable. Steps 8 and 9 had an ICC coefficient of 0.81 and 0.78 on the first observation day, respectively, whereas step 2 only had an ICC coefficient of 0.14. An ICC coefficient of less than 0.3 was calculated for five of the nine steps on the first observation day, suggesting poor interrater reliability for these steps.

Intrarater Reliability

The percent agreement and ICC coefficients for intrarater reliability are presented in Table 3. In general, the percent agreement achieved for individual steps was higher within an evaluator than between evaluators (Table 2). The overall percent agreement by step ranged from 74 to 97 percent. Out of all steps in which an ICC coefficient could be calculated (40 out of 45 steps), an ICC coefficient of 0.6 or above indicating good reliability between the first and second observation day was achieved for 68 percent of the steps by individual evaluators. An ICC coefficient of less than 0.3 was calculated for 23 percent of the steps.

The five evaluators differed in their ability to consistently evaluate individual steps of MDI technique. For example, evaluator one had excellent reliability with a 95 percent overall agreement between the first and second observation days. All of the ICC coefficients were greater than 0.6 with the exception of one step. Evaluator one also had perfect reliability as indicated by an ICC coefficient of 1.0 for four of the nine steps. In contrast, evaluator four had an 82 percent agreement overall between the two observation days and an ICC coefficient greater than 0.6 for only four of the steps.

Steps 3 (exhale normally) and 6 (continue to inhale slowly and deeply) had the lowest overall percent agreement, 76 percent and 74 percent, respectively. Not all of the evaluators, however, had difficulty in consistently evaluating these steps. Three of the five evaluators had agreement no better than chance between the two days for step 3 (ICC coefficient = 0.0, 0.06), whereas the remaining two evaluators achieved an ICC coefficient of 1.0 and 0.87.

**Discussion**

Achieving high interrater and intrarater reliability in the assessment of MDI technique is difficult. This study incorporated videotaped MDI performances into a 2-h training session in an attempt to train five health professionals to consistently evaluate MDI technique. Videotaped MDI performances were used to give the evaluators practice in adequately evaluating MDI technique in an interactive session to enhance assessment skills and ulti-

### Table 3—Interrater Reliability: Percent Agreement and ICC Coefficient By Evaluator and Step

<table>
<thead>
<tr>
<th>Step</th>
<th>Evaluator 1</th>
<th></th>
<th></th>
<th>Evaluator 2</th>
<th></th>
<th></th>
<th>Evaluator 3</th>
<th></th>
<th></th>
<th>Evaluator 4</th>
<th></th>
<th></th>
<th>Evaluator 5</th>
<th></th>
<th></th>
<th>Overall, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICC</td>
<td>%</td>
<td>ICC</td>
<td>%</td>
<td>ICC</td>
<td>%</td>
<td>ICC</td>
<td>%</td>
<td>ICC</td>
<td>%</td>
<td>ICC</td>
<td>%</td>
<td>ICC</td>
<td>%</td>
<td>Overall, %</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.0</td>
<td>100</td>
<td>0</td>
<td>57</td>
<td>.77</td>
<td>93</td>
<td>1.0</td>
<td>100</td>
<td>0</td>
<td>57</td>
<td>81</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.0</td>
<td>100</td>
<td>.65</td>
<td>93</td>
<td>0</td>
<td>79</td>
<td>NE*</td>
<td>100</td>
<td>NE*</td>
<td>100</td>
<td>94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.0</td>
<td>100</td>
<td>0</td>
<td>71</td>
<td>.06</td>
<td>64</td>
<td>0</td>
<td>50</td>
<td>.87</td>
<td>93</td>
<td>76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.0</td>
<td>100</td>
<td>1.0</td>
<td>100</td>
<td>.65</td>
<td>93</td>
<td>.77</td>
<td>93</td>
<td>1.0</td>
<td>100</td>
<td>97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.65</td>
<td>93</td>
<td>NE*</td>
<td>93</td>
<td>NE*</td>
<td>93</td>
<td>.29</td>
<td>79</td>
<td>NE*</td>
<td>100</td>
<td>91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.87</td>
<td>93</td>
<td>.15</td>
<td>57</td>
<td>.70</td>
<td>86</td>
<td>.17</td>
<td>57</td>
<td>.48</td>
<td>79</td>
<td>74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.87</td>
<td>93</td>
<td>1.0</td>
<td>100</td>
<td>.73</td>
<td>86</td>
<td>.45</td>
<td>71</td>
<td>.82</td>
<td>93</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.86</td>
<td>93</td>
<td>.59</td>
<td>79</td>
<td>1.0</td>
<td>100</td>
<td>.86</td>
<td>93</td>
<td>.87</td>
<td>93</td>
<td>91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>.46</td>
<td>86</td>
<td>.65</td>
<td>93</td>
<td>.65</td>
<td>93</td>
<td>.82</td>
<td>93</td>
<td>.65</td>
<td>93</td>
<td>91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*NE = ICC coefficient could not be calculated because there was not sufficient variation in the scores.*
imately result in high rater reliability. Despite this training session, individual evaluators still varied in the assessment of MDI technique. These findings corroborate the results of a previous study conducted by O’Connell et al.18 These authors reported that six trials were required before a high interrater reliability was achieved between multiple evaluators.18 Results from these studies carry important implications for clinicians and researchers involved in the assessment of MDI technique.

The direct benefit of using videotaped MDI demonstrations to train healthcare professionals to consistently evaluate MDI technique can not be determined from this study, because we did not use a control group. The interrater reliability observed in this study is similar to that reported by O’Connell et al.18 Although different definitions of correct inhaler technique were used in these two studies, four steps were identified that were defined similarly and thus the interrater reliability could be compared. In general, our ICC coefficients were similar to the kappa coefficients achieved in the first demonstration trial (labeled Trial 2) of the O’Connell study; kappa values of 1.0, 0.12, 0.54, and 0.55 were reported for shaking, exhaling, close lips on inhaler (or hold 2.5 to 5 cm from mouth), and breath holding, respectively.18 The major discrepancy in interrater reliability between these studies occurred for the initial shaking of the inhaler, in which our reliability was markedly lower. Despite the low kappa coefficients after the first trial in the O’Connell et al trial, a kappa score of 0.8 or greater was achieved for all steps by the fifth trial (Trial 6), except for inhaling slowly (κ = 0.6). The authors concluded that the revision of the checklist of steps to be more concise, discussions before each trial, and practice resulted in high interrater reliability after the fifth trial.

One limitation of this study is that we did not directly assess interrater reliability at the end of the training session. Although a consensus was reached for each patient demonstration, this was accomplished as a result of discussion and replaying the demonstration if necessary. The training did not continue until the evaluators were able to agree on a demonstration on the initial evaluation without discussion. Therefore, the evaluators may not have achieved total agreement in assessing MDI technique at the conclusion of the training session, which would explain the low interrater reliability observed during the study. In any case, the low interrater reliability achieved after a single training session illustrates the difficulty in achieving consistency in MDI evaluation among multiple evaluators. In practice, most healthcare professionals involved in inhaler training and evaluation are unlikely to undergo a training session as extensive as used in this trial. Furthermore, agreement may have been higher than what would normally occur because evaluators were aware that they were being evaluated as part of a research study.

Certain steps were more difficult for evaluators to agree on. For example, it was often difficult to determine if a patient exhaled properly. For some patients, their exhalation was not accompanied by detectable chest movement or audible breath sounds; therefore, exhalation was difficult to assess. The other major areas of difficulty were in determining if the patient activated in the first part of the inhalation, continued to inhale slowly and deeply and held their breath for 5 to 10 s. Frequently, it was difficult to distinguish between when a patient stopped inhaling and where the breath held began. Recently, devices have become available to assist in the assessment of MDI technique (eg, aerosol inhalation monitors); however, it is not known whether these are useful in addressing the problems mentioned.

Lack of concentration may explain the low reliability which was observed for shaking the inhaler before the first inhalation. The interrater reliability between the initial shake (57 percent agreement, ICC coefficient = 0.26) was considerable lower than the shake between inhalations (79 percent agreement, ICC coefficient = 0.81). Since the evaluators achieved high reliability for the second shake, it is unlikely that the low reliability of the first shake was a result of difficulty in assessing this step. More probable is that the evaluators were not paying attention in the beginning of the demonstration, or they forgot to record the first shake.

This is the first study to evaluate intrarater reliability. Definite differences existed in the ability of the five evaluators to consistently evaluate MDI technique between the first and second observation days. The reasons for low intrarater reliability for some steps for some evaluators is not clear. The evaluation days were separated by only 7 to 10 days; therefore, it was unlikely that the evaluators forgot the criteria used to judge the steps. Another potential factor influencing intrarater reliability is that the evaluator may have remembered a problem they had in assessing a specific step on the first observation day, made a special point to critically assess that step on the second day, and perhaps changed their score. Additionally, with experience the evaluators may have developed an increased awareness of technique and scored demonstrations differently on the second observation day. Perhaps intrarater reliability will improve as evaluators acquire more experience in evaluating MDI technique, as was demonstrated in a previous trial for intrarater reliability.18
These results indicate the importance of reporting the ICC coefficients (or kappa) along with the percent agreement. The ICC coefficient allows a person to evaluate the actual ability of the evaluators after taking into account the amount of agreement due to chance. One limitation of using ICC or kappa is that it requires some variation in the scoring of subjects to calculate the coefficient. In this study, a high percent agreement (93 to 100 percent) was calculated in five cases; however, the ICC coefficient could not be calculated, because there was no or limited variation in the scoring of the step for all patients. For example, in three cases, an evaluator had 100 percent agreement for a step between the first and second observation day. In all cases, no variation existed in the scores for that specific step, ie, the evaluator scored the step as either correct or incorrect for all subjects for both days; therefore, an ICC coefficient could not be calculated. It is unlikely that the reliability of scoring the step on two different occasions was no better than chance, but rather the step was easier to score, more patients performed the step correctly, or the evaluator was not very discriminating between small differences among the subjects. To answer this question definitively would require a larger sample size with variability in scoring of the step.

In summary, consistent assessment of MDI technique is important in practice and research. The use of videotaped MDI performances in a single 2-hour training session was not successful in achieving high interrater and intrarater reliability. The need for training and assessing evaluation skills of those involved with inhaler education is necessary. The role that inconsistent evaluation and subsequent retraining plays in the high prevalence of suboptimal inhaler technique is not known. Since consistent MDI evaluation is difficult to achieve, researchers should consider incorporating at least two evaluators to assess MDI technique to minimize the influence of observer error and thus improve accuracy of results. Subsequent published reports in this area of practice and research should include some measure of interrater reliability.

REFERENCES
3 Newman SP, Pavia D, Clarke SW. How should a pressurized β-adrenergic bronchodilator be inhaled? Eur J Respir Dis 1981; 62:3-21
4 Lindgren S, Bake B, Larsson S. Clinical consequences of inadequate inhalation technique in asthma therapy. Eur J Respir Dis 1987; 70:93-8
5 De Tullio PL, Corson ME. Effect of pharmacist counseling on ambulatory patients’ use of aerosolized bronchodilators. Am J Hosp Pharm 1987; 44:1802-06
10 O’Bye KA, Jim CK, Gee JP, Cowen ME, Quigley AE. An education program that improves the psychomotor skills needed for metaproterenol inhaler use. DICP 1982; 16:945-48
11 Armitage JM, Williams SJ. Inhaler technique in the elderly. Age Aging 1988; 17:275-78
17 Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. Oxford: Oxford University Press, 1989; 79-95