A METHOD FOR THE CLINICAL EVALUATION OF ANTIEMETIC AGENTS

J. WELDON BELLVILLE, M.D., IRWIN D. J. BROSS, PH.D., WILLIAM S. HOWLAND, M.D.

MANY studies have evaluated antiemetic agents by means of the double-blind technique. As in any clinical pharmacologic study of subjective effects, one of the requisites is that neither the person administering the drug and observing the reaction, nor the patient receiving the drug know the medication given. This is necessary to reduce bias. An excellent review of the double-blind technique and some of the factors influencing clinical evaluation of drugs has recently been published by Modell and Houde.[1] The double blind technique itself is not a method of complete evaluation, nor is it a test in itself. It is a method to reduce bias entering in the experimental situation. The use of a placebo in clinical pharmacology has become standard procedure. The placebo effect necessarily depends on the patient knowing that he received a medication. If the medication is administered without the patient being aware of it, placebo effect is not likely to be observed, and a placebo in this instance merely serves as a blank for the observer.

In addition to equating bias, it is desirable to control those variables that may obscure or misrepresent therapeutic effects. One method of balancing the effects of variables is to try to distribute the recognized variables evenly among the drugs. To further eliminate unrecognized variables, a procedure of randomization is employed. Randomization of drug administration is a requisite of good experimental design. This balances the uncontrollable variables in the experimental situation, so as to be reasonably certain that the effect measured is indeed the drug effect.

Before investigating antiemetics it is worthwhile to consider what questions need be answered. These might be several: Is this a good prophylactic drug; that is, will it prevent nausea and vomiting? Is this a good therapeutic drug; in other words, is it effective in eliminating nausea and vomiting after its onset? These questions are independent of clinical problems such as the desirability of preventing vomiting in a particular patient.

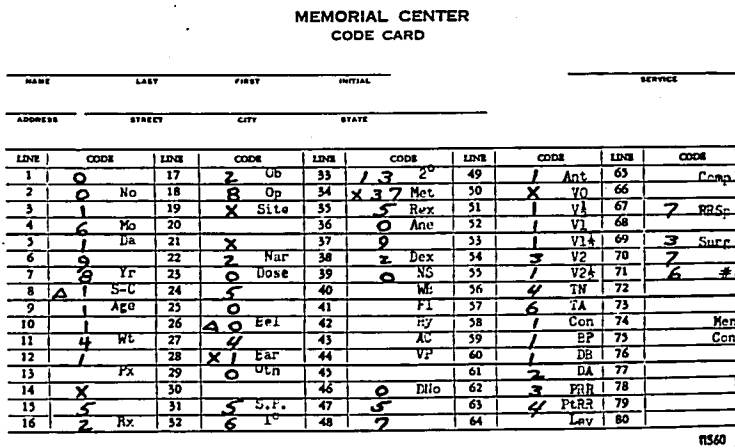
What indices of drug action are to be used? For many years no evidence was considered acceptable which was not the result of objective measurements. It is now realized that subjective verbal reports provide a simple means of clinically evaluating the effectiveness of many forms of therapy. Subjective criteria are used by physicians evaluating drug effects in daily practice. Proper experimental design and statistical analysis will enable subjective information to provide quantitative data with a fairly high degree of accuracy. In our study the scale is somewhat unique in that it is partially subjective and partially objective.

Finally, before any conclusions can be drawn, we must have some assurance that the method employed is sensitive enough to detect the effects under consideration. One measure of this is the ability to show a significant difference between a standard medication and a placebo. Although it is desirable to have a dose effect curve, this may not always be possible.

METHOD

The patients used in this study were those arriving in the postoperative recovery room of the Memorial Hospital for Cancer and Allied Diseases. The group given the antiemetic drugs were among patients of surgeons who had granted permission for routine prophylactic treatment for postoperative nausea and vomiting. Those below 15 years of age and those who had complications, such as bronchospasm, hypotension, or excessive bleeding, during the operation were excluded. All drugs were administered in 1 ml. of solution intramuscularly upon the patient's arrival in the

Accepted for publication May 27, 1959. Dr. Bellville is Assistant Attending Anesthesiologist, Memorial Hospital, and Assistant, Sloan-Kettering Institute; Dr. Bross is Head, Research Design and Analysis Service, Sloan-Kettering Institute, and Dr. Howland is Chairman, Department of Anesthesiology, Memorial Center, and Head, Section of Experimental Anesthesia, Sloan-Kettering Institute, New York, New York.



MEMORIAL CENTER
CODE CARD

NAME LAST FIRST INITIAL SERVICE

ADDRESS STREET CITY STATE

LINE	CODE	LINE	CODE	LINE	CODE	LINE	CODE	LINE	CODE
1	0	17	2 Ob	33	1 3 2°	49	1 Ant	65	Comp
2	0 No	18	8 Op	34	x 3 7 Met	50	X VO	66	
3	1	19	X Site	35	5 Rex	51	1 V1	67	7 RRSp
4	6 Mo	20		36	0 Ane	52	1 V1	68	
5	1 Da	21	X	37	9	53	1 V1½	69	3 Surg
6	9	22	2 Nar	38	2 Dex	54	3 V2	70	7
7	8 Yr	23	0 Dose	39	0 NS	55	1 V2½	71	6 #
8	Δ 1 S-C	24	5	40	WE	56	4 TN	72	
9	1 Age	25	0	41	F1	57	6 TA	73	
10	1	26	4 0 Epi	42	HY	58	1 Con	74	Men
11	4 Wt	27	4	43	AC	59	1 EP	75	Con
12	1	28	X 1 Bar	44	VP	60	1 DB	76	
13	Px	29	0 Oth	45		61	2 DA	77	
14	X	30		46	0 Dilo	62	3 PRR	78	
15	5	31	5 S.F.	47	5	63	4 PtRR	79	
16	2 Rx	32	6 I°	48	7	64	Lav	80	

11560

FIG. 1. Typical study card used in antiemetic studies for entering data on each patient.

recovery room, without regard to his state of consciousness. The drugs were coded and administered in a randomized Latin square sequence [2] by means of a modified envelope technique. One full-time nurse observer noted whether the patient had intra-abdominal or extra-abdominal surgery and the primary anesthetic agent administered. On this basis, the patients were placed into one of six arbitrary groups. As patients were sequentially placed in a particular group, the code number of the drug assigned to the next patient in that group was ascertained from a study book and then this drug was administered. The patients were scored upon arrival in the recovery room and every half hour thereafter for two and a half hours as follows: (x) asleep, (0) no nausea or vomiting, (1) nausea, (2) retching, and (3) vomiting. Patients who were unresponsive but who vomited or retched were scored as vomiting or retching (fig. 1, lines 50–55). This information was entered on the study card along with pertinent data relative to the medical history (lines 13–17), anesthesia (lines 22–45), operation (lines 18–21, 58–61), and postoperative recovery (lines 62–66). The material was later transferred to IBM punch cards.

A parallel group of patients received no drugs and served as controls. These patients fulfilled as nearly as possible all the criteria for the patients given drugs, except that permission of the attending surgeon to administer study drugs was not requested. Since the incidence and severity of nausea and vomiting were not significantly different, the combined group of control plus placebo patients was used for comparisons and preparations of the ridit scale.

The studies were designed so that two dose levels of a standard, test drug and placebo were administered. After a sufficient volume of information on placebo was obtained, the design was modified so that one dose of the standard and two dose levels of each of two unknown drugs were employed. These drugs were prepared in identically appearing ampules and administered under double blind conditions. Twenty or forty drug code numbers were employed at one time so as to decrease observer bias.

INTERROGATING THE PATIENTS

The patients were seen at half-hourly intervals for two and a half hours, and by means of indirect questioning the degree of nausea

and vomiting was scored. To elicit the subjective effect of nausea, they were asked "How do you feel?" or "Is anything bothering you?" If the answer to one of these questions was in the affirmative, then further questioning was necessary such as: "What is bothering you?" The suggestion of nausea was avoided in questioning, which was made as much a part of the postoperative recovery room routine as possible. If patients were retching or vomiting, the scoring was simple.

Measurement of Response

For the majority of patients the report is simple—there was nothing relative to the study to report. However, for the patients who do show some adverse response, the report can be complex—the duration and severity of the symptoms varies widely. Potentially, at least, the detailed basic data might be of value in discriminating between the effectiveness and mode of action of different drugs or in determining optimum dosage levels. The problem is to find measures or indices of response which will make effective use of these basic data.

The construction of measures of patient response is not simple, because we have different and conflicting objectives. We should like an index which is easy to calculate and simple to interpret. One such index would be the proportion of cases where any adverse symptom occurred (at any time period). But we should also like an index which would lose as little information as possible. The proportion of cases (incidence of nausea and vomiting) where any adverse symptom occurred would evidently lead to the discard of much of the detailed information. Hence we are led to consider other ("ancillary") indices which will recover the information lost by the simple proportion.

At first sight a quest for ancillary indices might appear an unnecessary refinement because the patient series numbers in the hundreds and we might therefore suppose that it would not hurt to lose part of the information. However, a closer look makes it clear that the large number of patients is illusionary—the *effective* size of the series is a fraction of this number. For example, with two fairly effective agents tested in 200 patients, the conclusions will hinge on the dozen or so patients who experience adverse symptoms (*i.e.*, the *effective* series size is of the order of magnitude of 20 and not 200). Hence we cannot afford to waste information, and an effort to develop ancillary measures of response seems justified.

Since we do not want ancillary measures that repeat the story that the simple proportion tells us, we can eliminate from consideration those patients who fail to show an adverse response. In effect, we separate the question of effectiveness of an agent into two questions: Does the agent reduce the incidence of adverse symptoms? Does it lessen the severity and duration of the symptoms that do occur? The simple proportion answers the first question, and we now wish a measure of response to answer the second.

There are two different approaches ("strategies") that we might use to construct a second index. The first might be called the "mechanical" strategy and the second the "clinical."

One "mechanical" approach would be an index somewhat analogous to the pain-relief-hour measure currently employed in analgesic trials.[3] We give an arbitrary "score" to the response in each time period (example: vomiting = 3, retching = 2, nausea = 1, none = 0) and then we add the scores for the 5 relevant time periods (since the drug is given at the first time-point the response here would not reflect drug action). This "mechanical" index, called "score," would be 3 for a patient who vomited once and had no other symptoms. The Score would also be 3 for a patient who reported nausea on three occasions (1 + 1 + 1 = 3). Theoretically the scale would run from 1 to 15, but the highest value observed in 292 patients with postoperative nausea or vomiting was 13 (vomiting on 4 occasions and nausea at the other time-point).

It is easy to find fault with such a "mechanical" index. The scores are arbitrary and there are plausible reasons for considering a patient who vomited once to be sicker than a patient who reported nausea on three occasions. The Score itself is a number without special meaning. It depends heavily on protocol details such as the number of time periods and the distinction between retching and vomiting. Although it was not feasible to follow patients

TABLE 1

DEFINITION OF CATEGORIES

I. Nausea reported at one observation.

II. Nausea reported at two observations.

III. Retching observed at one observation and nausea reported at one observation or nausea reported at three observations.

IV. Vomiting observed at one observation or retching observed at two observations.

V. Vomiting observed at one observation plus nausea reported at one observation.

VI. Vomiting observed at one observation plus retching observed at one observation.

VII. Vomiting observed at one observation plus retching and nausea.

VIII. Vomiting observed at two or more observations.

for more than 2½ hours because many of them left the recovery room at this point, a longer follow-up might be possible elsewhere. With changes in protocol the Scores calculated by different investigators would not be comparable. Then, too, the Score is affected by practical problems such as non-response of patients who cannot easily be roused.

The second, or "clinical," strategy would avoid some objections and produce others. In this approach we set up a graded series of categories which ranges from minimal response to responses highly unfavorable from the clinical standpoint. These categories would be operationally defined in terms of the patient report card. The eight categories actually employed in this study are defined in table 1

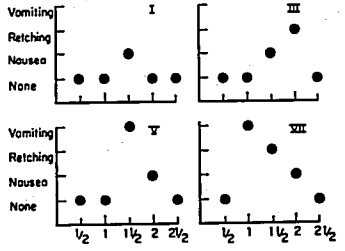


FIG. 2.- Examples of placement of patient response reports into categories.

and some examples are given (fig. 2). In theory the "clinical" approach is more flexible than the "mechanical" in that such things as the *sequence* of response (*e.g.*, steady improvement or worsening of the patient's status) might be taken into account. The clinical ratings, like the Scores, can be readily criticized. The operational definition may not eliminate the factor of personal judgment, so that the ratings might be more subjective than the scores. The issue of "mechanical" versus "judgmental" measures of response is often debated but rarely resolved. A direct empirical test of the measures of response is likely to be a better way to settle the issue.

An obstacle to such an empirical test is the different nature of the two proposed indices (one is numerical, the other consists of categories). To overcome this difficulty both indices were converted to *ridits*, a simple tech-

TABLE 2

CALCULATION OF RIDITS (COMPUTING FORM)

"Score"	(1)	(2)	(3)	(4)	(5)
1	22	11	0	11	.054
2	31	15.5	22	37.5	.185
3	57	28.5	53	81.5	.401
4	28	14	110	124	.611
5	28	14	138	152	.749
6	11	5.5	165	171.5	.845
7	10	5	177	182	.897
8	6	3	187	190	.936
9	6	3	193	196	.965
10	2	1	199	200	.985
11	1	.5	201	201.5	.993
12	0	0	202	202	.995
13	1	.5	202	202.5	.997
Total	203		203		

Instructions:

Column (1): The frequency distribution in the identified distribution (placebo + control reference class).

Column (2): One-half the corresponding entry in Column (1).

Column (3): The cumulate of Column (1) (displaced one category downward).

Column (4): Column (2) + Column (3).

Column (5): The entries in Column (4) divided by the grand total (203). The numbers are the ridits.

TABLE 3

CALCULATION OF RIDITS (COMPUTING FORM)

"Category"	(1)	(2)	(3)	(4)	(5)
I	20	10	0	10	.049
II	30	15	20	35	.172
III	25	12.5	50	62.5	.308
IV	44	22	75	97	.478
V	24	12	119	131	.645
VI	26	13	143	156	.768
VII	16	8	169	177	.872
VIII	18	9	185	194	.956
Total	203		203		

Instructions:

Column (1): The frequency distribution in the identified distribution (placebo + control reference class).

Column (2): One-half the corresponding entry in Column (1).

Column (3): The cumulate of Column (1) (displaced one category downward).

Column (4): Column (2) + Column (3)

Column (5): The entries in Column (4) divided by the grand total (203). The numbers are the ridits.

nique which had previously proved useful in a problem of comparing dissimilar indices.[4, 5, 6] An advantage of the ridit transformation is that arbitrary numbers and classifications acquire a valuable interpretation in terms of probabilities.[7] In ridit analysis a specified series of patients is chosen as the reference set ("identified distribution") and all comparisons are automatically made with respect to this set. A preliminary examination showed that because of their close similarity, control series could be combined with the natural reference set, the placebo series. This combined series was taken as the "identified distribution" and the ridits calculated as shown in tables 2 and 3. A complete description of ridit calculations may be found elsewhere.[7]

Ridit results are conveniently presented in the form of confidence interval graphs. The first step is to calculate the average ridit for each series of patients. If a test agent has no effect (as measured by the given index) the average ridit in the series will be about the same as that in the controls (*i.e.*, 0.50). If the agent has a favorable effect, the average ridit will be less than 0.50 and the departure from 0.50 indicates the strength of the effect (in terms of probabilities). Thus, an average ridit of 0.25 means that a random individual in the test agent series has a probability of 0.25 (or only one chance in 4) of being sicker than a corresponding random individual from the control series. Confidence intervals are an easy way to tell whether differences in average ridits are statistically significant. The interval is calculated by adding to (and subtracting from) the average ridit the reciprocal of $\sqrt{3}$ (number of observations). If two confidence intervals do not overlap, the difference in the averages is statistically significant at the 5 per cent level.

The results for the index based on the sum of the "scores" for the five time periods is shown in figure 3. It will be noted that the average ridit for the 7.5 mg. dose of triflupromazine (Vesprin) is close to the control value of 0.50. However, the 15 and 30 mg. doses of triflupromazine fall considerably below the control value, suggesting some reduction in the severity and duration of the symptoms. However, the confidence intervals tell us that we cannot

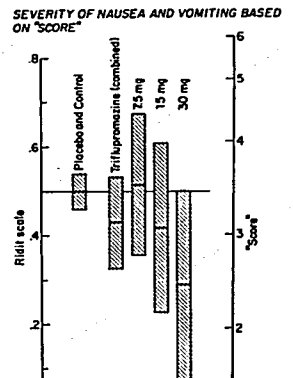


FIG. 3. Severity of postoperative nausea and vomiting based on "scores." Solid line represents mean value and cross-hatched bar represents 95 per cent confidence limits.

be sure that these are real effects because the intervals are so wide that they overlap the control interval. The main reason for this overlap is that the 30 mg. ridit is based on only the 7 patients who became ill.

The results for the index based on "clinical rating" categories are shown in figure 4. It will be noted that the general appearance of this figure is quite like that of figure 3 although the two indices differ in nature and rationale. This time, however, the 30 mg. triflupromazine confidence interval does not quite overlap the control interval, so that there is a significant difference. There is a "moral" to be derived from these results that we suspect applies to many clinical trials. It would be possible to argue the respective merits and demerits of the two indices on "rational" grounds. However, an empirical examination of performance shows that such an argument would be pointless because the two are really so similar.

The conclusion that we draw from the performance of the indices is that they appear to have value. Furthermore, this is an illustration of the practical advantage of the ridit transformation. As previously pointed out,[7] the ridit compensates for unequal lengths of subjective scales and overlapping of subjective scales ("slippage") which is often otherwise a problem. Since the ridit scale based on "score" is the simplest to employ and since it seems to provide valuable ancillary information, we plan to use it in all future antiemetic studies. However, it remains to be seen whether or not it can make an important contribution to the problem of discriminating between effective agents or setting optimum doses.

SEVERITY OF NAUSEA AND VOMITING BASED ON "CATEGORIES"

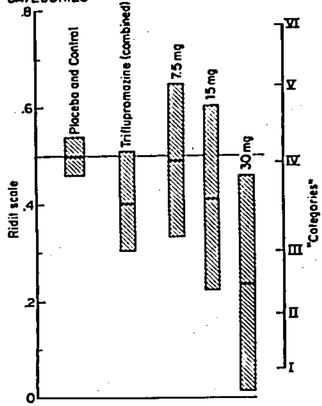


FIG. 4. Severity of postoperative nausea and vomiting based on "categories." Solid line represents mean value and cross-hatched bar represents 95 per cent confidence limits.

RANDOMIZATION

Randomization of the order of drug administration is an essential feature of this protocol. The purpose of this technique is to balance out those variables which have not been otherwise controlled. In combination with other protocol procedures, such as the double-blind technique, randomization represents a kind of "insurance policy" against the systematic biases that have plagued clinical trials in the past. It should not be supposed, however, that any one device is a panacea for the ailments of clinical trials. Hence, it is useful to check back on the methods to see how they would have worked out in actual studies.

Table 4 presents some percentages and averages for relevant variables such as sex, premedication, and primary agent in six drug series. For the most part, the numbers are similar in the different drug series. In other words, the six series are comparable with respect to the factors listed in table 4. Some care is needed in interpreting these results because the series were not collected simultaneously. Initially 30 mg. of triflupromazine was employed but subsequently the 7.5 mg. level was substituted. The placebo series includes data collected prior to the start of the main study. As a consequence it would be possible for time effects to appear in table 4, effects which would not necessarily be balanced out by randomization. The similarity between the drug series suggests that the population composition has not changed markedly in the time interval covered by this study.

In a continuing drug-testing program it would be desirable to make periodic runs on relevant factors as a cross-check on randomiza-

TABLE 4

PER CENT DISTRIBUTION OF CERTAIN VARIABLES AMONG PATIENTS TREATED WITH DRUGS

		Placebo	Cyclizine		Trifluopromazine		
			50 Mg.	100 Mg.	7.5 Mg.	15 Mg.	30 Mg.
Total		331	263	274	203	266	129
Habitus	Obese	18.1	19.4	15.0	15.7	16.5	26.4
	Normal	76.7	75.7	77.4	81.8	77.4	67.4
	Thin	5.1	4.9	7.7	2.5	6.0	6.2
Primary agent	Ether	27.8	26.2	25.5	23.1	24.8	34.1
	Cyclopropane	21.7	21.7	23.0	18.2	21.4	27.1
	Thiopental + N_2O	49.9	50.6	50.3	56.6	51.5	38.7
Premedication	Morphine	18.7	16.7	17.5	21.7	18.8	8.5
	Demerol	79.1	79.1	79.6	76.3	78.9	88.4
	Barbiturate	67.1	66.5	61.7	60.6	60.9	71.3
	Scopolamine	90.0	87.8	87.2	89.6	85.0	84.5
	Atropine	10.0	11.8	12.8	10.3	13.2	14.7
Intubated		45.6	41.8	45.3	42.3	45.1	50.4
Intra-abdominal		19.3	20.1	19.0	19.2	19.5	21.7
Female		70.1	65.4	67.1	71.4	66.2	69.0
White		98.2	95.8	97.4	96.5	99.2	95.3
Average age in years		48.5	50.0	49.9	49.8	50.4	49.2
Average weight in kilograms		66.3	66.1	64.2	63.3	63.9	67.7

tion and on stability over time. If temporal stability can be established, both with respect to sample composition and response to placebo or standard medication, then valid comparisons can be made both within *and* between studies.

We would predict that other investigators should be able to repeat these investigations in other patient populations, use different scoring scales and with the ridit transformation obtain essentially the same results. Moreover, if studies of severity of postoperative nausea and vomiting are carried out in populations in which the incidence of vomiting is higher than that we reported, ancillary information on severity of postoperative sickness will be even more valuable.

SUMMARY

We have presented a method for the clinical evaluation of antiemetic agents. This paper defines a protocol that we have followed and found useful in evaluating antiemetic agents. The method is simple, requires no complex equipment and does not interfere with the normal routine of the recovery room. The sensitivity of the method has been demonstrated and the patient population defined.

The analysis of ancillary information designed to answer the question, "Does the agent lessen the severity and duration of the symptoms that do occur?" has been presented. Although two "strategies," a clinical and a mechanical, were employed to classify severity of postoperative sickness, they appear to measure the same thing despite a different theoretical approach. The ridit scale was found to be advantageous as a means for expressing these results.

The decrease in incidence of vomiting we have obtained applies to the patient population at Memorial Hospital under the conditions of our study. These absolute figures cannot be strictly applied to other patient populations; however, we believe that drugs that appear superior by our testing technique will also be found superior in other clinical situations.

REFERENCES

1. Modell, W., and Houde, R. W.: Factors influencing clinical evaluation of drugs, J. A. M. A. 167: 2190, 1958.

2. Snedecore, G. W.: Statistical Methods, Ames, Iowa, Iowa State College Press, 1956.
3. Wallenstein, S. L., and Houde, R. W.: Changes in pain intensity as means of estimating analgetic power, Fed. Proc. 12: 1243, 1953.
4. Houde, R. W., and Wallenstein, S. L.: Clinical Studies of Narcotics at Memorial Cancer Center, Bull. Drug Addiction & Narcotics, p. 1684, Appendix M, 1957.
5. Bross, I. D. J., and Feldman, R.: Ridit Analysis of Automotive Crash Injuries, Division of Automotive Crash Injury Research, Cornell University Medical College, New York, 1956.
6. Greiner, T., Gold, H., and Bross, I. D. J.: Method for evaluation of laxative habits in human subjects, J. Chronic Dis. 6: 244, 1957.
7. Bross, I. D. J.: How to use ridit analysis, Biometrics 14: 18, 1958.

PAINFUL FASCICULATIONS In 125 patients suxamethonium produced postoperative muscle soreness in 36 per cent of cases. The apparent degree of fasciculation was not related to the amount of postoperative pain, but the prior administration of a depolarizing relaxant or slow administration of suxamethonium decreased pain. Early ambulation seemed to make the muscle soreness more noticeable. Pain was more prominent in the muscles of the trunk, abdomen, and shoulder girdle than in the limbs. (*Leatherdale, R. A. L., Mayhew, R. A. J., and Hayton-Williams, D. S.: Incidence of "Muscle Pain" After Short Acting Relaxants, Brit. Med. J. 1: 904 (April 4) 1959.*)

CUFFED TUBE Overinflation of a cuff caused fatal rupture of the trachea in one patient. Experiments on cadavers showed that a cuff pressure of 320 mm. Hg may cause rupture of the trachea. Routine inflation of a cuff may result in a pressure of 300 mm. Hg. (*Hackl, H., and Koenig, C.: Experimental Investigations Concerning Resistance of the Trachea Against Inflatable Cuffs, Der Anaesthesist 8: 134 (May) 1959.*)

TRACHEAL FLORA Orotracheal intubation under aseptic technique causes an increase in the number of bacteria in the larynx and trachea. Oropharyngeal germs, not present in the larynx and the trachea before intubation, may be found. Use of a lubricant containing a sulfonamide significantly reduces the likelihood of this contamination. (*Beck, H., and Preisler, O.: Laryngeal and Tracheal Flora Before and After Intubation, Der Anaesthesist 8: 110 (April) 1959.*)