Basic statistics for clinicians: 4. Correlation and regression

Gordon Guyatt,*† MD; Stephen Walter,* PhD; Harry Shannon,* PhD; Deborah Cook,*† MD; Roman Jaeschke,*† MD; Nancy Heddle,‡ MSc

Abstract • Résumé

Correlation and regression help us to understand the relation between variables and to predict patients' status in regard to a particular variable of interest. Correlation examines the strength of the relation between two variables, neither of which is considered the variable one is trying to predict (the target variable). Regression analysis examines the ability of one or more factors, called independent variables, to predict a patient's status in regard to the target or dependent variable. Independent and dependent variables may be continuous (taking a wide range of values) or binary (dichotomous, yielding yes-or-no results). Regression models can be used to construct clinical prediction rules that help to guide clinical decisions. In considering regression and correlation, clinicians should pay more attention to the magnitude of the correlation or the predictive power of the regression than to whether the relation is statistically significant.

La corrélation et la régression aident à comprendre le rapport entre des variables et à prédire l'état de patients en fonction d'une variable particulière d'intérêt. La corrélation porte sur la force du rapport entre deux variables dont ni l'une ni l'autre n'est considérée comme la variable que l'on essaie de prédire (la variable cible). L'analyse de régression porte sur la capacité d'un ou de plusieurs facteurs, appelés variables indépendantes, d'aider à prédire l'état d'un patient en fonction de la variable cible ou dépendante. Les variables indépendantes et dépendantes peuvent être soit continues (prendre tout un éventail de valeurs), soit binaires (être dichotomiques, c'est-à-dire donner des résultats présence–absence). On peut utiliser des modèles de régression pour construire des règles de prédiction cliniques qui aident à guider les décisions cliniques. Lorsqu'ils examinent la régression et la corrélation, les cliniciens doivent accorder une plus grande attention à l'ordre de grandeur de la corrélation ou de l'efficacité prédictive de la régression qu'à l'importance statistique de la relation.

Clinicians are sometimes interested in the relation between different factors or "variables." How well does a relative's impression of a patient's symptoms and well-being predict the patient's own report? How strong is the relation between a patient's physical well-being and emotional function? In answering these questions, our goal is to enhance our understanding and consider the implications for action. If the relation between patients' perceptions and those of patients' relatives is not a strong one, the clinician must obtain both perspectives on a situation. If physical and emotional function are only weakly related, then clinicians must probe both areas thoroughly.

Clinicians may be even more interested in making predictions or causal inferences than in understanding the relation between phenomena. Which clinical features of patients with chest pain presenting to the emergency department predict whether they have a myocardial infarction? What determines how dyspneic we feel when we exercise or when we suffer from a cardiac or respiratory illness? Can we predict which critically ill patients will tolerate weaning from a ventilator and which will not?

We refer to the first issue — understanding the magnitude of the relation between different variables or phenomena — as "correlation." We call the statistical techniques for exploring the second issue — making a prediction or causal inference — "regression." In this final article in our series

*From the departments of *Clinical Epidemiology and Biostatistics, †Medicine and ‡Pathology, McMaster University, Hamilton, Ont.*

Dr. Cook is a recipient of a Career Scientist Award from the Ontario Ministry of Health. Dr. Walter is the recipient of a National Health Scientist Award from Health Canada.

***Reprint requests to:** Dr. Gordon Guyatt, Rm. 2C12, McMaster University Health Sciences Centre, 1200 Main St. W, Hamilton ON L8N 3Z5*

This is the final article in a series of four that began in the Jan. 1, 1995, issue of CMAJ.

← For prescribing information see page 619

we will provide illustrations of the use of correlation and regression in medical literature.

Correlation

Traditionally, we measure the exercise capacity of patients with cardiac and respiratory illnesses with the use of a treadmill or cycle ergometer. About 20 years ago, investigators interested in respiratory disease began to use a simpler test that is more closely related to day-to-day activity.[1] In this test, patients are asked to cover as much ground as they can in a specified time (typically 6 minutes) walking in an enclosed corridor. There are several reasons why we may be interested in the strength of the relation between the 6-minute walk test and conventional laboratory measures of exercise capacity. If the results of these tests are strongly related, we could substitute one test for the other. In addition, the strength of the relation could tell us how well exercise capacity, determined by laboratory measures, predicts patients' ability to undertake physically demanding activities of daily living.

What do we mean by the strength of the relation between two variables? A relation is strong when patients who obtain high scores on the first variable also obtain high scores on the second, those who have intermediate scores on the first variable also show intermediate values on the second, and those who have low scores on one measure score low on the other. By contrast, if patients who have low scores on one measure are equally likely to have high or low scores on another, the relation between the two variables is poor or weak.

We can gain a sense of the strength of the correlation by examining a graph that relates patients' scores on the two measures. Fig. 1 presents a scatterplot of the results of the walk test and of the cycle ergometer exercise test. The data for this graph, and for the subsequent analyses involving walk-test results, are taken from three studies of patients with chronic airflow limitation.[2-4] Each point on the scatterplot is for an individual patient and presents two pieces of information: the patient's walk-test score and cycle ergometer exercise score. The walk-test results are continuous; however, the cycle ergometer results tend to take only certain values because patients usually stop the test at the end of a particular level. From Fig. 1, one can see that, in general, patients who have a high score on the walk test tend to have a high score on the cycle ergometer exercise test, and patients who have a low score on the cycle ergometer test tend to have a low score on the walk test as well. Yet one can find patients who are exceptions, scoring higher than most other patients on one test and not as high on the other.

These data represent a moderately strong relation between two variables, the walk test and the cycle ergometer exercise test. The strength of the relation can be summarized in a single number, the correlation coefficient (r). The correlation coefficient can range from −1.0 (the strongest possible negative relation — the patient with the highest score on one test has the lowest score on the other) to 1.0 (the strongest possible positive relation). A correlation coefficient of 0 denotes no relation at all between the two variables: patients with a high score on one test have the same range of scores on the other test as those with a low score on the first test. The scatterplot of data with a correlation coefficient of 0 looks like a starry sky (without the constellations).

The correlation coefficient assumes a straight-line relation between the variables. However, there may be a relation between the variables that does not take the form of a straight line. For example, values of the variables may rise together, but one may rise more slowly than the other for low values and more quickly than the other for high values. If there is a strong relation, but it is not a straight line, the correlation coefficient may be misleading. In our example, the relation does appear to approximate a straight line, and the value of r for the correlation between the walk test and the cycle ergometer test is 0.5.

This value for r indicates a moderately strong correlation, but is it strong enough? It depends on how we wish to apply the information. If we were thinking of substituting the walk test for the cycle ergometer test (after all, the walk test is much simpler to carry out) we would be disappointed. A correlation of 0.8 or higher is required for us to feel comfortable with that kind of substitution. If the correlation is any lower than 0.8, there is too great a risk that a patient with a high score on the walk test would have mediocre or low score on the cycle ergometer test or vice

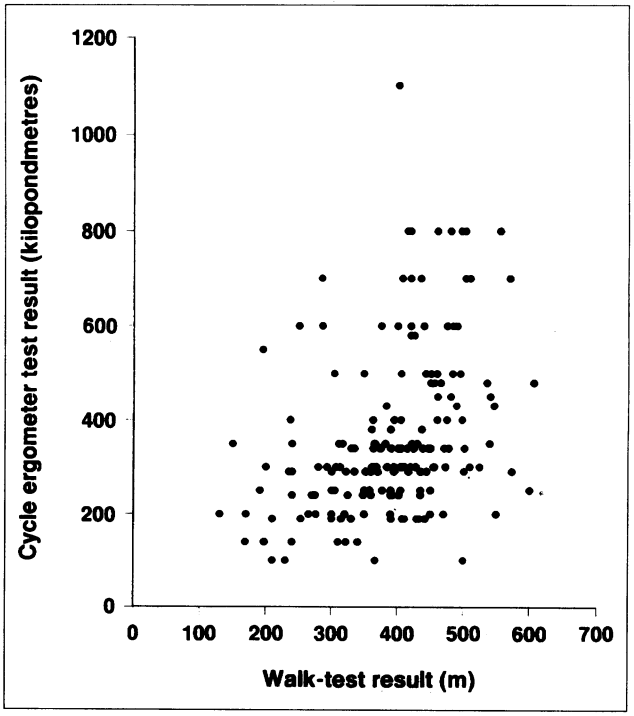

Fig. 1: Scatterplot of the results of the 6-minute walk test and the cycle ergometer exercise test for 179 patients. Each point gives the results for one patient.

versa. However, if we assume that the walk test provides a good indication of exercise capacity in day-to-day life, the moderately strong correlation suggests that the result of the cycle ergometer test also tells us something, although not as much, about day-to-day exercise capacity.

You will often see a p value provided with a correlation coefficient (the first article in this series discusses the interpretation of p values). This p value is determined from a hypothesis test, with the null hypothesis being that the true correlation between the two measures is 0. Thus, the p value represents the probability that, if the true correlation were 0, a relation as strong as or stronger than the one we actually observed would have occurred by chance. The smaller the p value, the less likely it is that chance explains the apparent relation between the two measures.

The p value depends not only on the strength of the relation but also on the sample size. In this case, we had data on the results of the walk test and the cycle ergometer test from 179 patients and a correlation coefficient of 0.5, which yields a p value of less than 0.0001. A relation can be very weak, but if the sample is large enough the p value may be small. For instance, with a sample of 500, we reach the conventional threshold for statistical significance (p = 0.05) when the correlation coefficient is only 0.10.

In a previous article in this series we pointed out that, in evaluating treatment effects, the size of the effect and the confidence interval tend to be much more informative than p values. The same is true of correlations: the magnitude of the correlation and the confidence interval are the key values. The 95% confidence interval for the correlation between the results of the walk test and of the laboratory exercise test is 0.38 to 0.60.

Regression

As clinicians, we are often interested in prediction: we wish to know which patient will get a disease (such as coronary artery disease) and which will not, and which patient will fare well (returning home after a hip fracture rather than remaining in an institution) and which will fare poorly. Regression analysis is useful in addressing these sorts of issues. We will once again use the walk test to illustrate the concepts involved in statistical regression.

Predicting walk-test scores

Let us consider an investigation in which the goal is to predict patients' walk-test scores from more easily measured variables: sex, height and a measure of lung function (forced expiratory volume in 1 second [FEV_1]). Alternatively, we can think of the investigation as an examination of a causal hypothesis. To what extent are patients' walk-test scores determined by their sex, height and lung function? Either way, we have a target or response variable that we call the dependent variable (in this case the walk-test score) because it is influenced or determined by other variables or factors. We also have the explanatory or predictor variables, which we call independent variables — sex, height and FEV_1.

Fig. 2, a histogram of the walk-test scores for 219 pa-

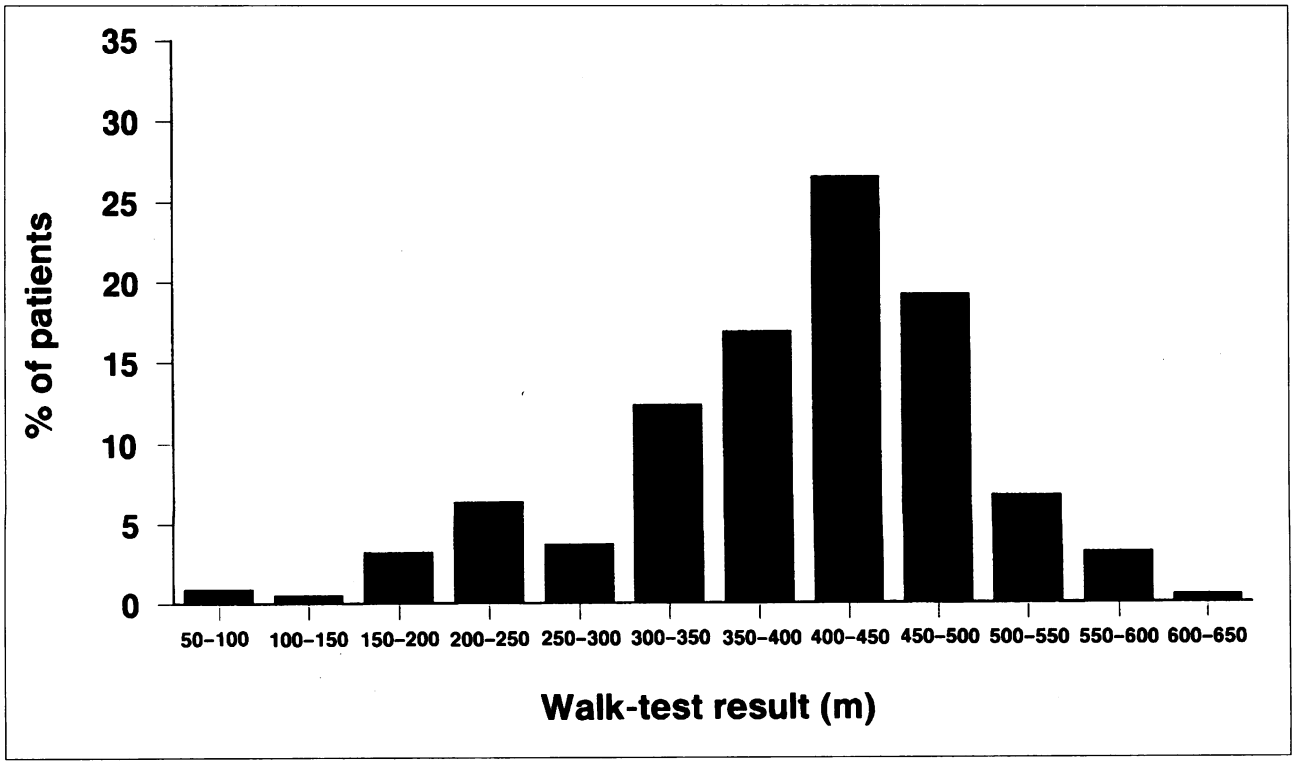

Fig. 2: Distribution of 6-minute walk-test results in a sample of 219 patients.

tients with long-term lung disease, shows that these scores vary widely. If we had to predict an individual patient's walk-test score without any other information, our best guess would be the mean score for all patients (394 m). For many patients, however, this prediction would be well off the mark.

Fig. 3 shows the relation between FEV_1 and walk-test scores. There is a relation between the two variables, although it is not as strong as that between the walk-test score and the exercise-test score, examined earlier (Fig. 1). Thus, some of the variation in walk-test scores seems to be explained by, or attributable to, the patient's FEV_1. We can construct an equation that predicts the walk-test score as a function of FEV_1. Because there is only one independent variable, we call this a univariate or simple regression.[5]

In regression equations we generally refer to the predictor variable as x and the target variable as y. The equation assumes a straight-line fit between the FEV_1 and the walk-test score, and specifies the point at which the straight line meets the y-axis (the intercept) and the steepness of the line (the slope). In this case, the regression equation is $y = 298 + 108x$, where y is the walk-test score in metres, 298 is the intercept, 108 is the slope of the line and x is the FEV_1 in litres. In this case, the intercept of 298 has little practical meaning: it predicts the walk-test score of a patient with an FEV_1 of 0 L. The slope of 108 does, however, have meaning: it predicts that, for every increase in FEV_1 of 1 L, the patient will walk 108 m farther. The regression line corresponding to this equation is shown in Fig. 3.

We can now examine the correlation between the two variables, and whether it can be explained by chance. The correlation coefficient is 0.4, and, since p is 0.0001, chance is a very unlikely explanation for this relation. Thus, we conclude that FEV_1 explains or accounts for a statistically significant proportion of the variation in walk-test scores.

We can also examine the relation between the walk-test score and the patients' sex (Fig. 4). Although there is considerable variation in scores among men and among women, men tend to have higher scores than women. If we had to predict a man's score, we would choose the mean score for the men (410 m), and we would choose the mean score for the women (363 m) to predict a woman's score.

Is the apparent relation between sex and the walk-test score due to chance? One way of answering this question is to construct a simple regression equation with the walk-test score as the dependent variable and the sex of the patient as the independent variable. As it turns out, chance is an unlikely explanation of the relation between sex and the walk-test score ($p = 0.0005$).

As these examples show, the independent variable in a regression equation can be an either/or variable, such as sex (male or female), which we call a dichotomous variable, or a variable that can theoretically take any value, such as FEV_1, which we call a continuous variable.

In Fig. 5 we have divided the men from the women, and for each sex we have separated the patients into groups with a high FEV_1 and a low FEV_1. Although there is still a range of scores within each of these four groups, the range is narrower. When we use the mean of any group as our best guess for the walk-test score of any member of that group, we will be closer to the true value than if we had used the mean for all patients.

Fig. 5 illustrates how we can take into account more than one independent variable in explaining or predicting the dependent variable. We can construct a mathematical model that explains or predicts the walk-test score by simultaneously considering all of the independent variables; this is called a multivariate or multiple regression equation.

We can learn several things from such an equation. First, we can determine whether the independent variables from the univariate equations each make independent contributions to explaining the variation. In this example, we consider first the independent variable with the strongest relation to the dependent variable, then the variable with the next strongest relation and so on. FEV_1 and sex make independent contributions to explaining walk test ($p < 0.0001$ for FEV_1 and $p = 0.03$ for sex in the multiple regression analysis), but height (which was significant at the $p = 0.02$ level when considered in a univariate regression) does not.

If we had chosen the FEV_1 and the peak expiratory flow rate as independent variables, they would both have shown significant associations with walk-test score. However, the FEV_1 and the peak expiratory flow rate are very strongly associated with one another; therefore, they are unlikely to

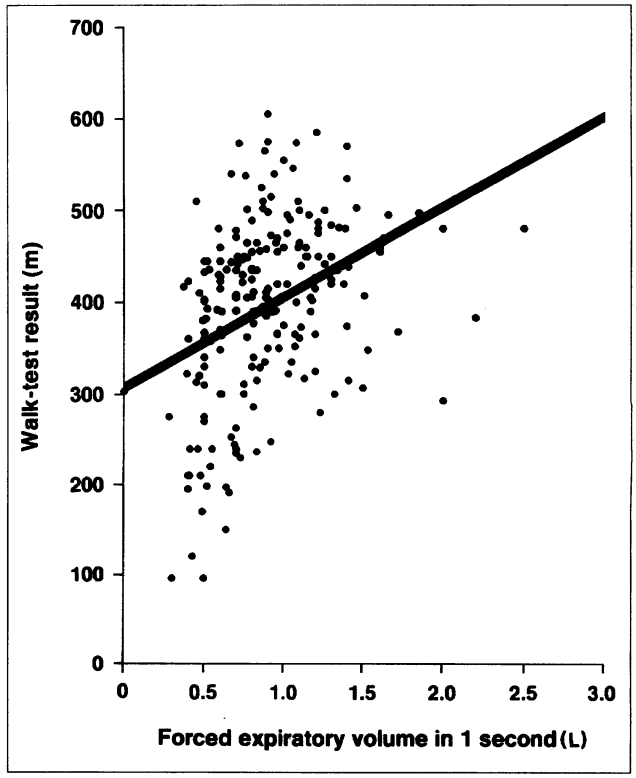

Fig. 3: Scatterplot of the forced expiratory volume in 1 second and of the 6-minute walk-test results for 219 patients. Each point gives the results for one patient.

provide independent contributions to explaining the variation in walk-test scores. In other words, once we take the FEV_1 into account, the peak flow rates are not likely to be of any help in predicting walk-test scores; likewise, if we first took the peak flow rate into account, the FEV_1 would not provide further explanatory power in our model. Similarly, height was a significant predictor of walk-test score when considered alone, but it was no longer significant in the multiple regression because of its correlation with sex and FEV_1.

We have emphasized that the p value associated with a correlation provides little information about the strength of the relation between two values; the correlation coefficient is required. Similarly, the knowledge that sex and FEV_1 independently explain some of the variation in walk-test scores tells us little about the power of our predictive model. We can get some sense of the model's predictive power from Fig. 5. Although the distributions of walk-test scores in the four subgroups differ appreciably, there is considerable overlap. The regression equation can tell us the proportion of the variation in the dependent variable (that is, the differences in walk-test scores among patients) associated with each of the independent variables (sex and FEV_1) and, therefore, the proportion explained by the entire model. In this case, the FEV_1 explains 15% of the variation when it is the first variable entered into the model, sex explains an additional 2% of the variation once the FEV_1 is in the model already, and the overall model explains 17% of the variation. We can therefore conclude that many other factors we have not measured (and perhaps cannot measure) determine how far people with long-term lung disease can walk in 6 minutes. Other regression analyses have found that patients' experience of the intensity of their exertion as well as their perception of the severity of their illness may be more powerful determinants of walk-test distance than their FEV_1.[6]

In this example, the dependent variable — the walk-test score — was continuous. Because this regression analysis

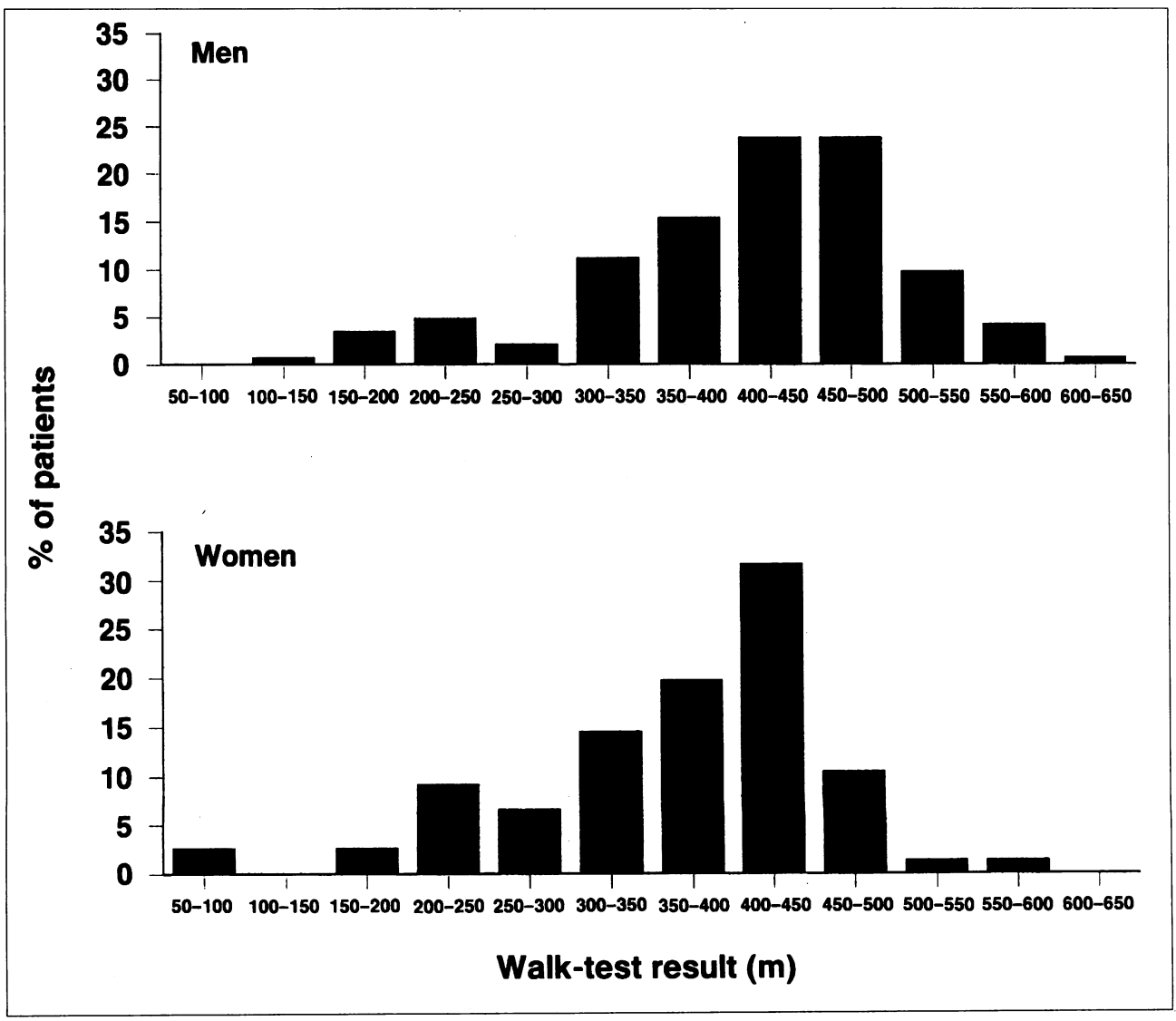

Fig. 4: Distribution of the 6-minute walk-test results in men (top) and women (bottom) from the sample of 219 patients.

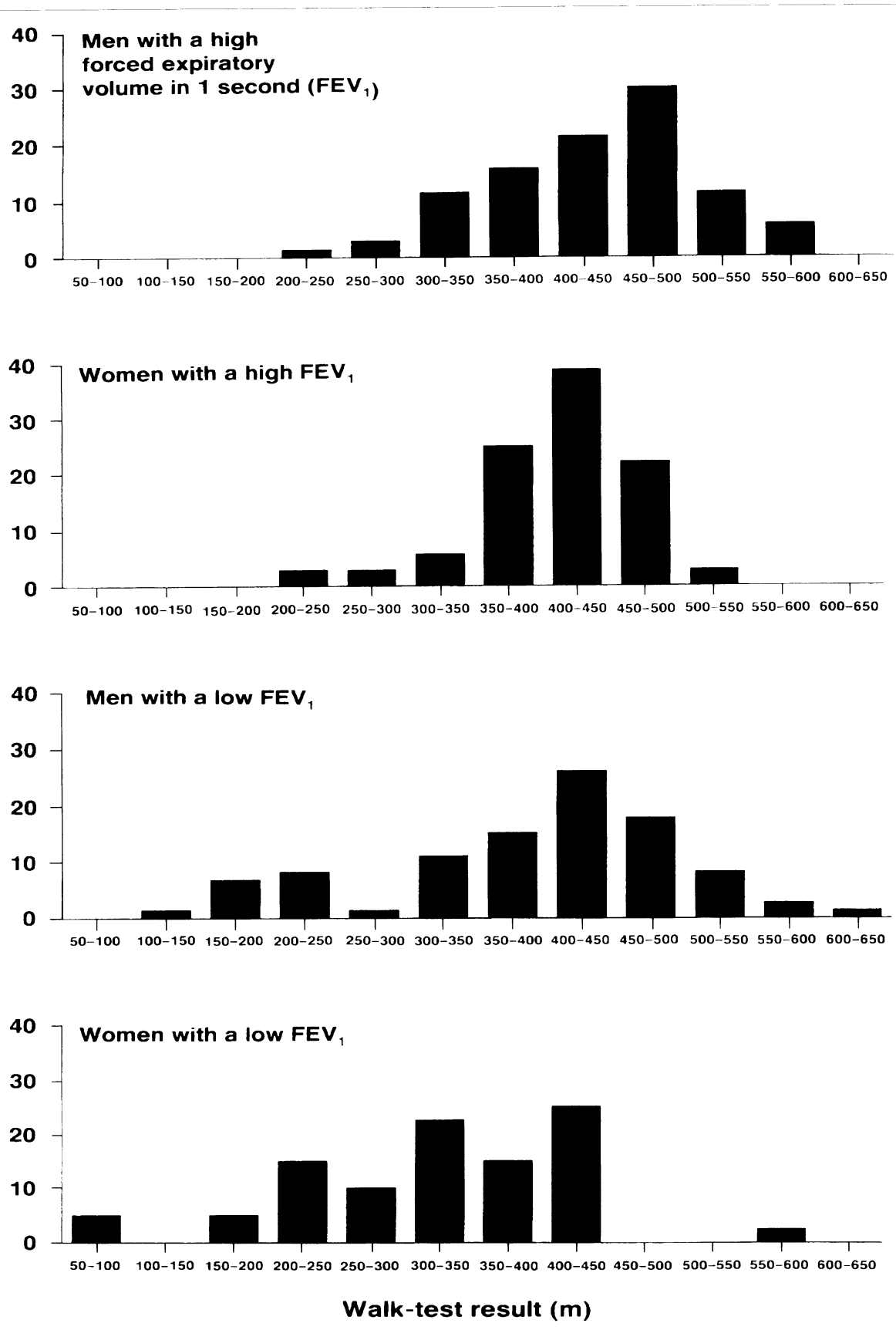

Fig. 5: Distribution of the 6-minute walk-test results in men and women with a high FEV_1 (top), and in men and women with a low FEV_1 (bottom) from the sample of 219 patients.

assumes a straight-line fit between the independent and dependent variable, and the dependent variable is continuous, we refer to the analysis as "linear regression." In our next example, the dependent variable is dichotomous. Investigators sometimes use the term "logistic regression" to refer to such models because they are based on logarithmic equations.

Predicting clinically important gastrointestinal bleeding

We have recently considered whether we could predict which critically ill patients were at risk of clinically important gastrointestinal bleeding.[7] In this example, the dependent variable was whether patients had had a clinically important bleed. When the dependent variable is dichotomous we use a logistic regression. The independent variables included whether patients were breathing independently or required ventilator support and the presence or absence of coagulopathy, sepsis, hypotension, hepatic failure and renal failure.

In the study we followed 2252 critically ill patients and determined which of them had clinically important gastrointestinal bleeding. Table 1, which contains some of the results, shows that in univariate logistic regression analyses many of the independent variables were significantly associated with clinically important bleeding. For several variables, the odds ratio (discussed in a previous article in this series), which indicates the strength of the association, was large. However, when we constructed a multiple logistic regression equation, only two of the independent variables — ventilator support and coagulopathy — were significantly and independently associated with bleeding. All of the other variables that had predicted bleeding in the univariate analysis were correlated with either ventilation or coagulopathy and were not statistically significant in the multiple regression analysis. Of the patients who were not supported by a ventilator, 0.2% (3/1597) had an episode of clinically significant bleeding, whereas 4.6% (30/655) of those being supported by a ventilator had such an episode. Of those with no coagulopathy 0.6% (10/1792) had an episode of bleeding, whereas of those with coagulopathy 5.1% (23/455) had such an episode.

Our main clinical interest was identification of a subgroup with a risk of bleeding low enough that prophylaxis could be withheld. In an analysis separate from the regression analysis, but suggested by its results, we divided the patients into two groups, those who were neither supported by a ventilator nor had coagulopathy, in whom the incidence of bleeding was only 0.14% (2/1405), and those who were either supported by a ventilator or had coagulopathy, of whom 3.7% (31/847) had an episode of bleeding. Prophylaxis may reasonably be withheld from patients in the former group.

Conclusion

Correlation examines the strength of the relation between two variables, neither of which is necessarily considered the target variable. Regression examines the strength of the relation between one or more predictor variables and a target variable. Regression can be very useful in formulating predictive models such as the risk of myocardial infarction in patients presenting with chest pain,[8] the risk of cardiac events in patients undergoing noncardiac surgery,[9] or the risk of gastrointestinal bleeding in critically ill patients. Such predictive models can help us make clinical decisions. Whether you are considering a correlation between vari-

Odds ratios and *p* values for risk factors for clinically important gastrointestinal bleeding in critically ill patients, calculated with the use of simple and multiple logistic regression analysis

Risk factors	Odds ratio (*p* value)	
	Simple regression analysis	Multiple regression analysis*
Respiratory failure	25.5 (< 0.0001)	15.6 (< 0.0001)
Coagulopathy	9.5 (< 0.0001)	4.3 (0.0002)
Hypotension	5.0 (0.03)	2.1 (0.08)
Sepsis	7.3 (< 0.0001)	NS
Hepatic failure	6.5 (< 0.0001)	NS
Renal failure	4.6 (< 0.0001)	NS
Enteral feeding	3.8 (0.0002)	NS
Administration of steroids	3.7 (0.0004)	NS
Transplantation of an organ	3.6 (0.006)	NS
Therapy with anticoagulants	3.3 (0.004)	NS

*NS = not significant.

ables or a regression analysis, you should consider not only the statistical significance of the relation but also its magnitude or strength, in terms of the proportion of variation explained by the model or the extent to which groups with very different risks can be specified.

We thank Derek King, BMath, for conducting the original analyses reported in this article and for preparing the figures.

References

1. McGavin CR, Gupta SP, McHardy GJR: Twelve-minute walking test for assessing disability in chronic bronchitis. *BMJ* 1976; 1: 822–823
2. Guyatt GH, Berman LB, Townsend M: Long-term outcome after respiratory rehabilitation. *Can Med Assoc J* 1987; 137: 1089–1095
3. Guyatt GH, Keller J, Singer J et al: A controlled trial of respiratory muscle training in chronic airflow limitation. *Thorax* 1992; 47: 598–602
4. Goldstein RS, Gort EH, Stubbing D et al: Randomized controlled trial of respiratory rehabilitation. *Lancet* 1994; 344: 1394–1397
5. Godfrey K: Simple linear regression. *N Engl J Med* 1985; 313: 1629–1636
6. Morgan AD, Peck DF, Buchanan DR et al: Effect of attitudes and beliefs on exercise tolerance in chronic bronchitis. *BMJ* 1983; 286: 171–173
7. Cook DJ, Fuller HD, Guyatt GH et al: Risk factors for gastrointestinal bleeding in critically ill patients. *N Engl J Med* 1994; 330: 377–381
8. Pozen MW, D'Agostino RB, Selker HP et al: A predictive instrument to improve coronary-care-unit admission practices in acute ischemic heart disease. *N Engl J Med* 1984; 310: 1273–1288
9. Detsky AS, Abrams HB, McLaughlin JR et al: Predicting cardiac complications in patients undergoing non-cardiac surgery. *J Gen Intern Med* 1986; 1: 211–219

Apr. 3–5, 1995: Current Treatments Conference: Diagnostic and Treatment Problems in Primary Care
Victoria
Coastal Conferences Ltd., 1459 Jamaica Rd., Victoria BC V8N 2C9; tel 604 477-7559, fax 604 595-9594

Apr. 6–7, 1995: Therapeutic Camps for Children and Adolescents (sponsored by the Recreation Discipline and the Children's Outpatient Department, Royal Ottawa Hospital)
Ottawa
Robert Wilson, conference facilitator, Wilcom Services Inc., 59 Horner Dr., Nepean ON K2H 5G1; tel 613 596-6064, fax 613 596-0711

Apr. 10–11, 1995: 5th Annual Palliative Care Conference — Palliative Care . . . Towards Consensus in Practice (in collaboration with the Canadian Association of Nurses in AIDS Care, the Canadian Association of Nurses in Oncology, the Community Hospice Association of Ontario, the Metropolitan Toronto Palliative Care Council, the Ontario Medical Association, Section of Palliative Care, and the Ontario Palliative Care Association)
Toronto
Humber College, Business and Industry Services, 205 Humber College Blvd., Etobicoke ON M9W 5L7; tel 416 675-5077, fax 416 675-0135

Apr. 20–22, 1995: 6th International Congress on Dermatology and Psychiatry: Getting in Touch
Amsterdam, the Netherlands
Bureau PAOG Amsterdam, Tafelbergweg 25, 1105 BC Amsterdam, the Netherlands; tel 011 20 566-4801, fax 011 20 696-3228

Apr. 23–27 1995: An Update on Geriatrics — Alzheimer's disease, depression, therapeutics and medical ethics
Jerusalem, Israel
Study credits available.
Dr. A. Mark Clarfield, Sarah Herzog Hospital, c/o Ortra Ltd., 2 Kaufman St., PO Box 50432, Tel Aviv 61500, Israel; tel 011 972 3 517-7888, fax 011 972 3 517-4433

Apr. 23–27, 1995: Probing the Future: Canadian Society of Clinical Chemists and Canadian Association of Medical Biochemists 39th Annual Scientific Congress
Whistler, BC
CSCC–CAMB 39th Annual Conference, PO Box 1570, 190 Railway St., Kingston ON K7L 5C8; tel 613 531-8899, fax 613 531-0626

Du 23 au 27 avr. 1995 : «Probing the Future» : 39e congrès annuel de la Société canadienne des clinico-chimistes et de l'Association canadienne des médico-biochimistes
Whistler, C-B
Congrès annuel de la SCCC-AMBC, CP 1570, 190, rue Railway, Kingston ON K7L 5C8; tél 613 531-8899, fax 613 531-0626

Apr. 24–25, 1995: Approaching the Dream: Clinical and Cultural Perspectives
Paris, France
Ontario Council for Leadership in Educational Administration, 252 Bloor St. W, Ste. 12–115, Toronto ON M5S 1V5; tel 416 944-2652, fax 416 944-3822

Apr. 26–29, 1995: Canadian Association of Speech-Language Pathologists and Audiologists
Ottawa
Linda J. Garcia, CASLPA Conference '95, Programme d'audiologie et d'orthophonie, University of Ottawa, 545 King Edward Ave., Ottawa ON K1N 6N5; tel 613 564-9918, fax 613 564-9919

Apr. 28–29, 1995: 8th Annual Scientific Conference on Lyme Borreliosis and Other Tick-Borne Diseases
Vancouver
Lyme Disease Foundation, Inc., 1 Financial Plaza, Hartford CT 06103; tel 203 525-2000 or 800 886-LYME, fax 203 525-TICK

Apr. 30, 1995: 7th Annual Symposium on Treatment of Headaches and Facial Pain
New York
Dr. Alexander Mauskop, director, New York Headache Center, 301 E 66th St., New York NY 10021; tel 212 794-3550

For prescribing information see page 594 →