Chapter 15.1: Correlation and Regression

Shanil Ebrahim; Stephen D. Walter; Deborah J. Cook; Roman Jaeschke; Gordon Guyatt

Introduction

Investigators are sometimes interested in the relationship among different measures or variables. They may pose questions related to the correlation of these variables. For example, they might ask, “How well does the clinical impression of symptoms in a child with asthma relate to the parents' perception?” “How strong is the relationship between a patient's physical and emotional functions?”

By contrast, other investigators may be primarily interested in predicting individuals at high risk of having a subsequent event. For instance, can we identify patients with asthma who are at high risk for exacerbations that require hospitalization?

Still other investigators may seek the causal relations among biologic phenomena. For instance, they might ask, “What determines the extent to which a patient with asthma will experience dyspnea when exercising?” Finally, investigators also may pose causal questions that could directly inform patient management. For example, “Does use of long-acting β-agonists in asthma really increase the likelihood of dying?”

Clinicians may be interested in the answers to all 3 sorts of questions—those of correlation, prediction, and causation. To the extent that the relationship between child and parental perceptions is weak, clinicians must obtain both perspectives. If physical and emotional functions are only weakly related, then clinicians must probe both areas thoroughly. We may target patients at high risk of subsequent adverse events with prophylactic interventions. If clinicians know that hypoxemia is strongly related to dyspnea, they may be more inclined to administer oxygen to patients with dyspnea. The clinical implications of the causal questions are more obvious. We may avoid long-acting β-agonists if they really increase the likelihood of dying.

We refer to the degree of association among different variables or phenomena as *correlation*.[1] If we want to describe the relationship among different variables and subsequently use the value of a variable to predict another or make a causal inference, we use a technique called *regression*.[1] In this chapter, we provide examples to illustrate the use of correlation and regression in the medical literature.

Correlation

Correlation is a statistical tool that permits researchers to examine the strength of the relationship between 2 variables when neither variable is necessarily considered the *dependent variable*.

Traditionally, we perform laboratory measurements of exercise capacity in patients with cardiac and respiratory illnesses by using a treadmill or cycle ergometer. Approximately 30 years ago, investigators interested in respiratory disease began to use a simpler test that is related more closely to day-to-day activity.[2] In the walk test, patients are asked to cover as much ground as they can during a specified period (typically 6 minutes) walking in an enclosed corridor. For several reasons, we may be interested in the strength of the relationship between the walk test and conventional laboratory measures of exercise capacity. If the tests relate strongly enough to one another, we might be able to substitute one test for the other. In addition, the strength of the relationship might inform us of the potential of laboratory tests of exercise capacity to predict patients' ability to undertake physically demanding activities of daily living.

What do we mean by the strength of the relationship between 2 measures? One finds a strong positive relationship between 2 measures when patients who obtain high scores on the first also tend to obtain high scores on the second, when those in whom we find intermediate scores on the first also

tend to have intermediate values on the second, and when patients who tend to score low on one measure score low on the other measure.[3] One also can have strong negative relationships: those who score high on one measure score low on the other.[3] If patients who score low on one measure are equally likely to score low or high on another measure, the relationship between the 2 variables is poor, weak, or nonexistent.[3]

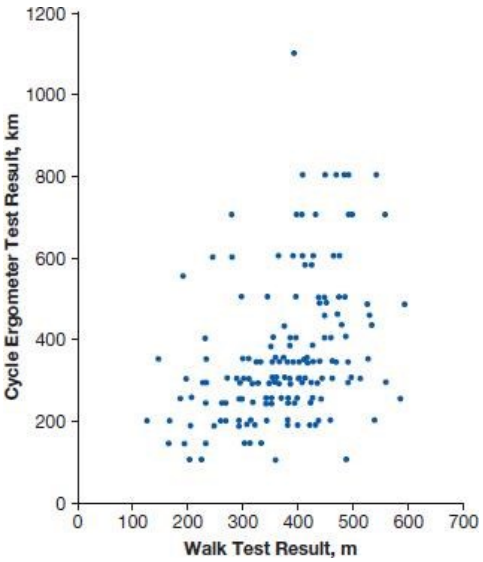
We can gain a sense of the strength of the correlation by examining a visual plot relating patients' scores on the 2 measures. Figure 15.1-1 presents such a plot relating walk test results (on the x-axis) to the results of the cycle ergometer exercise test (on the y-axis). The data for this plot, and those for the subsequent analyses using walk test results, come from 3 studies of patients with chronic airflow limitation.[4-6] Each dot in Figure 15.1-1 represents an individual patient and presents 2 pieces of information: the patient's walk test score and cycle ergometer exercise time. Although the walk test results are truly continuous, the cycle ergometer results tend to take only certain values because patients usually stop the test at the end of a particular level rather than part way through a level.

Examining Figure 15.1-1, you can see that, in general, patients who score well on the walk test also tend to score well on the cycle ergometer exercise test, and patients who score poorly on the cycle ergometer tend to score poorly on the walk test. However, you can find patients who represent exceptions, scoring better than most other patients on one test but not as well on the other test. These data, therefore, represent a moderate relationship between 2 variables: the walk test and the cycle ergometer exercise test.

FIGURE 15.1-1

Relationship Between Walk Test Results and Cycle Ergometer Exercise Test Results

Adapted from Guyatt et al,[1] by permission of the publisher. Copyright © 1995, Canadian Medical Association.



One can summarize the strength of a relationship between 2 continuous (also called interval) variables in a single number, the *Pearson correlation coefficient*. The Pearson correlation coefficient, which is denoted by *r*, can range from –1.0 to 1.0. A correlation of 1.0 or –1.0 occurs when there is a perfect linear relationship between the 2 scores, such that one is completely predictable from the other. A correlation coefficient of –1.0 corresponds to a perfect negative relationship, whereby higher scores on test A are associated with lower scores on test B. A correlation coefficient of 1.0 corresponds to a perfect positive relationship, whereby higher scores on test A are associated with higher scores on test B. A correlation coefficient of 0 denotes no relationship between the 2 variables (ie, the scores on test A and the scores on test B fall in a random pattern). Typically, when calculating a correlation coefficient, a linear relationship between the variables is assumed. There may be a relationship between the variables, but it may not take the form of a straight line when viewed visually. For example, even if larger scores on the variables are found together, one may increase more slowly than the other for low values but will increase more quickly than the other for high values. If there is a strong relationship but it is not a linear one, the correlation coefficient may be misleading.

In the example depicted in Figure 15.1-1, the relationship appears to approximate a straight line, and the *r* value for the correlation between the walk test and the cycle ergometer is 0.50. Should the clinician be pleased and comfortable or displeased and uncomfortable with this moderately strong correlation? It depends on how we wish to apply the information. If we were thinking of using the walk test value as a substitute for the cycle ergometer—after all, the walk test is much simpler to perform—we would be disappointed. A correlation of 0.8 or higher (although the threshold is arbitrary) would be required for us to be confident in that kind of substitution. If the correlation were too low, there would be too much risk that a person with a high walk test score would have mediocre or low performance on the cycle ergometer test or that a person who did poorly on the walk test would do well on the cycle ergometer test. On the other hand, if we assume that the walk test gives a good indication of exercise capacity in daily life, the moderate correlation suggests that the cycle ergometer result tells us something (less than the walk test, but still something) about day-to-day exercise capacity.

In getting a sense of the magnitude of a correlation, in addition to the possibility of substituting one variable for another (requiring a very high correlation) or one variable giving us some indication of status on another (requiring a lower correlation), think of the proportion of variability in one variable that is explained by the other. The square of the correlation represents the proportion of variance explained (eg, if the correlation is 0.4, variable A explains 16% of the variance in variable B; if the correlation is 0.8, variable A explains 64% of the variance in variable B).

You often will see a *P* value in association with a correlation coefficient (see Chapter 12.1, Hypothesis Testing). When correlation coefficients are considered, the *P* value is usually associated with the typical *null hypothesis* that the true correlation between the 2 measures is 0. Thus, the *P* value represents the probability that, if the true correlation were 0, an apparent relationship as strong as or stronger than that actually observed would have occurred as a result of chance. The smaller the *P* value, the less likely it is that chance explains the apparent relationship between the 2 measures.

The *P* value depends not only on the strength of the relationship but also on the sample size. In this case, we had data on both the walk test and the cycle ergometer from 179 patients, and with a correlation of 0.50, the associated *P* value is less than .001. A relationship can be very weak, but if the sample size is sufficiently large, the *P* value may be small. For instance, with a sample size of 500, we reach the conventional threshold *P* value of .05 at a correlation of only 0.10. At the same time, for any given sample size, a stronger correlation will be associated with a lower *P* value.

In evaluating *treatment effects*, the size of the effect and the *confidence interval* (CIs) around the effect tend to be much more informative than *P* values (see Chapter 10, Confidence Intervals: Was the Single Study or Meta-analysis Large Enough?).[7] The same is true of correlations, in which the magnitude of the correlation and the CI around the correlation are the key parameters.

The 95% CI around the correlation between the walk test and laboratory exercise tests ranges from 0.38 to 0.60. A lower limit of 0.38 in the CI signifies a modest correlation.

Regression

Regression examines the strength of a relationship between 1 or more predictor variables and a target variable. As clinicians, we are often interested in prediction. We would like to be able to predict which persons will develop a disease (such as coronary artery disease) and which persons will not; we would like to be able to predict which patient will do well and which patient will do poorly. We also are interested in making causal inferences in situations in which *randomized clinical trials* are not possible. Regression techniques are useful in addressing both sorts of issues.[8]

Regression Modeling With Continuous Target Variables

In any regression, we have a target outcome or response variable that we call the dependent variable because it is influenced or determined by other

variables or factors. When this dependent variable is a continuous variable—such as a 6-minute walk test score that can take a large number of values—a *linear regression* is typically used.[9] Sometimes, individuals treat target variables that take 1 of a number of discrete values, such as the 10 or so levels that a patient might achieve on a conventional exercise test, as if they were continuous.

Regressions also involve explanatory or predictor variables that we suspect may be associated with, or causally related to, the dependent variable. These independent variables can be binary (either/or; also called dichotomous), such as sex (male or female). They may be categorical, with more than 2 categories, such as marital status (single, married, divorced, or widowed). Finally, they may be continuous, such as forced expiratory volume in 1 second (FEV_1).

When there is a single predictor variable and a single dependent variable, we call the regression approach a *bivariable* or *simple regression*.[10] When we are examining more than 1 independent variable, we call the regression approach a *multivariable* or *multiple regression*. The term univariable is reserved for descriptive statistical tests that involve only 1 variable and no independent variable, typically used to describe a sample or to expand a sample to a wider population.[11]

Let us assume we are trying to predict patients' walk test scores using easily measured variables: sex, height, and FEV_1 as a measure of lung function. Alternatively, we can think of the investigation as examining a causal hypothesis: to what extent are patients' walk test scores determined by sex, height, and pulmonary function? Either way, the dependent variable here is the walk test result, and the independent variables are sex, height, and FEV_1.

Figure 15.1-2, a histogram of the walk test scores of 219 patients with chronic lung disease, shows that walk test scores vary widely among patients. If we had to predict an individual's walk test score without any other information, our best guess would be the mean score of all patients (394 m). For many patients, however, this prediction would be well off the mark.

Figure 15.1-3 shows the relationship between FEV_1 and the walk test. Note that there is a relationship between the 2 variables, although the relationship is not as strong as the relationship between the walk test and the exercise test depicted in Figure 15.1-1. Thus, some of the differences, or variation, in walk test scores seems to be explained by, or attributable to, the patient's FEV_1. We can construct an equation using FEV_1 to predict walk test scores.

Generally, when we construct regression equations, we refer to the predictor (independent) variable as x and the target (dependent) variable as y. The regression equation in this example assumes a linear fit between the FEV_1 and the walk test data 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 is expressed as follows:

$$y = 298 + 108x$$

where y is the value of the walk test, 298 is the intercept, 108 is the slope of the line, and x is the value of the FEV_1 in liters. In this case, the intercept of 298 has little practical meaning; it predicts the walk test distance of a patient with an FEV_1 of 0. The slope of 108, however, has some meaning; it predicts that for every increase in FEV_1 of 1 L, the patient will walk 108 m farther. We show the regression line corresponding to this formula in Figure 15.1-3.

Having constructed the regression equation, we can examine the correlation between the 2 variables, and we can assess whether the correlation might be explained by chance. The correlation is 0.40, suggesting that chance is a very unlikely explanation ($P < .001$). Thus, we conclude that FEV_1 explains or accounts for a statistically significant proportion of the variability, or variance, in walk test scores.

We also can examine the relationship between the walk test score and patients' sex (Figure 15.1-4). Although there is considerable variability within the sexes, men tend to have higher walk test scores than women. If we had to predict a man's score, we would choose the mean score of the men (410 m); to predict a woman's score, we would choose the women's mean score of 363 m.

We can ask the question, “Does the apparent relationship between sex and walk test score result from chance?” One way of answering this question is to construct another simple regression equation with walk test as the dependent variable and patient's sex as the independent variable. As it turns out, chance is an unlikely explanation of the relationship between sex and the walk test ($P < .001$).

In Figure 15.1-5, we have separated the men from the women, and for each sex, we have divided them into groups with high and low FEV_1 results. Although there is a range of scores within each of these groups, the range is narrower than among all women or all men and even more so than all patients. When we use the mean of the men as our best guess of the walk test score of a man and the mean of the women as our best guess of the walk test score of a woman, we will, on average, be closer to the true value than if we had used the mean for all patients.

Figure 15.1-5 illustrates how we can take more than 1 independent variable into account at the same time 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, thus creating a multivariable regression equation.

FIGURE 15.1-2

Distribution of Walk Test Results in the Total Sample of 219 Patients

Adapted from Guyatt et al,[1] by permission of the publisher. Copyright © 1995, Canadian Medical Association.

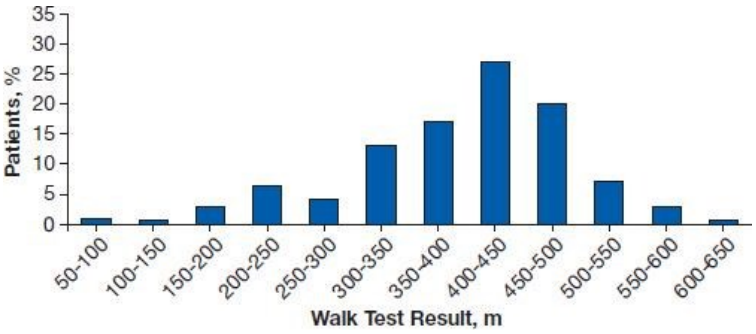


FIGURE 15.1-3

Relationship Between FEV_1 and Walk Test Results in 219 Patients

Abbreviation: FEV_1, forced expiratory volume in 1 second.

Adapted from Guyatt et al,[1] by permission of the publisher.

Copyright © 1995, Canadian Medical Association.

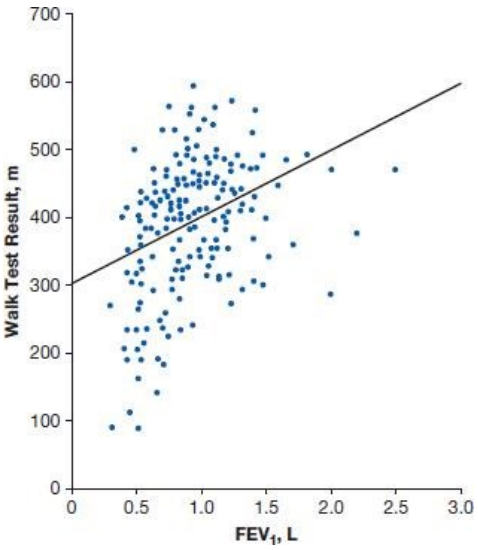


FIGURE 15.1-4

Distribution of Walk Test Results in Men and in Women (Sample of 219 Patients)

Adapted from Guyatt et al,[1] by permission of the publisher. Copyright © 1995, Canadian Medical Association.

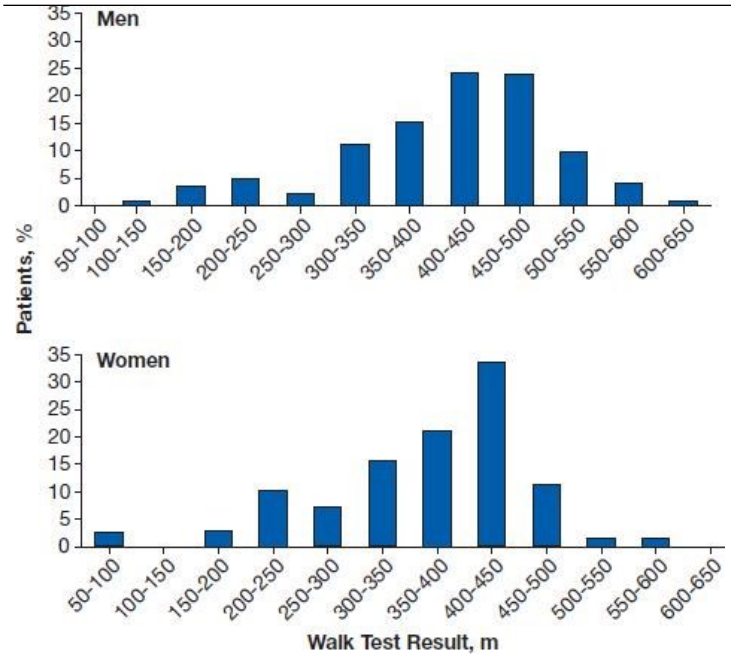
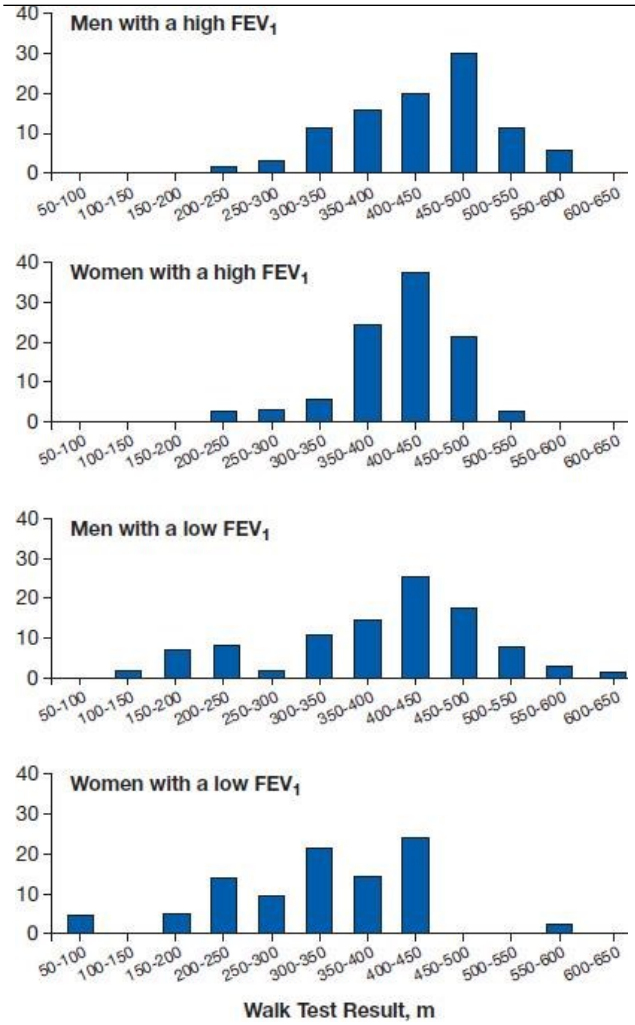


FIGURE 15.1-5

Distribution of Walk Test Results in Men and Women With High and Low FEV_1 (Sample of 219 Patients)

Abbreviation: FEV_1, forced expiratory volume in 1 second.

Adapted from Guyatt et al,[1] by permission of the publisher. Copyright © 1995, Canadian Medical Association.



Multivariable regression equations allow us to determine whether each of the variables that were associated with the dependent variable in the bivariable equations makes contributions to explaining the variation. Independent variables that are strongly associated with one another (such as age and year of birth) usually will not make strong separate contributions to predicting the dependent variable. Multivariable regression approaches provide us with models in which each variable makes its own independent contribution to the prediction.[12]

For example, FEV_1 and sex both make independent contributions to explaining walk test results ($P < .001$ for FEV_1 and $P = .03$ for sex in the multivariable regression analysis), but height (which was significant at the $P = .02$ level when considered in a bivariable regression) does not make a comparable contribution to the explanation.

If we had chosen both the FEV_1 and peak expiratory flow rates as independent variables, they would both reveal significant associations with walk test score. However, because FEV_1 and peak expiratory flow rates are associated strongly with one another, they are unlikely to provide independent contributions to explaining the variation in walk test scores. In other words, once we take FEV_1 into account, peak flow rates are not likely to be of any help in predicting walk test scores, and if we first took peak flow rate into account, FEV_1 would not provide further explanatory power to our predictive model. Similarly, height was a significant predictor of walk test score when considered alone but was no longer significant in the multivariable regression because of its correlation with sex and FEV_1.[1]

We have emphasized how the *P* value associated with a correlation provides little information about the strength of the relation between 2 values; the

correlation coefficient itself is required. Similarly, knowing that a number of independent variables in a multivariable model explain some of the variation in the dependent variable tells us little about the power of our predictive model.

Regression equations can tell us much more: the proportion of the variation in the dependent variable that is explained by the model. If a model explains less than 10% of the variability, it is not very useful. If it explains more than 50% of the variability, it will be extremely useful. Intermediate proportions of variability explained are of intermediate value.

Returning to our example, Figure 15.1-5 gives us some sense of the model's predictive power. Although the distributions of walk test scores in the 4 subgroups differ appreciably, considerable overlap remains. In this case, 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, and the total model explains 17% of the variation. We therefore can conclude that there are many other factors that we have not measured—and, perhaps, that we cannot measure—that determine how far people with chronic lung disease can walk in 6 minutes. Other investigations that use regression techniques have found that patients' experience of the intensity of their exertion, as well as the perception of the severity of their illness, may be more powerful determinants of walk test distance than is their FEV_1.[13]

Regression Modeling With Dichotomous Target Variables

Frequently, we are interested in predicting a patient's status on a dichotomous dependent variable, such as death or myocardial infarction, in which the outcome is present or absent. We use the term *logistic regression* to refer to such models.

Some time ago, we addressed the question of whether we could predict which critically ill patients are at risk of clinically important upper gastrointestinal tract bleeding.[14] The dependent variable was whether patients had a clinically important bleeding episode. The independent variables included whether patients were breathing independently or required mechanical ventilation and the presence of coagulopathy, sepsis, hypotension, hepatic failure, or renal failure.

Table 15.1-1 gives some of the results from this study, in which we documented the frequency of major bleeding episodes in 2252 critically ill patients. The table indicates that in bivariable logistic regression equations, many independent variables (respiratory failure, coagulopathy, hypotension, sepsis, hepatic failure, renal failure, enteral feeding, corticosteroid administration, organ transplantation, and anticoagulant therapy) were significantly associated with clinically important bleeding. For a number of variables, the *odds ratio* (see Chapter 7, Therapy [Randomized Trials]), which indicates the strength of the association, is quite large.

When we constructed a multiple logistic regression equation, however, only 2 of the independent variables, mechanical ventilation and coagulopathy, were significantly and independently associated with risk of bleeding. All of the other variables that predicted bleeding in the bivariate analysis were correlated with mechanical ventilation or coagulopathy and, therefore, did not reach conventional levels of *statistical significance* in the multiple regression model. Of those not requiring mechanical ventilation, 3 of 1597 (0.2%) experienced a bleeding episode; of those who received ventilatory support, 30 of 655 (4.6%) experienced a bleeding episode. Of those with no coagulopathy, 10 of 1792 (0.6%) bled; of those with coagulopathy, 23 of 455 (5.1%) experienced a bleeding episode.

Our primary clinical interest was to identify a subgroup with a sufficiently low bleeding risk that prophylaxis might be withheld. Separate from the regression analysis but suggested by its results, we divided the patients into 2 groups: those who were neither mechanically ventilated nor had a coagulopathy and in whom the incidence of bleeding was only 2 of 1405 (0.14%) and those who were either ventilated or had a coagulopathy and of whom 31 of 847 (3.7%) had a bleeding episode. We concluded that prophylaxis may reasonably be withheld in the former low-risk group.

TABLE 15.1-1

Odds Ratios and *P* Values According to Simple (Bivariable) and Multiple (Multivariable) Logistic Regression Analysis for Risk Factors for Clinically Important Gastrointestinal Bleeding in Critically Ill Patients

Risk Factor	Simple Regression		Multiple Regression	
	OR	*P* Value	OR	*P* Value
Mechanical ventilation	25.5	<.001	15.6	<.001
Coagulopathy	9.5	<.001	4.3	<.001
Hypotension	5.0	.03	2.1	.08
Sepsis	7.3	<.001	NS	
Hepatic failure	6.5	<.001	NS	
Renal failure	4.6	<.001	NS	
Enteral feeding	3.8	<.001	NS	
Corticosteroid administration	3.7	<.001	NS	
Organ transplant	3.6	.006	NS	
Anticoagulant therapy	3.3	.004	NS	

Abbreviations: OR, odds ratio; NS, not significant.

Adapted from Guyatt et al,[1] by permission of the publisher. Copyright © 1995, Canadian Medical Association.

Conclusion

Correlation is a statistical tool that permits researchers to examine the strength of the relationship between 2 variables when neither variable is necessarily considered the dependent variable. Regression, by contrast, examines the strength of the relationship between 1 or more predictor variable and a target variable. Regression can be very useful in formulating predictive models to assess risks; for example, the risk of subsequent death in patients presenting with acute coronary syndrome,[15] the risk of cardiac events in patients undergoing noncardiac surgery,[16] or the risk of bleeding in critically ill patients.[14] Such predictive models can help us make better clinical decisions. Such models are also vital for examining causal associations, particularly with rare harmful events, in *observational studies* when *randomization* is not possible. Regardless of whether you are considering an issue of correlation or regression, you should note not only whether the relationship among variables is statistically significant but also the magnitude or strength of the relationship in terms of the proportion of variation explained, the extent to which groups with very different risks of the target event can be specified, or the odds ratio associated with a putative harmful *exposure*.

References

1. Guyatt G, Walter S, Shannon H, Cook D, Jaeschke R, Heddle N. Basic statistics for clinicians: 4. Correlation and regression. *CMAJ*. 1995;152(4):497--504. [PubMed: 7859197]

2. McGavin CR, Gupta SP, McHardy GJ. Twelve-minute walking test for assessing disability in chronic bronchitis. *Br Med J*. 1976;1(6013):822-–823. [PubMed: 1260350]

3. Streiner DL. *A Guide for the Statistically Perplexed: Selected Readings for Clinical Researchers*. Toronto, Ontario: University of Toronto Press; 2013:187.

4. Guyatt GH, Berman LB, Townsend M. Long-term outcome after respiratory rehabilitation. *CMAJ*. 1987;137(12):1089-–1095. [PubMed: 3676969]

5. Guyatt G, Keller J, Singer J, Halcrow S, Newhouse M. Controlled trial of respiratory muscle training in chronic airflow limitation. *Thorax*. 1992;47(8):598-–602. [PubMed: 1412115]

6. Goldstein RS, Gort EH, Stubbing D, Avendano MA, Guyatt GH. Randomised controlled trial of respiratory rehabilitation. *Lancet*. 1994;344(8934):1394-–1397. [PubMed: 7968075]

7. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 2. Interpreting study results: confidence intervals. *CMAJ*. 1995;152(2):169-–173. [PubMed: 7820798]

8. Katz MH. Multivariable analysis: a primer for readers of medical research. *Ann Intern Med*. 2003;138(8):644-–650. [PubMed: 12693887]

9. Sedgwick P. Statistical question: correlation versus linear regression. *BMJ*. 2013;346:f2686.

10. Godfrey K. Simple linear regression in medical research. *N Engl J Med*. 1985;313(26):1629-–1636. [PubMed: 3840866]

11. Winker MA, Lurie SJ. Glossary of statistical terms. In: *AMA Manual of Style: A Guide for Authors and Editors.* 10th ed. New York, NY: Oxford University Press; 2007. http://www.amamanualofstyle.com/view/10.1093/jama/9780195176339.001.0001/med-9780195176339-div1-215. Accessed January 7, 2014.

12. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med*. 2004;66(3):411-–421. [PubMed: 15184705]

13. Morgan AD, Peck DF, Buchanan DR, McHardy GJ. Effect of attitudes and beliefs on exercise tolerance in chronic bronchitis. *Br Med J (Clin Res Ed)*. 1983;286(6360):171-–173. [PubMed: 6401516]

14. Cook DJ, Fuller HD, Guyatt GH, et al; Canadian Critical Care Trials Group. Risk factors for gastrointestinal bleeding in critically ill patients. *N Engl J Med*. 1994;330(6):377-–381. [PubMed: 8284001]

15. Eagle KA, Lim MJ, Dabbous OH, et al; GRACE Investigators. A validated prediction model for all forms of acute coronary syndrome: estimating the risk of 6-month postdischarge death in an international registry. *JAMA*. 2004;291(22):2727-–2733. [PubMed: 15187054]

16. Detsky AS, Abrams HB, McLaughlin JR, et al. Predicting cardiac complications in patients undergoing non-cardiac surgery. *J Gen Intern Med*. 1986;1(4):211-–219. [PubMed: 3772593]