Committee on Cancer Staging System. *Eur J Cardiothorac Surg* 2013; **44**: e207–e211.
54 Nafteux P, Lerut T, De Hertogh G, Moons J, Coosemans W, Decker G *et al*. Can extracapsular lymph node involvement be a tool to fine-tune pN1 for adenocarcinoma of the oesophagus and gastro-oesophageal junction in the Union Internationale Contre le Cancer (UICC) TNM 7th edition? *Eur J Cardiothorac Surg* 2014; **45**: 1001–1010.
55 Enlow JM, Denlinger CE, Stroud MR, Ralston JS, Reed CE. Adenocarcinoma of the esophagus with signet ring cell features portends a poor prognosis. *Ann Thorac Surg* 2013; **96**: 1927–1932.
56 Nafteux PR, Lerut TE, Villeneuve PJ, Dhaenens JM, De Hertogh G, Moons J *et al*. Signet ring cells in esophageal and gastroesophageal junction carcinomas have a more aggressive biological behavior. *Ann Surg* 2014; **260**: 1023–1029.
57 Izzo JG, Wu TT, Wu X, Ensor J, Luthra R, Pan J *et al*. Cyclin D1 guanine/adenine 870 polymorphism with altered protein expression is associated with genomic instability and aggressive clinical biology of esophageal adenocarcinoma. *J Clin Oncol* 2007; **25**: 698–707.
58 Ong CA, Shapiro J, Nason KS, Davison JM, Liu X, Ross-Innes C *et al*. Three-gene immunohistochemical panel adds to clinical staging algorithms to predict prognosis for patients with esophageal adenocarcinoma. *J Clin Oncol* 2013; **31**: 1576–1582.

**Statistical nugget**

# Statistical models: an overview

**J. Ranstam and J. A. Cook**

*BJS Statistical Editors*

**DOI:** 10.1002/bjs.10240

A simple statistical model consists of an outcome and a sole explanatory (or predictor) variable. This type of model is sometimes described as bivariable as it includes only two variables. It is often used to estimate an 'unadjusted' or 'crude' effect, that is the influence of a single factor on the outcome of interest without taking account of other factors that may also influence the outcome. To produce an 'adjusted' estimate, other explanatory variables can be incorporated into the model (a multivariable model). Different models can be used for various types of outcomes (such as logistic regression for a binary outcome), although all make assumptions regarding the relationship between the variables in the model in order to estimate their effects. Less commonly it may be useful or necessary to model two or more outcomes together by assuming a joint probability distribution and constructing a model accordingly (using a multivariate model)[1].

Multivariable models have two major uses: to explain or to predict. The first purpose is to define (with allowance for multiple factors) parameter estimates. In an observational study, an estimate of the factor of interest with adjustment for other important factors is often carried out this way. It may not fully resolve the issue owing to unobserved factors and lead to a biased (although hopefully less biased) estimate. In a randomized trial, the reason for adjusting is mainly to reduce the uncertainty of the estimates (increase the precision) by adjusting for pre-specified strongly prognostic factors, accounting for randomization (stratification) factors, and adjusting for chance baseline imbalance in continuous endpoints when studying these endpoints' change from baseline.

The second purpose is to predict the outcome typically for an individual, such as recurrence of disease following treatment (or classify individual properties). Development of a multivariable model in order to predict (predictive or prognostic model) differs from those above. The best prediction model is simply the model that predicts best[2]. Predictive performance can be assessed in different ways. We will discuss the principles for model choice in each of these situations in more detail later.

## References

1 Hidalgo B, Goodman M. Multivariate or multivariable regression? *Am J Public Health* 2013; **103**: 39–40.
2 Shmueli G. To explain or to predict? *Statistical Science* 2010; **25**: 289–310.