Leading articles

Derivation versus validation

Assessing the probable clinical course of children presenting to their care is one of the day to day dilemmas facing the practising paediatrician. With accurate prognostication, invasive or expensive treatment may be targeted towards those most likely to benefit. There is little point in expending vital resources on, or administering invasive treatments to, children who are likely to recover without such intervention. Children with certain characteristics may tend to do better or worse than others. For example, younger children or those with specific clinical signs might be expected to deteriorate more rapidly. Past experience often acts as a guide for the experienced paediatrician. Algorithms (or prognostic models) can be developed to provide a means of transferring expert knowledge to the novice. For example, Apgar scores are routinely used to assess the health of newborn babies, and APACHE scores can be used as a measure of prognostication among admissions to paediatric intensive care.[1] In this issue, Brogan and Raffles[2] present an algorithm for identifying children presenting to A&E with fever and petechiae who are at increased risk of significant bacterial sepsis. It is interesting to note that of the many prognostic models that are published each year,[3 4] relatively few are sufficiently validated and fewer still find their way into clinical practice.[5]

Determining which patients will benefit from treatment is only one use of prognostic algorithms. They also provide a means of informing parents of the likely outcome and can be used in research to give baseline measures of severity in different groups.

Derivation of prognostic models

An experienced clinician accumulates knowledge via the patients that he or she has cared for from initial presentation through to final outcome. What has happened to these cases will inform the future decisions that the clinician makes. The process can be formalised by documenting information on newly referred patients who are then tracked until their outcome is known. When a suitably sized sample has been collected, statistical techniques may be used to build an algorithm designed to predict poor outcome in future patients.

The most common way of deriving a prognostic model from existing data is via regression analyses. The development of prognostic models is extensively covered in many medical statistics texts. From the competing models one may be chosen that is both practical and statistically acceptable. In particular, models may be preferred on the grounds that they require only routinely collected and reliable data and do not have substantially less predictive ability than alternative models that require invasive or non-routine data. Decisions may need to be made quickly and algorithms should be simple and user friendly.

One problem with the development of algorithms for prognostication is that derivation will be driven by the available dataset. Quirks individual to that dataset may appear to be prognostic and be encapsulated in the predictive algorithm. For example, time between symptoms and presentation may be unrelated to outcome but by chance those patients with the poorest outcome in the available dataset tended to present soon after the first appearance of symptoms. A prognostic model designed to be applied at presentation which incorporates time from symptoms as a factor would under diagnose those presenting late. Unusual features or extreme but random differences between prognostic groups within the development dataset may not be replicated elsewhere, leading to the creation of a non-transportable model. Where there are many variables competing for inclusion this problem may be extreme. Hence, analyses that are not pre-specified but are data dependent are liable to give a better fit than is obtained when the model is applied elsewhere. The problem is further compounded when cut off points for continuous variables are selected to give the best prognostication based on the development dataset.[6] For example, respiratory rate was dichotomised at 40, 50, 60, 70, and 80/min to identify the cut off point giving the best sensitivity and specificity for predicting hypoxia in acutely ill infants.[7] Although arbitrary thresholds for continuous variables are not generally recommended,[5] they are often preferred because of the simplicity that they confer on the final algorithm.

The converse problem is that of under fitting. Important prognostic variables may not be identified in the derivation dataset and there are several reasons this could occur. Random or chance variation may mean that the values of a truly prognostic variable are not significantly different between prognostic groups within the available dataset. For example, suppose time from first symptoms to presentation is predictive of outcome but that, in the development dataset, it just happens that the patients with the worse outcome were unusual in that they tended to present early. Alternatively prognostic factors may not be identified because the patient set used for derivation is limited in some way. For example, age may be highly prognostic but does not enter into the algorithm because the model was derived from data collected only from children in a very narrow age range.

Methods of model checking are available with most statistical computer packages and tend to be well covered in most regression texts. However, these methods merely detect whether the chosen model adequately describes the trends in the dataset on which it has been developed. They cannot inform on whether the model is suitable for use in clinical practice or an accurate description of the population trends.

Validation

> The actual evidence that the application of a prognostic model alters medical practice and improves the outcome of patients has to be established additionally which is in accordance with phase IV studies of diagnostic tests.[8]

Model validation is the process whereby the derived (or fitted) model is shown to be suitable for the purpose for which it was developed. It addresses the question of whether the model has wider applicability. As the aim of most published papers is to present results that will be generally useful, these are questions of major importance that cannot be overlooked. Surprisingly few medical statistics textbooks discuss techniques for model validation and most do not even mention its importance or relevance to model interpretation.

In addition to being user friendly and statistically sound, a prognostic algorithm needs to be generalisable to be clinically useful. A model based on variables which have low reliability will not tend to be valid in a sample other than that for which it was developed. One of the reasons published algorithms may not find their way into standard practice is the lack of evidence that they are applicable to patients from establishments other than where they were developed. Generalisability needs to be established by testing the prognostic algorithm in numerous and diverse settings.[9] In the words of a recent article[10] on the subject of model validation, "Usefulness is determined by how well a model works in practice, not by how many zeros there are in the associated p-values."

Techniques for model validation

It is well recognised that deriving and validating a model on the same dataset will by definition lead to over optimistic estimates of the model's accuracy. An alternative approach is to split the dataset into two parts, one part for derivation and the other for validation. A major drawback is that the precision of the fitted parameters will clearly be reduced as only a portion of the dataset is used for model derivation.[11] A variety of methods have been advocated for dividing the dataset. Automated procedures exist for choosing two halves that are homogeneous but these will also clearly give over optimistic estimates of model validity. If some measure of internal validity is required then it is recommended that the data are split in a non-random way. Data from different time periods could be used[3] and this is the same as validation using a more recent cohort or prospective validation using an algorithm derived from a retrospective dataset. Alternatively, one of the less arbitrary "leave one out" approaches may be used.[12 13]

External validity

The model should be externally validated by assessing its applicability to data collected at another centre or by different individuals. Altman and Royston[10] present a series of examples of models that have been derived, internally validated, and then validated elsewhere. They note that authors tend to confirm the validity of their own models but that others are less successful at doing so. This finding could be the result of a form of publication bias; if the authors' internal validation was weak it is doubtful that they would attempt to publish the results. Alternatively there may be real differences between centres and the model, while being internally valid, is not transportable and hence of limited use.

Statistical versus clinical validity

> In general authors show no appreciation of a distinction between statistically and clinically valid models.[10]

A model that is statistically valid will yield unbiased predictions when applied to new datasets. This quality however does not necessarily mean that the model has clinical validity. To be clinically valid the model must be accurate enough to serve the purpose for which it was developed. For example, abnormal white cell might be significantly associated with increased risk of significant bacterial sepsis among children presenting to A&E with rash. This finding, even if replicated across hospitals and hence statistically valid, may be of little clinical relevance if only a few extra children with poor outcome are identified as a result. There are different clinical implications if application of the algorithm means that an additional child in 50 or an additional child in four with poor outcome is identified early. Consideration must of course also be given to whether there are other children who achieve a worse outcome, perhaps because treatment is withheld that would normally have been administered, as a result of applying the algorithm. Similarly a statistically invalid model is not necessarily clinically invalid. The algorithm may not be consistent in the extent to which it identifies children with poor outcome but it may always identify enough to warrant its implementation on a regular basis. Scores need to be reliable enough for the purpose they were developed for, even if this reliability is relatively low.[14] There will be a trade off between the additional workload entailed in applying the model and the likely benefit derived by the patients.

Conclusion

> Simple diagnostic models may be more portable than more complex models.[15]

When Kennedy *et al* made the above observation they were commenting on the statistical aspects of deriving a model which is then used elsewhere, but it is of course true in more ways than one. Models based only on factors which are highly related to outcome will be more likely to be similarly predictive in another setting. Complex models incorporating multiple factors, for some of which any predictive value may be highly specific to that dataset, are less likely to be similarly predictive elsewhere. From a practical viewpoint simple models are more likely to be readily incorporated into clinical practice with minimal disruption.

Model derivation and validation are two separate and important parts of the same process, the identification of clinically useful models. It is to be remembered that the final test of a model should always be whether it is accurate and generalisable enough for the purpose for which it was derived.

ANGIE WADE

Department of Epidemiology and Public Health,
Institute of Child Health, 30 Guilford Street,
London WC1N 1EH, UK
Awade@ich.ucl.ac.uk

1 Keengwe IN, Stansfield F, Eden OB, Nelhans ND, Dearlove OR, Sharples A. Paediatric oncology and intensive care treatments: changing trends. *Arch Dis Child* 1999;**80**:553–5.
2 Brogan PA, Raffles A. The management of fever and petechiae: making sense of rash decisions. *Arch Dis Child* 2000;**83**:506–7.
3 Muhe L, Oljira B, Degefu H, Enquesellassie F, Weber MW. Clinical algorithm for malaria during low and high transmission seasons. *Arch Dis Child* 1999;**81**:216-20.
4 Kumar R, Singh SN, Kohli N. A diagnostic rule for tuberculous meningitis. *Arch Dis Child* 1999;**81**:221-4.
5 Wyatt JC, Altman DG. Prognostic models: clinically useful or quickly forgotten? *BMJ* 1995;**311**:1539–41.
6 Buettner P, Garbe C, Guggenmoos-Holzmann I. Problems in defining cut-off points of continuous prognostic factors: example of tumor thickness in primary cutaneous melanoma. *J Clin Epidemiol* 1997;**50**:1201–10.
7 Rajesh VT, Singhi S, Kataria S. Tachypnoea is a good predictor of hypoxia in acutely ill infants under 2 months. *Arch Dis Child* 2000;**82**:46–9.
8 Windeler J. Prognosis—what does the clinician associate with this notion? *Stat Med* 2000;**19**:425–30.
9 Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;**130**:515–24.
10 Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;**19**:453–73.
11 Roecker EB. Prediction error and its estimation for subset-selected models. *Technometrics* 1991;**33**:459–68.
12 Allen DM. Mean square error of prediction as a criterion for selecting variables. *Technometrics* 1971;**13**:469–75.
13 Efron B, Tibshirani R. *An introduction to the bootstrap*. London: Chapman and Hall, 1993:255.
14 Polerman KH, Thijs LG, Girbes ARJ. Interobserver variability in the use of APACHE II scores. *Lancet* 1999;**353**:380.
15 Kennedy RL, Burton AM, Fraser HS, McStay LN, Harrison RF. Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: derivation and evaluation of logistic regression models. *Eur Heart J* 1996;**17**:1181–91.