Prognosis and prognostic research: application and impact of prognostic models in clinical practice

Karel G M Moons,[1] Douglas G Altman,[2] Yvonne Vergouwe,[1] Patrick Royston[3]

An accurate prognostic model is of no benefit if it is not generalisable or doesn't change behaviour. In the last article in their series **Karel Moons and colleagues** discuss how to determine the practical value of models

[1]Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, Netherlands
[2]Centre for Statistics in Medicine, University of Oxford, Oxford OX2 6UD
[3]MRC Clinical Trials Unit, London NW1 2DA
Correspondence to: K G M Moons k.g.m.moons@umcutrecht.nl
Accepted: 6 October 2008

Cite this as: *BMJ* 2009;338:b606
doi: 10.1136/bmj.b606

This article is the last in a series of four aiming to provide an accessible overview of the principles and methods of prognostic research

Prognostic models are developed to be applied in new patients, who may come from different centres, countries, or times. Hence, new patients are commonly referred to as different from but similar to the patients used to develop the models.[1-4] But what exactly does this mean? When can a new patient population be considered similar (enough) to the development population to justify validation and eventually application of a model? We have already considered the design, development, and validation of prognostic research and models.[5-7] In the final article of our series, we discuss common limitations to the application and generalisation of prognostic models and what evidence beyond validation is needed before practitioners can confidently apply a model to their patients. These issues also apply to prediction models with a diagnostic outcome (presence of a disease).

Limitations to application

Extrapolation versus validation

Most prediction models are developed in secondary care, and it is common to want to apply them to primary care.[1 8-10] The predictive performance of secondary care models is usually decreased when they are validated in a primary care setting.[1 9] One example is the diagnostic model to predict deep vein thrombosis, which had a negative predictive value of 97% (95% confidence interval 95% to 99%) and sensitivity 90% (83% to 96%) in Canadian secondary care patients.[11] When the model was validated in Dutch primary care patients, the negative predictive value was only 88% (85% to 91%) and sensitivity 79% (74% to 84%).[12] The question arises whether primary and secondary care populations can indeed be considered to be different but similar.

A change in setting clearly results in a different case mix, which commonly affects the generalisability of prognostic models.[4 9 13 14] Case mix is here defined as the distribution of the outcome and predictive factors whether included in the model or not. Primary care doctors often selectively refer patients to specialists. Secondary care patients can thus largely be considered to be a subpopulation of primary care patients, commonly with a narrower range of patient characteristics, a larger fraction of patients in later disease stages, and worse outcomes.[9] Consequently, application of a secondary care model in general practice requires extrapolation. This view suggests that applying a primary care model to secondary care would have a limited effect on predictive performance, although this requires further research.

Another common generalisation, or rather extrapolation, is from adults to children. Various prognostic models have been developed to predict the risk of postoperative nausea and vomiting in adults scheduled for surgery under general anaesthesia. When validated in children, the models' predictive ability was substantially decreased.[15] The researchers considered children as a different population and developed and validated a separate model for children that included other predictors.[16] In contrast, the Intensive Care National Audit and Research Centre model to predict outcome in critical care was initially developed with data from adults but also has good accuracy in children.[17]

In general, models will be more generalisable when the ranges of predictor values in the new population are within the ranges seen in the development population. The above examples show that we cannot assume that prediction models can simply be generalised from one population or setting to another, although it may be possible. Therefore, accuracy of any prediction model should always be tested in a formal validation study (see third article in this series[7]).

Adequate prediction versus application

Just because a model is well used does not mean it has adequate prediction. For example, the Framingham risk model discriminates only reasonably in certain (sub)populations, with a receiver-operating characteristic (ROC) curve area of little over 0.70.[18] The model is nevertheless widely used. The same applies to various intensive care prediction models—for example, the APACHE scores and the simplified acute physiology scores (SAPS).[19 20] A likely reason is the relevance of the outcomes that these rules predict: risk of cardiovascular disease (Framingham) and mortality in critically illness (APACHE, SAPS). Another reason for the wide use of such models is their face validity, such that doctors trust these models to guide their practice rather than their own experience.

Whether the predictive accuracy of a model in new patients is adequate is also a matter of judgment and depends on available alternatives.[21] For instance, a prognostic model to predict the probability of spontaneous ongoing pregnancy in couples with unexplained subfertility has good calibration but rather low discriminative ability (ROC area even below 0.70) but remains the best model available.[22] Hence, the model was used to identify couples with intermediate probability of spontaneous ongoing pregnancy for a clinical trial.[23]

Finally, the role of prognostic models and prognostic factors in clinical practice still depends on circumstances. A positive family history of subarachnoid haemorrhage increases the risk of subarachnoid haemorrhage 5.5 times, but only 10% of cases of subarachnoid haemorrhage occur in people with a family history. Thus screening for subarachnoid haemorrhage in people with a family history is not recommended as it will identify relatively too few cases.[24]

Usability

Constraints on the usability of the prognostic model can also limit the application. Application of prognostic models requires unambiguous definitions of predictors and reproducible measurements using methods available in clinical practice. For example, one of the predictors in the deep vein thrombosis model described above is "alternative diagnosis just as likely as deep vein thrombosis."[11] General practitioners may be less experienced in properly coding this predictor for a patient, leading to misclassification that potentially compromises the rule's predictive performance. Another example of an ambiguous predictor definition is "history of nausea and vomiting after previous anaesthesia" in the prognostic model for postoperative nausea and vomiting.[25] A negative answer could mean that the patient has had anaesthesia before but not experienced symptoms or that the patient has never had anaesthesia. Also, children will have had previous anaesthesia less often than adults. As a consequence, this predictor may have a different effect in children.

Similarly, the definition of the outcome variable may vary across populations. Occurrence of neurological sequelae after childhood bacterial meningitis was defined in a development population as mild cases (for example, hearing loss), severe cases (for example, deafness), or dead.[26] The prognostic model was validated in a population that included children with mainly mild neurological sequelae. The model showed poor performance in the validation population, possibly because of the different distribution of outcomes.[27] In addition, the follow-up time differed between the two populations (the maximum duration of follow-up was 3.3 years in the development population and 10 years in the validation population).

Consecutive stages to produce a usable multivariable prognostic model

- *Development studies*[5 6]—Development of a multivariable prognostic model, including identification of the important predictors, assigning the relative weights to each predictor, and estimating the model's predictive performance (eg, calibration and discrimination) adjusted if necessary for overfitting
- *Validation studies*[7]—Validating or testing the model's predictive performance in new subjects, preferably from different centres, with a different case mix or using (slightly) different definitions and measurements of predictors and outcomes
- *Impact studies*—Quantifying whether use of a prognostic model in daily practice improves decision making and, ultimately, patient outcome using a comparative design

Changes over time

As we discussed in the first article in this series,[5] changes in practice over time can limit the application of prognostic models. Improvements in diagnostic tests, biomarker measurement, or treatments may change the prognosis of patients. For example, spiral computed tomography can better visualise the pulmonary circulation than older computed tomography.[28] As a consequence, a patient with pulmonary embolism detected by spiral computed tomography and treated accordingly may have a better prognosis on average than a patient with an embolism detected by conventional computed tomography.

Changes over time may even lead to the situation that prognostic models are no longer used to estimate outcome risks and to influence patient management. For example, the suggestion that everyone older than 55 is given a "polypill" to reduce the risk of cardiovascular diseases[29] may make models to predict these diseases redundant.

Evidence beyond validation studies

Adjusting and updating prognostic models to improve performance

Newly collected data from prediction research are often used to develop a new prognostic model rather than to validate existing models.[2 3 7 14] For example, there are over 60 models to predict outcome after breast cancer[30] and about 25 models to predict long term outcome in patients with neurological trauma.[31] If researchers do perform a formal validation study of a published model and find poor performance, they often then re-estimate the associations of the predictors with the outcome in their own data. Sometimes even the entire selection of important predictors is repeated. This is unfortunate, since predictive information captured in developing the original model is neglected. Furthermore, validation studies commonly include fewer patients than development studies, making the new model more subject to overfitting and thus even less generalisable than the original model.[4 14]

When a prognostic model performs less well in another population, adjusting the model using the new data should first be considered to determine whether it will improve the performance in that population.[4 13 14] The adjusted model is then based on both the development and validation data, further improving its stability and generalisability. Such adjustment of prognostic models is called updating. Updating methods vary from simple recalibration to more extensive methods referred to as model revision.[4 13 14] Recalibration includes adjustment of the intercept of the model and overall adjustment of the associations (relative weights) of the predictors with the outcome. Model revision includes adjustment of individual predictor-outcome associations and addition of new predictors. Interestingly, simple recalibration methods are often sufficient.[4 14] The extent to which this

Comparison of characteristics of validation study and impact study for prognostic models

Characteristic	Validation study[7]	Impact study
Control group	No	Yes. Index group includes doctors exposed to or using the prognostic model; control group is usual care (without using the model)
Design	Prospective cohort (preferred); retrospective cohort	Cluster randomisation (preferred); before and after
Outcome	Usually occurrence of event (eg, death, complication, treatment response) after a certain time or follow-up period	(Change in) doctors' decisions or behaviour
		Patient outcome (eg, events, pain, quality of life)
		Cost effectiveness of care
Follow-up	Yes	No, if outcome is doctors' decisions or behaviour
		Yes, if outcome is patient outcome or cost effectiveness of care
Statistical analysis and reporting	Model's calibration and discrimination	Comparison of outcome between index and control group—eg, using relative risks, odds ratios, or difference in means
	Defining particular risk groups by introducing thresholds	
	Improving or updating a model (if needed)	

process of model validation and adjustment has to be pursued before clinical application, will depend on the context. General rules are as yet unavailable.

Impact of prognostic models

Prognostic models are developed to provide objective estimates of outcome probabilities to complement clinical intuition and guidelines.[5 8 10 21] The underlying assumption is that accurately estimated probabilities improve doctors' decision making and consequently patient outcome. The effect of a previously developed, validated, and (if needed) updated prognostic model on behaviour and patient outcomes should be studied separately in so called impact studies (box).

Validation and impact studies differ in their design, study outcome, statistical analysis, and reporting (table). A validation study ideally uses a prospective cohort design and does not require a control group.[7] For each patient, predictors and outcome are documented, and the rule's predictive performance is quantified.

By contrast, impact studies quantify the effect of using a prognostic model on doctors' behaviour, patient outcome, or cost effectiveness of care compared with not using such model (table). They require a control group of healthcare professionals who provide usual care. The preferred design is a randomised trial.[3] If behaviour changes of professionals is the main outcome, a randomised study without follow-up of patients would suffice. Follow-up is required if patient outcome or cost effectiveness is assessed. However, since changes in outcome depend on changes in doctors' behaviour, it may be sensible to start with a randomised study assessing the model's impact on therapeutic decisions, especially when long follow-up times are needed to assess patient outcome. The same applies to diagnostic procedures[32] and therapeutic interventions for which effects are realised by changing behaviour and decisions—for example, exercise therapy to reduce body mass index.

Impact studies may use an assistive approach—simply providing the model's predicted probabilities of an outcome between 0% and 100%—or a decisive approach that explicitly suggests decisions for each probability category.[3 33] The assistive approach clearly leaves room for intuition and judgment, but a decisive approach may have greater effect.[3 34 35] Introduction of computerised patient records that automatically give predictions for individual subjects, enhances implementation and thus impact analysis of prognostic models in routine care.[35 36]

Randomising individual patients in an impact study may result in learning effects because the same doctor will alternately apply and not apply the model to subsequent patients, reducing the contrast between both randomised groups. Randomisation of doctors (clusters) is preferable, although this requires more patients.[37] Randomising centres is often the best method as it avoids exchange of experiences between doctors within a single centre.

An alternative design is a before and after study with the same doctors or centres, as was used to evaluate the effect of the Ottawa ankle rule on physicians' behaviour and cost effectiveness of care.[38 39] A disadvantage of this design is the sensitivity to temporal changes in therapeutic approaches. Although impact studies are scarce, are a few good examples exist.[40-42]

When to apply a prognostic model

Do all prognostic models require a three step assessment (box) before they are used in daily care? Does a model that has shown adequate prediction for its intended use in validation studies—adequately predicting the outcome—still require an impact analysis using a large, multicentre cluster randomised study? The answers depend on the rate of (acceptable) false positives and false negative predictions and their consequences for patient management and outcome. For models with (near) perfect discrimination and calibration in several validation studies the answer may be no, though such success is rare. An example is a model to predict the differential diagnosis of acute meningitis. It showed an area under the ROC curve of 0.97 in the development population[43] and of 0.98 in two validation populations.[44 45]

For models with less perfect performance, only an impact analysis can determine whether use of the model is better than usual care. Impact studies also provide the opportunity to study factors that may affect implementation of a prognostic model in daily care, including the acceptability of the prognostic model to clinicians and ease of use.

An intermediate step using decision modelling techniques or Markov chain models can be helpful. These evaluate the potential consequences of using the prognostic model in terms of subsequent therapeutic decisions and patient outcome.[46] If such analysis does not indicate any potential for improved patient outcome, a formal impact study would not be warranted.

Concluding remarks

Many prognostic models are developed but few have their predictive performance validated in new patients, let alone an evaluation of their impact on decision making and patient outcome.[3 47 48] Thus it seems right that few such models are actually used in practice. Recent methodological advances enable the adjustment of prognostic models to local circumstances to give improved generalisability. With these innovations,

SUMMARY POINTS

Prognostic models generalise best to populations that have similar ranges of predictor values to those in the development population

When a prognostic model performs less well in a new population, using the new data to modify the model should first be considered rather than directly developing a new model

Application of prognostic models requires unambiguous definitions of predictors and outcomes and reproducible measurements using methods available in clinical practice

Impact studies quantify the effect of using a prognostic model on physicians' behaviour, patient outcome, or cost effectiveness of care compared with usual care without the model

Impact studies need different design, outcome, analysis, and reporting from validation studies

correctly developed and evaluated prediction models may become more common.

Many questions remain unresolved. How much validation, and perhaps adjustment, is needed before an impact study is justified? Is it feasible for a single model to apply to all patient subgroups, across all levels of care and countries? These issues require further research. Finally, we reiterate that unvalidated models should not be used in clinical practice, and more impact studies are needed to determine whether a prognostic or diagnostic model should be implemented in daily practice.

Funding: KGMM and YV are supported by the Netherlands Organization for Scientific Research (ZON-MW 917.46.360). PR is supported by the UK Medical Research Council. DGA is supported by Cancer Research UK.

Contributors: This series was conceived and planned by DGA, KGMM, PR, and YV. KGMM wrote the first draft of this paper. All the authors contributed to subsequent revisions. KGMM is guarantor.

Competing interests: None declared.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130:515-24.
2 Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;19:453-73.
3 Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med* 2006;144:201-9.
4 Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004;23:2567-86.
5 Moons K, Royston P, Vergouwe Y, Grobbee D, Altman D. Prognosis and prognostic research: what, why and how? *BMJ* 2009;338:b375.
6 Royston P, Moons K, Altman D, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. *BMJ* 2008;338:b604.
7 Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;338:b605.
8 McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. *JAMA* 2000;284:79-84.
9 Knottnerus JA. Between iatrotropic stimulus and interiatric referral: the domain of primary care research. *J Clin Epidemiol* 2002;55:1201-6.
10 Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA* 1997;277:488-94.
11 Wells PS, Anderson DR, Bormanis J, Guy F, Mitchell M, Gray L, et al. Value of assessment of pretest probability of deep-vein thrombosis in clinical management. *Lancet* 1997;350:1795-8.
12 Oudega R, Hoes AW, Moons KG. The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. *Ann Intern Med* 2005;143:100-7.
13 Van Houwelingen JC. Validation, calibration, revision and combination of prognostic survival models. *Stat Med* 2000;19:3401-15.
14 Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008;61:76-86.
15 Eberhart LH, Morin AM, Guber D, Kretz FJ, Schauffelen A, Treiber H, et al. Applicability of risk scores for postoperative nausea and vomiting in adults to paediatric patients. *Br J Anaesth* 2004;93:386-92.
16 Eberhart LH, Geldner G, Kranke P, Morin AM, Schauffelen A, Treiber H, et al. The development and validation of a risk score to predict the probability of postoperative vomiting in pediatric patients. *Anesth Analg* 2004;99:1630-7.
17 Harrison DA, Rowan KM. Outcome prediction in critical care: the ICNARC model. *Curr Opin Crit Care* 2008;14:506-12.
18 Liao Y, McGee DL, Cooper RS, Sutkowski MB. How generalizable are coronary risk prediction models? Comparison of Framingham and two national cohorts. *Am Heart J* 1999;137:837-45.
19 Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 1991;100:1619-36.
20 Le Gall JR, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *JAMA* 1993;270:2957-63.
21 Wyatt JC, Altman DG. Prognostic models: clinically useful or quickly forgotten? *BMJ* 1995;311:1539-41.
22 Hunault CC, Habbema JD, Eijkemans MJ, Collins JA, Evers JL, te Velde ER. Two new prediction rules for spontaneous pregnancy leading to live birth among subfertile couples, based on the synthesis of three previous models. *Hum Reprod* 2004;19:2019-26.
23 Steures P, van der Steeg JW, Hompes PG, Habbema JD, Eijkemans MJ, Broekmans FJ, et al. Intrauterine insemination with controlled ovarian hyperstimulation versus expectant management for couples with unexplained subfertility and an intermediate prognosis: a randomised clinical trial. *Lancet* 2006;368:216-21.
24 Rinkel GJ. Intracranial aneurysm screening: indications and advice for practice. *Lancet Neurol* 2005;4:122-8.
25 Van de Bosch J, Moons KGM, Bonsel GJ, Kalkman CJ. Does measurement of preoperative anxiety have added value in the prediction of postoperative nausea and vomiting? *Anesth Analg* 2005;100:1525-32.
26 Oostenbrink R, Moons KGM, Derksen-Lubsen G, Grobbee DE, Moll HA. Early prediction of neurological sequelae or death after bacterial meningitis. *Acta Paediatr* 2002;91:391-8.
27 Biesheuvel CJ, Koomen I, Vergouwe Y, Van Furth M, Oostenbrink R, Moll HA, et al. Validating and updating a prediction rule for neurological sequelae after childhood bacterial meningitis. *Scand J Infect Dis* 2006;38:19-26.
28 Holbert JM, Costello P, Federle MP. Role of spiral computed tomography in the diagnosis of pulmonary embolism in the emergency department. *Ann Emerg Med* 1999;33:520-8.
29 Wald NJ, Law MR. A strategy to reduce cardiovascular disease by more than 80%. *BMJ* 2003;326:1419.
30 Altman D. Prognostic models: a methodological framework and review of models for breast cancer. In: Lyman GH, Burstein HJ, eds. *Breast cancer. Translational therapeutic strategies*. New York: Informa Healtcare, 2007:11-25.
31 Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. *BMC Med Inform Decis Mak* 2006;6:38.
32 Deeks JJ. Assessing outcomes following tests. In: Price CP, Christenson RH, eds. *Evidence-based laboratory medicine: principles, practice, and outcomes*. 2nd ed. Washington: AACC Press, 2007:95-111.
33 Gordon-Lubitz RJ. Risk communication: problems of presentation and understanding. *JAMA* 2003;289:95.
34 Michie S, Johnston M. Changing clinical behaviour by making guidelines specific. *BMJ* 2004;328:343-5.
35 Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ* 2005;330:765.
36 James BC. Making it easy to do it right. *N Engl J Med* 2001;345:991-3.
37 Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. *BMJ* 2004;328:702-8.
38 Stiell I, Wells G, Laupacis A, Brison R, Verbeek R, Vandemheen K, et al. Multicentre trial to introduce the Ottawa ankle rules for use of radiography in acute ankle injuries. *BMJ* 1995;311:594-7.
39 Cameron C, Naylor CD. No impact from active dissemination of the Ottawa ankle rules: further evidence of the need for local implementation of practice guidelines. *CMAJ* 1999;160:1165-8.
40 Foy R. A randomised controlled trial of a tailored multifaceted strategy to promote implementation of a clinical guideline on induced abortion care. *BJOG* 2004;111:726-33.
41 Meyer G, Kopke S, Bender R, Muhlhauser I. Predicting the risk of falling—efficacy of a risk assessment tool compared to nurses' judgement: a cluster-randomised controlled trial. *BMC Geriatr* 2005;5:14.
42 Marrie TJ, Lau CY, Wheeler SL, Wong CJ, Vandervoort MK, Feagan BG. A controlled trial of a critical pathway for treatment of community-acquired pneumonia. *JAMA* 2000;283:749-55.
43 Spanos A, Harrell FE Jr, Durack DT. Differential diagnosis of acute meningitis. An analysis of the predictive value of initial observations. *JAMA* 1989;262:2700-7.
44 McKinney WP, Heudebert GR, Harper SA, Young MJ, McIntire DD. Validation of a clinical prediction rule for the differential diagnosis of acute meningitis. *J Gen Intern Med* 1994;9:8-12.
45 Hoen B, Viel JF, Paquot C, Gerard A, Canton P. Multivariate approach to differential diagnosis of acute meningitis. *Eur J Clin Microbiol Infect Dis* 1995;14:267-74.
46 Steyerberg EW, Keizer HJ, Habbema JD. Prediction models for the histology of residual masses after chemotherapy for metastatic testicular cancer. *Int J Cancer* 1999;83:856-9.
47 Graham ID, Stiell IG, Laupacis A, McAuley L, Howell M, Clancy M, et al. Awareness and use of the Ottawa ankle and knee rules in 5 countries: can publication alone be enough to change practice? *Ann Emerg Med* 2001;37:259-66.
48 Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH, Abboud PA, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. *JAMA* 1999;282:1458-65.