METHODOLOGIC PERSPECTIVES

Assessing Prognosis from Nonrandomized Studies: An Example from Brain Arteriovenous Malformations

J. Raymond
O. Naggara
F. Guilbert
D.G. Altman

SUMMARY: Two recent publications from Helsinki and Toronto that investigated the natural history of brain AVMs are the background topic for reviewing some principles and pitfalls of prognostic studies. Multivariable prognostic research involves 3 steps: developing the prognostic model, validating its performance in other individuals, and assessing its clinical impact on patients' outcomes. Unfortunately, the predictive ability of the model can be poor when it is applied to a new population, and clinical impact studies are rarely performed. Models that have not been validated should not be used to inform clinical decisions. Unfortunately, for rare outcomes in rare diseases, clinical data are limited. Although the 2 studies on brain AVMs may represent the best data currently available, they still included few patients with events and there are several methodologic concerns undermining the reliability of results. The estimates of risk of rupture per year are uncertain. Multiplying those uncertain numbers by the life expectancy of individuals can inflate error beyond control. Hence relying on these estimates to make clinical decisions may be dangerous.

ABBREVIATIONS: AVM = arteriovenous malformations

Brain AVMs are relatively rare central nervous system lesions that can cause significant long-term morbidity and mortality. Although they are believed to be congenital malformations, most patients present after a long delay (20–50 years). The most common presentation is intracranial hemorrhage (40%) or epilepsy (40%); less common clinical presentations are nonspecific headaches and a progressive neurologic deficit. With the availability of noninvasive neurovascular imaging studies, increasing proportions of AVMs are incidental findings.

Treatment options include surgery, stereotactic radiation, endovascular embolization, or a combination of these. While microsurgical removal may provide an immediate cure for superficial AVMs in noneloquent brain, resection of malformations in certain locations with a large nidus, deep draining veins, and high-flow shunts may carry a relatively high risk of morbidity. Embolization is performed either to render surgery easier or less morbid or to reduce the AVM size to make it more likely to respond to radiation therapy. Hence, therapy may be initiated with a number of sessions of embolization during a few months. Therapy may be completed by radiation therapy, which takes 2–3 years to sclerose the AVM in ≤80% of cases if the nidus is ≤2 cm. Therapy is sometimes suspended or interrupted because of a complication (transient or permanent). All current therapeutic options involve risks and benefits that have never been evaluated in randomized trials.[1] However, A Randomized Trial of Unruptured Brain AVMs, a comparison between conservative management of unruptured AVMs and any treatment (surgical, endovascular, or radiation therapy, alone or in combination), is currently recruiting.[1]

One difficulty in clinical decisions and assessment of prognosis with or without treatment is the diversity of lesions that can occur anywhere in the brain (in silent or eloquent areas). Size can vary from microscopic to giant (>6 cm); lesion complexity may vary from 1 abnormal vessel to dozens of vessels. The patients may present additional secondary risk factors such as acquired aneurysms on arteries feeding the AVM, on veins draining the AVM, or inside the nidus.

Prognostic Studies to Inform Clinical Decisions?

In the absence of evidence, 1 basis for treatment choice (which does not meet standards of evidence-based medicine and assumes the long-term efficacy of treatment) is to compare the risks of treatment with the natural history of the disease. However, what is the natural history of brain AVMs? Two recently published articles shed some light on this matter.[2,3] Both studies evaluated time to hemorrhage (survival without rupture) by using Kaplan-Meier curves, logrank tests, and multivariable Cox proportional hazards analyses.

The first study from Helsinki looked at 238 patients followed for a median of 7.4 years.[2] Of these, 77 experienced a hemorrhage, for an average annual risk of 2.4%. The risk was highest during the first 5 years after diagnosis, decreasing markedly thereafter. Risk factors predicting subsequent AVM hemorrhage were previous rupture, large size, deep and infratentorial locations, and deep venous drainage.

The other study, from Toronto, included 678 prospectively enrolled patients followed for a mean of 2.9 years, during which time 89 had hemorrhages.[3] It showed that hemorrhage rates were 4.6% per year for the entire cohort ($n = 678$), 7.5% per year for AVMs with initial hemorrhagic presentation ($n = 258$), 4.2% per year for initial seizure presentation ($n = 260$), 4.0% per year for patients not harboring aneurysms ($n = 556$), 6.9% per year for patients with associated aneurysms ($n =$

METHODOLOGIC PERSPECTIVES

From the Interventional Neuroradiology Research Unit (O.N., F.G, J.R.), Department of Radiology, International Consortium of Neuroendovascular Centres, University of Montreal, Centre Hospitalier de l'Université de Montreal, Notre-Dame Hospital, Montreal, Quebec, Canada; Department of Neuroradiology (O.N.), Paris-Descartes University, Institut National de la Santé et de la Recherche Médicale U894, Centre Hospitalier Sainte-Anne, Paris; and Centre for Statistics in Medicine (D.G.A.), University of Oxford, Oxford, United Kingdom.

Paper previously presented at: 10th Congress of World Federation of Interventional and Therapeutic Neuroradiology, June 29-July 3, 2009; Montreal, Quebec, Canada.

Please address correspondence to Jean Raymond, MD, Interventional Neuroradiology, CHUM-Notre-Dame Hospital, 1560 Sherbrooke East, Pavilion Simard, Room Z12909, Montreal, PQ, Canada H2L 4M1; e-mail: jean.raymond@Umontreal.ca

DOI 10.3174/ajnr.A2516

122), and 5.4% per year for AVMs with deep venous drainage (n = 365). Hemorrhagic presentation was the only statistically significant independent predictor of future hemorrhage (HR, 2.15; P = .01). A potentially disturbing finding for endovascular therapists, the hemorrhagic risk was not noticeably different in patients who underwent partial AVM embolization (n = 211; HR, 0.875; P = .32).

How reliable are these estimates? In the absence of direct evidence, can we multiply the observed yearly rates by life expectancy to yield a lifetime risk for individual patients? In the presence of the wide divergence between the 2 studies, which number should be used? Can resulting lifetime estimates justify risky preventive interventions? Addressing these clinical questions with multivariable analyses and prognostic models, especially those based on nonrandomized studies, is the general topic of this article. The danger of extrapolation is also an important issue to be addressed.

General Issues in Estimating Prognosis

Observational studies are commonly used to explore potential prognostic factors in the occurrence of events or diseases or to evaluate prognosis in patients diagnosed with a specific disease or condition. The aim is to develop a model (or an equation) that expresses risk in relation to multiple risk factors and to estimate the risk for an individual patient. How much faith should we generally put in the results of such studies? A particular concern is the fact that many patients will have undergone treatments, which of course can impact prognosis, and that the choice of treatment was very likely influenced by the same prognostic factors we wish to study. In the absence of randomized allocation of treatments, it becomes impossible to disentangle the effects of the prognostic factor from the effects of treatment. As Byar claimed, "I have yet to see an analysis . . . that had really good information on why some patients got one treatment and others got another."[4]

As history has repeatedly shown, for example in the assessment of the role of hormone replacement therapy in the prevention or causation of cardio- and cerebrovascular diseases, the reasons for being at risk for the occurrence of a particular disease or outcome and the reasons for opting for a certain treatment, habit, or practice may share a confounding factor that will misleadingly impact the results and interpretation of the study.[5]

There are many ways to attempt to reduce or compensate for the effects of confounding factors in the evaluation of statistical data. These statistical adjustments involve many assumptions, including the following: 1) knowledge of which variables to take into account, 2) reliable data on those variables for each patient, and 3) using those variables appropriately to make the adjusted treatment comparison. According to Moses,[6] unfortunately "we are likely to fail on all three accounts." If those concerns are likely to be less serious when the aim is purely to evaluate prognosis because the impact of treatment is less than the range of prognoses associated with patient factors such as extent of disease or age, they become paramount when we wish to compare the effectiveness of treatments from a nonrandomized study. Although it is valuable to know the overall risk among all patients in a cohort, in most circumstances, risk will vary according to factors we would like to identify. While in some cases, risk may be related to a single variable, more often risk prediction incorporates multiple variables.

The general aim is to develop a "prognostic model" that yields an equation that will hopefully enable the estimation of risk in relation to multiple risk factors.[7] Statistical methods primarily focus on producing a model that, in some sense, best fits the data from the available sample of patients from the past. However, the primary goal should be the identification of a model that predicts outcome for the future patients, those who will be the object of clinical decisions. A model may make excellent predictions for past patients but may provide rather poor projections for future clinical decisions. In other words, the model needs to predict well enough to be clinically more useful than misleading. Hence multivariable prognostic research involves 3 essential steps: 1) developing the prognostic model,[7] 2) validating its performance in other individuals,[8] and 3) assessing its clinical impact on the outcomes of patients.[9]

In developing a statistical model and assuming that suitable data are available, one must make important decisions. First one must select a set of clinically relevant candidate predictors for possible inclusion in the model, evaluate data quality, and decide what to do about missing data. Then a strategy for selecting the important variables in the final model must be chosen. One also needs to select the best way of modeling continuous variables. Hence results will reflect not only the data but specific choices of the investigator as well, and these are often arbitrary and debatable. This process is well-illustrated in the article from Helsinki in which several models, leading to various results are proposed to identify risk factors for hemorrhage in patients with brain AVMs.[2]

A prognostic modeling study makes multiple assumptions, including the following: 1) The sample is representative of all patients for whom the resulting model will be used. 2) No important predictors are missing. 3) Predictors are measured without error. 4) Any missing data are missing at random. 5) The effect of each predictor is additive on the modeling scale (there is no interaction). 6) The effect of continuous predictors is modeled correctly. The internal validity of prognostic studies can be assessed by checking a number of study characteristics and methodologic choices, summarized in the Table.

The study should be sufficiently powered to reduce the play of chance. Models derived from small samples will tend to be overoptimistic about predictive performance. One must remember here that the power of the study, too frequently insufficient, is driven by the number of events, not the number of patients. A widely used rule of thumb is that there should be at least 10 events per variable of interest for the study to be powered appropriately to the number of variables that will be examined. Even if that criterion is met, there will be considerable uncertainty about predictions from a study with a small number of events. In our examples, there were only 77 events in the Helsinki study and 89 in the Toronto study.[2,3] It should be remembered that though the modeling uses data from all patients in the cohort, the numbers with particular features, which split the population into various specific subgroups, can be very small.

Framework for assessing the internal validity of studies of prognosis[a]

Study Feature	Qualities Sought
Patients	Recruited at a common point (usually early) in the course of the disease Ideally complete (all eligible patients included) Source and selection of patients is explained Diagnostic and inclusion criteria well-described and defined Clinical and demographic characteristics fully described
Follow-up	Sufficiently long to be meaningful
Outcome	Appropriate Unbiased Assessed blinded to prognostic information Fully defined and known for all patients
Candidate prognostic variables	Fully defined, including method of measurement Available for all patients
Analysis	Continuous predictors analyzed appropriately Statistical adjustment for all important factors
Treatments	Randomized or at least standardized Fully described and explained

[a] Modified from Altman.[10]

Validation

A prognostic model is valuable when there is evidence that it performs well for patients not used to develop the model. A hierarchy of increasingly stringent strategies may be used for validation. First, internal validation can be carried out by using bootstrap/cross-validation or data-splitting methods.[8] Validation can also be tested on a second dataset from the same center, but from a different timeframe, in a prospective fashion. Finally and preferably, external validation on data from different centers, perhaps with different investigators, can be performed by using retrospective or prospective cohorts.

For various statistical or clinical reasons, a prognostic model may perform poorly when it is applied to other patients.[7-10] The predictions of the model may not be reproducible because of deficiencies in the study design or modeling methods used in the study in which the model was derived; if the model was overfitted; or if an important predictor is absent from the model. Poor performance in new patients can also arise from differences between the setting of patients in the new and derivation samples, including factors reflecting differences in health care systems, methods of measurement, or patient characteristics, including disease severity.[11,12] Models that have not been validated in other patients should ideally not be used in clinical practice because they are more likely to mislead. Unfortunately, the prognostic studies discussed here have not been validated.

Some Specific Concerns about the Natural History Studies of Brain AVMs

The first assumption that is violated in these 2 AVM studies is that all patients are at the same stage of the disease, usually an inception cohort very early in the course of the disease. A serious problem for chronic diseases is defining "time zero."[4] In both studies, it is unclear when the clock started (first admission, diagnosis, referral?). Theoretically at least, all these lesions are congenital malformations. One could that argue the clock starts at birth. Most studies will simply start counting follow-up at the time of diagnosis, but some lesions can present with revealing symptoms such as seizures and have a longer observation period without rupture than others, such as deep or posterior fossa lesions, which can present only with the first rupture. Finally, if treatment varies in relation to prognostic variables then the study cannot deliver an unbiased assessment of prognostic ability. The Toronto study[3] did not exclude patients who were partially treated with embolization (211/678, 31%). It is likely that the decision to treat is related to the same prognostic variables that are being studied (hemorrhagic presentation for example). The observed outcomes thus relate to a mix of treated and untreated patients, complicating interpretation unless treatment is ineffective. If treatment during the observational period is problematic, excluding treated patients is also problematic. The Helsinki study claims recruitment of patients during a period characterized by a policy of conservative management, yet more than half of the patients were treated and excluded from the cohort unless they had at least 1 month of follow-up between diagnosis and treatment.[2] The remaining patients are unlikely to be representative of all patients with AVMs.

There are other points to note in these studies. Not all patients had data on all variables of interest, so the sample size was reduced for the multivariable analysis. Also, because there was low power, some important variables may not have been identified in the multivariable analysis.

These methodologic concerns will have contributed to distorting, perhaps severely, these prognostic models of the so-called natural history of brain AVMs. Unfortunately, it is impossible to know by how much.

Risk Varying with Time in Time-to-Event Analyses

Statistical models exploring predictors assume that the hazard ratios are constant with time (hence allowing comparisons in the presence or absence of the putative predictor). For example, the relative hazard for patients with or without hemorrhage at presentation is presumed constant across the many years of follow-up. Available studies are too small to evaluate this assumption. There is no assumption that the actual or absolute hemorrhagic risk is constant. Indeed both AVM studies suggest that the risk of hemorrhage was highest in the early years after enrollment and decreased thereafter,[2,3] though the studies differ in the magnitude of those risks. Again the reliability of those conclusions is weak given the small number of events. Neither article presented any confidence intervals around their estimates of risk.

If we are to estimate the lifetime risk of hemorrhage, we need to observe patients during a very long period. If risk varies with time, assuming a constant risk could seriously under- or overestimate lifetime risks. The estimated risks are unadjusted for other patient variables. Although estimates of absolute risk can be obtained from regression models, they rarely are. They cannot be determined from published summary statistics.

Other Difficulties with Prognosis of Brain AVMs

Multiple important difficulties are other potential sources of bias. In addition to the issues discussed above, with studies collecting patients for decades, we would expect changing trends with time, for example in diagnostic accuracy (with the introduction and availability of modern imaging), in referral patterns to tertiary centers, in selection of patients for treatment or observation, and in indications for imaging or for treatment. These trends alter the very nature of the patients with AVMs included in the prognostic study and make the assumption that resulting numbers can apply to future patients very unlikely.

The available prognostic studies[2,3] provide limited evidence to support clinical decisions because the models have not been validated in different populations. We can, for example, compare results of those studies. They seem to agree that presentation with rupture or with a deep location increases the risk of future hemorrhages, but absolute yearly rates of hemorrhage diverge by a factor of 2. Furthermore, extrapolation of risks observed during a relatively small number of years to lifetime risks by multiplying the observed rate by the number of years the patient is expected to live is, to say the least, uncertain. When a measurement is included in a regression model, extrapolation for patients outside the range of the original data can be seriously misleading.[13] In general, observational studies cannot replace trials to justify risky preventive interventions.[14] Even if a reliable estimate of a lifetime risk of hemorrhage was available, clinicians would still need to compare this prognosis with the risks and efficacy of treatment for the very same patient, a comparison that will remain uncertain in the absence of data from randomized trials.

Conclusions

Few prognostic models are used in clinical practice, probably because most have not been validated. Demonstration that a prognostic model is valuable requires evidence that the model performs well for patients not used to develop the model. In the absence of validation studies, prognostic information should be treated with great care.

References

1. Mohr JP, Moskowitz AJ, Stapf C, et al. **The ARUBA trial: current status, future hopes.** *Stroke* 2010;41:e537–40
2. Hernesniemi JA, Dashti R, Juvela S, et al. **Natural history of brain arteriovenous malformations: a long-term follow-up study of risk of hemorrhage in 238 patients.** *Neurosurgery* 2008;63:823–9, discussion 829–31
3. da Costa L, Wallace MC, Ter Brugge KG, et al. **The natural history and predictive features of hemorrhage from brain arteriovenous malformations.** *Stroke* 2009;40:100–05
4. Byar DP. **Problems with using observational databases to compare treatments.** *Stat Med* 1991;10:663–66
5. Sackett DL. **The arrogance of preventive medicine.** *CMAJ* 2002;167:363–64
6. Moses LE. **Measuring effects without randomized trials? Options, problems, challenges**. *Med Care* 1995;33(suppl):AS8–14
7. Royston P, Moons KGM, Altman DG, et al. **Prognosis and prognostic research: developing a prognostic model.** *BMJ* 2009;338:b604
8. Altman DG, Vergouwe Y, Royston P, et al. **Prognosis and prognostic research: validating a prognostic model.** *BMJ* 2009;338:b605
9. Moons KGM, Altman DG, Vergouwe Y, et al. **Application and impact of prediction models in clinical practice.** *BMJ* 2009;338:b606
10. Altman DG. **Systematic reviews of evaluations of prognostic variables.** In: Egger M, Smith DG, Altman DG, eds. *Systematic Reviews in Health Care*. 2nd ed. London, UK: BMJ Publishing Group; 2001:228–47
11. Altman DG. **Systematic reviews of evaluations of prognostic variables.** *BMJ* 2001;323:224–28
12. Yap CH, Reid C, Yii M, et al. **Validation of the EuroSCORE model in Australia.** *Eur J Cardiothorac Surg* 2006;29:441–46, discussion 446. Epub 2006 Feb 13
13. Altman DG, Bland JM. **Generalisation and extrapolation.** *BMJ* 1998;317:409–10
14. Byar DP. **Why data bases should not replace randomized clinical trials.** *Biometrics* 1980;36:337–42