Continuous variables – be they outcomes, exposures or covariates – are common in clinical studies. They are frequently modified into categorical variables during their analysis. Pocock et al. [1] found that 84% of epidemiological articles from leading journals categorized continuous variables. Such a categorization could be done for several reasons [2]. It is commonly perceived that categorization makes it easier to report and interpret final results (‘X doubles the risk of Y’ vs. ‘The risk of Y doubles when X increases by 10 units’). Researchers may be uncomfortable assuming a linear relationship between a continuous variable and the outcome but are unfamiliar with methods of handling non-linearity. Researchers and analysts may have less experience in dealing with continuous variables and prefer to make them behave like the more familiar categorical ones. Finally, it is also possible that physicians and epidemiologists, who frequently categorize continuous measures during their routine life (hypertensive or not, dyslipidemic or not, etc.), instinctively transplant this training from the clinic or field to their analysis.

However, categorizing continuous variables can cause problems. The first is information loss. Zhao and Kolenol [3] found that analyses with categorized continuous variables required greater than 40% more patients for the same power as that achieved using continuous variables. Selvin [4] derives a formula to calculate the efficiency loss due to categorizing a continuous variable. Becher et al. [5] found that models with a categorized exposure variable removed only 67% of the confounding controlled when the continuous version was used.

Categorizing continuous variables may not only miss the message, it can also get it wrong. Under some circumstances, categorizing continuous variables can give biased results. In a simulation study, Taylor and Yu [6] found that categorizing one continuous variable can artificially make another variable appear associated with the outcome. Selvin [4] showed that the cutpoint chosen during the categorization of continuous variables significantly changed the calculated odds ratio. Royston et al. [2] found that the significant association of the ‘S phase fraction’ with cancer outcomes repeatedly came and went depending on which cutpoint was used to define ‘abnormal’. Ragland [7] showed similar findings with prevalence ratios of hypertension. Information loss and bias from categorizing continuous variables explain why statisticians frequently warn us to leave continuous variables alone [2, 8].

It appears that this advice was lost to investigators – ourselves included – who have developed risk stratification schemes for patients with atrial fibrillation (AF). Quantifying stroke risk in AF is essential for patient management: high-risk patients require oral anticoagulants while low-risk patients (who stand to have a minimal absolute benefit from treatment) can avoid such a therapy. Patient age is significantly associated with stroke...
risk [9]. Most AF risk stratification schemes have categorized patient age using arbitrary or data-driven cutpoints.

A study in Neuroepidemiology may illustrate the errors of our ways. Frost et al. [10] conducted a population-based administrative database analysis of all adult Danes discharged from hospital between 1977 and 2002 with a discharge diagnosis code for AF or atrial flutter. Patients with a previous diagnosis of AF, previous stroke or valvular heart disease (also identified by diagnostic codes in the hospitalization administrative database) were excluded. Patients were followed until they were admitted to hospital with any stroke or were censored (for death, emigration or end of study). Discharge codes for AF and stroke were validated in a small reabstraction study. They then constructed a series of survival models that adjusted for important covariates (hypertension, diabetes, hyperthyroidism, coronary artery disease and congestive heart failure) in which age was modeled as: 3 or more age categories; a linear term; a linear spline with different cutpoints, or a cubic polynomial (i.e. $\beta_1 \cdot \text{age}^2$ with or without $\beta_2 \cdot \text{age}^3$). They compared the explanatory value of the models using the global $\chi^2$ goodness-of-fit test and found that the model expressing age as a cubic polynomial best explained the data in both men and women, seemingly doing a much better job than models in which age was categorized into 3 groups.

The study is an efficient reminder of potential information loss when we categorize continuous variables prior to analysis. But like all clinical studies, it raises a number of important questions. Although models with smaller global $\chi^2$ goodness-of-fit statistic measures have greater explanatory capability than those with greater values, we do not know whether the difference is statistically significant or meaningful. While the study examined several modeling strategies, we wonder whether other modeling strategies for patient age (including natural splines, fractional polynomials or nonparametric techniques) might have had better results. Patient treatment with anticoagulant therapy – which decreases the risk of stroke by 62% [11] – was not considered in the model. The study did not quantify the ‘cost’ of categorizing patient age by, for example, measuring residual confounding in the categorized models compared to that with the cubic polynomial [5].

Finally, we do not know whether treating patient age as a continuous rather than a categorical variable meaningfully changes stroke risk prediction in AF patients. To clinicians, this is the most useful application of such statistical models and is used for deciding who should and should not receive oral anticoagulation therapy. Of course, an accurate model is best when making such a decision. However, the predicted stroke risk from ‘new and improved’ models must differ extensively from previous models for treatment choice in a group of patients to meaningfully change. Otherwise, the cruder models in which age was treated as a categorized variable would probably be more useful since they are invariably easier to use at the bedside when estimating the risk score for a particular patient.

References