The current publication by Chen and colleagues\(^1\) presents a model aimed at predicting overall survival and disease-free survival in patients with gastric cancer. They argue that the current staging system, which uses routinely collected data, is limited in its ability to predict outcomes and, importantly, to predict potential benefits of chemotherapy. They focus on building a tumor-associated collagen signature (TACS) using data from the tumor microenvironment. Collagen is a major component of that tumor microenvironment, which can influence both cancer progression and response to chemotherapy.

The authors used a sophisticated statistical modeling approach that allows consideration of large numbers of potential predictor variables, with limited overfitting of the models. (When models are overfit to a specific data set, they can be excellent for prediction in the current sample but not nearly as good at prediction in other populations.) They developed the model in a cohort from a single institution in China and validated the model (more on this later) in a separate cohort from another hospital.

The model included 9 features in the TACS. A higher TACS was associated with worse outcomes, regardless of treatment. When compared with a model that contained only the clinicopathological variables, the model that included TACS showed a statistically significantly better ability to predict outcomes. There are some nuances to these findings that require a bit of technical background.

First, it is helpful to keep in mind the difference between prognostic models, which are aimed at predicting natural outcomes among identified subgroups, and explanatory models, which are usually aimed at drawing causal inferences about the relationships between particular variables and outcomes. In developing prognostic models, the goal is accurate prediction—can we correctly determine, in advance, who will have a favorable or unfavorable outcome? In these models, we tend not to worry about causal interpretation, so issues such as confounding are not that pertinent. In contrast, explanatory models worry a great deal about confounding in the interest of determining which variables cause the outcomes. A third model, relevant to this study, is the so-called predictive model (as opposed to a prognostic model). Predictive models aim to predict outcomes from treatment rather than the natural outcomes predicted by prognostic models. Factors identified in predictive models are sometimes called treatment biomarkers. A key point is that, for all 3 types of models, constructing and interpreting the models should not be based solely on \(P\) values.

One of the more useful recent articles on prognostic models\(^2\) proposes a framework for developing such models. The steps involved in model development, according to that framework, are (1) defining the problem explicitly; (2) selecting data sets that can inform the problem; (3) extracting and constructing variables to be considered for the models; (4) developing (also known as learning) the predictive model; and (5) validating the performance of the model.

Chen et al\(^1\) did an excellent job of fulfilling these criteria. Where the nuance enters has more to do with the interpretation of the results. The \(C\) statistic, which the authors present for their various models, is a measure of discrimination, ie, the ability of the model to distinguish between those who will go on to have the event of interest and those who will not. This is also called the area under the receiver operator characteristic curve (AUC). The AUC is "the probability that a randomly chosen patient with the outcome will be assigned a higher risk of the outcome by the model than a randomly chosen patient without the outcome."\(^2\) A related, but different, term is calibration, which measures how closely the predicted probabilities from the model match the observed probabilities.
Chen et al1 present calibration curves, which, on visual inspection, appear to confirm the model's ability to make accurate prognoses. The C statistics reported by the authors did show a statistically significant improvement in discrimination with the addition of TACS to the clinicopathological variables. The question is, what magnitude of change in the C statistic, regardless of statistical significance, is important clinically? For disease-free survival, the C statistic was 0.80 for the full model vs 0.78 for the clinicopathological model alone ($P = .03$). For overall survival, the corresponding numbers were 0.81 vs 0.80 ($P = .03$). Results for the validation cohort showed a similar pattern. The small increment in the C statistic raises the question of whether the addition of TACS to the smaller model is worth the extra effort and expense. As a nonclinician and a clinician, we are not convinced.

When developing both prognostic and predictive models, it is always essential to confirm the performance of the model in data sets that were not used to develop the model. In general, the concern is that models (even with the sophisticated technology used by the authors) tend to perform better in the data sets on which they were developed than on other data sets. Validation can be internal, using a split-sample from the same population, usually chosen randomly from the full sample. The model is developed on one of the samples and then tested on the other. This can be a somewhat weak validation because we know (by definition) that any variability between results is purely random (and therefore is not expected to be large). There are varying degrees of external validation, depending on how different the validation population is from the original sample. Chen and colleagues1 used data from another hospital in China to perform their validation, and the results were reassuring that the model applied equally well to the new sample as it did to the development sample. One might wish to do further validation using data from more recent time periods, from other countries, or both.

From a practical perspective, thinking about how best to treat patients, our perspective points to predicting response to chemotherapy. Chen and colleagues1 used statistical interaction tests to address this question. Interactions test whether the association between chemotherapy and outcome is the same, in this case, for patients with high vs low TACS values. Specifically, their analyses showed that the protective association between chemotherapy and outcome was stronger in those with low TACS than in those with high TACS. The extent of this difference in associations depended on stage. For patients with stage 2 disease, there were protective associations, numerically, in the high TACS group, but they were not statistically significant and smaller in magnitude than the associations in the low TACS group (eTable 12 in the Supplement of the study by Chen et al). For patients with stage 3 disease, the hazard ratios for chemotherapy were close to the null value of 1.0 in the high TACS group, whereas there remained a favorable association in the low TACS group.

The ability to predict response to therapy is valuable. If clinicians can focus chemotherapy on the patients most likely to benefit, they can avoid the toxic effects of chemotherapy in those least likely to benefit. Further analyses would be needed to understand whether TACS provides unique value in this prediction, above and beyond what one might expect based on the clinicopathological findings. (This interaction analysis was not conducted in the study by Chen et al.)

In summary, Chen and colleagues1 have done an excellent job of showing that TACS may be a useful addition in helping understand gastric cancer prognosis. The additional value of TACS seems small in this setting, however. Furthermore, patients with low TACS seem to be more likely to respond to chemotherapy, but this needs further prospective validation in independent data sets. Lessons learned from the statistical modeling within this study could be applicable to an emerging biomarker of interest, namely circulating tumor DNA sequencing, which is performed postoperatively in various cancers including gastric cancer to identify minimal residual disease, to risk stratify patients, and to potentially determine therapy in the context of already understood prognostic and predictive clinicopathologic variables.