Developing a Risk Prediction Model

I'm going to now talk about model development and the considerations that I'd be thinking about for that. So, model development typically requires three things; a need for a model, some data that we can use to develop the model, and some methods that we can use to develop a suitable model using the data to address the intended purpose.

Because if we're developing a prediction model we need to think about, first of all, is a model actually needed? So, what are we using the model for? What's the actual purpose? So, the problem, I think, is that building prediction models is, for people that know a little bit of statistics or machine learning and have some data, it's a very quick and easy thing to do. We have one variable that we call the outcome and we predict it from the others.

But, of course, for a model that we're actually going to use in practice, we have to think very carefully about what a model might be needed for. So, either a sort of concurrent outcome, in which case we’re building a diagnostic model. So, developing a model that helps us diagnose a current condition of a patient. Or we're interested in future outcomes; a prognostic model.

And the key thing is that the prediction that we make has to be useful for something. So, with cancer risk models, this is often the decision that's taken based on the risk that we record. If, for example, somebody is at high risk of breast cancer, then we might increase the intensity of screening for that individual, an actual change in the care pathway for that patient based on this model. So, we're not just measuring the risk of an outcome for the sake of it.

So, the second thing is, if we decide we need a model, then we need to identify some data upon which to build such a model. So, we need to find some patients on which we've measured some risk factors, some variables, and for whom we know what actually happened to them, so we know their outcomes.

Then the next step is we need to somehow prepare our data, thinking about what our model should look like and the data that we have available to build that model, with a view to what data are we going to have in future on which to base such a model.

And we also need to think about issues with the data such as missing-ness. So, if we have variables with lots of missing-ness, do we want to input or do we want to not use that variable? So, what we're now doing essentially is building the algorithm. The final step that's usually included in the development stage is what we call internal validation. There are key points at the bottom as well, I'll be looking for the papers that report how this development process happened. And this should all be very clearly reported, and there’s guidelines, such as tripods, to ensure that.