*Nature Reviews Cancer* | AOP, published online 3 May 2013; doi:10.1038/nrc3530

PROGNOSTIC MODELS

# Rising to the challenge

> access to defined data sets helps to limit potential bias

The development of models that can accurately predict patient survival on the basis of a number of clinical, molecular and genomic variables is much sought after. Can using the 'wisdom of the crowds' help?

Adam Margolin, Stephen Friend and colleagues set up the Sage-Bionetworks–DREAM Breast Cancer Prognosis Challenge (BCC). The 354 registered participants developed computational models that predict overall survival in patients with breast cancer based on data from the METABRIC data set (1,981 samples), which contains clinical information such as age and tumour size, along with mRNA expression and DNA copy number variation (CNV) data. The BCC participants were given Web access to data from 1,000 samples to develop and train their models, and these were submitted for evaluation on two withheld data sets consisting of the remaining 981 samples from the METABRIC data. The concordance index (CI) of 1,400 submitted models was calculated, and 83 models were further assessed against a new data set (OsloVal).

The winning model was conceived by Dimitris Anastassiou, Wei-Yi Cheng and Tai-Hsien Ou Yang and used an approach that they had previously developed called the 'Attractor metagene' model. Attractor metagenes are the weighted average of signatures of co-expressed genes that are identified by an iterative approach using data-rich gene expression sets from several tumour types. These authors were keen to find out whether three metagenes — a CIN metagene associated with chromosomal instability; a mesenchymal transition (MES) metagene associated with migration and invasion; and a LYM metagene associated with lymphocyte-specific recruitment — also worked with the METABRIC data.

From the METABRIC data the authors identified similar metagenes, which they termed the CIN feature, the MES feature and the LYM feature for breast cancer. The ten genes in the CIN feature were associated with a poor prognosis and were also associated with tumour grade. The MES feature contained ten genes associated with tumour stage: tumours that have not reached invasive stage 1 carcinoma *in situ* do not express genes in the MES feature. The ten genes in the LYM feature were associated with a good prognosis in patients with oestrogen receptor (ER)-negative breast cancer, but conferred a poor prognosis in patients with ER-positive disease and lymph node invasion.

The authors' ensemble model for the BCC contained the CIN feature score, the MES feature score based on data from tumours of less than 30 mm in size and with no positive lymph nodes, the LYM feature data generated from ER-negative patients and a new metagene that contains the expression values of two genes (*FGD3* and *SUSD3*) that are associated with a good prognosis. When combined with the number of positive lymph nodes and age at diagnosis, this model could predict overall survival with a CI of 0.7526 (that is, the model could be used to correctly predict who will survive longer for every three of four pairs of patients in the OsloVal data set). Interestingly, the CNV data were not used, as they did not improve the predictive capability of the model when the CIN feature genes were included.

PHOTODISC

Although it remains to be seen how much the 'wisdom of the crowds' will be used in future cancer-relevant studies, both sets of authors state that access to defined data sets helps to limit potential bias that is associated with developing, training and validating models on restricted data sets and may help to minimize overfitting of the data.

*Nicola McCarthy*

**ORIGINAL RESEARCH PAPERS** Margolin, A. A. *et al.* Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl Med.* **5**, 181re1 (2013) | Cheng, W.-Y., Yang, T.-H. O. & Anastassiou, D. Development of a prognostic model for breast cancer survival in an open challenge environment. *Sci. Transl Med.* **5**, 181ra50 (2013)