BJC

British Journal of Cancer (2016) 114, e5 | doi: 10.1038/bjc.2016.1

Comment on 'Circulating cell-free miRNAs as biomarker for triple-negative breast cancer'—Methodological challenges in combining miRNAs as circulating biomarkers

P Verderio*[,1], S Bottelli[1], M Lecchi[1], M Plebani[1], M Gariboldi[2,3], S Pizzamiglio[1] and C M Ciniselli[1,4]

[1]Unit of Medical Statistics, Biometry and Bioinformatics, Fondazione IRCCS Istituto Nazionale dei Tumori, via G. Venezian 1, 20133 Milano, Italy; [2]Molecular Genetics of Cancer, Fondazione Istituto FIRC di Oncologia Molecolare, via Adamello 16, 20139 Milano, Italy; [3]Department of Experimental Oncology and Molecular Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori, via G. Venezian 1, 20133 Milano, Italy and [4]Department of Clinical Sciences and Community Health, Università degli Studi di Milano, via G. Venezian, 1, 20133 Milano, Italy

Sir,

We read with great interest the work published by Shin *et al* (2015), which highlights the potential relevance of circulating cell-free miRNAs as biomarkers for the detection of triple-negative breast cancer (TNBC). Of importance, the authors identified three miRNAs (miR-16, miR-21 and miR-199-5p) as potential diagnostic biomarkers for TNBC. The information provided is of interest as the identification of miRNA signatures for TNBC, as well as for other types of cancer (Calin and Croce, 2006), is of increasing relevance. However, we found some worthwhile issues that need to be discussed. The authors' conclusions seem to be based only on results obtained from a univariate analysis performed for each of the above mentioned miRNAs. Specifically they performed a receiver–operator characteristics (ROC) curve to assess their ability to discriminate TNBC patients from healthy controls. Results showed a considerable discriminatory performance for each of the three miRNAs. Although the authors reported in the statistical analysis section the following sentence: 'Multivariate logistic regression model was established and leave one-out cross validation to find the best logistic model', no results were provided in multivariate terms. The lack of assessment of the more intriguing level of diagnostic accuracy achievable by combining the three miRNAs in a composite score is a relevant drawback of the paper. This topic, that actually represents one of the most critical steps in developing a miRNA-based signature in cancer research, implies some methodological considerations directly related to the multivariate regression models theory (Harrell, 2001). Multivariate regression models allowing simultaneous association of miRNAs and predictors with clinical outcome, such as logistic regression for presence/absence of disease, are common building blocks of biomarker-based risk prediction tools. It should be considered that in such scenario the number of observations is not generally of the order of magnitude greater than the number of variables. Results from the multivariate regression models may thus be affected by the small number of events per variable (Verderio, 2012). As a consequence, the model may produce over-optimistic estimation of the combined area under the curve (AUC) on the original data, but fails when applied in an independent data set (Verderio *et al*, 2010). In addition, to better generate prediction and generalisation to new data, the model should be defined according to the principle of parsimony, which is essential in discriminating the structural part (signal) of empirical data from the idiosyncratic (noise) one (Vandekerckhove *et al*, 2015). Although different approaches had been described in the literature to find the optimal linear combination of putative miRNAs to maximise the AUC (Su and Liu, 1993; Pepe and Thompson, 2000; Kang *et al*, 2013; Yan *et al*, 2015), we believe that it is urgent to delineate a procedure that is methodologically as robust as flexible to cover this fundamental step.

To this end, we are developing a comprehensive procedure that, starting from a set of potential miRNAs, identifies a more powerful and parsimonious composite score. Briefly, the best combination of the potential miRNAs is reached by resorting to penalised maximum likelihood estimation (PMLE) regression methods (Harrell, 2001) that can provide more reliable results in the presence of large numbers of input variables. A more parsimonious final model was then obtained using a step-down procedure as suggested by Ambler *et al* (2002).

As example, for illustration purpose only, we applied our procedure in a similar context of Shin *et al* (2015), to data on circulating miRNAs in plasma from 20 hepatocellular carcinoma (HCC) patients and 20 healthy donors (GSE50013) retrieved from the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/gds). By applying our NqA algorithm (Verderio *et al*, 2014), four miRNAs were identified as potential diagnostic biomarkers for HCC. As reported in Table 1, the AUC value observed for each of these miRNAs ranged from 0.739 to 0.841. Interestingly, by combining these miRNAs with the PMLE approach, we observed a sensible increment of the predictive capability with an AUC value of 0.953. In addition, we obtained a more parsimonious model based only on three miRNAs (AUC = 0.923) without the loss of discriminatory power. A similar AUC value (AUC = 0.920) was observed by applying the least absolute shrinkage and selection operator (LASSO) method (Tibshirani, 1996). Notably, the two approaches retained the same three miRNAs.

In conclusion, this example shows that a more appropriate way to get the information for the evaluation of miRNAs as biomarkers could be interpreting their predictive role in a multivariate fashion or following Collins *et al* (2015), that 'Prediction is inherently multivariable'.

This suggests the need of resorting to statistical procedures, generally based on advanced methods, in order to properly embrace the complexity of the data with the ultimate aim of better predicting the presence/absence of disease.

Table 1. Estimated AUC and 95% confidence interval (CI) of the considered model

Model	AUC (95%CI)
Univariate logistic models	
hsa-miR-1274B-002884	0.761 (0.606; 0.916)
hsa-miR-483-5p-4395449	0.841 (0.722; 0.961)
MammU6-4395470	0.811 (0.677; 0.946)
U6 snRNA-001973	0.739 (0.585; 0.893)
Multivariate logistic models	
Full PMLE[a]	0.953 (0.893; 1.000)
Reduced PMLE[a]	0.923 (0.845; 1.000)
LASSO[b]	0.920 (0.841; 0.999)

[a]Penalised maximum likelihood estimation.
[b]Least absolute shrinkage and selection operator.

ACKNOWLEDGEMENTS

This work was supported by Grants from Associazione Italiana per la Ricerca sul Cancro (AIRC, Grant 10529 to MA Pierotti and 12162 to G Sozzi).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Ambler G, Brady AR, Royston P (2002) Simplifying a prognostic model: a simulation study based on clinical data. *Stat Med* **21**(24): 3803–3822.

Calin GA, Croce CM (2006) MicroRNA signatures in human cancers. *Nat Rev Cancer* **6**(11): 857–866.

Collins GS, Reitsma JB, Altman DG, Moons KGM (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* **162**(1): 55–63.

Harrell FE Jr (2001) *Regression Modeling Strategies*. Springer-Verlag: New York.

Kang L, Liu A, Tian L (2013) Linear combination methods to improve diagnostic/prognostic accuracy on future observations. *Stat Methods Med Res*; e-pub ahead of print 16 April 2013; doi:10.1177/0962280213481053.

Pepe MS, Thompson ML (2000) Combining diagnostic test results to increase accuracy. *Biostatistics* **1**(2): 123–140.

Shin VY, Siu JM, Cheuk I, Ng EK, Kwong A (2015) Circulating cell-free miRNAs as biomarker for triple-negative breast cancer. *Br J Cancer* **112**(11): 1751–1759.

Su JQ, Liu JS (1993) Linear combinations of multiple diagnostic markers. *JASA* **88**(424): 1350–1355.

Tibshirani R (1996) Regression shrinkage and selection via the LASSO. *J R Stat Soc B* **58**(1): 267–288.

Vandekerckhove J, Matzke D, Wagenmakers E (2015) Model comparison and the principle of parsimony. In *The Oxford Handbook of Computational and Mathematical Psychology*, Busemeyer JR, Wang Z, Townsend JT, Eidels A (eds). Oxford University Press: New York, pp 300–319.

Verderio P, Mangia A, Ciniselli CM, Tagliabue P, Paradiso A (2010) Biomarkers for early cancer detection - methodological aspects. *Breast Care (Basel)* **5**(2): 62–65.

Verderio P (2012) Assessing the clinical relevance of oncogenic pathways in neoadjuvant breast cancer. *J Clin Oncol* **30**(16): 1912–1915.

Verderio P, Bottelli S, Ciniselli CM, Pierotti MA, Gariboldi M, Pizzamiglio S (2014) NqA: an R-based algorithm for the normalization and analysis of microRNA qPCR data. *Anal Biochem* **461**: 7–9.

Yan L, Tian L, Liu S (2015) Combining large number of weak biomarkers based on AUC. *Stat Med* **34**: 3811–3830.

*Correspondence: Dr P Verderio; E-mail: paolo.verderio@istitutotumori.mi.it

Published online 21 April 2016