New approaches but the same flaws in the search for prognostic signatures

Running title: Development and validation of gene expression profiles

Ramon Salazar¹ and Josep Tabernero²

¹Department of Medical Oncology and Translational Research Laboratory, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL), L’Hospitalet de Llobregat, Barcelona, Spain.

²Medical Oncology Department, Vall d’Hebron University Hospital and Institute of Oncology (VHIO), Universitat Autònoma de Barcelona, Barcelona, Spain

The authors have no relevant affiliations or financial interest in or financial conflict with the subject matter or materials discussed in the manuscript

SUMMARY:
A six-gene prognostic colorectal cancer hypoxia score was generated from expression data from in vitro experiments and microarray datasets and was validated in two distinct patient cohorts. The approach followed by the authors is original and biologically sound but could be limited by potential biases and other methodological limitations.

TEXT
In this issue of Clinical Cancer Research, Derkevel and colleagues¹ describe a novel mechanistic approach to the development of a new prognostic gene signature for stage II and III colorectal cancer (CRC) in the clinical setting. CRC continues to be the third most common cancer worldwide². As with most cancers, survival is directly dependent on the stage of the disease.

The American Joint Committee on Cancer classification is an invaluable tool, for most neoplastic diseases, to assess a particular patient’s risk based on disease stage and to evaluate the need for adjuvant treatment. Nevertheless, the limitations of the current staging system may hamper our ability to provide the best clinical care to our patients as the clinical decision to administer adjuvant chemotherapy is mainly determined by this clinicopathological staging tool and does not take molecular biology features into consideration³. Adjuvant treatment in CRC is current standard treatment for stage III but still controversial in stage II. Current prognostic factors used in the clinical setting are insufficient to identify those patients with stage II CRC at high risk of recurrence or those patients with stage III CRC at low risk, thus leading to potential under-treatment or over-treatment with adjuvant chemotherapy. With the aim of solving this clinical conundrum, a large number of gene-expression profiling-based assays have been developed over recent years as prognostic signatures using supervised
analytical methods. These approaches involve an initial training step in pre-defined groups of patients, for instance, relapse versus non-relapse, to identify gene expression signatures with predictive capability. A large number of prognostic gene profiles have been discovered and developed effectively in the clinical setting, but only two of them belong to commercial platforms that have succeeded in validating their signature in clinically relevant, multivariate, independent analyses.

Dekervel et al. report on the development of yet another prognostic gene expression signature that could potentially have clinical implications in the crowded field of CRC prognostic signatures. The mechanistic approach provides a proof of concept affirming the importance of the tumor microenvironment and, in particular, hypoxia in the biological behavior of the tumor. The steps performed which led to the development of the colon cancer hypoxia score (CCHS) are innovative and may yield promise. Expression data of an \textit{in vitro} hypoxia experiment were combined with microarray data sets. Subsequently, the six-gene score obtained was validated in two independent patient cohorts. Patients with low CCHS showed significantly better disease-free survival at three years (77.3%) with respect to high CCHS patients (46.4%) (Log rank, \textit{p}=0.006) in the primary clinical validation patient cohort from the author’s centre. Patients with a favorable CCHS were also more likely to be alive at three years (estimated overall survival 81.5\% vs 66.6\%, \textit{p}=0.044) and at five years (71.9\% vs 59.1\%, \textit{p}=0.036). In addition, CCHS (high versus low) could significantly predict disease-free survival within stage II and III separately. This was independently confirmed in an external public dataset of 90 stage II patients where the CCHS identified two subgroups of stage II patients with distinct DFS at three years (86.9\% vs 52.2\%, \textit{p}<0.001) and five years (86.9\% vs 52.2\%, \textit{p}<0.001). However, the approach followed by the authors could be limited by potential biases in the selection of the public data used in the training set, and standard rules for clinical validation still need to apply. A few methodological issues merit further detailed comments. First, differentially expressed hypoxia genes were analyzed in one single CRC cell line, CaCo-2. This single-cell approach limits gene selection to the individual molecular background of this particular model, which may or may not be representative of the far more heterogeneous clinical setting. The authors are aware of this, since they added \textit{TP53} at a later point in the clinical training of their signature based on prior literature interpretation but not on their data (\textit{TP53} gene is mutated with undetectable expression levels in the CaCo-2 cell line). Second, training the model \textit{in silico} also involves important limitations. Clinical information is limited to what is publicly available, which is usually heterogeneous and biased towards particular populations. Datasets usually differ in patient characteristics, inclusion criteria and outcome.
definition. Amongst the four-gene expression omnibus (GEO) series used for the training in this paper, one has scarce clinical information and even lacks a recurrence outcome report (GSE13294, Table 1). Nevertheless, this is the only one of the 4 datasets used that has information on MSI status, which was arguably used as a surrogate variable for clinical outcome. MSI status is unknown in the remaining 3 series, which makes the author’s assumptions underlying use of MSI as a surrogate parameter for clinical outcome even more unorthodox. In addition, the majority of the published prognostic signatures have reportedly performed well in two of the four datasets used in the training of the CCHS (GSE17537 and GSE5206), but not in the other public datasets, which casts doubt on the usefulness of these series for such purposes. As a corollary, an obvious risk of this in silico training approach is data set selection bias, which is a critical point in any genomic high-throughput analysis design and subsequent data interpretation. In addition, the inclusion of the p53 expression in the signature training process is based only on literature and not data, which is a questionable decision as the real role of p53 expression in prognosis of loco-regional stage CRC is not clear and not recommended in any guidelines or expert consensus publication.

Regarding clinical validation of the score, the authors analyzed a series of 71 patients with stage II and 55 patients with stage III CRC from their own center in Leuven. They extracted tumor RNA from FFPE samples and analyzed it with the nCounter technique (NanoString Technologies, Seattle, WA). They then applied a multivariate analysis involving CCHS and relevant clinicopathological variables but not MSI status, which is a recognized molecular prognostic biomarker used routinely in the clinical setting. Stage (III versus II), extramural venous invasion (yes versus no), and CCHS (high versus low) were the 3 significant independent prognostic factors with hazard ratios (HRs) of 2.58, 3.21 and 2.58, respectively. Another major limitation of this analysis is that the authors do not separate this analysis by stage of the disease, and some of the candidate clinicopathological biomarkers are more relevant to prognosis of patients with stage II than stage III. The AMC-AJCCII-90 data set (GSE33113), which consisted of 90 patients with stage II CRC treated at the Academic Medical Center in Amsterdam, was used as a second independent clinical validation of the CCHS. This public dataset includes information describing clinical outcome such as relapse (yes/no), age, gender and primary location but no other potentially prognostic clinicopathological features or MSI status. Tumor samples from this series were fresh-frozen and analyzed with the Affymetrix Human Genome U133 Plus 2.0 Array. CCHS did indeed perform well in this series but a multivariate analysis with other relevant clinicopathological prognostic factors and MSI status was not feasible and would be required to infer potential future clinical utility of the score.
Importantly, CCHS was validated in these two independent groups of patients using optimal non pre-specified cut-off values (ROC curves) without internal cross-validation for each dataset. The authors justify this because of the distinct techniques used to measure the gene expression (nCounter vs. Affymetrix microarray), which is a fair argument. However, this downgrades their study to a previous step in the validation process (see Figure), closer to what is defined by Simon et al. as “translation of platforms” or “assay reproducibility” rather than to clinical validation \(^9\).

Despite all the limitations described above, one strength of the work by Dekervel et al. is its robustness on different platforms and tumor specimens (fresh-frozen and FFPE samples). In addition, many methodological issues are shared by the vast majority of prognostic signatures previously published. A paper exploring the performance of 31 published gene expression signatures in eleven CRC gene expression datasets concludes that although most of the published signatures show significant statistical association with prognosis, their ability to accurately classify independent samples into high-risk and low-risk groups is limited. This could be explained both by methodological weaknesses and intra-tumor heterogeneity. It is well known that CRC tumors are composed not only of tumor epithelial cells but also of cancer associated stromal fibroblasts (CAFs), endothelial cells or inflammatory cells, among others. This heterogeneity in tumor cell populations might dilute the prognosis signal of reported gene expression signatures\(^7\).

Finally, the rationale behind this innovative approach to signature discovery is biologically plausible because hypoxia can be one of the drivers of epithelial mesenchymal transition (EMT) and it is well known that EMT plays an important role in cancer progression \(^10,11\). In addition, recent hierarchical clustering of whole-genome expression data has provided new insights into the biologic and prognostic heterogeneity that supports this concept: subtyping studies on several groups also confirmed that upregulation of the EMT phenotype, matrix remodeling and TGF-Beta pathway is clearly associated with increased risk of relapse and/or reduced overall survival. Poor prognosis has also been observed in stem-like subtypes with overexpressed genes \textit{associated with} mesenchymal and stem cells. By contrast, subtypes associated with an epithelial phenotype showed a better prognosis\(^12\). In conclusion, this manuscript by Dekervel et al, reports on a new candidate prognostic score based on a hypoxia induced phenotype and a novel methodological approach that requires validation. Despite the originality of the approach, a number of limitations are apparent that need to be addressed in future studies.

Acknowledgements: The authors would like to thank Rebeca Sanz-Pamplona for critical reading of the manuscript and Stephen Kelly for English editing.
References:


FIGURE: Pathway/Steps for developing and validating prognostic multigene classifiers. *CaCo-2 Cell Line experiment ** Smith VMC, Jorissen, Aronow *** Smith MCC ****Leuven cohort *****AMC-AJCC-II does not meet this criteria
1: Developing a genomic classifier by a supervised approach

1A Biological hypothesis

E.g. In vitro hypoxia model

High throughput

1B Agnostic approach

Clinical outcome

2: Internal validation

2A Split sample into training and test

Gene list optimization

2B Cross validation: e.g. LOOCV

3: Translation of platforms (assay reproducibility): e.g. rtPCR, RNA seq

External independent datasets

Pre-specified scoring method

4: Clinical validation

Multivariate analysis vs other clinically relevant biomarkers

Figure 1:
New approaches but the same flaws in the search of prognostic signatures

Ramon Salazar and Josep Tabernero

Clin Cancer Res Published OnlineFirst February 28, 2014.