

December 16, 2010

Steven McKinney, Ph.D.
Statistician
Molecular Oncology and Breast Cancer Program
British Columbia Cancer Research Centre
675 West 10th Ave, Floor 4
Vancouver B.C.
V5Z 1L3
Canada
Email: smckinney@bccrc.ca
Tel: 604-675-8000 x7561

Christine M. Micheel, Ph.D.
Program Officer
Board on Health Care Services and National Cancer Policy Forum
Institute of Medicine
500 5th Street, NW, 767;
Washington, DC 20001
USA
Voice: 202 334-1402 and 202 603-3105
Fax: 202 334-2647
Email: cmicheel@nas.edu

Dear Dr. Micheel,

I have been following with interest and concern the development of events related to the three clinical trials (NCT00509366, NCT00545948, NCT00636441) currently under review by the Institute of Medicine (Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials).

I have reviewed many of the omics papers related to this issue, and wish to communicate my concerns to the review committee. In brief, my concern is that the methodology employed in the now retracted papers, and many others issued by the Duke group, all use a flawed statistical analytical paradigm. Essentially the paradigm involves fitting a statistical model to all available study data, then splitting the data into subsets, labeling one of them a "training" set, another a "validation" or "test" set, and showing that the statistical model works well for both sets. The analysis paradigm is described as a statistical train-test-validate exercise in several published papers, though it is technically not a true train-test-validate exercise as the model under evaluation involves predictor components derived from the full data set.

I believe that this issue needs to be investigated as part of the Institute of Medicine's review, because concerned readers who have written letters to journal editors have not

been successful in educating a wider audience (in particular journal editors and reviewers of biomedical journals) as to the problematic aspects of the analysis method that are repeatedly used by the Duke group. The issue at hand is not just one researcher who committed errors in one analysis, but rather the systematic use of a flawed analytical paradigm in multiple papers discussing personalized medicine in a widening scope of medical scenarios.

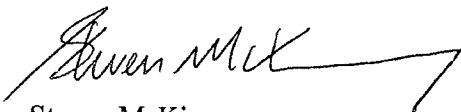
The statistical properties of this analytical paradigm, in particular its type I error rate, have not to my knowledge been reviewed or published. I respectfully request the IOM committee to include this issue in its agenda for the upcoming review, as findings from this committee will provide a broader educational opportunity, allowing journal editors and reviewers to have a better understanding of the statistical properties of the analyses repeatedly developed and submitted for publication by the Duke University investigators.

As a citizen of the United States and a taxpayer, and as a practicing biomedical applied statistician, I am especially concerned about the possibility that the funding garnered for such potentially flawed studies is detracting from other groups' ability to obtain funding to perform valid research in the valuable arena of personalized medicine. Additionally, the use of human subjects in ongoing studies involving this methodology is ethically problematic.

I will discuss this issue in greater detail in the attached Appendix to this letter of concern, and provide citations to the literature illustrating the various aspects involved.

Thank you for your consideration of this matter.

Yours sincerely

A handwritten signature in black ink, appearing to read "Steven McKinney", with a long horizontal stroke extending to the right.

Steven McKinney

Attachments: Appendix – Details of points of concern regarding the statistical analytical paradigm repeatedly used in personalized medicine research papers published by Duke University investigators.

Appendix – Details of points of concern regarding the statistical analytical paradigm repeatedly used in personalized medicine research papers published by Duke University investigators

In 2001 West et al. [1] published some details of a statistical analytical method involving “Bayesian regression models that provide predictive capability based on gene expression data”. In the Statistical Methods section of this paper they state that the “Analysis uses binary regression models combined with singular value decompositions (SVDs) and with stochastic regularization by using Bayesian analysis (M.W., unpublished work), as discussed and referenced in Experimental Procedures, which are published as supporting information on the PNAS web site.”

Given the current state of affairs, it is of concern that so many papers have been published using this methodology when some undetermined amount of the underlying theory is unpublished.

In the supporting information on the PNAS website, the authors state “Statistical Methods. The analysis uses standard binary regression models combined with singular value decompositions (SVDs), also referred to as singular factor decompositions, and with stochastic regularization using Bayesian analysis (1). It is beyond the scope here to provide full technical details, so the interested reader is referred to ref. 2, which extends ref. 3 from linear to binary regression models; these manuscripts are available at the Duke web site, www.isds.duke.edu/~mw. Some key details are elaborated here.”

It is unclear why it should be “beyond the scope” to include details of the analytical methods in the supporting information materials – typically this is precisely the place to provide such details. Fortunately the reference “ref. 2” cited above (West, M., Nevins, J. R., Marks, J. R., Spang, R. & Zuzan, H. (2000) *German Conference on Bioinformatics*, in press.) is still available as an online publication in the electronic journal *In Silico Biology* (reference [2] below).

In the online journal article, the authors provide additional details about the analytical method, including the fact that “In a first step we fitted the regression model using the entire set of expression profiles and class assignments.” (see the section titled “Probabilistic tumor classification”). This is a key point, and is precisely why the investigators’ continued publications claiming to have “validated” their analysis is false and deserves thorough statistical evaluation as part of the IOM review of these issues. When predictor variables derived from the entire set of data are used, it can not be claimed that subsequent “validation” exercises are true cross-validation or out-of-sample evaluations of the model’s predictive capabilities, as the Duke investigators repeatedly state in publications.

In the same paragraph, the authors state “Note, that if we draw a decision line at a probability of 0.5 we obtain a perfect classification of all 27 tumors. However the analysis uses the true class assignments $z_1 \dots z_{27}$ of all the tumors. Hence, although the plot demonstrates a good fit of the model to the data it does not give us reliable

indications for a good predictive performance. One might suspect that the method just “stores” the given class assignments in the parameters $\gamma_1 \dots \gamma_n$. Indeed this would be the case if one uses binary regression for n samples and n predictors without the additional restraints introduced by the priors. That this suspicion is unjustified with respect to the Bayesian method can be demonstrated by out-of-sample predictions.”

I believe this is the key flaw in the reasoning behind this statistical analytical method. The authors state without proof (via theoretical derivation or simulation study) that this Bayesian method is somehow immune to the issue of overfitting a model to a set of data. This is the aspect of this analytical paradigm that truly needs a sound statistical evaluation, so that a determination as to the true predictive capacity of this method can be scientifically demonstrated.

The authors state further in the Discussion section that “Clearly, the methodology is not limited to only this medical context nor is it specialized to diagnostic questions only. We have applied our model to the problem of predicting the nodal status of breast tumors based on expression profiles of tissue samples from the primary tumor. The results are reported in West et al., 2001. Due to the very general setting of our model, we expect it would be successful for a large class of diagnostic problems in various fields of medicine.”

Interestingly, the supporting material cited in [1] actually references [1]. This again is an issue of concern.

Also now of concern is the realization of the authors’ prediction, that they expect the method to be applicable to a large class of diagnostic problems in various fields of medicine. Indeed the authors have used this methodology in a widening scope of medical fields, as will be outlined below. That this methodology has been accepted for publication in many journals over many years, before its statistical properties have truly been investigated, is indeed an issue of concern.

I believe that part of the reason that journal editors and reviewers have not questioned the methodology is that the method uses primarily Bayesian statistical models, which are not as widely taught or understood in biological and medical higher education. It is difficult for many non-statisticians to follow the statistical logic and mathematical aspects of such complex Bayesian methods.

Thus the authors clearly describe that their paradigm is to fit a model to an entire data set, derive a set of predictors from that model, then use those predictors along with others on subsets of the entire data set. They state that the excellent performance of such models is validated, when it appears that what is actually demonstrated is that a model overfitted to an entire data set performs well on subsets of that entire data set. This issue should in my opinion be a key issue of concern in this IOM review of this omics methodology.

In early 2006, an apparently seminal paper was published by the Duke investigators in the journal *Nature* (Bild et al. [3]). This was followed by a publication in the *New*

England Journal of Medicine (Potti et al. [4]) and another in *Nature Medicine* (Potti et al. [5]). All of these papers cite West et al. [1] and use the methodology therein. References [3] and [5] discuss breast cancer, and reference [4] discusses lung cancer. All use the same analytical paradigm, fitting an initial model to all available data to develop predictors (called “metagenes” in [3] and [4], then “gene expression signatures” in [5]) based on a singular value decomposition of the entire data set; then using these predictors on various subsets of the data involved and calling some portion of this subset analysis a “validation” exercise.

At this point, researchers at the M.D. Anderson clinic explored the possibility of adapting this analytical paradigm, and asked statisticians Keith Baggerly and Kevin Coombes to review the publications. Their investigations are of course key in shedding light on this issue. In 2007 Baggerly and Coombes published a letter in the Correspondence section of *Nature Medicine* (Coombes et al. [6]). Coombes et al. state “Their software does not maintain the independence of training and test sets, and the test data alter the model. Specifically, their software uses ‘metagenes’: weighted combinations of individual genes. Weights are assigned through a singular value decomposition (SVD). Their software applies SVD to the training and test data simultaneously, yielding different weights than when SVD is applied only to the training data (Supplementary Report 9). Even using this more extensive model, however, we could not reproduce the reported results.” and further state that “When we apply the same methods but maintain the separation of training and test sets, predictions are poor (Fig. 1 and Supplementary Report 7). Simulations show that the results are no better than those obtained with randomly selected cell lines (Supplementary Report 8).”

Thus Coombes et al. have performed some initial analysis that sheds light on the true type I and type II error rates of this methodology. What is unclear from the work of Coombes et al. is the degree of departure from the null condition of no difference between groups of interest in the data sets used, so that the power of the statistical method can be properly evaluated. This is why a careful systematic study of this methodology, using known null data (data with equivalent distributional characteristics between groups of interest) and known non-null data (data with increasing levels of differential characteristics between groups of interest) is required, so that power characteristics of the methodology can be measured under null and non-null conditions. Further, such analysis needs to properly evaluate model performance on true out-of-sample data.

In 2007, the Duke group published another heavily cited paper (Hsu et al. [7], recently retracted on November 16, 2010). SVD components developed for this paper were termed “gene expression signatures”. All of these papers share the attribute that excessive claims of model accuracy are repeatedly asserted, with purported evidence from exercises termed “cross-validation”.

Recently, additional papers from the Duke investigators have been published concerning viral infection (Zaas et al. [8]) and bacterial infection (Zaas et al. [9]). Statnikov et al. [10] submitted a letter challenging this methodology once again, stating “We suggest several approaches to improve the analysis protocol that led to discovery of the acute

respiratory viral response signature. First, to obtain an unbiased estimate of predictive accuracy, genes should be selected using the training set of subjects as opposed to selecting genes from the entire data set as was done in the study of Zaas et al. (2009). The latter gene selection procedure is known to typically lead to overoptimistic predictive accuracy estimates. Second, the cross-validation procedure employed by Zaas et al. should be modified to prohibit the use of samples from the same subjects both for developing signature and estimating its predictive accuracy, as this is another potential source of over-optimism.”

The Duke investigators, as with all previous challenges, offer only verbal refutations to these points, with no formal statistical evaluation via simulation or otherwise to address the true distributional properties of this method.

More recent papers from the Duke investigators that should be reviewed include Chen et al. [11] and Chen et al. [12]. Additional complex Bayesian methods continue to be combined around the same analytical paradigm, and it is beyond the capability of many journal editors and reviewers to understand and deconstruct the arguments offered by the Duke investigators.

Additionally, with the apparent weight of so many seemingly accurate analysis results, resources such as research grants from federal agencies are being utilized without proper understanding of the value of the returns. Moreover, several studies on humans (the clinical trials currently scheduled for review, and the viral infection studies described in references [8] and [9]) have been conducted based on methodology with as yet unknown statistical properties. This is an issue of major concern, and a review of the statistical properties of the methodology used throughout these studies along with a publication of guidelines for evaluation of whether or not human trials involving this methodology should be permitted would be very valuable to the research community.

References:

1. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA Jr, Marks JR, Nevins JR. "Predicting the clinical status of human breast cancer by using gene expression profiles". *Proc Natl Acad Sci U S A*. 2001 Sep 25;98(20):11462-7. Epub 2001 Sep 18.
2. R. Spang, H. Zuzan, M. West, J.R. Nevins, C. Blanchette and J.R. Marks. "Prediction and uncertainty in the analysis of gene expression profiles." *In Silico Biology* 2 (2001). URL <http://www.bioinfo.de/isb/2002/02/0033/>
3. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, Olson JA Jr, Marks JR, Dressman HK, West M, Nevins JR. "Oncogenic pathway signatures in human cancers as a guide to targeted therapies". *Nature*. 2006 Jan 19;439(7074):353-7. Epub 2005 Nov 6.
4. Potti A, Mukherjee S, Petersen R, Dressman HK, Bild A, Koontz J, Kratzke R, Watson MA, Kelley M, Ginsburg GS, West M, Harpole DH Jr, Nevins JR. "A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer." *N Engl J Med*. 2006 Aug 10;355(6):570-80.
5. Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, Cragun J, Cottrill H, Kelley MJ, Petersen R, Harpole D, Marks J, Berchuck A, Ginsburg GS, Febbo P, Lancaster J, Nevins JR. "Genomic signatures to guide the use of chemotherapeutics." *Nat Med*. 2006 Nov;12(11):1294-300. Epub 2006 Oct 22.
6. Coombes KR, Wang J, Baggerly KA. "Microarrays: retracing steps." *Nat Med*. 2007 Nov;13(11):1276-7; author reply 1277-8.
7. Hsu DS, Balakumaran BS, Acharya CR, Vlahovic V, Walters KS, Garman K, Anders C, Riedel RF, Lancaster J, Harpole D, Dressman HK, Nevins JR, Febbo PG, Potti A. "Pharmacogenomic strategies provide a rational approach to the treatment of cisplatin-resistant patients with advanced cancer." *J Clin Oncol*. 2007 Oct 1;25(28):4350-7.
8. Zaas AK, Chen M, Varkey J, Veldman T, Hero AO 3rd, Lucas J, Huang Y, Turner R, Gilbert A, Lambkin-Williams R, Øien NC, Nicholson B, Kingsmore S, Carin L, Woods CW, Ginsburg GS. "Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans." *Cell Host Microbe*. 2009 Sep 17;6(3):207-17. Epub 2009 Aug 6.
9. Zaas AK, Aziz H, Lucas J, Perfect JR, Ginsburg GS. "Blood gene expression signatures predict invasive candidiasis." *Sci Transl Med*. 2010 Mar 3;2(21):21ra17.
10. Statnikov A, McVoy L, Lytkin N, Aliferis CF. "Improving development of the molecular signature for diagnosis of acute respiratory viral infections." *Cell Host Microbe*. 2010 Feb 18;7(2):100-1; author reply 102.
11. Chen M, Carlson D, Zaas A, Woods C, Ginsburg G, Hero Iii A, Lucas J, Carin L. "Detection of Viruses via Statistical Gene-Expression Analysis." *IEEE Trans Biomed Eng*. 2010 Jul 19. [Epub ahead of print]
12. Chen B, Chen M, Paisley J, Zaas A, Woods C, Ginsburg GS, Hero A 3rd, Lucas J, Dunson D, Carin L. "Bayesian inference of the number of factors in gene-expression analysis: application to human virus challenge studies." *BMC Bioinformatics*. 2010 Nov 9;11:552.

