Cancer Biomarkers 2 (2006) 87–88
IOS Press

Foreword

Ziding Feng
Biostatistics Program, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

Welcome to the special issue of the Journal of Disease Markers entitled "Quantitative methods for biomarker discovery and validation". In this special issue, seven leading biomarker research groups covered variety of topics from general study design principles to normalization of spectral data.

Understanding design issues in validating biomarker is crucial for study success and efficient use of resource. Richard Simon lends his many years experience in clinical trials and genomics studies at NCI. Though his article focused on validation of pharmacogenomic biomarker classifiers for treatment selection, the basic concept applies more generally. He demonstrated that with proper study design such as marker based treatment selection or targeted design (enrichment design), tremendous savings in terms of the number of events required could be achieved. He emphasized the rigorous practice in biomarker validation. For example, specifying classification rule to be validated and not just repeatedly "validate" markers by correlating markers with outcomes. Does that sound familiar in many studies?

Building a disease prediction classifier using high dimensional genomic or proteomic markers and appropriately dealing with false positive findings is a well known challenge. Don Berry and his colleagues used split data set to develop and validate a prognostic index for biomarker study. Splitting data set is not a new idea but the novelty lies in the way of screening and combining many markers and validate them using the split data and the evaluation of its efficiency and false positive rate. Readers should not apply this approach blindly for any sample size as the authors correctly pointed out that this approach is beneficial, i.e. the reduction in false positive rate is big and the loss in efficiency is small, when the sample size is reasonably large ($n = 776$ in their application) and when the number of the markers to be considered is large.

Yudong He gave a very comprehensive review of the analytical system Rosetta group has been using in their genomic studies. I call it "system" because it includes an array of methods, many of them quite innovative, for gene expression profiling and for drug discovery and development. Due to its comprehensiveness, the paper is quite long. However, it will well worth your efforts. It may help you to build your own analytical system.

Proteomic research often deals with spectral data in which normalization is absolutely important but often subjectively done. Illustrated by examples for Raman spectra, Near infrared spectra, and MALDI-TOP spectra, Tim Randolph proposed a novel scale-based normalization of spectral data. It is an extension of the wavelet-based multi-scale analysis method he developed and a generalization of the popular "standard normal variate" transformation. One advantage of this approach is its relative objectivity brought by wavelet method.

For continuous marker, what is the most appropriate statistics to report? Ian Saunders from CSIRO proposed that D, an effectiveness parameter, should always be reported. A simple conclusion is that for a useful biomarker there must be a difference between affected and unaffected individuals more than twice the between-individual variability. Similar effectiveness measure has been used in other fields but the extension here is to illustrate the relationships between D and other more familiar measures AUC, sensitivity and specificity, and risk ratio.

Longitudinal biomarkers have important place in cancer screening (e.g. PSA for prostate cancer and CA-125 for ovarian cancer) but its advantage over simple cross-sectional measures depends on marker's tempo-

ISSN 1574-0153/06/$17.00

ral stability, frequency of screening, and how fast tumor grows. Martin McIntosh and his colleagues quantified these relationships under a Parametric Empirical Bayes screening rule.

The contribution by Xingde Li and his colleagues is unique in the sense that it is not a statistical methodology paper but an interesting paper investigating the feasibility of non-invasive optical coherence tomography (OCT) for *in vivo* imaging of microanatomical changes in the epidermis and dermis during early carcinogenesis using a mouse skin model. The data is image taken longitudinally over five time points. The platform and its application is novel and the finding is interesting that early structure changes during carcinogenesis were clearly delineated *in vivo* using OCT. The relevance for this special issue is to illustrate that new technology is rapidly adopted by investigators with unprecedented data challenging quantitative scientists to help interpreting data. For example, this data has four dimensions (x-, y- axis plus intensity and time). How to quantify information from such data and find the features that distinguish disease from non-diseased? We will not be bored for many years.

I hope you enjoy your reading and get something out of it. I thank the editorial office of Disease Markers for their help and all contributors to make this special issue a reality.