The Curse of Dimensionality: A Blessing to Personalized Medicine

TO THE EDITOR: We wish to provide an alternative consideration to the recommendations raised in the editorial by Sikic et al1 in Journal of Clinical Oncology relating to the study by Bhojwani et al.2 The study reports gene expression profiles that predict early response and long-term outcome in children with acute lymphoblastic leukemia. Bhojwani et al identified 41 genes out of a 38,500-gene expression data set that were predictive of long-term patient outcome treated on a single protocol and a 24-gene set predictive of early therapeutic response. Of most significance to the field however, was their conclusion that gene expression signatures provide no greater prognostic value compared with current factors such as patient age, WBC count, or karyotype.

Most clinical stratification protocols stream patients into a limited number of treatment strategies with individuals being managed within risk groups. However, to realize the potential for personalized medicine, patients in need of specific clinical intervention should be distinctively identified out of the crowd. So, by personalized medicine are we aiming to improve risk stratification with patient cohorts being divided into ever smaller risk groups, or do we consider that medicine can be truly tailored for individual patients? The former is based on the assumption that we can classify complex disease into specific subgroups where patients within each group have similar genetic activity and are homogenous with respect to treatment response. This fails to recognize that individual patients within subgroups will have unique clinical responses to treatment strategies. If the latter, we then need to be able to realistically compare individual patients to each other. This will require identifying individual genetic differences, as much as similarities.

Sikic et al1 identified the fundamental problem is the “traditional paradigm of sifting data to identify one or a few markers to use prospectively for prognosis of outcomes or prediction of therapies,” that is, reductionism. Traditionally, reductionism guides observational deductive science, data derived from complex biologic samples being broken down to single points of study. Many interesting disease-related genes or biomarkers have been discovered as a result. The premise of investigations such as Bhojwani et al2 is that gene expression microarray datasets can be reduced using statistical or informatics approaches to identify a smaller set of genes which define patients at risk of poor clinical response. We have reported that while such gene expression signatures can be identified which segregate different acute lymphoblastic leukemia patient subgroups in one study, such signatures could not be validated in independent cohorts.3,4 To date, reductionist approaches have yet to realize signatures that have universal diagnostic, prognostic, or therapeutic applications which would be the basis for personalized medicine.

The nature of microarray data and its acquisition means that it is subject to the curse of dimensionality, the situation where there are vastly more measurable features (genes) than there are samples (Fig 1A). Reductionist analysis (Fig 1B) of highly dimensional microarray data, such as used by Bhojwani et al2 lead to feature selection approaches that are liable to extreme type I error, and will not identify enough features to provide for individual differences within the patient cohort. For leukemia, this has been address by the Microarray Innovations in Leukemia (MILE) international multicenter study,5 which aims to expand and standardize the number of samples analyzed (Fig 1C) and to assess the clinical accuracy of gene expression profiling across leukemia subtypes. It is anticipated that a more robust predictor will be derived from reductionist analyses seeking gene expression similarities within disease subgroups. However, despite the large number of genes which we can derive expression data from in a single experiment, it only informs us about one cellular process, transcription. It is unclear exactly why particular genes are good indicators of patient outcome as it is not known whether these genes directly cause the metabolic effects or whether they are in the coregulating

![Fig 1](https://example.com/fig1.png)

Fig 1. The expression of m genes (features) determined in n samples (patients). (d) indicates the number of dimensions in the feature space the data is visualized. (A) raw data; (B) feature selection; (C) increasing samples number; and (D) expanding dimensions of feature space are shown.
ensemble. For personalization of treatment, the large number of subtle differences in gene expression, interacting biologic pathways and genetic variations between individual patients which influence clinical response need to be determined. The principal level of complexity in current high-throughput genetic and genomic data is driven by its sheer volume and not by the complex biology of the cancer cells from which the data is derived. Hence, removing the curse of dimensionality by increasing the number of sample analyzed does not remove the biologic complexity of the disease nor make it more straightforward to interrogate.

The complexity of genomic data derived from equally complex diseases like cancer necessitates a paradigm shift in our approach to translational research utilizing data derived from these high-throughput next generation technologies. Such a paradigm shift will impact researchers, clinicians, trials, and pathologists alike. It is our proposal that this paradigm shift needs to be inductive and constructionist in nature, building models of integrated genetic, genomic, epigenetic, and clinical datasets which represent individual patients allowing us to predict their clinical course. West et al. call researchers to "embrace the complexity of genomic data for personalized medicine." So how do we embrace the complexity?

Understanding the nature of complex diseases and interacting networks of genetic features is found through data integration and systems biology. The predictive potential for systems biology, especially applied to medicine, is its ability to capture not just the similarities between individuals in a cohort, but the differences between them, enabling individualization of patients. There is a need for tools to make sense of the huge number of genes involved and allow researchers to phrase meaningful hypotheses for testing. Consequently, to embrace the complexity of microarray-derived data we first need to make sense of the data. The computational discipline of sense making provides a theoretical framework to support analytic reasoning, the iterative and collaborative process of applying human judgments to reach conclusions from a combination of questions, judgments, and evidence. Technologies for sense making make possible the holistic interrogation of complex and voluminous data, often using visual analytic tools that display the data in feature space and producing dynamic graphical representations which capture all the biologic dimensions for an individual patient as a single feature (Fig 1D). As we look to the human genome to provide information for patients, their disease and their ability to respond to treatments at our disposal, an increasingly complex picture emerges. Yet, each patient’s picture is different from the next. Indeed the high dimensionality of this information is a reflection of how a patient is unique within the cohort and hence, needs to be exploited to gain maximum insight into the complex data available. Subsequently building into this model an understanding of the specific domain (e.g., pediatric cancer, clinical patient management) and the underlying biologic principles (e.g., regulatory pathways, gene function, treatment modalities) will allow the clinician to determine the true biologic basis for how a patients responds in the clinic.

This approach presents the opportunity to allow analysts, in this case, clinicians, to have a true discourse with the patient-derived integrated omic data that will glean insight into how to treat individual patients, creating confident human judgments about specific cases in the clinical setting leading to personalized medicine. Achieving this at the point of care requires us to take on the challenge to turn the curse of dimensionality into a blessing.

Daniel R. Catchpoole
The Tumour Bank, The Children’s Hospital at Westmead, Westmead, New South Wales, Australia

Paul Kennedy
The University of Technology Sydney, Sydney, New South Wales, Australia

David B. Skillicorn
The School of Computing, Queen’s University, Kingston, Ontario, Canada

Simeon Simoff
The School of Computing and Mathematics, The University of Western Sydney, Parramatta, New South Wales, Australia

Authors’ Disclosures of Potential Conflicts of Interest
The author(s) indicated no potential conflicts of interest.

References

DOI: 10.1200/JCO.2010.30.1986; published online ahead of print at www.jco.org on November 1, 2010