CANCER GENOMICS & PROTEOMICS *4*: 111-120 (2007)

Review

An Exploration into Study Design for Biomarker Identification: Issues and Recommendations

JACQUELINE A. HALL, ROBERT BROWN and JIM PAUL

Centre for Oncology and Applied Pharmacology, Cancer Research UK Beatson laboratories, University of Glasgow, Garscube estate, Glasgow, G61 1BD, U.K.

Abstract. *Genomic profiling produces large amounts of data and a challenge remains in identifying relevant biological processes associated with clinical outcome. Many candidate biomarkers have been identified but few have been successfully validated and make an impact clinically. This review focuses on some of the study design issues encountered in data mining for biomarker identification with illustrations of how study design may influence the final results. This includes issues of clinical endpoint use and selection, power, statistical, biological and clinical significance. We give particular attention to study design for the application of supervised clustering methods for identification of gene networks associated with clinical outcome and provide recommendations for future work to increase the success of identification of clinically relevant biomarkers.*

With the advent of DNA microarrays large volumes of data are being generated from genomic profiling of tumour samples. Currently a major aim in cancer research is to reliably identify tumour markers or marker combinations that are strongly prognostic of clinical outcome of the patient, a problem known as "Class Prediction" (1). Armed with these novel candidate prognostic markers it is hoped that patient management will be improved leading to better survival rates and less morbidity. Additionally, these techniques offer the opportunity for greater understanding of the complex disease process which could open doors to developing novel therapies.

In the literature, many prognostic factors have been proposed but most studies have been hypothesis generating and await independent confirmation and assessment of clinical relevance (2-5). Many disparate reports have been published with limited power and comparisons between studies are difficult due to differences in methodology. These problems were recently illustrated in a review of publications of TP53 as a prognostic marker in ovarian cancer suggesting that we need to re-evaluate our methodology and study design for biomarker discovery (6). Study design is becoming an increasingly important issue for maximising returns from large prospective profiling initiatives.

With high throughput study of the human genome the simple, "one-gene-one outcome" hypothesis in complex diseases such as cancer is changing (7). Much attention has focussed on the challenge of identifying multi-gene signatures; with each gene contributing a small but significant effect but when combined have prognostic value (8). In trying to answer these more complex questions, study design issues become especially important when applying data mining methods to a relatively small set of samples with very large numbers of candidate genes (1).

The design of a study may differ for the particular questions posed, therefore clear hypotheses need to be defined up front so that the study may be appropriately designed. There are many considerations and decisions to make as part of this process. Some of these questions are summarized in Table I.

For many of these design issues guidelines or recommendations do not exist. In this article we highlight some of these key design issues and we illustrate with examples how they can affect ultimate success of the study.

Choice of Study Endpoint

One of the principal design issues for prognostic factor studies is selection of the most appropriate clinical endpoint for biomarker identification. Although this may sound simple, different endpoints of clinical outcome confer

Correspondence to: Jacqueline A. Hall, Centre for Oncology and Applied Pharmacology, Cancer Research UK Beatson laboratories, University of Glasgow, Garscube estate, Glasgow, G61 1BD, U.K. e-mail: j.hall@beatson.gla.ac.uk

Key Words: Cancer biomarkers, tumour profiling, study design, supervised clustering, review.

Table I. *Decision points in biomarker study design. A summary of some important points for consideration in biomarker study design, the chevron indicates points that will be discussed in detail.*

Decision points in biomarker study design
• Is it a hypothesis generating or hypothesis testing study?
➢ What is considered as a "significant" association?
➢ Is the aim to identify biological, prognostic or predictive biomarkers?
➢ What endpoint is most suitable for discovery of this marker type? What biases does use of the endpoint incur and how should we use the endpoint information?
➢ What is the patient population that would benefit from the marker and therefore which patients should be included/excluded from the biomarker study?
➢ Do we wish to identify an individual marker or a set of markers and what difference does this make to our study design?
➢ Do we want to uncover information on description of the biological system *e.g.* gene networks or are we aiming for the best predictor of clinical outcome?
➢ Given the method of analysis what sample size do we require to detect associations?
➢ How can we increase the power of our study or use fewer patient samples to detect an effect? And what impact does this incur on the results of the study *e.g.* pre-selection of patient population and pre-filtering?
• What is the validation strategy?
• What is the gain in predictive value of the marker in comparison to only using existing clinical criteria?
• Does identification of the marker lead to further hypotheses? If so do we need to design a new study to test this?

different information and are often loosely defined and assumed to be interchangeable when they may in fact reflect a slightly different patient population or be affected by other confounding factors. Different endpoint definitions, such as what constitutes "progression" (clinical, radiological, CA125) makes interpretation and cross comparison of studies difficult. Even using robust endpoints such as overall survival does not avoid issues of other confounding factors, such as salvage therapy, that may influence survival time (Figure 1).

The choice of endpoint may be influenced by the type of biomarker sought. Categorical endpoints may be used for markers discriminating between clinico-pathological groups like grade. If we were interested in identifying a marker of response to chemotherapy we have a choice of a categorical endpoint (response to chemotherapy) or a continuous one of progression-free survival (PFS). We can appreciate that prognosis endpoints like PFS and overall survival (OS) may be further removed from the underlying biology of tumour cell death in response to chemotherapy, however response endpoints remain subjective, non-quantitative and may not be a robust surrogate for survival endpoints (9).

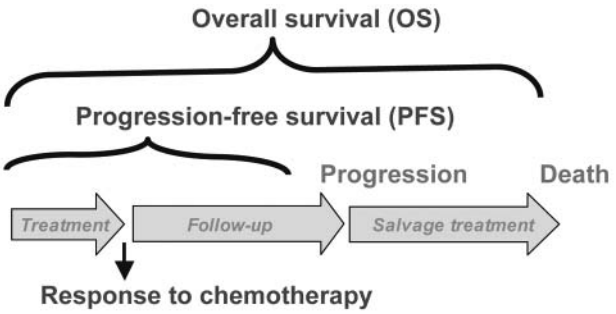

Figure 1. *An illustration depicting how different clinical endpoints like progression free survival (PFS), overall survival (OS) and response to chemotherapy, reflect different stages of a cancer patients' disease management and therefore confer different information on patient outcome.*

A second issue with the choice of clinical endpoints is that of patient exclusion that may lead to bias in the clinical population under study. This may impact on the "generalizability" of any marker subsequently found, an example being response to chemotherapy in ovarian cancer patients. Solid tumour response criteria such as RECIST (response evaluation criteria in solid tumours) or SWOG (south west oncology group) are based on measurement of the tumour from pre-treatment debulking surgery. Patients who have had optimal debulking may well have no residual baseline disease to measure making them unevaluable for this endpoint. This can lead the study population to be biased towards larger tumours or those that are more difficult to surgically remove, which may in turn correlate with other biological properties of the tumour. In a large study of 1077 ovarian cancer patients, radiological response was not observed to correlate with stage of disease ($p=0.089$) although stage is clearly associated with PFS ($p<0.001$) and OS ($p<0.001$) (10). This may be because only larger tumours were able to be assessed for response. If known prognostic factors are not associated with this endpoint it is highly probable that important prognostic genes will also be missed. Radiological response was only available for 55% of patients in the study and begs the question of its relevance to the general population of ovarian cancer patients?

How Do we Use the Clinical Information?

More recently a second issue in regards to clinical endpoints has arisen, this being what is the best way to use this clinical information? Many sophisticated data mining methods have been developed for identifying multiple features for discrimination between distinct groups. With survival data we have the specific problem of a continuous survival time which may be censored, *i.e.* the date of progression or death

for the patient may be unknown, but we do know they were alive at a certain time point. Several studies to date have classified survival data into "good" and "bad" prognosis groups so that data mining methods may be applied for biomarker identification (11, 12). This dichotomisation is likely inefficient leading to loss of information in the continuous variable. It also leads to the problem of how to select a cut point for dichotomisation, it could be driven by existing clinical knowledge or by the data itself (13). However, as shown for categorisation of covariates, the selection of a cut point for categorising outcome could affect the degree of association observed and may therefore influence our interpretation of the study (14). Using simulated survival data for a continuous covariate such that the variable had a strong association with the continuous survival data ($p=1.64x10^{-4}$), categorisation of this survival data at different percentiles we see the association of the continuous variable with categorised outcome changes depending on the cut point used: 40th percentile $p=6.76x10^{-3}$, 50th percentile (median) $p=1.75x10^{-4}$, 60th percentile $p=7.17x10^{-3}$ (all these associations were obtained from fitting a logistic regression model of the continuous covariate on the categorised survival data). Furthermore, censoring complicates categorisation since the time the event occurred is unknown, simply categorising the time the patient was last seen could lead mis-classification of cases into the "poor" prognosis group if they were lost to follow up at an early stage. This could subsequently affect model estimation and with algorithms searching for gene-combinations the combined effect of multiple model parameters being mis-estimated could be very large. To avoid this, we suggest the use of data mining methods with appropriate statistical methodologies specifically developed for censored survival data, namely Cox regression. Recently there has been much activity in this area of algorithm development and methods such as supervised clustering and partial least squares Cox regression have been described (15). The advantage of former methods is that they give insight into functional gene groups that relate to clinical outcome aiming at a more transparent and interpretable model as opposed to a "black-box" system where little information on the underlying relationship is obtained. This class of methods, called semi-supervised clustering, aims to simultaneously perform clustering and identify genes that associate with clinical outcome. The most successful of these methods being supervised principal components, an algorithm based on singular value decomposition which creates a multigene model that comprises of a linear combination of selected genes (15). Data mining methods based on Cox regression technique also have the flexibility to allow for incorporation of existing clinical or pathological criteria that are known to have prognostic value into the model building process where incorporation of clinical covariates may be more challenging for other data mining techniques.

What is a Significant Association?

In the application of complex data mining methods to high dimensional data we must take precautions again making erroneous conclusions. Care must be taken in the design of the analysis to prevent false positive discovery, a situation that is exacerbated by a large number of potential candidates in comparison to the number of samples. Methods such as permutation tests, controlling the false discovery rate and Bonferroni adjustments have all been described for this purpose and make stringent correction of the p-values to adjust for multiple testing (1).

Of late, much attention has focussed on the issue of stability of gene lists selected by multi-gene classifiers built from different sub-samples of the same patient cohort (16). The suggestion being that if the same genes are identified by the model over multiple training/ test set cohorts these candidates represent "significant" effects and associations. A recent study of node negative breast cancer by gene expression profiling identified a 70 gene signature associated with the presence of metastasis at five years (9). On re-analysis of this data Ein-dor *et al.* demonstrated that there were multiple alternative 70 gene prognostic profiles that could have been extracted from the original data collected by Van't Veer and colleagues and that the data structure, *i.e.* correlations between the genes, can affect the particular gene selection of the algorithm (16). We should take note that the set of genes selected by an algorithm applied to a particular set of patients do not represent the only model that could be built from this data. In other words, this is not to say that since many 70 gene prognostic profiles could have been selected that the one that was selected will not be prognostic. Although the Van't Veer 70 gene signature was not the most stable 70 gene list it still has confirmed prognostic capability in an independent study (17).

Ein-dor *et al.*'s estimation that thousands of samples may be required to consistently select the same genes may very well be correct, however attainment of a "stable" gene list is not equivalent to identifying significant effects and therefore it may not be necessary for identifying prognostic candidates (18). The presence of a gene in one gene list and its absence in another does not necessarily mean it is a false positive result but in all likelihood reflects the complexity of data structure in terms of correlations between genes and the methods used to select out candidates. We suggest the milestone for identification of truly significant prognostic markers should be their prognostic ability in a well defined validation study in independent data.

What Sample Size is Required to Achieve Adequate Power in this Study?

A second situation of inaccuracy is that a marker is not identified as prognostic when in fact it is. This lack of identification of a candidate is related to the sample size *i.e.* the power of the study to detect the effect of the marker. Several factors affect the power of a study and these have been quite well characterised for single markers of prognosis. For example for a given marker, the frequency of the marker in the population affects the power to detect it, the most favourable situation being if the factor is found to be positive in half of the population (a 50:50 ± split) (19). If this ratio decreases to 35% of the population testing positive (35:65) 10% more samples are required to reach the same power as for the 50:50 split and for a marker seen in only 20% of the population (20:80) a 56% increase in the sample size is needed to still attain this same power.

It is understood that correlation between covariates affects the power to recover associations and standard power calculations have been extended to include corrections for this (19). For example if we wish to demonstrate that a new individual marker carries prognostic information over and above existing prognostic factors both the existing and the novel factor must be included in the model. If there is a degree of correlation between the two variables in the model this will affect the power of the test to identify the novel markers prognostic capability. A correlation of 0.2 between factors will require an increase in sample size of 4% to reach the same power whereas a correlation of 0.4 requires an extra 19% of samples.

Power considerations will change with respect to the type of marker under study. In the case of identifying predictive markers, since we are effectively looking for differential prognostic capability within two treatment groups- those receiving treatment of interest and those not, we will effectively need to increase the sample size to account for this. Here we exemplify with a strong potential biomarker of response to DNA damaging chemotherapy, the TP53 tumour suppressor gene (20, 21). Given that this gene is involved in DNA damage response, patients with wild type TP53 may respond well to DNA damaging therapies like cisplatin, whereas patients with mutated TP53 may respond better to treatments that involve other mechanisms of cell death independent of TP53, like the taxanes. Taking an example of a large study investigating the prognostic and predictive capability of TP53 status with respect to the use of cisplatin and taxane agents in ovarian cancer (22), performing some basic power calculations using the observed frequency of mutated p53 (48%), proportion of censored cases and the sample size used we can see that comparing the power of the study to investigate the prognostic potential and the predictive potential differs markedly. For a hazard ratio (HR) of 2, where patients with abnormal TP53 have twice the risk of progression compared to TP53 wild type, the power to detect prognostic capability is 99.9%. However for detecting a two-fold difference in the relative risk of the two patient populations (*e.g.* treatment A marker HR=2 *vs.* treatment B HR=4) the power is 90%; meaning if the association was true under these conditions it would be detected in 99.9 or 90 times out of 100. The difference in power between the two types of question becomes more acute if the difference in risk is less. For detecting prognostic capability of a HR of 1.5 the power in this study is still 95.9% but to detect 1.5-fold difference in HRs between patient groups (*e.g.* HR 2 *vs.* HR 3) the power to find such a predictive marker is only 47.5%.

A Single Marker or Multigene Classifier? What Extra Requirements are Needed in the Application of More Complex Algorithms? Functional Significance of Genes *versus* the Best Prognostic Model of Clinical Outcome

In situations where we are searching for a gene profile that is not only prognostic, but also contains coherent sets of genes (in the sense that they are clustered in some way) this leads to restriction on the gene selection process that can compromise study power. For example, application of supervised clustering methods (supervised principal components and partially supervised gene shaving) to the original Van't Veer data produced models that were less prognostic than the original Van't Veer 70 gene classifier. This is illustrated by the respective odds ratios (OR) of the classifiers produced from supervised clustering methods compared to the Van't Veer 70 gene classifier for the prediction of poor prognosis patients in independent data (the Van't Veer 70 gene classifier had an OR of 66 and associated $p=1x10^{-3}$, compared to the classifier selected by supervised principal components that selected a model with an OR of 18 and $p=0.02$ and also the classifier of partially supervised gene shaving that had an OR of 8.4 and $p=0.147$, manuscript in preparation). It is likely that by constraining the genes selected to clusters of correlated genes this affects the power of the algorithm to detect associations with survival in comparison to unconstrained approaches which simply find the optimal combination of genes for outcome classification.

How Can we Increase the Power of the Study or Use Fewer Patient Samples to Detect an Effect? How Does this Affect the Final Results of the Study?

Several methods could be used for reducing multiplicity and increasing power when investigating many genes simultaneously but these methods are not without drawbacks and this requires consideration. One way to reduce the issue

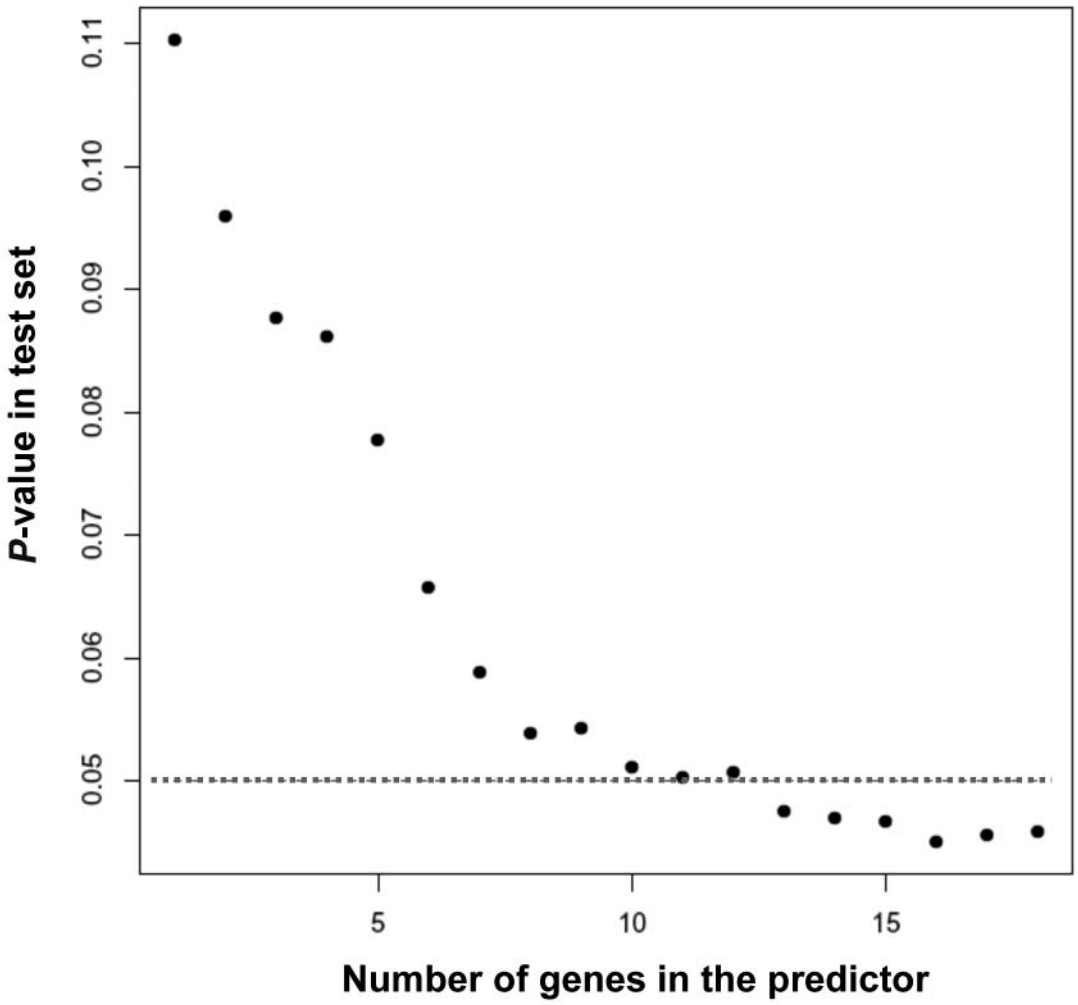

Figure 2. *A graph depicting the decreasing prognostic capability of a multi-gene classifier as individual genes are removed from the model. 100 simulations of data consisting of a training set of 900 cases and a test set 500 cases. The p-values reported are from the likelihood ratio test. Genes were removed sequentially by setting their weights (coefficients) in the singular value decomposition to 0 in order of their absolute weight, smallest first (i.e. in order of their contribution to the model, the least important removed first). The average association between the 20 genes was a Pearson correlation coefficient of 0.4.*

of multiplicity is to reduce the number of genes available for selection during model building (23). This could be approached by clustering similar genes into a smaller number of gene groups and perform the model building on gene groupings. Another method could be pre-filtering the data with respect to an association of each feature with clinical outcome. However, by pre-filtering data we could introduce a bias into the analysis and potentially lose interesting and important genes. Simulations demonstrate that exclusion of a gene in a multigene classifier can subsequently affect the performance of the algorithm (Figure 2). Here we simulated a dataset that was split into a training and test set. This simulated data contained a gene cluster that constituted of a linear combination of 20 genes that was strongly associated with survival in the training data ($p=2.6\text{x}10^{-6}$). This was generated by singular value decomposition of 20 selected correlated genes and the gene combination variable was used to simulate survival data. Using this gene combination in a Cox regression we measured its association with the simulated survival data. The effect of incomplete identification of all 20 genes on the model is seen in the *p*-value in the test set. Clearly the full 20 gene model is strongly associated with clinical outcome in the test data, however once the method is only recovering half of the genes in the cluster (even with no false positive identification) significance at the 5% level is not reached in the test set.

Even though the main reason for application of supervised clustering methods may be to uncover interesting biological mechanisms for use in hypothesis generation, it is still

important that we assess the statistical significance of the association with clinical outcome for selecting interesting models. If we do not do this, we risk investigating a set of noise genes that are not related to survival.

If feature selection is performed on the basis of a univariate association of each feature with clinical outcome that could lead to exclusion of subtle gene effects that could be contributing in part to a "gene combination" signature. This inappropriate filtering could easily occur if the study is underpowered. In our above simulation, given it is the multi gene model of twenty genes that associates with outcome we may not expect each individual gene in the model to be significant in a univariate analysis. In fact only half of the 20 features in the above simulation would be considered as having an association with clinical outcome at the 5% level of statistical significance and this would be even less with corrections for multiple testing.

Can we Use Fewer Patients in our Study to Detect the Same Effect?

Recently there has been some suggestion that pre-selection of the patient population on the basis of their clinical survival can aid the identification of prognostic factors through reducing "noise" in the data in the form of "intermediate survival" cases thereby increasing the power of the study (less samples are used to identify an effect) (24). For example, instead of selecting a random cohort of patients from the patient population, we could select a group of very good prognosis patients and a set of very bad prognosis patients (herein we refer to this as extreme sample selection). We believe that this selection process forms part of a variance-bias trade off. With extreme sample selection it may be easier to identify models and the resulting models may appear to be less variable but in fact they are consistently mis-estimating the effect, or in other words, are biased. We illustrate this by simulations of a single variable strongly associated with simulated survival data in a large dataset (N=4000 samples). For each simulation we estimated the "true" regression coefficient by fitting a Cox regression model for the variable in all 4000 samples. We then took a random sub-sample of 200 cases for model building and also 200 extreme cases (100 good prognosis cases selected from the top 1/3 surviving patients and 100 bad prognosis from the bottom third). We proceeded to build models and obtain estimates of the HR for each sample selection method. We then tested both models in the same test set of 500 randomly sampled (independent) cases. We observed that over 400 such simulations that the model built with extreme sample selection was over estimating the HR compared to the random sample selection and this was a highly significant effect ($p<2e^{-16}$, extreme patients model: mean HR=1.61, random selection model: mean HR=1.53, "true" HR=1.52). The variability of the HR for the extreme patient model was lower than for the random selection model suggesting consistent over-estimation of the effect (HR) and this was reflected in the prognostic ability of the model in independent data. The extreme patient model had a worse accuracy in terms of pseudo R^2 value (0=bad, 1=good). The R^2 measure was significantly higher for the random selection model ($p<2e^{-16}$). Given that this is a simple scenario of a single variable, it is likely that mis-estimation would have an even larger effect for a multi-gene classifier where multiple coefficients must be estimated. This leads us to the conclusion that such sample selection leads to a "quick and dirty" model that may be useful for logical checking of hypotheses (*e.g.* gene A is up in "good" prognosis and down in "poor" prognosis patients) but is not appropriate for prognostic modelling.

What Patients Should be Included in the Study?

Evidence is building that suggests cancer is a heterogeneous disease and any given cancer may be comprised of molecularly distinct tumour subtypes. An example of this being breast or ovarian cancer where multiple subtypes and histologies of disease are present which are associated with different clinical characteristics (25, 26). Further advances in molecular markers for accurately describing disease subtypes, such as markers for grade or histology (27), would allow accurate and consistent sub-population identification that would be highly beneficial for biomarker study design. Given these issues we must take into consideration the patient population we intend to study and ensure that it is well defined and homogenous with respect to the criteria. The issue of patient inclusion/exclusion criteria is a well accepted part of clinical trial design and this philosophy should be extended to biomarker studies. Tighter inclusion criteria in future studies could reduce inter-patient heterogeneity sufficiently to identify factors of clinical importance for that subgroup. It is still important however, that the study population be representative of the target clinical population, this will allow extrapolation of findings directly with a higher success rate.

Is the Aim of the Study to Identify Markers of Biological or Clinical Significance, and How Do these Goals Relate?

Striving for statistical significance should not be the sole aim for prognostic factor identification. We also need to address the biological and clinical significance of identified biomarkers. To test the clinical significance of a marker it is important to test for independent prognostic capability over and above other known clinico-pathological criteria. For

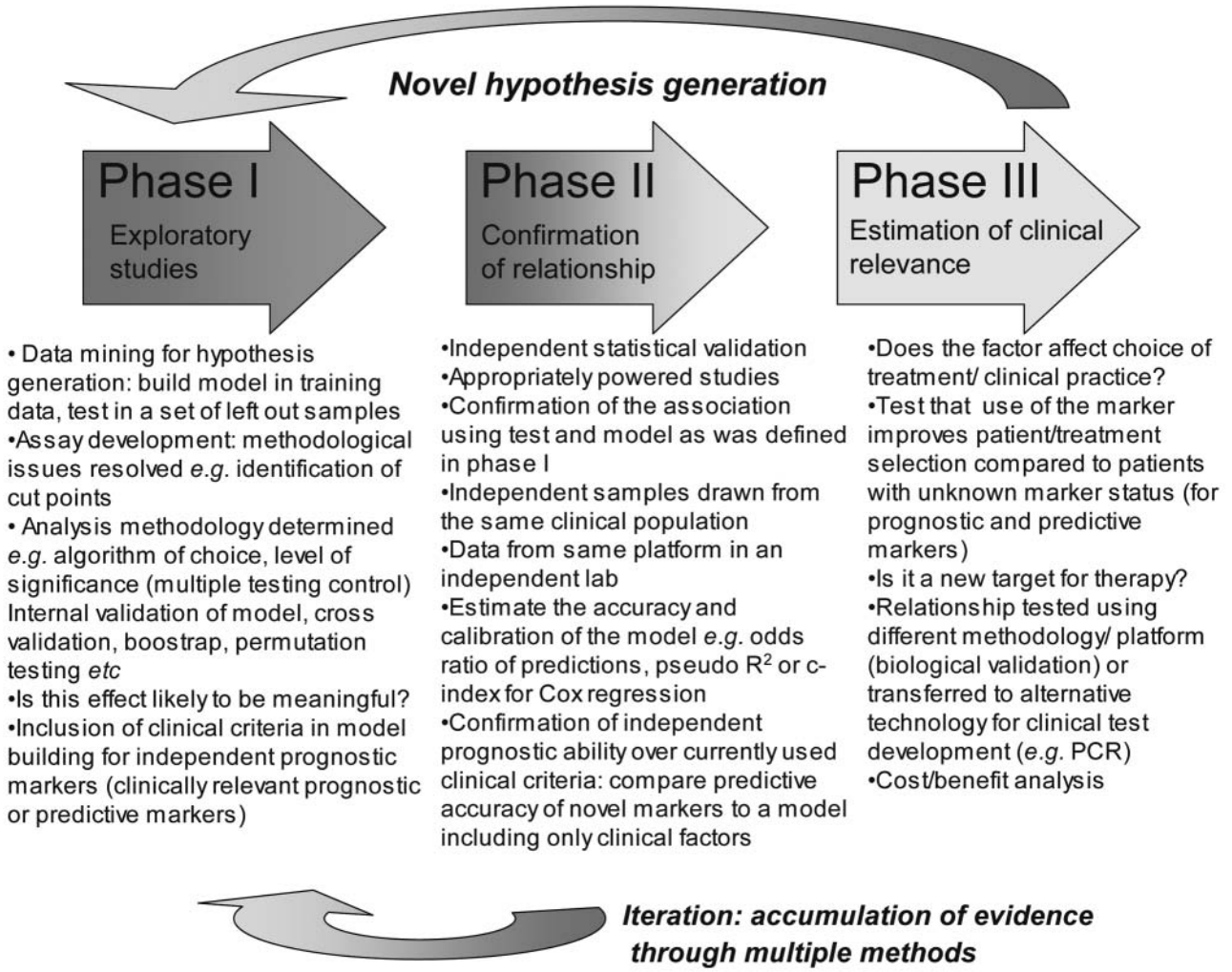

Figure 3. *A flow diagram representing how a staged approach to biomarker studies could be used to robustly identify and test new clinically relevant markers. The three main types of biomarker studies, biological, prognostic and predictive of response to therapy, largely pass through similar stages, although details in study design at each stage may differ.*

example, recently Sotiriou *et al.* identified a prognostic signature that was related to grade in breast cancer and given that grade associates with clinical outcome, it is not surprising that this grade profile is also prognostic (28). This marker profile provides interesting biological insights into the biological mechanisms behind the pathology and it is clear that a study aiming at identifying biologically significant markers is a very valuable approach. However for a marker to add novel, clinically relevant, prognostic information it needs to confer information over and above, or be more accurate than, known or used clinical criteria. It may also be of benefit to address the addition in prognostic capability that the biomarker adds to currently used clinical criteria by comparison of the prognostic capability of a multivariate model including the marker relative to a model including solely clinical criteria (29). Biological importance and clinical relevance are related but not equivalent, however iteration between the two goals will likely lead to successful prognostic candidate identification. We could view the whole process of prognostic factor identification as an interactive one, from data generation in the laboratory through data mining to generate hypotheses and back to the laboratory for further testing. In this process novel candidates, interactions or pathways could be identified and form the focus for more laboratory based research. Markers identified though this kind of iterative process are likely to be stronger candidates for successful clinical application as we will have a greater understanding of the functional biological mechanism and if we have converged upon these candidates through multiple approaches we gain confidence through an accumulation of

evidence that the marker has an effect that is clinically relevant (29). In such an iterative process, it is important, that clear aims and hypotheses are set in advance at each stage so that any study and analysis may be clearly defined to answer the specific question. It should be noted if the aim of the study changes then a different approach to the design and analysis may be justified.

After hypothesis generation and initial estimation of the effect of the biomarker when a clear hypothesis has been established, we can use this information for the design of the validation study. The estimated hazard ratio from the initial analysis can be used for determining sample size required for validation and care must be taken that the cases used in validation are drawn from the same clinical population. Internal validation and error rate estimation may have been performed using the original training data alone for example bootstrapping and cross validation, the benchmark for estimation of the capacity of the biomarker is through applying it to independent data. For clinical utility is may also be necessary for specimens to be processed and assayed at different times, in different labs and by different operators to determine the reproducibility of the assay as a whole. Once evidence has accrued of the utility of the marker, confirmation on a different assay platform and further development into clinical tests for example blood based assays can be investigated.

Conclusion

For taking biomarker studies forward we would do well to adopt a staged approach to biomarker studies, akin to clinical trials (30). Figure 3 illustrates the issues that are addressed at the different stages of discovery and validation.

Is it clear that robust and complete data collection is necessary for reliable and consistent determination and validation of both potential prognostic factors and also the study end-point. High quality information should also be collected on existing prognostic factors in the cohort. These recommendations would be best achieved within the context of a prospective clinical trial. In accordance with well-designed and -executed studies, we must ensure high quality dissemination of the results. It is hoped that the community will be quick to adopt the newly released REMARK criteria guidelines for the conduct and reporting of prognostic factors studies (3).

References

1 Simon R *et al*: Pitfalls in the Use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst *95(1)*: 14-18, 2003.

2 Hilsenbeck S, Clark G and McGuire W: Why do so many prognostic factors fail to pan out? Breast Cancer Res Treat *22(3)*: 197-206, 1992.

3 McShane LM *et al*: REporting recommendations for tumour MARKer prognostic studies (REMARK). Br J Cancer *93(4)*: 387-391, 2005.

4 Altman D and Lyman G: Methodological challenges in the evaluation of prognostic factors in breast cancer. Breast Cancer Res Treat *52*: 289-303, 1998.

5 Paik S *et al*: Real-world performance of HER2 testing – national surgical adjuvant breast and bowel project experience. J Natl Cancer Inst *94(11)*: 852-854, 2002.

6 Hall J, Paul J and Brown R: Critical evaluation of p53 as a prognostic marker in ovarian cancer. Expert Rev Mol Med *2004*: 1-20, 2004.

7 Chanock S and Wacholder S: One gene and one outcome? No way. Trends Mol Med *8(6)*: 266-269, 2002.

8 Paik S: Incorporating genomics into the cancer clinical trial process. Seminars in oncology *28(3)*: 305-309, 2001.

9 Oye RK and Shapiro MF: Reporting results from chemotherapy trials. Does response make a difference in patient survival? JAMA *252(19)*: 2722-2725, 1984.

10 Vasey PA *et al*: Phase III randomized trial of docetaxel-carboplatin versus paclitaxel-carboplatin as first-line chemotherapy for ovarian carcinoma. J Natl Cancer Inst *96(22)*: 1682-1691, 2004.

11 Shipp MA *et al*: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med *8(1)*: 68-74, 2002.

12 van 't Veer LJ *et al*: Gene expression profiling predicts clinical outcome of breast cancer. Nature *415(6871)*: 530-536, 2002.

13 Verduijn M *et al*: Dichotomization of ICU length of stay based on model calibration, in artificial intelligence in medicine. Springer: Berlin / Heidelberg, pp. 67-76, 2005.

14 Farewell V, Tom BDM, and Royston P: The impact of dichotomization on the efficiency of testing for an interaction effect in exponential family models. J Am Stat Assoc *99(467)*: 822, 2004.

15 Bair E and Tibshirani R: Semi-supervised methods to predict patient survival from gene expression data. PLoS Biology *2(4)*: e108, 2004.

16 Ein-Dor L *et al*: Outcome signature genes in breast cancer: is there a unique set? Bioinformatics *21(2)*: 171-178, 2005.

17 Buyse M *et al*: Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst *98(17)*: 1183-1192, 2006.

18 Ein-Dor L, Zuk O and Domany E: Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. PNAS *103(15)*: 5923-5928, 2006.

19 Schmoor C, Sauerbrei W and Schumacher M: Sample size considerations for the evaluation of prognostic factors in survival analysis. Stat Med *19*: 441-452, 2000.

20 Vogelstein B, Lane D and Levine A: Surfing the p53 network. Nature *408*: 307-310, 2000.

21 Lowe S *et al*: p53 status and the efficacy of cancer therapy *in vivo*. Science *266(5186)*: 807-810, 1994.

22 de Graeff P *et al*: Factors influencing p53 expression in ovarian cancer as a biomarker of clinical outcome in multicentre studies. Br J Cancer *95*: 627-633, 2006.

23 Simon R: Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n). Association of Computing Machinery, SIGKDD Explorations *5(2)*: 31-36, 2003.

24 Liu H, Li J and Wong L: Use of extreme patient samples for outcome prediction from gene expression data. Bioinformatics *21(16)*: 3377-3384, 2005.
25 Sorlie T *et al*: Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. BMC Genomics *7(1)*: 127, 2006.
26 Schwartz DR *et al*: Gene expression in ovarian cancer reflects both morphology and biological behavior, distinguishing clear cell from other poor-prognosis ovarian carcinomas. Cancer Res *62(16)*: 4722-4729, 2002.
27 Ouellet VM *et al*: Tissue array analysis of expression microarray candidates identifies markers associated with tumor grade and outcome in serous epithelial ovarian cancer. Int J Cancer *119(3)*: 599-607, 2006.
28 Sotiriou C *et al*: Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst *98(4)*: 262-272, 2006.
29 Katz E and Kattan M: How to judge a tumor marker. Nature Clin Practice Oncol *2*: 482-483, 2005.
30 Simon R and Altman D: Statistical aspects of prognostic factor studies in oncology. Br J Cancer *69(6)*: 979-985, 1994.

Received February 13, 2007
Accepted February 20, 2007