EUROPEAN JOURNAL OF CANCER XXX (2008) XXX–XXX

available at www.sciencedirect.com

journal homepage: www.ejconline.com

Lost in translation: Problems and pitfalls in translating laboratory observations to clinical utility

*Richard Simon**

National Cancer Institute, Division of Cancer Treatment and Diagnosis, 9000 Rockville Pike, MSC 7434, Bethesda, MD 20892, USA

ARTICLE INFO

Article history:
Received 12 February 2008
Accepted 23 September 2008

Keyword:
Breast

ABSTRACT

Developments in whole genome biotechnology have dramatically increased the opportunities for developing more effective therapeutics and for targeting them to patients who require them and who can benefit from them. This can have profound benefits for patients and for the economics of health care. There are, however, many obstacles to overcome in achieving this revolution. The effectiveness of translational research in oncology is seriously limited by many factors, both structural and scientific. Some of the obstacles involve the failure of biomedical organisations to develop and fund new models of inter-disciplinary collaboration needed to attract and support the best and brightest quantitative scientists to predictive medicine. Many of the challenges are scientific, requiring paradigm changes in the way drugs are developed and in the way clinical trials are designed and analysed. Some of these issues are addressed here, specifically in the context of developing molecular diagnostics in a manner that moves retrospective correlative science to prospective predictive medicine.

Published by Elsevier Ltd.

1. Introduction

Translational research in oncology involves translating basic research discoveries into products and interventional strategies that reduce the burden of cancer. Development of anti-oestrogens, angiogenesis inhibitors, imatinib and trastuzumab might be taken as examples of successful translational research. In spite of the substantial advances in basic research, improvements in prevention, early detection and treatment have been more modest. The previous examinations of this gap had focused primarily on organisational and funding issues.[1–3] Here, we will explore some problems and pitfalls in the way that translational research is conducted and will provide some suggestions for improvement.

2. Translational research is usually based on incomplete understanding of biological mechanisms

One factor that makes effective translational research difficult is limited understanding of tumour biology. It can be argued that in spite of the progress in basic research, today we do not adequately understand the pathogenesis of any type of human tumour. Consequently, it is often not clear what findings are ripe for translating. For the examples mentioned in the introduction, in spite of the limitations in understanding the biology of the tumours involved, the oestrogen receptor, VEGF, bcr-abl fusion kinase and HER2 turned out to be important enough to form the basis for the development of impor-

* Tel.: +1 301 496 0975; fax: +1 301 402 0560.
E-mail address: rsimon@nih.gov
0959-8049/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.ejca.2008.09.009

tant therapeutics. In fact, many useful medical interventions have been developed without good understanding of biology; e.g. the development of rabies vaccine by Pasteur. Although our reductionist approach to understanding cancer biology is appropriate, for the short and mid term future, successful translational research will have to take place in the context of very incomplete understanding of tumour biology.

Biology is often compared unfavourably with physics with regard to the development of fundamental laws. Biologists often excuse this lapse on the basis that biology is much more complex. Another perspective, however, is that physics has been immensely successful in providing the basis of our technological society because physicists have focused on predictive laws, rather than on trying to understand the why of those laws. Both Newton's laws and the laws of quantum mechanics are phenomenological laws that provide accurate predictions and enable important developments, but few physicists would claim to understand why those laws work.

This distinction between prediction and mechanistic understanding has a parallel in gene expression profiling of tumours. It is often much easier to develop a classifier that predicts accurately than it is to understand the biology of the tumour. Although classifier development is often done poorly, with the right specimens, study designs and biostatistical methods, it can be rightly done in a predictable time frame. On the other hand, scientists spend careers trying to understand biological systems much simpler than mammalian tumours. Accurate and robust predictive classifiers should not be rejected because we do not understand the underlying biology or because the particular genes used in the prediction may not be unique.

3. Translational research requires identifying key therapeutic targets

We have vigorous biotechnology and pharmaceutical sectors for developing potent inhibitors of identified therapeutic targets and effective infrastructures for conducting high quality clinical trials. A bottleneck to progress, however, is the identification of the key molecular targets. Most tumours are genetically heterogeneous and many abnormally expressed or mutated genes may not represent good therapeutic targets because they are characteristic of only a subset of the tumour cells. Depending on the predominance of the target, the treatment may cause tumour shrinkage, but not substantial prolongation of life.

We need better information about the key oncogenic mutations that lead to the development of an invasive tumour. A treatment that specifically inhibits the protein product of an oncogenic mutation should, if delivered early enough, be effective against all tumour cells, and non-toxic to normal cells. One still might select for cells that are resistant to a given drug because a subsequent mutation interferes with the target binding, but that type of resistance is more easily overcome, as experience with Gleevac has demonstrated. Success may, however, be dependent on early treatment. Even human solid tumours may undergo the equivalent of a 'blast crisis' in which a destabilised genome undergoes an 'information meltdown' with such substantial genomic heterogeneity that effective treatment is almost impossible.[4]

Although it usually takes more than one mutational event for oncogenesis, the number of such mutations may be small. Zhang and Simon[5] estimated the number of mutational events occurring at rates of point mutations and the loss of heterozygosity for mammalian cells leading to breast cancer. Their estimates were based on a mathematical model of breast epithelia dynamics and data on the age-incidence of breast cancer in the United States (US) population and in carriers of BRCA1 or BRCA2 mutations. They interpreted their findings as indicating that 2–3 mutations initiate a process which leads inexorably to sporadic breast cancer. Numerous additional mutations subsequently develop during invasion and metastasis of a proliferating and genomically unstable population of tumour cells. For BRCA1 or BRCA2 mutation carriers, their model indicated that loss of the normal allele and one additional mutational event initiated the process that leads inexorably to invasive breast cancer.[6]

4. The architecture of translational research

Effective translational research involves many components and partnerships. For our limited purposes, we will briefly comment on some considerations during a pre-clinical development phase, a clinical development phase and a clinical validation phase.

4.1. *Pre-clinical development*

The pre-clinical development phase may be extensive because it involves the bridging research and development needed to begin translating a biological discovery into a drug, diagnostic or intervention that may be used in patients. This phase begins after the basic research discovery, and requires focus on both the initial discovery and the intended application. Although important discoveries in basic research may occur by serendipity, effective translational research usually requires focus on the discovery and on the type of product desired. No one individual has sufficient breadth of knowledge to perform this effectively. Technology specialists cannot be expected to have sufficient expertise in oncology applications, and unless they partner with those who have such knowledge, successful translational research is unlikely. Lack of such partnership is often a key reason why many promising technologies are not effectively developed. This type of partnership is sometimes possible in large companies, where the required breadth of expertise is available, but even in such settings there may be departmental barriers or lack of appreciation of the importance of integrated development starting at the pre-clinical development phase.

Academic scientists are usually funded and rewarded for discovery, rather than to pursue focused translational research as members of a large inter-disciplinary team. Funding agencies such as the National Institutes of Health are not experienced in funding and monitoring focused translational team research.

Because of the structural limitations of conducting and funding translational research focused on translating a

Please cite this article in press as: Simon R, Lost in translation problems and pitfalls in translating..., *Eur J Cancer* (2008), doi:10.1016/j.ejca.2008.09.009

defined discovery to a product for use in a defined medical context, many discoveries go untranslated unless they are of interest to the industry.

4.2. *Clinical development*

The clinical development of a drug, diagnostic or technology also requires clear focus on the intended application. Pusztai et al.[7] identified 939 publications over a 20-year period on prognostic factors for patients with breast cancer. Other than the traditional staging variables, only four factors, oestrogen receptor, progesterone receptor and HER2 amplification and OncoType Dx™ recurrence score were recommended by ASCO guidelines. Several reasons might be mentioned for this apparent waste of effort. First, few of the markers studied were properly validated; nearly all of these studies were *developmental* studies rather than *validation* studies. Most investigators are interested in developing new prognostic factors or in using them in new ways, rather than validating the prognostic models published by others. Second, most of the studies were performed using *convenience samples* of available specimens. These specimens often are from a heterogeneous collection of patients who have received a variety of treatments. It is generally difficult to use such results in the therapeutic decision making for individual patients. Finally, most of the publications were based on research assays without demonstration of robustness or analytical validity. Academic investigators are not well suited or rewarded for such reproducibility studies. Without a diagnostic company to develop a robust assay for a test with a clear and important medical application, the publication is unlikely to be part of successful translational research.

One important area of translational research today is biomarker development. There are at least four common types of applications of biomarkers, and the nature of the application has a fundamental bearing on the kind of developmental and validation studies needed. Traditionally 'biomarker' referred to a biological measurement that was reflective of disease status, increasing as the disease progresses and decreasing as it regresses. Such a biomarker measured sequentially could be used to monitor treatment effect for purposes of patient management or drug development. Unfortunately, treatment effect is not the same as treatment effectiveness. For example, the size of a measurable tumour might be thought of as a biomarker. Reduction in size of the lesion may reflect the effect of treatment, but might not result in prolongation of patient survival or in improvement in patient symptoms. Establishing that a biomarker is a valid surrogate of clinical benefit is very difficult. It requires a series of randomised clinical trials demonstrating the concordance of treatment difference on the clinical end-point with treatment difference as measured with regard to the candidate biomarker.[8] Such demonstration would generally have to be established in the context of a specific type of cancer and perhaps in the context of a specific class of drugs. Regulatory agencies naturally require strong evidence for establishing that a biomarker is a valid surrogate of clinical benefit if the biomarker is to be used as the basis of drug approval. In therapeutics development, it is often more expedient to use clinical end-point than to attempt to establish a biomarker as a valid surrogate.

A disease status biomarker may be useful in the early stages of new drug development even if it is not a valid surrogate of clinical benefit. A biomarker might be called a *partial surrogate* if the change in it is necessary but perhaps not sufficient for patient benefit. Such biomarkers could be used in the early clinical development in a variety of ways; e.g. for dose/schedule selection, for identifying an appropriate patient population for a phase III trial, and for deciding whether phase III evaluation is warranted. Although there may be uncertainty in whether the biomarker is truly a partial surrogate, it may be appropriate to use the biomarker for designing a phase III trial in which a regimen will be tested for a defined population using a clinical end-point that is an accepted measure of patient benefit.

There is increasing awareness of the importance of baseline prognostic and predictive biomarkers that aid in the treatment selection. The Oncotype Dx™ recurrence score is an example of a therapeutically relevant prognostic biomarker.[9] It provides prognostic information for patients with node negative oestrogen receptor positive breast cancer receiving Tamoxifen. Because it was developed and validated for such a clearly defined set of patients, it can be used to identify patients whose prognosis on Tamoxifen is sufficiently favourable that they may elect not to receive cytotoxic chemotherapy. The Mamaprint prognostic score is a prognostic index with a similar therapeutic application although it was developed for a more mixed group of patients receiving no systemic therapy.[10–12]

Predictive biomarkers are pre-treatment measurements, which can be used to predict the likelihood that a given patient will benefit from a given treatment. For example, HER2 amplification is a predictive marker for benefit from trastuzumab. Because cancer is a life threatening disease, we frequently treat the majority for the benefit of the minority. For example, an adjuvant treatment that increases the long-term disease-free survival rate from 80% to 85% would be considered valuable although it involves treating 100 patients to benefit 5.[13] If we could predict, which patients were likely or unlikely to benefit from the treatment, then we could potentially spare patients unnecessary adverse effects, triage patients to treatments most likely to benefit them, and reduce health care costs associated with ineffective treatments. Such predictive classifiers can also potentially be used to improve the efficiency of clinical trials.[14,15]

The potential value of predictive biomarkers for patients, clinical development and health care economics supports an increased emphasis on the development of such biomarkers. Substantial knowledge about the therapeutic target facilitates the process of developing a useful predictive biomarker. In the case of trastuzumab, HER2 over-expression was initially measured at the protein level using immunohistochemistry. Although that assay was quite imperfect, it played an important role in facilitating the early approval of trastuzumab. Even the current FISH test for gene amplification does not provide perfect predictive accuracy of which patients will benefit from trastuzumab. A predictive biomarker can be of tremendous benefit for patient management and for clinical drug development even if it is far from perfect. Simon and Maitournam showed that the number of patients needed to randomise for a clinical trial of new drug versus control regi-

Please cite this article in press as: Simon R, Lost in translation problems and pitfalls in translating..., *Eur J Cancer* (2008), doi:10.1016/j.ejca.2008.09.009

men can be dramatically reduced if the patient population can be enriched for patients likely to benefit from the new drug even if the predictive biomarker is quite imperfect.[14,15] Their computer programs for sample size determination for targeted trials compared to standard un-targeted trials are available for on-line use at http://linus.nci.nih.gov/brb.

In developing a predictive biomarker for a drug with multiple targets or uncertain targets, a reverse-genomics approach can be used. The traditional translational research was hypothesis driven and dependent on assay development for each protein to be studied. Using gene expression profiling, however, one can potentially avoid having to be smart enough to know in advance the genes or proteins to use as predictive biomarkers. Instead, one can use a training set of tumour expression profiles for responders and non-responders to develop a predictive classifier of the tumours that are likely to respond. Dobbin et al.[16,17] studied the relationship between sample size and achievable predictive accuracy, and recommended a minimum of 20–30 responders and a similar number of non-responders for classifier development. Their sample size planning methods are available on-line at the site indicated above. One can develop a predictive classifier of the patients likely to respond to the new drug or a predictive classifier of the patients unlikely to respond to the standard therapy. One can alternatively develop a predictive classifier of the patients more likely to respond to the new regimen than to the standard therapy.[18,19]

With gene expression profiling, it is more appropriate to talk in terms of a predictive classifier or a predictive index rather than a predictive biomarker because the prediction is based on combining the expression levels of multiple genes in a mathematically defined manner. A predictive classifier is not just a set of genes that have been found differentially expressed between responders and non-responders. In some cases, authors of publications reporting the ability to predict response to a given treatment do not provide the details about the predictive function used; i.e. the relative weights for the different components of the classifier or the classification threshold. Consequently, it is impossible for others to either use or independently try to validate their classifier.

Dupuy and Simon[20] reviewed the cancer literature of studies relating gene expression profiles to patient outcome, either response to treatment, survival or disease-free survival. They found that 50% of the publications had at least one flaw so serious as to raise questions about the validity of the conclusions. The three most common serious flaws they found were misleading use of cluster analysis, lack of adjustment for the multiplicity of analysing thousands of genes and erroneous use of partial cross-validation. They pointed out that cluster analysis rarely has a valid role in the development of predictive classifiers. Its wide use in the literature reflects a lack of proper statistical guidance or collaboration in the conduct of expression profiling studies. Cancer research organisations need to better appreciate the fundamental changes occurring in the nature of biomedical research, and make major commitments to departments for providing professional biostatistical collaboration as an integral part of translational research. There has been some misunderstanding that the greatest challenge in effectively utilising the new tools of whole genome biotechnology is managing the volume of data, rather than providing biostatistical collaboration to plan and analyse studies that generate biological insight and medical utility.[21]

The clinical development of therapeutics is itself much more complex today in the era of molecular targeting. Although most molecularly targeted drugs are sufficiently toxic that they are administered at their maximum tolerated dose, assays are needed to determine whether they inhibit their targeted pathways *in vivo*, and this should be assessed in phase I trials. Because there are more approved drugs available today, patients eligible for single agent phase II trials are heavily pre-treated, and may represent an invalid context for evaluating the potential of the drug. If the drug is administered in combination with the existing drugs, then evaluation of the contribution of the new drug will not be apparent in single arm studies. Even if the new drug is administered as a single agent, its potential cytostatic effect will not be observable in single arm studies unless a sensitive biomarker of tumour proliferation is available. Molecularly targeted drugs are also more likely to be effective for only a subset of patients whose tumour is driven by the pathway targeted by the drug. Unless those tumours can be identified based on knowledge of the molecular target or pre-clinical findings, much larger phase II trials may be required. For these reasons, traditional small single arm phase II studies of heavily previously treated patients are much less adequate for translational research today. The tradition of relying on individual cancer centres for clinical development of new therapeutics may no longer be viable. New designs, including randomised designs and neo-adjuvant treatment need to be considered.

4.3. *Clinical validation*

The principles of clinical validation are well established for therapeutics; randomised clinical trials of a defined regimen for a defined patient population using an end-point that is a direct measure of clinical benefit. Exploratory analyses of dose-schedules, patient subsets and unvalidated surrogate end-points are more appropriate for developmental studies. Similar principles have not been established for predictive and prognostic biomarkers, and there has been a substantial confusion about what it means to 'validate' such biomarkers.[22] The evaluation of multivariate predictive classifiers based on gene expression profiling or serum proteomics has been particularly problematic.[20,23–25] We shall therefore touch on some of these issues here.

When dealing with predictive classifiers based on genome-wide assays, it is essential to distinguish the data used to develop the classifier from the data used to evaluate the classifier. This careful partitioning is not observed in traditional statistical analysis where the number of cases is many times the number of variables. In fact, for traditional statistical regression analyses it is recommended to have at least 10 times as many cases as variables. In developing predictive classifiers with gene expression data or serum proteomic data, the number of variables is orders of magnitude greater than the number of cases, and the traditional statistical model development approaches may give misleading results.

The cardinal principle for evaluating a predictive classifier is that the data used for evaluating the classifier should

Please cite this article in press as: Simon R, Lost in translation problems and pitfalls in translating..., *Eur J Cancer* (2008), doi:10.1016/j.ejca.2008.09.009

not be used in any way for building the classifier. The simple *split-sample* method achieves this by partitioning the study cases into two parts. The separation is often done randomly, with half to two-thirds of the cases used for developing the classifier and the remainder of the cases in the test set. The cases in the test set should not be used *in any way*, until a *single completely specified predictive model* is developed using the training data. At that time, the test cases are simply classified using the single completely specified classifier. For example, with a gene expression profile classifier, the classifier is applied to the expression profiles of the cases in the test set, and each of them is classified, e.g. as a responder or a non-responder to the therapy. The patients in the test set have received the treatment in question, and so one can count how many of those predictive classifications were correct and how many were incorrect. In using the split-sample method properly, a single classifier should be defined on the training data. It is not valid to develop multiple classifiers and then use their performance on the test data to select among the classifiers.[26] It is also completely invalid to use the full dataset of all cases to select the genes that will be used for classification and then fit a model with those genes using the training cases.[23] These invalid practices are not uncommon in the literature of even high profile journals[20] Many journals do not appear to have editorial boards sufficiently knowledgeable in statistical genomics to select qualified referees for submitted papers. Inability of journals to adequately pre-screen manuscripts creates great demands on the limited number of qualified referees. The number of major publications relating gene expression or serum proteomic profiles to outcome without involvement of experienced biostatistical collaborators suggests that insufficient resources are being provided for inter-disciplinary collaboration in this area.

There are more complex forms of dividing the data into training and testing portions. These *cross-validation* or *re-sampling* methods utilise the data more efficiently than the simple division described above.[27] Cross-validation generally partitions the data into a large training set and a small test set. A classifier is developed on the training set, and then applied to the cases in the test set to estimate the error rate. This is repeated for numerous training-test partitions, and the prediction error estimates are averaged. Molinaro et al. showed that for small datasets (e.g. less than 100 cases), leave-one-out cross-validation or 10-fold cross-validation can provide much more accurate estimates of prediction accuracy than the split-sample approach.[27] Michiels et al.[28] suggested that multiple training-test partitions be used, rather than just one. The split-sample approach is mostly useful, however, when one does not have a well-defined algorithm for developing the classifier. When there is a single training set-test set partition, one can perform numerous analyses on the training set to develop a classifier and use biological considerations of which genes to include before deciding on the single classifier to be evaluated on the test set. With multiple training-test partitions, however, this type of flexible approach to model development cannot be used. If one has an algorithm for classifier development, it is generally better to use one of the cross-validation approaches to estimate error rate because the split-sample approach, or the replicated split-sample approach, does not provide as efficient a use of the available data.

In order to honour the key principle of not using the same data to both develop and evaluate a classifier, it is essential that for each training-test partition the data in the test set are not used in any way. Hence, a model should be developed from scratch in each training set. This means that multiple classifiers are developed in the process of doing cross-validation, and those classifiers will in general involve different sets of genes. It is invalid to select the genes beforehand using all the data, and then to just cross-validate the model building process for that restricted set of genes. Radmacher et al.[29] and Ambroise and McLachlan[30] demonstrated that such pre-selection results in severely biased estimates of prediction accuracy. In spite of this known severe bias, this error was one of the most common serious errors found in the literature review by Dupue and Simon.[20] It is also made in many biased reports touting the merits of new kinds of classifiers.[31]

Some authors use complete cross-validation to produce estimates of prediction accuracy for a variety of predictive classifiers. Then they select the classifier having the smallest cross-validated error. That estimate is itself biased, being the minimum of a set of random quantities.[26] The optimisation of tuning parameters or classifier types ideally should be considered part of the classifier development algorithm, and the entire algorithm should be included in each loop of the cross-validation used to estimate predictive accuracy.

Developmental studies should use either the split-sample method or complete cross-validation to provide an unbiased estimate of prediction accuracy. Although the study may be too small for these estimates to be precise, the study should demonstrate that prediction accuracy is better than can be obtained by chance without using the expression data. For complete cross-validation, this can be accomplished by generating the distribution of cross-validated prediction accuracy under permutations of the outcome data as suggested by Radmacher et al.[29] This approach is preferable to the test suggested by Michiels et al.[28]

The predictive classifiers are constructed in *developmental* studies. *Validation* studies should test pre-specified classifiers. The estimates of predictive accuracy obtained by data splitting or cross-validation are types of *internal validations*. Taking one set of data collected and assayed under carefully controlled research conditions, and splitting it into training and testing sets are not the same as evaluating the predictive accuracy of a classifier on a new set of patients from different centres with tissue collection and assay performance more representative of real-world conditions. Developmental studies are often too limited in size, structure and the nature of the cases to establish medical utility of a predictive classifier.

Validating a predictive classifier means validating that the classifier predicts accurately for independent data. It does not mean that the same genes would be selected in developing a classifier with independent data. This point is often misunderstood, and is a source of inappropriate criticism of expression profiling studies.[32] The expression levels among genes are highly correlated. It has been long known for regression analysis that in such settings there are many models that predict about equally well. This is even more the case for genomic studies where the number of candidate variables is

Please cite this article in press as: Simon R, Lost in translation problems and pitfalls in translating..., *Eur J Cancer* (2008), doi:10.1016/j.ejca.2008.09.009

large relative to the number of cases. It would take enormous numbers of cases to distinguish the small differences in predictive accuracy among such models,[33] but it is a very inappropriate criterion for sample size planning. Dobbin and Simon have shown that much smaller sample sizes are generally needed to obtain predictive classifiers with accuracy within 5–10% points to the accuracies that could be achieved with unlimited cases.[16,17]

Some investigators attempt to perform external validation of the predictive accuracy of a classifier using expression data, already available, from a separate study. In some cases, the second study used a different platform for measuring gene expression. Any adjustments for platform differences must be made with great care, however, not to introduce bias into the validation. Validation of predictive accuracy using external data should involve using a single completely specified classifier developed without using the validation data in any way. The completely specified classifier should be simply applied to the expression profiles of the validation cases to make predictions of outcome. Those predictions are then compared to the true outcomes, and predictive accuracy is assessed. When the validation data are from a different platform, true external validation of the original predictive classifier is not possible. An intermediate kind of validation can be attempted by partitioning the validation data into a training set and a test set. The expression profiles in the training set are used to modify the original predictive classifier for use on the new platform. The outcomes of the cases in the training set of the validation data should *not* be used in this modification process. If they are used, then the process is not really a validation of the original classifier. The modified model is then applied to the test cases in the validation data with no further modifications. In this way, there should be no question about whether the process produced biased estimates of predictive accuracy.[25]

The external validation study should be designed to establish the medical utility of the classifier, not just its predictive accuracy. In a new drug development, medical utility may mean that a treatment regimen including the new drug is more effective than the control regimen for classifier positive patients, but not for classifier negative patients. This can be addressed by a randomised clinical trial comparing the two regimens, and sized for adequate separate analysis of classifier positive and classifier negative patients. Stratifying the randomisation using the binary classifier ensures that specimens will be available and classified for all randomised patients. Stratification also ensures that one has a prespecified binary classifier with pre-specification of threshold for distinguishing test positive from test negative cases. In some cases, it may not be ethically appropriate to include classifier negative patients as there may be compelling biological basis for believing that they will not benefit from the new regimen. In these cases, the enrichment designs of Simon and Maitournam[14,15,34] can be used. Designs which do not exclude classifier negative patients have been described by Sargent et al.,[35] Pusztai and Hess,[36] Freidlin and Simon,[18] Jiang et al.,[19] Simon and Wang,[37] and Song and Chi.[38]

Evaluating a predictive or prognostic classifier to guide application of an approved and widely used treatment can be difficult because it may be difficult to conduct a study that restricts the use of an effective regimen. Establishing medical utility may require establishing that the predictive classifier is more effective than the standard practice guidelines for providing treatment selection; e.g. results in better patient outcome (or a similar outcome with less adverse events). Establishing medical utility depends on the available treatment options and the current standards of care. A key step in developing a useful predictive classifier is identifying a key therapeutic decision setting that can be potentially improved based on genomic data.

An independent validation study could be a prospective clinical trial in which patients are randomised to treatment assignment based on the standards of care versus treatment assignment with the aid of the classifier. This design requires that the classifier be determined only in half of the patients. It is often very inefficient, however, because many patients will receive the same treatment either way they are randomised. A better alternative is to perform the assay up front for all patients, and then randomise only those for whom the classifier specified treatment differs from the practice guidelines. Even for this design, however, the required sample size is likely to be very large. In some cases, external validation is possible using archived specimens from an appropriate randomised clinical trial that was performed before the treatment in question had become established. This was done for the case of Oncotype DX[9,39]. With this approach to validation, the classifier and statistical analysis plan should be prospectively specified. In using archived specimens, there is often a question of whether the available specimens are representative, and one must establish externally the robustness and analytical validity of classifier measurement. Nevertheless, this approach is very efficient, and it provides a strong motivation for archiving tumour tissue for all patients in major randomised clinical trials.

5. Conclusions

The effectiveness of translational research in oncology is limited by many factors, both structural and scientific. As clinical research has become more complex and less well supported, translational research has become more difficult than ever. Developments in whole genome biotechnology have dramatically increased the opportunities for targeting therapeutics to patients who require them and who can benefit from them. This can have profound benefits for the economics of health care. There are, however, many obstacles to overcome in achieving this revolution. Some of these obstacles are regulatory, and some involve structural limitations of academic medicine, funding agencies and industry. Some of the obstacles involve the failure of biomedical organisations to develop and fund new models of inter-disciplinary collaboration needed to attract and support the best and brightest quantitative scientists to predictive medicine. Many of the challenges are scientific, requiring paradigm changes in the way the drugs are developed and in the way clinical trials are designed and analysed. Many of these problems are not new and have been recognised before. However, the opportunities to use genomics and biotechnology to reduce mortality from cancer and to have major impact on the economics of healthcare are

Please cite this article in press as: Simon R, Lost in translation problems and pitfalls in translating..., *Eur J Cancer* (2008), doi:10.1016/j.ejca.2008.09.009

unprecedented. The opportunities and importance of achieving success are sufficiently great that a critical re-examination of these obstacles is in order. Progress will likely come from those organisations with leadership, vision and resources to re-structure themselves in new ways that permit them to nurture broad inter-disciplinary teams of basic, clinical and quantitative scientists focused on translating the most important development in basic research to products for key clinical applications.

Conflict of interest statement

None declared.

REFERENCES

1. Hait WN. Translating research into clinical practice: deliberations from the American Association for Cancer Research. *Clin Cancer Res* 2005;**11**:4275–7.
2. Hait WN. Sustaining the clinical in clinical translational research. *Clin Cancer Res* 2006;**12**:1–2.
3. Paul M. Translational investigators: life sciences' application engineers. *Nat Biotechnol* 2007;**25**:817–8.
4. Eigen M, Schuster P. *The hypercycle – a principle of natural self-organization*. Berlin: Springer; 1979.
5. Zhang X, Simon R. Estimating the number of rate-limiting genomic changes for human breast cancer. *Breast Cancer Res Treat* 2005;**91**:121–4.
6. Simon R, Zhang X. On the dynamics of breast tumor development in women carrying germline BRCA1 or BRCA2 mutaions. *International Journal of Cancer* 2008;**122**:1916–7.
7. Pusztai L, Ayers M, Stec J, Hortobagyi GN. Clinical application of cDNA microarrays in oncology. *Oncologist* 2003;**8**:252–8.
8. Torri V, Simon R, Russek-Cohen E, Midthune D, Friedman M. Relationship of response and survival in advanced ovarian cancer patients treated with chemotherapy. *J Natl Cancer Inst* 1992;**84**:407–14.
9. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of Tamoxifen-treated, node-negative breast cancer. *New Engl J Med* 2004;**351**:2817–26.
10. Van't-Veer LJ, Dai H, Vijver MJVD, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;**415**:530–6.
11. Buyse M, Loi S, Veer Lvt, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst* 2006;**98**:1183–92.
12. Simon R. Development and evaluation of therapeutically relevant predictive classifiers using gene expression profiling. *J Natl Cancer Inst* 2006;**98**(17):1169–71.
13. Miller JD. Finding clinical meaning in cancer data. *J Natl Cancer Inst* 2007;**99**:1832–5.
14. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin Cancer Res* 2005;**10**:6759–63.
15. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials: supplement and correction. *Clin Cancer Res* 2006;**12**:3229.
16. Dobbin K, Simon R. Sample size planning for developing classifiers using high dimensional DNA expression data. *Biostatistics* 2007;**8**:101–17.
17. Dobbin KK, Zhao Y, Simon RM. How large a training set is needed to develop a classifier for microarray data? *Clin Cancer Res* 2008;**14**:108–14.
18. Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin Cancer Res* 2005;**11**:7872–8.
19. Jiang W, Freidlin B, Simon R. Biomarker adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *J Natl Cancer Inst* 2007;**99**:1036–43.
20. Dupuy A, Simon R. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst* 2007;**99**:147–57.
21. Simon R. Bioinformatics in cancer therapeutics-hype or hope? *Nat Clin Pract Oncol* 2005;**2**:223.
22. Simon R. When is a genomic classifier ready for prime time? *Nat Clin Pract Oncol* 2004;**1**(1):2–3.
23. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 2003;**95**:14–8.
24. Coombes KR, Morris JS, Hu J, Edmonson SR, Baggerly KA. Serum proteomics profiling – a young technology begins to mature. *Nat Biotechnol* 2005;**23**:291–2.
25. Coombes KR, Wang J, Baggerly KA. Microarrays: retracing steps. *Nat Med* 2007;**13**:1276–7.
26. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform* 2006;**7**:91.
27. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 2005;**21**(15):3301–7.
28. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple validation strategy. *Lancet* 2005;**365**:488–92.
29. Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. *J Comput Biol* 2002;**9**:505–12.
30. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci* 2002;**99**:6562–6.
31. Lai C, Reinders MJT, Veer LJvt, Wessels LFA. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC Bioinform* 2006;**7**:235.
32. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DSA, Nobel AB, et al. Concordance among gene-expression based predictors for breast cancer. *New Engl J Med* 2006;**355**:560–9.
33. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci* 2006;**103**:5923–8.
34. Maitournam A, Simon R. On the efficiency of targeted clinical trials. *Stat Med* 2005;**24**:329–39.
35. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *J Clin Oncol* 2005;**23**(9):2020–7.
36. Pusztai L, Hess KR. Clinical trial design for microarray predictive marker discovery and assessment. *Ann Oncol* 2004;**15**:1731–7.
37. Simon R, Wang SJ. Use of genomic signatures in therapeutics development. *Pharmacogenom J* 2006;**6**:1667–73.
38. Song Y, Chi GYH. A method for testing a prespecified subgroup in clinical trials. *Stat Med* 2007;**26**:3535–49.
39. Paik S. Development and clinical utility of a 21-gene recurrence score prognostic assay in patients with early breast cancer treated with Tamoxifen. *Oncologist* 2007;**12**:631–5.

Please cite this article in press as: Simon R, Lost in translation problems and pitfalls in translating..., *Eur J Cancer* (2008), doi:10.1016/j.ejca.2008.09.009