MAIN PAPER

Pharmaceutical Statistics

(wileyonlinelibrary.com) DOI: 10.1002/pst.532

Published online 9 December 2011 in Wiley Online Library

A statistician's perspective on biomarkers in drug development

Martin Jenkins,[a]* Aiden Flynn,[b] Trevor Smart,[c] Chris Harbron,[a] Tony Sabin,[d] Jayantha Ratnayake,[e] Paul Delmar,[f] Athula Herath,[g] Philip Jarvis,[a] James Matcham,[d] and on behalf of the PSI Biomarker Special Interest Group

Biomarkers play an increasingly important role in many aspects of pharmaceutical discovery and development, including personalized medicine and the assessment of safety data, with heavy reliance being placed on their delivery. Statisticians have a fundamental role to play in ensuring that biomarkers and the data they generate are used appropriately and to address relevant objectives such as the estimation of biological effects or the forecast of outcomes so that claims of predictivity or surrogacy are only made based upon sound scientific arguments. This includes ensuring that studies are designed to answer specific and pertinent questions, that the analyses performed account for all levels and sources of variability and that the conclusions drawn are robust in the presence of multiplicity and confounding factors, especially as many biomarkers are multidimensional or may be an indirect measure of the clinical outcome. In all of these areas, as in any area of drug development, statistical best practice incorporating both scientific rigor and a practical understanding of the situation should be followed. This article is intended as an introduction for statisticians embarking upon biomarker-based work and discusses these issues from a practising statistician's perspective with reference to examples. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: biomarkers; statistics; personalized health care; safety biomarkers; surrogate endpoint

INTRODUCTION

Within drug discovery and development, biomarkers play a crucial role in understanding the mechanism of action of a drug, identifying efficacy or toxicity signals at an early stage of development and in identifying patients likely to respond to treatment.

The aim of this paper is to review statistical issues in biomarker-based drug development, focusing on common issues from a practicing statistician's perspective. The paper is structured as follows. First, the concept and common uses of biomarkers will be introduced and the appropriate preparations and statistical considerations for fit-for-purpose biomarker validation will be discussed. The challenges frequently faced by statisticians in biomarker usage will be examined, including specific issues relating to multivariate endpoints and signatures. Particular considerations for the use of biomarkers in personalized health care (PHC) and toxicity will also be covered. The final section will discuss the appropriateness of formal biomarker qualification, closing with more general thoughts for statisticians working in this field.

1. WHAT ARE BIOMARKERS AND WHAT ARE THEY USED FOR?

The biomarker definitions working group[1] of the National Institutes of Health established the following definition of a biomarker. ***Biological marker (biomarker): A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.***

The term 'biomarker' covers a very wide range of data types, uses and applications across all stages of pharmaceutical development. Typical examples include measurements taken from biological samples, such as cytokines in blood, or from images, such as via FDG-PET. Some examples are given in Table I.

Several classification methods have been proposed for biomarkers based upon their intended application [2] or the strength of available evidence [3, 4]. However, in statistical terms, the key distinction is between the biomarker used at a single time point (often baseline), or as a dynamic endpoint which changes in

[a] *AstraZeneca UK Ltd., Alderley Park, Macclesfield, UK*

[b] *Exploristics, Belfast, UK*

[c] *Pfizer, Sandwich, Kent, UK*

[d] *Amgen Limited, Cambridge Science Park, Cambridge, UK*

[e] *Roche Products Limited, Shire Park, Welwyn Garden City, UK*

[f] *Hoffman–La Roche, Basel, Switzerland*

[g] *MedImmune, Granta Park, Cambridge, UK*

**Correspondence to: Martin Jenkins, AstraZeneca UK Ltd., Alderley Park, Macclesfield, UK.*
E-mail: martin.jenkins@astrazeneca.com

This article is published in Pharmaceutical Statistics as a special issue on Focusing on the PSI Special Interest Groups, edited by John Stevens, Centre for Bayesian Statistics in Health Economics, ScHARR, Regent Court, 30 Regent Street, Sheffield, South Yorkshire, S1 4DA, UK.

Table I. Examples of biomarker use.

Biomarker	Current Use	Classification	Qualification
HER2, EGFR, K-RAS mutations	Directing treatment in oncology	Predictive Biomarker	Defines indication in label, diagnostic development required
P450 enzymes (CYP2D6, CYP2C9, CYP2C19 polymorphisms)	Known to affect drug metabolism (e.g., for NSAIDs)	Predictive Biomarker	Can appear in label as risk factor. Prior testing suggested, dose adjustment
UGT1A1, TMPT, HLA-B*5701 polymorphisms	Predisposition to certain toxicities (e.g., liver, bone marrow)	Predictive Biomarker	Can appear in label as risk factor. Prior testing suggested, dose adjustment
AB1-42	Diagnosis of prodromal Alzheimer's Disease	Prognostic marker	Used to enrich clinical trial populations. Example of qualification procedure
Gene signature chips (e.g., Oncotype, Mammaprint)	Prognosis prediction in oncology	Prognostic marker (also predictive in certain cases)	Diagnostic qualification process applies
CRP, IL-6, TNFa in blood samples	Proof of principle in inflammatory diseases	Pharmacodynamic biomarker	Formal qualification not required, but fit for purpose assay validation
FDG-PET (SUVmax) Functional imaging	Proof of concept (e.g., in tumour metabolism)	Pharmacodynamic biomarker	Formal qualification not required, but collaborative opportunities
LDL cholesterol	Confirmatory trials in coronary heart disease	Surrogate Endpoint	Appears in label, used for approval. Any such new markers require qualification
HbA1c	Represents glycemic control in diabetics	Surrogate Endpoint	Appears in label, used for approval. Any such new markers require qualification

response to an intervention. Various types of biomarker usage are illustrated in a simple fashion in Figure 1.

1.1. Prognostic and predictive biomarkers

When the intention is to use a biomarker to explain the variation in outcomes, be this in terms of disease prognosis, treatment response or occurrence of toxicities, then it will be considered as a covariate in a model of this clinical outcome. The biomarker will act as a predictor, with the aim of explaining the variation in these responses. This prediction will be of most practical value if it can be deduced from a baseline measurement, but in theory predictions could be made on the basis of any time point or combination of time points prior to the measurement of response. As an illustration, the number of circulating tumour cells (CTCs) in peripheral blood at baseline can give an indication of survival prognosis in prostate cancer, but the number of cells in the most recent blood draw can be even more informative for the management of patients [5].

Such fixed time-point usage of biomarkers could be described by the term 'cross-sectional' biomarkers. These could be divided into two major sets, 'prognostic' and 'predictive' biomarkers. These often take the form of categorical variables so as to enable the definition of biomarker sub-groups (for example those who are positive or negative expressers of a genetic marker), although a continuous variable could equally well be used in a prognostic capacity. In statistical terms, there is an important distinction between prognostic and predictive markers as this impacts the choice of statistical model applied during the analysis of biomarker data. Prognostic markers can be highlighted using models with biomarker as a fixed main effect whereas predictive markers can be identified using models with an interaction between biomarker and treatment.

Prognostic biomarkers are biomarkers that predict the prognosis or the likely outcome of the disease independent of the mode of treatment. In contrast, predictive biomarkers are biomarkers that predict the likelihood of response to a particular treatment

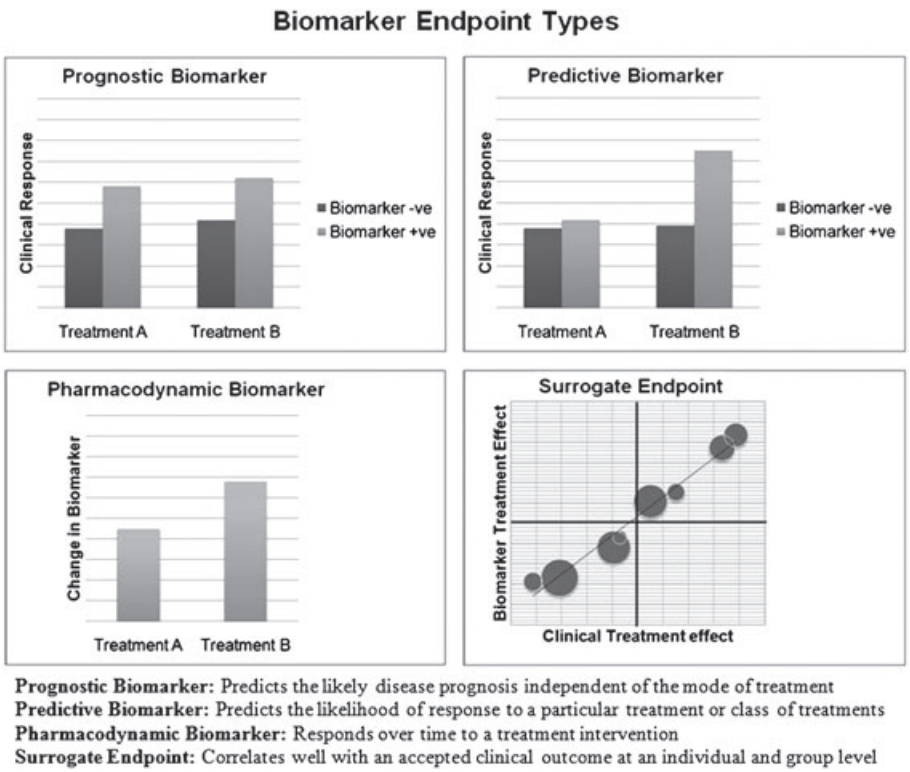

Figure 1. Biomarker Endpoint Types.

or a class of treatments. As an illustration, if expression of a prognostic biomarker (such as observing CTCs above a given cutoff), indicates a favorable outcome, then the treatment effect in a randomized trial of two treatments might still be expected to be the same in those patients who express the marker and those who do not. In comparison, if the biomarker were predictive of response to the investigative treatment, as for K-ras wild type in colorectal cancer patients treated with cetuximab [6], then the biomarker-defined group (in this case those without a mutation) demonstrate a larger treatment effect than other patients.

Biomarkers may be both prognostic and predictive (as EGFR and K-ras are in nonsmall-cell lung cancer, predicting response to EGFR Tyrosine Kinase Inhibitors [7]) and a potential source of predictive markers for specific drugs may be recognized prognostic markers. Increasingly, though predictive markers such as these are developed alongside the treatment and are related to its well understood mechanism of action.

Prognostic markers have utility in the management of patients, but also the early stages of pharmaceutical development, such as target discovery or target validation. They are also useful in the segmentation of populations, such as for setting the inclusion criteria for early stages of a clinical development program. In the mid to late stages of the clinical development program, predictive biomarkers lend themselves to targeted clinical development, as discussed in Section 5, and are often codeveloped as diagnostics to identify those who are likely to respond to the treatment.

1.2. Longitudinal Biomarkers

When the change in a biomarker is the parameter that is to be understood, explained or controlled, then a biomarker will be considered as an endpoint. The biomarker could be used in this sense as a marker of the drug activity to demonstrate proof of principle and used for optimizing the dosing schedule of the drug during the earlier phases of the development program. For example inflammatory markers such as CRP or ESR may be used to select a dose in rheumatoid arthritis treatment, or can form part of a clinical composite such as DAS28, used for the same purpose. A marker that responds to treatment intervention may be said to demonstrate a pharmacodynamic response and can be considered as a potential dynamic endpoint which could demonstrate the 'therapeutic efficacy' of the agent.

It is common to find pharmacodynamic endpoints used to address secondary or exploratory objectives as 'add-ons' in large clinical trials. Their role here may be to elucidate various scientific hypotheses about the treatment, provide differentiation from competitors or to give additional confidence to a development decision. However, providing sufficient evidence of their utility and relevance is established, as discussed in Section 2, biomarkers can be used to make key prioritization or investment decisions in their own right, especially during the translational or the proof of concept stages of development.

Pharmacodynamic biomarkers could allude to either or both of the safety or efficacy of a treatment. Safety biomarkers (including those historically referred to as lab variables in clinical development) are measured throughout preclinical and clinical stages of development and figure prominently in decision making, safety monitoring, safety surveillance and pharmacovigilance as described in Section 6.

1.3. Surrogate endpoints

Pharmacodynamic biomarkers can be used as an early outcome measure that may demonstrate some clinical benefit for the patient and indeed a certainly degree of correlation with clinical outcomes may be desired. Those markers which correlate well

with a widely accepted clinical outcome at both an individual and group level could potentially act as a surrogate endpoint and substitute for a recognized clinical endpoint, in the way that LDL cholesterol acts as a surrogate for major cardiovascular events in the licensing of statins. Sufficient evidence of this relationship must be demonstrated and these cases are discussed in Section 7. It is important to appreciate that a pharmacodynamic biomarker can still play a crucial role in pharmaceutical development even if the evidentiary standards required to demonstrate surrogacy are not met, as is generally the case.

The objectives of a biomarker's use should be clear to avoid any misunderstanding as the same biomarker could be used in a cross-sectional capacity in one study and in a pharmacodynamic capacity in another. It is the context of use that defines these terms rather than the biomarker itself. It should not be assumed that successful use of a marker in one sense automatically demonstrates its appropriateness for other applications. A common misconception is that the evidence for a successful prognostic marker confers on it the ability to act as a surrogate. For example, if a biomarker is an established prognostic marker, such as prostate specific antigen (PSA) in prostate cancer[5], this does not imply in itself that this biomarker would be a surrogate for survival, even though it has been shown that men with low PSA survive longer on average. Similarly a differential reduction in a prognostic marker because of a treatment does not imply that a clinical effect from this treatment is inevitable. It may be that the mechanism of the drug effect on this marker is downstream of the effect which is needed to activate response. Whilst biomarkers can be put to a wide variety of uses within pharmaceutical development, caution should be applied and the utility of a biomarker seen in the appropriate context.

1.4. Biomarker platforms

There are a large number of different technologies and sample types that can be used to generate biomarkers, ranging from imaging modalities such as CT, MRI or PET to molecular biomarkers measuring, for example, gene expression at the mRNA level, protein concentrations, SNPs or metabolites in sample types ranging from blood or saliva to cerebrospinal fluid or tumour biopsies. These different technologies may measure single markers, or, in the case of 'omic' type technologies, may measure many thousands of markers simultaneously.

Many variables may be accompanied by specific preprocessing methods (for example, relative expression in RT-PCR or the scaling and normalization of metabonomic data). Biomarkers may take the form of continuous measures, ordinal scores (for example in histopathology) or composites (as discussed in Section 4) and the context of use will inform the most suitable form of scoring. Unless this necessitates the identification of small number of biomarker-based categories it may otherwise be of interest to retain as much information as possible rather than simplifying a measure, as might be carried out in histopathology when choosing between the possible measures of staining area and intensity.

2. WHAT PREPARATIONS ARE NEEDED TO USE A BIOMARKER FOR INTERNAL DECISION MAKING?

To confidently utilize a biomarker for robust decision making, its value needs to have been assessed. Both the science behind the biomarker and the ability to use the biomarker in a practical clinical trial setting should be considered. Often, potential biomarkers, which have shown promise in a preclinical setting, may be evaluated in feasibility or methodology studies or via adding them on to existing trials or using archived samples.

2.1. Assessment of science

There can be a false assumption that a small study will be able to establish the scientific rational as well as estimate the treatment or population differences and the associated variability. Instead, this should be justified prior to running the study, either on the basis of prior experience or literature reviews. Information can be gained from careful meta-analyses using appropriate models[8]. However, for many novel biomarkers, there will be a limited number of papers relevant to the planned context of use in drug development and whatever evidence exists must be assessed to establish if the assumptions can be justified. Pre-clinical data may be valuable in building the package of evidence supporting the biomarker and many clinical markers made by translated from pre-clinical counterparts. Key opinion leaders who are promoting the use of the biomarker are also often used at this point, but care may be needed if views or published reports could potentially be biased[9, 10]. If the science cannot be established, then the use of the biomarker should be questioned.

2.2. Assessment of design

There are two main aspects of study design to assess when using a biomarker; the practical feasibility of running the study and the statistical characteristics of the marker. There are often feasibility issues that can affect the measurements or bias the results. For example,

- One measurement may compromise another, such as taking a blood sample just prior to assessing blood flow.
- In a time course assessment, having many measurements and assessments planned concurrently would require timings to be compromised, such as when measuring cognitive function in a first in human study at the same time as trying to accurately capture the pharmacokinetic profile of the compound.
- Being asked the same question repeatedly in a short time frame, for example in a pain challenge model, could induce false differences.
- If a scoring method is subjective, then measures should be taken to minimize bias, for example, with the use of scripted questions, blinding or standard operating procedures.

A small pilot study can be valuable in highlighting any practical issues, but the statistical aspects associated with the biomarker can best be assessed in a study specifically designed for this purpose. Clear objectives need to be set and these may be more wide-ranging than purely estimating a treatment effect. These goals may include the following:

- Comparing potential biomarkers or forms of endpoint;
- Assessing study designs, considering parallel group and crossover options;
- Estimating variability and effect sizes;
- Considering the analysis approach.

Ultimately, a common question is 'What sample size is required to run a biomarker study and have confidence that the right decision is made at the end?'

If there is a drug that can be used as a positive control, then a study could be set up as if it were being used to assess a new compound, so that the variability and effect size can then be estimated. However, this is often not possible, and so, it may be more appropriate to take other approaches such as to compare healthy volunteers with patients or use a challenge to evoke a change that a new drug could reverse. Functional magnetic resonance imaging (fMRI) can be used as an illustration. In the pain therapy area, gabapentin, taken for pain relief, has been used to assess the viability of fMRI in assessing future compounds [11], whereas pain challenge models in healthy volunteers, such as temporal summation [12], are also highly relevant. However, when studying sexual function disorders, it is more practical to compare healthy females with patients when assessing a new methodology such as fMRI [13]. For some situations, there is no simple way of assessing the desired difference in the study, and informed judgement, for example, on the basis of animal models, will need to be used to estimate a relevant difference. In such situations, methodology studies can still be key in estimating the variance of the biomarker [14].

When assessing the potential of the biomarker, the same design should be used as planned for the subsequent clinical trial. For example, if a four-period crossover design is planned, using laser Doppler imaging to assess blood perfusion, then this design should be studied rather than a simpler two-period crossover as dropouts or period effects, such as learning or anxiety effects, may be important. When comparing crossovers with parallel group studies, knowledge of inter and intra-subject variability is important to assess consistency across subjects and inform sample sizing. The use of any covariate, such as a baseline value, should be appropriately implemented to avoid any introduction of bias [15].

Once information is gained, either from including the biomarker in a specific study or another existing study, an assessment can be made as to whether the marker could be used in a clinical trial to make internal decisions with confidence. If data from one study alone has been used to obtain estimates of variance and desirable differences then this may carry a greater level of risk, compared with the use of a meta-analysis. The level of confidence needs to be fit for purpose, but all aspects of assay development should be considered, as has been well summarized by Lee *et al.* [16]. For early development, this may be less rigorous than for Phase III confirmatory studies, but this preparation should be sufficiently robust such that biomarkers can be confidently used to select those compounds with early promise from those with less value.

3. WHAT ARE THE CURRENT CHALLENGES OF BIOMARKER ANALYSIS?

Despite significant advancements in biological understanding and statistical methodology, inference from biomarker analysis is posed with many practical and statistical challenges that are fundamental to the experimental analysis. Sources of variability, missing data, bias and confounders should be considered.

3.1. Sources of variability, bias, and confounding

Unbiased estimation of the relationship between the biomarker and treatment effect depends on multiple factors and sources of variability may arise from the biospecimen or relate to the technology applied for quantitation of the marker [17, 18]. Lack of quality standards and routine quality assessments of biospecimens directly contribute to the measurement errors [19]. There should be adequate consideration on experimental design in identifying factors that may confound [20, 21] such as the following:

- Patient-related factors:
  - Individual factors such as genetic composition, race, age, and comorbidities;
  - Social and habitual factors, such as smoking (even when passive [22]), physical fitness and diet;
  - Biological variation within a single patient, such as diurnal variation;
- External factors:
  - Pre-analytical variation in sample collection and fixation practices;
  - Variation between laboratories, readers and batch runs;
  - Technical precision of the assay;

The levels of inter or intra-patient variability or the analytical (observer, batch or laboratory) variability in biomarker assessments can often be higher than the assay variability, leading to inconsistent results for the same biomarker across different biomarker studies [23, 24]. Studies reporting new methods in the absence of existing validated assays, such as high-dimensional gene signatures using reverse transcription PCR (RT-PCR)[25] or the use of in-house assays such as HercepTest^{TM}, can yield variable results, too [26, 27]. It can be difficult to determine if differences are meaningful or because of random error when there is this lack of a gold standard.

Analytical variability can be accounted for if complete information on the reader, batch and laboratory are available to the biomarker statistician, allowing the components of variation to be assessed. However, pre-analytical variation (from sample draw to assay) is harder to quantify and potentially confounds the analytical assay variability further [28]. These factors can lead to misclassification of biomarker levels resulting in false positive or false negative associations. The likelihood of these eventualities should be considered when powering trials using biomarker defined subgroups [29]. Therefore, explorations to understand variations arising from batch runs, sampling practices, natural diurnal variation or latent biological variability of a given biosample will be imperative to provide unbiased estimates. Robust statistical methods for analysis of biomarkers measured with batch/experiment-specific errors are provided by Long *et al.* [30].

In certain biomarker utility trial designs, different treatment options may be assigned to patients depending on their biomarker expression. In such a study, while treatment assignment may be randomized (for example, between an assignment that considers the biomarker and one that does not [31]), comparisons between the biomarker defined groups are a non-randomized comparison. Bias can arise, especially in open-label studies, for several reasons [32]:

- Verification bias because of the choice of locally available assessment methods [33]
- Patient selection bias, for example, patients with breast cancer family history, may not consent to BRCA DNA testing [34]
- Treatment allocation bias, if treatment assignment is based on a subjective assessment of the biomarker, as in some histo-pathological endpoints

Blinding may also be difficult in such a setting, and so, any co-interventions applied differently to study groups will also have an effect on the overall outcome. Ability to deal with such bias

is limited to the amount of information available to the statistician. Another practical consideration is the length of time over which sampling and analysis may take place. Signalling pathways are highly cross-networked and various synergisms, antagonisms and induced resistance can obscure the relationship between a marker and response to the drug [35]. The effect is pronounced in long term studies where the patients' molecular profile is subject to change, resistance mechanisms can develop, and the treatment response may no longer be the same as when the patients were first exposed to the treatment. The emergence of acquired EGFR resistance mutations and other molecular and histological aberrations on EGFR-TKI treatment is one example [36].

Retrospective analyses offer further challenges and are often not adequately powered if the trial was not designed with these objectives in mind. It is important for the statistician to communicate the implication of underpowered analysis and set realistic expectations with clinicians. Further, if the objectives are not prospectively defined then the validity of such analyses can be questioned. Uncertainties can exist in situations where samples have been collected retrospectively, but biomarker analyses are carried out prospectively according to defined objectives. Regulatory opinion on such activities has varied, as in the case of panitumumab [37]. Although predefined analyses are the ideal, such retrospective–prospective analyses may require further debate.

3.2. Sources of missing data

It is not uncommon to observe a greater proportion of missingness in biomarker data compared with clinical data. Missing data can arise because of many reasons, especially where biological sampling is concerned:

- Inaccessibility of tissues (as in lung cancer)
- Low consent rates for optional samples
- Lack of residual tissue in complete responders
- Poor quality of fixation in archival samples [28]
- Patient drop-outs due to non-response or toxicity

Such practical problems [38] result in the challenge of small evaluable sample sizes, where numbers can be too small for convincing validation of the modelling results in the full target population. The likely missing data mechanisms should be considered and appropriate assumptions made, especially if missing data imputation methods are adopted. For example, sample analysis errors may not be related to the biomarker value, but the willingness to consent to a sample may be related to a subject's wellbeing.

Unfortunately it is not uncommon to find that data quality issues can occur due to a lower priority being given to the handling of exploratory endpoints compared to recognized clinical ones. Data management practices should be considered to be just as important for biomarker data and the existence of data standards is highly beneficial in the course of a development program.

Clearly, there are many complexities for a statistician working with biomarkers and further challenges remain that could be the focus of future research (Table II).

4. HOW DO WE HANDLE LARGE NUMBERS OF BIOMARKERS?

Recent years have seen the development of a number of new high-dimensional technologies allowing the study of several thousand separate markers from a single biological sample. These include genetics, genomics, proteomics, and metabonomics, and more recently, Next Generation Sequencing of the entire human genome. Scientifically, these have allowed us to understand in detail at a molecular level the processes of disease and response to treatment and identify biomarkers that can identify or predict these changes. For example OncotypeDx® and MammaPrint® are commercially available tests based on 21 and 70 genes, respectively, for predicting the likelihood of recurrence of breast cancer.

4.1. Adjustments because of multiplicity of endpoints

When the purpose of an analysis is to identify those biomarkers of interest amongst a large number of potential markers, the major statistical concern is multiplicity. Given the number of genes or proteins often studied, we would expect to observe some strong correlations with outcome by chance alone, whilst traditional methods for addressing multiplicity, such as the Bonferroni correction, will be too conservative. The False Discovery Rate (FDR)

Table II. Future challenges.

i.	There is often little chance to observe biomarkers in patient populations prior to using them in early clinical development. What opportunities exist to reengineer the traditional clinical programme so that the learning phase around a biomarker may be supplemented, without delaying the overall clinical programme for the drug?
ii.	New markers of drug toxicity are emerging preclinically. What study designs can be used to translate these into man given that the sort of preclinical studies used to validate them are not feasible to conduct in humans?
iii.	Many technologies suffer from a 'batch effect' and show a correlation with sample quality. What pre-processing techniques can be developed to counteract these effects, so that we are focusing on measuring true biology rather than artifacts?
iv.	Next Generation Sequencing has moved the degree of dimensionality by several orders of magnitude and is now becoming affordable for regular use. How, both practically and analytically, can we cope with this deluge of data?
v.	Effective ways, possibly Bayesian, could be developed to build existing pathway knowledge, both in terms of the interrelationship of genes and their impact, into data analyses.
vi.	When identifying potential biomarkers for PHC, interaction analyses typically have low power, especially in the presence of many potential markers. How can such markers best be reliably identified?
vii.	As PHC becomes more common the processes for gaining approval of a drug, diagnostic and biomarker may become more burdensome. How can the demonstration of this evidence be streamlined?

[39,40] provides an alternative view by estimating the proportion of genes that are false positives. Permutation testing of a summary statistic can also be a powerful tool to generate a global test of the association between the set of markers being studied and outcome.

Significance Analysis of Microarrays (SAM) [41] utilizes data on many thousands of genes by adapting standard test statistics to incorporate an additional common parameter into the variability term within the test statistic for each gene, allowing information on variability to be shared across genes. This is especially valuable in situations with small sample numbers, where an individual test's variability may be poorly estimated on a limited number of degrees of freedom.

Filtering genes or proteins prior to analysis can remove a large number of false positives, whilst still leaving the majority of interesting results, reducing the false discovery rates for these genes of interest. Approaches to filtering can include a combination of simple statistical methods such as requiring a certain range or maximal level of expression; technology based methods measuring the quality of the measurements of each gene [42]; or incorporating biology by restricting the analysis to a set of candidate genes with a pathway link to the biology being studied.

As with univariate markers, the interpretation of results to identify genes of interest is often a combination of statistical significance and a scientifically relevant effect. The 2D FDR [43] provides a way of formalizing this, which has been further developed to a 3D FDR [44] by also incorporating a quality measure.

4.2. Composites and multivariate analyses

Combining large numbers of biomarkers can also be highly informative. The adage 'a picture is worth a thousand words' is as true with highly multidimensional data as with much simpler data, the challenge is that the pictures become harder to draw. As an exploratory analysis technique, Principal Components Analysis (PCA) offers a quick overall view of the main features within the data. This can aid in identifying any groupings, indicates covariates such as batch or reagent effects or other processing parameters that will need to be incorporated into the analysis, and gives a very good indication of the chances of observing a treatment effect from either univariate or multivariate modelling. Clustering can provide much of the same information, but needs to be interpreted with care, considering that it is based on the assumption that there are clusters within the data. Typically there is an underlying continuous population and clustering can often hide more subtle relationships between what come to be defined as clusters.

A wide variety of supervised multivariate predictive modeling techniques are available, including regression-based approaches such as Partial least Squares (PLS) [45], proximity-based methods such as Nearest Neighbours [46], tree-based methods such as Random Forests [47], distance-based approaches such as Support Vector Machines (SVM) [48] and many more.

As well as generating predictive multivariate models as 'black-boxes,' understanding the variables driving response within these models is also key. In practice, this may enable development of a simpler or more interpretable, but still equally predictive, model that may be easier to communicate, implement and work with in the future. For example, the OncotypeDx assay is expressed as a linear combination of six summary biological areas. For some methods, these can be naturally visualized through parameter estimates. Random forests implements a variable importance methodology that can generically be applied to multivariate models derived using any technique.

To address questions around best practice for predictive biomarker modeling, the FDA recently published the results of its Micro-Array Quality Control (MAQC II) initiative [49], which compared the predictive performance of models independently generated by a number of analysis teams on a common collection of data sets. Key conclusions included that although different approaches may generate highly distinct models, the predictive performance of these models may be remarkably similar irrespective of the differing methodologies applied. In fact, overall the major driver of model performance was found to be the experience and proficiency of the analysis team. MAQC II also highlighted that different clinical endpoints represent very different levels of classification difficulty, so that whilst some data sets readily generated highly predictive models, others appeared to have little or no predictive content despite being submitted to a battery of different approaches. MAQC-II pointed out that to provide clinical benefit it is not sufficient for a gene expression model to demonstrate some predictivity, it must demonstrate benefit beyond what would be possible using more readily available clinical variables alone.

Multiplicity can be a challenge throughout the entire discovery and development process. A number of candidate biomarkers may be studied within a single clinical trial. These markers may often be measuring the same biology, for example, different genes on a pathway known to be relevant to the drug's mode of action, or even measuring different modalities of the same gene, for example, DNA, mRNA, and protein, so can often be correlated with each other or confounded with other covariates. This makes careful planning of the analysis and interpretation of these results, including bringing in additional scientific information to strengthen the belief in any observed relationships, critical to be able to make robust conclusions that can be reproduced in future trials.

Although high dimensional data has required the development of new statistical approaches, classical statistical approaches are more important than ever, as discussed in Sections 2 and 3. Study design is critical, as is visualization, understanding data quality and its impact on subsequent analyses. These are typically more challenging than in a univariate situation where a simple plot may tell the story, but their complexity and the subtle influence they can have on conclusions make following these good statistical principles even more important.

5. WILL BIOMARKERS MAKE PERSONALIZED MEDICINE A REALITY?

5.1. Personalized health care examples and potential

Personalized health care offers the potential to identify patients more likely to derive benefit from treatment and as such is of great interest to Pharmaceutical companies, Regulatory Authorities and Health care Providers. The use of biomarkers is an integral part of the development of PHC and has experienced substantial growth in recent years [50]. Many companies are now routinely collecting biomarkers within their development programs. Aside from leading to safer and more effective treatments, the use of biomarkers may also help to reduce the numbers compounds failing in late stages of development by helping to explain unexpected variability in response to treatment.

The application of biomarkers to PHC has led to more informed prescribing practices for a number of drugs such as trastuzumab [51] (based on HER2), panitumumab [37] (EGFR), warfarin [52] (CYP2C9) and abacavir [53] (HLA-B*5701) amongst others. Biomarkers have also enabled the continued development of new drugs by facilitating the identification of a smaller population of patients [54]. However, the number of successes in PHC remains limited, highlighting the need for further technological, methodological and operational research. The development and application of new analysis methods and tools will enable more efficient identification of clinically relevant biomarkers, the design of confirmatory and exploratory biomarker studies and the informed strategy for biomarker integration across drug development programs.

The use of Biomarkers in PHC typically refers to predictive factors in determining the outcome of treatment, as discussed in Section 1, and these should be studied in a combined model that includes all subgroups to provide consistency with any overall models. This retains the ability to study treatment effects within each group. It should be recognized that biomarker-defined subgroups are generally not randomized comparisons and responses may relate to other clinical covariates that correlate with the biomarker. As such it is often of further interest to investigate the relationship between biomarkers and known clinical predictors.

5.2. Biomarker study designs

A number of study designs have been proposed for assessing the utility of prognostic and predictive biomarkers in PHC [55–57]. The optimal design depends on the level of prior knowledge relating to the biomarker effects and the proposed application. Statisticians are key in interpreting earlier phase findings on the prevalence of biomarkers and accuracy of assays and incorporating this knowledge into the design of confirmatory studies as the size and efficiency of designs can vary greatly [29, 58, 59]. Where biomarker effects are known or well understood, the purpose of the study is to confirm the effects and is typically prospectively designed.

The targeted or *enriched design* involves a prescreening step whereby patients are selected for the study based on biomarker status. Patients who test negative are excluded from the study whereas biomarker positive patients are randomized to one of the treatment groups. This can result in smaller studies when the effect of treatment is greater in the positive group, although a wide number of patients may still need to be screened.

The *stratification design* is less restrictive as all subjects are allocated to groups on the basis of biomarker status at screening, then randomized to treatment. This design gives more complete knowledge as information on the treatment effect in the marker negative group is collected, allowing the operating characteristics, such as sensitivity and specificity, of the biomarker to be estimated. Both of the aforementioned designs are practical when the study is required to test a single hypothesis relating to a known marker. However, it is common to have multiple objectives involving the evaluation of a treatment effect in the entire study population as well as in subpopulations. One approach for this type of study is based on an adaptive design [60]. Enrolled patients are randomized to treatment groups and the treatments are compared as part of a primary objective of the study. If the study fails to meet the primary objective (i.e., there is no difference in the treatments), then patients are subdivided into groups on the basis of biomarkers and a comparison of treatments is performed within these groups. As multiple statistical tests are performed in this approach, the overall false-positive error rate and needs to be controlled. Other adaptive approaches have been suggested where the sub-group has not been identified at the outset (adaptive signature design [61]) or where the appropriate cut-off for defining biomarker status has not been established (adaptive threshold design [62]).

Whilst much research has taken place on the design of confirmatory studies, most biomarker research remains exploratory for generating hypotheses that can be tested in future studies. However, the limited sample size of such studies and need to control the false positive detection rate when many potential markers are considered severely restricts the ability to detect biomarkers with moderate effects. One of the key challenges with regards to the use of biomarkers in PHC is the lack of power to identify new biomarkers in exploratory studies. Statisticians have a major role to play in optimizing exploratory research though the development and application of new analysis methods, study designs and the use of other data resources and in the detailed reporting of studies of predictive markers [26].

5.3. Diagnostic development

In many ways, the development and approval process for a diagnostic is similar to that of a new therapeutic agent, whereby the utility of the diagnostic must be demonstrated prior to approval. The statistician plays an important role in determining the performance characteristics of the diagnostic in the proposed target population such as the sensitivity, specificity, positive and negative predictive value, as well as capturing the conditions under which the diagnostic has been evaluated and developed. There are also a number of diagnostic-specific activities that must be undertaken [63], including the definition of the optimal cut-off for biomarkers on a continuous scale, and the evaluation of the repeatability and reproducibility of the biomarker assay. The statistician can ensure the validity of the entire development process by understanding and quantifying the factors that may impact the performance of the diagnostic. Where a companion diagnostic is being developed in conjunction with a drug and is a requirement for approval, good coordination of the codevelopment process is critical. As biomarkers may be identified during the course of a drug development program diagnostics development often lags therapeutic development and can lead to a delay in the drug approval unless codevelopment programmes are in place that meet the needs of both diagnostic and drug development.

6. WHAT CAN BIOMARKERS TELL US ABOUT DRUG TOXICITY?

Specific, sensitive and predictive safety biomarkers (SBMs) are required in pharmaceutical research and development to allow early detection of toxicity and the assessment of human risk. SBMs represent a crucial element of a comprehensive PHC approach to limit exposure of susceptible patients to drugs.

The strategy and approaches to the identification and qualification of biomarkers for safety is as follows:

1. Preclinical qualification of biomarkers, including the mechanistic understanding of the relationship between the biomarker and organ damage. Emphasis is toward markers that have a clear signal prior to damage becoming irreversible.

2. Translation of models and appropriate thresholds from the preclinical species to humans.
3. Clinical qualification of the biomarker in at-risk populations and in prospective studies covering the course of symptom development.

The role of safety biomarkers is to minimize potential risk and is the antithesis of efficacy biomarkers that are used to maximize potential benefit.

A lack of specific and sensitive mechanistic SBMs and the development of their respective assays for human samples are regularly delaying drug development programs. This is especially the case when a histo-pathological signal such as testicular toxicity is seen in preclinical toxicology studies, but cannot be adequately monitored in humans.

Many of the SBMs that have been used in clinical trials and in clinical practice for many years are markers of severe organ damage. For example, serum creatinine is a late marker of nephrotoxicity that does not reflect rapid changes in renal function and up to 2/3 of the nephron function must be lost before this marker shows significantly increased levels indicative of renal injury [64]. The correct assessment of kidney function is important both for dosage adjustment of renally excreted drugs and for early detection of drug nephrotoxicity. This is largely reversible if the offending agent is discontinued. Similarly bilirubin and the enzymes AST and ALT lack specificity as markers of liver function and cannot discriminate between transient effects and the development of fulminant liver disease [65, 66].

The realization that companies working in isolation could not deliver predictive SBMs led regulators, particularly the FDA and EMEA, to encourage the creation of enabling frameworks under which private industry could partner with regulatory authorities to advance the development and qualification of safety biomarkers for drug development [67–71].

6.1. Preclinical qualification

It is important during the biomarker evaluation process that stakeholders seek agreement on which critical experiments are needed to test that a biomarker meets specific performance claims; how new biomarkers and traditional comparators will be measured and how the resulting data will be merged, analyzed and interpreted [72]. This includes reaching consensus on the histopathological evaluations for assessing biomarker performance, the optimization of sample collection and preparation; and the implementation of blinded studies to minimize bias while performing histomorphological assessment.

The standard for the determination of organ (e.g., kidney) toxicity in the rat is the careful examination of organ toxicity by a qualified toxicological pathologist. Although highly accurate, this determination is not perfect, as pathologists cannot examine every possible section of an organ, molecular signals may precede the ability to observe structural damage, and some level of variability on an individual animal-by-animal basis between the subjective evaluations of pathologists is expected.

When assessing the relative performance of safety biomarkers statistical considerations may include the following [72]:

- Analyses should be based upon on a set of samples for which all the biomarkers being compared have been measured.
- Multiple supportive analyses may be appropriate based upon the level of confidence in injury determination. Performing an analysis non-selectively using all data is objective, but in addition an 'exclusion' subset where only samples with the highest level of confidence in the injury determination are considered (for example excluding animals treated with a known organ toxicant) may be valuable. Although experience suggests that generally both types of analyses yield similar comparative performance among biomarkers, a full data analysis can yield higher thresholds, and thus lower sensitivity, than exclusion analysis for the same level of specificity. As such the 'exclusion' analysis might be preferable for setting thresholds, thus also allowing for the detection of markers that signal before the onset of histopathological effects.
- If testing a hypothesis specified a priori then it is not necessary to adjust for multiple testing. However in instances where there is not this prespecification adjustments for multiple testing are necessary.
- Receiver operating characteristic (ROC) curves are useful to depict sensitivity and specificity across all possible decision rules. ROC curve area under curve (AUC) is a useful metric for the statistical assessment of relative performance as it is easily interpreted and allows for statistical significance tests indicating that one marker outperforms another.

6.2. Preclinical to clinical translation

Preclinical qualification is relatively straightforward and uses data generated using known organ toxicants. New compounds that cause a similar degree of biomarker change and associated pathology as known toxicants with insufficient safety margin will not progress to clinical studies. It is clearly not acceptable in the pursuit of safety biomarker qualification to place patients in clinical trials at additional risk.

The approach to assessing translation with the Animal Model Framework (AMF) project being conducted under the auspices of the ABPI [73] is to combine preclinical and clinical data to determine the sensitivity, specificity and predictive value of safety pharmacology core battery models assessing cardiovascular, central nervous system and respiratory effects. This collaborative effort is necessary because the majority of compounds that progress to man have negative signals in both the preclinical and clinical models. For example, only 15 of the 109 compounds that were assessed in the conscious dog telemetry model and evaluated in the single ascending dose first in man study had a clinically relevant change observed in QTc [74]. As such, when considering a threshold such as a 6% change in mean QTc relative to vehicle, confidence in the ability to judge the specificity (estimated at 80%) is much higher than that in the level of sensitivity (75%).

The primary objective of the AMF project is to quantify the properties of the current Safety Pharmacology models, with a secondary objective of refining the thresholds used to define a preclinical positive. In particular, the positive and negative predictive values associated with different thresholds are to be considered. Although sensitivity and specificity are properties of the test system, the negative and positive predictive values depend upon the prevalence of true clinical positives and negatives among the compounds being assessed in the preclinical model [75].

6.3. Clinical qualification

The clinical qualification of safety biomarkers in man is less straightforward because verification of organ toxicity in most circumstances is not possible. In this case, methods that can assess

the performance of SBMs in predicting organ toxicity in the absence of a gold standard are required [76, 77] and the use of Bayesian methods is likely [78, 79].

There are already public–private precompetitive partnerships established with an interest in searching, validating and qualifying safety biomarkers for use in predicting drug-induced organ injury (kidney, liver, and vascular) in the development of new medicines. Several examples are underway including the Predictive Safety Testing Consortium [67], as part of the Critical Path initiative and the SAFE-T consortia [70] under the Innovative Medicines Initiative.

Currently, laboratory parameters are used to indicate kidney injury (serum creatinine, blood urea nitrogen) and liver injury (AST, ALT), but the performance of these as safety biomarkers is poor. For example, by the time that serum creatinine indicates drug induced kidney injury, a high degree of loss of kidney function has already occurred [80, 81]. It is hoped that the work of the SAFE-T and PSTC group can help to validate and qualify new safety biomarkers for use in identifying unsafe drug candidates at an earlier stage.

Qualifying safety biomarkers will focus on evaluating a set of biomarkers in studies of drugs that are known to cause specific injuries. The objective will be to identify individual or panels of biomarkers that are more sensitive to drug induced injury without losing the specificity of the current clinical laboratory markers. It is also hoped to identify biomarkers that indicate injuries to specific areas and functions of each organ [82].

7. ARE THERE SITUATIONS WHERE BIOMARKERS REQUIRE FORMAL QUALIFICATION?

In certain situations, it may be desirable for biomarkers to have achieved some form of peer reviewed acceptance in order to demonstrate that the evidence collected is meaningful and reliable. Both the FDA (*qualification process for drug development tools*) [83] and EMA (*qualification of novel methodologies and biomarkers*) [84] have introduced guidance, at least in draft, for a form of biomarker qualification. These processes can result in a regulatory opinion on the acceptability of a biomarker for a particular use and can give advice on the path up to this point. Clearly the notion of demonstrating a good base of evidence for biomarker-based claims is sensible. However, as most biomarker usage is purely to aid understanding during the development program there are only certain circumstances where this formal process might be utilized, such as in the contexts of PHC or toxicity screening.

7.1. Is demonstrating surrogacy necessary?

The idea of a 'surrogate biomarker' and the required levels of evidence, has been much discussed [3] and was previously regarded as the aspiration for a qualified biomarker. This is often difficult or unrealistic [85, 86] and so objectives are now changing. To demonstrate that a biomarker could substitute for a clinical endpoint it is necessary to show that treatment effects on the biomarker are related to treatment effects on the recognized endpoint at a group, as well as an individual level [87]. As such a large number of randomized controlled trials where both endpoints are measured are required and this can be challenging to compile. This does not imply that the relevance of many biomarkers has not been demonstrated, just that this level of qualification is rarely necessary as biomarkers are seldom used as primary endpoints in confirmatory trials. The context of biomarker use may also not be one of broad application given that the endpoint may be mechanistically related to only a small class of compounds.

Nonetheless, there may still be reason to investigate the association between a biomarker and a relevant clinical endpoint. It can often be of interest to consider how the effect size seen in a small study with a short term biological endpoint may translate to a proposed future study, for example when moving from a tumour size-based endpoint to a survival based endpoint in oncology trials. An idea of this relationship will aid design and planning and allow calculations of the likely chance of success in the next phase of development [88, 89]. To develop this understanding some of the methods developed to demonstrate surrogacy can prove useful for the statistician to be aware of, except that the body of trials considered would likely be restricted purely to the drug class and specific disease indication of interest (cancer type in this instance), rather than seeking to show a very broad relationship.

Statistical methods for the demonstration of surrogacy have been well reviewed elsewhere[90, 91]. Methods have developed from the work of Prentice [92] and Freedman [93] to now focus on meta-analytic techniques [94–96], it having been recognized that early methods were restrictive. Buyse, Molenburghs, Burzykowski *et al.* [97–99] have developed meta-analytic techniques that can be utilized for the sort of informal comparisons described in the previous text.

Given the many challenges [100], successful examples of surrogates are rare and although HIV offers the examples of CD4 or RNA copy number [2, 101] many recognized surrogates such as blood pressure are endpoints that have been in use since before such qualification methods were suggested. Oncology has attracted attention for several meta-analyses [95, 102] as surrogates for overall survival could give great benefit, however the strength of relationship for endpoints such as progression-free-survival varies by class and indication, emphasizing the context-dependent nature of surrogate evaluation.

Glycated haemoglobin (HbA1c), which represents glycemic control over a period of time in diabetes patients, offers an illustration of some of the challenges. Although HbA1c is a recommended surrogate endpoint in diabetes [103], it could not demonstrate the cardiovascular risk related to rosiglitazone [104], and the drug was subsequently restricted. It should be recognized what a surrogate endpoint can accurately represent and safety outcomes data can still be paramount. Such cardiovascular event data is now often a requirement in many areas.

7.2. Qualification of other biomarker types

Although internal pharmacodynamic markers do not require formal assessment there are other circumstances where qualification could be required or encouraged. Biomarkers may still play a role in a confirmatory trial or drug label if they are used in a PHC capacity to identify those who benefit from a treatment. (Examples are listed on the FDA pharmacogenomic website [105], such CYP2C9 for defining poor metabolizers in the celecoxib and warfarin labels).

In many situations, the marker itself is inherently related to the assay used to measure it and new devices must be qualified. Recent FDA guidance [106] discusses simultaneous submissions of therapeutic products and *in vitro* companion diagnostics, but does not clarify the expectations on biomarker qualification in

this instance. This could be clarified, however the FDA biomarker process does appear to be aimed towards endpoints of wide applicability. Given that a predictive marker is likely to be related to drug mechanism this could not be said to fall into such a class of endpoints. In fact, sharing such information could jeopardize competitive advantage from the sponsor submitting the biomarker qualification application.

One area where these criteria of broad applicability and wide interest are met is for safety biomarkers. Clearly, a parameter that could predict toxicity issues, such as the liver toxicities discussed in Section 6, reliably or in advance would be of mutual interest to sponsors and regulators and, hence, these are good candidates for qualification processes. Given that the body of evidence is large and that the adverse events considered are often rare, precompetitive collaborations offer the best chance of compiling the materials required for this and the preclinical nephrotoxicity work such as that by ILSI/HESI offer a template [72, 107, 108].

Documents on such EMA consultations are publically available on the EMA website, where details of the first qualification opinion on a marker in humans have been released as a draft consultation [109]. This concerns cerebral spinal fluid markers to enable the selection of Alzheimer's disease patients for inclusion in clinical trials in Prodromal AD. Such use of a baseline marker to define disease characteristics is more conducive to the use of a qualification process compared with a predictive marker related to drug mechanism. The example also demonstrates how it is possible for a single sponsor to drive this process. Although examples of formally qualified biomarkers are rare and few have been discussed at length with regulators at this early stage, clearly the idea of evidentiary standards should be endorsed by all and fit-for-purpose validation should practiced whatever the situation. There may only be certain situations where this kind of formal qualification is necessary and further clarification from regulators of when it might be expected or encouraged could be beneficial. Guidance could also be extended to suggest what evidence should be submitted in addition to the procedures for doing so and how this would fit alongside other channels for receiving scientific advice.

CONCLUSIONS

Biomarker research is an area of critical importance within the pharmaceutical discovery and development process, but also an area of high uncertainty given the novelty of many of the applications and technologies used. As biomarker research is typically data-intensive, statisticians play a vital role in the development of new biomarkers and require a good working knowledge of the common issues, which they are well placed to understand and explain. As with any aspect of pharmaceutical discovery or development, the key concepts of statistical best practice should be maintained and high standards should be demanded, even in exploratory work. A consideration of first principles can be instructive and assist in tackling common issues. Keeping this in mind, we have provided a list of key points that statisticians should consider (Table III).

Much attention has focused on biomarker terminology and the potential of biomarkers as surrogate endpoints or to realize PHC. Collaboration can be vital in potentially achieving these aims. However key is that regardless of whether formal qualification or usage is sought for a biomarker, researchers should still follow the principles of fit for purpose validation, keeping the objectives in mind at all times. The ideas of understanding the characteristics of a potential endpoint and investigating subsequently observed data should remain familiar to statisticians in any context.

Table III. Key considerations for statisticians in biomarker development.

- Ensure the primary question of interest is well defined; even in exploratory research clear objectives are essential as it may not be safe to assume that the biomarker taxonomy is used in the same way by all. The future potential purposes that the marker will be put to should shape objectives and suggest appropriate study designs.
- Ensure biomarker studies have a reasonable chance of success through prospective planning, even where biomarkers are considered to be exploratory. With this in mind, it is useful to specify a meaningful biomarker effect size.
- Consult expert colleagues in order to get a good knowledge of likely distributions and pre- processing methods and to shape the choice of endpoint and scoring methodology prior to studies being run.
- Gather prior knowledge on sources of variability and causes of missing data when designing studies. Demand high standards of study conduct and sample and data management to avoid such scenarios and adjust modeling and missing data assumptions accordingly.
- As well as practical experience, estimates of variability and likely effect sizes should be gained via methodology studies, which should reflect the practice of intended formal use. Preclinical and translation sources are also valuable.
- Analyse a biomarker using the appropriate methodology for the given endpoint type and objective. As with other endpoints there is no universal approach and the characteristics of the endpoint should be considered.
- Particular attention should be paid to issues of multiplicity and model validation in situations where large numbers of candidate biomarkers are proposed.
- A skeptical and questioning point of view can be healthy to prevent results being interpreted out of context. Consider results in relation to the prior knowledge and restrict conclusions to sensible levels. (Manage expectations in terms of surrogacy for example).
- In exploratory studies where numerous candidate biomarkers are being evaluated, consider how to add confidence to the results though the use of other data resources, integration with biological information and cross-validation.

Although the application and technologies underlying biomarkers may be new, many of the issues, such as the effects of measurement error, missing data or confounders, can be addressed using principles and approaches familiar to statisticians experienced in data analysis within a pharmaceutical environment: understanding the questions and issues being asked, appropriate experimental or study design to address these questions, followed by robust and correct analysis of any data generated, taking into account any particular features of the data arising from the technology or the study and accurate communication of the results and the implications of any conclusions.

Acknowledgements

Sections were drafted by individual authors: P. Delmar & A. Herath on Section 1, T. Smart on Section 2, T. Sabin & J. Ratnayake on Section 3, C. Harbron on Section 4, A. Flynn on Section 5, P. Jarvis and J. Matcham on Section 6 and M. Jenkins on Section 7. All authors contributed to the reviews and finalization of the paper, which was coordinated and edited by M. Jenkins. The views expressed in this paper are those of the authors alone and should not be read as representing the position or views of our employers.

REFERENCES

[1] Atkinson AJ, *et al*. (Biomarkers Definitions Working Group). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics* 2001; **69**(3):89–95.

[2] Mildvan D, Landay A, De Gruttola V, Machado S, Kagan J. An approach to the validation of markers for the use in AIDS clinical trials. *Clinical Infectious Diseases* 1997; **24**(5):764–74.

[3] Lassere MN, *et al*. Definitions and validation Criteria for biomarkers and surrogate endpoints: development and testing of a quantitative hierarchical levels of evidence schema. *Journal of Rheumatology* 2007; **34**:607–615.

[4] Altar CA, *et al*. A prototypical process for creating evidentiary standards for biomarkers and diagnostics. *Clinical Pharmacology and Therapeutics* 2008; **83**(2):368–371.

[5] De Bono JS, Scher HI, Montgomery RB, Parker C, Miller MC, Tissing H, Doyle GV, Terstappen LWWM, Pienta KJ, Raghavan D. Circulating tumor cells predict survival benefit from treatment in metastatic castration-resistant prostate cancer. *Clinical Cancer Research* 2008; **14**(19):6302–6309.

[6] Van Cutsem E, *et al*. Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. *The New England Journal of Medicine* 2009; **360**:1408–1417.

[7] Oldenhuis CNAM, Oosting SF, Gietema JA, De Vries EGE. Prognostic versus Predictive value of Biomarkers in Oncology. *European Journal of Cancer* 2008; **44**:946–953.

[8] Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of Royal Statistical Society A* 2009; **172**:137–159.

[9] Kyzas PA, Denexa-Kyza D, Ioannidis JPA. Almost all articles on cancer prognostic markers report statistically significant results. *European Journal of Cancer* 2007; **43**:2559–2579.

[10] Ioannidis JPA, Panagiotouo OA. Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. *Journal of the American Medical Association* 2011; **305**(21):2200–2210.

[11] Iannetti GD, Zambreanu L, Wise RG, Buchanan TJ, Huggins JP, Smart TS, Vennart W, Tracey I. Pharmacological modulation of pain-related brain activity during normal and central sensitization states in humans. *PNAS (Proceedings of the National Academy of Science of the United States of America)* 2005; **102**:18195–18200.

[12] Arendt-Nielsen L, Frokjaer JB, Staahl C, Graven-Nielsen T, Huggins JP, Smart TS, Drewes AM. Effects of Gabapentin on Experimental Somatic Pain and Temporal Summation. *Molecular Pain* 2007; **32**:382–388.

[13] Arnow BA, Millheiser L, Garrett A, Lake Polan A, Glover GH, Hill KR, Lightbody A, Watson C, Banner L, Smart T, Buchanan T, Desmond JE. Women with hypoactive sexual desire disorder compared to normal females: A functional magnetic resonance imaging study. *Neuroscience* 2009; **158**:484–502.

[14] Keegan J, Horkaew P, Buchanan TJ, Smart TS, Yang GZ, Firmin DN. Intra- and interstudy reproducibility of coronary artery diameter measurements in magnetic resonance coronary angiography. *Journal of Magnetic Resonance Imaging* 2004; **20**: 160–166.

[15] Kenward MG, Roger JH. The use of baseline covariates in crossover studies. *Biostatistics* 2010; **11**:1–17.

[16] Lee JW, *et al*. Fit-for-purpose method development biomarker and validation for successful biomarker measurement. *Pharmaceutical Research* 2006; **23**(2):312–328.

[17] Moore HM, *et al*. Advancing Cancer Research Through Biospecimen Science. *Cancer Research* 2009; **69**(17):6770–6772.

[18] Sparrow J, Hickson S. Analyzing Laboratory Assay Data for Accuracy. *SAS Conference Proceedings*, Pittsburgh, PA, September, 14-17, 2008.

[19] Dancey JE, *et al*. Guidelines for the Development and Incorporation of Biomarker Studies in Early Clinical Trials of Novel Agents. *Clinical Cancer Research* 2010; **16**:1745.

[20] Freidlin B, McShane LM, Korn EL. Randomized clinical trials with biomarkers: design issues. *Journal of the National Cancer Institute* 2010; **102**:152–160.

[21] De Noo ME, Tollenaar RA, Deelder AM, Bouwman LH. Current status and prospects of clinical proteomics studies on detection of colorectal cancer: hopes and fears. *World Journal of Gastroenterology* 2006; **12**(41):6594–6601.

[22] Matt GE, Quintana PJ, Liles S, Hovell MF, Zakarian JM, Jacob 3rd P, Benowitz NL. Evaluation of urinary trans-3'-hydroxycotinine as a biomarker of children's environmental tobacco smoke exposure. *Biomarkers* 2006; **11**(6):507–523.

[23] Mattsson N, Blennow K, Zetterberg H. Inter-laboratory variation in cerebrospinal fluid biomarkers for Alzheimer's disease: united we stand, divided we fall. *Clinical Chemistry and Laboratory Medicine* 2010; **48**(5):603–607.

[24] Fenech M, *et al*. Intra- and inter-laboratory variation in the scoring of micronuclei and nucleoplasmic bridges in binucleated human lymphocytes: Results of an international slide-scoring exercise by the HUMN project. *Mutation Research* 2003; **538**:185–186.

[25] Clark-Langone KM, Wu JY, Sangli C, Chen A, Snable JL, Nguyen A, Hackett JR, Baker J, Yothers G, Kim C, Cronin MT. Biomarker discovery for colon cancer using a 761 gene RT-PCR Assay. *BMC Genomics* 2007; **8**:279.

[26] Hinestrosa MC, Dickersin K, Klein P, Mayer M, Noss K, Slamon D, Sledge G, Visco FM. Shaping the future of biomarker research in breast cancer to ensure clinical relevance. *Nature Reviews Cancer* 2007; **7**:307–315.

[27] O'Leary TJ. Standardization in Immunohistochemistry. *Applied Immunohistochemistry & Molecular Morphology* 2001; **9**(1):3–8.

[28] Betsou F, ISBER Working Group on Biospecimen Science, *et al.* Human Biospecimen Research: Experimental Protocol and Quality Control Tools. *Cancer Epidemiology, Biomarkers & Prevention* 2009; **18**:1017–1025.

[29] Carroll KJ. Biomarkers in drug development: friend or foe? A personal reflection gained working within oncology. *Pharmaceutical Statistics* 2007; **6**(4):253–260.

[30] Long Q, *et al*. Robust statistical methods for analysis of biomarkers measured with batch/experiment-specific errors. *Statistics in Medicine* 2010; **29**:361–370.

[31] Cree IA, Kurbacher CM, Lamont A, Hindley AC, Love S. A prospective randomized controlled trial of tumour chemosensitivity assay directed chemotherapy versus physician's choice in patients with recurrent platinum-resistant ovarian cancer. *Anti-Cancer Drugs* 2007; **18**(9):1093–1101.

[32] Veneis P, McMichael AJ. Bias and confounding in molecular epidemiological studies: special considerations. *Carcinogenesis* 1998; **19**(12):2063–2067.

[33] Punglia RS, D'Amico AV, Catalona WJ, Roehl KA, Kuntz KM. Effect of verification bias on screening for prostate cancer by measurement of prostate-specific antigen. *The New England Journal of Medicine* 2003; **349**:335–342.
[34] Klaren HM, Van't Veer LJ, Leeuwen FE, Rookus MA. Potential for bias in studies on efficacy of prophylactic surgery for BRCA1 and BRCA2 mutation. *Journal of the National Cancer Institute* 2003; **95**(13): 941–947.
[35] Arpino G, Wiechmann L. Crosstalk between the estrogen receptor and the HER tyrosine kinase receptor family: molecular mechanism and clinical implications for endocrine therapy resistance. *Endocrine Reviews* 2008; **29**(2):217–233.
[36] Sequist LV, *et al*. Genotypic and histological evolution of lung cancers acquiring resistance to EGFR inhibitors. *Science Translational Medicine* 2011; **3**(75):75ra26.
[37] Amado R, Wolf M, Peeters M, Van Cutsem E, Siena S, Freeman D, Juan T, Sikorski R, Suggs S, Radinsky R, Patterson S, Chang D. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *Journal of Clinical Oncology* 2008; **26**(10):1626–1634.
[38] Hodgson D, Whittaker R, Herath A, Amakaye D, Clack G. Biomarkers in oncology drug development. *Molecular Oncology* 2009; **3**(1): 24–32.
[39] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 1995; **57**(1):289–300.
[40] Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* 2002; **64**:479–498.
[41] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS (Proceedings of the National Academy of Science of the United States of America)* 2001; **98**:5116–5121.
[42] Talloen W, Clevert DA, Hochreiter S, Amaratunga D, Bijnens L, Kass S, Gohlmann HWH. I/NI-calls for the exclusion of non-informative genes: A highly effective filtering tool for microarray data. *Bioinformatics* 2007; **23**:2897–2902.
[43] Ploner A, Calza S, Gusnanto A, Pawitan Y. Multidimensional local false discovery rate for microarray studies. *Bioinformatics* 2006; **22**(5):556–565.
[44] Harbron C. A flexible probe level approach to improving the quality and relevance of Affymetrix microarray data. Presentation at the *Non-Clinical Statistics Conference*, Leuven, Belgium, September 25th 2008.
[45] Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* 2001; **58**:109–130.
[46] Parry RM, Jones W, Stokes TH, Phan JH, Moffitt RA, Fang H, Shi L, Oberthuer A, Fischer M, Tong W, Wang MD. k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *The Pharmacogenomics Journal* 2010; **10**: 292–309.
[47] Breiman L. Random Forests. *Machine Learning* 2001; **45**(1):5–32.
[48] Cortes C, Vapnik V. Support-Vector Networks. *Machine Learning* 1995; **20**:273–297.
[49] Shi L. *et al*. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology* 2010; **28**(8): 827–838.
[50] Frueh F, Amur S, Mummaneni P, Epstein RS, Aubert RE, DeLuca TM, Verbrugge RR, Burckart GJ, Lesko LJ. Pharmacogenomic biomarker information in drug labels approved by the United States food and drug administration: prevalence of related drug use. *Pharmacotherapy* 2008; **28**(8):992–998.
[51] Vogel CL, Franco SX. Clinical experience with trastuzumab (Herceptin). *The Breast Journal* 2003; **9**(6):452–462.
[52] Tan GM, Wu E, Lam YY, Yan BP. Role of warfarin pharmacogenetic testing in clinical practice. *Pharmacogenomics* 2010; **11**(3): 439–448.
[53] Hughes AR, Mosteller M, Bansal AT, Davies K, Haneline SA, Lai EH, Nangle K, Scott T, Spreen WR, Warren LL, Roses AD, CNA30027, CNA30032 study teams. Association of genetic variations in HLA-B region with hypersensitivity to abacavir in some, but not all, populations. *Pharmacogenomics* 2004; **5**(2): 203–211.
[54] Mega JL, Close SL, Wiviott SD, Shen L, Hockett RD, Brandt JT, Walker JR, Antman EM, Macias W, Braunwald E, Sabatine MS. Cytochrome p-450 polymorphisms and response to clopidogrel. *The New England Journal of Medicine* 2009; **360**(4): 354–362.
[55] Friedlin B, McShane LM, Korn E. Randomized clinical trials with biomarkers: design issues. *Journal of the National Cancer Institute* 2010; **102**:152–160.
[56] Simon R. Advances in clinical trial designs for predictive biomarker discovery and validation. *Current Breast Cancer Reports* 2009; **1**: 216–221.
[57] Buyse M, Michiels S, Grothey A, Matheson A, De Gramont A. Integrating biomarkers in clinical trials. *Expert Review of Molecular Diagnostics* 2011; **11**(2):171–182.
[58] Hoering A, LeBlanc M, Crowley JJ. Randomized phase III clinical trial designs for targeted agents. *Clinical Cancer Research* 2008; **14**(14):4358–4367.
[59] George S. Statistical issues in translational cancer research. *Clinical Cancer Research* 2008; **14**(19):5954–5958.
[60] Wang S-J, O'Neill RT, Hung HMJ. Approaches to Evaluation of Treatment Effect in Randomised Clinical Trials with Genomic Subset. *Pharmaceutical Statistics* 2007; **6**(3):227–244.
[61] Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for Sensitive Patients. *Clinical Cancer Research* 2005; **11**(21):7872–7878.
[62] Jiang W, Freidlin B, Simon R. Biomarker adaptive threshold design: a Procedure for Evaluating Treatment with Possible Biomarker-Defined Subset Effect. *Journal of the National Cancer Institute* 2007; **99**(13):1036–1043.
[63] FDA. Drug diagnostics co-development concept paper (Draft, August 2005). Available at: http://www.fda.gov/downloads/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/UCM116689.pdf.
[64] Schetz M, Dasta J, Goldstein S, Golper T. Drug-induced acute kidney injury. *Current Opinion in Critical Care* 2005; **11**:555–565.
[65] Ozer JS, Chetty R, Kenna G, Palandra J, Zhang Y, Lanevschi A, Koppiker N, Souberbielle BE, Ramaiah SK. Enhancing the utility of alanine aminotransferase as a standard reference for drug-induced liver injury. *Regulatory Toxicology and Pharmacology* 2010; **56**: 237–246.
[66] FDA. Guidance for Industry: Drug-Induced Liver Injury: Premarketing Clinical Evaluation, 2009. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM174090.pdf.
[67] The Predictive Safety Testing Consortium. Critical Path Institute. Available at: http://www.c-path.org/pstc.cfm.
[68] European Medicines Agency. Evaluation of Medicines for Human Use Innovative Drug Development Approaches Final Report from the EMEA/CHMP-Think-Tank Group on Innovative Drug Development, 22 March 2007. Available at: http://www.emea.europa.eu/pdfs/human/itf/12731807en.pdf.
[69] Institute of Medicine (IOM). *Accelerating the Development of Biomarkers for Drug Safety: Workshop Summary*. the National Academic Press: Washington DC, 2009. Available at: http://www.nap.edu/catalog/12587.html.
[70] SAFE-T Consortium. Safer And Faster Evidence-based Translation. http://www.imi-safe-t.eu/.
[71] Mattes WB, Walker EG, Abadie E, Sistare FD, Vonderscher J, Woodcock J, Woosley RL. Research at the interface of industry, academia and regulatory science. *Nature Biotechnology* 2010; **28**:432–433.
[72] Sistare FD *et al*. Towards consensus practices to qualify safety biomarkers for use in early drug development. *Nature Biotechnology* 2010; **28**:446–454.
[73] Valentin JP, Bialecki R, Ewart L, Hammond T, Leishmann D, Lindggren S, Martinez V, Pollard C, Redfern W, Wallis R. A framework to assess the translation of safety pharmacology data to humans. *Journal of Pharmacological and Toxicological Methods* 2009; **60**: 152–158.
[74] Ewart L. Conscious Dog Cardiovascular Telemetry Predictive Value to Man as Defined by the Animal Model Framework. *Presentation at the Safety Pharmacology Society 11th Annual Meeting*, Innsbruck, Austria, September 19–22, 2011.

[75] Altman DG, Bland JM. Diagnostic tests 2: predictive values. *British Medical Journal* 1994; **309**:102.
[76] Jafarzadeha SR, Johnson WO, Utts JM, Gardner IA. Bayesian estimation of the receiver operating characteristic curve for a diagnostic test with a limit of detection in the absence of a gold standard. *Statistics in Medicine* 2010; **29**:2090–2106.
[77] Hsieh H-N, Su H-Y, Zhou X-H. Interval estimation for the difference in paired areas under the ROC curves in the absence of a gold standard test. *Statistics in Medicine* 2009; **28**:3108–3123.
[78] Joseph L, Gyorkos TW, Coupal L. Bayesian Estimation of Disease Prevalence and the Parameters of Diagnostic Tests in the Absence of a Gold Standard. *rican Journal of Epidemiology* 1995; **141**(3): 263–272.
[79] Choi Y-K, Johnson WO, Collins MT, Gardner IA. Bayesian Inference for Receiver Operating Characteristic Curves in the Absence of a Gold Standard. *Journal of Agricultural, Biological, and Environmental Statistics* 2006; **11**(2):210–229.
[80] Bonventre JV, Vaidya VS, Schmouder R, Feig P, Dieterle F. Next-generation biomarkers for detecting kidney toxicity. *Nature Biotechnology* 2010; **28**(5):436–440.
[81] Waikar SS, Betensky RA, Bonventre JV. Creatinine as the gold standard for kidney injury biomarker studies? *Nephrology Dialysis Transplantation* 2009; **24**:3263–3265.
[82] Matheis K *et al*. A generic operational strategy to qualify translational safety biomarkers. *Drug DiscoveryToday*. In Press.
[83] FDA. Guidance for Industry: Qualification Process for Drug Development Tools, October 2010. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM230597.pdf.
[84] Qualification of Novel methodologies for Drug Development: Guidance to applicants, Jan 2009. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2009/10/WC500004201.pdf.
[85] Fleming TR, DeMets DL. Surrogate endpoints in clinical trials: are we being misled? *Annals of Internal Medicine* 1996; **125**:605–613.
[86] Fleming T. Surrogate endpoints and FDA's accelerated approval process. *Health Affairs* 2005; **24**(1):67–78.
[87] Baker S, Kramer B. A perfect correlate does not a surrogate make. *BMC Medical Research Methodology* 2003; **3**:16.
[88] O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. *Pharmaceutical Statistics* 2005; **4**:187–201.
[89] Julious SA, Swank DJ. Moving statistics beyond the individual clinical trial: applying decision science to optimize a clinical development plan. *Pharmaceutical Statistics* 2005; **4**:37–46.
[90] Weir C, Walley R. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Statistics in Medicine* 2006; **25**: 183–203.
[91] Burzykowski T, Molenberghs G, Buyse M. *The evaluation of surrogate endpoints (statistics for biology and health)*. Springer: Berlin Heidelberg New York, 2005.
[92] Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* 1989; **8**:431–440.
[93] Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine* 1992; **11**:167–178.
[94] Hughes M, Daniels M. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* 1997; **16**:1965–1982.
[95] Shi Q, Sargent D. Meta-analysis for the evaluation of surrogate endpoints in cancer clinical trials. *International Journal of Clinical Oncology* 2009; **14**:102–111.
[96] Li Y, Taylor J. Predicting treatment effects using biomarker data in a meta-analysis of clinical trials. *Statistics in Medicine* 2010; **29**: 1875–1889.
[97] Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* 1998; **54**: 1014–1029.
[98] Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 2000; **1**(1):49–67.
[99] Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics* 2006; **5**:173–186.
[100] Buyse *et al*. Biomarkers and surrogate end points—the challenge of statistical validation. *Nature Reviews Clinical Oncology* 2010; **7**:309–317.
[101] Hughes *et al*. CD4 cell count as a surrogate endpoint in HIV clinical trials: a meta-analysis of studies of the AIDS Clinical Trials Group. *AIDS* 1998; **12**(14):1823–32.
[102] Duffy S, Treasure FP. Potential surrogate endpoints in cancer research—some considerations and examples. *Pharmaceutical Statistics* 2010; **10**(1):34–39.
[103] Nathan DM *et al*. International expert committee report on the role of the A1C assay in the diagnosis of diabetes. *Diabetes Care* 2009; **32**(7):1327–1334.
[104] Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *The New England Journal of Medicine* 2007; **356**:2457–2471.
[105] Table of pharmacogenomic biomarkers in drug labels. Available at: http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm.
[106] Draft guidance for industry and food and drug administration staff – In vitro companion diagnostic devices, July 2011. Available at: http://www.fda.gov/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm262292.htm.
[107] Warnock D, Peck C. A roadmap for biomarker qualification. *Nature Biotechnology* 2010; **28**(5):444–445.
[108] Dieterle F, Sistare F, Goodsaid F *et al*. Renal biomarker qualification submission: a dialog between the FDA-EMEA and Predictive Safety Testing Consortium. *Nature Biotechnology* 2010; **28**(5): 455–462.
[109] Qualification opinion of Alzheimer's disease novel methodologies/biomarkers for BMS-708163. Reference EMA/CHMP/SAWP/102001/2011. Draft consultation 2011. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2011/02/WC500102018.pdf.