Action following the discovery of a global association between the whole genome and adverse event risk in a clinical drug-development programme

John Whitehead[1], Patrick Kelly[2], Yinghui Zhou[3], Nigel Stallard[4], Helene Thygesen[1] and Clive Bowman[5]

[1]*Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics, Lancaster University, UK*
[2]*School of Public Health, The University of Sydney, Australia*
[3]*Section of Quantitative Biology and Applied Statistics, The University of Reading, UK*
[4]*Warwick Medical School, The University of Warwick, UK*
[5]*Genetics Research, GlaxoSmithKline*

Observation of adverse drug reactions during drug development can cause closure of the whole programme. However, if association between the genotype and the risk of an adverse event is discovered, then it might suffice to exclude patients of certain genotypes from future recruitment. Various sequential and non-sequential procedures are available to identify an association between the whole genome, or at least a portion of it, and the incidence of adverse events. In this paper we start with a suspected association between the genotype and the risk of an adverse event and suppose that the genetic subgroups with elevated risk can be identified. Our focus is determination of whether the patients identified as being at risk should be excluded from further studies of the drug. We propose using a utility function to determine the appropriate action, taking into account the relative costs of suffering an adverse reaction and of failing to alleviate the patient's disease. Two illustrative examples are presented, one comparing patients who suffer from an adverse event with contemporary patients who do not, and the other making use of a reference control group. We also illustrate two classification methods, LASSO and CART, for identifying patients are risk, but we stress that any appropriate classification method could be used in conjunction with the proposed utility function. Our emphasis is on determining the action to take rather than on providing definitive evidence of an association.

Keywords: Decision procedure; Pharmacogenetics; Pharmacovigilance; Safety monitoring; Utility

Short title: Association between the whole genome and adverse event risk

4 September 2008

1. INTRODUCTION

The incidence of adverse events (AEs) is carefully monitored throughout the development of any new drug. Occurrence of too many AEs of sufficient severity relative to the condition under treatment might lead to cessation of the whole programme of clinical evaluation. However, now that routine genotyping has become available, it is possible to investigate whether emerging AEs are associated with a particular genetic subgroup of patients, opening up the possibility of continuing the programme once that subgroup has been excluded.

In this paper we consider a drug development programme in which an association between the genotype and the risk of an AE is suspected. Suspicion might have been aroused in a variety of ways. An unexpectedly high rate of AEs might have led to a global comparison of the genomes of the patients who have had an adverse event with those of all or some of the patients who have not had an adverse event. A number of genomewide association methods have been proposed [1-4] which could be used. Alternatively, sequential testing might have been conducted from the outset as new adverse events arose, perhaps using the approach of [5].

Once an association between genotype and the risk of AEs is suspected within an ongoing drug development project, it is imperative that continuation of the programme of clinical research be reviewed. It is not appropriate at this stage to demand rigorous and definitive statistical proof of the association, nor necessary to establish a detailed scientific understanding of its nature. The pragmatic question is whether the research can continue to recruit from all, or from some, of the original patient population. The answer will depend on the relative costs of experiencing the AEs and of withholding the treatment, and on estimates (albeit rough) of the risks of AEs within various genetic groups of patients.

We shall refer to patients who suffer AEs in a drug development programme as “cases”, while patients who do not are “controls”. When the latter come from the same clinical studies as the cases then they will be referred to as “concurrent controls”. Sometimes it is more practical to compare cases with subjects from a reference population that has already been genotyped and whose data are readily available from an existing database: such patients will be referred to as “reference controls”.

It could be that all patients in the drug development programme are routinely genotyped. Then cases can be compared directly with concurrent controls and the probabilities of AEs within specific genetic groups can be estimated from the programme data alone. Alternatively, genetic data might be available (or acquired) from all of the cases, but from only a subset of the concurrent controls, or patients with AEs might be compared with reference controls. Subsequent analysis will then depend on the assumption that the controls used are otherwise comparable with the cases, and external information about the overall risk of an AE will be required. Given these two conditions, case-control techniques can be used to estimate the probabilities of AEs within genetic groups.

Sometimes, suspicion of an association between the genotype and the risk of an AE will be based on a test which controls the type I error rate. This is desirable, as the chances of unnecessarily excluding genetic groups from clinical research are then limited. An example of a test that controls type I error through monitoring of this association throughout the development programme, is given by [5]. However, unplanned analyses with uncontrolled error rates might become ethically necessary, and they will lead to the same imperative for a review of study inclusion criteria.

In this paper we propose a method for deciding which genetic subgroups, if any, should be excluded from further clinical study, and whether their exclusion is sufficient to allow continuation of drug development. The proposed method is based on utility functions reflecting the cost-benefit balance associated with excluding specific genetic groups. Once genetic subgroups that should be excluded have been identified, the new exclusion rule will apply to all new patients recruited to clinical studies, and their eligibility will have to be assessed prior to randomisation. A policy will also be needed for patients who fall into the exclusion group who have already been randomised to treatment. This will concern whether existing patients are to be retrospectively genotyped and whether any action is needed for patients who are now off treatment. This issue is important, and could affect the final analysis of ongoing trials, but it lies outside the remit of this paper.

In order to determine whether any genetic subgroup should be excluded, we must first identify the covariates that appear to be associated with higher risk. Any classification method could be used for this purpose. If only a limited number of covariates are being considered, then standard logistic regression analysis could be applied. However, if many potential covariates under consideration, such as hundreds of thousands of SNPs, then more sophisticated classification methods will be needed.

The methods described in this paper have been motivated by the availability of whole genome data, but they are applicable to a wide range of factors with potential for association with the risk of AEs. For other high dimensional assessments (genomics, transcriptomics, proteomics and so on) the situation would be similar to that described here. More conventional risk factors, such as age and gender, could be added to the list of covariates.

Section 2 is devoted to the development of the utility function for deciding whether to exclude patients, and this forms the core of the paper. Section 3 describes possible approaches for identifying genetic subgroups with differing AE risks which can be used in the application of the proposed utility function approach. Two options are considered, the LASSO method and CART, but these are by no means the only possibilities. These classification methods are used in the illustrations presented in Section 4, while Section 5 is a discussion of the methodology and its implications.

As mentioned above, we are not concerned with making scientifically rigorous and definitive claims about associations between the genome and AEs. That is an important but different problem, [6-7]. In the context of ongoing clinical research, if it appears that some genetic subgroups are at high risk of AEs, then it is essential that the options of excluding those patients are urgently explored. Possible conclusions will be to exclude no-one, to exclude everyone and thus terminate the drug development programme, or to exclude a specific subset.

2. EXCLUSION OF PATIENTS BASED ON A UTILITY FUNCTION

Suppose that an association between the genome and the risk of an AE is suspected, and that it is necessary to identify which patients should be excluded from further clinical studies. It may be that no patients are excluded or that all of them are. The exclusion criterion will be defined in terms of genotype, and S will denote the set of genotypes associated with those patients to be excluded. The form of S is determined to maximise a utility function, which will now be defined. Consequently, all *subsequent* patients in the programme will have to be genotyped prior to entry.

Let D be the event of "having the disease". More precisely, D denotes the event that a patient's condition is not alleviated in the way that would be expected if

the novel drug were administered and was efficacious. Reality is simplified by supposing that all patients can be classified as either experiencing event D or event $\bar{D}$ (not D). On the other hand, it will be allowed that both patients receiving the drug and those not receiving it may have event D. The event of having an AE (of a specific, pre-defined type) is denoted by A. In this section it will be supposed that the event A can occur for patients receiving the drug and for those not receiving it, with the probability of A depending both on whether or not patients receive the drug and on their genotype.

The cost of event D will be denoted by d and that of event A by a. These notional costs can be fixed subjectively, and what will turn out to be important is their ratio a/d: how serious the adverse event is relative to failure to control the disease. The value of this ratio should be determined from discussion with experts in the clinical area. It will be supposed that the cost of experiencing both D and A is d + a, although a more general, non-additive approach could easily be developed. We also assume that the events D and A occur independently, conditional on treatment.

Suppose that patients can be classified into k genetic groups with distinct risks of event A. Some methods for doing this will be described in Section 3. Consider a patient who lies in the genetic group g. In terms of the events D and A, the expected cost of administering the drug is d P(D| g, E) + a P(A| g, E), where E is event of administering the drug (E). The expected cost of withholding the drug is d P(D| g, $\bar{E}$) + a P(A| g, $\bar{E}$). The patient should be excluded if the former exceeds the latter. Let h(g) denote the probability that a patient in the drug development programme has genotype g. Then the utility, per patient, of excluding all patients in genetic group g for all g in the set S is

$$U = \sum_{g \in s} h(g) \;\; d\,P(D \mid g, E) + a\,P(A \mid g, E) - d\,P(D \mid g, \bar{E}) - a\,P(A \mid g, \bar{E}) \;\; .$$

Thus U is the difference, per patient, between the expected cost of administering the drug and the expected cost of withholding the drug for patients with genotypes in the set S. The set S should be chosen to maximise U, which can be achieved by excluding patients in genetic group g if $P(A \mid g, E) - P(A \mid g, \bar{E}) > d/a \quad P(D \mid g, \bar{E}) - P(D \mid g, E)$. If D depends on treatment, but not on genotype, then dependence on g can be removed from the right-hand side of this expression, and a patient in genetic group g will be excluded if

$$P(A \mid g, E) - P(A \mid g, \bar{E}) > \frac{d}{a} \; P(D \mid \bar{E}) - P(D \mid E) \; . \qquad (1)$$

The difference $P(D \mid \bar{E}) - P(D \mid E)$ represents the treatment effect. This might be estimated from existing data. Alternatively, if only limited data are available when the association between genotype and AE risk is first suspected, an anticipated treatment effect based on expert opinion or previous relevant data might have to be used instead.

The probabilities h(g) associated with each genetic group are estimated by the proportions of each genetic group within the combined population of cases and controls. When reference controls or just a sample of concurrent controls are used, then the validity of these estimates depends on the assumption of comparability between the controls and patients in the drug development programme. If genetic data are available from all cases and concurrent controls, then estimates of $P(A \mid g, E)$ and $P(A \mid g, \bar{E})$ can be obtained directly. If the genetic data are available from cases and a sample of concurrent controls or of reference controls, giving a population Π, then direct estimates can be obtained for $P(A \mid g, \Pi)$. The following approach, familiar from drawing inferences from matched case-control studies [8] can

then be used to estimate the P(A | g, E). Suppose that the genetic groups identified form the set $\{g_1,\ldots, g_k\}$. It is shown in Appendix 1 that

$$\frac{P\ A|\ g_i,E\ \ P\ \bar{A}|\ g_j,E}{P\ A|\ g_j,E\ \ P\ \bar{A}|\ g_i,E}=\frac{P\ A|\ g_i,\Pi\ \ P\ \bar{A}|\ g_j,\Pi}{P\ A|\ g_j,\Pi\ \ P\ \bar{A}|\ g_i,\Pi}. \qquad (2)$$

Now denote $P(A|\ g_j, E)$ by p_j, j = 1, ..., k, and denote the right-hand side of (2) by OR_{ij}. It follows that

$$\begin{aligned} p_1\ 1-p_2\ &=OR_{12}p_2\ 1-p_1\\ &\ \ \vdots\\ p_1\ 1-p_k\ &=OR_{1k}p_k\ 1-p_1 \end{aligned} \qquad (3)$$

and it is also true that

$$h\ g_1\ \ p_1+\ldots+h\ g_k\ \ p_k=P\ \ A|E\ . \qquad (4)$$

Now OR_{12}, ..., OR_{1k} will be known from the classification method, P(A | E) is known from external sources and $h(g_1)$, ..., $h(g_k)$ are estimated from the genome data. Thus, solving for p_2, …, p_k from (3) and substituting into (4), we obtain an equation in p_1 alone, that can be solved using a simple numerical search. The same method can be used with E replaced by $\bar{E}$ to estimate the $P(A\,|\,g,\bar{E})$. Note that if $P(A|g_j, \Pi) = 0$ or 1 then $p_j = 0$ or 1 respectively, and (3) has then to be solved for the remaining unknown probabilities. Of course, when reference controls are used, the unrepresentative mix of cases and controls will also have an effect on the estimation of the $h(g_i)$: it might be desirable to used a weighted average of separate estimates from cases and controls, weighted according to the external value of P(A | E).

A very simple case of this approach occurs when a single SNP is used to identify three genetic groups based on whether the genotype at that SNP is aa, Aa or AA. Expression (1) can then be used to identify whether patients from any of these three groups should be excluded. There may be a well known candidate SNP which

can be investigated in this way. A naive extension of this idea is to explore which groups of patients should be excluded on the basis of each of a number of SNPs, considered separately, one at a time, in this way. This will give one exclusion set for each SNP. Then any patient falling in *any* of these exclusion sets would be excluded. Although this might appear to be a sensible strategy, it is based on separate marginal analyses and is thus oblivious to the joint influences of the SNPs. The identification methods described in Section 3 overcome this limitation by taking into account the whole cross-classification of genetic groups, rather than just the marginal classifications.

3. METHODS FOR IDENTIFYING GENETIC GROUPS AND ESTIMATING THEIR ADVERSE EVENT RISKS

Computation of the utility defined above requires division of the population into a number of genetic subgroups that are homogeneous with regard to AE risk, while differing between one another in this respect. Also needed are the estimates of the AE risks for each genetic group. The problem to be overcome is the sheer number of potential prognostic factors arising from the genome scan, usually far more than there are patients. Many methods have been suggested for overcoming this difficulty [9], and our concern here does not lie with creating new methods nor reviewing those that already exist. Readers may choose their own favourite approach for this step. In this section we describe two methods that could be used, in preparation for the illustrations of Section 4. In common with most such procedures, neither make explicit use of potential dependence (linkage disequilibrium) between the covariates arising from different SNPs, but they do lead to models which are valid in the presence of such dependence. Throughout this section the term y_i takes the value 1 if

the i^{th} patient has suffered an AE and 0 otherwise, i = 1, 2, …, n and $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})'$ denotes the corresponding genotype data, where z_{ij} denotes the genotype for the i^{th} patient at locus j, taking a value of 0, 1, or 2 depending on whether the patient's genotype at that locus is aa, Aa or AA respectively. In a regression model to fit a dominant or recessive genetic model, z_{ij} will be replaced by x_{ij} taking just one of the values 0 or 1. In a general genetic model that has no restrictions, z_{ij} can be replaced by two independent variables $x_{i,j(a)}$ and $x_{i,j(b)}$ to define the genetic structure. In the examples of Section 4 $x_{i,j(a)}$ takes the value 0 for aa and 1 otherwise, while $x_{i,j(b)}$ takes the value 0 for aa or Aa and 1 otherwise.

3.1 The least absolute shrinkage and selection operator (LASSO)

The LASSO method was proposed for normally distributed data by Tibshirani [10] for estimation in linear models when the number of independent variables is large. The approach has been extended to the case in which the number of independent variables exceeds the number of observations [11, 12]. The method consists of maximising the likelihood, subject to a penalty term concerning the size of the fitted coefficients. The LASSO achieves accurate prediction by shrinking some coefficients and setting others to zero. A logistic regression version of the LASSO method was described by Genkin et al. [13] in which the linear model $\text{logit}\ p(\mathbf{x}_i) = \eta(\mathbf{x}_i) = \alpha + \boldsymbol{\beta}'\mathbf{x}_i$ is fitted, where $p(\mathbf{x}_i) = P(Y_i = 1 \mid \mathbf{x}_i)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of unknown coefficients. The logistic LASSO estimate of the coefficients $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ is the value of $(\alpha, \boldsymbol{\beta})$ maximising the penalized log-likelihood

$$S_\lambda(\alpha, \boldsymbol{\beta}) = \ell\ \alpha, \boldsymbol{\beta}\ - \lambda \sum_{j=1}^{p} \left|\beta_j\right|$$

where $\lambda \geq 0$ is called the *tuning parameter* and ℓ is the log-likelihood function:

$$\ell(\alpha,\beta) = \sum_{i=1}^{n} y_i \eta(\mathbf{x}_i) - \log\left[1+\exp\ \eta(\mathbf{x}_i)\ \right].$$

The higher the value of λ, the fewer the number of terms appearing in the model. For the general genetic model, it may be felt desirable to ensure that either both independent variables relating to a single locus, $x_{i,j(a)}$ and $x_{i,j(b)}$, are in the model or that neither is. Yuan and Lin [14] extended the LASSO method to incorporate this restriction, so that $(\hat{\alpha},\hat{\beta})$ is chosen to maximise

$$S_\lambda(\alpha,\beta) = \ell\ \alpha,\beta\ - \lambda\sum_{j=1}^{q}\ \beta_{j,a}^2 + \beta_{j,b}^2\ ^{1/2}$$

where q = p/2 (that is, there are q SNPs and p = 2q coefficients). It is debatable whether such pairs of independent variables need to be kept together and either option could be taken: in the example of Section 4.1 below, we have not made this restriction. The routine `penalized` [15] in the software R [16] has been used in this paper to implement the LASSO method.

3.2 Classification and Regression Trees

A Classification And Regression Tree (CART) is an alternative tool for uncovering structure in data whenever a single response is to be related to many explanatory variables [17-18]. In R, CART may be implemented using the `tree` routine [19]. In the setting of this paper, the i^{th} patient provides the binary response y_i and the vector of genetic variables $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})'$, which can either be assumed to be continuous or factor explanatory variables. Each patient is classified into one of k disjoint genetic groups, $\{g_1, ..., g_k\}$ on the basis of the values of some chosen subset of the genetic variables. The model is fitted using binary recursive partitioning, whereby the data are successively split along coordinate axes of the predictor variables. Each split is

referred to as a "node" of the tree, and the final groups are referred to as "terminal nodes". At each node, the split which maximises some measure of the difference between the response variables in the left and the right branches, is selected. Splitting continues until the terminal nodes either become homogeneous, that is to say the within-node deviance becomes too small to be worthwhile splitting further, or they contain too few observations (that is fewer than some pre-specified limit). Both of these criteria are set manually, and for the R `tree` routine, the default values are 0.01 and 10, respectively. The terminal nodes are called "leaves", while the initial node is called the "root". The resulting classification can be displayed in the form of a binary tree.

3.3 Missing Values

Missing values in the data might occur in the genotype data due to, for example, poor DNA quality, differential PCR amplification or uncertain genotype calling. The R routine `penalized` cannot be used when the explanatory variables have missing values. In order to prepare the dataset for use of the LASSO on the examples in Section 4, the K nearest neighbours (KNN) method of Troyanskaya et al. [20] was used for imputation. If the value of the x_{iv}, the v^{th} explanatory variable for the i^{th} patient is missing, then this method finds the K other patients, who do have values for this explanatory variable, who are most similar to the i^{th} patient with respect to the other explanatory variables. Similarity is assessed in terms of a distance measure: here the Euclidean distance, $d_{im} = \sqrt{\sum_j x_{ij} - x_{mj}}^2$ was adopted. An average of values of the x_{jv} for values of j corresponding to the K most similar patients is then used as an estimate for the missing value of x_{iv}. Note that imputed values may be non-integer and will then have no literal genetic interpretation.

For the CART approach, the handling of missing values is straightforward. Missing values can be classified as another level in the factor, or they can be ignored when splitting [19]. The latter approach is used for the example in Section 4.

4. EXAMPLES

The examples in this section are based on data extracted from two GlaxoSmithKline (GSK) trials of abacavir in the treatment of HIV infection (CNA30027 and CNA30033). They concern a subset of the patients who received the active drug and were genotyped and they have already been used in the illustration of reference [5]. The adverse event of interest is taken to be incidence of hypersensitivity. The data comprise genotype information from 856 loci for 593 controls and 530 cases. The trial scenario has been modified and simplified for illustrative purposes, and in the following subsections two extreme sampling situations are described in order to bring out the essential features of the approach. Because of these manoeuvres, conclusions drawn from the examples may not be relevant to the use of abacavir, and so we refer only to "active drug" in most of what follows.

According to expression (1), future patients should be excluded from clinical studies if $P(A \mid g, E) - P(A \mid g, \bar{E}) > d/a \quad P(D \mid \bar{E}) - P(D \mid E)$. Hypersensitivity cannot occur amongst untreated patients, although it may be falsely diagnosed. For the purpose of the illustration in this section we suppose that no such diagnoses were found in the placebo group, or that such patients have been discounted in the analysis, so that $P(A \mid g, \bar{E}) = 0$. The methodology of Section 2 allows a more delicate analysis to be undertaken, in which $P(A \mid g, \bar{E})$ is also estimated. Apart from $P(A \mid g, E)$, the remaining terms in this expression will be determined from external sources. The values of $P(D \mid E)$ and $P(D \mid \bar{E})$ cannot usually be found from the clinical trial data, as

these are likely to be too limited and too short-term to support reliable estimates. Data from other studies might be used, or hypothesised values used in the trial design. Here we will rely on therapeutic response rates to abacavir in three trials (CNA30021, 30024 and CNA3005) specified on the US package inserts as of June 2007. These range from 49% to 69%. We shall take $P(D|\bar{E}) - P(D|E)$ = 0.5, assuming therapeutic response is rare in the placebo group. It will be assumed that a = 5d, indicating that incidence of hypersensitivity is five times as bad as failing to experience a therapeutic response. Thus patients should be excluded if $P(A \mid g, E) > 0.1$. In the example below, the classification methods have been implemented in R, using the default options for LASSO and CART, unless otherwise stated. In the first example below LASSO is used for the classifying of patients into genetic groups, while CART is used for the second example. This is done for illustrative purposes only, showing that any type of classification method can be used in conjunction with the utility approach; either classification method could have been used for both examples below.

4.1 Applying the LASSO in an illustration based on using all concurrent controls

For the first example, consider a prospective drug development program in which patients are randomised between active drug and placebo. All patients in the programme are routinely genotyped. At some stage, concern is raised about the rate of hypersensitivity occurring amongst patients receiving active treatment. The concern might have come from a formal monitoring procedure, from a Data and Safety Monitoring Board or from less formal considerations. However it arose, it has become necessary to consider exclusion of patients on a genetic basis. Suppose that concern is reached when 52 patients on active drug have suffered hypersensitivity and

593 have not. The overall probability of an adverse event is therefore 52/(52 + 593) = 0.081. To illustrate this scenario, all of the patients who did not suffer AEs in the available dataset are used to represent 593 concurrent controls and 52 cases are selected from the 530 patients who did suffer an AE.

The genetic groups and their associated AE risks will be found using the LASSO method in this illustration. The KNN imputation method was applied to the data before model fitting began. Then a general genetic model was fitted, with two dummy indicator variables being associated with each SNP. For this illustration, the tuning parameter was fixed to be 9 in this illustration, which identified four SNPs 336, 410, 825 and 847 to be influential. In each case only one of the associated dummy variables was selected (as explained in Section 3, we chose not to require the pair of variables at each locus to be included or excluded together). The following model was fitted:

$$\text{logit}\{P(A \mid g, E)\} = -2.691 + 2.166x_{336a} + 0.018x_{410a} + 0.014x_{825b} - 0.079x_{847a},$$

splitting patients into 16 genetic groups. As the 52 cases are being compared with 593 concurrent controls on the active drug, and all patients in the trial are being genotyped, the fitted model concerns the population E. Hence, fitted probabilities can be used directly in the computation of the utility of exclusion without recourse to equations (2) and (3). The fitted values of $P(A \mid g_i, E)$ and corresponding estimates of the $h(g_i)$ are listed in Table 1.

As explained at the beginning of this section, patients should be excluded if $P(A \mid g_i, E) > 0.1$, which means only patients with $x_{336a} = 0$, that is patients with allele of aa at locus 336, will be included. Applying the exclusion rule to the dataset used to create it, we find that there are 23 patients amongst the 52 cases and 14 patients amongst the 593 controls who would be excluded according to this strategy, and the

utility of the policy would be U = 0.0164a. This is the utility per patient, so that the total utility if (say) 1000 more patients were now treated in the clinical research programme would be 16.4a: this strategy would be equivalent to saving more than 16 per 1000 patients from suffering the AE. Evaluation of the exclusion rule can be more accurately assessed using cross-validation: we used leave-one-out cross-validation, leaving each of the 645 subjects out in turn, creating an exclusion rule as described above, and then classifing the remaining subject (the `cvl` function from the penalised package was used). The cross-validation estimated that 24 patients amongst the 52 cases (46% sensitivity) and 19 patients amongst the 593 controls (97% specificity) would have been excluded. Additionally, there were in fact 530 cases (patients with AEs) in the dataset, of whom we have used 52. Amongst the 478 cases who were not used to build the model, a further 174 would be excluded by the rule that has been developed.

Applying the exclusion policy will result in the inclusion of an estimated 93.9% of future patients (found by summing the h(g) over $g \notin S$), who would face an estimated risk of 0.062 of an AE (found by summing $h(g) \times P(A \mid g, E)$ over $g \notin S$ and dividing by the sum of h(g) over $g \notin S$), compared with an estimated risk of 0.081 without the exclusion policy. Thus by excluding 6.1% of patients, the risk of AE is predicted to fall by 23%.

4.2 Applying CART in an illustration based on reference controls

The second example concerns the situation described in [5] where, as each new case occurs, they are compared to a fixed group of reference controls according to a sequential procedure in which the overall type I error rate is controlled. Applying this method to the GSK data, the procedure is found to stop after observation of the 18th

case. Here, the 593 non-AE patients were used assuming these were reference controls and cases were taken randomly one at a time from the 530 AE patients. The remaining 512 AE-patients are used to see how effective the methods are at preventing would be "new" cases from receiving the drug.

Since we are assuming that the controls are fixed, the population Π (the 18 cases and 593 controls) is not representative of E in terms of AE risk. Consequently, the probability of an adverse event for a given genotype group, g_i, calculated from the 18 cases and 593 controls represents $P(A \mid g_i, \Pi)$, rather than $P(A \mid g_i, E)$, and so we need to use Equations (3) and (4) to calculate the $P(A \mid g_i, E)$ before applying the exclusion criteria. To do this we require estimates of genotype frequencies, $h(g_i)$, and an estimate of the overall adverse event rate, $P(A \mid E)$. The values of $h(g_i)$ can be calculated from any relevant source of information, for the purpose of this illustration we calculate it using a weighted estimate, adding $0.08 \times$ the proportion cases in group g_i to $0.92 \times$ the proportion of controls in group g_i. Such an average reflects the case-control mix likely to be found in the general patient population, rather than the mix within the genotyped subjects available for the analysis. The value for $P(A \mid E)$ would in practice be found from external information, or from the clinical patient data observed so far, using AE information on all patients regardless of whether they have been genotyped or not. Here, the value 0.08 is used for illustration as the overall probability of an adverse event. (US prescribing information for abacavir describes a hypersensitivity rate of 8% in nine clinical studies between November 1999 and February 2002.) A short-cut to the calculation that is available in this specific situation was used, and this is described in Appendix 2.

Table 2 shows the classification into genotype groups produced using CART. The $P(A \mid g_i, \Pi)$ are the estimated probabilities produced by CART, and the $P(A \mid g_i, E)$

are derived from them using equations (3) and (4). CART classifies patients into 11 groups. The $P(A|g_i, \Pi)$ values for groups g_1, g_2, g_3, g_5, g_6 and g_9 are equal to zero, consequently the corresponding $P(A|g_i, E)$ values are also equal to zero. All other $P(A|g_i, E)$ values exceed 0.1. Thus future patients in groups g_4, g_7, g_8, g_{10} and g_{11} should be excluded. The corresponding regression tree is shown in Figure 1. There are 17 patients amongst the 18 cases and 12 patients amongst the 593 fixed controls who would have been excluded according to this strategy. Applying leave-one-out cross-validation results in the exclusion of 4 patients amongst the 18 cases (22% sensitivity) and 44 patients amongst the 593 fixed controls (93% specificity) who would have been excluded.

Again we may use the remaining cases to indicate how the proposed method would perform amongst future patients who would suffer AEs if treated with the drug. Amongst the remaining 512 cases, 95 would be excluded. Summing the $h(g_i)$ values in Table 2 over the genetic groups that would be excluded shows that application of an estimated 9.9% of all future patients would be excluded.

5. DISCUSSION

This paper describes a pragmatic approach to the situation in which a suspicion of an association between the genome and the risk of an AE has arisen. The future development of the drug will depend, not only on the existence of such an association, but also on its magnitude and on the relative costs of suffering the AE and of being denied the real or potential benefit of the drug. Throughout this paper, the context has been a clinical drug development programme, and the objective one of modifying the exclusion criteria for patients yet to be randomised in order to reduce the incidence of AEs and allow continuation of the programme. The possibility that such a

modification might lead to bias in the evaluation of the experimental drug has to be considered carefully. Problems might arise in meta-analyses of trials before and after the change, but as investigators should always learn from past trials in designing new ones such difficulties may be inherent in any meta-analysis. When the changes are instituted during the course of a trial, possibilities of bias must be addressed. It might be that the principal endpoint used to assess efficacy is independent of the incidence of the AEs being considered, or at least any association might be regarded as so tenuous that independence can safely be assumed. Otherwise, the possibility of bias has to be recognised. It might be possible to overcome the problem using the methods of adaptive design, but this issue lies outside the scope of this paper.

The emphasis of this paper is on the use of a utility function to determine the set of patients who should be excluded from further clinical studies, and probably from any future administration of the drug. In order to compute the utility it is necessary to identify genetic groups and to estimate the AE risks associated with them. For this step, a variety of methods are available: the LASSO and CART methods considered above being just two candidates.

The classification into genetic groups is taken as a pragmatic tool, and the estimation of the associated AE risks is clearly approximate. The approach here will not stand up in terms of accuracy and reliability to comparison with approaches that can be applied retrospectively to completed studies in which there is no acute ethical need to decide who should and should not be included. The risk estimates suffer from the problem that they are estimated from the same data that have been used to identify the genetic groups, leading to underestimation of risk following implementation of the exclusion policy. The shrinkage element of the LASSO approach goes some way to

allowing for this, and cross-validation can been used to estimate exclusion rates more accurately.

Both the LASSO and the CART approaches suffer from the problem of how to choose their respective settings. The tuning parameter for the LASSO method has to be fixed, and in the example this was done in an arbitrary way. One automatic criterion of choice is the K-fold cross-validation method described by [21, 22]. When applied to the example of Section 4.1, it results in the choice $\lambda = 7.74$ and the identification of 2^8 genetic groups! Estimation of h(g) and P(A| g, E) for so many groups is impractical given the limited data available. In CART, the arbitrary settings concern when to stop splitting the nodes of the tree [17]. For either approach, it might be worth exploring the combination of the cross-validation criterion with a penalty based on the number of genetic groups identified, but this lies outside the scope of the present paper. In this paper, settings have been chosen to yield a small and thus manageable set of genetic groups, and in practice this would appear to be a feasible approach. An alternative approach to classification would be to use random forests: a method which automatically incorporates internal validation [23, 24].

Detection of a genetic association with AEs will not necessarily lead to action. It may be that there is no genetic group for which the utility of exclusion is positive. This could be the case if the cost of not controlling the disease is judged to be greater than the cost of the AE, for example when the disease is late stage cancer and the AE is simply a mild headache. On the other hand, it may be that all genetic groups should be excluded: that is, the clinical development programme should be terminated. The advantage of the method presented here is that it provides rationally based quantitative input into discussions which might otherwise be based only on informal clinical judgement.

A naive method based on considering each SNP separately was described in Section 2. This was applied to the examples of Sections 4.1 and 4.2, and in each case it led to the exclusion of all future patients and thus closure of the drug development programme. Rather than basing decisions on expected utility, as has been done here, other criteria such as minimax [25] could be considered. The minimax criterion seeks a decision to minimise the maximum loss, or in the terminology used here, to maximise the minimum utility. For every policy, the minimum utility is $-\,d+a\,\sum_{g\in S}h\;g\;$, as the worst thing that might happen if denied treatment is that a patient might suffer both D and A, while the best thing that might happen if given treatment is to suffer neither D nor A. The minimax criterion would then lead to the exclusion of no-one.

The illustrations used in this paper are of necessity artificial. The method has yet to be implemented in practice. When it is, then the nature and completeness of data collection will be affected by the knowledge of their importance for safety monitoring and hopefully improved. There are countless methodological improvements that could be made to the theory underlying the approach, but more valuable in its development would be experience from its use in real studies.

Once a genetic association with AEs is suspected, and action to exclude certain genetic groups has been taken, the question arises of how further monitoring should proceed. Could the approach be used again later in the drug development programme to identify further genetic groups for exclusion? As each round of exclusions is of necessity based on limited data, should ways be found for readmitting excluded groups later in the process?

REFERENCES

1. Dudbridge F, Koeleman BPC. Rank truncated product of p-values, with application to genomewide association scans. *Genetic Epidemiology* 2003; 25: 360-366.

2. Wille A, Hoh J, Ott J. Sum statistics for the joint detection of multiple disease loci in case-control association studies with SNP markers. *Genetic Epidemiology* 2003; 25: 350-359.

3. Xiong M, Zhao J, Boerwinkle E. Generalized T^2 test for genome association studies. *American Journal of Human Genetics* 2002; 70: 1257-68.

4. Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated product method for combining p-values. *Genetic Epidemiology* 2002: 22:170-85.

5. Kelly P, Stallard N, Zhou Y, Whitehead J, Bowman C. Sequential genomewide association studies for monitoring adverse events in clinical evaluation of new drugs. *Statistics in Medicine* 2006; 25: 3081-3092.

6. Roses AD. Pharmacogenetics and Drug Development: The path to safer and more effective drugs. *Nature Reviews Genetics* 2004; 5: 643-655.

7. Roses AD. Genome-wide screening for drug discovery and companion diagnostics. *Expert Opinion in Drug Discovery* 2007; 2: 489-501.

8. Breslow NE, Day NE. *Statistical Methods in Cancer Research Volume 1 – the Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon, France, 1980.

9. Fielding AH. *Cluster and Classification Techniques for the Biosciences*. Cambridge University Press: Cambridge, 2007.

10. Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society Series B* 1996; 58**:** 267-288.

11. Efron B, Hastie T, Johnstone I, Tibshirani R. Least Angle Regression. *Annals of Statistics* 2004; 32: 407-451.

12. Park MY, Hastie T. L_1-regularisation path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B* 2007; 69: 659-677.

13. Genkin A, Lewis DD, Madigan D. Large-Scale Bayesian logistic regression for text categorization. *Technometrics* 2007; 49**:** 291-304.

14. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society Series B* 2006; 68: 49-67.

15. Goeman J. *The penalized package*. http://cran.r-project.org/doc/packages/penalized.pdf. 2007.

16. R Development core team. *The R Manuals.* http://cran.r-project.org/manuals.html 2008.

17. Breiman, L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.

18. Ripley BD. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.

19. Ripley B. *Tree: Classification and Regression Trees*. http://cran.r-project.org/doc/packages/tree.pdf, 2007.

20. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botsterin D and Altman R. Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001; 17**:** 520-525.

21. Graven P, Wahba G. Smoothing noisy data with spline functions. *Numerische Mathematik* 1979; 31**:** 377 - 403.

22. Hastie T, Tibshirani R Friedman J, *The Elements of Statistical Learning*. 2001: Springer-Verlag.

23. Izmirlian G. Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial. *Annals of the New York Academy of Sciences* 2004; 154–174.

24. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 2006; 7: 3.

25. Lindley DV. *Making Decisions*. London: Wiley, 1971.

APPENDIX 1: Indirect estimation of adverse event probabilities

Suppose that the cases and controls cannot be used to provide estimates of the risk of AEs. This might be because differential proportions of cases and controls are genotyped, or because the controls form a fixed group with genotype data available before the drug-development programme begins. In the latter case, the fixed controls will be assumed to be of the same genetic mix as the patients to be entered into clinical trials, but they will not have been exposed to the drug and so will not have had AEs.

Estimation of the $P\ A \mid g, E$ for the various genetic groups can then be achieved following an approach used in epidemiology for matched case-control studies. Consider two genotypes, g_1 and g_2. Suppose that the data from the cases and from the available controls are pooled. From this pool Π, we can estimate $P\ A \mid g_i, \Pi$, i = 1, 2, using LASSO, CART or some other approach. Now

$$P\ A \mid g_i, \Pi\ = \frac{P\ g_i \mid A, \Pi\ P\ A \mid \Pi}{P\ g_i \mid \Pi},$$

so that

$$\frac{P\ A \mid g_1, \Pi}{P\ A \mid g_2, \Pi} = \frac{P\ g_2 \mid \Pi\ P\ g_1 \mid A, \Pi}{P\ g_1 \mid \Pi\ P\ g_2 \mid A, \Pi}.$$

A similar equation holds for the event $\bar{A}$, and so

$$\frac{P\ A \mid g_1, \Pi\ P\ \bar{A} \mid g_2, \Pi}{P\ A \mid g_2, \Pi\ P\ \bar{A} \mid g_1, \Pi} = \frac{P\ g_1 \mid A, \Pi\ P\ g_2 \mid \bar{A}, \Pi}{P\ g_2 \mid A, \Pi\ P\ g_1 \mid \bar{A}, \Pi}.$$

The same arguments apply to the population E of patients exposed to the active drug, so that

$$\frac{P\ A \mid g_1, E\ P\ \bar{A} \mid g_2, E}{P\ A \mid g_2, E\ P\ \bar{A} \mid g_1, E} = \frac{P\ g_1 \mid A, E\ P\ g_2 \mid \bar{A}, E}{P\ g_2 \mid A, E\ P\ g_1 \mid \bar{A}, E}.$$

As we are assuming that the genetic make up of patients in Π is the same as that of those in E, it follows that

$$\frac{P\ A \mid g_1, E\ P\ \bar{A} \mid g_2, E}{P\ A \mid g_2, E\ P\ \bar{A} \mid g_1, E} = \frac{P\ A \mid g_1, \Pi\ P\ \bar{A} \mid g_2, \Pi}{P\ A \mid g_2, \Pi\ P\ \bar{A} \mid g_1, \Pi}.$$

This means that odds ratios estimated for the pooled population Π, will be valid for the population E. Probabilities will not, because the proportion of patients with A in the population Π is determined by the experimental design (in particular by the size of the chosen fixed group of controls, and does not reflect the risk of A).

APPENDIX 2: Indirect estimation of adverse event probabilities in the context of CART

Let the external estimate of the risk of an AE be $\pi = P(A \mid E)$. Suppose that, in the data available, there are $n_{ca,j}$ cases and $n_{co.j}$ controls within genetic group g_j, and that $N_{ca} = n_{ca,1} + \ldots + n_{ca,k}$ and $N_{co} = n_{co,1} + \ldots + n_{co,k}$. A direct estimate of the probability of an AE in group g_j would be $p'_j = n_{ca,j}/(n_{ca,j} + n_{co.j})$, and a weighted estimate of the probability that a patient lies in group g_j would be $h(g_j) = \pi n_{ca,j}/N_{ca} + (1 - \pi)n_{co.j}/N_{co}$. Put

$$p_j = qn_{ca,j}/(qn_{ca,j} + n_{co,j}),\ j = 1, \ldots, k, \tag{A1}$$

where $q = N_{co}\pi/\{N_{ca}(1 - \pi)\}$. As the direct odds ratios from CART are given by $OR_{ij} = p'_i\,(1 - p'_j)/\{p'_j\,(1 - p'_j)\}$, it is clear that the p_j satisfy equation (3). Furthermore

$$\begin{aligned} h\ \ g_j\ \ p_j &= \left\{\frac{\pi n_{ca,j}}{N_{ca}} + \frac{1-\pi\ \ n_{co,j}}{N_{co}}\right\}\left\{\frac{qn_{ca,j}}{qn_{ca,j} + n_{co,j}}\right\} \\ &= \left\{\frac{\pi N_{co} n_{ca,j} +\ 1-\pi\ \ N_{ca} n_{co,j}}{N_{ca} N_{co}}\right\}\left\{\frac{\pi N_{co} n_{ca,j}}{\pi N_{co} n_{ca,j} +\ 1-\pi\ \ N_{ca} n_{co,j}}\right\} \\ &= \frac{\pi n_{ca,j}}{N_{ca}}, \end{aligned}$$

so that the sum of these terms over j is equal to π, as required by equation (4). Thus, the required estimates p_j can be found by direct evaluation from (A1), rather than from simultaneous indirect solution of equations (3) and (4).

Table 1: Results from the application of LASSO to the data of Section 5.1

genetic group, g_i	specification of g_i				$h(g_i)$	$P(A\|g_i,E)$	Controls excluded (from 593)	Cases excluded (from 52)	Cases excluded (from 478)
	x_{336a}	x_{410a}	x_{825b}	x_{847a}					
g_1	0	0	0	0	0.118	0.064			
g_2	0	0	0	1	0.113	0.059			
g_3	0	0	1	0	0.138	0.064			
g_4	0	0	1	1	0.088	0.060			
g_5	0	1	0	0	0.113	0.065			
g_6	0	1	0	1	0.068	0.060			
g_7	0	1	1	0	0.195	0.065			
g_8	0	1	1	1	0.106	0.061			
g_9	1	0	0	0	0.012	0.372	3	4	28
g_{10}	1	0	0	1	0.010	0.353	5	2	19
g_{11}	1	0	1	0	0.008	0.375	2	3	32
g_{12}	1	0	1	1	0.002	0.357	1	0	27
g_{13}	1	1	0	0	0.007	0.376	0	4	14
g_{14}	1	1	0	1	0.003	0.358	2	0	19
g_{15}	1	1	1	0	0.012	0.379	0	7	20
g_{16}	1	1	1	1	0.007	0.361	1	3	15

$$U = a\{-(0.012 + 0.010 + 0.008 + 0.002 + 0.007 + 0.003 + 0.012 + 0.007) \times 0.1 + (0.012 \times 0.372 + 0.010 \times 0.353 + 0.008 \times 0.375 + 0.002 \times 0.357 + 0.007 \times 0.376 + 0.003 \times 0.358 + 0.012 \times 0.379 + 0.007 \times 0.361)\}$$
$$= 0.0164\ a.$$

In addition, imputation assigns the values $x_{336a} = 1$, $x_{410a} = 0.25$, $x_{825b} = 1$ and $x_{847a} = 0$ to one case, and this patient would therefore also be excluded.

Table 2: Results from the application of CART to the data of Section 5.2

Genetic group, g_i	specification of g_i	$h(g_i)$	$P(A \mid g_i, \Pi)$	$P(A \mid g_i, E)$	Controls excluded (from 593)	Cases excluded (from 18)	Cases excluded (from 512)
g_1	$x_{336}=0$; $x_{821}=2$; $x_{751}=0,2$	0.681	0.00	0.000	-		-
g_2	$x_{336}=0$; $x_{821}=2$; $x_{751}=1$; $x_{556}=0,1$	0.116	0.00	0.000	-		-
g_3	$x_{336}=0$; $x_{821}=2$; $x_{751}=1$; $x_{556}=2$; $x_{61}=0,2$	0.023	0.00	0.000	-		-
g_4	$x_{336}=0$; $x_{821}=2$; $x_{751}=1$; $x_{556}=2$; $x_{61}=1$	0.017	0.60	0.811	2	3	5
g_5	$x_{336}=0$; $x_{821}=0,1$; $x_{437}=0,2$	0.049	0.00	0.000	-		-
g_6	$x_{336}=0$; $x_{821}=0,1$; $x_{437}=1$; $x_{568}=0,1$; $x_{809}=0$	0.016	0.00	0.000	-		-
g_7	$x_{336}=0$; $x_{821}=0,1$; $x_{437}=1$; $x_{568}=0,1$; $x_{809}=1$;	0.011	0.20	0.417	4	1	2
g_8	$x_{336}=0$; $x_{821}=0,1$; $x_{437}=1$; $x_{568}=2$	0.020	0.80	0.920	1	4	3
g_9	$x_{336}=1,2$; $x_{64}=1$	0.016	0.00	0.000	-		-
g_{10}	$x_{336}=1,2$; $x_{64}=0,2$; $x_{169}=1,2$	0.018	0.29	0.534	5	2	40
g_{11}	$x_{336}=1,2$; $x_{64}=0,2$; $x_{169}=0$	0.033	1.00	1.000	0	7	45

$$U = a\{-(0.017 + 0.011 + 0.020 + 0.018 + 0.033) \times 0.1 + (0.017 \times 0.811 + 0.011 \times 0.417 + 0.020 \times 0.920 + 0.018 \times 0.534 + 0.033 \times 1.00)\}$$
$$= 0.0893a.$$

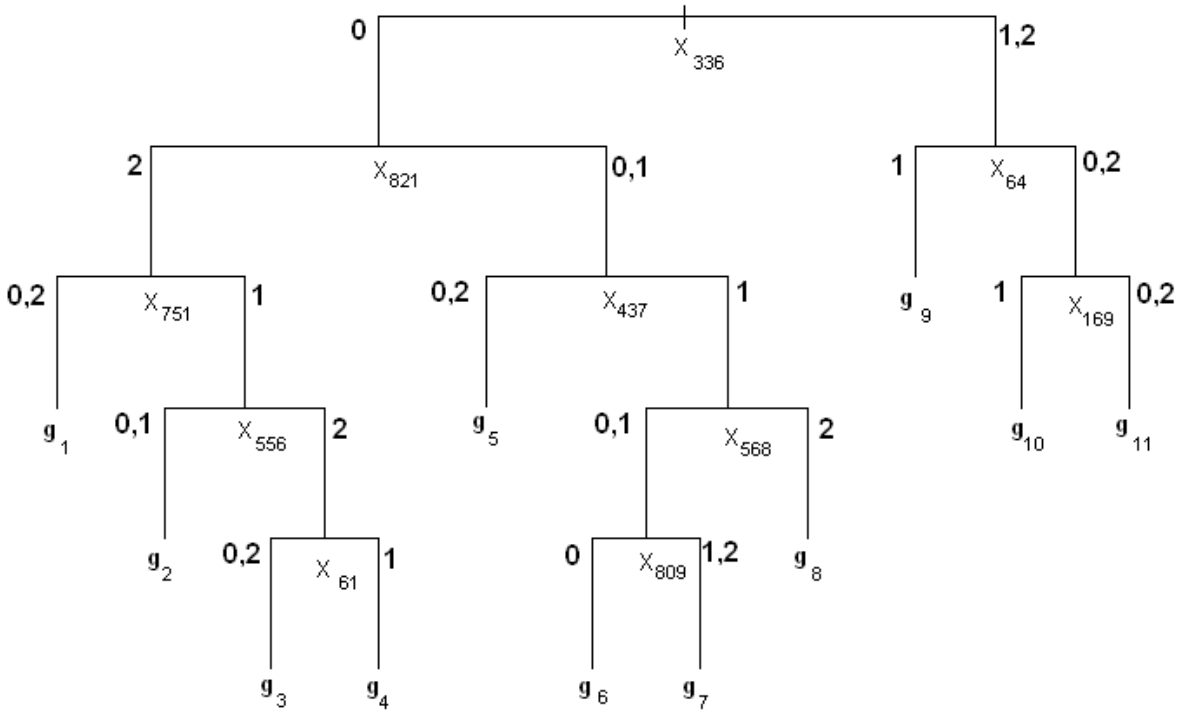

Figure 1: The tree diagram arising from the application of CART to the data of Section 5.2. The probability of being a case or control for the terminal nodes, g_i are given in Table 2.