Subject Section

Distinguishing prognostic and predictive biomarkers: An information theoretic approach

Konstantinos Sechidis\(^1\)*, Konstantinos Papangelou\(^1\), Paul D. Metcalfe\(^2\), David Svensson\(^2\), James Weatherall\(^2\) and Gavin Brown\(^1\)

\(^1\)School of Computer Science, University of Manchester, Manchester, M13 9PL, UK and
\(^2\)Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK.

*To whom correspondence should be addressed.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The identification of biomarkers to support decision-making is central to personalised medicine, in both clinical and research scenarios. The challenge can be seen in two halves: identifying predictive markers, which guide the development/use of tailored therapies; and identifying prognostic markers, which guide other aspects of care and clinical trial planning, i.e. prognostic markers can be considered as covariates for stratification. Mistakenly assuming a biomarker to be predictive, when it is in fact largely prognostic (and vice-versa) is highly undesirable, and can result in financial, ethical and personal consequences. We present a framework for data-driven ranking of biomarkers on their prognostic/predictive strength, using a novel information theoretic method. This approach provides a natural algebra to discuss and quantify the individual predictive and prognostic strength, in a self-consistent mathematical framework.

Results: Our contribution is a novel procedure, $INFO^+$, which naturally distinguishes the prognostic vs predictive role of each biomarker and handles higher order interactions. In a comprehensive empirical evaluation $INFO^+$ outperforms more complex methods, most notably when noise factors dominate, and biomarkers are likely to be falsely identified as predictive, when in fact they are just strongly prognostic. Furthermore, we show that our methods can be 1-3 orders of magnitude faster than competitors, making it useful for biomarker discovery in ‘big data’ scenarios. Finally, we apply our methods to identify predictive biomarkers on two real clinical trials, and introduce a new graphical representation that provides greater insight into the prognostic and predictive strength of each biomarker.

Availability: R implementations of the suggested methods are available at https://github.com/sechidis

Contact: konstantinos.sechidis@manchester.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

A prognostic biomarker is a clinical or biological characteristic that provides information on the likely patient health outcome (e.g. disease recurrence) irrespective of the treatment. On the other hand, a predictive biomarker indicates the likely benefit to the patient from the treatment, compared to their condition at baseline (Ruberg and Shen, 2015).

The difference between these two types of marker is clearly important, yet, surprisingly it is often not recognised. Ballman (2015) states that there ‘is considerable confusion about the distinction between a predictive biomarker and a prognostic biomarker.’ A specific example is highlighted by Clark (2008) when examining clinical biomarkers used routinely to make treatment decisions for non-small cell lung cancer, such as gender and histology – the key finding is that:

‘... gender and histology are actually prognostic, rather than predictive factors. Before biomarkers or clinical characteristics are included in guidelines for selecting patients for specific treatments, it is imperative that the prognostic effects of these factors are distinguished from their ability to predict a differential clinical benefit from the specific treatment.’

© The Author(s) 2018. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
In reality, biomarkers will almost always have some degree of prognostic value, and some degree of predictive value – but will also likely be dominated by one or the other. Using a biomarker for treatment assignments (i.e. in a predictive context), when in fact it provides mostly a prognostic signal, can have personal, financial, and ethical consequences — the inverse holds with different, though equally valid, consequences. A prognostic biomarker that is incorrectly labelled as predictive may result in overestimating the benefits of the treatment for a subset of the population and prescribing it to specific patients while in fact it should be available to all. Consequently, this may force the price of the drug up, as it is now considered as a treatment tailored to a specific portion of the population. The opposite applies if a predictive biomarker is incorrectly labelled as prognostic. In this case, the differential effect of the treatment to subsets of the population will be missed. The drug may be wrongly considered to have the same effect in all patients, affecting its price accordingly.

The challenge of finding markers with prognostic character is explored extensively in biostatistical and Machine Learning literature alike (Saey et al., 2007). On the other hand, discovery of predictive biomarkers has seen much less attention in Machine Learning, e.g. (Su et al., 2008). The biomedical literature on subgroup identification (Ondra et al., 2015) includes predictive biomarker ranking as an intermediate step, with SIDES (Lipkovich et al., 2011), Virtual Twins (Foster et al., 2011), and Interaction Trees (Su et al., 2009) as recent examples in this direction. However, little attention has been paid to the challenge of explicitly distinguishing between markers with mixed predictive/prognostic value. A statistical tool that could explicitly distinguish and quantify the predictiveness and prognostic value of a biomarker may be useful in study design and clinical interpretation of predictive models. In our work, we propose a unified approach that provides a language highly suited to biomarker discovery and related tasks around personalised medicine.

Our strategy is to align the challenges of data-driven biomarker discovery with that of information theoretic feature selection (Brown et al., 2012). The information theoretic viewpoint provides an ‘algebra’ with well-defined semantics, enabling one to discuss and quantify the predictiveness/prognostic value of a biomarker — each separately, but in a self-consistent mathematical framework. This also formalises the problem as finding the biomarkers that optimise a well-defined objective function, capturing the joint influence of patient characteristics and the treatment regime, on the outcome. Stated informally, this is:

\[ f(X, T) = \alpha_1 X_1 + \alpha_2 X_2 + (\beta_1 X_3 + \beta_2 X_2) T, \]

where \( \alpha_1 \) and \( \alpha_2 \) coefficients define the prognostic elements, and \( \beta_1 \) and \( \beta_2 \) define the predictive. When \( Y \) is continuous, \( f(X, T) \) can model the conditional mean \( E[Y|X, T] \), alternatively, if \( Y \) is binary, \( f(X, T) \) can be the logit of the conditional probability: logit\[P(Y = 1|X, T)]. If data is generated according to the model above, \( X_1 \) is regarded as solely prognostic as it directly influences the outcome \( Y \) — also known as a main effect. The variable \( X_3 \) is solely predictive as it influences the outcome only via an interaction with the treatment variable — also known as an interaction effect. However, note that variable \( X_2 \) has both predictive and prognostic strength.

Following standard notation (Lipkovich et al., 2017), assuming linear interactions, the outcome function can be written as:

\[ f(X, T) = h(X) + z(X) T, \]

where \( h(\cdot) \) and \( z(\cdot) \) are arbitrary functions of the covariates. We stress again that this is for the purposes of exposition, and all novel methods/results in the paper apply without such assumptions. There are two challenges here — firstly ranking variables in \( X \) by how much they influence \( Y \) via \( h(X) \) (i.e. main effects) — we refer to this as diagnostic ranking. And, secondly, ranking those that influence \( Y \) via \( z(X) \) (i.e. interaction effects) — we refer to this as predictive ranking.

Assuming a binary treatment \( T \in \{0, 1\} \), simple algebra yields that \( z \) is the change in outcome \( Y \) due to treatment, known as treatment effect:

\[ z(X) = f(X, 1) - f(X, 0). \]

The estimation of treatment effect (via first estimating \( f(X, \cdot) \)) motivates a set of methods that utilises the potential outcomes (or counterfactual) modelling (Rubin, 1974). For example, a popular method in this class is Virtual Twins (VT) (Foster et al., 2011), which non-parametrically estimates \( f(X, T) \) using random forests (RF) (Breiman, 2001). Predictive rankings can then be derived by ranking the biomarkers on their dependence with the estimated variable \( z(X) \). Another class of methods for predictive ranking is based on recursively partitioning data using interaction
tests between treatment and covariate – two popular methods here are SIDES (Lipkovich et al., 2011) and Interaction Trees (Su et al., 2009).1

We take a different approach, by reformulating the problem as an optimisation of an information theoretic objective. We introduce a new method for deriving predictive rankings, without the assumption of linear models, or binary \( T \) as above. Our method is directly applicable to multi-arm trials (i.e. more than two treatment groups) and captures higher-order biomarker interactions. In the following sections we introduce our framework.

### 2.2 A Natural Objective Function

Shannon (1948) in his seminal work “A Mathematical Theory of Communication” introduced information theory to quantify the amount of information and the capacity of the communication channel. One of the most fundamental concepts is mutual information. Through time, information theoretic approaches based on mutual information used to solve challenging problems in various research areas, e.g. Physics (Lloyd, 1989), Bioinformatics (Steuer et al., 2002) and Machine Learning (Zeng, 2015).

In simple terms, the mutual information \( I(X; Y) \) captures the extent to which two random variables \( X, Y \) depend on each other, or in other words the reduction of uncertainty in one variable \( Y \) given the values of the other \( X \). Mutual information has various interesting properties. For example, it can be associated with both upper and lower bounds on the Bayes error (Zhao et al., 2013). As a result, optimising information theoretic measures to solve challenging problems, i.e. feature selection (Brown et al., 2012), can lead to methods with competitive performance.

In clinical trial data, a natural way to select a set of biomarkers is to maximise the shared mutual information between the target \( Y \) and the joint random variable of the treatment \( T \) and the optimal feature set \( X^* \).2 Using the mutual information chain rule (Cover and Thomas, 2006), this objective can be decomposed in the following way:

\[
X^* = \arg \max_{X \in \mathbb{X}} I(T; Y) = \arg \max_{X \in \mathbb{X}} \left( I(X; T) + I(X; Y | T) \right)
\]

The first part captures the biomarkers with prognostic strength, while the second captures the biomarkers with predictive strength.3 By optimising these two parts independently we can derive two different objectives for the two different biomarker sets:

\[
X^*_{\text{pro}} = \arg \max_{X \in \mathbb{X}} I(X; Y), \quad (1) \quad X^*_{\text{Pred}} = \arg \max_{X \in \mathbb{X}} I(T; Y | X)
\]

Brown et al. (2012), in the context of feature selection, presented two heuristics for optimising eq. (1), which consider sequentially features one-by-one for adding or removal; the forward selection and the backward elimination respectively. For example, in forward selection at each step \( k \) we select the feature \( X^*_{k} \) that maximises the conditional mutual information (CMI):

\[
X^*_{k} = \arg \max_{X_k \in \mathbb{X}_k} I(T; | X_k; Y | X_{\bar{k}}) \quad (3)
\]

where \( X_{\bar{k}} \) is the set of the features already selected, and \( X_k \) the unselected. The CMI criterion can be used to rank the biomarkers on their prognostic strength and it is provably a hill-climber on the likelihood (Brown et al., 2012). In the following section we extend this result to derive predictive biomarker rankings.

#### 2.3 Information theoretic predictive rankings

In order to rank the biomarkers on their predictive strength, we should derive an optimisation procedure for the predictive part, eq. (2). A common heuristic approach to optimise an objective like this is to sequentially consider biomarkers one-by-one for adding or removal. This heuristic can be seen as a greedy iterative optimisation of the desired objective function. To present this procedure we will use a time-index notation for the feature sets, where \( X_{\bar{\tau}} \) represent the selected features at timestep \( \tau \), while \( X_{\bar{\tau}} \) the unselected ones.

**Definition 1** (Predictive biomarker forward selection step with CMI). The forward selection step adds the biomarker \( X^*_{k} \) which maximises the conditional mutual information between \( T \) and \( Y \) in the context of the variable of the currently selected set \( X_{\bar{\tau}} \) and \( X^*_{k} \). The operations performed are:

\[
X^*_{k} = \arg \max_{X_k \in \mathbb{X}_k} I(T; Y | X_{\bar{\tau}} X_k),
\]

\[
X_{\bar{\tau}+1} \leftarrow X_{\bar{\tau}} \cup X^*_{k},
\]

\[
X_{\bar{\tau}+1} \leftarrow X_{\bar{\tau}} \setminus X^*_{k}
\]

Using our formalisation of the problem and the results of Brown et al. (2012) the following theorem holds.

**Theorem 1.** The predictive forward selection heuristic adds the biomarker that causes the largest increase in the predictive part.

While for the backward elimination we have the following definition:

**Definition 2** (Predictive biomarker backward elimination step with CMI). The backward elimination step removes the biomarker \( X^*_{k} \) which minimises the conditional mutual information between \( T \) and \( Y \) in the context of the variable of the currently selected set \( X_{\bar{\tau}} \) without \( X^*_{k} \). The operations performed are:

\[
X^*_{k} = \arg \min_{X_k \in \mathbb{X}_k} I(T; Y | X_{\bar{\tau}} \setminus X_k)
\]

\[
X_{\bar{\tau}+1} \leftarrow X_{\bar{\tau}} \setminus X^*_{k}
\]

\[
X_{\bar{\tau}+1} \leftarrow X_{\bar{\tau}} \cup X^*_{k}
\]

Using the results of Brown et al. (2012) the following theorem holds.

**Theorem 2.** The predictive backward elimination heuristic removes the marker that causes the minimum possible decrease in the predictive part.

For simplicity from now on we will focus on the forward selection procedure, where at each step we select the feature not ranked so far \( X^*_{k} \in X_{\bar{\tau}} \) that maximises the following score:

\[
X^*_{k} = \arg \max_{X_k \in \mathbb{X}_k} I_{\text{Predict-CMI}}(X_k) = \arg \max_{X_k \in \mathbb{X}_k} I(T; Y | X_{\bar{\tau}} X_k)
\]

We will call this criterion as PRED-CMI, since it measures the predictive strength of each biomarker \( X_k \) by measuring the conditional mutual information between treatment and outcome given the joint variable between \( X_k \) and the currently ranked biomarkers \( X_{\bar{\tau}} \). In the following section we propose low-dimensional approximations, that will allow us to reliably estimate the conditional mutual information when the dimensionality of \( X_{\bar{\tau}} \) becomes prohibitively high.

#### 2.4 Lower-dimensional criteria for predictive rankings

In both scenarios, deriving prognostic rankings using CMI, and deriving predictive rankings using PRED-CMI, we need to tackle an important challenge: as the number of selected features grows, the dimension of \( X_{\bar{\tau}} \) also grows, and this makes our estimations less reliable. To overcome this problem low-dimensional criteria need to be derived.

For deriving prognostic rankings, the machine learning literature for feature selection is vast of low-order criteria. Brown et al. (2012) showed that a criterion that controls relevancy, captures feature interactions through redundancy and complementarity and provides a very good trade-off in terms of accuracy, stability and flexibility is the Joint Mutual

---

1. Section 1 of Supp. material provides details on all of the competing methods.
2. Using the same analysis as in (Brown et al., 2012) it can be proved that this criterion is equivalent to maximise the conditional likelihood of the outcome given the features and treatment, i.e. \( \log p(Y | XT) \).
3. Section 2 of Supplementary material provides some background on information theory, while Section 3 motivates why these two terms capture the prognostic and predictive strength by connecting the estimation of the mutual information with the deviance of the log-linear models.
Information (JMI) criterion (Yang and Moody, 1999): \( J_{\text{JMI}}(X_k) = \sum_{X_j \in X_g} I(X_k; Y | X_j) \).

Now we will tackle the challenge of deriving low-order criteria for the predictive rankings. The simplest way is to measure the conditional mutual information of \( T \) and \( Y \) given each biomarker independently without any regards of the others. This criterion can be seen as a univariate information theoretic way to derive predictive rankings, and from now on we will call it \( \text{INFO} \). The score that \( \text{INFO} \) uses to rank the biomarkers is:

\[
J_{\text{INFO}}(X_k) = I(T; Y | X_k).
\] (5)

While this is a low-dimensional criterion — we simply need to estimate the conditional mutual information between three variables — it fails to capture the dependencies between the biomarkers, i.e. collinearity/multicollinearity. We can illustrate these dependencies better by using the information theoretic identity \( I(A; B|C,D) = I(A; B|D) - I(C; B|AD) + I(C; B|AD) \) to re-write the PRED-CMI criterion as

\[
J_{\text{PRED-CMI}}(X_k) = I(T; Y | X_k) - I(X_{g}^k; Y | X_k) + I(X_{g}^k; Y | T X_k).
\] (6)

The first term captures the predictive strength of the biomarker \( X_k \) being considered for inclusion in \( X_g \). The second term capture the three-way interaction between the existing biomarker set \( X_{g} \), the target \( Y \) and the biomarker \( X_k \). Finally, third term captures the four-way interaction between \( X_{g} \), \( Y \), \( T \) and the biomarker \( X_k \). The last two are those that capture the complementarity between \( X_{g} \) and \( X_k \). We refer to these three terms as: the predictive relevancy, the predictive redundancy and the predictive complementarity of the biomarkers.

As we see, \( \text{INFO} \) is an approximation of PRED-CMI that captures only the first term, which measures the predictive strength of biomarker \( X_k \), but it fails to account for terms that capture the redundancy between the biomarkers. These, the predictive redundancy and predictive complementarity, are high dimensional functions of \( X_{g} - \) we can approximate them (to second-order interactions) with a sum as follows.

\[
J_{\text{INFO}^+}(X_k) = \sum_{X_j \in X_g} I(T; Y | X_j X_k).
\] (7)

We refer to this as \( \text{INFO}^+ \), and using the same identity as before, it can be re-written in the following ranking-equivalent form:

\[
J_{\text{INFO}^+}(X_k) \propto I(T; Y | X_k) \left[ 1 - \sum_{X_j \in X_g} \left( I(X_j Y | X_k) - I(X_j Y | T X_k) \right) \right],
\]

where \( |X_g| \) is the number of biomarkers already ranked. As we see from the last expression, \( \text{INFO}^+ \) captures all the three desirable terms for predictive biomarker ranking: predictive relevancy, predictive redundancy, and predictive complementarity. Furthermore, as the size of \( X_g \) grows, \( \text{INFO}^+ \) does not involve high dimensional conditional information estimations as the ones in PRED-CMI, eq. (6). Instead it averages over all possible lower dimensional (pairwise) terms, \( I(X_j Y | X_k) \) and \( I(X_j Y | T X_k) \) \( \forall X_j \in X_g \). By averaging over all possible pairwise terms, \( \text{INFO}^+ \) captures second-order biomarker interactions. In theory this could be extended to arbitrary higher order interactions, but data constraints will always limit this.

Algorithm 1 describes our approach for deriving predictive biomarker rankings. The provided algorithm is in a user-friendly form for illustrative purposes, but can easily be optimised to be 2-3 orders of magnitude faster than a direct translation. Note that in some scenarios i.e. subgroup identification (Foster et al., 2011), and especially in high-dimensional settings, we might be interested in only a few biomarkers (Lipkovich and Dmitrienko, 2014). To this end, in contrast to existing methods (i.e. SIDES/VT/IT) which rank all biomarkers, our \( \text{INFO}^+ \) forward step-wise procedure can return only the top-\( K \), without the need to rank all of them. This capability removes a significant computational burden (see Section 3.1.10).

As we already mentioned, our methods rank the biomarkers by estimating conditional mutual information quantities. To this end we can use any off-the-shelf estimator suggested in the literature\(^4\). Since in clinical trials we often encounter small-samples, in our implementation we used a shrinkage estimator suitable for ‘small \( n \), large \( p \)’ scenarios (Hauser and Strimmer, 2009). This approach can be extended to handle various types of covariates, i.e. categorical, continuous and mixed, and various types of outcomes, i.e. categorical, continuous and survival. Furthermore, when we have mixed type of data direct comparison of the mutual information values might be problematic. To overcome this issue, we use normalised versions of the conditional mutual information, which take into account the diverse characteristics of each covariate (Vinh et al., 2010).

**Algorithm 1 Greedy forward selection for \( \text{INFO}^+ \) ranking**

**Input**: Clinical trial data \( X, T, Y \) and size of the returned ranking \( K \)

**Output**: List of top-\( K \) predictive biomarkers \( X_g \)

1: \( X_g = X \) \( \triangleright \) Set of candidate biomarkers
2: Set \( X_g \) to empty list \( \triangleright \) List of selected biomarkers
3: for \( k := 1 \) to \( K \) do
4:   Let \( X_g^k \in X_g \) maximise \( J_{\text{INFO}^+}(X_k) = \sum_{X_j \in X_g} I(T; Y | X_j X_k) \)
5:   \( X_g^k \in X_g \) \( \triangleright \) Add biomarker \( X_g^k \) to the list
6: \( X_g = X_g \setminus X_g^k \) \( \triangleright \) Remove biomarker \( X_g^k \) from the candidate set
7: end for

### 3 Results

This section presents a comprehensive study in comparing our information theoretic methods with state-of-the-art approaches for biomarker rankings that capture their predictive strength. We will compare our methods (\( \text{INFO} \) and \( \text{INFO}^+ \)) against various methods, which can be discussed/characterised from the perspective of the statistical tools each is using: penalised linear regression methods (such as MCR (Tian et al., 2014)), counterfactual modelling methods (such as VT (Foster et al., 2011)) and recursive partitioning methods (such as SIDES (Lipkovich et al., 2011) and IT (Su et al., 2008))\(^5\).

#### 3.1 Experiments with simulated data

We simulate a large number of different scenarios and Section 3.1.1 presents all the necessary details of the simulation models. Section 3.1.2 presents the evaluation measures that we will use. Finally, Sections 3.1.3-3.1.10 explore empirically a series of interesting questions for the performance characteristics of the different methods.

#### 3.1.1 Simulation models

With our simulated models we capture a wide variety of different scenarios. For each model we simulate data with various size \( n \) and dimensionality \( p \). For each dataset we assumed equal allocations of patients to intervention and placebo arms, i.e. a clinical trial with 1:1 randomisation. Following Lipkovich et al. (2011) experimental setting, most of our models emulate the challenging scenario of ‘failed’ clinical trials, where the overall treatment effect in a population is nonexistent. We simulated data using different logistic regression models, categorised in three levels of difficulty: easy, medium and hard with the different functional forms \( f(X, T) = \log[p(Y = 1 | T, X)] \).\(^6\) Table 1 summarises all the above models in increasing challenge for identifying predictive biomarkers. In the experiments of the main paper we focused on categorical covariates, so in all scenarios, after we generated the data, each covariate was discretised in 2-5 levels using an equal-width strategy (Sec. 7 of Supplementary

\(^4\) Section 4 of the Supplementary material provides a short review and details on how to estimate conditional mutual information.

\(^5\) Section 5 of the Supplementary material provides the necessary details regarding the implementation of the competing methods.

\(^6\) Sec. 6 of Supp. material presents in detail the simulation models.
Table 1. Different scenarios of increasing challenge in identifying predictive biomarkers. **Fully-separate pred/prog biomarkers** is where there are no biomarkers with both predictive and prognostic strength, so a method cannot find a predictive biomarker by simply picking up on its prognostic nature. **Correlated covariates** creates situations where we might mistakenly pick up a noisy/prognostic biomarker, as it may be correlated to the predictive one for which we are searching. **Interaction terms** creates situations where two biomarkers interact to cause the outcome, which needs to be accounted for in the biomarker discovery algorithm. The presence of **Subgroups** creates situations where clearly defined groups of patients have enhanced treatment effect. More ticks equate to a more challenging scenarios.

<table>
<thead>
<tr>
<th>Fully separate pred/prog biomarkers?</th>
<th>Correlated biomarkers?</th>
<th>Interaction terms?</th>
<th>Subgroups?</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-1</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M-2</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>M-3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>M-4, M-5</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>M-6, M-7</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

material presents experimental results using continuous covariates). Furthermore, in order to have a better control over the effect of the prognostic (i.e. main-effects) and predictive part (i.e. interaction-effects), for each part we use the same functional form but with different variables. By following this approach we can control the relative strength of the predictive part using a coefficient \(\theta\).

### 3.1.2 Evaluation measures

One natural evaluation measure is to check how accurate are the different methods on correctly placing the predictive biomarkers in the top of the rankings. Let us define as \(X_{\text{Pred}}\) the set of biomarkers with predictive strength, with size \(q = |X_{\text{Pred}}|\). Also, \(\bar{X}_{\text{Pred}}\) the set of the top-\(q\) biomarkers returned by any of the methods that produce predictive rankings.

We define as true positive rate (TPR) the fraction of predictive biomarkers correctly ranked in the top-\(q\) positions of the list:

\[
\text{TPR} = \frac{|X_{\text{Pred}} \cap \bar{X}_{\text{Pred}}|}{q}
\]

Since our main objective is to introduce an information theoretic method for disentangling predictive and prognostic strength, it will be interesting to see how many times prognostic biomarkers are mistakenly placed in the top of the predictive rankings. Let us define as \(X_{\text{Prog}} \setminus X_{\text{Pred}}\) the set of markers with solely prognostic strength, while \(X_{\text{Irre}}\) are the irrelevant markers, i.e. the ones that they are neither prognostic nor predictive.

The false negative rate (FNR = 1 - TPR) can be decomposed in two terms:

\[
\text{FNR} = \frac{|X_{\text{Prog}} \setminus X_{\text{Pred}}| \cap \bar{X}_{\text{Pred}}|}{q} \quad + \quad \frac{|X_{\text{Irre}} \cap \bar{X}_{\text{Pred}}|}{q} = \text{FNR}_{\text{Prog}} + \text{FNR}_{\text{Irre}}
\]

TPR captures how accurate are the algorithms in correctly identifying the predictive biomarkers, while FNR\(_{\text{Prog}}\) how often they tend to select as predictive biomarkers those that contain only prognostic information.

We generate test data from the simulation models, and rank the biomarkers on their predictive strength using the methods presented above. In this synthetic situation, we have the ground truth set of biomarkers that have any degree of predictive strength: \(X_{\text{Pred}}\), and the same for prognostic markers \(X_{\text{Prog}}\). We use this to calculate the TPR and FNR\(_{\text{Prog}}\), for each ranking. Finally, we report the average results over multiple simulated datasets. For the ranking methods that use an estimate of the classification error to produce a variable importance score, such as VT, and in order to avoid overfitting, we use out-of-bag estimates (Foster et al., 2011).

### 3.1.3 Does markers’ prognostic strength affect the predictive ranking?

Before exploring in depth the performance of the different methods for deriving predictive biomarkers, the first thing we should explore is whether the ranking they produce is biased towards the prognostic strength of each biomarker. To answer this question we generate 200 datasets from the M-1 model with \(p = 30\) biomarkers and without any predictive information, i.e. setting \(\theta = 0\). Under this model, we have no predictive biomarker, five prognostic \(X_1, \ldots, X_5\), and the rest are irrelevant. To rank the biomarkers on their predictive strength we use three different methods (INFO+, VT, SIDES), and we derive the ranking score as follows: the most important marker takes score 30, the second most important 29 till the least important which takes score 1. Since there is no predictive biomarker, we expect that on average the score of each biomarker should be the same, \(\approx 15.5\).

Figure 1 shows that VT is biased towards the prognostic biomarkers, i.e. \(X_1, \ldots, X_5\), since on average, these biomarkers get higher score and they are on the top of the list. SIDES is also biased towards prognostic markers, but in smaller extent than VT. Our method, INFO+, is not biased towards the prognostic strength, since it produces equal scores for each biomarker. This also shows that ranking biomarkers on their conditional mutual information \(I(T; Y | X)\), captures the predictive strength, and not any prognostic information.
and it performs almost always worse than random selection. This highlights that VT is somewhat biased towards the biomarkers with strong prognostic effect. Furthermore, from Figure 2 we observe that the recursive partitioning methods (SIDES/IT) perform very similar in all scenarios, while our INFO+ method outperforms all of the rest in almost every setting, and it achieves a better trade-off between TPR/FNR

Remark 2: VT is biased towards predictive biomarkers that also carry prognostic information. INFO+ achieves better performance by disentangling the predictive and prognostic information of each biomarker.

3.1.5 What happens when we have correlated and interacted markers?

In this section we motivate the necessity of multivariate methods, such as INFO+, that capture higher-order biomarker interactions. We will compare INFO+ with two univariate approaches: our information theoretic INFO, and MCR, which, due to the linear modelling, does not capture higher order biomarker interactions. For the purpose of this section we will focus on three models M-2, M-3 and M-4 with diverse characteristics.

Model M-2 does not contain higher order interactions and the biomarkers are uncorrelated. In this case we expect that higher-order methods do not provide any actual benefit, and Figure 3 verifies it. All three methods have similar performance in terms of TPR, and this holds for various values of the predictive strength θ. However, this is not the case when we have correlated features (model M-3). In this case INFO+ outperforms the univariate methods, and this trend is even stronger when we also have interaction terms in the model (model M-4). In the latter scenario the univariate methods completely fail, even with strong predictive signals.

Remark 3: INFO+ captures interactions between biomarkers without the need to explicitly model the functional form of the predictive part.

3.1.6 What happens when we change the sample size?

In this section we compare the different methods in terms of their efficacy with the sample size. We focus on the medium difficulty model M-5 and we explore how the different methods perform as we vary the sample size. We simulate from small trials of n = 100 subjects, up to larger ones with n = 2000. Figure 4 presents our findings. Our method achieves higher TPR, increasing faster with n, and similarly shows a more rapid decrease in FNRProg, outperforming the competitors.

Remark 4: INFO+ is the most sample efficient method, i.e. it converges faster with the sample size.

3.1.7 What happens when we increase dimensionality?

Another interesting hypothesis to explore is how the above methods perform when we have a large number of covariates/biomarkers. To explore this we use the medium difficulty model M-6 and on Figure 5 we present how the different methods perform for various dimensionalities, p = \{50, 100, 200, 400\} covariates. As we see INFO+ consistently outperforms the other methods in terms of TPR, for both low and high dimensional trials, while it controls very well FNRProg.

Remark 5: INFO+ is the most efficient method in the presence of large number of noisy variables.
3.1.8 What happens when we have different size of subgroups?
As we already mentioned, an important usage of predictive biomarkers is to define subgroups of people with an enhanced treatment effect (Lipkovich et al., 2017). Defining these subgroups is crucial for personalised medicine, and in this section we will explore how the methods perform in the presence of such subgroups. We will focus on models M-6 and M-7, which have subgroups with diverse characteristics.

Figure 6 shows that when we have subgroups that are defined by a small number of biomarkers, such as two in M-6, our method achieves better TPR than the other two. This trend is more marked in medium to strong predictive signals (i.e. $\theta \geq 1$), while for weak signals all the methods perform similarly. The same holds for more complex scenarios, i.e. M-8, where the subgroup is defined by a three-variable interaction term.

![Comparison of TPR for models with different number of biomarkers](image)

**Fig. 6.** Comparing VT/SIDES/INFO+ for problems with subgroups with enhanced treatment effect. This is the average TPR over 200 simulated datasets for various values of the predictive strength $\theta$. The sample size is 2000 and the dimensionality $p = 30$ biomarkers.

**Remark 6:** INFO+ achieves competing performance in ranking biomarkers in the presence of subgroups with an enhanced treatment effect.

3.1.9 What happens in trials with a significant overall treatment effect?
So far our models (M-1–M-7) simulated scenarios of ‘failed’ clinical trials, where the treatment effect in the population is nonexistent, and there was a significant effect only within a small subgroup of the population. In this section we will focus on the performance of the algorithms when there is a significant effect of the treatment across the whole population, i.e. a successful trial. To emulate this scenario we use two medium difficulty models M-8 and M-9, with diverse characteristics: the first one has a common treatment effect for all the examples, while the second has a stochastic subject-specific treatment effect, generated independently of the covariates (Section 6 of Supplementary material provides more details).

Figure 7 presents how the different methods perform for various strengths of the predictive signal. Our method outperforms the other methods in terms of TPR, especially for medium and high predictive effects, while achieving lower FNR than the other two. This trend is more marked in medium to high-dimensional trials.

![Comparison of TPR for models with different predictive signals](image)

**Fig. 7.** Comparing VT/SIDES/INFO+ for models that simulate successful trials, where there is a treatment effect on the outcome independently of the covariates. This is the average TPR over 200 simulated datasets for various values of the predictive strength $\theta$. Small values of $\theta$ mean that the prognostic signal is stronger than the predictive, while the opposite holds for large values of $\theta$. For $\theta = 1$ both signals have the same strength. The sample size is 2000 and the dimensionality $p = 30$ biomarkers.

3.1.10 Which algorithm is more computationally efficient?
Lastly, it will be interesting to compare the performance of the methods in terms of their computational complexity. For this set of experiments we compare the average CPU time that each method needs to return the rankings, and see how it scales with the sample size and the dimensionality.

All the experiments were run on a PC with Intel ®Core(TM) i5-2400 CPU @ 3.10GHz and 8GB RAM, on a 64-bit Windows 7 OS.

Furthermore, we will use our optimised computational implementation of INFO+. In the user-friendly INFO+ implementation presented in Alg. 1, every time we select a marker we estimate from scratch the INFO+ score, or in other words we need to estimate $|X|_\theta$ conditional mutual information terms for each unselected biomarker (Alg. 1 Line 4). But we can optimise this process by storing the score of each unselected biomarker, and update it in every iteration. With this optimisation, instead of estimating $|X|_\theta$ terms for every unselected biomarker, we estimate just one.

Figure 8a shows that our optimised version of INFO+ outperforms all of the other methods for all sample sizes. Furthermore, by our step-wise forward selection we can improve the computational cost, by just asking to return the top-$K$ biomarkers instead of the full ranking. This is very useful in practice, where we have high-dimensional data where only few variables contain meaningful information. Figure 8b shows the execution time for various values of top- $K$ biomarkers, using our optimised version of INFO+. As we observe, we can save computational time by just returning the most important predictive biomarkers instead of ranking all of them. This result can be very useful in high dimensional trials.

![Comparison of execution time for INFO+ with different top-K](image)

**Fig. 8.** Comparing VT/SIDES/INFO+ in terms of their execution time. For INFO+ we have two versions, one that it returns a full ranking of all biomarkers, and one that it is just return the top-$K$ most important biomarkers. For all the experiments we simulated data from M-1 with predictive strength $\theta = 1$. For (a) we fixed dimensionality $p = 30$ and we simulate various sample sizes, while for (b) we fixed sample size $n = 2000$ and we simulate various dimensionalities.

**Remark 8:** Our optimised implementation of INFO+ is the most computationally efficient way to derive full rankings. Furthermore, by our forward step-wise procedure, INFO+ is suitable for exploring the ranking of the top-$K$ most influential biomarkers, something very useful for high-dimensional trials.
3.2 Applications to real clinical trials

Now we will present two applications of our methods in real clinical trials, and introduce a new graphical representation that provides more insight into the prognostic and predictive strength of each biomarker.

3.2.1 Lung cancer study – IPASS trial

It is of interest to explore how the suggested methods perform on a real clinical trial data, which has a known predictive biomarker. We explore the IPASS study (Mok et al., 2009): a Phase III, multi-center, randomized, open-label, parallel-group study comparing gefitinib (Iressa, AstraZeneca) with carboplatin (Paraplatin, Bristol-Myers Squibb) plus paclitaxel (Taxol, Bristol-Myers Squibb) as a first-line treatment for clinically selected patients from East Asia, who had advanced non small-cell lung cancer (NSCLC). The primary end point was progression-free survival (PFS).

It is known that gefitinib inhibits the epidermal growth factor receptor (EGFR), and is now indicated for the first-line treatment of patients with NSCLC whose tumours have specific EGFR mutations. Figure 9 presents the main finding of IPASS study (Mok et al., 2009): the presence in the tumor of a mutation of the EGFR gene is strongly predictive for better outcome with gefitinib. We therefore expect EGFR mutation status to appear as a strongly predictive biomarker. A detailed description of the trial can be found in Sec. 8 of the Supp. material.

On the full 1217 subjects, all three of VT/SIDES/INFO+, identify EGFR mutation status as the most predictive biomarker — however, an interesting question is to explore how they perform with minor perturbations in the data. Table 2 presents, for each method, the top-3 biomarkers with the highest score, averaged over 500 bootstrap samples along with required computational time. As we see, INFO+ is an order of magnitude faster than the competing methods. Of the biomarkers seen in Table 2, it is reassuring that INFO+ suggests EGFR mutation status to be the most predictive; as discussed above gefitinib inhibits EGFR, which was noted to have a significant interaction with the treatment indicator in the original study (Mok et al., 2009). Ethnicity is also related to the likelihood of EGFR mutation status; it is unsurprising that this has been pulled out by VT as a possible predictive biomarker, while our method, INFO+, manages to capture this interaction.

Table 2. Top-3 predictive biomarkers in IPASS for each competing method.

<table>
<thead>
<tr>
<th>Position</th>
<th>INFO+</th>
<th>VT</th>
<th>SIDES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>EGFR mut. status (X&lt;sub&gt;2&lt;/sub&gt;)</td>
<td>Weight (X&lt;sub&gt;2&lt;/sub&gt;)</td>
<td>WHO perf. status (X&lt;sub&gt;1&lt;/sub&gt;)</td>
</tr>
<tr>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>Weight (X&lt;sub&gt;2&lt;/sub&gt;)</td>
<td>EGFR mut. status (X&lt;sub&gt;2&lt;/sub&gt;)</td>
<td>EGFR mut. status (X&lt;sub&gt;2&lt;/sub&gt;)</td>
</tr>
<tr>
<td>3&lt;sup&gt;rd&lt;/sup&gt;</td>
<td>Age (X&lt;sub&gt;1&lt;/sub&gt;)</td>
<td>Ethnicity (X&lt;sub&gt;7&lt;/sub&gt;)</td>
<td>Bilirubin (X&lt;sub&gt;11&lt;/sub&gt;)</td>
</tr>
<tr>
<td>Time (sec)</td>
<td>0.97</td>
<td>18.52</td>
<td>6.76</td>
</tr>
</tbody>
</table>

Now we will present a visualisation tool, PP-graphs, that captures both the prognostic and predictive strength of biomarkers. It is our hope that this may provide useful information to healthcare professionals, in controlling false discoveries in clinical trials.

**Description of PP-graphs**: A PP-graph (Figure 10) is a scatter plot, where each point represents a biomarker, while coordinates (x, y) capture its prognostic and predictive strength respectively. For example, one way to measure prognostic strength is to rank the biomarkers, and then use a normalised score that takes values in [0, 1], where 1 is the score for the most-prognostic biomarker. To derive a prognostic ranking we can use the dataset \( \{x_i, y_i\}_{i=1}^n \) and any method that ranks biomarkers on their dependence with the output. For example, we can use any information theoretic method (Brown et al., 2012), such as MIM/JMI, or we can use RF and rank the biomarkers on their variable importance score. On the other hand, to derive predictive rankings we can use the dataset \( \{x_i, t_i, y_i\}_{i=1}^n \) and any method presented so far for deriving predictive rankings, such as INFO+ INFO++VT/SIDES/IT/MCR. Furthermore, we can capture the sample variations on the ranking scores by using a resampling methodology. For example, instead of estimating the scores only once from the whole dataset, we can average over the scores of a large number of bootstraps.

The red area (vertical shaded region) represents the top-K prognostic biomarkers, while the green (horizontal shaded region) the top-K predictive. For example for the PP-graphs of Figure 10 we used k = 1, which corresponds to the score cut-off value of \((p−k)/p = (23−1)/23 = 0.96\), where p = 23 is the total number of biomarkers in IPASS trial. The biomarkers being in the red and the green areas, are the ones that ranked, on average, in the first position of the prognostic and predictive ranking respectively. The intersection of these two areas – top right area – will contain the biomarkers that are both prognostic and predictive.

Fig. 9. Kaplan-Meier curves for the probability of progression-free survival (PFS) for: (a) overall population, where we see that the study met its primary objective and showed the superioriety of gefitinib as compared with carboplatin-paclitaxel for PFS (Hazard Ratio (HR) = 0.74, p < 0.001). Furthermore, EGFR mutation carries predictive information: (b) in the mutation positive subgroup patients treated with gefitinib had significantly longer PFS than the ones treated with carboplatin-paclitaxel (HR = 0.48, p < 0.001), while (c) in mutation negative subgroup, patients in carboplatin-paclitaxel group had longer PFS than the ones in gefitinib (HR = 2.82, p < 0.001). Note: as this is an unplanned analysis, all p values are nominal, and they have been used as descriptive measures of discrepancy and not as inferential tests of null hypotheses.
**PP-graphs for RF biomarker discovery in IPASS:** Figure 10a shows the PP-graph of RF based methods. For the prognostic axis we used RF to rank the biomarkers, while for the predictive axis VT, which is a counterfactual predictive method based on RF. Furthermore, we plot the average predictive/prognostic normalised ranking scores over 500 bootstrap samples of IPASS dataset. Using VT, the top four predictive biomarkers (y-axis) are X6, X2, X7 and X11 (Table 5 of Supplementary material provides the names of the biomarkers). However, we could ask the question whether these biomarkers are also prognostic, and, by using RF, we observe that X5, X11, X7 and X13 are the most prognostic biomarker (x-axis). Knowing the result from Section 3.1.4, that VT may be biased towards strongly prognostic biomarkers, we might now change our investigation: instead of pursuing X5 we should perhaps prioritise X2.

**PP-graphs for information theoretic biomarker discovery in IPASS:** Instead we can use the information theoretic methodology proposed in this paper, in which the predictive (y-axis) and prognostic (x-axis) rankings are derived from a self-consistent manner. For example, in Figure (b), the ranking in the y-axis is derived by using INFO+, while the ranking in the x-axis by using JMI (Sec. 2.4). Both of these approaches capture higher order interactions, by using low dimensional approximations. This PP-graph shows that our suggested INFO+ approach correctly ranks as the most important predictive biomarker X2 (green area).

### 3.2.2 Cardiovascular disease study – AURORA trial

Another interesting scenario to explore is how our methods perform in a trial where there is no known predictive biomarker. We explore the AURORA study (Fellström et al., 2009): a randomized, double-blind, placebo-controlled, multicenter trial in which 2776 patients with end-stage renal disease were randomly assigned 1:1 to double-blind treatment with rosvastatin at a dose of 10 mg or placebo. The primary endpoint was the time to a major cardiovascular event (MACE) defined as a nonfatal myocardial infarction, nonfatal stroke, or death from cardiovascular causes. All myocardial infarctions, strokes, and deaths were reviewed and adjudicated by a clinical end-point committee whose members were unaware of the randomized treatment assignments, in order to ensure consistency of the event diagnosis. A detailed description of the trial can be found in Sec. 9 of the Supp. material. For full details of the trial see (Fellström et al., 2009).

Figure 11a presents Kaplan-Meier curves of the cumulative incidence of the primary end point (MACE) in the overall population, where we see that the study failed to meet its primary objective: treatment with rosvastatin was not associated with a reduction in major adverse cardiac events (HR = 0.95, p =0.516). Furthermore, rosvastatin had no benefit in any examined subgroup, more details can be found in (Fellström et al., 2009). It is important to have a structured way to explore the data in such trials, in which any hypotheses arising out of the data may be handled in a controlled manner.

Table 3 presents, for each predictive biomarker discovery method VT/SIDES/INFO+, the top-3 biomarkers with the highest score, averaged over 500 bootstrap samples. None of the suggested variables have been previously identified as predictive, although Age has previously been identified as prognostic in a post hoc analysis (Schneider et al., 2013).

### Table 3. Top-3 predictive biomarkers in AURORA for each competing method.

<table>
<thead>
<tr>
<th>Position</th>
<th>INFO+</th>
<th>VT</th>
<th>SIDES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>Lymphocytes (X24)</td>
<td>Age (X1)</td>
<td>Hematocrit (X21)</td>
</tr>
<tr>
<td>2nd</td>
<td>Apolipoprotein B (X30)</td>
<td>BMI (X7)</td>
<td>Haemoglobin (X22)</td>
</tr>
<tr>
<td>3rd</td>
<td>Leukocyte conc. (X23)</td>
<td>Pulse pressure (X15)</td>
<td>Leukocyte conc. (X23)</td>
</tr>
</tbody>
</table>

As in the IPASS trial, it is also informative to explore the prognostic strength of each biomarker. Figure 12 presents the PP-graphs for AURORA trial. As earlier, the red area (vertical shaded region) represents the top-K prognostic-biomarkers, while the green (horizontal shaded region) the top-K predictive. For the PP-graphs of Figure 12 we used again k = 1, which corresponds to the score cut-off value of (p−k)/p = (44−1)/44 = 0.9773, where p = 44 is the total number of biomarkers in the trial. The biomarkers being in the red and the green areas, are the ones that ranked, on average, in the first position of the prognostic and predictive ranking respectively. Figure 12a shows that only VT ranks a biomarker in the predictive area. VT ranks X1 (Age) as the most predictive biomarker, but the same biomarker also carries the most prognostic information. Taking into account the previously observed bias of VT to prognostic biomarkers, we might conclude that age is a false positive. On the other hand, Figure 12b shows that our suggested method, INFO+, does not rank any biomarker close to the predictive region (green area) – a result in agreement with the trial findings.

At this point it is useful to explore more the biomarker that INFO returned as the most predictive, the percent of lymphocytes (X24) in the blood. Interestingly, in the subgroup of 994 patients with low percentage (< 65%) (Figure 11b) the ones receiving rosvastatin they had longer MACE-free survival than the ones receiving placebo (HR = 0.78, p =0.037).

The INFO+ method has identified inflammatory status (lymphocytes & leukocytes) as predictive markers, which is a new and unvalidated hypothesis, which did not surface in the AURORA trial. For example, the subgroup of Figure 11b was 994 patients, a non-trivial subgroup size in a trial of this nature. The patient population is, though, known to be complicated, and further research is necessary to assess the plausibility of the suggested biomarkers. There are previously noted prognostic associations for cardiovascular events in the literature (Xiang et al., 2018), but no...
investigation of predictive nature with Rosuvastatin. Such a future investigation seems plausible to yield interesting results, but we do not claim any association from this dataset/paper alone—as always, methods such as INFO+, are exploratory rather than confirmatory.

4 Conclusions

The primary contribution of this work is a formalism for data-driven ranking of predictive vs prognostic biomarkers. With an information theoretic approach, we can disentangle the prognostic vs predictive strength of a biomarker, naturally allowing for issues such as correlated biomarkers. We presented a novel procedure for predictive biomarker discovery, INFO+, which we evaluated over a wide gamut of synthetic data, increasing in difficulty. Our results demonstrate that INFO+ captures higher order interactions between biomarkers without the need to explicitly model the functional form of the predictive part. Furthermore, it disentangles better the predictive and prognostic information of each biomarker, and as a result, provides a better performance in terms of TPR/FNR/PP. The INFO+ approach also requires 1-3 orders of magnitude less in computation, compared to appropriate baselines, making it feasible to explore datasets larger than ever before. Furthermore, we introduced a new visual representation, the $P$-graph, that captures both the prognostic and the predictive strength of a set of biomarkers. We expect that this tool will prove beneficial in visualising and interpreting biomarker investigations for clinical trials. Finally, by formalising the problem of predictive biomarker discovery in information theoretic terms, we can potentially extend this work to other challenging scenarios, such as misclassification bias (Sechidis et al., 2017) or partially labelled data (Sechidis and Brown, 2018).

Acknowledgements

We would also like to thank Iain Buchan, Matthew Sperrin and Andrew Brass for their useful feedback on earlier versions of this work, and all the anonymous reviewers for their useful comments. We would also like to thank Daniel Dalevi for helping us with AURORA trial.

Funding

KS research was funded by the AstraZeneca Data Science Fellowship at the University of Manchester. KP was supported by the EPSRC through the Centre for Doctoral Training Grant [EP/I028099/1]. GB was supported by the EPSRC LAMBDA project [EP/N035127/1].

References


