Disentangling Prognostic and Predictive Biomarkers Through Mutual Information

K. Sechidis [a], E. Turner [a], P. D. Metcalfe [b], J. Weatherall [b] and G. Brown [a]

[a] *School of Computer Science, University of Manchester*
[b] *Advanced Analytics Centre, Global Medicines Development, AstraZeneca*

Abstract. We study information theoretic methods for ranking biomarkers. In clinical trials there are two, closely related, types of biomarkers: predictive and prognostic, and disentangling them is a key challenge. Our first step is to phrase biomarker ranking in terms of optimizing an information theoretic quantity. This formalization of the problem will enable us to derive rankings of predictive/prognostic biomarkers, by estimating different, high dimensional, *conditional mutual information* terms. To estimate these terms, we suggest efficient low dimensional approximations. Finally, we introduce a new visualisation tool that captures the *prognostic* and the *predictive* strength of a set of biomarkers. We believe this representation will prove to be a powerful tool in biomarker discovery.

1. Introduction

We present an information theoretic approach to disentangle predictive and prognostic biomarkers. In clinical trials, a *prognostic biomarker* is a clinical or biological characteristic that provides information on the likely outcome irrespective of the treatment. On the other hand a *predictive biomarker*, is a clinical or biological characteristic that provides information on the likely benefit from treatment. One of the key challenges in personalised medicine is to discover predictive biomarkers which will guide the analysis for tailored therapies, while discovering prognostic biomarkers is crucial for general patient care [10]. We should clarify that our work focuses on hypothesis generation (exploratory analysis), instead of hypothesis testing (confirmatory analysis) [6].

In our work we will focus on a clinical dataset $\mathscr{D} = \{y_i, x_i, t_i\}_{i=1}^n$, where, y is a realization of a binary target variable Y, t is a realization of binary treatment indicator T (i.e. $T = 1$ if patient received experimental treatment, 0 otherwise), and x is a p-dimensional realization of the feature vector X, which describes the joint random variable of the p categorical features (or biomarkers). To make the distinction between prognostic and predictive biomarkers more formal we will follow a strategy introduced by various previous works [4, 6]. Let us assume that the true underlying model is the following logistic regression with up to second order interaction terms:

$$\text{logit}P(y = 1|t,x) = \alpha + \underbrace{\sum_{i=1}^{p} \beta_i x_i + \sum_{i,j=1}^{p} \beta_{i,j} x_i x_j}_{\text{Prognostic term}} + \gamma t + \left(\underbrace{\sum_{i=1}^{p} \delta_i x_i + \sum_{i,j=1}^{p} \delta_{i,j} x_i x_j}_{\text{Predictive term}} \right) t.$$

Covariates with non-zero β coefficients are prognostic, while the ones with non-zero δ coefficients are predictive. Our work proposes an information theoretic framework for deriving two different rankings of the biomarkers, one that captures their *prognostic*

strength, and one that captures their *predictive* strength. On top of that, we introduce a visualisation tool that captures both the *prognosticness* and the *predictiveness* of a set of biomarkers. This tool enables us to identify potentially undiscovered biomarkers, worthy of further investigation.

2. Background on Biomarker Ranking

Here we connect the problem of biomarker discovery, in context of the machine learning problem of feature selection and the clinical trials problem of subgroup identification.

2.1. Prognostic Biomarker Discovery and Feature Selection

We now demonstrate that the problem of selecting *prognostic biomarkers* is equivalent to feature selection using a supervised dataset $\{y_i, x_i\}_{i=1}^n$. There are many different methods for feature selection, but we will focus on information theoretic approaches, where, firstly we *rank* the features and then we *select* the top-k ones that contain most of the useful information. The underlying objective function is to find the smallest feature set X^* that maximizes $I(X^*; Y)$, or in other words that the shared information between X^* and Y is maximized: $X^* = \arg\max_{X_\theta \in X} I(X_\theta; Y)$. Brown et al. [2] derived a greedy optimization process which assesses features based on a simple scoring criterion on the utility of including a feature. At each step we select the feature X_k that maximizes the conditional mutual information (CMI): $J^{\mathrm{CMI}}(X_k) = \hat{I}(X_k; Y | \mathbf{X}_\theta)$, where $\mathbf{X}_\theta$ is the set of the features already selected. As the number of selected features grows, the dimension of X_θ also grows, and this makes our estimates less reliable. To overcome this problem *low order* criteria have been derived. For example, by ranking the features independently on their mutual information with the class, we derive a ranking that takes into account the *relevancy* with the class label. Choosing the features according to this ranking corresponds to the *Mutual Information Maximization* (MIM) criterion; where the score of each feature X_k is given by: $J^{\mathrm{MIM}}(X_k) = I(X_k; Y)$. This approach does not take into account the *redundancy* between the features. By using more advanced techniques, we can take into account both the relevancy and the redundancy between the features themselves, *without* having to compute very high dimensional distributions. Brown et al. [2] showed that a criterion that controls relevancy, redundancy, conditional redundancy and provides a very good tradeoff in terms of accuracy, stability and flexibility is the *Joint Mutual Information* (JMI) criterion [12]: $J^{\mathrm{JMI}}(X_k) = \sum_{X_j \in \mathbf{X}_\theta} \hat{I}(X_k; Y | X_j)$. Through heuristic, this guarantees to increase the likelihood at each step.

While the above framework has been suggested for supervised scenarios, our aim is to explore how it can been extended to be useful in clinical trial scenarios, i.e. $\mathscr{D}$. The extra treatment variable T provides interesting dynamics, but before showing our suggested extension, we will briefly present the literature on predictive biomarkers and subgroup identification.

2.2. Predictive Biomarker Discovery and Subgroup Identification

The problem of deriving *predictive biomarkers* is closely related to the problem of subgroup identification [6]. In clinical trials, patient populations cannot be considered homogeneous, and thus the effect of treatment will vary across different subgroups of the population. Exploring the heterogeneity of subject responses to treatment is very critical for drug development, which is underlined by a draft FDA (Food and Drug Administra-

tion) guidance [10]. As a result consideration of patient subgroups is necessary in multiple stages of trial development. Berry [1] gives the following definition: subgrouping is a partition of the set of all patients into disjoint subsets or subgroups and it is usually determined by a small number of measurable covariates, which are the predictive biomarkers. In the traditional subgroup identification problem the set of predictive biomarkers is relatively small, i.e. 2-3 biomarkers [7].

In the literature there are many different methods for subgroup identification. A popular one is *recursive partitioning* of the covariate space, using criteria that capture the interaction between T and Y [7, 8, 11]. Another solution builds upon the *counterfactual modelling* idea: firstly by deriving a new variable for each patient that captures the treatment effect and then using this variable to select or rank the covariates. For example, Foster et al. [4] can be seen as exploring the covariate space which maximizes the odds-ratio between T and Y. In the following section, we will show that starting from a natural objective function, we can derive predictive biomarkers by exploring areas that maximize the mutual information between T and Y.

3. An Information Theoretic View on Biomarker Ranking

Our work extends the feature ranking framework from supervised to clinical trial data. The treatment variable T provides extra useful information, and a natural way to capture this is by the following criterion: to maximize the shared mutual information between the target Y and the joint random variable of the treatment T and the optimal feature set X^*, or in information theoretic notation: $X^* = \text{argmax}\, I(X_\theta T; Y)$. By using the chain rule [3], these objective can be decomposed as follows in the following way:

$$X^* = \arg\max_{X_\theta \in X} I(X_\theta T; Y) = \arg\max_{X_\theta \in X} \Big(\underbrace{I(X_\theta; Y)}_{\text{Prognostic}} + \underbrace{I(T; Y | X_\theta)}_{\text{Predictive}} \Big)$$

The first term, captures the features with prognostic power, while the second captures the features with predictive power. By optimizing these two terms independently we can derive two different objectives for the two different features set: $X^*_{\text{Prog}} = \arg\max_{X_\theta \in X} I(X_\theta; Y)$

and $X^*_{\text{Pred}} = \arg\max_{X_\theta \in X} I(T; Y | X_\theta)$. Similar to [2], to optimize these two objectives, we can derive a greedy optimization process, where are each step we select the feature X_k that maximizes the following terms:

$$J_{Prog}(X_k) = I(X_k; Y | X_{Prog}), \qquad J_{Prog}(X_k) = I(T; Y | X_k X_{Pred}).$$

where $\mathbf{X}_{Prog}$ are the features already been ranked as prognostic, while $\mathbf{X}_{Pred}$ as predictive. As the number of selected features grows, the dimension of $\mathbf{X}_{Prog}$ and $\mathbf{X}_{Pred}$ also grows, and this makes the estimates less reliable. To overcome this issue, we derive low-order approximations, such as the one presented in Section 2.1.

Theorem 1. *The first two order approximations are given by:*

$$J_{Prog}^{1^{st}}(X_k) = I(X_k; Y), \qquad\qquad J_{Pred}^{1^{st}}(X_k) = I(T; Y | X_k).$$

$$J_{Prog}^{2^{nd}}(X_k) = \sum_{X_j \in X_{Prog}} I(X_k; Y | X_j), \qquad\qquad J_{Pred}^{2^{nd}}(X_k) = \sum_{X_j \in X_{Pred}} I(T; Y | X_k X_j).$$

Proof sketches: For prognostic, the proof is identical to [2], while for the predictive we can prove these approximations by combining the results of [2] with the chain rule [3].

For example, by making assumptions similar to the ones of MIM, we can derive the 1st-order criteria for deriving prognostic and predictive rankings respectively. These criteria do not take into account interactions between features, and as a result fail to capture the *redundancy*. To overcome this limitation so we can use higher order criteria, such as JMI, which explores 2nd-order interaction terms between features.

4. Predictive–Prognostic (PP) Graphs

We now present a visualisation tool that captures both the *prognostic* and the *predictive* power of a set of biomarkers (PP-graphs). We believe that this representation will provide useful information over both the prognostic and predictive power of each biomarker, and it will be helpful for controlling false discoveries in clinical trials. For example, in subgroup-identification (Section 2.2), we define interesting subgroupings by using predictive biomarkers. Many methods, such as the counterfactual modelling, i.e. Virtual-twins suggested by [4], derive as predictive, biomarkers that are strongly prognostic. Using a PP-graph we get more insight over the prognostic and predictive power of each biomarker and this may help in eliminating these type of errors.

Now we will show these graphs through a motivating example. We will use the same data generation model as in [4]. Let us assume that we simulate randomized trials with 1000 patients, and the Xs are generated as independent $X_j \sim N(0,1), j = 1...15$. We consider logit models for data generation:

$$\text{logit}P(y = 1|t,x) = -1 + 0.5(x_1 + x_2 - x_7 + x_2x_7) + 0.1t + 1.5tI(x_1 > 0 \cap x_2 < 0 \cap x_3 > 0).$$

Thus, the patients with $(x_1 > 0 \cap x_2 < 0 \cap x_3 > 0)$ will have an enhanced treatment effect. As a result the three variables, X_1, X_2 and X_3, are the predictive biomarkers. Furthermore, X_1, X_2 and X_7 are the three prognostic biomarkers and the other nine biomarkers are irrelevant.

Figure 1 shows three PP-graphs. In the *x*-axis we have the normalised score of each biomarker derived by a prognostic ranking. We normalised scores to take values from $[0,1]$, where 1 is the score for the most-prognostic biomarker. In the *y*-axis we have the normalised scores for the predictive ranking. The red area (vertical shaded region) represents the top-*k* prognostic-biomarkers, while the green (horizontal shaded region) the top-*k* predictive, for these specific PP-graphs we used $k = 3$, which corresponds to the score cut-off value of $(p - k)/p = (15 - 3)/15 = 0.80$. The intersection of these two areas – orange area (top right shaded corner)– should contain the biomarkers that are both prognostic and predictive. We plot the average predictive/prognostic rankings over 100 sample datasets, using Virtual-twins [4] and our two approaches suggested in Theorem 1. For estimating mutual information, the features were discretized in 4 equal width bins. As we observe, Virtual-twins [4], tends to push a prognostic biomarker (i.e. X_7) into the predictive area –*false positive*. The 1st-order approach classifies X_1 only as prognostic and not as predictive –*false negative*. While, our 2nd-order criterion distinguishes perfectly between predictive and prognostic.

5. Conclusions and Future Work

In this work we focused on disentangling rankings of the biomarkers that quantify their predictive and their prognostic power. We presented an information-theoretic approach, where we started from a clearly specified objective function and we suggested lower-order approximations. Lastly, we introduced a new graphical representation that captures the dynamics of biomarker ranking.
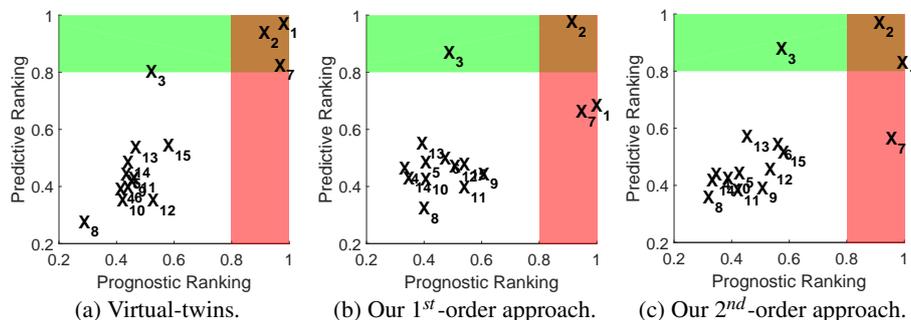
Figure 1. **P-P graphs** when: X_1, X_2 and X_3 are truly predictive, X_1, X_2 and X_7 are truly prognostic, and the rest nine biomarkers are irrelevant. Note that our our 2nd-order approximation distinguishes perfectly between predictive and prognostic.

For future work we are planning to apply our methodologies in exploring the overall survival according to epidermal growth factor receptor biomarker in an lung cancer study [5].

References

[1] D. A. Berry. Subgroup analyses (letter). *Biometrics*, 46(4):1227–1230, 1990.

[2] G. Brown et al. Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. *JMLR*, 13:27–66, 2012.

[3] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley, 2006.

[4] J. C. Foster et al. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880, 2011.

[5] Masahiro et al. Fukuoka. Biomarker analyses and final overall survival results from a phase iii, randomized, open-label, first-line study of gefitinib versus carboplatin/paclitaxel in clinically selected patients with advanced non–small-cell lung cancer in asia (ipass). *Journal of Clinical Oncology*, 29(21):2866–2874, 2011.

[6] I. Lipkovich and A. Dmitrienko. Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. *Journal of Biopharmaceutical Statistics*, 24(1):130–153, 2014.

[7] I. Lipkovich et al. Subgroup identification based on differential effect searchA recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 30(21):2601–2621, 2011.

[8] W.-Y. Loh et al. A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in medicine*, 34(11):1818–1833, 2015.

[9] D. N. Reshef et al. Detecting novel associations in large data sets. *Science*, 334 (6062):1518–1524, 2011.

[10] S. J. Ruberg and L. Shen. Personalized medicine: Four perspectives of tailored medicine. *Statistics in Biopharmaceutical Research*, 7(3):214–229, 2015.

[11] X. Su et al. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research (JMLR)*, 10:141–158, 2009.

[12] H. H. Yang and J. Moody. Data visualization and feature selection: New algorithms for nongaussian data. In *NIPS*. 1999.