THE BEST TREATMENT FOR EVERY PATIENT

New algorithms to predict treatment benefit in cancer using genomics and transcriptomics

JOSKE UBELS

The best treatment for every patient

New algorithms to predict treatment benefit in cancer using genomics and transcriptomics

Joske Ubels

ISBN: 978-94-6416-208-0

Design by: Deniz Lehnert
Layout: Joske Ubels
Printed by: Ridderprint | www.ridderprint.nl

The Best Treatment for Every Patient

New algorithms to predict treatment benefit in cancer using genomics and transcriptomics

De beste behandeling voor elke patiënt

Nieuwe algoritmes om met genomics en transcriptomics baat bij behandeling te voorspellen bij kanker

Proefschrift

ter verkrijging van de graad van Doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. R.C.M.E. Engels

en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op

dinsdag 1 december 2020 om 15.30 uur

door

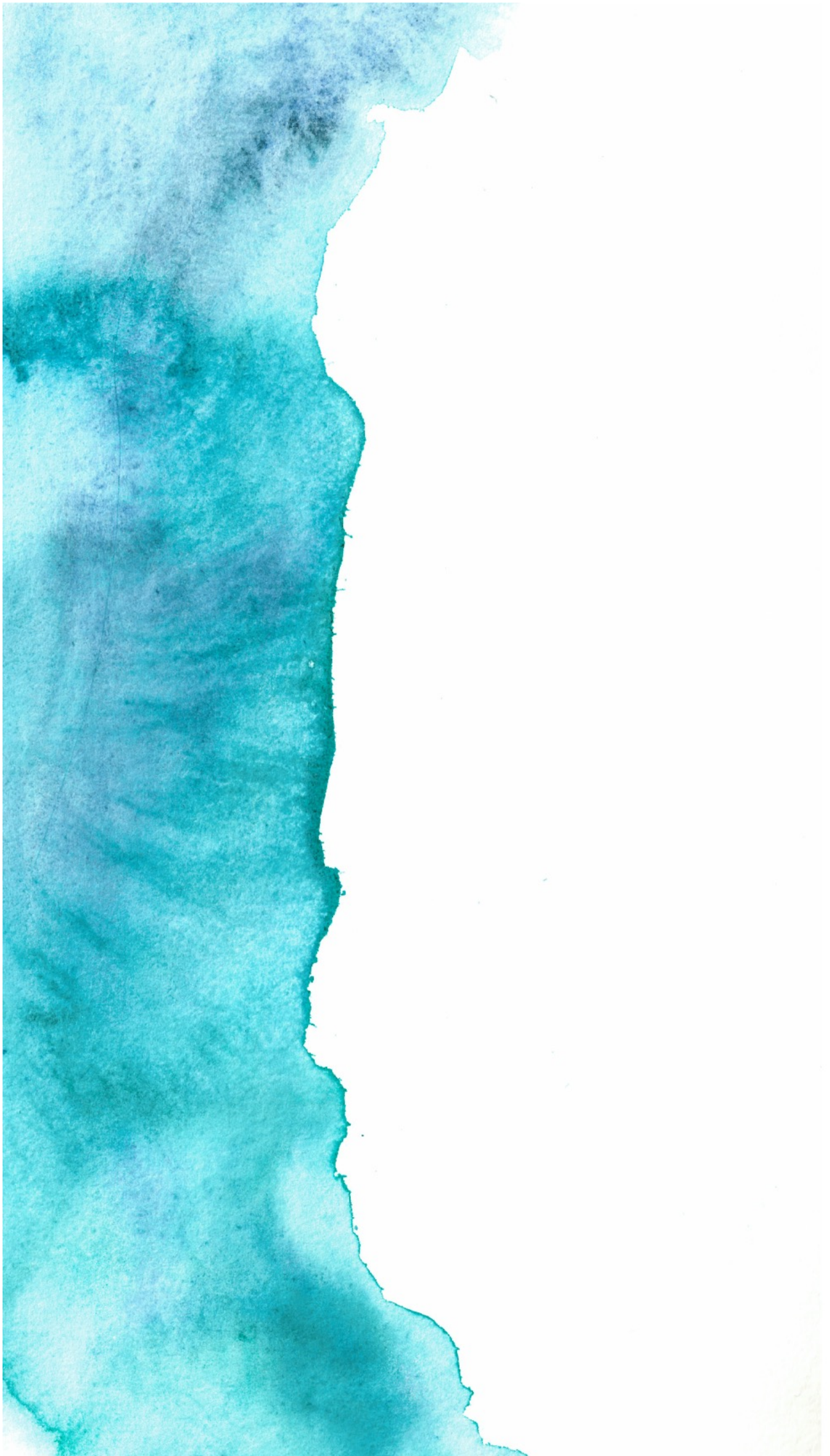
Joske Ubels
geboren te Amstelveen.

Erasmus University Rotterdam

Promotiecommissie:

Promotor:	Prof. dr. P. Sonneveld
Overige leden:	Prof. dr. I.P. Touw Prof. dr. ir. M.J.T. Reinders Prof. dr. P.J. van der Spek
Copromotor:	Dr. J. de Ridder

Contents



Chapter 1

General Introduction

1

Ever since scientists were first able to read a DNA sequence, techniques to do so have developed explosively. The transcription of DNA to mRNA and subsequent translation into protein determines to what extent a gene plays a role in the functioning of the cell. For this reason, techniques to measure gene expression (i.e. the abundance of the mRNA of a certain gene in a sample) were developed. The first description of a microarray approach to measure gene expression on a genome wide level, rather than in individual genes, was described in 1995 (Schena et al. 1995). From this point onwards, progress was rapid and with more data available, ever more correlations between gene expression and disease progress could be discovered. A research area in which the application of gene expression measurements particularly exploded is cancer research.

Cancer

Broadly defined, cancer is the malignant proliferation of cells. In other words, cancer arises when cells start dividing when they should not. A fully developed human body consists of an estimated 37.2 trillion cells (Bianconi et al. 2013). From the moment the ovum is fertilized and starts dividing to form a foetus, the division of each cell is tightly regulated. Whether a cell divides or not is influenced by many factors, arising from both within the cell as well as its environment. A host of mechanisms are involved in this complicated process: mechanical factors, hormones, signalling molecules and nutrient receptors, among other things.

When all signals align and a cell starts to divide to form a new cell, the roughly 3 billion DNA bases in our genome need to be copied in order to provide the new cell with an identical copy of the genetic material. This process is not error free. It has been estimated that per 100,000 bases one error occurs (Arana and Kunkel 2010). If these errors would persist and be passed on to the new cell (and then to subsequent progeny), they could lead to dysregulated activity within these cells and eventually disease. There are therefore many safeguards against passing on aberrant DNA; fidelity of the copy is checked both during transcription and before the final cell division. When errors are detected that cannot be corrected a cell can induce apoptosis, a controlled cell death.

Nevertheless, with 37.2 trillion cells and 3 billion bases per cell, sometimes errors will slip through and be passed on to the next generation of cells. However, most often, in

order to disturb the function of a certain gene, both alleles of the genes need to contain an error. This is known as Knudsons "two hit-hypothesis" (Knudson 1971). Even if this happens and leads to a cancerous cell it will not necessarily cause disease; the cell can be destroyed by the immune system before proliferating and forming a tumor.

Then what does need to happen before cancer develops, given all the safeguards? In 2000 Hanahan and Weinberg defined 6 hallmarks of cancer that can be used to understand and categorize the steps that are required for carcinogenesis to be initiated (Hanahan and Weinberg 2000). Two hallmarks are about taking the brakes off proliferation: *resisting cell death* and *evading growth suppressors*. This for example means disrupting the checks for accurate DNA replication before cell division. Two more hallmarks are about accelerating proliferation: *enabling replicative immortality* and *sustaining proliferative signalling*. A normal, healthy cell has a finite number of divisions it is able to perform, while a cancer cell needs to be able to divide indefinitely. Moreover, a cell is usually dependent on signals from its environment to kickstart the division; a cancer cell needs to sustain its own signals to achieve ongoing proliferation. Lastly, cancer is characterized by its ability to leave the site of origin and spread through the body. It therefore needs to *activate invasion and metastasis*. To have access to nutrients and oxygen a cancer cell needs to *activate angiogenesis* in order to form new blood vessels. The follow up paper in 2011 introduced four other hallmarks and also described the need for cancer cells to evade the immune system (Hanahan and Weinberg 2011).

According to the hallmarks of cancer each cancer cell needs to exhibit all of these hallmarks to develop into disease. However, there are many different ways a cell can acquire one or more hallmarks since dysregulating different parts of the control system can have the same downstream effect. This dysregulation is usually caused by changes in the DNA of key genes regulating the cell behavior. Some genes controlling the cell cycle need to be under-expressed, i.e less present than in a healthy cell. On the other hand, cells driving cell division can be over-expressed. The fact that there are different roads a cell can take to become a cancer cell, means that the same type of cancer can exhibit different behavior and response to treatment in different patients.

1

When reading out the DNA sequence and measuring mRNA became easier and cheaper, tumors that were always considered to be the same disease, started to be subtyped and were shown to have a vastly different genetic architecture. Breast cancer was the first type of cancer where this was extensively shown. Perou et al. already described 6 different intrinsic subtypes in 2000 based on gene expression measurements (Perou et al. 2000).

Not long after Perou et al., Van 't Veer et al. took the next step and described how gene expression measurements could be used to predict survival in breast cancer at the moment of diagnosis (Veer et al. 2002). This 70-gene model could predict if a breast cancer patient had a high or low risk of experiencing a metastasis of the primary tumor within 5 years. This proved that the different genetic background of tumour influences the progression of disease. Many different gene expression signatures in many different cancer types would follow (Raponi et al. 2006; Barrier et al. 2006; Bullinger and Valk 2005; Kuiper et al. 2012).

Machine learning

These growing datasets also called for new analysis methods and machine learning started to play a bigger part in biological and medical research. The term "machine learning" was coined by computer scientist Arthur Samuel. His 1959 paper on an algorithm that can play checkers starts with describing his studies on machine learning as "concerned with the programming of a digital computer to behave in a way which, if done by human beings or animals, would be described as involving the process of learning" (Samuel 1959).
Samuels checker-playing program is seen as the first machine learning program (Schaeffer 2006). It clearly demonstrates an aspect of machine learning that is often explicitly included in later definitions: they can build models based on available data to perform a certain task on new data - without being explicitly programmed to do so. That is, the algorithm is told what its ultimate goal is (winning at checkers, in the case of Samuel) and the boundaries of the problem (the rules of checkers). However, how it should behave within these boundaries to achieve its goals is something it has to learn, as this behaviour is not explicitly programmed. Moreover, the program has to learn this in a way that makes its solutions applicable to situations on the board it has never seen

before. What goes for checkers, goes for all machine learning problems. A model that learns to predict cancer progression in an available dataset is useless if it cannot also predict this in a newly diagnosed patient with gene expression patterns it has never seen before.

Machine learning approaches can generally be divided into three categories: supervised learning, unsupervised learning and reinforcement learning. Unsupervised learning aims to learn a structure in the data, without being guided by labels or classifications. Clustering algorithms are a good example in this category; they attempt to group data points that are similar to each other within a cluster and separate data points that are very dissimilar into different clusters. The different subtypes that were discovered in breast cancer are an example of unsupervised learning. Figure 1a shows a gene expression matrix, figure 1b shows the same gene expression matrix when clustered through unsupervised learning. As can be seen in the clusters marked by the two blue rectangles, clusters can be formed through finding genes that are all highly expressed, but also through a combination of under-expression (green) and over-expression (red). In reinforcement learning, the algorithm takes a sequence of decisions and gets rewarded (or punished) for the outcome of this decision. It learns by updating its model and amending its decision in response to this reward. Samuels checker player is an example of reinforcement learning; certain moves (decisions) lead to better game outcomes (rewards) than others. Unsupervised and reinforcement learning will not be considered further here; most approaches to predict survival or progression in cancer and all algorithms presented in this thesis use supervised learning.

Supervised learning uses labelled data as input and learns a model that can accurately predict something about this label on unseen data; we use labels to define what the model should learn. The 70-gene breast cancer signature, for instance, labels patients as ‘poor prognosis’ if their survival was shorter than 5 years and ‘good prognosis’ otherwise. Labels that indicate class membership (like poor or good prognosis) are categorical, but labels can also be continuous; for example, reduction in tumor size. Approaches differ for both types of labels, but in all supervised methods the model combines certain features (i.e. what was measured) to predict the label of interest (i.e. what we cannot measure and want to know). Figure 1c shows supervised learning with a continuous label; the red line is the regression line that described the relation between

gene expression for a certain gene and the tumor size. This model can be extended to incorporate many variables. Often, we have measured more about the sample than is relevant. For example, we have measured expression for all genes, but the majority is not informative for survival time. Most approaches therefore include a feature selection step, to select the most relevant features. Feature selection can precede training or be incorporated in the training procedure.

High dimensional data and overtraining

An important challenge in machine learning, which is particularly salient in the analysis of gene expression data, is the curse of dimensionality. This challenge stems from the fact that we, in general, have many more features (genes) than samples (patients). When considering high dimensional datasets, it is likely to find untrue correlations: when you consider thousands of features, some will by chance correlate with the label even if no true signal is present in the data. It is important to take this into account when selecting features and training the model. In a high dimensional setting a machine learning model can easily overtrain, which means the model is not fitting an actual relationship between the gene expression patterns and the outcome, but starts to accommodate noise in the data. As a result, overtrained classifiers will not work on new and unseen data. Figure 1d shows how this can happen with categorical labels; the grey line represents an overtrained classifier. Instead of learning a general distinction between good and poor prognosis (the red line), it has fitted the specific datapoints present in this dataset. To assess whether a true pattern is found (i.e. a pattern that generalized to new and unseen data) an important concept is the separation of training and test data, where one dataset is used to fit a model and other, unseen data is used to assess the performance of the model. Of course, we usually do not have unlimited data available. To guard against overtraining we can use cross-validation. In cross-validation we split the training data in equal parts, for example three, also called folds (Figure 1e). We then train a model on the first two folds and test the model on the remaining third to obtain a better estimate of the expected performance on external data. We repeat this until all three folds have been used as test data once. When we do multiple repeats of this, the variables or models that perform well over all folds are most likely to be true and can be tested on true external data.

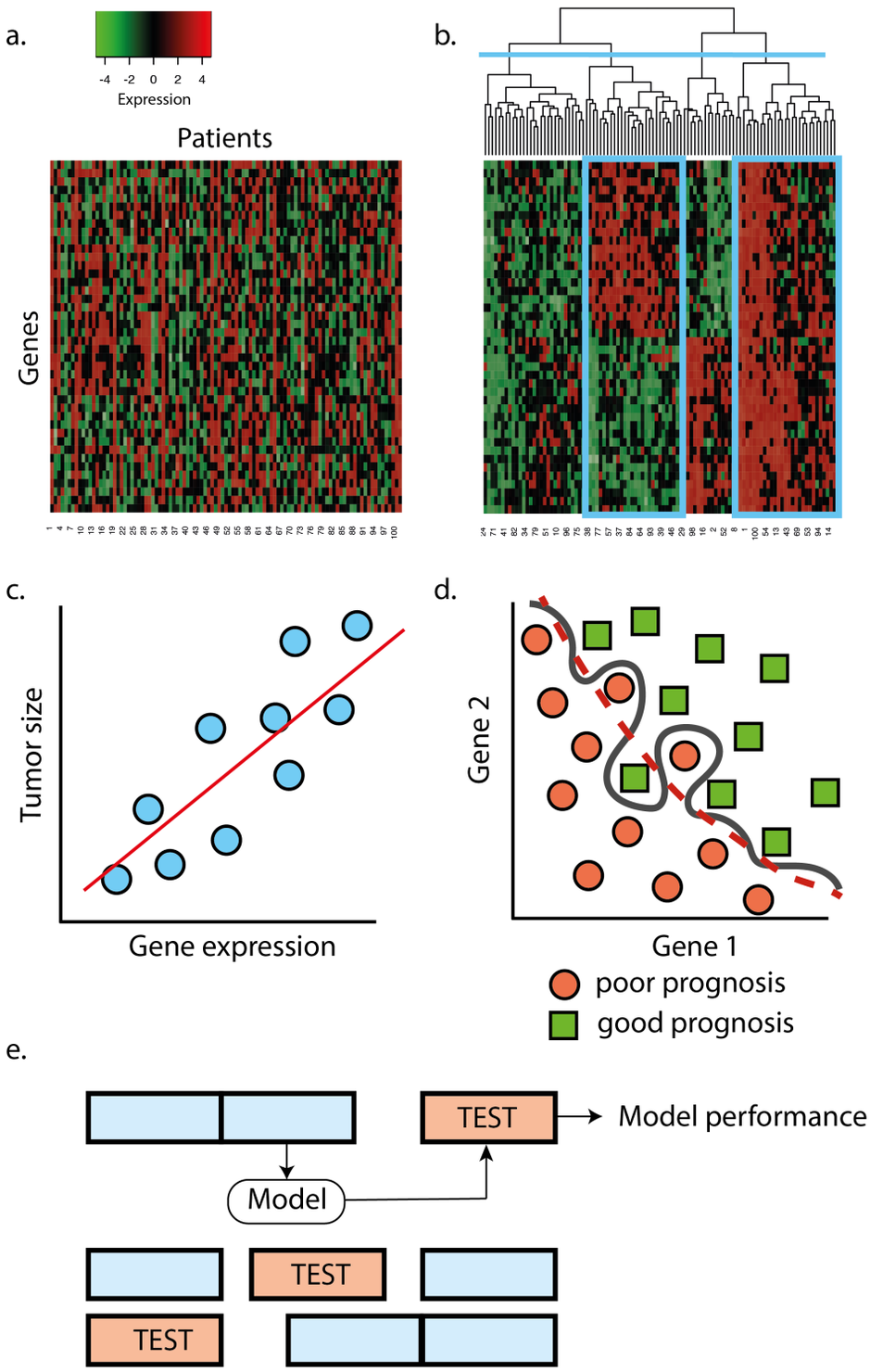


Figure 1. a. Unclustered gene expression. b. Unsupervised learning: clustering of gene expression. c. Supervised learning with continuous label: regression. d. Supervised learning with categorical label: classification. The grey line represents an overtrained classifier, the red dotted line a more generalizable classifier. e. Three-fold cross validation.

One can also try to directly prevent overtraining in the training of the model itself; one way is regularization. When applying regularization, the model contains parameters that penalize complicated models. If a model is allowed to incorporate enough features, it can fit any pattern. Imagine the model would incorporate each feature that was measured; it could describe the training data perfectly, while not learning general patterns. While regularization may lead to choosing simpler models with a slightly worse fit, such models are more likely to generalize to external data.

Another way of preventing overtraining is using ensemble classifiers and bootstrapping. In an ensemble classifier many weak classifiers are trained: classifiers of which the performance by
itself will not be satisfactory. The idea here is that a weak classifier will make many mistakes in assigning a sample to a class, but when we combine many weak classifiers that all make a different mistake, together they can still distinguish better between classes than any classifier on its own. We can make it more likely that these classifiers fit different effects in the data by bootstrapping. In bootstrapping we sample randomly from the data (typically with replacement) to generate a training dataset which encompasses a random subset of the variables and samples of the full dataset. Because we do this for each classifier separately, all classifiers have access to a slightly different part of the data. This simultaneously assures they cannot overfit on the dataset as a whole and that each classifier will make different mistakes. Which approach to prevent overtraining is best depends on the type of data and classification problem, though many successful approaches use a combination of all mentioned techniques.

Personalized medicine

If tumors behave differently based on the differences in mutations and gene expression patterns, a logical next step is to investigate whether these differences can be used to inform treatment. In 2001, it was estimated that for treatment across cancer types only one in four patients sees a beneficial effect (Spear, Heath-Chiozzi, and Huff 2001). While these numbers have improved somewhat with the rise of targeted therapies, it is clear that even today we treat patients with drugs that will not benefit them.

The practice to tailor treatment to the individual patients is known as personalized medicine. Broadly, we can differentiate between two approaches in personalized medicine. The first approach entails looking for specific mutations or aberrations present in the tumor that can be targeted with drugs. One of the first examples of such an approach was applied in chronic myelogenous leukemia, a type of blood cancer. A common aberration in CML creates a so-called fusion gene between the *BCR* gene and the *ABL* gene. This fusion gene encodes a protein that drives the rapid division of leukocytes. In the late 90's a drug was developed - imatinib - that specifically inhibited this fusion gene and enormously improved survival for patients whose tumor harbors this particular fusion gene (Druker et al. 2001). By now more drugs that target cancer specific mutations have been introduced, like vemurafenib for *BRAF* mutations and crizotinib targeting *ALK* positive tumors (Chapman et al. 2011; Shaw et al 2013). While this has led to great advances in cancer survival, there are many cancer patients for whom the tumor is not characterized by a cancer-specific, targetable mutation (Priestley et al. 2019) and that therefore do not benefit from this strategy.

The second approach in personalized medicine is based on the presence of patient or tumor characteristics that can predict whether they will benefit from generic treatment, i.e. treatment not targeted to a cancer-specific aberration. Sometimes this can be achieved by simply associating known prognostic markers with treatment benefit. For example, it was shown that patients identified as low-risk by the 70-gene breast cancer signature could safely forego chemotherapy (Cardoso et al. 2016). There have also been more specific machine learning approaches to predict a patient's response to a treatment, both using mutational data and gene expression (Le et al. 2017; Tanoue 2012; O'Connell et al. 2010). Response to treatment can also be determined by non-tumor characteristics, like how the body metabolizes the drug before it reaches the tumor. The field of pharmacogenomics has identified many germline variants - common DNA variants inherited from your parents - that have an influence on how a drug is metabolized. Certain variants known to influence treatment are already routinely used to determine for example effective dose (van der Wouden et al. 2019). Of course, even targeted treatments do not benefit every patient that receives them; here the two approaches combine.

1 Survival analysis

To investigate whether one treatment leads to a better patient outcome than another treatment, survival analysis is often employed. This enables an assessment of whether a patient is statistically significantly less likely to die from the disease when for example being treated with a certain treatment. To perform survival analysis we need to define an endpoint: this is the outcome we are interested in. This can be death, but also, for example, metastasis of the cancer. A big challenge in analyzing survival data is the fact that some patients will be censored. When patients are enrolled in a trial and follow up is performed for 10 years some patients will die during this period and some will be known to be alive at the end of trial. However, there will also be a group for whom no information is available: they have left the trial or contact was lost for some reason. The patients for whom we have not recorded a date of death will be censored; we record the last date on which they were known to be alive. The challenge is using the data from censored patients; even if follow-up was not completed, there is useful information in the fact that a patient was still alive after a certain time. The most commonly used model is Cox proportional hazard model (Cox 1972). Here the partial log likelihood is optimized over the β by:

$$\ell(\beta) = \sum_{i:C_i=1} (X_i - \log \sum_{j:Y_j \geq Y_i} \theta_j)$$

Optimizing the likelihood means the model finds the β that is most likely to give rise to the observed data. The i indicates the censoring status; if this is 1 a date of death was recorded, if it is a 0 this was not the case. Simply put, the Cox model describes which β best explains the observed sequence of deaths. This enables us to take censored patients into account up to the point of censoring; if a patient is censored at 5 years, we know for sure everyone with an event before 5 years died before them. The X in the formula represents the variable under consideration; when evaluating treatment effect this is the treatment variable. If a treatment had no effect at all, the β will be (near) zero.

When we use the Cox model to estimate treatment effect this is often captured in the hazard ratio. The hazard ratio is the exponent of the β; when a treatment has no effect,

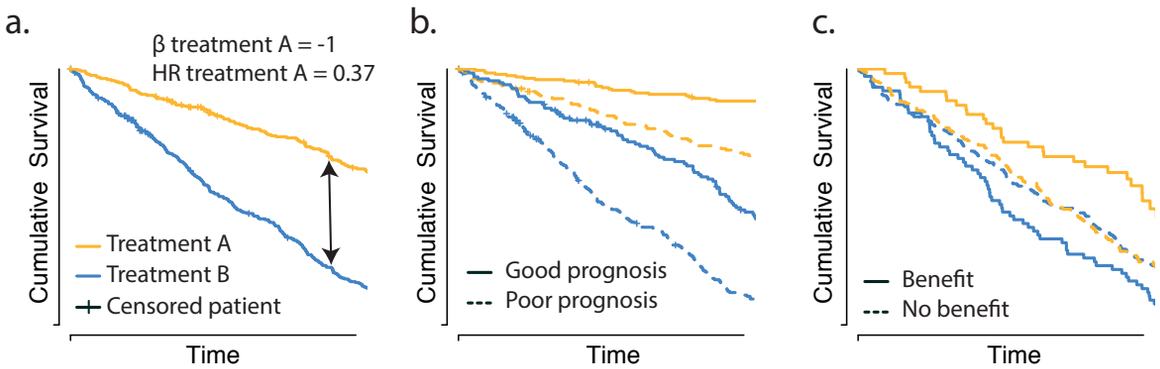


Figure 2. a. An example Kaplan Meier plot of a treatment A that confers a survival advantage. b. An example plot of a prognostic classifier. c. An example plot of a predictive classifier.

the HR is 1 (i.e. the exponent of 0). Survival data is often visualized in Kaplan Meier plots, an example of which is shown in Figure 2a. The β describes the difference between the two treatment groups, here with treatment A as reference. The β is this example is -1, which means that when a patient receives the treatment their hazard of dying is lower. A β above 0 would signify the patient has a higher hazard when treated with treatment A. Censored patients are represented with a vertical mark. The Cox model makes several assumptions about the data, the most important of which are that a) the hazards between the different groups are proportional over time and that b) the censoring is uninformative. Proportional hazards mean that the difference in risk between groups is the same at any point in time - i.e. if treatment A reduces risk two-fold this should be true in year 1 but also in year 5, etc. When the lines in a Kaplan Meier plot cross, this assumption is violated. Uninformative censoring means that the variable under consideration should not influence whether a patient is censored. If one treatment group has much more censoring and this is somehow due to the treatment itself, this cannot be modelled accurately within the Cox model and it will bias the estimate of treatment effect.

In this thesis we mostly employ Cox proportional hazards modeling to estimate treatment effect, but it is not confined to treatment estimates. It can for example also be used to estimate the effect of gene expression on survival; it can incorporate multiple variables at once and a fitted Cox model can then also be used to predict outcome for a new patient. Survival analysis has been combined with machine learning, where the survival data functions as a label. For example, regularized cox models (i.e. models with a penalty on model complexity) were developed that can be used to model survival on

high dimensional data sets like gene expression data (Simon et al. 2011). Another popular approach for high dimensional data is training a Random Survival Forest. A random forest is a machine learning approach that can be used on both discrete and continuous labels and trains an ensemble of decision trees (Breiman 2001). It is particularly suitable for high dimensional datasets as it prevents overtraining both by bootstrapping and forming an ensemble classifier (discussed in the Machine Learning section). As the name suggests, Random Survival Forests extend this approach to survival data with censoring present. Rather than predict a particular label, Random Survival Forests aims to divide the samples in subsets with a maximum survival difference (Ishwaran and Lu 2019).

The difficulties of treatment benefit

Due to the rapid developments in cancer treatment, there is an increasing number of cancer treatments available to choose from and often it is not clear which will be the best choice. The classifiers previously discussed either predicted prognosis (regardless of treatment) or predicted response to a single treatment. Figure 2b shows a Kaplan Meier plot for a prognostic classifier. While this classifier identifies patients with a better survival, the benefit from treatment A is present in both classes. Had this classifier been trained and validated on a population with solely patients who were treated with treatment A it would be impossible to distinguish between a predictive effect for treatment A specifically or a general prognostic effect. In this example the poor prognosis group still survives better than the good prognosis group when treated with treatment B; all patients should receive treatment B. When multiple treatments are available and a choice has to be made between them, the current classifiers are not sufficient. Arguably the most clinically relevant question is which treatment will benefit a patient most; i.e. which treatment would lead to the longest survival. However, patients who benefitted from a certain treatment cannot be identified straightforwardly, since we can only observe the response to the treatment the patient actually received. Even if a good response was achieved, it does not mean the patient benefited specifically from this treatment. Possibly any other treatment would have achieved the same results. Conversely, even if a patient had a short survival time, the given treatment could still have been the best choice - maybe the response would have been even worse on any other treatment. We can thus not label a patient as benefiting or not from the

observed survival. Traditional supervised machine learning approaches cannot be employed; these approaches rely on predefined labels.

We thus need to employ other approaches to predict treatment benefit. In all following work we define treatment benefit as a patient surviving longer on the treatment of interest than they would have done on a comparator treatment. Figure 2c visualizes treatment benefit in a Kaplan Meier plot: we identify a 'benefit' class with a larger benefit than the population as a whole and a 'no benefit' class where treatment A does not lead to a better survival.

One approach is to investigate if known prognostic markers are also linked to treatment benefit, as was done in the case of the 70-gene breast cancer signature. However, these associations can only be investigated after the genes or markers were identified using survival information only (or possibly an unsupervised approach). It is to be expected that methods taking treatment specific survival into account in the discovery will be superior to after the fact analysis.

Another approach is to model on two treatments separately, but to only retain variables that have an opposite effect in both treatment arms. The drawback here is that the model does not get an opportunity to specifically look for a combination of variables that achieve this. It is not necessarily expected that there will be one single marker that can separate these groups. Of course, a good response to one treatment would not automatically mean a bad response to another treatment and markers would be difficult to find in separate analyses.

We show in **Chapter 2** that we cannot successfully train a model on labels derived directly from survival and treatment information. In this thesis we will present multiple approaches to predict treatment benefit using survival outcome and treatment annotation without having to define training labels.

Counterfactual reasoning

When we talk about treatment benefit, we are trying to answer the question "what would have happened had we given this specific patient a different treatment?". The

1

answer to this type of what-if question is known as a counterfactual. Counterfactual reasoning has mostly been used in establishing causality, i.e. which variable is causal for the difference in events when the "if" is changed. It should be noted that establishing causality is not necessarily the goal of machine learning; accurate prediction is. These two things can and perhaps ideally do occur together, but it is not necessary.

Counterfactual approaches can explain models or important variables by investigating what would have changed the prediction of an already existing model (Mothilal, Sharma, and Tan 2020). This is visualized in Figure 3a; would a different treatment have led to the same poor outcome? The problem with these approaches is that a model needs to exist already or that at the very least candidate variables need to be known. With a defined model it can be investigated what the change in predicted outcome is when the value of a variable (like the treatment variable) is changed. In a gene expression setting tens of thousands of variables are available and it is likely only a small part of those are relevant to benefit from the treatment under investigation.
We thus need a method to answer the what-if question without already having a model. An example of attempting this is an approach using so-called 'virtual twins' (Foster, Taylor, and Ruberg 2011). In the context of a clinical trial a 'virtual twin' can be modelled for each clinical trial participant, where the twin undergoes the counterfactual

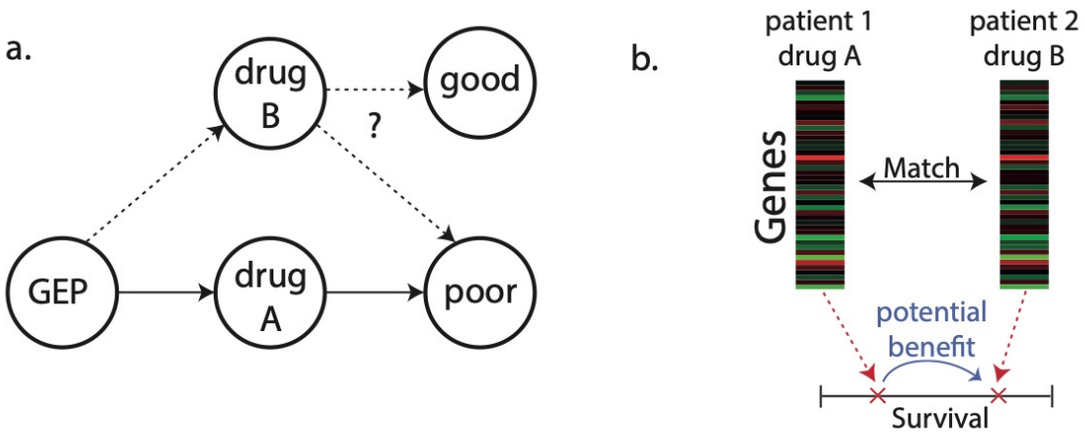


Figure 3. a. Example of a counterfactual model. We have observed a poor outcome when a patient with a certain gene expression profile (GEP) was exposed to drug A. The model now needs to predict whether drug B would have led to the same outcome. b. Example of how matched patients can be used; patient 2 has a similar gene expression profile as patient 1, but was exposed to a different drug and experienced a longer survival. This represent the potential benefit for patient 1, had they been treated with drug B.

condition of the real participant (Vittinghoff et al. 2010). Here again, the assumption is that the estimation of both responses arises from the same (linear) model and that the measured variables are independent of the alternative option you are modelling. Figure 3b shows this matching of patients based on gene expression profiling, where patient 1 and 2 are very similar, but received a different treatment. However, the virtual twin approach was proposed in a setting of low dimensionality, considering less than 100 variables. It has been shown since that the assumptions made in this approach do not hold in high dimensional settings (Lu et al. 2018). Other imputation-like approaches, where the outcome on the treatment not received is regarded as a missing data point, have mostly been applied in a setting with a limited number of variables that are all likely to be of influence. In a high dimensional setting like gene expression - or even more difficult, germline variation - we are dealing with many irrelevant variables, but no way of determining which are irrelevant before building the model. We do not know which genes should be used to identify matched patients. We thus need new methods to be able to apply counterfactual reasoning in high dimensional datasets.

Multiple Myeloma

Chapter 2 and **Chapter 3** deal with predicting treatment benefit in multiple myeloma. Multiple myeloma is a cancer of the plasma cells that develops in the bone marrow (Rajkumar 2018). Plasma cells are a fully differentiated white blood cell and play an important role in the immune defense by producing immunoglobulins, i.e. the antibodies that enable the immune system to recognize pathogens. Multiple myeloma can develop slowly, sometimes being present as smouldering multiple myeloma over the course of decades, to suddenly spike and cause symptoms (Kyle et al. 2007). Multiple myeloma is also a very heterogeneous disease. A few chromosomal aberrations are often found in multiple myeloma, but most only occur in a minority of the patients (Nahi et al. 2011). Mutations in DNA are also sparse and there is no clear mutational event to define multiple myeloma (Walker et al. 2018). A lot of effort has gone into distinguishing patients with high or low risk variants of the disease and most of these are defined by gene expression (Szalat, Avet-Loiseau, and Munshi 2016). It is still an incurable disease, though survival expectancy at the moment of diagnosis has increased significantly in the past two decades due to novel treatment being introduced in the clinic (Rajkumar 2018).

Two major treatment classes now used in the clinic are proteasome inhibitors (PI) and immunomodulatory drugs (IMIDs) and in this thesis we focus on predicting benefit to PIs. The rationale behind PIs is that MM cells overproduce immunoglobulins, which are proteins. The proteasome is the main way a cell has to get rid of unwanted proteins and this system is overburdened in MM cells. When the proteasome is inhibited, proteins start to accumulate in the cell, eventually triggering apoptosis when this situation cannot be resolved. MM cells are more reliant on the proteasome than other, healthy cells, providing a therapeutic window for PI treatment (Moreau et al. 2012).

An open problem is whether the risk profiles and different gene expression patterns across MM patients can also be informative for which treatment is ideal. Currently these markers are not used to decide on an ideal treatment and we thus have to look beyond the known markers. Multiple myeloma represents a good test case for the prediction of alternative treatment response from gene expression data; clinical trials are available and it is known gene expression is of influence on disease trajectory and there is an unmet need for tools to aid in treatment decisions. A clinical trial setting is ideal for training a model like this, since treatment assignment is random. As discussed, in counterfactual reasoning it is assumed that the variables in the model are independent of the condition to be modelled; this can be safely assumed in a clinical trial.

Understanding treatment benefit

Predicting treatment specific survival is one part of the challenge and very important in clinical decision making. The next question that inevitably presents itself is *why* certain patients respond better than others to a certain treatment. Could a well-performing model shed some light on this?

A usual step to gain insight in the mechanism behind the predicted benefit is to investigate the genes included in the model that can predict treatment response, but more often than not these do not present a clear picture of mechanisms of treatment response. Classifiers trained for the same purpose, with similar performances, show very little overlap in genes used (Tang et al. 2017). For the 70-gene breast cancer classifier mentioned earlier, it was shown that a similar classifier can be built when these 70 genes are excluded from the analysis (Ein-Dor et al. 2005). The fact that a good prediction performance can be achieved by many different genes is at least partly caused by the

great redundancy in gene expression information. This in turn is due to the fact genes act in pathways and regulate each other giving rise to highly (inversely) correlated gene expression patterns. For classification purposes, it can be irrelevant which of these genes are included in the model, since they provide the same information - as shown by Ein-Dor et al. Substituting one for the other will not change the model performance. However, for the biological interpretation and understanding the role of these genes in determining patient benefit to treatment these genes are not equal.

Some classification approaches take this aspect into account, and include pathways and known relationships between genes in the model. However, it has been shown these methods can achieve similar performances when using random networks as when true biological networks are used (Staiger et al. 2012), rendering the importance of the biological links doubtful. These networks can also be biased towards well-studied genes; if a gene is known to be important in cancer development, more research will study it and more relationships will be discovered. Thus there are a few well known genes, that are annotated in many different contexts, while other genes are not annotated at all (Haynes, Tomczak, and Khatri 2018). This limits the new mechanisms that can be discovered to what is already known. Moreover, disease can also change how genes interact with each other; interactions in healthy tissue can be very different to interactions in cancerous tissue and interactions can differ between cancer types. A possible approach is to learn new gene networks that are specific to the disease or even the treatment. This can be done in a data-driven manner, so it is not biased by gene annotation of pathways in health cells. In this thesis we use both known biological annotation (**Chapter 2** and **Chapter 4**) and present a method to learn new networks, specific to treatment benefit (**Chapter 3**).

Contribution of this thesis

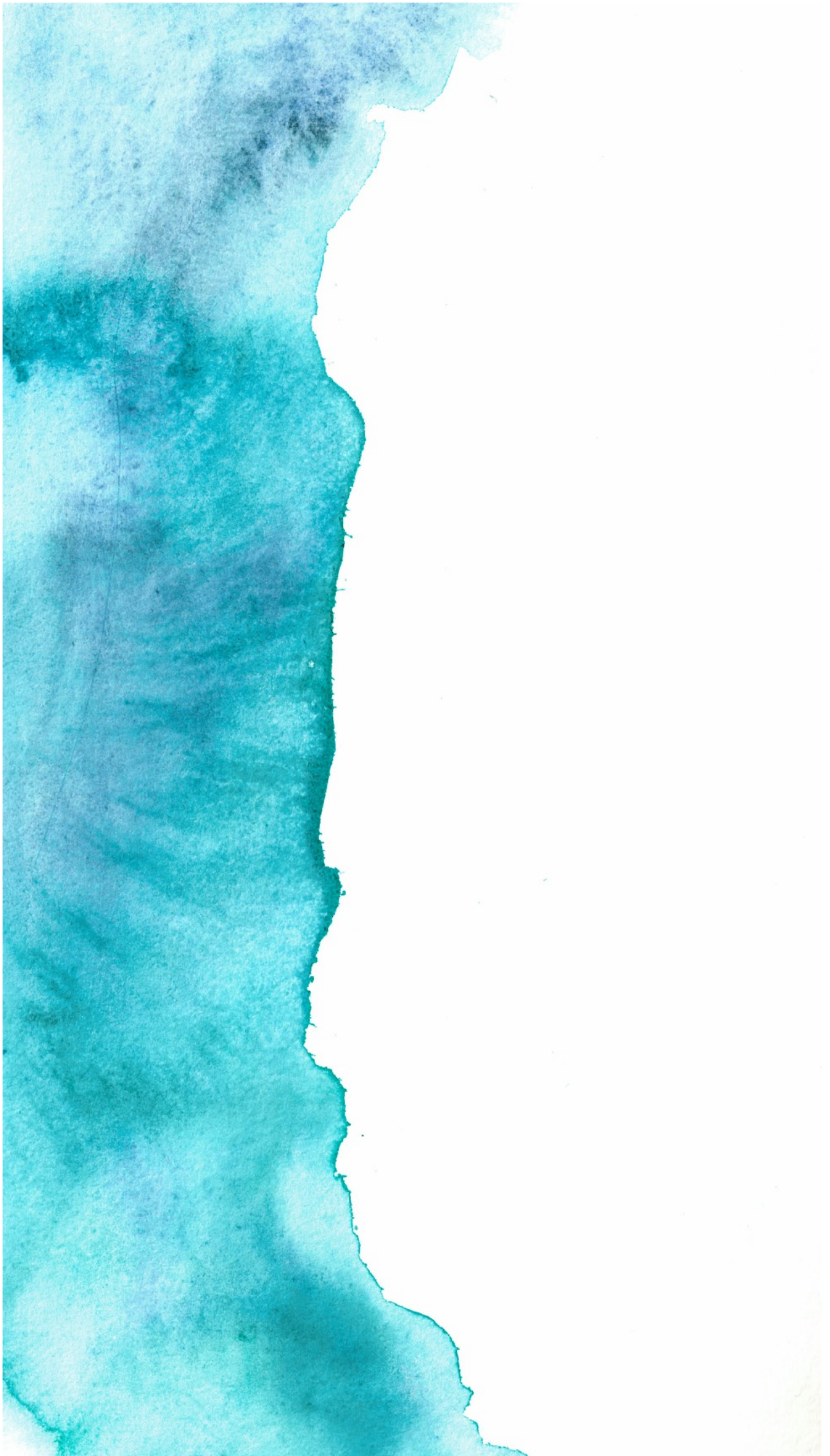
There is a gap between the machine learning approaches available to predict treatment response and the clinical reality, where we are interested in answering the question: which drug is the best choice for this patient? There have been several approaches developed in the field of counterfactual reasoning, but none that can handle the high dimensional nature of gene expression data. In this thesis we present several different approaches to model what the outcome would have been for a patient had they received

1

a different treatment. With these we can predict treatment benefit in a clinically relevant way. Much of the work rests on our concept Simulated Treatment Learning, which uses the idea that genetically similar patients who received a different treatment can be used to model response to the alternative treatment. In **Chapter 2** we present GESTURE (Gene Expression-based Simulated Treatment Using similaRity between patiEnts), an algorithm that evaluates which gene sets (here formed by Gene Ontology annotation) are most relevant to treatment benefit and combines them in an ensemble classifier to predict treatment benefit for new patients. We show its performance in a multiple myeloma dataset, predicting benefit to both bortezomib (a PI) and lenalidomide (an IMID), representing two major treatment classes in multiple myeloma. In **Chapter 3** we present STLsig, which uses Simulated Treatment Learning to form disease and treatment specific gene networks that can predict treatment benefit. We demonstrate its utility in predicting treatment benefit to PI treatment in multiple myeloma and moreover showing that the genes in the signature are unique (i.e. a new, similar performing signature cannot be found with the same method when the genes are removed). This offers perspective for biological interpretation. In **Chapter 4** we adapt GESTURE to predict chemotherapy benefit in breast cancer. This offers additional challenges, as we do not have access to randomized trial data and the event rate is much lower. Here we also find the limitations of such a setting, as we can build a classifier that validates in cross validation, but not in external data. In **Chapters 2 - 4** we use tumour gene expression data to predict treatment benefit, but in **Chapter 5** we use SNP data (i.e. germline variation) to predict treatment benefit. We introduce RAINFOREST (tReAtment benefIt prediction using raNdom FOREST) and use it to predict benefit to cetuximab in metastatic colorectal cancer. This method is based on random forests, but does not need predefined labels and can identify a subset of patients who benefit from the addition of cetuximab, while the population as a whole.

Together, we provide an array of tools that can be used to predict treatment benefit in high dimensional settings and we show their utility in a variety of settings. This can help make personalized medicine a reality in cancer treatment.

1



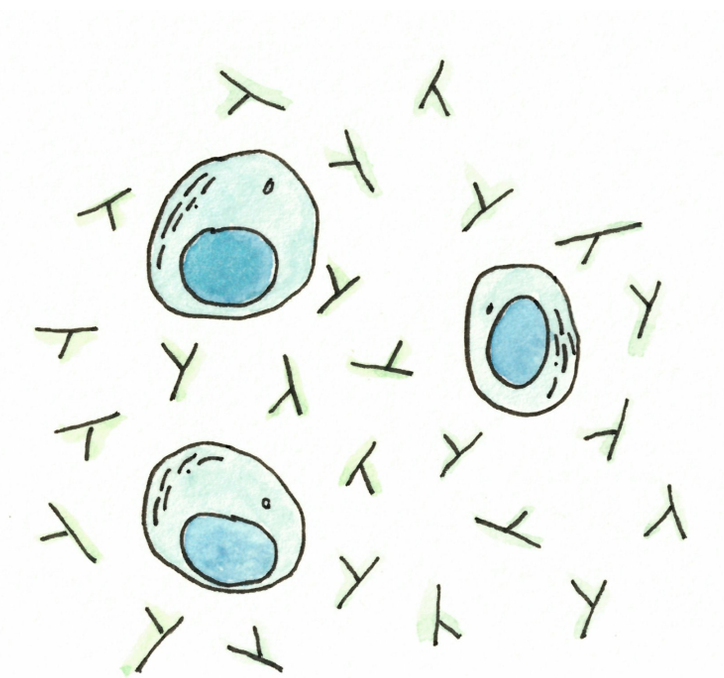
Chapter 2

Predicting treatment benefit in multiple myeloma through simulation of alternative treatment effects

Joske Ubels[1,2,3], Pieter Sonneveld[2], Erik H. van Beers[3], Annemiek Broijl[2], Martin H. van Vliet[3]*, Jeroen de Ridder[1]*

1. Department of Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands 2. Department of Hematology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands 3. SkylineDx, Rotterdam, The Netherlands

Adapted from Nat Commun 9, 2943 (2018). https://doi.org/10.1038/s41467-018-05348-5



Abstract

Many cancer treatments are associated with serious side effects, while they often only benefit a subset of the patients. Therefore, there is an urgent clinical need for tools that can aid in selecting the right treatment at diagnosis. Here we introduce Simulated Treatment Learning (STL), which enables prediction of a patient's treatment benefit. STL uses the idea that patients who received different treatments, but have similar genetic tumor profiles, can be used to model their response to the alternative treatment.

We applied STL to two Multiple Myeloma gene expression datasets, containing different treatments (bortezomib and lenalidomide). We find that STL can predict treatment benefit for both; a two-fold progression free survival (PFS) benefit was observed for bortezomib for 19.8% and a three-fold PFS benefit for lenalidomide for 31.1% of the patients. This demonstrates that STL can derive clinically actionable gene expression signatures that enable a more personalized approach to treatment.

Introduction

The successful treatment of cancer is hampered by genetic heterogeneity of the disease. Differences in the genetic makeup between tumors can result in a different response to treatment (Burrell et al. 2013). As a result, despite the existence of a wide range of efficient cancer treatments, many therapies only benefit a minority of the patients that receive them (Block et al. 2015). Because many therapies may be associated with serious adverse effects, there is a great clinical need for tools to predict - at the moment of diagnosis - which patient will benefit most from a certain treatment.

To address this, substantial efforts have been made to identify clinical and molecular markers, such as gene expression signatures, that can predict a favorable or adverse prognosis (Santos et al. 2015). Traditionally, this is achieved by defining subtypes (e.g. through unsupervised learning approaches) based on molecular markers such as genotype or gene expression. For many of these subtypes an association has been determined to survival or drug response (Lièvre et al. 2006; Bernard et al. 2009; Walther 2009).

More direct approaches use supervised learning, such as (logistic) regression, to identify markers associated with survival. In this setting, a class label is defined for each patient based on their survival or some other outcome measure, such as the risk of experiencing a relapse. The training procedure then focuses on predicting these labels as accurately as possible to ultimately produce a classifier that can predict outcome for a new patient. One of the first successful examples of such approach resulted in a 70-gene prognostic expression signature for breast cancer (Van ‘t Veer et al. 2002). A phase III clinical trial recently revealed that patients predicted to have good survival based on this signature can safely forego chemotherapy without compromising outcome (Cardoso et al. 2016), thus preventing overtreatment of these patients. These examples demonstrate that prognostic predictors can have value in predicting benefit to treatment.

Despite these successes, prognostic signatures are fundamentally limited in their ability to predict treatment benefit. This is because prognostic signatures are determined without taking treatment into account, i.e. they are not trained to distinguish patients that survive long as a result of the treatment. For this reason, patients classified in the

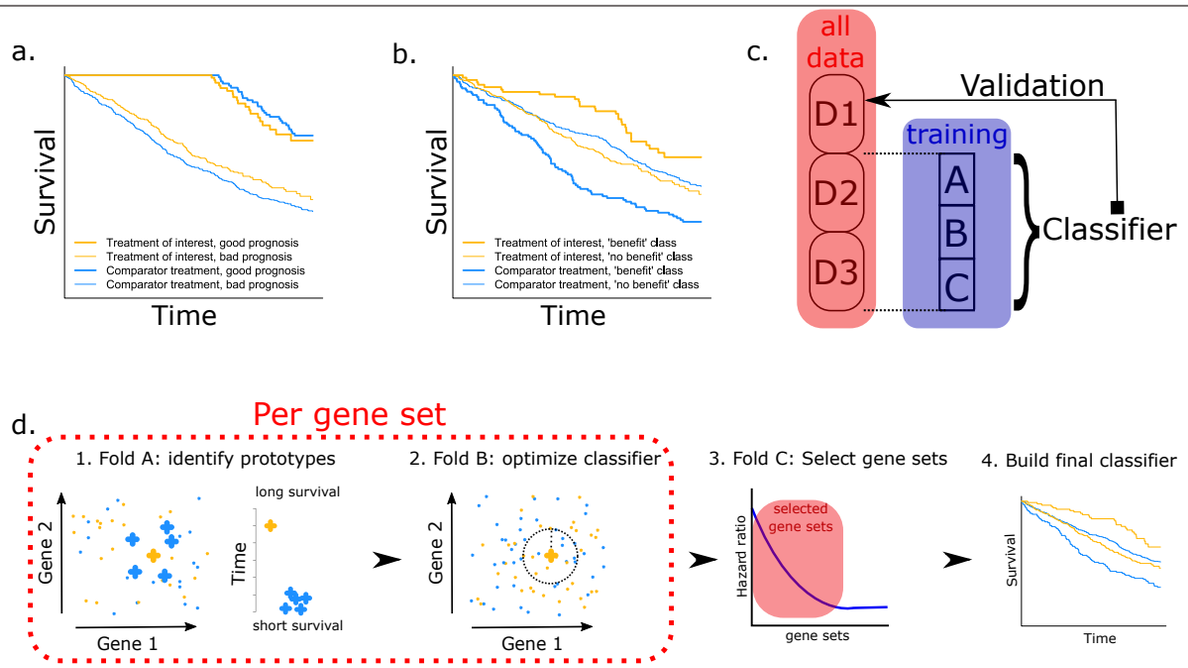


Figure 1. Illustration of the difference between prognostic and predictive classifiers and an overview of the approach **a.** Example of the Kaplan Meier curve for a prognostic classifier. **b.** Example of the Kaplan Meier curve for a predictive classifier. **c.** Division of dataset into training and test sets. D1, D2 and D3 are all used once to validate the classifier trained on the remaining two thirds of data. **d.** Flow of the GESTURE algorithm. In step 1 the prototypes with a longer than expected survival difference are identified on fold A. In step 2 the number of prototypes and corresponding decision boundary used in the classifier are optimized on fold B. In step 3 the performance of the classifier on fold C across all repeats is used to select the combination of gene sets to be used in the final classifier. In step 4 a classifier for these gene sets is defined on all training data. This classifier will be validated on the fold D not included in the training data.

'long survival' class may in fact survive just as long on any treatment available. Conversely, patients in the 'short survival' class could actually have benefit from treatment because they would have had an even shorter survival on another treatment. In **Figure 1a** and **1b** we illustrate this in the setting of a randomized trial with two treatment arms. Figure 1a shows the result for a prognostic classifier which results in a survival difference between the two classes that is similar in both treatment arms. However, to achieve treatment benefit prediction we should identify a subset of patients that specifically benefit from one of the two treatments, that is, where the difference in survival between the two treatments is larger than in the population as a whole (**Figure 1b**). It should be noted that it is possible that a prognostic classifier happens to identify a difference between treatment arms as well, but this is not an aim in the training procedure. We hypothesized that a method that is specifically geared towards

optimizing the identification of a subset of patients with a greater treatment benefit will achieve better results.

Treatment benefit is commonly measured by the Hazard Ratio (HR), which describes a patient's hazard to experience an event, for example death or progression of disease, relative to another set of patients who received a different treatment. Some recently published predictive classifiers have only shown to find a difference in response or survival between two groups of patients who all received the same treatment (Bhutani et al. 2017; Vansted et al. 2018; Ting et al. 2017). These signatures are not constructed to be predictive, since they do not necessarily provide a treatment decision; the prognosis may well be the same in every treatment group. To be truly predictive, a subgroup with a difference in survival between two treatment arms needs to be identified.

Constructing classifiers that can achieve true treatment benefit prediction thus poses a unique challenge, as it is impossible to know how a patient would have responded to the alternative treatment. As a result, class labels based which can be used to train a classifier are not available and existing classification schemes are not applicable (as demonstrated in the Results and discussion section).

To address the lack of suitable training labels, we introduce the concept of Simulated Treatment Learning (STL), a method to derive classifiers that can predict treatment benefit. STL can be applied to gene expression datasets with two treatment arms and survival data. STL uses genetic similarity, defined based on gene expression in the tumor, between patients from different treatment groups to model how a particular patient would have responded to the alternative treatment.

In this work we focus on predicting treatment benefit for Multiple Myeloma (MM), a clonal B-cell malignancy that is characterized by abnormal proliferation of plasma cells in the bone marrow. Median survival of MM patients is 5 years (Howlader et al. 2016). In the last two decades many novel therapies have been introduced for MM, resulting in an improved survival (Kumar et al. 2008; Munshi and Anderson 2013). Bortezomib and lenalidomide were crucial in achieving these improved survival rates. However, despite these advances, not all patients benefit from these novel agents and there are

insufficient tools to predict treatment response or survival. Between MM patients heterogeneity in gene expression profiles is observed (Lohr et al 2014; Keats et al. 2012). For these reasons, genetic signatures that can predict treatment benefit for MM patients are of high clinical value, making it an ideal test case for STL.

There are some preliminary indications that predictive signatures may exist for MM. Some of the various prognostic factors known in MM were later found to be predictive as well. For instance, it was shown that patients with the chromosomal aberration del17p, known to be prognostic, benefitted more from the proteasome inhibitor bortezomib than patients without del17p (Neben et al. 2012). Furthermore, expression levels of tumor suppressor RPL5, located on chromosome 1, were also found to correlate with bortezomib response (Hofman et al. 2017). Both these abnormalities have been found to be recurrently present in MM plasma cells and were later found to be prognostic and predictive. STL enables us to directly discover predictive markers, without relying on previously discovered (prognostic) markers.

We implement the STL concept in the algorithm GESTURE (Gene Expression-based Simulated Treatment Using similaRity between patiEnts), which makes it possible to derive a gene expression signature that is able to distinguish a subset of patients with improved treatment outcome from the treatment of interest, but not from the comparator treatment.
We show that GESTURE can predict treatment benefit for two major treatments in multiple myeloma, bortezomib and lenalidomide. The final classifier finds a subgroup containing 19.8% of the patients that have a two-fold progression free survival (PFS) benefit when treated with bortezomib and a three-fold PFS benefit for lenalidomide for 31.1% of the patients. Our results demonstrate that GESTURE can be used to robustly derive clinically actionable gene expression signatures that enable a more personalized approach to cancer treatment.

Results

Definition of treatment benefit class

We combined data from three randomized phase III clinical trials comprising of 910 patients with MM (see methods), who either received the proteasome inhibitor

bortezomib (n = 407) or not (n = 503). For each patient gene expression profiles were generated from purified myeloma plasma cells at diagnosis. An overall HR of 0.74 (95% CI 0.61 – 0.90, p = 0.0029, n = 910) is observed between the two treatment arms, in favor of the bortezomib arm. While this HR indicates significant treatment benefit for bortezomib, we asked whether this was driven by a small benefit for all patients, or if a subgroup of patients can be identified showing a large benefit from treatment with bortezomib, while the remainder of patients show a smaller or no benefit from bortezomib. With this research we aim to identify a subset of patients, the 'benefit' class, who benefit from the treatment of interest (bortezomib) relative to a comparator treatment arm which does not contain bortezomib. The patients not included in the 'benefit' class belong to the class 'no benefit' and would not benefit from receiving bortezomib. The classifier identifying this 'benefit' class could serve as a valuable diagnostic to determine which newly diagnosed patients would benefit from bortezomib (based) treatment.

Regular classifiers cannot predict treatment benefit

We first aimed to evaluate how well a regular (prognostic) classification approach is able to reach treatment benefit prediction. According to our definition of treatment benefit, a classifier should identify a subset of patients (class 'benefit') with a significantly better survival on the treatment of interest than the population as a whole. In a regular binary classification setting, training such classifier requires a labeled dataset, where the label indicates if the patient will or will not benefit from treatment. As discussed in the introduction, such labels are not available, since we cannot know how a patient would have responded to a different treatment. However, one reasonable assumption could be that patients who survive long in the treatment arm of interest do so because they benefited from the treatment, and, conversely, patients who survive short in the other treatment arm do so because they should have received the treatment of interest. Following this line of reasoning, we define the 'benefit' class as the 25% longest surviving patients in the bortezomib arm and the 25% shortest surviving non-bortezomib patients. Together, these two groups form the class 'benefit' (25% of all patients). All other patients from the two arms (75%) are labeled as class 'no benefit'.

Table 1 demonstrates that with some classifiers class 'benefit' can be predicted from the gene expression data reasonably well, with a cross-validation accuracy ranging from

0.58 for the random forest classifier to 0.81 for the support vector machine classifier. However, using an independent validation fold, we find that prediction of treatment benefit fails as no improvement in HR is found over the whole population. A similar absence of performance is observed when other percentages than 25% were chosen to define the class 'benefit' (**Supplementary Table 2, 3 and 4**).

The approach to derive labels directly from survival information is essentially similar to prognostic classification, and our results thus cast doubt on the utility of prognostic approaches in a predictive setting. However, this lack of performance may not be surprising, since the training labels already lead to unrealistically large HRs (<0.1), indicating that the labels are often wrong. Classifiers trained on such noisy labels are indeed unlikely to have predictive performance in independent validation data. It should moreover be noted that this approach does not take censoring of the patients into account.

As an alternative approach, we therefore also generated a large number (1000) of random labelings and evaluated the HR in the 'benefit' class of these randomly labeled datasets. Those labelings that resulted in a significant ($p<0.05$) HR below 0.5 were subsequently used to train a classifier. This greedy random search procedure enables taking into account censoring of patients (through the calculation of the HR) and leads to less extreme HRs in the training data. However, this approach also did not yield classifiers with a significant HR when applied to the validation fold **(Table 2)**. This demonstrates that it is not straightforward to derive labels for treatment benefit that can be accurately predicted from the gene expression dataset.

Overview of simulated treatment learning

The key idea of STL is that a patient's treatment benefit can be estimated by comparing its survival to a set of genetically similar patients that received the comparator treatment (**Figure 1d**, step 1). Patients with a large survival difference compared to genetically similar patients can then act as prototype patients; new patients with a similar gene expression profile are expected to also benefit from receiving the treatment of interest. Since similarity in gene expression profile is greatly influenced by the choice of input genes, we define this similarity according to a large number of gene sets. Training the prototype-based classifier requires optimizing two parameters per gene set: the number of prototypes to use and the decision boundary, defined in terms of the

Table 1. Classification accuracy in cross validation and HR in independent validation for the classifiers trained on labels based on the top 25% surviving bortezomib patients and the bottom 25% non-bortezomib patients.

	Classification accuracy	Validation HR	p-value
Nearest mean	0.58 (sd: 0.07)	0.96 (95% CI: 0.57 - 1.60)	0.86
Random forest	0.68 (sd: 0.03)	0.95 (95% CI: 0.54 – 1.68)	0.87
SVM	0.81 (sd: 0.06)	0.81 (95% CI: 0.31 – 2.13)	0.67

Table 2. Classification accuracy in cross validation and HR in independent validation for the classifiers trained on labels selected from randomly generated classifications with a significant HR under 0.5

	Classification accuracy	Validation HR	p-value
Nearest mean	0.50 (sd: 0.02)	0.81 (95% CI: 0.49 – 1.35)	0.42
Random forest	0.66 (sd: 0.02)	0.81 (95% CI: 0.50 – 1.41)	0.51
SVM	0.83 (sd: 0.06)	1.10 (95% CI: 0.52 – 2.34)	0.80

Euclidean distance to the prototype (**Figure 1d**, step 2). The STL classifier also needs to select the optimal gene sets to ultimately classify a patient. Importantly, the labels are now defined using the prototypes identified for the various gene sets, which means that in the STL approach there is no need to define labels before training the classifier. To train the classifier and select the best performing gene sets, the training data are split in three folds (A, B and C). Fold A is used to identify prototypes, fold B to optimize the decision boundary and fold C to estimate classifier performance.

To obtain unbiased estimates of the overall prediction performance, the entire dataset is divided in three equal folds, D1, D2 and D3, ensuring a similar HR between the treatment arms in all three folds. Training is performed on two folds, while the remaining fold is kept separate to serve as an independent validation set. This is rotated to obtain an unbiased prediction for each fold. The division of the data in D1, D2 and D3, and subsequently in folds A, B and C is shown in **Figure 1c**.

It is a priori unknown which genes will be relevant to defining patient similarity and predicting treatment response. We used 10,581 functionally coherent gene sets based on Gene Ontology annotation. Each gene set is used to train a separate classifier. The top-performing classifiers are subsequently combined into an ensemble classifier to determine the optimal number of gene sets to be used in the final classifier (**Figure 1d**, step 3, for details see Methods). For the gene sets included in this optimal number a single classifier is trained using all the training data. These classifiers are combined into the final ensemble classifier that is used to classify the patients in the validation set (**Figure 1d**, step 4).

STL finds a predictive classifier for bortezomib benefit

Figure 2a shows the cumulative progression free survival curves for two treatment arms, with an HR of 0.74 (95% CI 0.61 – 0.90, p = 0.0029, n = 910) between the treatment arms. **Figure 2b** shows the treatment arms and classes as identified by the STL classifier, when combining the class 'benefit' from the three validation folds. These three validation folds together comprise the whole dataset; the classification of each validation fold is predicted by separately trained classifiers. This enables us to show a validation performance for the whole dataset. The validation HRs for the 'benefit' and 'no benefit' class are 0.50 (95% CI 0.32 – 0.76, p = 0.0012, n = 180) and 0.78 (95% CI

0.63 – 0.98, p = 0.03, n = 730), respectively. In the entire population an HR of 0.74 (p = 0.0029, n = 910) is observed. These results show that a subgroup, comprising 19.8% of the population (n=180 out of 910), is identified by our method that benefits substantially more from bortezomib treatment than the population as a whole.

More importantly, the STL approach is able to discover and predict this subgroup using the gene expression data at diagnosis. In the bortezomib arm, the 'benefit' and 'no benefit' class exhibit similar survival curves. This is expected, since our classifier is trained to predict benefit with respect to the patient group not receiving bortezomib. As the Kaplan Meier in **Figure 2b** shows, the other treatment arm in the 'no benefit' class also has a similar survival, which means we expect these patients would have had a similar survival had they not received bortezomib. The ability to determine that a patient would not benefit from bortezomib is of equal importance as predicting benefit; preventing unnecessary treatment is an important aim of personalized medicine.

The HRs observed within each of the individual validation folds are similar to the HR obtained when combining all folds (0.51 (95% CI 0.28 – 0.92, p = 0.03, n = 89,), 0.39 (95% CI 0.14 – 1.08, p = 0.07, n = 30,) and 0.46 (95% CI 0.21 – 1.02, p = 0.06, n = 61) in folds D1, D2 and D3 respectively). We note that the HR is comparable in all folds, demonstrating a stable performance, although not statistically significant for fold D2 and D3 at $p < 0.05$ due to the fact that in D2 9.9% of patients and in D3 20.1% are included in the 'benefit' class and versus 29.4% in D1.

Traditionally, the performance of a classifier is assessed by computing its accuracy, which is done by comparing the labels predicted by the classifier with ground truth labels. Ground truth labels are labels that are known to be accurate because they can be directly observed, e.g. if a patient survives longer than 5 years or not. Since we do not know beforehand which patients benefited from bortezomib, we have no ground truth labels available and cannot compute the accuracy of our classifier. However, we can compare the class labels obtained with the three separate classifiers when applied to all 910 patients. We find that these three class assignments agree between the classifiers significantly more than expected by chance (i.e. 0/3 classifiers or 3/3 classifiers predict benefit; **Supplementary Figure 1**). A similar conclusion is reached by comparing the classification scores directly, which significantly correlate (all p-values < $1*10^{-4}$).

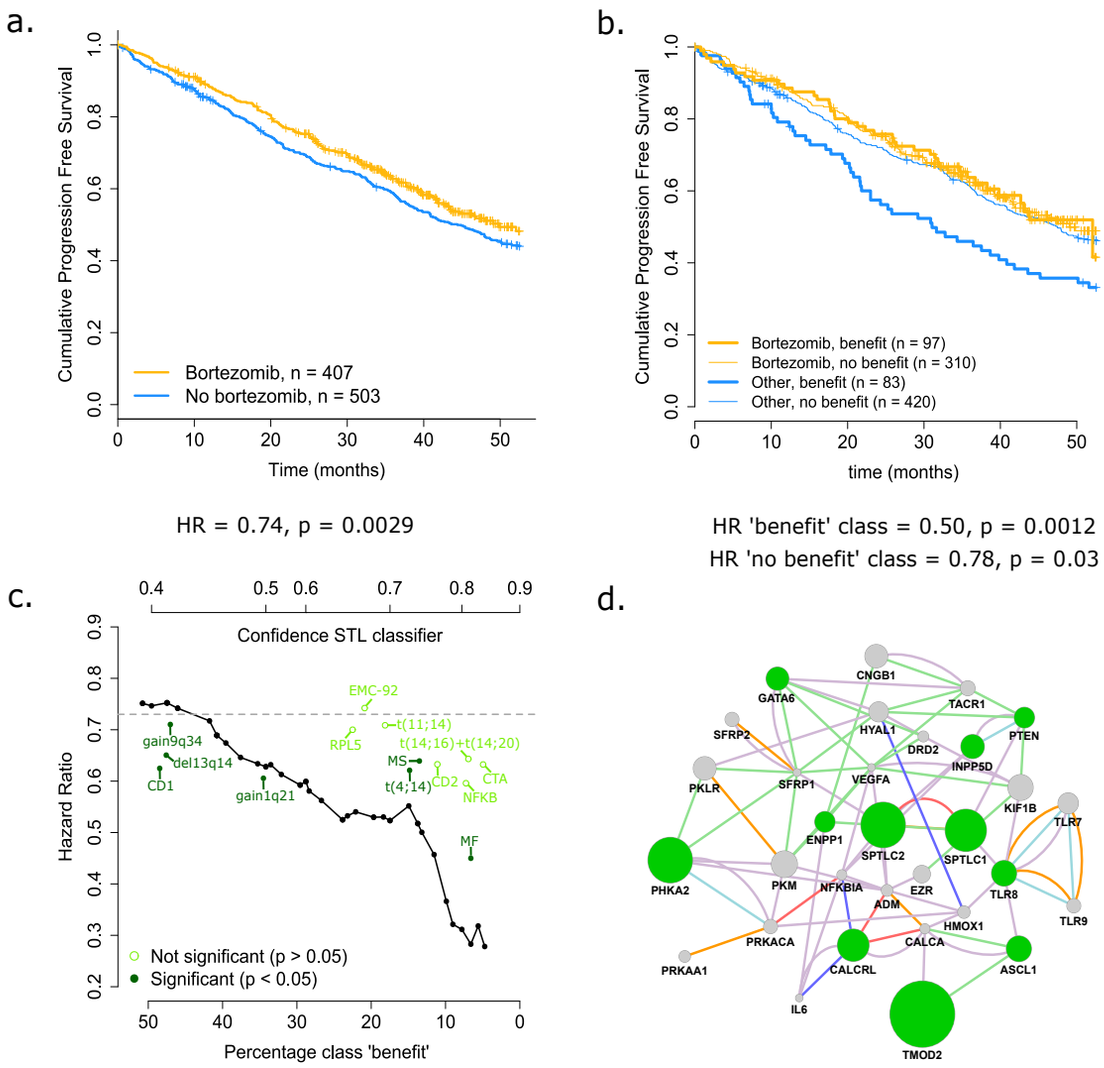


Figure 2. Overview of the bortezomib classifier results and comparison to known markers. **a.** Kaplan Meier of the entire bortezomib dataset, showing a HR of 0.74 (95% CI 0.61 – 0.90, p = 0.0029, n = 910,) between the treatment arms. **b.** Kaplan Meier of the combined classifications into a 'benefit' and 'no benefit' class of D1, D2 and D3. A HR of 0.50 (95% CI 0.32 – 0.76, p = 0.0012, n = 180,) is found between the treatment arms in the 'benefit' class and a HR of 0.78 (95% CI 0.63 – 0.98, p = 0.03, n = 730) in the 'no benefit' class. These results show that a subgroup, comprising 19.8% of the population (n=180 out of 910 total), is identified by our method that benefits substantially more from bortezomib treatment than the population as a whole; in the entire population an HR of 0.74 (95% CI 0.61 – 0.90, p = 0.0029, n = 910) is found. **c.** The HR found in the 'benefit' class (y-axis) when different operating points (x-axis) are used, compared with known predictive and prognostic markers. The gray dotted line indicated the HR found in the entire dataset, without classification. **d.** Relationships between the 31 genes in common between the D1, D2 and D3 classifiers. Node size corresponds to how much more a gene was observed in the selected gene sets than expected. Green nodes indicate that the gene is associated with a p-value < 0.05. Relationships are inferred from literature with the GeneMANIA algorithm (Warde-Farley et al. 2010). A purple edge indicates the genes are co-expressed, a green edge indicates a genetic interaction, a red edge a physical interaction, an orange edge a shared protein domain, a dark blue edge indicates co-localization and a light blue edge shows that both genes are annotated to the same pathway.

When considering the cases for which the three classifiers agree, we find that 503 patients are consistently classified as 'no benefit' and 57 patients as 'benefit'. Together, this demonstrates that, even though the classifiers do not agree on the class assignment for all patients (which is expected in practice for classifiers with less than 100% accuracy), they capture the same gene expression patterns.

The decision boundary of the classifiers are defined by the parameters k and gamma and a threshold T. We optimize the combination of k and gamma by an exhaustive grid search. We verified that the performance of our classifier is robust to small changes in these parameters (**Supplementary Note 1).** The operating point of the classier is determined by the number of individual classifiers in the ensemble that agree on the class label, and is thus directly related to the confidence of the ensemble classifier about the label 'benefit'. To ensure sufficient power and provide a treatment decision for a substantial group of patients, the operating point of the classifier was set to 20% in training (see methods). At this operating point, 19.8% of patients in the validation folds were actually assigned to the 'benefit' class. **Figure 2c** depicts the HR as a function of the confidence level of the classifier. We observe that, for higher confidence levels (yielding smaller sizes of the 'benefit' class) more extreme validation HRs are observed, demonstrating that there is a direct relation between classifier score and treatment benefit. This is consistent with the fact that the highest HR and largest class 'benefit' are found in fold D1 in validation, while the lowest HR and the smallest class 'benefit' are found in D2.

As a control experiment, we also ran the algorithm with shuffled treatment labels, destroying the relationship between the gene expression and the treatment specific survival. As expected, the classifier trained on this data shows no performance in the validation data, achieving an HR of 1.09 (95% CI 0.71 – 1.67, p = 0.69, n = 167) in the class 'benefit' and an HR of 0.95 (95% CI 0.77 – 1.18, p = 0.65, n = 743) in the class 'no benefit' (**Supplementary Figure 3**). This reinforces our observation that STL identifies a true effect, since the classifier shows no performance in random data.

STL classifier outperforms known markers

We compared the HRs found using the STL classifier with several known prognostic markers in MM, some of which also show predictive value (**Figure 2c**). The STL

classifier has a superior performance for operating points that result in assignment of up to 30% of the patients to the class 'benefit'. The markers that slightly outperform the STL classifier do so only for operating points that results in much larger sizes of the class 'benefit' and lead to smaller effect sizes. The grey line indicates the baseline HR found in the entire dataset. A clinically actionable classifier should reach a substantially larger benefit than this baseline, which is only attained by the STL classifier and the MF cluster for operating points <30%, where the STL classifier outperforms the MF biomarker.

Biological information is important for performance

To investigate if the biological knowledge contained in the Gene Ontology, used to define gene sets, truly aids classification performance, we also tested random gene sets with the same set size distribution. Using the random gene sets, final classification results in a significant HR of 0.56 (95% CI 0.34 – 0.90, p = 0.02, n = 148) when all three validation folds are combined (**Supplementary Figure 2**). This is not unexpected as combining random feature sets in an ensemble classifier is known to achieve good classification performance (Breiman 2001). Moreover, it has been shown previously that random gene signatures can perform on par in a prognostic setting (Venet et al. 2011). Nonetheless, the STL classifier trained using the GO gene sets outperforms the random gene set approach in both HR and p-value. Moreover, in contrast to the relatively stable performance across validation folds when using the GO gene sets, the performance of the random set approach varies greatly between the folds, ranging from an HR of 0.76 (95% CI 0.32 – 1.85, p = 0.55, n = 41) in D1 to an HR of 0.44 (95% CI 0.21 – 0.93, p = 0.03, n = 67) in D3.

Together, this demonstrates that the biological information contained in the Gene Ontology gene sets is important to the performance of the STL classifier.

Genes used to predict treatment benefit bortezomib

The classifiers built for D1, D2 and D3 use respectively 113, 218 and 111 GO gene sets to predict bortezomib benefit, encompassing a total of 1913 unique genes. There are 31 genes used in all three classifiers (**Figure 2d**). There are GO categories that include a large subset of these 31 genes, including "positive regulation of transcription from RNA polymerase II promoter", "cellular response to hypoxia" and "negative regulation of the apoptotic process". All these GO categories are associated with the pathogenesis of cancer. Both increased proliferation and the ability to evade apoptosis are hallmarks of

cancer (Hanahan and Weinberg 2011). It has also been established that cancer cells can adapt their metabolism to thrive in hypoxic conditions (Eales et al. 2016). For the 31 genes, we calculated they are selected more than expected by chance. GO sets are hierarchical (i.e. there is a larger parent category that can include several children categories) and genes can be annotated to multiple GO categories. Therefore, we have taken into account how many GO categories include a certain gene to establish if we observe a gene more often than expected in our classifiers. The expected count for a gene is based on the number of GO categories that include that gene, e.g. *PTEN* is included in 123 of the 10,581 gene sets, so in the 442 gene sets used across D1, D2 and D3 we would expect to observe *PTEN* approximately 5 times if it would occur at the same frequency as within our selected gene sets. Most genes in common between the three classifiers are observed more often than expected (degree of overrepresentation indicated by node size in **Figure 2d**), with 11 of 31 significantly overrepresented ($p < 0.05$). The most overrepresented genes are *TMOD2, PHKA2, SPTCL1* and *SPTCL2*. None of these genes are known to be associated with MM or response to bortezomib. However, investigation of the proteome of a cell line carrying a *SPTCL1* mutation showed an increased presence of Ig kappa chain C (Stimpson et al. 2015). Immunoglobulin light chain presence is used as a biomarker for MM and has been identified as a risk factor for progression (Dispenzieri et al 2008). *PTEN* is also found to be significantly overrepresented. *PTEN* is a known tumor suppressor and was found to be mutated in a various cancers (Yamada and Araki 2001). In MM, *PTEN* mutations are relatively uncommon and associated with advanced disease (Chang et al. 2006).

Impact of dataset of origin on validation performance

Our training dataset is a combination of three different datasets: Total Therapy 2, Total Therapy 3 (together forming the TT dataset) and HOVON65/GMMG-HD4 (H65). Both the bortezomib and the no bortezomib arm contain more than one treatment regimen (**Supplementary Table 1**). We trained and validated on a combination of the datasets (see Methods). To investigate the contribution of the different datasets to the final validation performance, we calculated the HR in class 'benefit' for the TT and H65 patients separately. Reassuringly, we observe a similar effect in class 'benefit' in both datasets, albeit not significant due to small sample size in the H65 dataset (HR = 0.69 (95% CI 0.36 – 1.32) , p = 0.26, n = 49, for H65 and HR = 0.38 (95% CI 0.21 – 0.69), p = 0.002, n = 131 for TT, **Supplementary Figures 4 and 5).** Also, the observed HR is much

smaller in the TT dataset. This may be expected, since the HR in the overall population is also smaller in TT than in H65 (the overall HR in TT is 0.62 (95% CI = 0.46 – 0.84), p = 0.002, n = 583 vs. an HR of 0.86 (95% CI = 0.66 – 1.13), p = 0.28, n = 327 in H65).
We hypothesized that heterogeneity helps to prevent overfitting to one specific dataset or treatment regimen. To test this, we also performed a cross validation within the two TT datasets only (the H65 dataset is too small for this with n = 327). Subsequently, we trained a classifier on the entire TT dataset (combining Total Therapy 2 and Total Therapy 3) and validated on H65. Cross validation within the TT dataset leads to an HR of 0.28 (95% CI 0.13 – 0.60, p = 0.00098, n = 86) in class 'benefit' and an HR of 0.71 (95% CI 0.51 – 0.98, p = 0.038, n = 497) in class 'no benefit' **(Supplementary Figure 6)**, which is a substantial improvement over the classifier trained on the combined dataset. In contrast, when the classifier is trained on the entire TT dataset, no performance is observed in the H65 dataset (an HR of 1.13 (95% CI 0.63 – 2.04), p = 0.68, n = 66 in class 'benefit' and 0.81 (95% CI 0.60 – 1.1), p = 0.18, n = 261 in class 'no benefit'), indicating that some dataset specific fitting has occurred. Importantly, dataset specific fitting does not necessarily indicate overtraining; the classifiers still validate on the completely independent hold out validation fold. These results do suggest that it is very important to match the training population with the population one intends to use the classifier in. If the population in which the classifier is intended to be applied is heterogeneous, the training dataset also needs to reflect this heterogeneity.
In the MM dataset under study here, one possible explanation for the lack of validation of the TT- based classifier on the H65 data is that the TT trials were conducted in the USA and included more additional treatment than the European H65 trial (see **Supplementary Table 1** for treatment details)**.** When the STL classifier is trained exclusively on the TT datasets, it could become specifically predictive for the TT regimen, rather than bortezomib, explaining why this classifier does not show a satisfactory performance in H65. When trained on the mixed dataset, the classifier does show performance in the H65 dataset, but still performs better within the TT dataset, which makes up a bigger part of the training data.

STL finds a predictive classifier for lenalidomide benefit

The STL method was developed based on the bortezomib dataset. Even though a strict separation of training and validation has been made, we cannot exclude the possibility of 'experimenter bias' (Holman et al. 2015), which is the result of making experimental

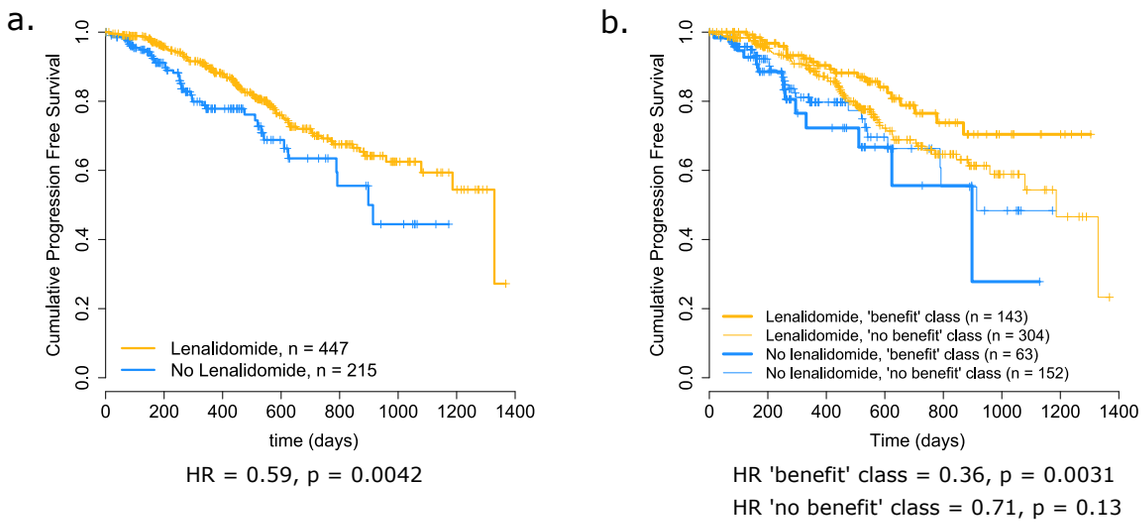


Figure 3. Overview of the lenalidomide classifier results **a.** Kaplan Meier curves for the entire lenalidomide dataset, showing an HR of 0.59 (95% CI 0.41 – 0.84, p = 0.0042, n = 662) between the treatment arms. **b.** Kaplan Meier curve of the combined classifications into a 'benefit' and 'no benefit' class of D1, D2 and D3. An HR of 0.36 (95% CI 0.18 – 0.71, p = 0.0031, n = 206) is found between the treatment arms in the 'benefit' class and an HR of 0.71 (95% CI 0.46 – 1.10, p = 0.13, n = 456) in the 'no benefit' class

choices based on the results on the training dataset and which can lead to a classifier that will only perform well on the specific dataset at hand.

To demonstrate that the STL method is not biased to just one dataset we applied it to a completely independent dataset obtained from the CoMMpass database (https://research.themmrf.org/). CoMMpass contains data from an observational MM study, meaning the trial did not interfere with the treating physician's choice of treatment. This is a good model for the setting in which an eventual predictive biomarker would be applied. Moreover, instead of microarrays, RNA-seq was used to obtain gene expression measurements, thus providing an additional axis of variation compared to microarray data. Overall, gene expression data and annotation was available for 662 patients, 447 of which received lenalidomide in the first line and 215 did not. An overall HR of 0.59 (p = 0.004) in favor of lenalidomide was observed, as seen in the Kaplan Meier in **Figure 3a.**

Similar as before, the dataset was divided into three equal folds and STL obtains classifiers that successfully predict benefit in all folds. Since the CoMMpass dataset is

smaller than the bortezomib dataset used before, we required the 'benefit' class to contain at least 30% of the patients, to ensure sufficient power. This results in a combined HR of 0.36 (95% CI 0.18 – 0.71, p = 0.0031, n = 206) over the entire dataset, as shown in **Figure 3b**. In total 31.1% of patients were classified as class 'benefit'. Again, the STL classifier was able to distinguish a subset of patients with significant treatment benefit in each fold with HRs of 0.27 (95% CI 0.07 – 1.06, p = 0.06, n = 72), 0.39 (95% CI 0.11 – 1.41, p = 0.15, n = 66) and 0.40 (0.14 – 1.15, p = 0.09, n = 68) in D1, D2 and D3, respectively. This demonstrates that STL also successfully identified a predictor for lenalidomide benefit.

Genes used to predict treatment benefit lenalidomide

The predictive classifiers for lenalidomide use 47, 5 and 119 gene sets in D1, D2 and D3 respectively, encompassing 3723 unique genes. Out of these, 5 genes are used in all three classifiers: *CYP11B2, SHH, HGNC, CAV1* and *SMO*, all of which are observed more frequently than expected. *SHH* and *CYP11B2* are significantly overrepresented ($p < 0.05$). *SHH* is a crucial part of the hedgehog signaling pathway, which has been previously found to play an important role in the pathogenesis of MM (Blotta et al. 2012). Neither of these genes has previously been associated with lenalidomide response, possibly representing an undiscovered mechanism influencing lenalidomide response in MM patients.

Discussion

Simulated Treatment Learning addresses an urgent clinical need because response rates to current cancer therapies are often poor and moreover frequently accompanied with serious side effects. STL offers an important step towards realistic personalization of cancer medicine administration by identifying gene expression markers that can be used to determine the most effective treatment for a cancer patient at the moment of diagnosis.

The STL classifier was successfully tested across different gene expression platforms, different treatments and different study types, demonstrating that STL is more generically applicable than one particular dataset. Since our work has focused on MM, an important next step is to investigate if STL is also successful in unraveling treatment

benefit for other diseases. If so, STL can play an important role in rescuing treatments that do not achieve a significant effect in the entire patient population but may still benefit a subset of the patients. For instance, STL can be an important post-hoc analysis for phase III clinical trials of novel treatments that have missed their endpoint, such as, for instance, nivolumab in the CheckMate-026 trial (Socinki et al. 2016). We do note that STL requires a relatively large number of samples to build the classifier, which may not always be available when a novel treatment first enters clinical trials. The generic concept of STL can be readily extended to include patient similarity definitions based on e.g. germline or somatic genomic profiles and other types of outcome measure such as categorical or binary measures.

Methods

Data and processing

We pooled gene expression and survival data from three phase III trials: Total Therapy 2 (TT2, GSE2658) Total Therapy 3 (TT3, GSE2658) and HOVON-65/GMMG-HD4 (H65, GSE19784). The TT2 dataset included 345 newly diagnosed multiple myeloma (NDMM) samples, treated either with thalidomide and melphalan (n = 173) or melphalan alone (n = 172). Average age is 56.3 (range: 24 - 76) and 57.1% of the patients is male. The TT3 dataset included 238 NDMM samples treated with bortezomib, thalidomide, dexamethasone, cyclophosphamide, cisplatin and etoposide (VTDPACE). Average age is 58.7 (range: 32 - 75) and 67.6% is male. The H65 dataset included 327 NDMM samples, treated either with vincristine, doxorubicin and dexamethasone (VAD, n = 158) or bortezomib, doxorubicin and dexamethasone (PAD, n = 169). Average age is 54.7 (range: 27 - 65) and 56.4% percent is male. In our analyses of the pooled data two treatment arms were considered: a bortezomib arm, which comprises the PAD arm from H65 and TT3, and a non-bortezomib arm, which comprises the VAD arm from H65 and TT2. Combined, these datasets include 910 patients, of which 407 received bortezomib and 503 did not.

All samples were profiled with the Affymetrix Human Genome U133 plus 2.0 array. Gene expression was MAS5 and log2 normalized. Batch effects resulting from pooling different datasets were corrected with ComBat (Johnson et al. 2007). Data was scaled

to mean 0 and variance 1 per probeset. Probesets with a variance of < 1 before scaling were discarded.

The data was split in fold D1 (303 samples), fold D2 (303 samples) and fold D3 (304 samples), stratifying for treatment arm and survival. Fold D1 is not used at any point in the training and serves as validation data, while Fold D2 and fold D3 are combined to serve as training data. After the STL classifier is successfully validated on fold D1, the folds are rotated to serve as additional validation folds to assess robustness. The training data for fold D2 consists of D1 and D3 and the training data for D3 consists of D1 and D2 (specification of which samples were used in which folds is available with the code in the GitHub repository).

After developing the STL method on the microarray dataset, we also applied it to the CoMMpass trial (NCT0145429) dataset generated by the Multiple Myeloma Research Foundation (MMRF). For 662 patients both RNAseq, survival data, and treatment information was available. Sequencing data is processed with the Cufflinks pipeline (researcher.themmrf.org). The dataset was split into a treatment arm where patients received lenalidomide as first line treatment (n = 447) and an arm where patients did not (n = 215). This data was also split into folds D1 (220 samples), D2 (221 samples) and D3 (221 samples), specification of which samples were used in which folds is available with the code in the GitHub repository.

Endpoint and survival analysis

Progression Free Survival (PFS) was used as endpoint, as this is the most direct readout of first line treatment related survival and therefore considered to be more relevant compared to overall survival. PFS times in the TT2 and H65 datasets were truncated to 52.53 months, corresponding to the longest follow-up time in the TT3 dataset.

Survival analyses were done using the Cox Proportional Hazards model (survival package, version 2.38.4)(Therneau 2015). For the microarray data, the survival analysis included a stratification for dataset of origin. This means the base hazard was estimated separately for the TT2/TT3 dataset and the H65 dataset. This is necessary to correct for the significant survival difference found between these datasets. Hazard Ratios (HR) and associated 2-sided p-values were calculated. P-values below 0.05 were considered

statistically significant. All HRs are computed as bortezomib vs no bortezomib and lenalidomide vs no lenalidomide, which means an HR below 1 signifies a benefit when receiving bortezomib or lenalidomide. All calculations were performed in R version 3.1.2.

Gene sets

For the bortezomib classifier we tested all Gene Ontology (GO) categories, as defined by the R Bioconductor package hgu133plus2.db (Carlson 2016)(accessed: 27 October 2015), with two or more probesets associated to them. This resulted in 10,581 gene sets. To test whether the biological information, contained in the GO annotation, aids the performance of the algorithm, 10,581 random gene sets matching the size of the actual selected GO categories were also tested.
For the lenalidomide classifier we tested all the GO categories with two or more genes associated to them, as defined by Bioconductor package biomaRt (Durinck et al 2009)(accessed: 19 June 2017). This resulted in 9,121 gene sets.

Algorithm

The STL classifier aims to predict if a patient does or does not benefit from a certain treatment of interest based on the gene expression profile of the patient. In order to train this classifier, a gene expression dataset is required that consists of two treatment arms and a continuous outcome measure. These data are first split into training and validation folds. The training data comprises of two thirds of the data, while one third (fold D) is kept apart to function as validation data. We define three separate folds D (D1, D2 and D3), such that each patient is included in the validation set once. The training data is subsequently split further into folds A, B and C for training.

We first define a ranked list of prototype patients on fold A (Step 1) that exhibit a better than expected prognosis on the treatment of interest compared to a set of genetically similar patients that received an alternative treatment. In Step 2, a decision boundary around a selection of prototype patients is determined on fold B. Patients that lie within this decision boundary are expected to show a favorable outcome when receiving the treatment of interest and are classified as benefitting (class 'benefit'). All other patients are considered class 'no benefit' and are not expected to benefit from receiving the treatment of interest. Because it is a priori unknown based on which genes patient

similarity should be defined, step 1 and 2 are performed for a large number of functionally coherent gene sets obtained from the Gene Ontology annotation, yielding one classifier per gene set. Step 1 and 2 are repeated 12 times to obtain a robust estimate of the performance per gene set. In each repeat, the training data is split into a different fold A, B and C. The performance is defined as the Hazard Ratio (HR) between treatments in class 'benefit', found in a fold C, which contains samples that were not used in step 1 and 2. All gene sets are ranked by their mean performance in fold C across repeats. In Step 3 we determine the optimal number of gene sets to combine into a final classifier. We found that defining performance and selecting the optimal number of gene sets on the same folds C leads to overtraining. Therefore, we run the entire algorithm a second time (Run 2), using 12 new repeats with different splits into fold A, B and C. The first run of 12 repeats is used to rank the gene sets. The combined performance of these ranked gene sets on the folds C from Run 2 is used to determine the optimal number *s* of gene sets. Similar to the boosting principle (Schapire 1999), the individual classifiers are combined into an ensemble to construct a more robust final classifier. The performance of this combined classifier is measured on fold C of Run 2. The gene sets are added to the classifier in order of their ranking, until an optimal performance is reached across all the repeats from Run 2. Since there are 12 repeats, each combination results in 12 HRs as measured on the folds C from run 12. To determine the optimal number of gene sets, we fit a local polynomial regression line on the median HRs for each combination of gene sets. The optimal number of gene sets *s* is reached when adding a gene set does not result in a lower HR. We then rank the gene sets based on their individual performance across the folds C of Run 2 and select the top *s* for inclusion in the final ensemble classifier. Finally, in Step 4, one final classifier is trained using the entire training dataset for these selected gene sets.
These steps are visualized in **Figure 1d** and are described in more detail below.

In Step 1, we perform prototype ranking on Fold A. For each patient receiving the treatment of interest, the treatment benefit is defined as

$$\Delta PFS_i = \frac{1}{n}\sum_{j \in O}(PFS_i - PFS_j), \quad (1)$$

where O is the set of the *n* most similar patients (based on Euclidean distance) that did not receive the treatment of interest. We use $n = 10$. In an approach similar to Harrell's C-statistic (Harrell et al. 1996), ΔPFS is only calculated for neighbor pairs where it is clear which patient experienced an event first; if both are censored or one patient is censored before the neighbor experienced an event, ΔPFS is not computed. When $n = 10$ is used, this on average leads to 7 neighbours being used in the calculation of ΔPFS. To correct for the fact that a patient with a long survival time will, on average, have a large ΔPFS irrespective of its relative treatment benefit compared to genetically similar patients, we define the z- normalized zPFS score as:

$$zPFS_i = \frac{\Delta PFS_i - \mu(RPFS_i)}{\sigma(RPFS_i)}, \quad (2)$$

where RPFS is a distribution of 1000 random ΔPFS scores, obtained by calculating ΔPFS for randomly chosen sets O, i.e. determining treatment benefit with respect to random patients instead of genetically similar patients. Based on the zPFS score all patients in fold A that were given the treatment of interest can be ranked.

In Step 2, we define the classifier on fold B. The classifier is defined by a subset of k top-ranked prototypes along with a decision boundary defined in terms of the Euclidean distance γ around a prototype. A patient is classified as class 'benefit' when it lies within γ of any of the top k prototypes. The optimal values for k and γ are those resulting in the lowest Hazard Ratio (HR) in class 'benefit' (the patient group in which the treatment of interest should have a better survival). We set an operating point that additionally constrains k and γ, such that class 'benefit' comprises at least a certain percentage of the dataset. This ensures sufficient statistical power to compute the significance of the HR in the 'benefit' class. The number of prototypes was restricted to 10 to prevent defining an extremely complicated classifier. The search grid for parameter γ was made dependent on the local density of the neighbors, and consisted of the sorted list of Euclidean distances between the prototype and its neighbors. The optimal k and γ combination is chosen so that the HR in class 'benefit' is minimal, while still associated with a p-value below 0.05. If no combination results in a p-value below 0.05, the minimal non-significant HR that results in a class 'benefit' of sufficient size is chosen.

In step 3, we rank and select the gene sets. First, the gene sets are ranked by their mean performance in fold C over all repeats from Run 1. After ranking, we run the algorithm a second time, with different divisions into fold A, B and C. We add gene sets to an ensemble classifier one by one based on this ranking. The performance of the combined gene sets is measured on each fold C of this second run. We find that defining the ranking on different folds than we use to measure combined performance prevents overtraining, although some bias is still expected to occur. Since the found HR can fluctuate between folds and gene set numbers, a regression line is fit through the median HRs found on folds C in the second run and the optimal number of gene sets is determined: the first combination of gene sets for which adding another gene set does not lead to an improvement of the HR larger than $1*10^{4}$.

After the optimal number of gene sets is determined in Step 3, the final classifier is defined in Step 4. The gene sets are ranked based on their mean performance in fold C in the second run. The top scoring gene sets are selected and for these gene sets a final classifier is trained. To this end, the complete training dataset is split into only two folds, since the third fold is no longer required. The classifiers defined by different gene sets are combined into an ensemble classifier by an equally weighted voting procedure, which means each classifier has an equal influence on the final classification. For an ensemble classifier containing *s* gene sets, this defines a classification score between 0 and *s* per patient. This score is thresholded by threshold T, which determines whether a patient is to benefit from the treatment of interest, where a patient with a score below the threshold is classified as not benefitting from treatment ('no benefit' class). The optimal threshold T is the one for which the HR between treatments is minimal in class 'benefit'. This combination of classifiers and threshold can be used to classify new and unseen patients and is validated on fold D.

Calculating overrepresentation of genes in the classifier

The same gene can be used multiple times in a single classifier and/or multiple times across the classifiers obtained for fold D1, D2 and D3. Both cases provide evidence of the importance of the gene for the treatment benefit prediction. To assess whether genes are selected more frequently than expected by chance across all three classifiers, we determine the degree of overrepresentation by dividing the observed count by the expected count. The expected count is calculated by $p * W$ where p is the fraction of the

gene sets containing the gene and *W* the total number of gene sets selected across all three classifiers. A p-value is determined using the binomial test.

Training regular classifiers

We defined the labels that were used to train the regular classifiers in two ways. First, labels were defined by assigning the 25% longest surviving bortezomib patients and the 25% shortest surviving non-bortezomib patients to the 'benefit' class and all others to the 'no benefit' class. A classifier was trained using folds A-C to predict these labels, using the HR in validation fold D1 as performance measure of the predictive power. For the nearest mean classifier, a double-loop cross-validation was used to optimize the number of genes (ranked based on t-score), using balanced accuracy as the performance measure.

A random forest classifier (R package randomForest, version 4.6.12)(Liaw and Wiener 2002) and a support vector machine (R package e1071, version 1.6.7)(Meyer et al. 2015) were also trained. For both these classifiers, the number of genes was optimized in cross validation. For the random forest classifier 2000 trees were trained per classifier and the bootstrap sample was sampled equally from both classes, to prevent the classifier being affected by the class imbalance. For the support vector machine, C-values from 1 to 100 were tested, in steps of 1. The gamma used is 1/P, where P is the number of input variables, i.e. the number of genes.
For all classifiers, the accuracy reported is the mean accuracy in cross validation for the optimal number of input genes.

Comparison with known prognostic markers

To the best of our knowledge, RPL5 is the only published gene expression based marker that predicts bortezomib benefit by comparing to another treatment group (Hofman et al. 2017). We tested RPL5 on the data from the Total Therapy studies, since it was trained on the HOVON-65 data. Since some predictive markers are discovered by testing markers previously known to be prognostic, we also compare with prognostic markers. FISH markers were called on the gene expression data, using previously developed classifiers (Van Vliet et al. 2013), since FISH data was not available for all patients. Unfortunately, there is no reliable gene expression classifier for del17p. We

tested if any predictive information was available in previously defined molecular subtypes in MM (Zhan et al. 2006) and in the prognostic gene signature EMC-92 (Kuiper et al. 2012).

Data availability

All survival and treatment data included in the bortezomib dataset are supplied in Supplement 1. The gene expression data from the Total Therapy II and Total Therapy III studies are accessible in the GEO database, accession number GSE2658. The gene expression data from the HOVON-65/GMMG-HD4 study is accessible in the GEO database, accession number GSE19784.

All survival, treatment and RNAseq data used for the lenalidomide dataset is accessible at research.themmrf.org.

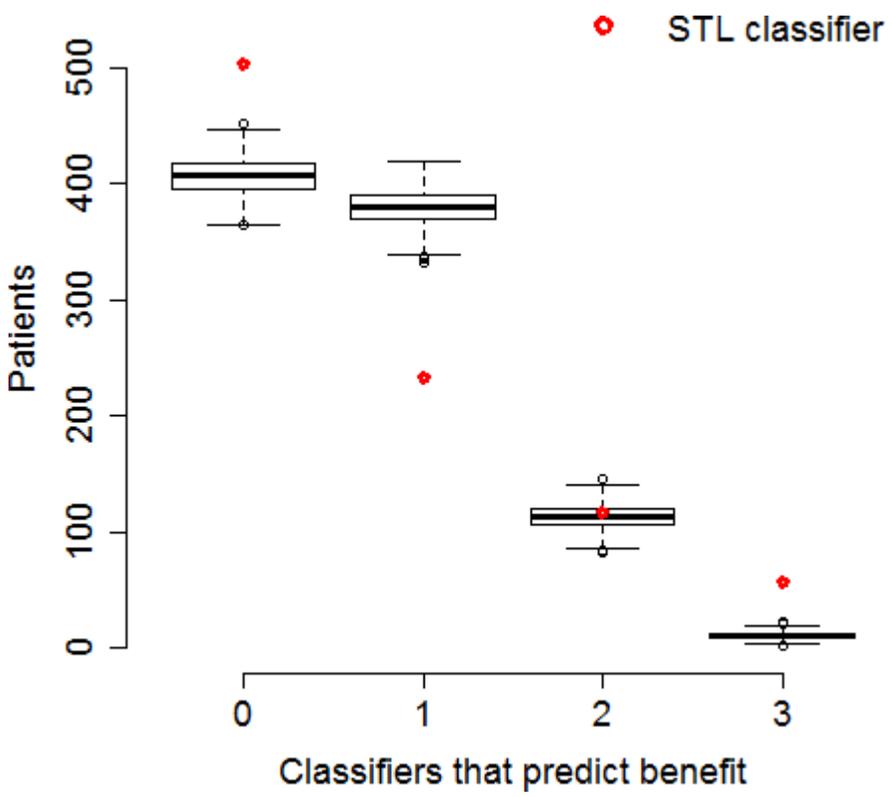
Code availability

All code needed to train and validate the classifier is available at github.com/jubels/GESTURE.

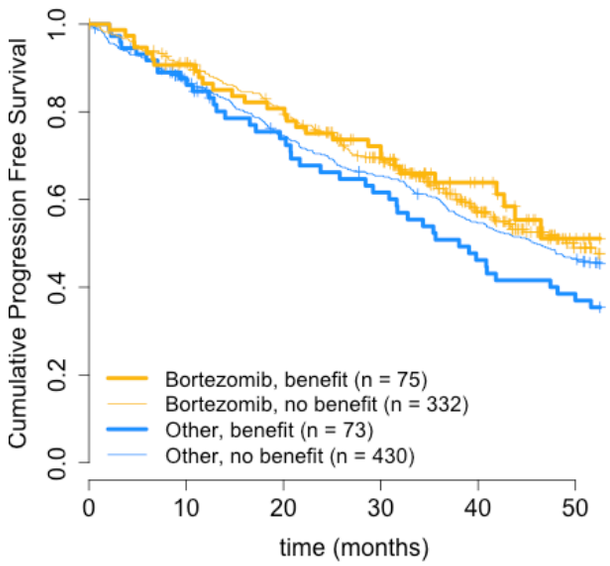
Acknowledgements

This work has been supported by a grant from the Van Herk Fellowship. The lenalidomide dataset was generated as part of the Multiple Myeloma Research Foundation Personalized Medicine Initiatives (https://research.themmrf.org and www.themmrf.org). We thank Rowan Kuiper for data aggregation and his advice on combining the datasets.

Supplementary material



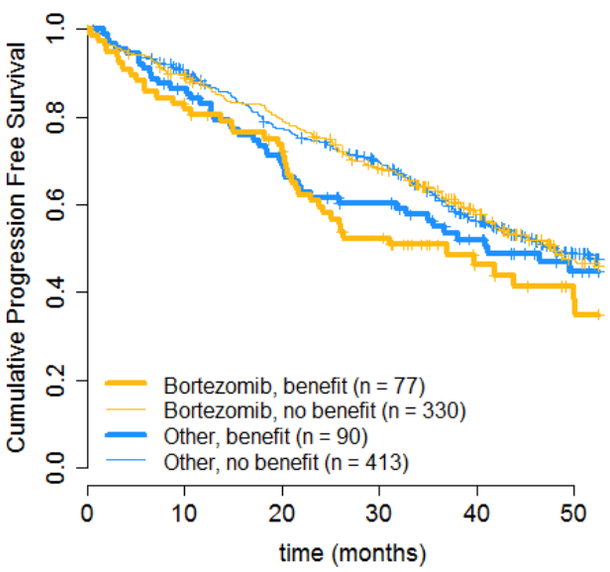
Supplementary Figure 1. We computed for how many patients the three classifiers trained in the different folds of the cross validation agree on class assignment. The values on the x-axis represent the number of classifiers that classified a patient as benefitting from treatment. A value of 0 means that all three classifiers classified a patient as 'no benefit' and the value of 3 (which is the maximum) means all classifiers agreed on the assignment to class 'benefit'. These are the red dots in the plot. We also generated 10 000 random labelings per training fold, with the same proportion of patients labeled 'benefit' and 'no benefit' as in the labelings found by STL to obtain a background distribution of the expected overlap by random chance (boxplot). Since the number of patients for which all three STL classifier agree (i.e. the patients with either a value of 0 or 3) is larger than expected by random chance, the concordance between the STL classifiers is significant.



HR 'benefit' class = 0.56, p = 0.02
HR 'no benefit' class = 0.77, p = 0.02

Supplementary Figure 2. Kaplan Meier of the classification of the bortezomib dataset using random gene sets. In the class 'benefit' an HR of 0.56 (95% CI 0.34 – 0.90, p = 0.02, n = 148) is found and in the class 'no benefit' an HR of 0.77 (95% CI 0.62 – 0.96, p = 0.02, n = 762).

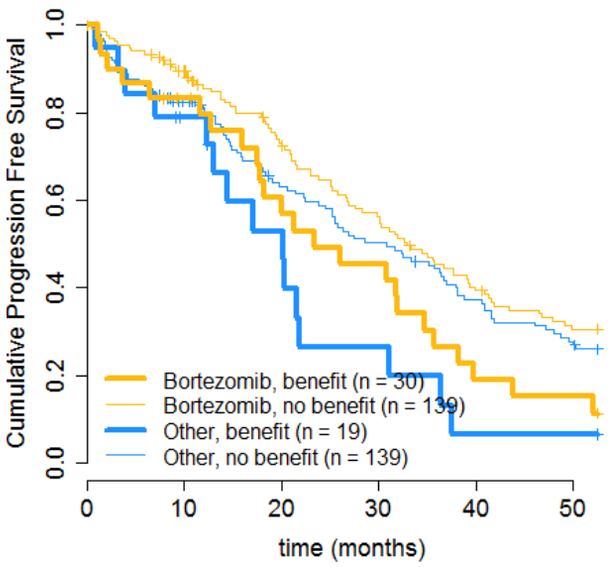
.



HR 'benefit' class = 1.09, p = 0.69
HR 'no benefit' class = 0.95, p = 0.65

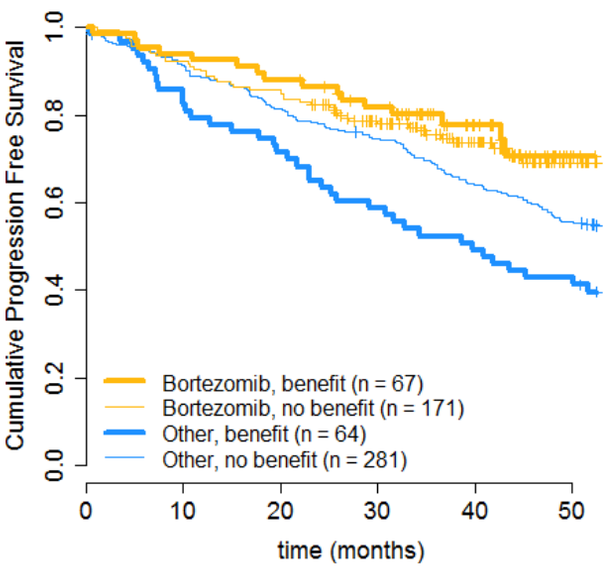
Supplementary Figure 3. Kaplan Meier showing the survival curves in validation when the treatment labels are shuffled, i.e. patients are in silico randomly assigned to the either the bortezomib or no bortezomib arm. An HR of 1.09 (95% CI 0.71 – 1.67, p = 0.69, n = 167) in the class 'benefit' and an HR of 0.95 (95% CI 0.77 – 1.18, p = 0.65, n = 743) in the class 'no benefit' is observed. It is expected that no performance is observed, since the relationship between the gene expression data and the treatment specific survival is destroyed.

2



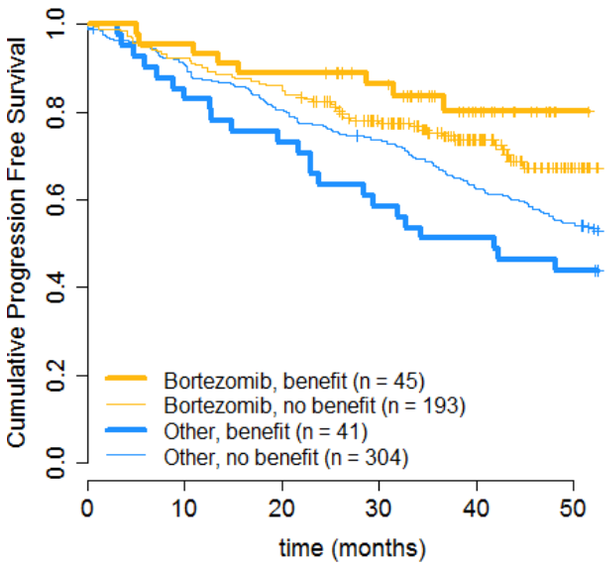
HR 'benefit' class = 0.69, p = 0.26
HR 'no benefit' class = 0.85, p = 0.27

Supplementary Figure 4. Validation performance of the STL classifier in the H65 dataset when the classifier is trained on the combined TT/H65 dataset. An HR of 0.69 (95% CI 0.36 – 1.32, p = 0.26, n = 49) is observed in class 'benefit' and an HR of 0.85 (95% CI 0.63 – 1.14, p = 0.27, n = 278).



HR 'benefit' class = 0.38, p = 0.002
HR 'no benefit' class = 0.71, p = 0.05

Supplementary Figure 5. Validation performance of the STL classifier in the Total Therapy dataset when the classifier is trained on the combined TT/H65 dataset. An HR of 0.38 (95% CI 0.21 – 0.69, p = 0.002, n = 131) is observed in class 'benefit' and an HR of 0.71 (95% CI 0.50 – 1.00, p = 0.05, n = 452) in class 'no benefit'.



HR 'benefit' class = 0.28, p = 0.00098

HR 'no benefit' class = 0.71, p = 0.038

Supplementary Figure 6. Kaplan Meier showing the survival curves when the STL classifier is trained within the Total Therapy (TT) datasets, excluding the data from the HOVON65 (H65) trial. An HR of 0.28 (95% CI 0.13 – 0.60, p = 0.00098, n = 86) is observed in class 'benefit' and an HR of 0.71 (95% CI 0.51 – 0.98, p = 0.038, n = 497) is class 'no benefit'. The HR found in class 'benefit' is far lower than the HR found in validation when TT and H65 are combined.

Supplementary Table 1. An X indicates a patient included in the study (rows) received that drug (columns).

	bortezomib	doxo-rubicin	dexa-methasone	thalido-mide	cyclophos-phamide	cisplatin	etoposide	vincristine
H65 - PAD arm	X	X	X					
TT3	X	X	X	X	X	X	X	
H65 - VAD arm		X	X	X				X
TT2		X	X	X	X	X	X	X

2

Supplementary Table 2. Performance of nearest mean classifier when different percentages are used to define class 'benefit'

	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
HR	1.05	1.02	0.72	0.85	0.96	0.65	0.64	0.74	0.75	0.65
p-value	0.85	0.96	0.26	0.53	0.86	0.08	0.09	0.23	0.23	0.08
Size class 'benefit'	0.30	0.20	0.27	0.35	0.36	0.41	0.42	0.44	0.46	0.49
Mean accuracy	0.41	0.50	0.49	0.49	0.58	0.56	0.56	0.55	0.55	0.56

Supplementary Table 3. Performance of random forest classifier when different percentages are used to define class 'benefit'

	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
HR	0.61	0.56	0.90	0.91	0.91	0.75	0.84	0.74	0.80	0.75
p-value	0.41	0.23	0.73	0.80	0.73	0.29	0.51	0.23	0.36	0.24
size class 'benefit'	0.11	0.13	0.20	0.16	0.25	0.30	0.37	0.41	0.44	0.50
accuracy	0.70	0.69	0.70	0.69	0.69	0.68	0.68	0.68	0.67	0.65

Supplementary Table 4. Performance of support vector machine when different percentages are used to define class 'benefit'. When using 5% no patients were assigned to class 'benefit' in validation, making it impossible to compute an HR.

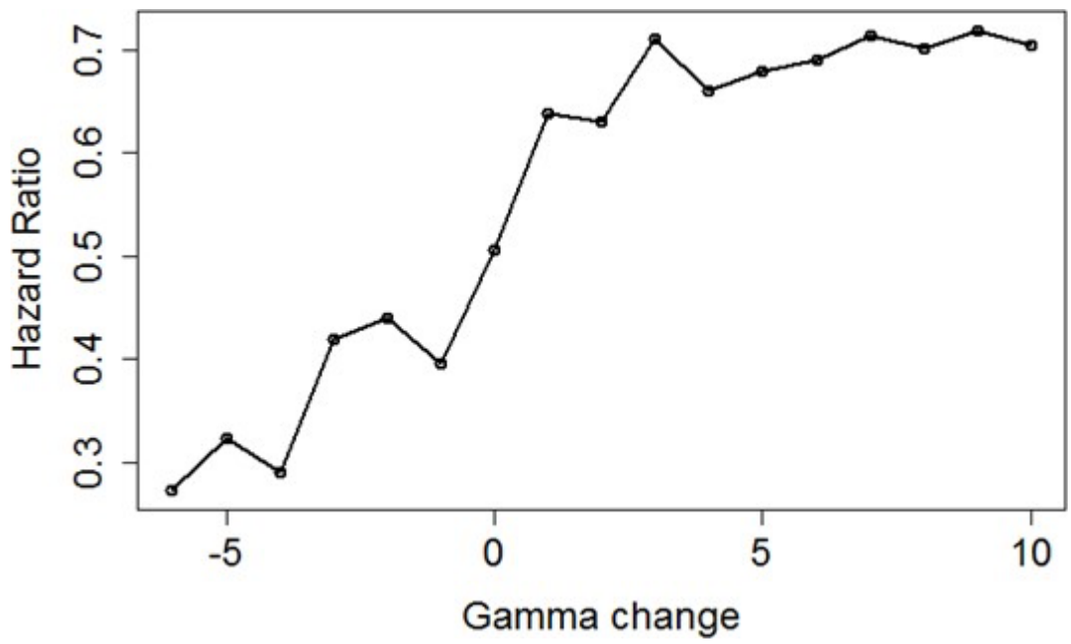
	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
HR	NA	9.47E+08	0.41	0.96	0.81	1.01	0.85	1.13	0.83	0.70
p-value	NA	1.000	0.53	0.95	0.67	0.97	0.57	0.66	0.46	0.16
size class 'benefit'	NA	0.013	0.02	0.05	0.10	0.23	0.30	0.34	0.40	0.50
accuracy	0.98	0.963	0.93	0.92	0.81	0.78	0.72	0.71	0.79	0.71

Supplementary Note 1

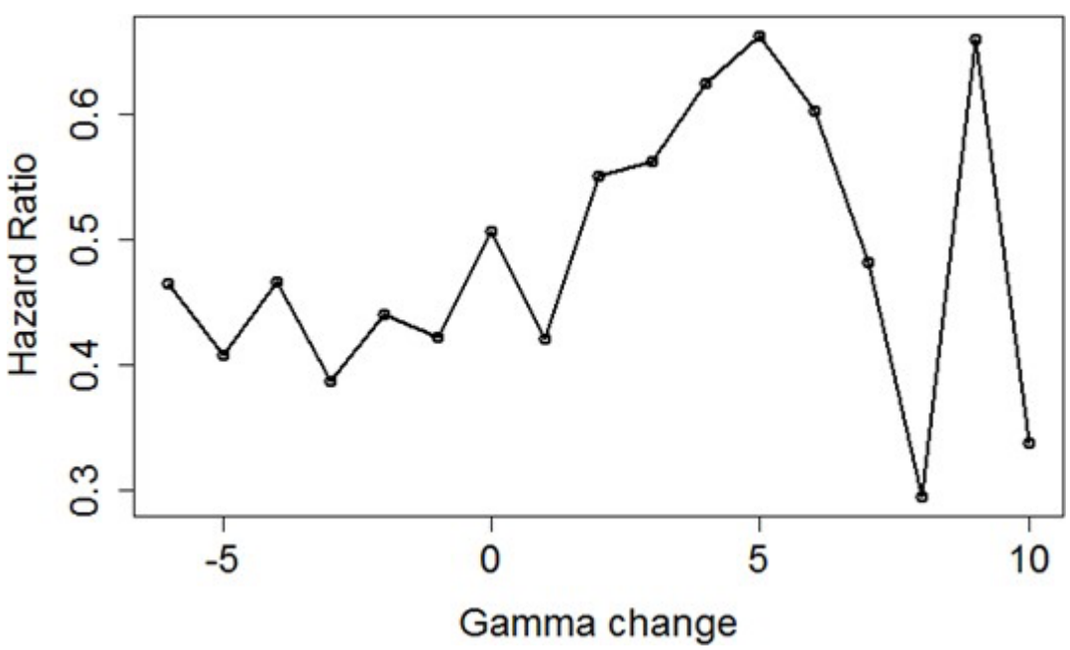
The parameters k and γ determine the classification boundary. For this reason, they are optimized using an exhaustive grid search which chooses the optimal combination. To investigate how sensitive this optimization is, we investigated how small changes to the parameters affect the HR found in validation. In essence, a smaller γ leads to a smaller class benefit. We show the effect of changing the γ parameter in two scenarios: leaving all other parameters as is (Supplementary Figure 7) and when also retraining the threshold T which determines how many classifiers need to agree on the 'benefit' classification (Supplementary figure 8). The classifier is robust to (small) changes in these parameters, which is a desirable feature of a robust classifier. As can be seen in Supplementary Figure 7, when γ decreases, the HR also decreases since a smaller class benefit is identified. This is consistent with our observation that a smaller class benefit leads to a lower HR (Figure 2c). When threshold T is also reoptimized, the HR stays relatively constant when γ is changed, since the threshold T is chosen so at least 20% of the patients are classified as class 'benefit'. Supplementary Figure 9 shows the number of patients who receive a different class assignment when γ is changed, again without reoptimizing threshold T (black line) and with reoptimization (red line). When

threshold T is reoptimized, few patients change classification, showing different settings for γ would identify the same patients as benefitting from bortezomib. We also investigated how sensitive the classifier is to changing the number of genesets in the classifier (with reoptimization of threshold T, Supplementary figure 10). The red line indicates the validation HR we originally found. As can be seen, there are many settings which achieve a similar or better validation performance, indicating the classifier is also not very sensitive to the exact number of gene sets included.

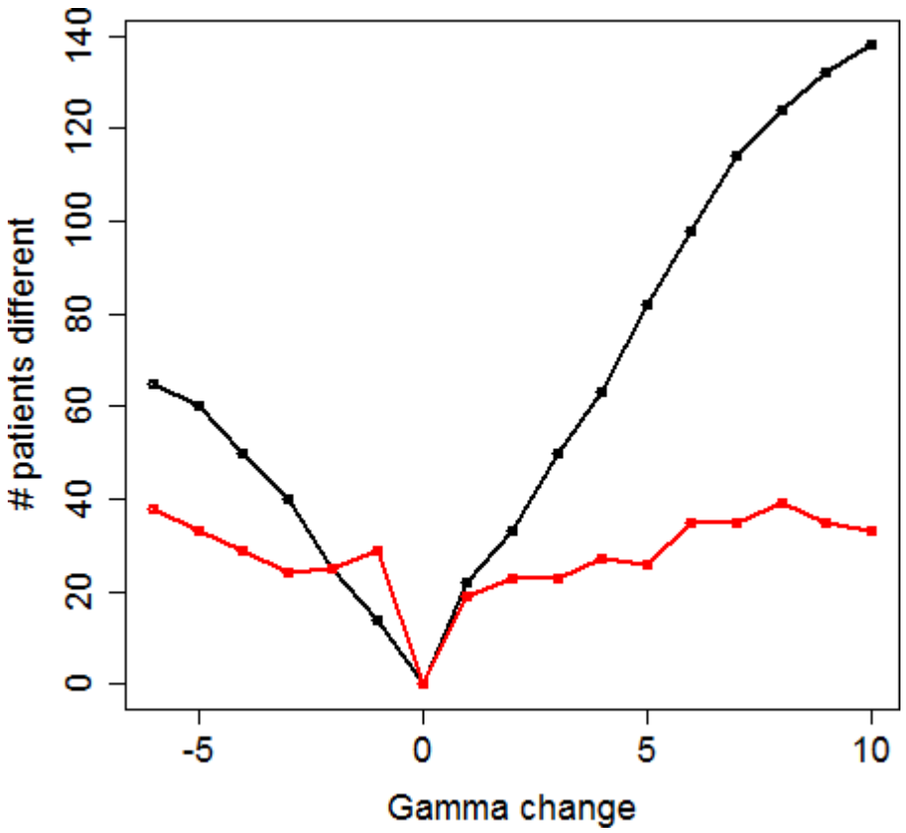
We also investigated the training HR found for all k and γ combinations of three of our top- performing genesets (Supplementary Figures 11 - 13). Note that these gene sets were the best performing in one of the folds and are not necessarily overrepresented in the final classifier. The y-axis show the different settings for k and the x-axis the different settings for γ. A yellow color indicates a low HR, a blue color a high HR and white indicates too few or too many patients were included in class 'benefit' when this combination was used. What can be seen is that a low setting for k (meaning few prototypes) leads to the most favorable HRs. Also here can be seen that small changes in k and γ do not lead to large changes in HR.



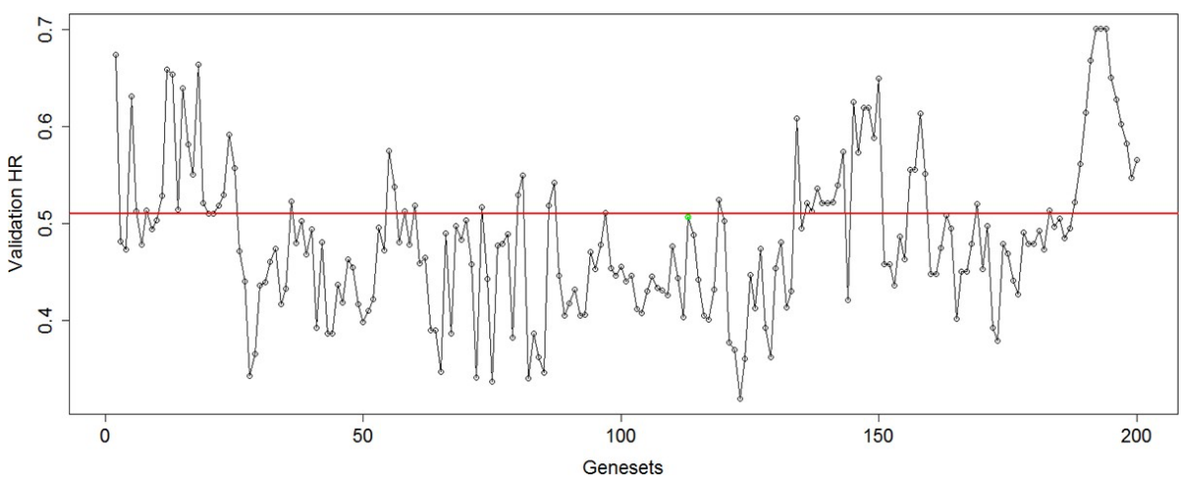
Supplementary Figure 7. The effect of changing γ on the HR when threshold T is not re-optimized. The y-axis shows the validation HR and the x-axis the change in γ.



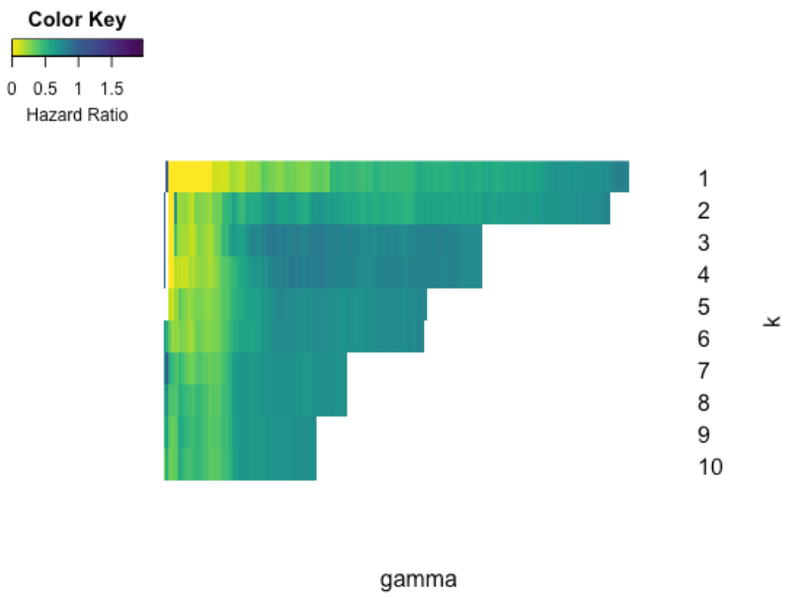
Supplementary Figure 8. The effect of changing γ on the HR when threshold T is re-optimized. The y-axis shows the validation HR and the x-axis the change in γ.



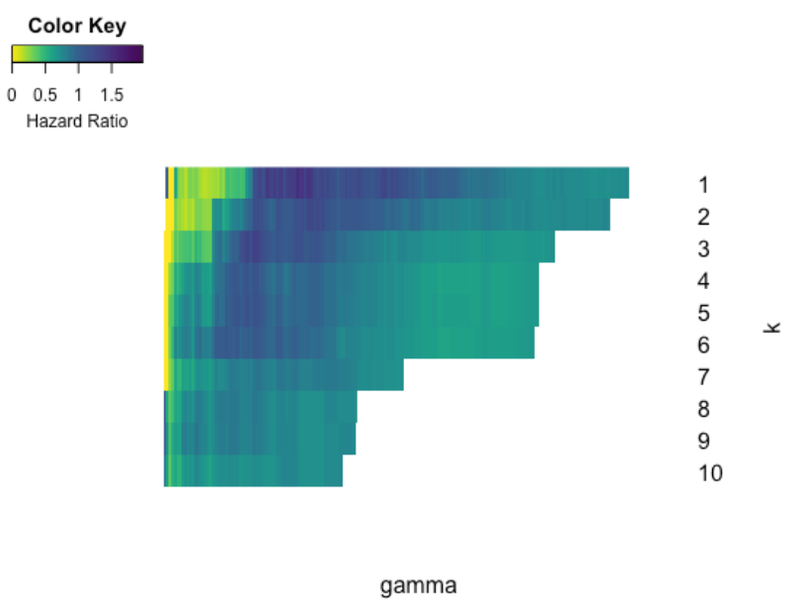
Supplementary Figure 9. The number of patients who change from class 'benefit' to class 'no benefit' or vice versa when γ is changed. The red line shows the difference when we re-optimize the threshold T, the black line when we do not.



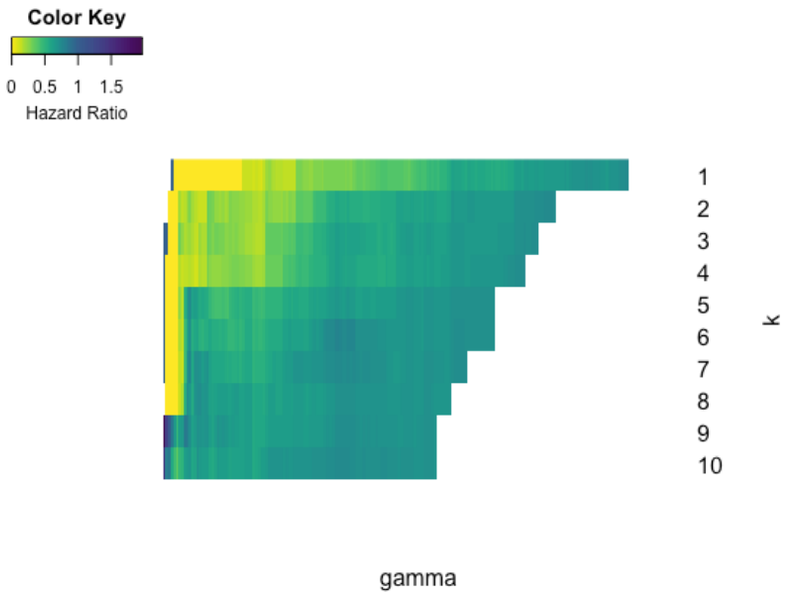
Supplementary Figure 10. The validation HR found when a different number of genesets is included in the final classifier. The red line indicates the validation HR found with the original classifier.



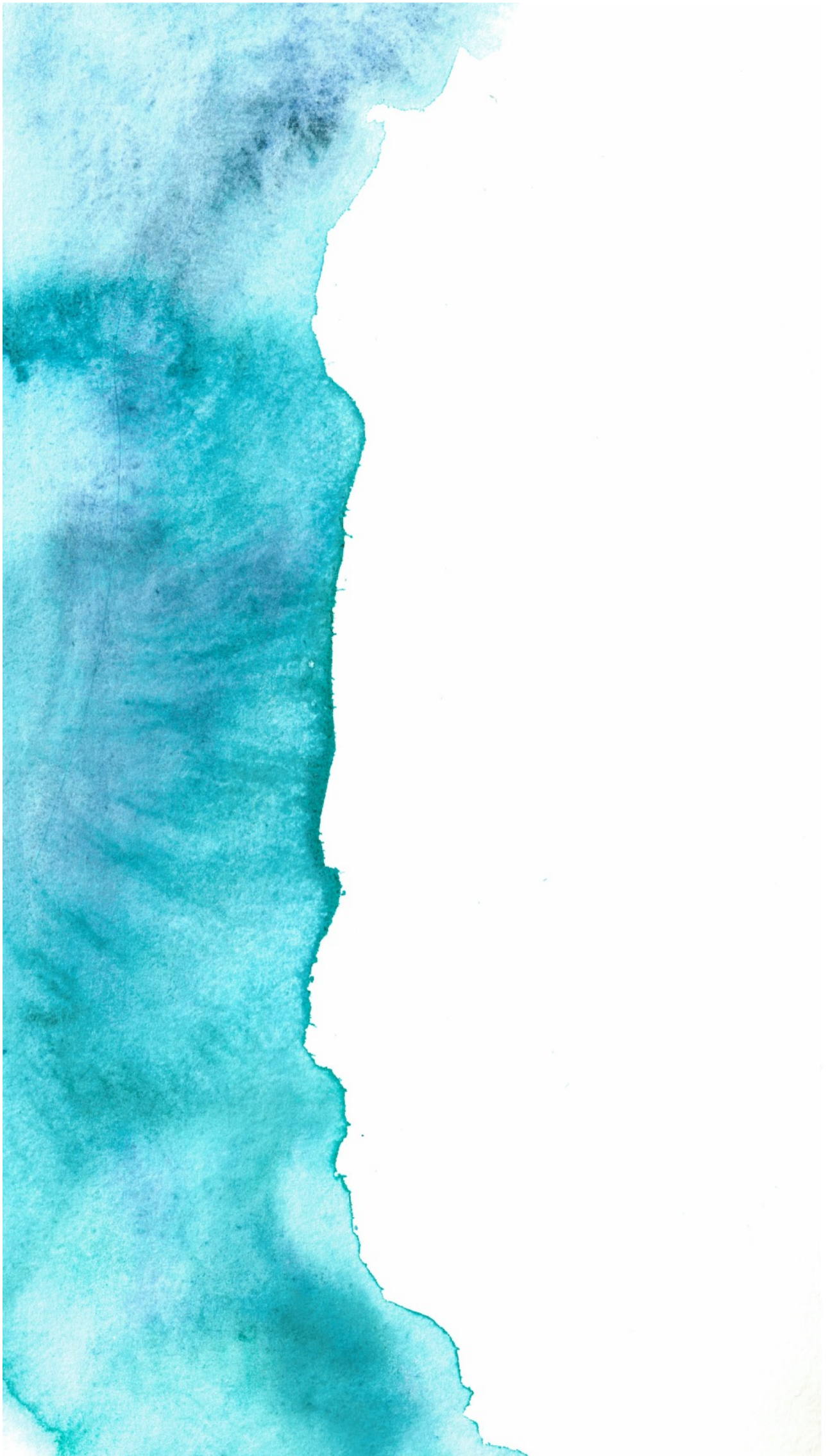
Supplementary Figure 11. Training performance for different combination of k and γ, using GO category olfactory bulb axon guidance. The y-axis show the different settings for k and the x-axis the different settings for γ. A yellow color indicates a low HR, a blue color a high HR and white indicates too few or too many patients were included in class 'benefit' when this combination was used.



Supplementary Figure 12. Training performance for different combination of k and γ, using GO category peptidoglycan receptor activity. The y-axis show the different settings for k and the x-axis the different settings for γ. A yellow color indicates a low HR, a blue color a high HR and white indicates too few or too many patients were included in class 'benefit' when this combination was used.



Supplementary Figure 13. Training performance for different combination of k and γ, using GO category IgG binding. The y-axis show the different settings for k and the x-axis the different settings for γ. A yellow color indicates a low HR, a blue color a high HR and white indicates too few or too many patients were included.



Chapter 3

Gene networks constructed through simulated treatment learning can predict proteasome inhibitor benefit in Multiple Myeloma

Joske Ubels[1,2,3,4], Pieter Sonneveld[3], Martin H. van Vliet[4], Jeroen de Ridder[1,2*]

1. Center for Molecular Medicine, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG, Utrecht, The Netherlands 2. Oncode Institute, Utrecht, The Netherlands 3. Department of Hematology, Erasmus MC Cancer Institute, Wytemaweg 80, 3015 CN, Rotterdam, The Netherlands 4. SkylineDx, Lichtenauerlaan 40, 3062 ME, Rotterdam, The Netherlands

Adapted from Clin Cancer Res September 10 2020 DOI: 10.1158/1078-0432.CCR-20-0742

Abstract

Proteasome inhibitors are widely used in treating Multiple Myeloma, but can cause serious side effects and response varies between patients. It is therefore important to gain more insight into which patients will benefit from proteasome inhibitors.

3

We introduce Simulated Treatment Learned signatures (STLsig), a machine learning method to identify predictive gene expression signatures. STLsig uses genetically similar patients who received an alternative treatment to model which patients will benefit more from proteasome inhibitors than from an alternative. STLsig constructs gene networks by linking genes that are synergistic in their ability to predict benefit.

In a dataset of 910 MM patients STLsig identifies two gene networks that together can predict benefit to the proteasome inhibitor bortezomib. In class 'benefit' we find a hazard ratio of 0.47 (p = 0.04) in favor of bortezomib, while in class 'no benefit' the hazard ratio is 0.91 (p = 0.68). Importantly, we observe a similar performance (HR class benefit = 0.46, p = 0.04) in an independent patient cohort. Moreover, this signature also predicts benefit for the proteasome inhibitor carfilzomib, indicating it is not specific to bortezomib. No equivalent signature can be found when the genes in the signature are excluded from the analysis, indicating they are essential. Multiple genes in the signature are linked to working mechanisms of proteasome inhibitors or MM disease progression.
STLsig can identify gene signatures that could aid in treatment decisions for MM patients and provide insight into the biological mechanism behind treatment benefit.

Introduction

For many anti-cancer drugs the response varies widely across patients. As many of these drugs are associated with serious side effects, it is essential to identify which drug will maximally benefit the patient. Tools that aid in such decisions, e.g. based on patient-derived genetic or transcriptomic profiles have only been developed for a few treatments and diseases. Most efforts in this direction focus on detecting specific mutations for which it is known that a targeted therapy exists (Syn et al. 2016). However, many patients do not carry any mutations that are known to be actionable and in practice only 7% of patients can be matched to a targeted therapy with the highest level of evidence (Zehir et al. 2017). Moreover, a range of efficacious therapies exist that are non-targeted. Consequently, there is a clear clinical utility for methods that can more generically predict - at the time of diagnosis - if a patient will benefit from a certain treatment or not.

Multiple myeloma (MM) is characterized by a malignant proliferation of plasma cells, both in the bone marrow and extramedullary sites. MM is considered incurable with a median survival of approximately 6 years (Rajkumar and Vincent Rajkumar 2018). Several driver mutations have been identified in MM (Walker et al. 2018), but in most patients no actionable mutations are observed and targeted therapies are therefore not commonly used in MM. Currently, proteasome inhibitors (PIs) are one of the most important components of treatment in MM and since their introduction in the clinic survival has significantly improved (Moreau et al. 2012). Due to higher immunoglobulin production, MM cells are thought to be more reliant on proteasomal degradation of proteins, making them vulnerable to proteasome inhibition (Laubach, Richardson, and Anderson 2011). After bortezomib, which was the first PI to be introduced in the clinic for MM, second generation proteasome inhibitors like carfilzomib and ixazomib have recently been approved.

Despite the success of PIs, there is still wide variability in response across patients. Substantial efforts have been made to discover what distinguishes responders from non-responders. For instance, several studies have implicated differential expression of genes involved in the unfolded protein response (Dong et al. 2009). Other studies describe complex changes in the entire energy metabolism as a potential discriminating

factor (Soriano et al. 2016). Several chromosomal aberrations have also been found to influence bortezomib response, although this effect is not fully understood (Smetana et al. 2013; Avet-Loiseau et al. 2010). Despite all efforts, there is currently no biomarker capable of determining which patients will benefit most from receiving a PI.

Most studies investigating PI response compare gene expression patterns of patients responding well or poor to a certain treatment (Laubach, Richardson, and Anderson 2011; Hofman et al. 2017; Yoshida et al. 2018; Narita et al. 2015). The identified genes can then be combined in a classifier to predict good or poor response in new patients. However, a clinically more interesting question is whether a patient will benefit more from a PI than from another treatment. This is a markedly different question than identifying good and poor responders within one homogeneous treatment group. After all, even patients with poor survival may have benefited from their treatment; their outcome could have been even worse on another treatment. Conversely, a patient with a good survival outcome could have experienced an equivalently good or better response on another treatment. It is therefore impossible to assign patients to class 'benefit' or 'no benefit' a priori, since response to another treatment cannot be observed. Standard methods. which rely on the existence of such class labels, are thus unsuitable for predicting treatment benefit.

Here we propose a novel method, Simulated Treatment Learning signatures (STLsig), to infer gene signatures that can predict treatment benefit for patients at the moment of diagnosis. We apply STLsig to find a gene expression signature capable of identifying patients for whom treatment with PIs results in better survival than an alternative treatment. Firstly, the gene signature should be capable of predicting PI benefit in an independent patient cohort, which has been shown challenging for prognostic classifiers (Bernau et al. 2014). A second important objective of STLsig is to identify a simple, interpretable model which contains genes that have biological relevance to the molecular mechanism underlying PI efficacy. To enable this, we leverage the core concept of Simulated Treatment Learning (STL), which we proposed previously (Ubels et al. 2018), that allows training classifiers without having a predefined labelling of patients. While our previous method was successful in identifying a model that can predict treatment benefit, these models rely on large numbers of Gene Ontology sets, making interpretation complex.

We propose a different approach which identifies small gene networks that can be used to predict PI benefit. To obtain a signature for treatment benefit, we form networks of genes that are complementary in their ability to predict benefit. STLsig is fully data driven and does not rely on any biological knowledge or predefined gene networks as input.

We demonstrate the utility of STLsig on a 910 sample dataset combining three different Phase III clinical trials with MM patients receiving either a treatment with or without the PI bortezomib (the HTT cohort). STLsig enables discovery of a 14-gene signature that can accurately identify a subset of patients benefiting from bortezomib. We validate this gene expression signature in independent data (the CoMMpass cohort) where we predict benefit for bortezomib or an alternative PI, carfilzomib, demonstrating that the signature is robust and generalizes to other data. Moreover, we show that no gene expression signature with a similar performance can be found when the signature genes are removed from the dataset. The genes included in the signature are thus essential for predicting PI benefit. Several of the genes in the signature are related to MM or the working mechanisms of PIs. To our knowledge, this is the first approach capable of discovering treatment benefit specific gene signatures without predefined labels.

Methods

Data

To develop the gene network and train the bortezomib benefit signature, we pool gene expression and survival data from three phase III trials (referred to as the HTT cohort): Total Therapy 2 (TT2, GSE2658), Total Therapy 3 (TT3, GSE2658) and HOVON-65/GMMG-HD4 (H65, GSE19784). The TT2 dataset includes 345 newly diagnosed multiple myeloma (NDMM) samples, treated either with thalidomide and melphalan (n = 173) or melphalan alone (n = 172). The TT3 dataset includes 238 NDMM samples treated with bortezomib, thalidomide, dexamethasone, cyclophosphamide, cisplatin and etoposide (VTDPACE). The H65 dataset includes 327 NDMM samples, treated either with vincristine, doxorubicin and dexamethasone (VAD, n = 158) or bortezomib, doxorubicin and dexamethasone (PAD, n = 169). In the HTT cohort we define a

bortezomib arm (n = 407), which comprises the PAD arm from H65 and TT3, and a non-bortezomib arm (n = 503), which comprises the VAD arm from H65 and TT2. We divide the HTT cohort in a train set (n = 606) and a test set (n = 304). We ensured the two treatment arms were distributed evenly between training and test data and that the HR between the treatments was similar.

All samples have been profiled with the Affymetrix Human Genome U133 plus 2.0 array. Gene expression is MAS5 and log2 normalized. Batch effects resulting from pooling different datasets are corrected with ComBat(Johnson, Li, and Rabinovic 2007). Data is scaled to mean 0 and variance 1 per probeset.

For validation of both the bortezomib model and carfilzomib model, we use the CoMMpass trial (NCT0145429) dataset generated by the Multiple Myeloma Research Foundation (MMRF). For 747 patients both RNAseq, survival data, and treatment information is available (CoMMpass Interim Analysis 13). Of these patients, 61 did not receive any PI in first line treatment, 530 received bortezomib and 156 received carfilzomib. Sequencing data is processed with the Cufflinks pipeline (for details see researcher.themmrf.org). For validation we combine the log2 normalized values from the HTT data and the FPKM values from CoMMpass. We scale the combined data to mean 0 and variance 1 and then perform ComBat batch correction, as performing mean-variance scaling before ComBat leads to better overlap between the datasets in the tSNE. In ComBat batch correction H65, TT2, TT3 and CoMMpass are defined as four separate batches.

For training the signature, we use Progression Free Survival (PFS) as an endpoint as we consider PFS a more direct measurement of treatment effect than overall survival. Cox proportional hazard models were fitted using the R package 'survival' (version 2.44).

Constructing and evaluating gene pairs

We select only probe sets that meet the following requirements: (i) variance across the samples > 2 in the training dataset before mean variance scaling, (ii) unambiguous mapping to one gene and (iii) matching gene in the CoMMpass dataset. This yields n = 3319 genes. We then construct all possible gene pairs from these 3319 genes, resulting in 5,506,221 gene pairs.

To train the gene signature we divide the HTT cohort (n = 910) into four folds, Fold A (n = 202), Fold B (n = 202), Fold C (n = 202) and Fold D (n = 304), fold assignment is provided in the supplementary information. Fold A, B and C are used to train the signature as described below, while fold D acts as hold out data to validate the signature and optimize a threshold to use in independent validation data.

To determine treatment benefit, we follow the core concept of STL laid out in our previous work (Ubels et al. 2018), where for each patient a score zPFS is defined that measures whether the patient survived longer than expected compared to patients with similar gene expression that received another treatment. More specifically, for genepair {*n,m*} and patient j we define:

$$\mu PFS_j^{n,m} = \frac{I}{K} \sum_{i \epsilon \Pi^j} PFS_j - PFS_i$$

where PFS_j is the progression free survival time of patient *j*, I =1 if patient *j* received the target treatment (a PI in this work) and I =-1, otherwise. Moreover, Π^j is the set of K nearest neighbors to patient *j* defined in terms of euclidean distance in the expression space spanned by genes *n* and *m* and only considering patients that received another treatment than patient *j*. Throughout this manuscript K=10. In the set Π^j, we discard patients for whom we cannot be sure whether they survived longer or not (i.e. if both patients are censored). This leads to an average of 7 patients being used in the calculation of μPFS_j. Subsequently, zPFS is normalized to a z-score by comparing *PFS* to a background distribution resulting from repeating this procedure M=1000 times with a random Π^j. The zPFS score describes how much smaller (or larger) the survival of patient *j* is compared to patients with similar gene expression but opposite treatment than expected by random chance.

To score gene pairs, a 2-fold cross validation is employed using fold A (n = 202) and fold B (n = 202). Within each fold, a kNN-regression model (k = 10) is trained, which is used to predict zPFS on the other fold. The gene pair score is defined as the Spearman correlation coefficient between the predicted zPFS and calculated zPFS across all

patients. The score for each gene pair is the mean correlation of the 2 folds. We repeat this procedure 5 times with a different split in folds.

Gene network construction

We construct gene networks separately for all 5 repeats and then construct a consensus network, which only contains the genes and edges found in all 5 repeats. To construct the gene networks, for each gene, we rank all gene pairs containing that gene on the mean Spearman correlation coefficient found. We then connect genes that are mutually synergistic. We achieve this by requiring that AB is among the top 5% of pairs including A and among the top 5% of pairs including B. However, if a single gene is informative for treatment benefit, gene pairs containing this gene could be highly ranked even if the second gene is uninformative. Including these gene pairs in our network and subsequent signature would introduce noise, which would both harm biological interpretation of the signature and potentially decrease the predictive performance in independent data. Therefore, we also require the mean correlation of the gene pair to be above the median correlation of all selected gene pairs. We evaluate all gene networks in the consensus network on their ability to predict benefit and select the best performing combination to construct the signature.

Gene network selection and gene signature construction

After gene network construction, gene networks are selected using forward feature selection. To rank gene networks, we determine the predictive performance for each gene network. To this end, we calculate zPFS for each patient and each gene network separately on fold A and B together (n = 404). The top 25% of patients (in terms of zPFS) are assigned to class 'benefit', while the remaining patients are assigned to class 'no benefit'. Subsequently, a Cox proportional hazards regression on the treatment variable is performed within the 'benefit' patient group as well as in the 'no benefit' patients. The performance of a gene network is defined by the difference between the Cox' regression β's in class 'benefit' and class 'no benefit'.

To select gene networks to use in the final model we perform forward feature selection using fold C, which comprises 202 patients not used in fold A and B. Gene networks are added sequentially based on their performance on fold A and B. Ranking of patients across more than one gene network is done based on the sum of the zPFS scores of the individual gene networks.

Validation of gene networks

To validate the signature in independent data, we use all training data (n = 606) as a reference set where zPFS is known. For each patient in the validation set we compute the euclidean distance to all patients in the reference set per gene network. We then use inverse distance weighting to calculate the estimated zPFS of a validation patient *j* by

$$\widehat{zPFS_j} = \frac{\sum_{i \in T} w_i * zPFS_i}{\sum_{i \in T} w_i}$$

where *T* comprises all patients in the reference dataset. Given a certain gene expression vector $\boldsymbol{x}$, weights w_i are given by

$$w_i = \frac{1}{d(x_j, x_i)}$$

where *d* is the Euclidean distance between the expression data of gene of patients *i* and *j*.

Results

Overview of the algorithm

STLsig relies on the idea that patients exhibiting similar gene expression profiles who received different treatments, can be used to model response to the treatment they did not receive. Similarity between patients should be defined by genes relevant to treatment benefit. STLsig therefore derives treatment specific gene networks, to form a gene expression signature capable of predicting treatment benefit. To train this signature we divide the HTT cohort in a test set (Fold D, n = 304) and a training set (n = 606), which is further subdivided into three equal parts, fold A, B and C. We then assess the ability to predict bortezomib benefit for all 5,506,221 gene pairs arising from the high variance genes (n = 3319) in the HTT training set.

For each patient *i* in fold A, we determine a z-score (zPFS) per gene pair describing the normalized mean survival difference of patient *i* with its genetically similar neighbours that received a different treatment than patient *i*. We then test the ability of the genepair to predict the zPFS score for patients in fold B. We also assess the performance of each gene pair when calculation of zPFS is performed on Fold B and predicted on Fold A. Performance of each gene pair is defined as the mean Spearman rank correlation coefficient between predicted and calculated zPFS values in both folds. A gene pair is

retained if it is synergistic, i.e. if the genes in the pair predict zPFS better together than when they are paired with other genes.
We form a consensus network by repeating the two-fold cross validation five times. Only gene pairs that are found to be synergistic in all repeats and that exceed the median correlation across all gene pairs and all repeats are retained. From this consensus network we extract gene networks, i.e. all connected components.

To evaluate each gene network, we recalculate zPFS for each patient using all genes in the network and classify the top 25% of the patients as class 'benefit' and the rest as class 'no benefit'. Subsequently, gene networks are ranked based on the difference between the Cox regression β's found in class 'benefit' and class 'no benefit'. To build the signature, we sequentially add each network based on this ranking and evaluate the performance of the combined networks on fold C. The steps of the algorithm are summarized in **Figure 1**.

Gene networks yield a 14-gene signature that can predict bortezomib benefit

The consensus network formed as described above contains 617 genes connected by 451 edges and consists of 167 gene networks, which includes 104 individual gene pairs. The largest gene network contains 42 genes; the mean number of genes per network is 3.7. The optimal signature is formed by combining the top two ranked gene networks, which are shown in **Figure 2**. With this signature we find a hazard ratio (HR) of 0.49 (p = 0.09, 95% CI 0.22 - 1.11) in class 'benefit' (n = 50) and an HR of 0.91 (p = 0.74, 95% CI 0.54 - 1.55) in class 'no benefit' (n = 152), on fold C of the HTT cohort. In order to assign a zPFS score to a new and unseen patient, for whom survival is unknown, we calculate the distance in gene expression space between this patient and every patient in the training data (the reference set). The predicted zPFS score of the new patient is the weighted sum of the zPFS scores of the patients in the reference set. Weights are determined by the inverse distance, i.e. the most similar patients in the reference set contribute most to the predicted zPFS (see 'Methods'). In this manner, we assess the ability of the 14-gene signature to predict benefit for the 304 patients from the HTT cohort not included in training (Fold D). The HR in favour of bortezomib found in fold D is 0.75 (p = 0.11, 95% CI 0.53 - 1.06). **Figure 3a** shows the HR in class benefit found

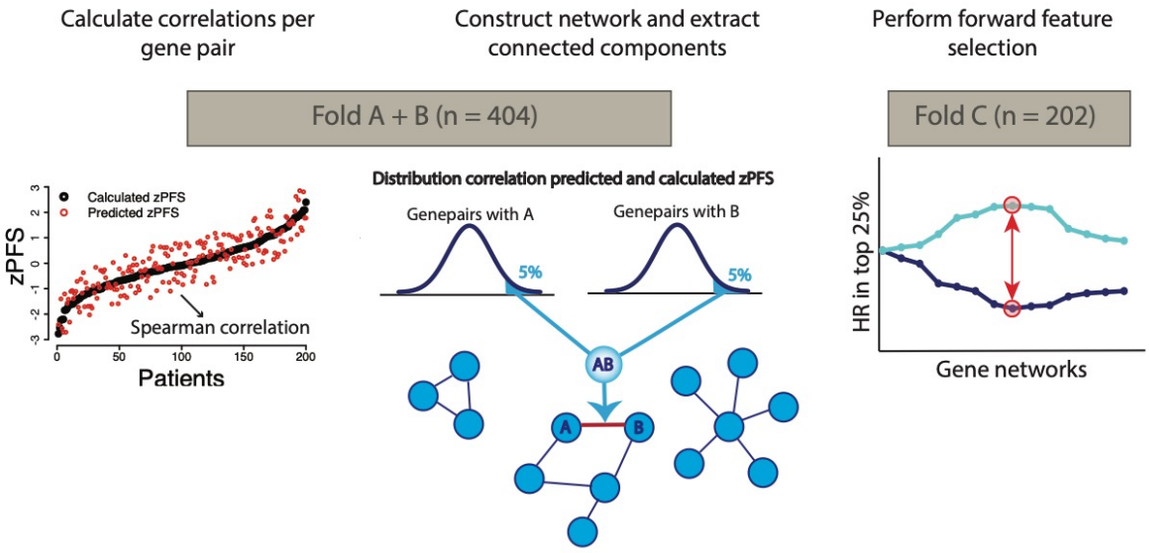


Figure 1. Overview of the construction and selection of the gene networks. First each gene pair is scored on the correlation between predicted and calculated zPFS. Gene networks are then formed by connecting synergistic genes, i.e. genes that are amongst the top 5% partners for each other based on correlation coefficient. The gene networks are then ranked based on difference between Cox regression β in class 'benefit' and 'no benefit'. The signature consists of the combination of gene networks that results in the largest difference in Cox' regression β between class 'benefit' and 'no benefit'.

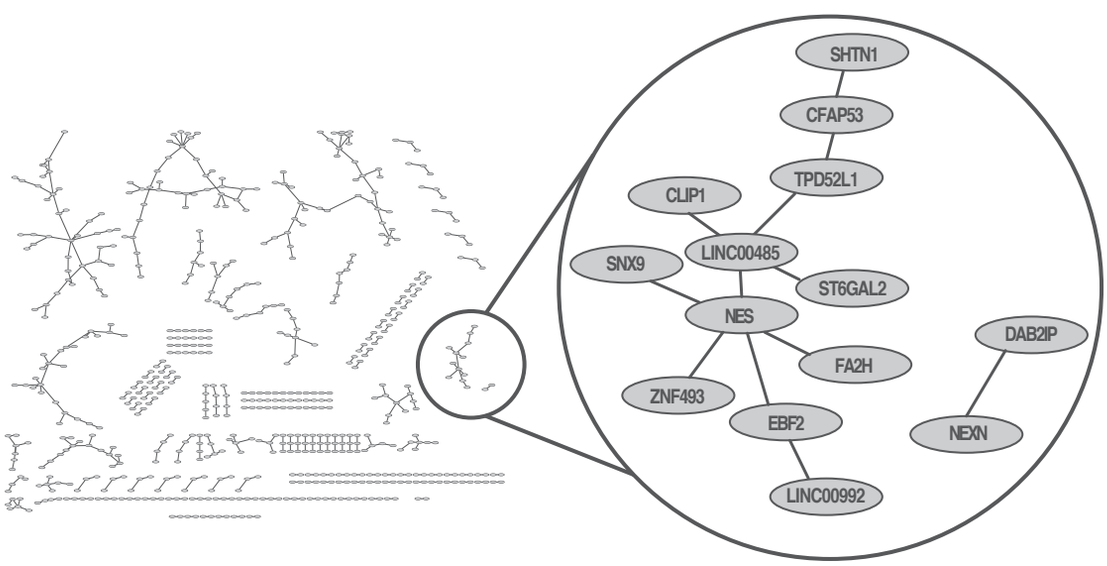


Figure 2. The constructed network with all gene networks. The highlighted networks are those selected by the feature selection procedure and contain the 14 genes in the signature.

with different zPFS thresholds. A range of thresholds result in an HR below the HR observed in the total dataset, indicating that the predicted zPFS is associated with bortezomib benefit. The optimal class 'benefit', i.e. the class 'benefit' associated with the lowest HR, comprises 30.6% of the patients which corresponds to a zPFS threshold of 0.3268. With this threshold we find an HR of 0.47 (p = 0.04, 95% CI 0.23 - 0.96) in class 'benefit' and an HR of 0.91 (p = 0.68, 95% CI 0.60 - 1.39) in class 'no benefit' (**Figure 3b**). This establishes that our signature can predict bortezomib benefit in unseen data from the same patient cohort, demonstrating that the signature can be used prospectively to inform treatment choice. Our results indicate that, despite the fact that nearly all MM patients receive a treatment regimen that includes a PI (Moreau et al. 2012), approximately 70% of patients do not see benefit.

The 14-gene signature achieves robust prediction performance in an independent patient cohort

Gene expression signatures often suffer from cohort-specific fitting and cross-validation within one dataset can thus lead to an overestimation of performance(Castaldi, Dahabreh, and Ioannidis 2011). To obtain a more robust estimate of performance it is essential to perform validation on an external and completely independent cohort. Therefore, we validate its performance in the CoMMpass trial, which represents an independent patient cohort which was profiled on a different platform (RNAseq). In contrast to the HTT dataset, which is a randomized clinical trial, the CoMMpass dataset is an observational study and thus represents clinical reality more closely. To bring the CoMMpass RNAseq data in the same space as the microarray reference dataset, we employ a ComBat batch correction (**Supplementary Figure 1** and **Methods**). We define a PI treatment arm (n = 686) and a no PI treatment arm (n = 61). The PI treatment arm contains both bortezomib and carfilzomib. Using the threshold optimized on fold D of the HTT cohort, we find an HR of 0.46 (p = 0.04, 95% CI 0.22 - 0.97) in class 'benefit' (n = 150) (**Figure 3c**). We also see a good performance when we use overall survival (OS) as an endpoint (**Table 1**). Our signature is thus capable of predicting benefit in a completely independent cohort and across platforms, indicating the signal picked up by our classifier is robust and generalizes to the broader MM patient population. We next assess the performance for each of the two PIs separately. When we evaluate benefit for the bortezomib patients (excluding carfilzomib patients from

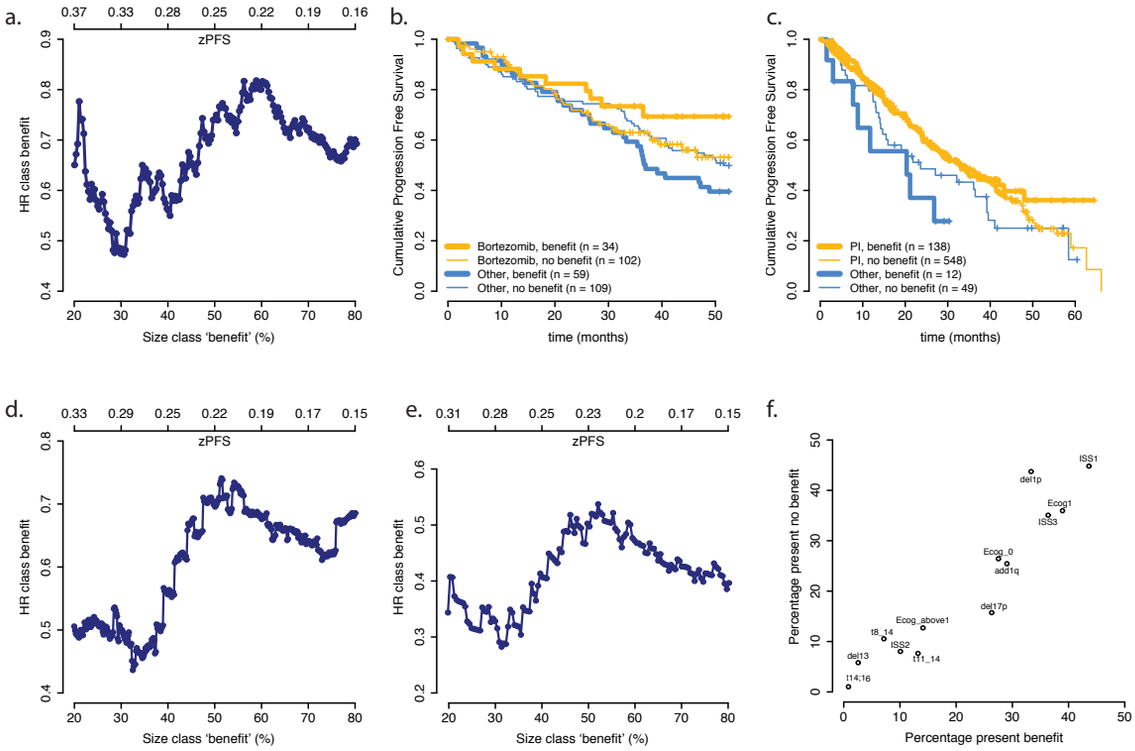


Figure 3. a. HR found in class 'benefit' using different zPFS thresholds on the hold out data. b. KM of bortezomib benefit prediction in the hold out data using the optimal zPFS threshold. c. KM of PI benefit prediction on CoMMpass using the optimal zPFS threshold from the hold out data. d. HR found in class 'benefit' for bortezomib in CoMMpass, using different zPFS thresholds. e. HR found in class 'benefit' for carfilzomib in CoMMpass, using different zPFS thresholds. f. the prevalence of certain patient characteristics in class 'benefit' and 'no benefit'.

the analysis), we find an HR of 0.49 (p = 0.06, 95% CI 0.23 - 1.03) in class 'benefit' (n = 124).

When predicting benefit for the carfilzomib patients we find an HR of 0.31 (p = 0.06, 95% CI 0.09 - 1.02) in class 'benefit' (n = 38). It should be noted the carfilzomib 'no benefit' group should be considered a 'less benefit' group, as there is still a significant HR in favor of carfilzomib in class 'no benefit', likely due to the low overall HR (0.42, p = 0.0004, 95% CI 0.26 - 0.68). Nevertheless, the fact that our signature can identify a patient group with substantially reduced HRs for carfilzomib treated patients indicates that it is more broadly applicable to PIs in general and not only bortezomib. All HRs for PFS and OS are shown in **Table 1**.

Table 1. Summary of HRs found in all analyses using PFS and OS as an endpoint

	HR whole population (PFS)	HR benefit (PFS)	HR no benefit (PFS)	HR whole population (OS)	HR Benefit (OS)	HR no benefit (OS)
Bortezomib (HTT cohort)	0.75 (0.53 - 1.06) p = 0.11 n = 304	0.47 (0.23 - 0.96) p = 0.04 n = 93	0.91 (0.60 - 1.39) p = 0.68 n = 211	0.71 (0.45 - 1.11) p = 0.13 n = 304	0.50 (0.20 – 1.23) p = 0.13 n = 93	0.82 (0.47 – 1.40) p = 0.46 n = 211
PI	0.70 (0.51 – 0.97) p = 0.04 n = 747	0.46 (0.22 - 0.97) p = 0.04 n = 150	0.79 (0.55 - 1.13) p = 0.2 n = 597	0.55 (0.34 – 0.85) p = 0.007 n = 747	0.18 (0.08 - 0.42) p = $8*10^{-5}$ n = 150	0.74 (0.44 - 1.23) p = 0.25 n = 597
Bortezomib (no Carfilzomib)	0.75 (0.54 – 1.04) p = 0.09 n = 591	0.49 (0.23 - 1.03) p = 0.06 n = 124	0.84 (0.58 - 1.21) p = 0.35 n = 467	0.60 (0.39 – 0.92) p = 0.02 n = 591	0.20 (0.09 - 0.48) p = 0.0003 n = 124	0.79 (0.48 - 1.33) p = 0.038 n = 467
Carfilzomib (no Bortezomib)	0.42 (0.26 – 0.68) p = 0.0004 n = 217	0.25 (0.06 – 0.93) p = 0.04 n = 37	0.47 (0.27 - 0.80) p = 0.005 n = 180	0.24 (0.11 – 0.53) p = 0.0004 n = 217	Inf*	0.37 (0.16 - 0.85) p = 0.02 n = 180
Lenalidomide	0.72 (0.58 – 0.88) p = 0.001 n = 747	0.79 (0.50 - 1.25) p = 0.31 n = 149	0.69 (0.54 - 0.86) p = 0.001 n = 598	0.56 (0.41 – 0.76) p = 0.0001 n = 747	0.74 (0.36 – 1.52) p = 0.42 n = 149	0.51 (0.37 – 0.73) p = 0.0002 n = 598
PI (excluding chemotherapy)	0.69 (0.49 – 0.97) p = 0.03 n = 515	0.50 (0.23 - 1.13) p = 0.10 n = 109	0.75 (0.51 - 1.11) p = 0.15 n = 406	0.52 (0.33 – 0.83) p = 0.006 n = 515	0.21 (0.08 – 0.53) p = 0.001 n = 109	0.67 (0.39 – 1.15) p = 0.15 n = 406

PI (no PI/lenalidomide combination)	0.91 (0.64 – 1.30) p = 0.60 n = 387	0.65 (0.27 – 1.57) p = 0.33 n = 71	1.02 (0.68 – 1.51) p = 0.93) n = 316	0.82 (0.51 – 1.32) p = 0.42 n = 387	0.29 (0.10 – 0.82) p = 0.02 n = 71	1.05 (0.61 – 1.81) p = 0.86 n = 316
Bortezomib (APEX)	0.74 (0.54 – 1.03) p = 0.07 n = 242	0.31 (0.11 - 0.87) p = 0.03 n = 25	0.78 (0.55 - 1.10) p = 0.15 n = 217	1.24 (0.82 – 1.82) p = 0.28 n = 242	0.42 (0.15 – 1.16) p = 0.09 n = 25	1.42 (0.93 – 2.16) p = 0.10 n = 242
Bortezomib/Lenalidomide vs Bortezomib	0.64 (0.51 – 0.81) p = 0.0002 n = 530	0.84 (0.51 – 1.39) p = 0.50 n = 112	0.59 (0.45 – 0.76) p = $7*10^{-5}$ n = 418	0.46 (0.32 - 0.66) p = $1*10^{-5}$ n = 530	0.63 (0.28 – 1.45) p = 0.28 n = 112	0.42 (0.28 – 0.63) p = $3*10^{-5}$ n = 418

* No events in carfilzomib arm

The percentage of patients classified as 'benefit' in the CoMMpass dataset is lower than on the HTT dataset. When we calculate the HR on the CoMMpass dataset using different zPFS thresholds to define class 'benefit', we find that for both bortezomib and carfilzomib the class 'benefit' associated with the lowest HR contains approximately 30% of the patients (**Figure 3d,e**), similar to what we observed in the HTT data. This shows that also in CoMMpass, different zPFS thresholds are associated with benefit and suggests approximately 30% of MM patients experience more benefit from PI treatment than the population as a whole.

Finally, we confirm that our model is specific for PI treatment by testing it on the immunomodulatory drug lenalidomide. We find an HR of 0.79 (95% CI 0.50 - 1.25) in class 'benefit' (n = 149), clearly showing the signature is specific for PI treatment.

The predictive performance of the 14-gene signature holds in single agent PI treatment

In clinical practice, the majority of patients receive a combination of treatments. To ensure the signal captured in our signature is PI specific, and not dependent on a specific treatment combination, we test the performance of our signature on data from the APEX trial(Lee et al. 2008) (GSE9782). In this trial, a single agent bortezomib

treatment was tested against high-dose dexamethasone in a relapse setting. Unfortunately, two of the genes in our signature (CFAP53 and linc00485) were not measured in this study, but we can apply the signature with the remaining 12 genes. With these genes, 10.3% of the patients are classified as benefit and we find an HR of 0.31 (95% CI 0.11-0.87, p = 0.03) in favor of bortezomib in class 'benefit', while in class 'no benefit' we find an HR of 0.78 (95% CI 0.55 - 1.10, p = 0.15)(**Supplementary Figure 2**). Secondly, while there are no single agent treated patients in CoMMpass, we find that we can still predict benefit when we remove patients who received both a PI and lenalidomide, albeit with a non-significant HR due to lower sample size (**Table 1**). The signal of our signature is thus not dependent on a combination of treatments.

The predictive performance of the 14-gene signature is relevant in current clinical practice

Chemotherapy is not regularly used to treat MM in the clinic anymore, but is present in the CoMMpass dataset. To test the performance of our model in a more clinically representative setting and show the performance generalizes to a more modern treatment regimen, we exclude all patients who received any type of chemotherapy (vincristine, doxorubicin, cyclophosphamide or melphalan, n = 232, patient numbers per treatment in Supplementary table 2). We find that, in this chemotherapy free cohort, we can still predict benefit to PI treatment with a similar effect size (HR = 0.50 (95% CI 0.23 - 1.13, p = 0.10)), as found in the whole dataset (**Supplementary Figure 3**). However, due to the smaller sample size this HR is not significant at p = 0.05. Bortezomib and lenalidomide are two of the most used drugs and are often given together. In the CoMMpass data many patients in the PI arm also received lenalidomide (see **Table 2** for patient numbers per treatment). The combination of bortezomib and lenalidomide is superior to both lenalidomide alone and to bortezomib alone. However, in class 'benefit' this combination is not superior to bortezomib alone (HR = 0.95, 95% CI 0.58 – 1.55, p = 0.84), suggesting the addition of lenalidomide is not beneficial if the patient already benefits from bortezomib treatment. This shows our signature is also relevant in treatment combinations and could guide when bortezomib alone is sufficient, thus reducing the treatment burden on the patient.

Table 2. Overview of the distribution of the combination of PI and lenalidomide treatment in the CoMMpass dataset

	Class 'benefit'	Class 'no benefit'	Total
Bor without Len	53 (35.6%)	219 (36.6%)	272 (36.4%)
Car without Len	8 (5.4%)	51 (8.5%)	59 (7.9%)
Len without Car/Bor	10 (6.7%)	46 (7.7%)	56 (7.5%)
BorLen	59 (39.6%)	199 (33.3%)	258 (34.5%)
CarLen	17 (11.4%)	80 (13.4%)	97(13.0%)
No Car/Bor/Len	2 (1.3%)	3 (0.5%)	5 (0.7%)
	149	598	747

Bor= bortezomib (PI arm), Car = carfilzomib (PI arm), Len = lenalidomide

Class 'benefit' cannot be characterized by known markers or models

Next we assess whether class 'benefit' can be characterized by known markers. To this end we first performed enrichment analysis of the routinely measured chromosomal aberrations (FISH markers), ECOG performance status or revised International Staging System (ISS) score in both classes. None of these were overrepresented in either class 'benefit' or 'no benefit' (**Figure 3f**).

Moreover, none of the markers have a predictive performance for PI benefit in the CoMMpass study that outperforms our signature (**Supplementary figure 4**). There have been extensive efforts to predict prognosis in MM using gene expression, for example with the GEP70 signature (Chapman et al. 2018; Shaughnessy et al. 2007). We find no correlation between our score and the GEP70 model (**Supplementary figure 5**). While we observe that the GEP70 low risk group in CoMMpass has more benefit form PI treatment (HR = 0.56, 95% CI 0.38 - 0.84, p = 0.005), we do not see this effect in the H65/GMMG-HD4 dataset (HR benefit = 0.90, 95% CI 0.67 - 1.20, p = 0.45). Our

signature can also still distinguish a benefit group (n = 86) within this low risk group in CoMMpass (HR benefit is 0.33, 95%CI 0.14 - 0.76, p = 0.009).
Recently, a 7-gene signature was published to distinguish standard and good response to bortezomib in the PADIMAC study (Chapman et al. 2018); none of these genes overlap with our signature genes. When we assess the ability of our signature to predict bortezomib response in the PADIMAC study, we find an AUC of 0.86 (**Supplementary Figure 6**), indicating that our signature is also capable of predicting response. Moreover, the 7-gene signature is reported to only be applicable in a non-transplant setting, while our signature does not have this limitation.

Selected genes and links between them are essential for performance

In prognostic classification it is known that many different signatures with similar performance can be found (Ein-Dor et al. 2005). This casts doubt on the usefulness of biologically interpreting the genes within a signature. We thus first investigate whether the genes in our signature are essential for performance.

We first permute the expression vector for each gene in the signature 100 times (while the other 13 genes remain unchanged) and apply this signature to fold D of the HTT cohort. The largest effect is observed for *DAB2IP*, with a mean difference in validation HR of 0.29 (se = 0.06). Correlation between genes influence the decrease in performance: for instance, shuffling *SHTN1* has the smallest impact on validation performance and its expression is significantly correlated with more genes than any other gene (with *TPD52L1*, *NES* and *ST6GAL2*, **Supplementary figure 7**). Therefore, losing its information has less impact. Nevertheless, we demonstrate all individual genes are important for the validation performance, as none can be shuffled without decreasing performance (**Figure 4a**).

Next, we assess the importance of the relationship between the genes by shuffling the edges between all genes included in the network ten times, while ensuring every gene remains linked to at least one other gene. We then infer a signature with STLsig, meaning a new combination of gene networks is selected to form the predictive signature. The mean HR found in the hold out data in class 'benefit' is 0.74 (se = 0.05), which is approximately equal to the HR found in the dataset without classification. The connection between genes is thus essential for the performance of the signature. Lastly,

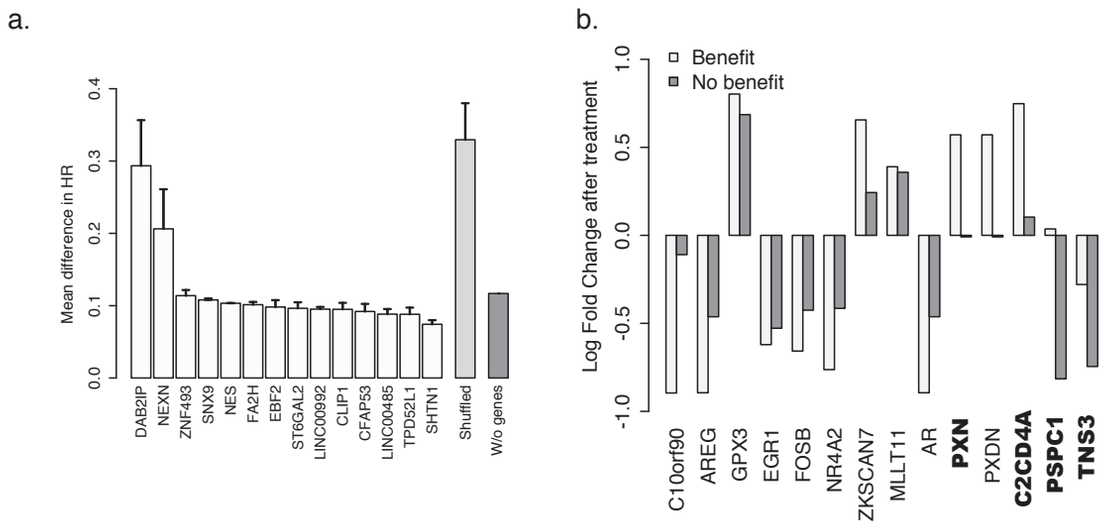


Figure 4. a. The decrease in performance (difference in HR) for i) shuffling of each gene separately, ii) shuffling links in the network and iii) when the 14 signature genes are excluded from the analysis. Error bars indicate standard error. b. Genes with a significant change in expression before and 48 hours after bortezomib treatment in only either class 'benefit' or 'no benefit'. Bold genes have a significant difference in response between class 'benefit' and 'no benefit', determined empirically by testing the difference with 1000 random labellings.

we remove all 14 signature genes from the dataset and rerun STLsig. The new signature, which contains 312 genes from 85 gene networks, results in an HR of 0.56 (p = 0.23, 95% CI = 0.23 - 1.41) in the training data, which is worse than the original signature. The performance on the patients in fold D, which requires optimizing a new zPFS threshold, also yields a worse performance (HR of 0.59; p = 0.06, 95% CI 0.34 - 1.01; n=130 in the 'benefit' class). Moreover, changing this threshold to yield a differently sized class 'benefit' does not yield performances that approach that of the original 14-gene signature (**Supplementary Figure 8**). Together, these results establish that the 14 identified genes are essential to the performance of the model.

Multiple signature genes are associated with MM or proteasome inhibition

Having established the genes in the signature are essential to the performance, we investigate how the genes in the signature may be involved in determining PI benefit. Interestingly, in addition to having the largest impact when its information is lost, *DAB2IP* is also the only gene that is significantly differentially expressed between class 'benefit' and 'no benefit' (p = 0.002). *DAB2IP* plays an essential role in the *IRE1*-

mediated ER stress response and inducing apoptosis via the *JNK* pathway (Luo et al. 2008). Apoptosis induced by ER-stress is one of the main working mechanisms of proteasome inhibitors (Moreau et al. 2012).

While none of the other signature genes are differentially expressed between class benefit and no benefit, several genes do have a clear link to cancer or MM specifically. For instance, NES is a stem cell marker that is not found in healthy plasma cells, but is found specifically in MM (Svachova et al. 2014). Moreover, *NES* has been associated with treatment response in MM. CLIP1 is involved in microtubule-kinetochore attachment and plays a role in proper chromosome alignment during mitosis (Amin et al. 2014) and has been associated with cancer progression and chemotherapy resistance (Sun et al. 2012), though not in relation to MM. SNX9 is described to play an important role in trafficking ADAM9 to the cell surface (Mygind et al. 2018). *ADAM9* is expressed in MM cells and induces IL6 production by osteoblasts, potentially creating a more permissive bone marrow environment for MM cell proliferation (Karadag, Zhou, and Croucher 2006). One of the described working mechanisms of bortezomib is the downregulation of the production of IL-6 in the bone marrow environment (Karadag, Zhou, and Croucher 2006; Roccaro et al. 2006). The gene *TPD52L1* is a negative regulator of *ATM* (Chen et al. 2013), which is involved in the DNA damage response and activated by bortezomib treatment (Hideshima et al. 2003). *ST6GAL2* has been described before to be significantly downregulated in carfilzomib-resistant cell lines (Zheng et al. 2017).

Together, this indicates that our signature is not only capable of predicting benefit but could also aid in understanding differential response to PI treatment.

Different cellular response to bortezomib in class benefit

For 142 patients in the HTT cohort tumor gene expression was measured again 48 hours after receiving bortezomib. To investigate whether the cellular response to bortezomib is different for patients classified as 'benefit', we performed a differential expression analysis before and after treatment separately in class 'benefit' and class 'no benefit' using SAM (Tusher, Tibshirani, and Chu 2001). Because of the low number of patients in class 'benefit' for whom a second measurement is available, we relax our definition of benefit and classify patients as 'benefit' if the calculated zPFS>0 (n = 71). We find 12

genes that are significantly differentially expressed before and after treatment in class 'benefit' but not in class 'no benefit'. We also find two genes that are significantly differentially expressed only in class 'no benefit' (**Figure 4b**). To identify the genes that truly represent a different cellular response in class 'benefit' and 'no benefit', we compute the difference in fold change between both classes. To ensure that this is not a random difference, we also compute this difference for all genes using 1000 random class labellings. We find four genes - *TNS3, PXN*, *C2CD4A* and *PSPC1* - where the difference between 'benefit' and 'no benefit' is larger than expected by random chance ($p < 0.05$ after Bonferroni correction for multiple testing). None of these genes have been linked to MM, though all have been connected to disease progression in other cancer types (Carter et al. 2013; Wu et al. 2010; Yao et al. 2015; Yeh et al. 2018). Interestingly, *TNS3*, *PXN* and *PSPC1* are all described to play a role in cell adhesion and a migratory phenotype (Yeh et al. 2018; Mouneimne and Brugge 2007). Cell adhesion mediated drug resistance (CAM-DR) has been described extensively in MM (Damiano et al. 1999; Landowski et al. 2003; Damiano and Dalton 2000). Moreover, it has been suggested that bortezomib can overcome CAM-DR (Hatano et al. 2009; Yanamandra 2006). A different regulation of cell adhesion in class 'benefit' could play a role in the observed benefit to PIs.

Discussion

In this work we propose STLsig, a method to identify interpretable signatures that robustly predict patient benefit to PIs from a gene expression measurement at time of diagnosis. The 14 gene signature, derived with our method, validates on an independent patient cohort which was moreover measured on a different platform, confirming the robust nature of the signature.

A clinical trial setting is most suitable for training the STLsig model. Here treatment is randomized and each treatment arm contains roughly the same number of patients. This is important for calculating zPFS and training the signature, as it ensures each patient has sufficiently similar neighbours in the gene expression space. Once the signature is trained, it can be validated on a less balanced dataset. We therefore used the HTT cohort as training data, rather than the newer CoMMpass dataset. It should be noted that the treatment combinations used in the HTT cohort are no longer

representative of clinical practice; since sufficiently long follow up is needed to train a model, we necessarily train on older data. Recently, it was shown that daratumumab, combined with bortezomib, thalidomide and dexamethasone (VTd), is superior to VTd (Moreau et al. 2019). This will arguably be the new standard, but since daratumumab is relatively new, suitable gene expression datasets are not available. It is clear that PIs continue to play an important role in MM treatment. In the CoMMpass dataset, we show that the performance of our signature remains stable in different treatment combinations and that - while it was trained on bortezomib - also can predict benefit to carfilzomib. Moreover, we show that the addition of lenalidomide to bortezomib based treatment only leads to better survival in the 'no benefit' group. This establishes our model is also relevant in a more modern, chemotherapy free setting. We also demonstrate our signature can be applied to patients for which the expression profiling was performed using RNAseq, demonstrating cross-platform robustness.

We have only considered gene expression patterns in this research since it has been shown that for classifiers aimed at predicting cancer survival, gene expression captures the majority of the signal (Aben et al., 2018). More specific to MM, Chapman et al found bortezomib response could not be reliably predicted from mutation events (Chapman et al. 2018). The mutational landscape in MM is quite sparse and we find no difference in mutation burden or in the specific genes that are mutated between class 'benefit' and 'no benefit' (**Supplementary Figure 9**). We also do not see a difference between class 'benefit' and 'no benefit' in the 63 driver genes that were recently identified (Walker et al. 2018) (**Supplementary Figure 10**). While MM is a very complex disease and this complexity can most likely not be captured in only two groups differentiated by gene expression patterns, the signature identified can aid in optimal treatment selection and thus has direct clinical applicability.

Several of the genes in the signature are already described to be involved in the proteasome system or disease progression in MM and we show these genes are essential for the predictive performance, as no equivalent signature can be found without them. These findings reinforce the importance of the selected genes and indicate the power of STLsig to further elucidate proteasome inhibitor specific mechanisms.

STLsig can readily be applied to other diseases and drugs. A very potent application could be to perform post-hoc analysis of clinical trial data for drugs which missed their endpoint. Such analysis could reveal a subset of patients who would still benefit from the drug, thus potentially extracting valuable information from failed clinical trials.

Taken together, we provide a powerful machine learning approach to aid in treatment decisions in the clinic, ensuring a more optimal treatment choice and ultimately improve patient outcomes.

Availability of data and material

The datasets supporting the conclusions of this article are available on GEO. Gene expression data from the HOVON-65/GMMG-HD4 study is available at GSE19784 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19784). Gene expression data from both Total Therapy 2 and Total Therapy 3 are available at GSE2658 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2658) . 30 patients from the Total Therapy 3 study used in the manuscript are not included in the GSE2658 dataset, these can be found in ArrayExpress dataset E-TABM-1138 (https://www.ebi.ac.uk/arrayexpress/experiments/E-TABM-1138/). The PFS survival data for all three studies are available at https://github.com/jubels/GESTURE, linked to the GEO and ArrayExpress IDs. All gene expression and survival data for the CoMMpass study is available at research.themmrf.org

All code needed to discover and validate the signature is available at https://github.com/jubels/STLsig. All code requires R and is platform independent.

Funding

J.U. is supported by a PhD fellowship from the Van Herk Charity foundation

Acknowledgements

The CoMMpass dataset was generated as part of the Multiple Myeloma Research Foundation Personalized Medicine Initiatives (https://research.themmrf.org and www.themmrf.org).

3

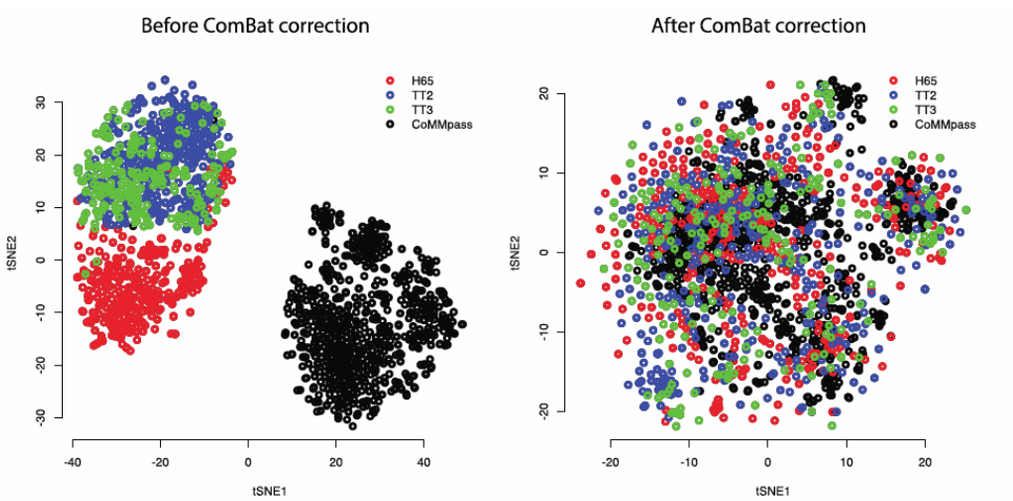
Supplementary material

Supplementary table 1 Log2 fold difference of signature genes between class 'benefit' and 'no benefit' in fold D of the HTT cohort

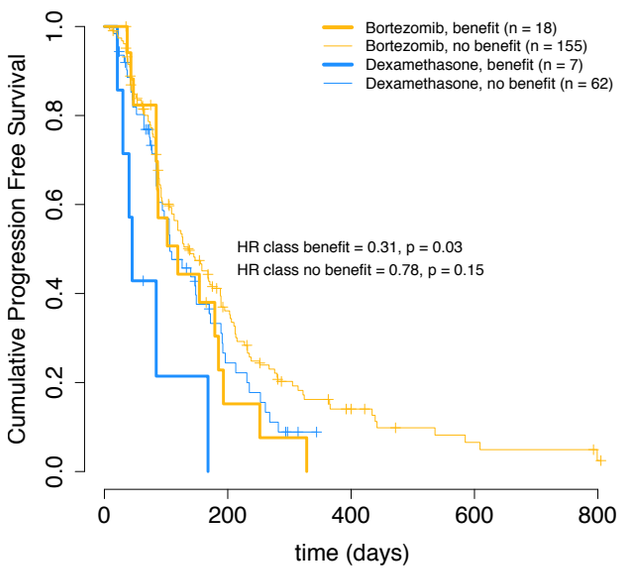
	Mean log2 fold difference	p-value
NEXN	0.27	0.16
DAB2IP	-0.65	0.002
CFAP53	-0.06	0.72
TPD52L1	0.20	0.34
SHTN1	0.38	0.05
ZNF493	-0.015	0.95
NES	0.10	0.77
CLIP1	0.23	0.26
LINC00485	0.15	0.49
ST6GAL2	0.26	0.14
EBF2	-0.14	0.44
LINC00992	0.15	0.44
FA2H	0.05	0.74
SNX9	-0.006	0.98

Supplementary table 2. Overview of the number of patients who received a form of chemotherapy

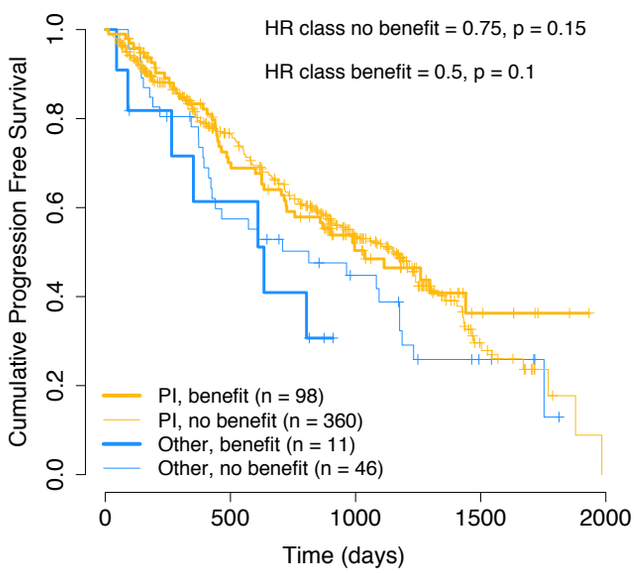
	Class benefit	Class no benefit	Total
Doxorubicin	0 (0%)	4 (0.7%)	4 (0.5%)
Cyclophosphamide	28 (18.8%)	168 (28.1%)	196 (26.2%)
Melphalan	12 (8.1%)	20 (1.7%)	32 (4.2%)
Vincristine	0	0	0



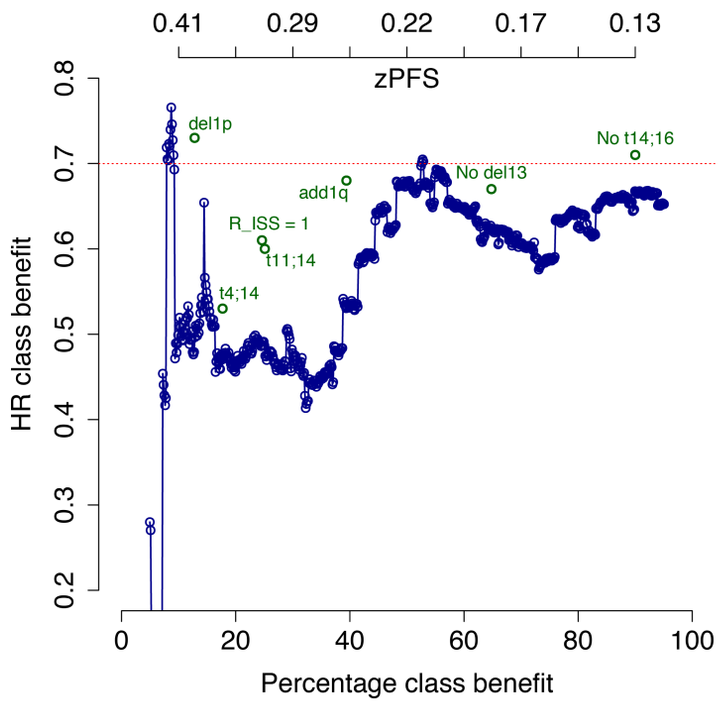
Supplementary figure 1. tSNE of the datasets before and after batch correction with ComBat.



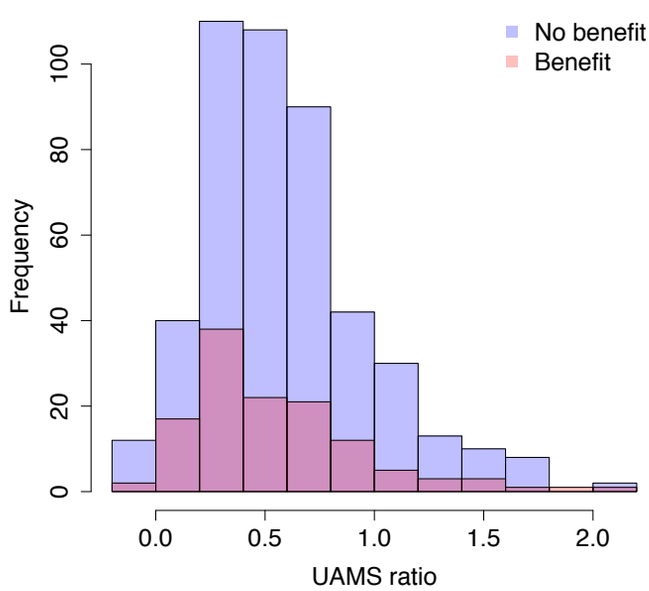
Supplementary figure 2. Kaplan Meier plot of the performance of the signature in the APEX study.



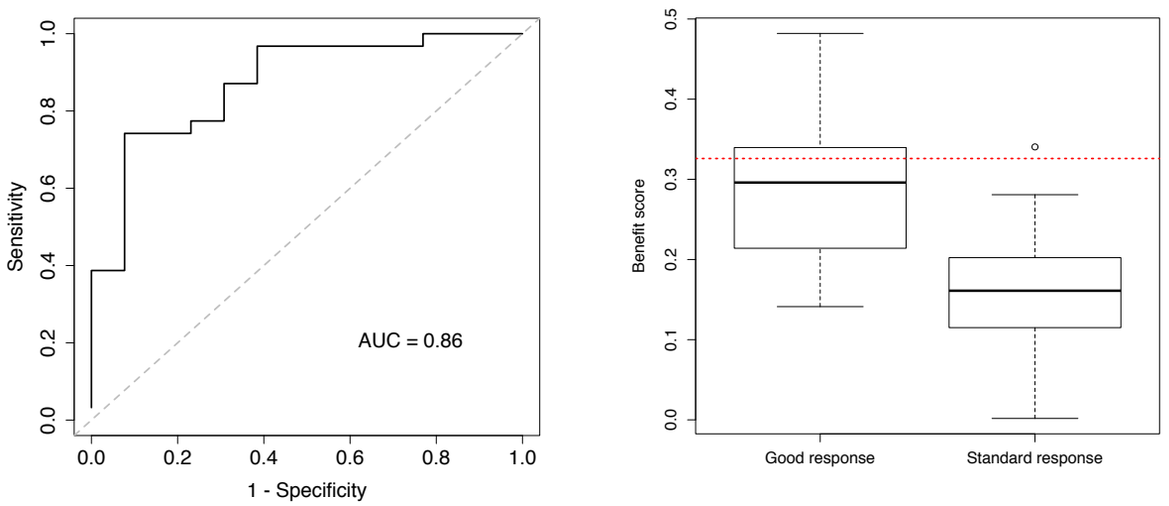
Supplementary figure 3. Kaplan Meier plot of the performance of the signature when patients who received chemotherapy are removed from the CoMMpass validation set.



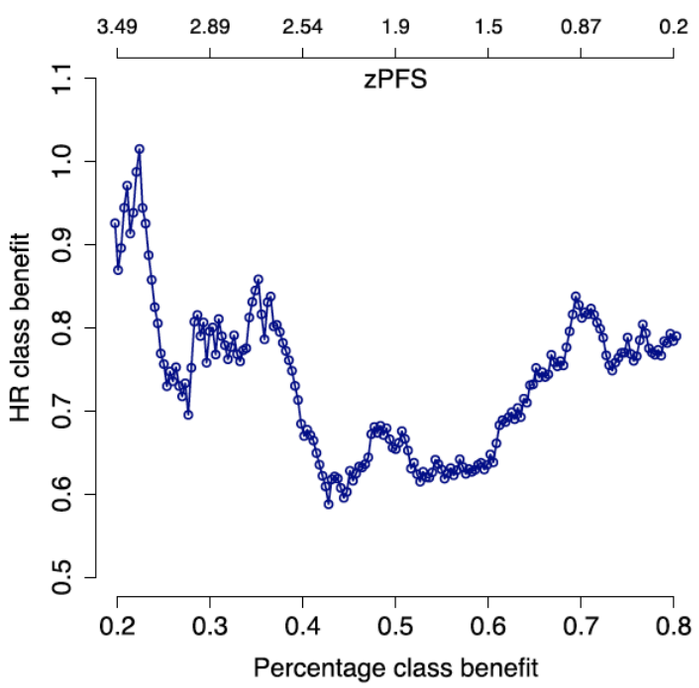
Supplementary figure 4. Performance of several known markers in predicting benefit to PI treatment. The blue line represents the performance of our signature at different size class 'benefit', the red dotted line represents the HR as found in the dataset as a whole.



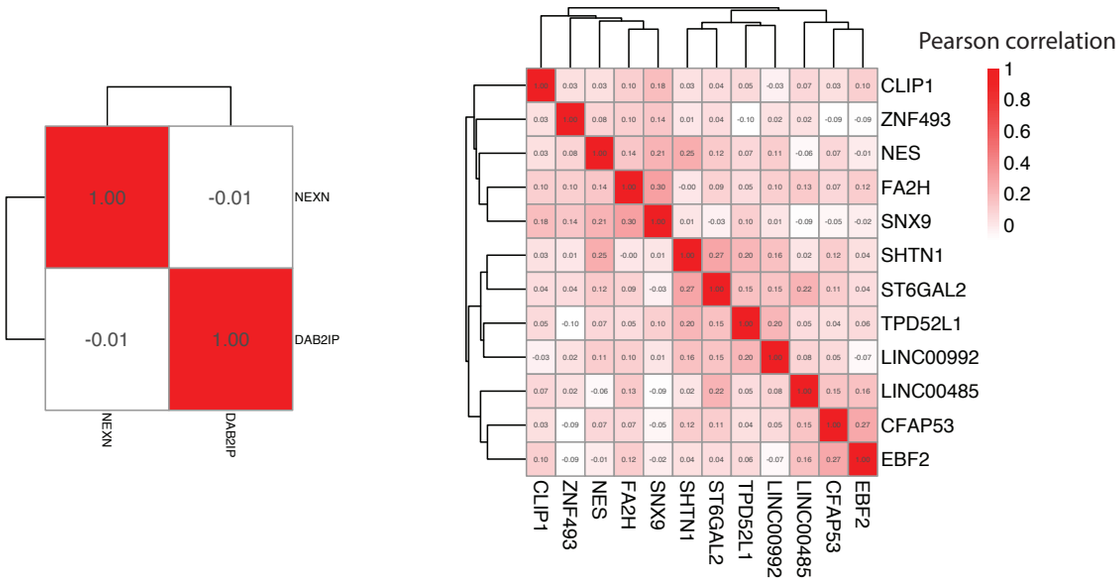
Supplementary figure 5. Distribution of UAMS ratio in class 'benefit' and class 'no benefit'.



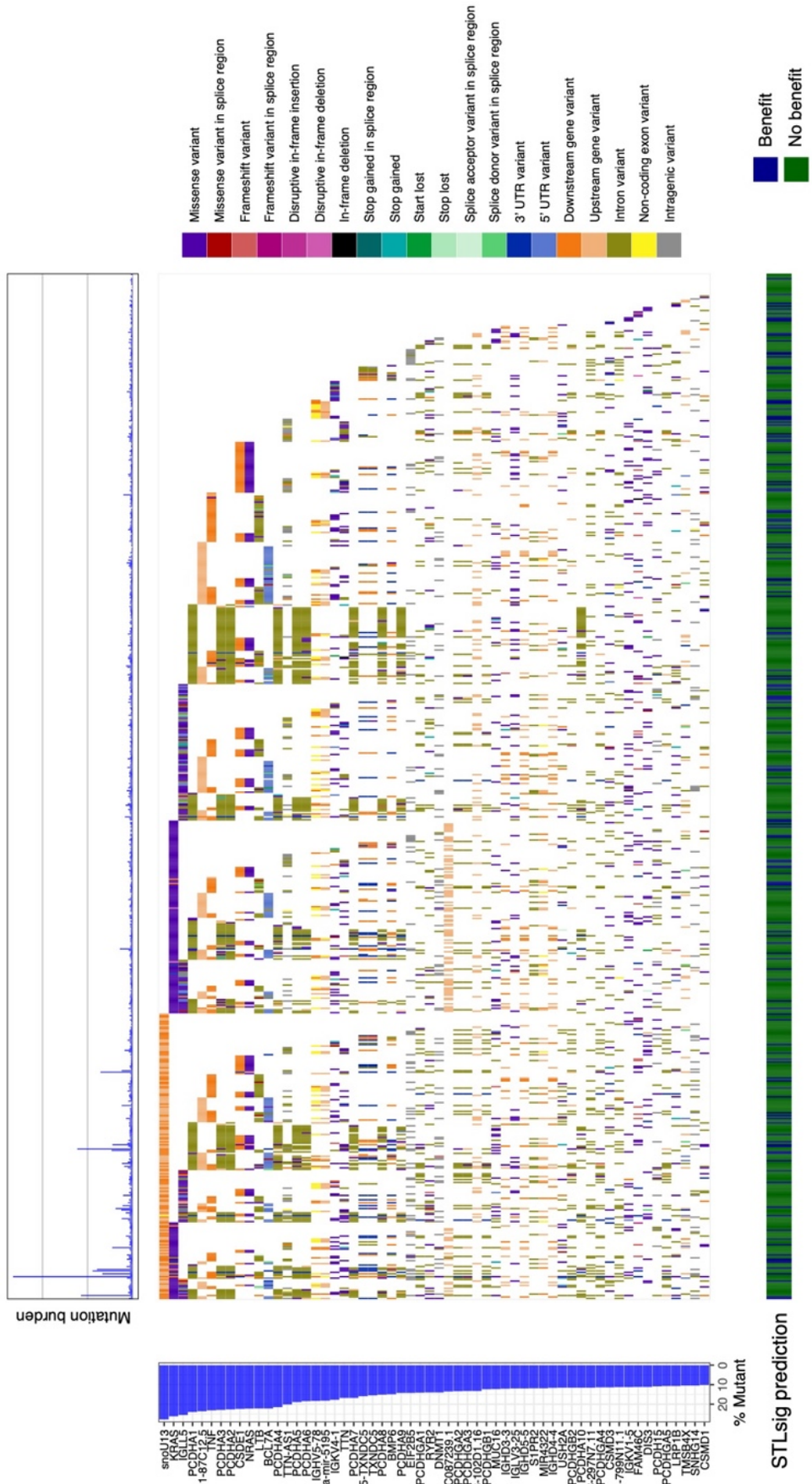
Supplementary figure 6. ROC curve of the performance of our signature in the PADIMAC dataset and boxplot of the benefit score for good and standard responders. The red dotted line represents the cutoff for class 'benefit'.



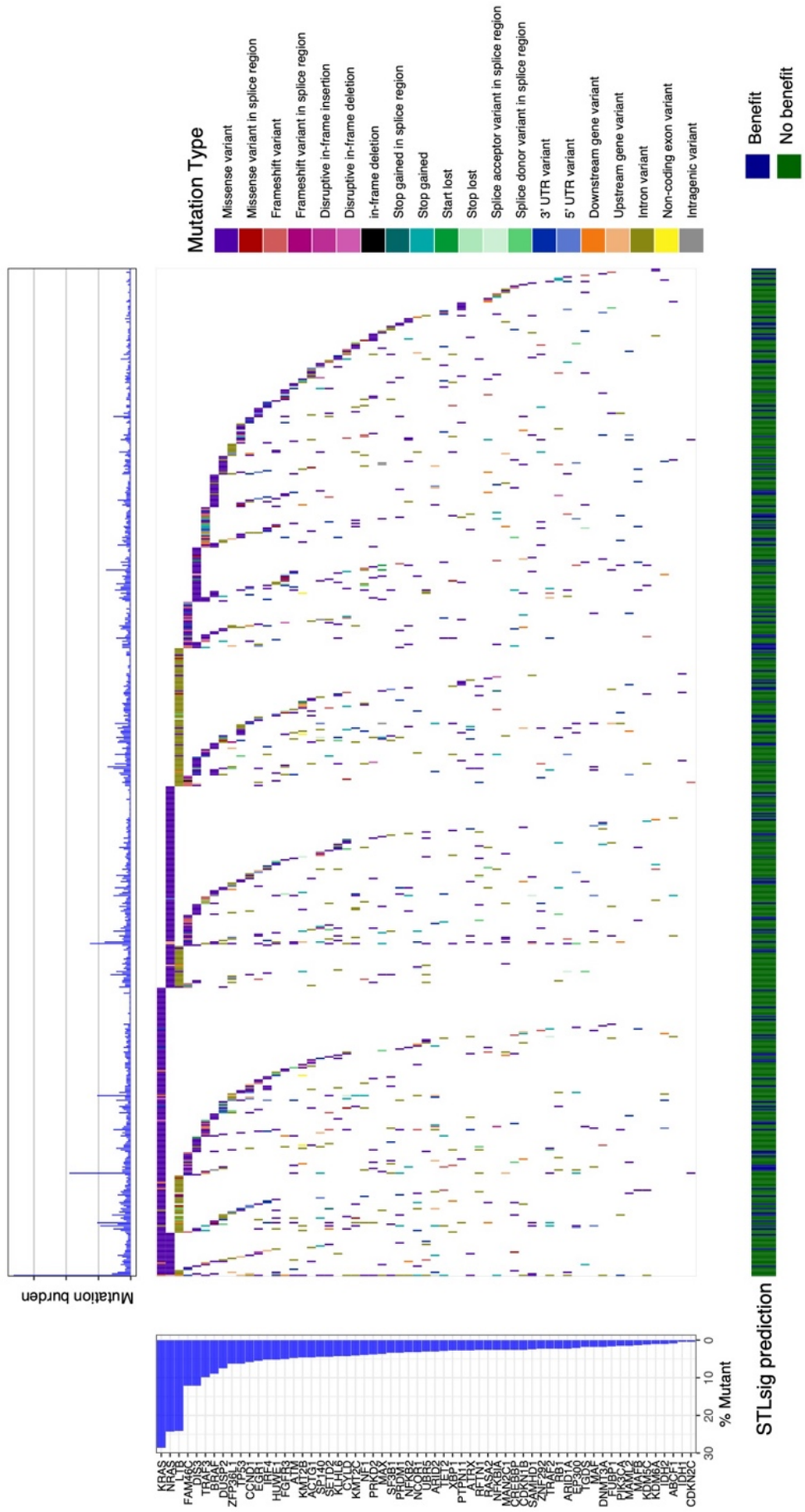
Supplementary figure 7 The HR found in class 'benefit' using different cutoffs for zPFS, as predicted by the signature found when excluding the original 14 genes from the analysis.



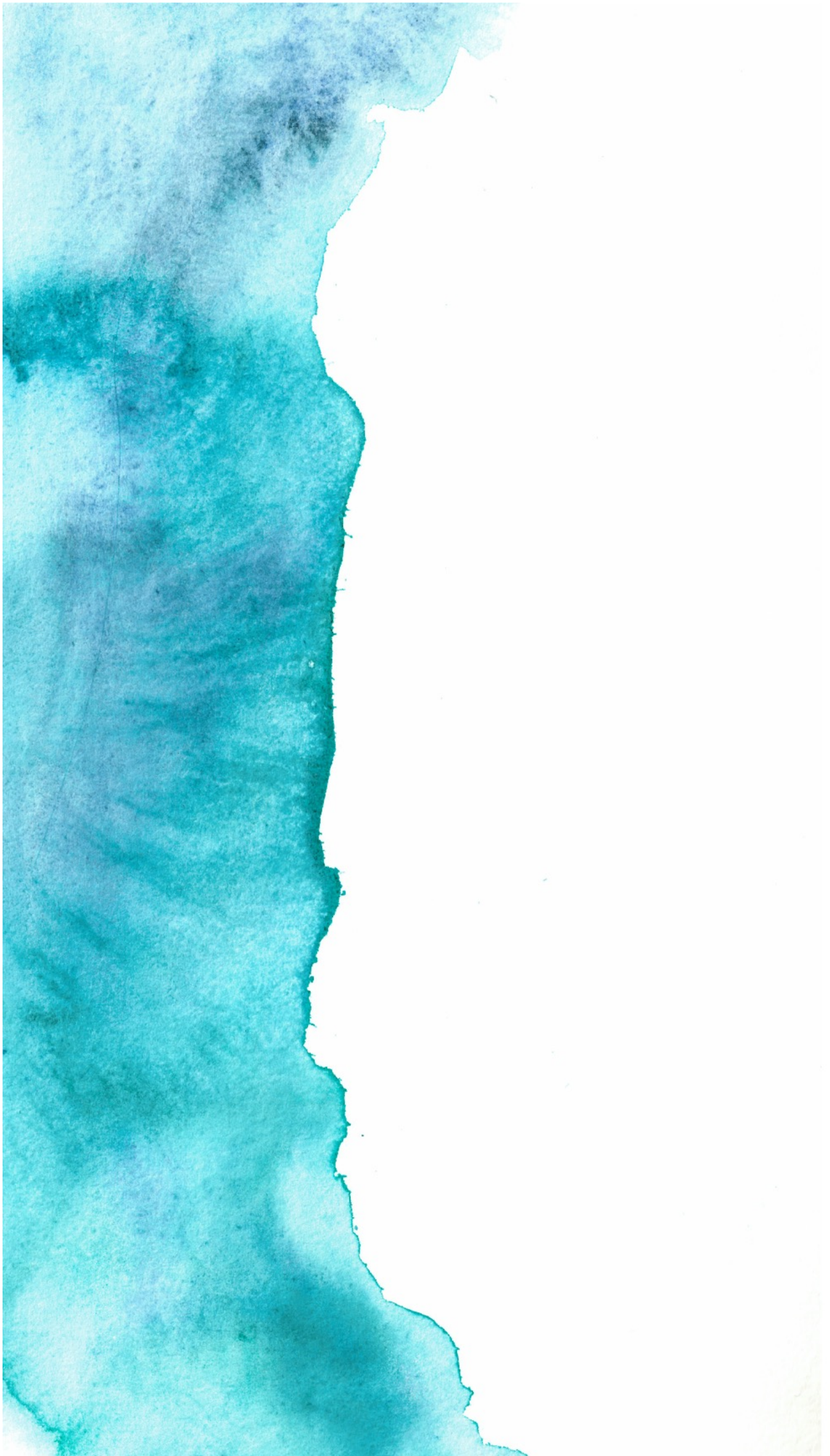
Supplementary figure 8. Correlation matrix of the Pearson correlation between the gene expression of the genes included in the signature, per gene network included in the signature.



Supplementary figure 9. All non-synonymous mutations found in more than 10% of the samples in the CoMMpass dataset per patient, with the lower bar indicating whether the patient is predicted to benefit or not.



Supplementary figure 10. Overview of the mutations found in the 63 driver mutations identified by Walker et al, with the lower bar indicating whether the patient is predicted to benefit or not.



Chapter 4

Predicting treatment benefit in data with low event rates and non-randomized treatment arms: chemotherapy benefit in breast cancer

Joske Ubels[1,2,3,4], Martin H. van Vliet[4], Jeroen de Ridder[1,2*]

1. Center for Molecular Medicine, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG, Utrecht, The Netherlands 2. Oncode Institute, Utrecht, The Netherlands 3. Department of Hematology, Erasmus MC Cancer Institute, Wytemaweg 80, 3015 CN, Rotterdam, The Netherlands 4. SkylineDx, Lichtenauerlaan 40, 3062 ME, Rotterdam, The Netherlands

Abstract

Selecting the best treatment for each patient remains a challenge in cancer. We have previously developed the GESTURE algorithm, which is designed to predict whether a patient will benefit more from a treatment than an alternative using tumor gene expression. We here adapt it to deal with survival data with few events, to predict chemotherapy benefit in breast cancer. We show it can successfully identify a subgroup of patients who benefit more from chemotherapy than the population as a whole. Importantly we also identify a group who does not see a significant benefit from chemotherapy and can thus be spared the side effects. The classifier does not validate in other external data with a different patient composition and treatment regimen, highlighting the importance of matching the patient population included in the training and validation set, and by extension the intended use population.

Background

Personalized medicine has received increasing attention in cancer. However, selecting the best treatment for each patient remains a challenge in almost all cancer types, especially when treatments targeting a specific mutation are not available.
We have previously developed GESTURE (Gene Expression-based Simulated Treatment Using similaRity between patiEnts), an algorithm that can predict treatment benefit using gene expression and survival data as input (Ubels et al. 2018). We define treatment benefit as having a superior survival on the treatment of interest to the survival had this patient been given an alternative treatment. This is a challenging problem, as we can only observe the response to the treatment a patient actually received. Even when a patient did not experience a good outcome on a certain treatment, they may still have benefited, as their outcome may have been even worse on another treatment. In GESTURE we therefore implement the concept of Simulated Treatment Learning (STL). STL relies on the idea that a genetically similar patient who received a different treatment can be used to model the response to a treatment the patient did not receive. We need to define this similarity between patients with genes that are relevant to treatment benefit. In GESTURE we use many different gene sets based on Gene Ontology (GO) annotation to define similarity; we then build a classifier out of the gene sets that are most successful at identifying so-called prototype patients. These are patients who experience more benefit from the treatment than similar patients who did not receive the treatment; new patients who are similar to such a prototype are also expected to benefit from the treatment.

We developed GESTURE and demonstrated its performance in Multiple Myeloma (MM), a type of plasma cell cancer. However, GESTURE is not specific to MM and can be applied to any dataset where gene expression, survival data and two treatment groups are available.

With roughly 1.7 new million cases per year, breast cancer is the most common cancer for women and one of the leading causes of death in women worldwide (Sharma 2019). Breast cancer was one of the first cancer types where gene expression was used to predict disease progression (Veer et al. 2002). Moreover, it was later shown in a prospective clinical trial that this 70 gene classifier (the MammaPrint) can predict

which women can safely forego chemotherapy (Audeh et al. 2019; Cardoso et al. 2016). Breast cancer represents a good test case for applying GESTURE in a new disease, as it is known that information relevant to treatment choice is captured in the gene expression. While the MammaPrint was developed to predict 5-year survival and was later found to be relevant to treatment benefit, GESTURE can train classifiers that are optimized to predict treatment benefit.

While GESTURE is not specific to MM, the algorithm as is does not achieve a satisfactory performance on the breast cancer data. This may be expected, as there are certain key differences between the clinical reality of both diseases. MM is an incurable disease, with a median survival of 5-6 years (Rajkumar and Vincent Rajkumar 2018). For breast cancer, the average 5-year survival rate is 90% (SEER statistic). Because of this higher survival rate and thus lower number of events, we have to adapt the optimization criterion GESTURE uses for training classifiers. For clarity, from here on we call the adapted version GESTURE-BC. GESTURE defines the best classifier as the one that can identify a subset of patients - class 'benefit' - with the largest difference in survival between the two treatment arms, as defined by the hazard ratio (HR). However, in breast cancer a subset without any events in the treated arm can easily be identified. A better HR cannot be achieved, even if the other treatment arm also has very good survival. Applying GESTURE to breast cancer thus leads to a class 'benefit' where patients in both treatment arms survived well, and a class 'no benefit' (i.e. all patients not in class 'benefit') where we also find a significant HR in favor of the treatment (Supplementary figure 1). Such a classifier cannot be used to aid in clinical decisions. We thus adjust the optimization criterion in GESTURE-BC to take both class 'benefit' and 'no benefit' into account, to arrive at a classifier that is clinically useful.

Another consideration is that STL hinges on the idea that similar patients are present in both treatment arms. The ideal setting for training a treatment benefit classifier is thus a clinical trial, where treatment is randomized and similar patients are by definition included in both treatment arms. However, in practice suitable clinical trials data is rarely available and data from clinical practice has to be used. Breast cancer has been well characterized both by gene expression and receptors present on the cell surface. An important consideration in treating breast cancer is the presence or absence of estrogen receptors (ER), progesterone receptor (PR) and human epidermal growth

factor receptor 2 (HER2). There are specific treatments targeting these and triple negative breast cancer (absence of all three receptors) carries the worst prognosis (Kaplan and Malmgren 2008). There are also treatment guidelines taking tumor size, lymph node status (i.e. whether cancer cells have infiltrated the lymph nodes) and cell differentiation into account (Waks and Winer 2019). In breast cancer it is thus most likely not true both treatment arms contain similar patients. When we train in such a cohort, it is likely the signal captured in the classifier is then specific to the setting where the data was gathered. For example, if only node positive patients received chemotherapy, the classifier cannot train on chemotherapy benefit for node negative patients. This classifier will then probably not generalize to a wider breast cancer patient population, where patients with other characteristics did receive chemotherapy.

When clinical trials are impossible or simply not (yet) available, one could potentially still leverage datasets with non-random treatment groups by computationally matching the patients characteristics between the two groups. A possible approach is the matching of patients to break the link between certain patient characteristics and the treatment variable, so the model fitted is not influenced by these correlations (Ho et al. 2007). In this approach we use the relevant patient characteristics (i.e. age, tumor size and node status) to calculate the probability a patient received chemo. We then match patients from both treatment arms that have a similar probability. With perfect matching, there is then no longer a correlation between these variables and the treatment, so they will not bias the classifier. While perfect matching is often not possible, this approach can still reduce the bias.

We here show GESTURE can be successfully adapted to fit the clinical reality of breast cancer and find a classifier that can predict chemotherapy benefit in cross-validation. We show its performance in 2273 ER positive and ER treated patients from the Sweden Cancerome Analysis Network - Breast cancer cohort. Our classification cannot characterized by factors currently in the treatment guidelines and we thus identify a new group of patients with chemotherapy benefit. We also show that training a classifier on matched data improves performance on unmatched data from the same population. However, neither a classifier trained on unmatched nor on matched data could show performance in an unrelated, older cohort, highlighting the importance of matching training and validation datasets.

Methods

Algorithm

GESTURE relies on the idea that patients who received different treatments, but have similar tumor gene expression profiles, can be used to model the expected response to an alternative treatment than the one received. Patients with a larger than expected survival difference with similar patients who received a different treatment, can be used as prototype patients. New patients who have a similar gene expression profile to a prototype patient can then be expected to also benefit from that particular treatment. The process of defining similarity and identifying prototype patients has been described in detail before (Ubels et al. 2018). Briefly, to identify prototype patients we first need to define similarity between all patients. Because it is unknown a priori which genes are relevant to treatment benefit and thus should be used to define this similarity, we use gene sets based on Gene Ontology (GO) annotation. We can then build a classifier based on each GO set separately and assess which gene set leads to the best performance. To build a classifier we divide the training data in three equal parts: fold A, fold B and fold C. First we compute the mean survival difference for each treated patient through:

$$\boldsymbol{\Delta} S_i = \frac{1}{n} \sum_{j \in O} (S_i - S_j)$$

where S_i is the overall survival for the treated patient and O the set of the n nearest patients (based on Euclidean distance) who did not receive the treatment of interest. For all training we set the n to 30. We normalize this survival difference by also sampling n random neighbours a 1000 times and calculating ΔS_i , to select patients with a larger survival difference with their neighbours than expected randomly.

The classifier then optimizes two variables on fold B: how many prototypes are used (k) and how close to a prototype a new patient should be to be considered class 'benefit' ($\boldsymbol{\gamma}$).

Previously, for the algorithm developed on MM data, the best classifier (representing one gene set) was defined as the classifier that resulted in the largest hazard ratio (HR) between the two treatment arms within class benefit. The HR in the class 'no benefit' (i.e. all patients not close to a prototype) was not taken into account. However, there are far fewer patients who experienced an event in the breast cancer dataset (9.0%

versus 48.2%). Furthermore, there is already a large HR between the chemotherapy and no chemotherapy arm (HR = 0.45, p = 5.6 * 10-6). When GESTURE only takes the HR in class 'benefit' into account when choosing the optimal values for *k* and $\boldsymbol{\gamma}$, it tends to define a class benefit with no events at all in the chemotherapy arm in the training procedure, since this leads to the best possible performance. However, this classifier is not useful in clinical practice, as we also find a large, significant HR in favor of chemotherapy in class 'no benefit' (**Supplementary Figure 1** shows the cross-validated performance of this classifier).

Therefore, we here define the best performance in GESTURE-BC as the minimum of $\beta_{benefit} - \beta_{no\ benefit}$where the β is the coefficient for the treatment variable in each class as calculated by Cox regression. We then test the optimal classifier on fold C. The β's found in fold C define the performance of this gene set in this repeat.
Since there can be large differences in performance of a gene set when a different division in folds is used, we repeat the procedure 48 times. We take the median performance over 6 repeats at a time and rank the gene sets, resulting in 8 separate rankings (48/6). The final ranking of the geneset is determined by its mean rank over these 8 rankings. We found that this method leads to a more robust ranking than either taking the mean of the 48 separate rankings or calculating the median performance over all repeats. Averaging the rankings, rather than the performance directly, reduces the impact of having a few extremely low HRs, while not disregarding them entirely. The whole algorithm is summarized in **Figure 1**.

This final ranking of gene sets is used to perform forward feature selection; gene sets are added sequentially to form an ensemble classifier. For each repeat we evaluate the performance of this ensemble classifier on fold C; the combination leading to the largest median difference between class 'benefit' and 'no benefit' is selected. For these gene sets a final classifier is trained on all training data to validate on hold-out data.

Data

We train GESTURE-BC on data from the Sweden Cancerome Analysis Network - Breast cancer (SCAN-B) initiative (Saal et al. 2015) (GEO accession: GSE96058). This study included women with breast cancer from centers around Sweden between 2010 and 2013 and measured tumor gene expression with RNAseq. This data is not from a

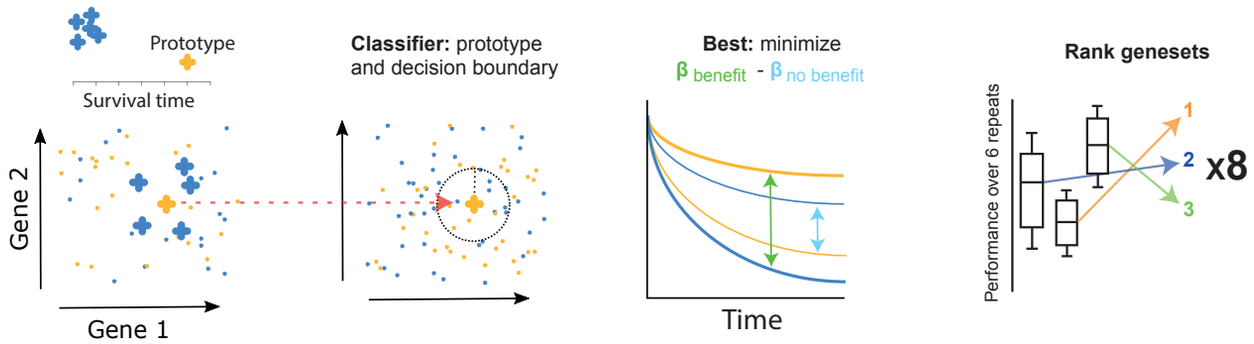


Figure 1. Summary of the training procedure for GESTURE in breast cancer (GESTURE-BC). First prototypes are identified and optimal parameters for the classifier are determined per gene set. The performance of the gene set is then defined on hold-out data by comparing the β's found in class 'benefit' and 'no benefit'. The gene sets are then ranked by mean rank over 8 repeats. The rank for each repeat is in turn determined over 6 repeats of cross validation.

randomized trial and this cohort thus represents current clinical practice. The publicly available data includes RNAseq for 2969 patients, patient characteristics are summarized in **Table 1**. The majority of patients are ER positive. Since survival and treatment strategies differ significantly between these two groups, we exclude all ER negative patients and ER positive patients who did not receive ER treatment from the analysis, which results in 2273 patients.

The only other dataset available which includes the necessary patient information is METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) (Curtis et al. 2012). This dataset includes 1981 patients from the UK and Canada, diagnosed between 1977 and 2005, for whom gene expression was measured with Illumina array.

Fold construction for cross validation

We divide the SCAN-B dataset into three equal folds, ensuring the balance between chemotherapy and no chemotherapy is the same in each fold. Moreover, we ensure the HR between the treatment arms does not differ more than 0.05 between the folds. The two folds used for training each classifier are divided in fold A, B and C in the same manner.

Table 1. Overview of the patient characteristics in the SCAN-B data.

	Chemotherapy (N=1190)	No Chemotherapy (N=1759)	Overall (N=2969)
Age (years)			
Mean (SD)	55.2 (11.3)	67.8 (11.6)	62.8 (13.1)
Median [Min, Max]	55.0 [24.0, 82.0]	68.0 [34.0, 96.0]	64.0 [24.0, 96.0]
Size (mm)			
Mean (SD)	21.9 (12.6)	18.5 (11.4)	19.9 (12.3)
Median [Min, Max]	20.0 [0, 125]	15.0 [1.00, 126]	17.0 [0, 126]
Missing	27 (2.3%)	5 (0.3%)	32 (1.1%)
Positive nodes			
No	544 (45.7%)	1254 (71.3%)	1816 (61.2%)
Yes	612 (51.4%)	450 (25.6%)	1064 (35.8%)
Missing	34 (2.9%)	55 (3.1%)	89 (3.0%)
ER			
Negative	167 (14.0%)	41 (2.3%)	214 (7.2%)
Positive	885 (74.4%)	1673 (95.1%)	2569 (86.5%)
Missing	138 (11.6%)	45 (2.6%)	186 (6.3%)
HER2			
Negative	840 (70.6%)	1633 (92.8%)	2490 (83.9%)
Positive	309 (26.0%)	67 (3.8%)	378 (12.7%)
Missing	41 (3.4%)	59 (3.4%)	101 (3.4%)
PGR			
Negative	217 (18.2%)	124 (7.0%)	347 (11.7%)
Positive	785 (66.0%)	1514 (86.1%)	2310 (77.8%)
Missing	188 (15.8%)	121 (6.9%)	312 (10.5%)

Matching patients between treatment arms

We calculate a propensity score for receiving chemotherapy for each patient using logistic regression with $chemotherapy \sim age + node\ status + tumor\ size$. All patients with missing values for one of these variables are excluded. We also exclude all patients older than 82, as there are no older patients in the chemotherapy arm. Based on the propensity score, we match each chemotherapy treated patient to one untreated patient. These pairs are chosen so total distance between all the pairs is minimized. This is implemented in the R matchIt package, using the "optimal" setting (Ho et al. 2011)

Construction of gene sets

We only use genes measured in both the SCAN-B and the METABRIC dataset, which results in 16,789 unique genes. We define GO sets with the R package goSTAG (Bennett and Bushel 2017) and keep all sets which included more than one and less than a thousand genes, which results in 9,578 gene sets. We use the FPKM values for SCAN-B and the log2 normalized data from METABRIC. We then perform a batch correction with ComBat (Johnson, Li, and Rabinovic 2007), with METABRIC and SCAN-B as the batches. We then perform a quantile normalization, so measurements from both datasets are directly comparable.

Results

Cross-validation on SCAN-B leads to a significant HR in class benefit

We perform 3-fold cross validation on the SCAN-B dataset. In the dataset as a whole an HR of 0.45 ($p = 6*10^{-6}$) in favor of chemotherapy is found (**Figure 2a**). When we classify the population with the GESTURE-BC classifier we find a class 'benefit' comprising 70.2% of the dataset with an HR of 0.31 ($p = 6*10^{-7}$) in favor of chemotherapy (**Figure 2b)**. The HR in class 'no benefit' (n =685) is 0.89 (p = 0.67). This performance is fairly stable across the three cross validation folds. In Fold 1 74.3% of the patients are classified as 'benefit', which results in an HR of 0.31 (p = 0.0005) in class 'benefit' and an HR of 1.46 (p = 0.50) in class 'no benefit'. The classifier validated on Fold 2, classifies 55.2% of the patients as 'benefit', which results in an HR of 0.27 (p = 0.007) in class 'benefit' and an HR of 0.74 (p = 0.45) in class 'no benefit'. The Fold 3 classifier classifies 81.0% of patients as 'benefit', which results in an HR of 0.32 (p = 0.01) in class 'benefit' and an HR of 0.74 (p = 0.54) in class 'no benefit'. The Kaplan Meiers of these classifications are

shown in **Supplementary Figure 2**. The classifiers validated on Fold 1, Fold 2 and Fold 3 use 10, 31 and 64 gene sets respectively.

The ensemble classifier is formed by classifying all patients with each gene set separately. The benefit score of a patient is then defined by the number of gene sets that classify the patient as 'benefit'. To define the final class 'benefit' a threshold *t* is set, where a patient is classified as class 'benefit' when their benefit score is above the threshold. This threshold *t* is optimized by the difference in β between the classes, with the constraint that both classes contain at least 20% of the patients and the HR in class 'benefit' is significant at $p < 0.05$. There is a trade-off between class size and HR in setting *t*.

When we vary the threshold determining class 'benefit' we find that a smaller class 'benefit' is associated with a lower HR (**Figure 2c**), since a higher threshold requires the classifier to be more confident about the classification (i.e. more individual gene sets need to classify the patient as 'benefit'). This shows the performance is not dependent on one specific threshold, but a high score also means more benefit.

GESTURE-BC classification is not characterized by known chemotherapy predictors

There are already many factors known to influence chemotherapy benefit. We first compare our classification with that of the MammaPrint. It has been shown that patients who are predicted to have a good prognosis by the MammaPrint can safely forego chemotherapy, i.e. see no benefit from it. The MammaPrint signature included 70 probes that code for 56 unique genes, which we can all match to a gene measured in the SCAN-B dataset. When we apply the MammaPrint to the SCAN-B dataset we do find a prognostic effect (HR poor prognosis = 1.67, 95% CI 1.13 - 2.46, p = 0.01), but we do not see a predictive effect (**Supplementary Figure 2**). This could be due to the fact that the population in SCAN-B was in part already treated in accordance with this risk classification: 42.8% of the poor prognosis group received chemotherapy, versus just 15.5% of the good prognosis group. For only 65.3% patients the classification of our classifier and MammaPrint overlap, which is according to statistical expectation given a class 'benefit' comprising 70.1% of the patients. Our classifier clearly identifies a different signal, which is to be expected given that we trained specifically on

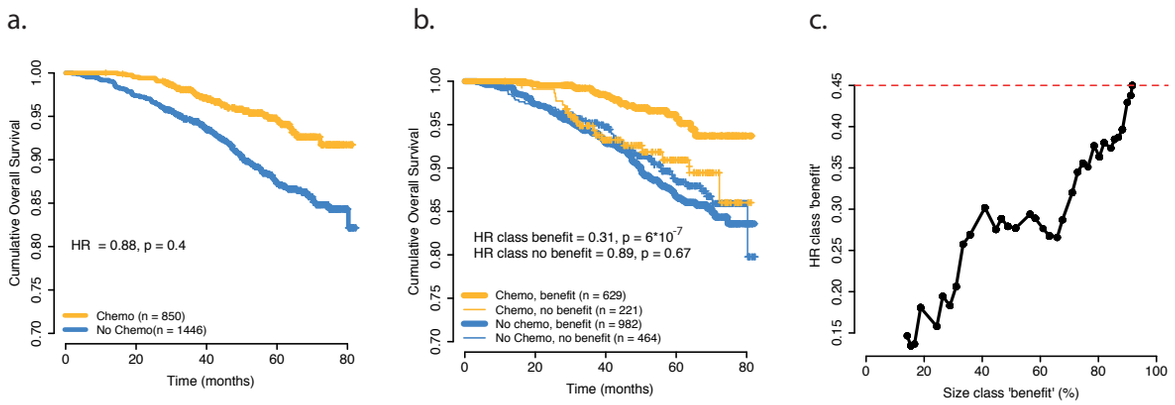


Figure 2. a. Kaplan Meier of the SCAN-B dataset, only including ER positive and ER treated patients. b. Kaplan Meier of the cross-validated performance of the GESTURE classifiers. c. Performance of the GESTURE classifiers using varying thresholds to define class 'benefit'. The dotted line represents the HR found in the dataset as a whole.

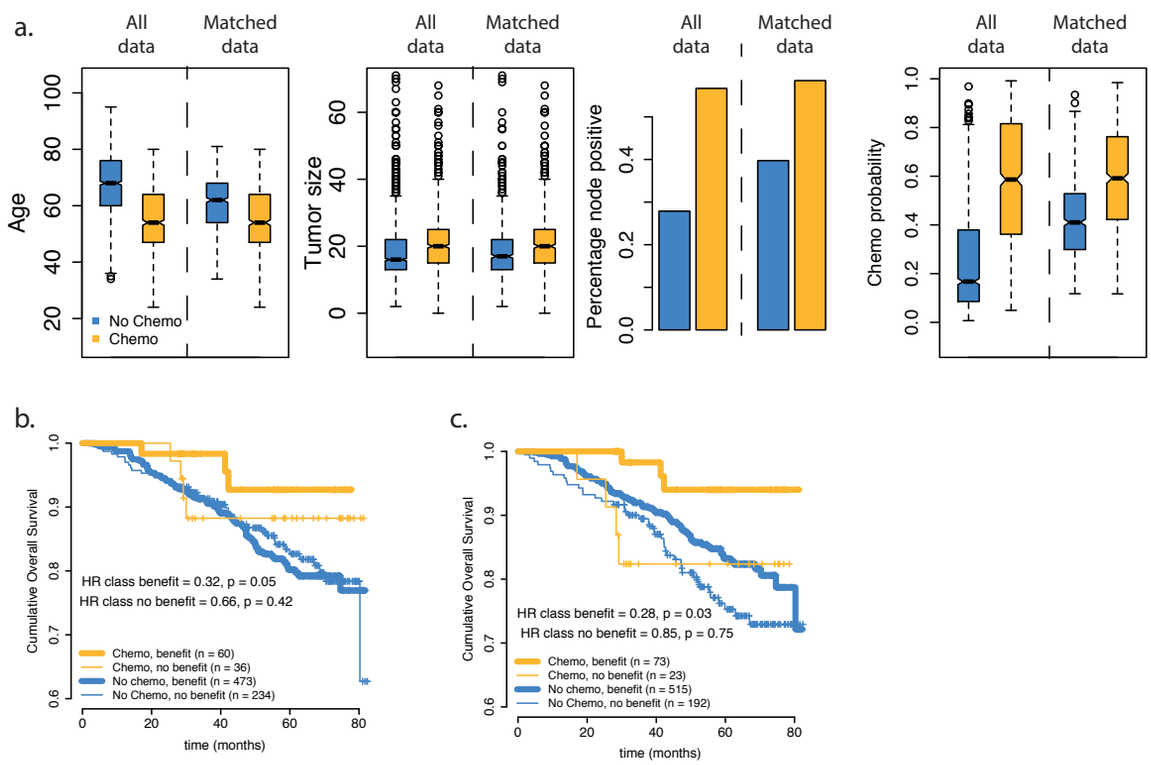


Figure 3. a. Overview of the difference in patient characteristics between the two treatment groups, in all data and in the matched dataset. b. Kaplan Meier of the performance of the classifiers trained on all data on the SCAN-B hold out data. c. Kaplan Meier of the performance of the classifier trained on matched data on the SCAN-B hold out data.

chemotherapy benefit rather than prognosis and the fact that most patients were already treated according to their MammaPrint classification.

Tumor characteristics like tumor grade, tumor size and lymph node status are also included in treatment guidelines. Tumor grade information is not available for the SCAN-B dataset, but tumor size and lymph node status is included. While tumor size is significantly greater in the chemotherapy treated group ($p = 4*10^{-7}$), this is not the case between class 'benefit' and 'no benefit' ($p = 0.97$).

The same holds true for lymph node status, with the percentage of lymph node positive patients similar in class 'benefit' and 'no benefit' (40.8% versus 36.5%). This shows our classifier does not identify patients according to known risk factors and adds new information that can be used clinically.

Classifier on matched data results in a better performance in unseen data

As seen in **Table 1**, patient characteristics vary between the treated and untreated patients, which may impede GESTURE-BC from finding the right predictive signal for predicting treatment benefit. To mitigate this issue, we created a matched population where the difference between the treatment arms is minimized. To this end, we calculate a propensity score per patient that describes the probability of receiving chemotherapy given the age, tumor size and node status of the patient. We then match each treated patient to an untreated patient minimizing the difference in this score over all patients pairs. **Figure 3a** shows the distribution of the characteristics in the data before and after matching. It can be observed that the difference in tumor characteristics cannot be fully equalized with the matching procedure. However, the difference in likelihood of receiving chemotherapy is much smaller in between the two matched groups. The HR in favor of chemotherapy in the matched dataset is 0.85 ($p = 0.49$), which is much higher than in the dataset as whole. In total 803 patients are not included in the matched dataset. These patient samples are used to compare the performance of the cross validated classifiers and the matched classifier. These 803 patients are not matched on patient characteristics and thus represent a better test case for performance in a clinical setting, where treatment is non-randomized. While the classifiers trained on all data do find a class 'benefit' with a larger benefit from chemotherapy than the population as a whole in these patients (**Figure 3b**), the

matched classifier performs better (**Figure 3c**), particularly in identifying a class 'no benefit'. This shows matching a population to simulate a more randomized setting could improve performance, even when validating in a non-matched setting.

GESTURE-BC classifier cannot identify a class 'benefit' in METABRIC data

We next assess the performance of the classifier trained on the matched dataset in the ER positive en ER treated patients included in the METABRIC data. Most patients in this dataset were diagnosed before 2000 and thus represent a different clinical reality. Moreover, only 12.0% of patients received chemotherapy and we find an HR of 1.57 (95 % CI 1.16 - 2.13, p= 0.003) against giving chemotherapy. When we classify these patients with the GESTURE classifier, we find an opposite effect, where class 'no benefit' in fact sees a greater benefit than the population as a whole (**Figure 4a**). Since there are only 21 patients in the no benefit class who received chemotherapy this HR is, however, not significant and could therefore be due to chance. Interestingly, when we train a classifier on METABRIC - matched in a similar manner as with SCAN-B - we see the same effect, with a lower HR in class 'no benefit' (**Figure 4b**). Unfortunately, the METABRIC dataset does not include enough chemotherapy patients to perform a cross validation. It is clear however that the SCAN-B classifier does not validate in the METABRIC dataset. While this could be due to overfitting on the SCAN-B dataset, cross validation did show some signal was captured in the classifier. It could also be that the METABRIC dataset, where an HR not in favor of chemotherapy is found, represents a different population in part due to the fact that these patients were included when clinical practise was different.

HAND2 is present in all classifiers

Finally, we investigate which genes are included in all four classifiers trained on the SCAN-B (three classifiers from cross validation and one trained on the matched data). Of the selected GO categories none are selected in all four classifiers. Also, there is little overlap among the genes within the GO categories selected (**Figure 4c**). The only gene present in all three classifiers trained in cross validation and the classifier trained on the matched data is HAND2. HAND2 is also present in the classifier trained on the METABRIC data. This gene plays an important role in limb development and is associated with progression in endometrial cancer (Jones et al. 2013), but has not been associated with breast cancer. It should be noted that the classifier validated on Fold 1 in the cross validation only included 10 gene sets and 35 genes in total, making overlap

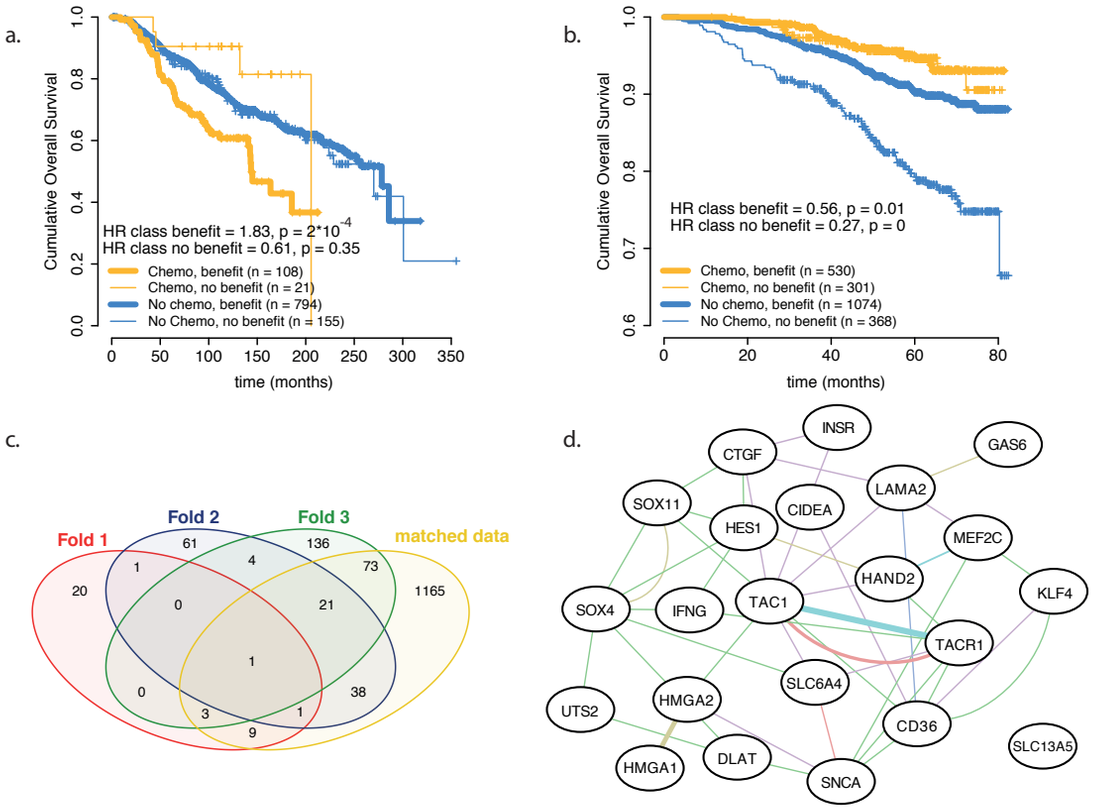


Figure 4. a. Kaplan Meier of the performance of the classifier trained on matched SCAN-B in the METABRIC dataset. b. Kaplan Meier of the performance of the classifier trained on the matched METABRIC dataset on the SCAN-B dataset. c. Overlap of genes between the four classifiers trained on the SCAN-B dataset. d. The overlapping genes in 3 out of 4 classifiers. The edges are inferred by GeneMania. A purple edge indicates co-expression, a green edge a genetic interaction, a light-blue edge a shared pathway, a dark-blue co-localization, a red edge a physical interaction and a yellow edge a shared protein domain.

less likely. The other three classifiers include 22 overlapping genes (including HAND1), which are shown in **Figure 4d**. The network was generated by GeneMania, which links genes based on the interaction described in literature (Warde-Farley et al. 2010). The network is highly enriched for the GO categories "RNA polymerase II core promoter sequence-specific DNA binding transcription factor activity" and "sequence specific DNA binding", both important for the regulation of gene expression. Multiple of these 22 genes have also already been implicated in disease progression and chemotherapy resistance in breast cancer. For example, overexpression of the transcription factor *KLF4* has been shown to be predictive of complete remission in response to neoadjuvant chemotherapy (Dong et al. 2014). Gas6 overexpression has been described to contribute

to chemoresistance (Wang et al. 2016). SOX4 and SOX11 are both related to disease progression in breast cancer (Zhang et al. 2012; Shepherd et al. 2016). While these 22 genes are not essential for performance - as they were not included in the classifier validated on Fold 1 - there are clear links with breast cancer progression.

Discussion

In this work, we show that the GESTURE approach, which we originally developed in the context of MM in which survival rates are poor, can successfully be adapted to breast cancer, which is characterized by much better survival rates. Our classifiers, which in total are based on 101 gene sets containing 368 unique genes, can predict chemotherapy benefit with an HR of 0.31 ($p = 6*10^{-7}$) in class 'benefit'. The classifier represents a different signal than known markers. We also show that though treatment is not randomized in the SCAN-B dataset, the performance of the classifier in non-randomized data can be improved by matching patients between the treatment arms in the training data.

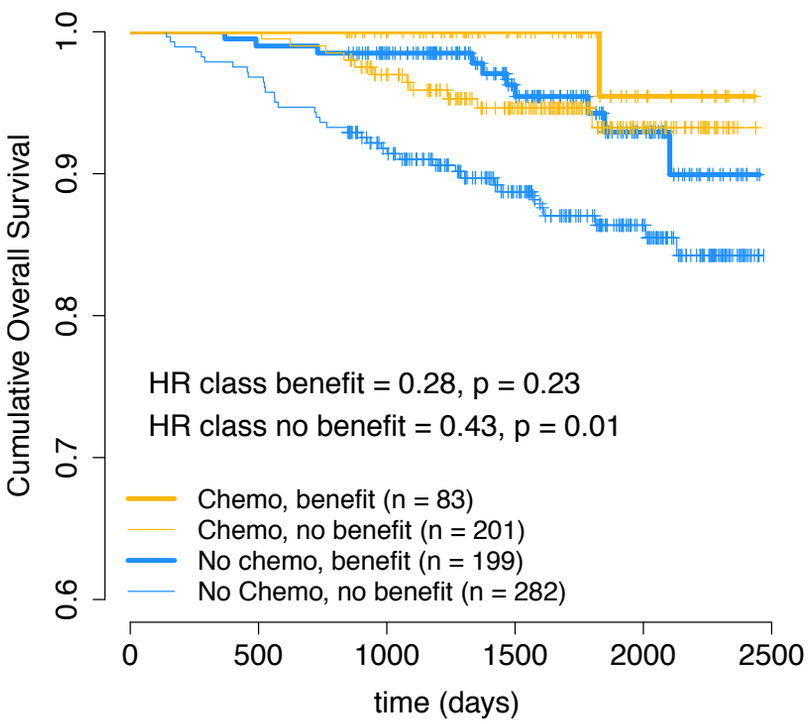
We find that when applied to an external dataset, the METABRIC dataset, the classifier does not show the same predictive behavior. However, it should be noted that the fact that there is no HR in favor of chemotherapy in this dataset already indicates it does not represent the same patient population and clinical setting. More specifically, the patients included in the METABRIC dataset were diagnosed roughly 15 years before the SCAN-B dataset and thus do not represent the same clinical practice. It may therefore not be surprising that a classifier trained on one of the two datasets does not validate on the other. This does highlight the necessity of training and validating in datasets that accurately reflect the patient population for which the classifier is intended. It represents a fundamental challenge in training these models; while the older dataset has longer follow-up, more events and thus more statistical power, the newer data most likely represents the intended use population better. The development of these models is further hampered by the limited number of datasets with all necessary annotations that are made available publicly.

While further validation in a representative test set is necessary, the cross validation and validation of the classifier trained on matched data does indicate that GESTURE

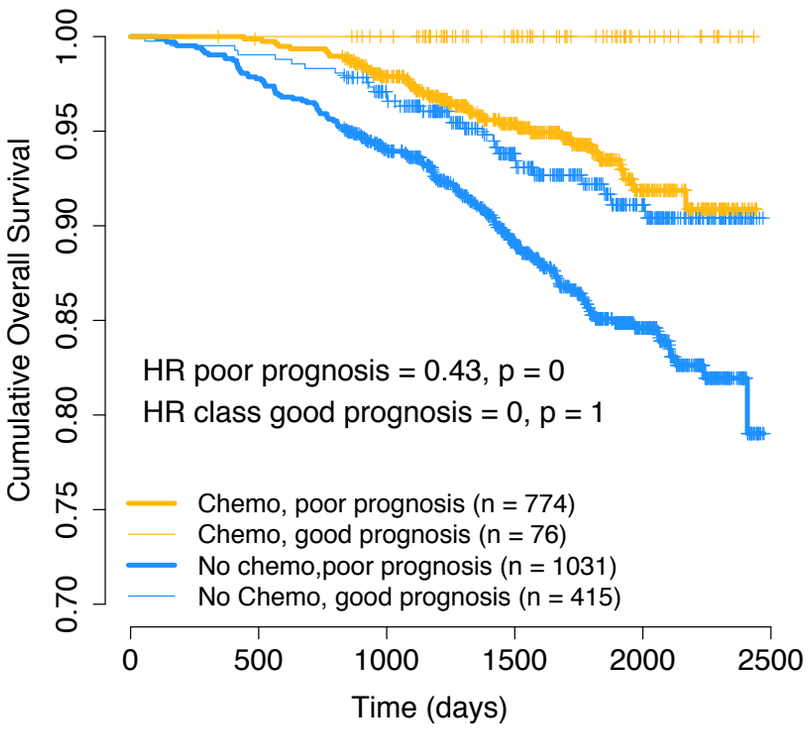
can be successfully adapted to breast cancer data and predict benefit to chemotherapy. GESTURE is thus more widely applicable than the setting it was developed in and can be adapted to different diseases. It could be an important tool in making personalized medicine a reality in more types of cancer.

4

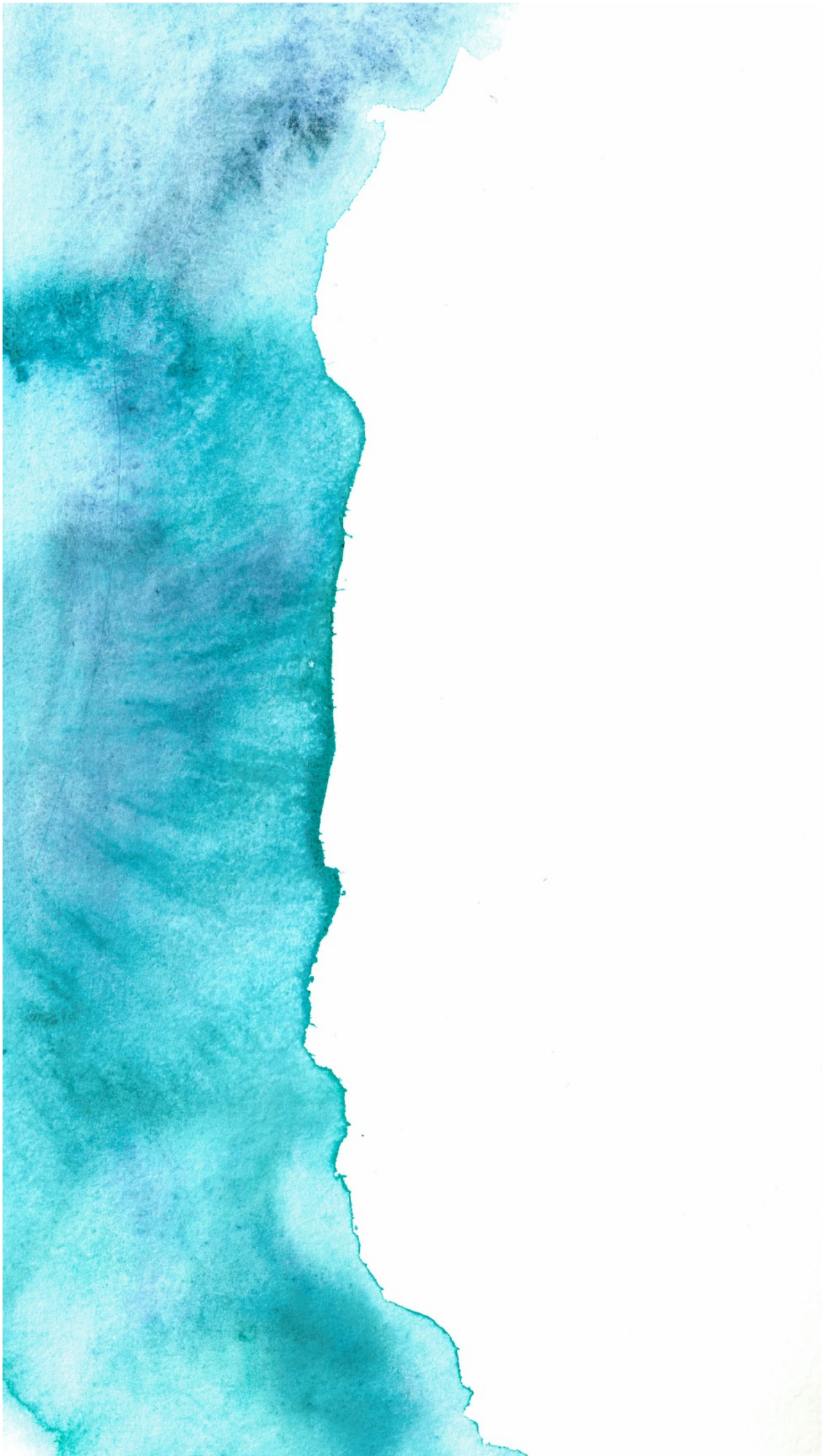
Supplementary material



Supplementary figure 1. Performance of the unadapted GESTURE algorithm in Fold 1.



Supplementary figure 2. Classification of the SCAN-B dataset with the MammaPrint classifier.



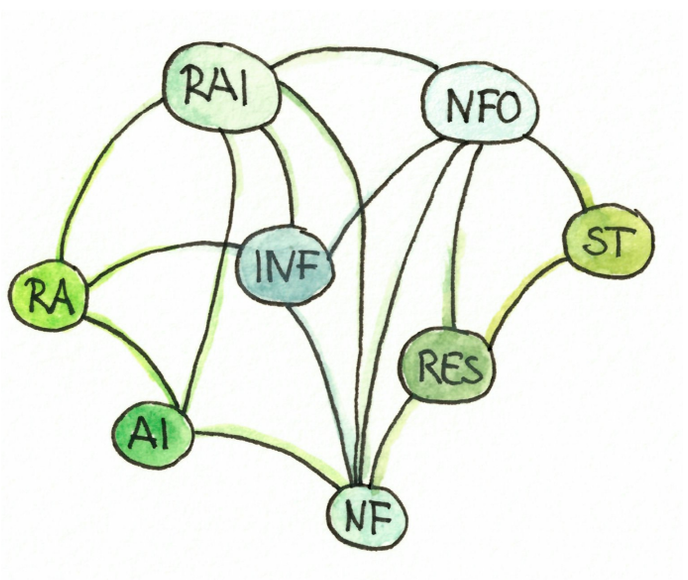
Chapter 5

RAINFOREST: A random forest approach to predict treatment benefit in data from (failed) clinical trials

Joske Ubels[1,2,3,4], Tilman Schaefers[1,4], Cornelis Punt[5], Henk-Jan Guchelaar[6] and Jeroen de Ridder[1,4]*

1. Center for Molecular Medicine, UMC Utrecht, Utrecht, The Netherlands 2. Erasmus MC Cancer Institute, ErasmusMC, Rotterdam, The Netherlands 3. SkylineDx, Rotterdam, The Netherlands 4. Oncode Institute, Utrecht, The Netherlands 5. Department of Medical Oncology, Amsterdam University Medical Center, University of Amsterdam, The Netherlands 6. Department of Clinical Pharmacy and Toxicology, Leiden UMC, Leiden, The Netherlands

Accepted for publication in Bioinformatics



Abstract

When phase III clinical drug trials fail their end-point, enormous resources are wasted. Moreover, even if a clinical trial demonstrates a significant benefit, the observed effects are often rather small and may not outweigh the side effects of the drug. Therefore, there is a great clinical need for methods to identify genetic markers that can identify subgroups of patients which are likely to benefit from treatment as this may i) rescue failed clinical trials and/or ii) identify subgroups of patients which benefit more than the population as a whole. When single genetic biomarkers cannot be found, machine learning approaches that find multivariate signatures are required. In the context of SNP profiles this is extremely challenging owing to the high dimensionality of the data. Here we introduce RAINFOREST (tReAtment benefIt prediction using raNdom FOREST), an adaptation of the random forest that can predict treatment benefit from patient SNP profiles obtained in a clinical trial setting.

We demonstrate the performance of RAINFOREST on the CAIRO2 dataset, a phase III clinical trial which tested the benefit of cetuximab treatment for metastatic colorectal cancer. While this trial concluded there was no benefit, we find that RAINFOREST is able to identify a subgroup comprising 27.7% of the patients that significantly benefit from treatment with a hazard ratio of 0.69 ($p = 0.04$) in favor of cetuximab. The method is not specific to colorectal cancer and could aid in reanalysis of phase III clinical trial data and provide a more personalized approach to cancer treatment, also for drugs where there is no clear link between a single variant and treatment benefit.

Introduction

Novel drugs are tested for efficacy in phase 3 clinical trials. Despite enormous investments in the development and research prior to the trial, approximately 54% of the phase 3 clinical trials still fail, most often due to a lack of efficacy of the drug tested (Hwang et al. 2016). Trials testing anti-cancer drugs have a higher failure rate than non-cancer drug trials. It was found that trials which adopt a biomarker strategy, i.e. attempt to identify a subset of patients most likely to benefit, have a significantly lower failure rate (Jardim et al. 2017). This is also true for trials evaluating targeted drugs. It is thus clear that even if a clinical trial does not reach its predefined endpoint, there could still be a subset of patients that do see benefit from the drug. Moreover, even if a clinical trial does indicate statistically significant benefit, this benefit may in fact be quite modest and driven by a subset of patients that have a larger benefit from the drug. For this reason, the benefit for all patients may be insufficient to warrant prescription of a drug with very serious side effects. In such cases, it is important to establish which subset of patients benefit more than the population as a whole and develop tools that can predict such treatment benefit at the moment of diagnosis.

It has become clear that the genetic background of both tumor and patient can influence drug response and several germline variants have been linked to the effectiveness of a number of drugs (anti-cancer and other). SNP panels enabling the use of these variants for personalized medicine are under active development (van der Wouden et al. 2019). For instance, for several chemotherapies, its sensitivity or toxicity has been linked to specific single nucleotide polymorphisms (SNPs) (Panczyk 2014; Sullivan et al. 2014; Yin et al. 2012). Despite this initial progress, for many drugs there is no clear relationship between response and a single variant or other simple molecular biomarker and more complex machine learning models are needed.

A major challenge is that genome wide germline variation datasets are very high dimensional, often including 100- to 1000-fold more features (SNPs) than samples (patients). As a result, machine learning models have a high risk of overtraining (Szymczak et al. 2009). One class of models, which has shown great promise in preventing overtraining in such situations, are Random Forests (RFs). Outside the cancer field, RFs have successfully been used to predict drug response using germline variation data (Athreya et al. 2019; Cosgun, Limdi, and Duarte 2011). RFs are ensemble

classifiers combining multiple decision trees. RFs are explicitly designed to prevent overtraining by using only a subset of the available training samples and randomly sampling a subset of the features at each split. Since the algorithm only has access to part of the dataset at a time, it is less likely to overtrain on the dataset as a whole, while predictive performance remains high due to the fact that many trees are combined in an ensemble. For instance, RFs have been successfully employed to predict optimal warfarin dose using genome wide germline variation data and shown to outperform alternative models (Cosgun, Limdi, and Duarte 2011).

Traditional machine learning methods like RFs enable the discovery of models that predict sensitivity for one specific treatment, i.e. distinguish between poor and good responders within one homogeneous treatment group. However, owing to recent progress in drug development for most cancers there are different treatment options available. A clinically more relevant question is thus whether an individual patient will benefit more from one treatment than another. In this work, we therefore define treatment benefit as having a better survival outcome on the treatment of interest than an alternative treatment. The difference between these treatments, often expressed in terms of hazard ratio (HR), should furthermore be greater than the difference observed in the population as a whole.

RFs have also been applied to survival analysis and used to identify (non-linear) prognostic factors in several cancer types, with modest success (Akai et al. 2018; Manilich et al. 2011). In essence, these random survival forests are similar to traditional RFs and also construct an ensemble classifier from individual decision trees, but the optimal split in these trees maximizes the survival difference between the two daughter nodes (Ishwaran et al. 2008).

In order to predict treatment benefit as we have defined it, traditional machine learning methods are unsuitable. Traditional class labels required for training machine learning models are not available. After all, we cannot know how a patient would have responded to a treatment they did not receive, and therefore we cannot know a priori (and thus label) a patient as class ‘benefit’ or class ‘no benefit’. More specifically, a patient responding well to a certain treatment could have had an even better response on an alternative treatment. Conversely, a poor response does not necessarily mean the

patient did not see any benefit from the treatment. This lack of training labels renders most regular machine learning approaches unsuitable. Likewise, survival analysis using random survival forests also does not solve the problem of a lack of training labels, as they simply aim to predict survival outcome instead of benefit to a certain treatment. An overview of the different aims of traditional machine learning, survival analysis and benefit prediction is provided in **Figure 1a**.

The machine learning method we propose can directly derive a benefit prediction model from germline genetic data gathered in a clinical trial in which patients were randomly assigned to one of two different treatment arms. To this end we propose an alternative formulation of the traditional RF classifier, called RAINFOREST (tReAtment benefIt prediction using raNdom FOREST). RAINFOREST implements the SurvDiff measure as an alternative to the Gini impurity, to decide on the best possible split in each decision tree. SurvDiff captures the survival difference between the treatment arms within a node. The SurvDiff measure enables training predictive decision trees by providing a split criterion which results in a 'benefit' and 'no benefit' branch in the tree. An overview of RAINFOREST and the SurvDiff measure is provided in **Figure 1b**.

We apply RAINFOREST to the CAIRO2-trial, a randomized phase III clinical trial designed to test whether patients with metastatic colorectal cancer benefit from addition of the EGFR inhibitor cetuximab to standard first-line treatment. This trial showed that the addition of cetuximab to a regimen of chemotherapy and bevacizumab results in a significantly shorter progression free survival (Tol et al, 2009). However, it is known that cetuximab response varies widely between patients. Previously, several somatic mutations in the tumor that influence cetuximab response have been identified (Salvatore et al. 2010; Khan et al. 2017). Moreover, in the context of the CAIRO2 trial a germline SNP was identified with the potential capability to predict treatment benefit (Pander et al. 2015), although this variant was not validated.

In this paper we demonstrate the capability of RAINFOREST on the CAIRO2 trial. We show that RAINFOREST can identify a subset of patients with significant benefit from cetuximab and that this approach outperforms both univariate analysis and a random forest trained on predefined labels.

Methods

Overview of RAINFOREST

A random forest model is an ensemble classifier consisting of individual decision trees trained on different subsets of the training data. More specifically, each tree in the forest only has access to a subset of the samples (sampled with replacement) and for each split in the tree a random subset of the features is sampled.

The optimization of each tree, i.e. choosing the optimal split for a node in the tree, is most often achieved by minimizing the Gini impurity. The Gini impurity is a measure of the probability that a sample would be incorrectly labeled in this split and is 0 when a node contains only samples with the same label. Problematically, in the context of predicting treatment benefit no predefined training labels are available, as we cannot know if a patient survived longer (or shorter) from treatment than on standard of care or some other treatment. We can therefore not use the Gini impurity for RF construction.

Treatment effect is most often determined through a Cox proportional hazards model (see next section for more details), based on which a hazard ratio (HR) is calculated. The HR associated with a treatment provides an estimate of the hazard of experiencing progression of disease relative to the hazard when another treatment would be given. A HR below 1 indicates benefit from receiving the treatment. In the absence of training labels that can be used to calculate accuracy, we use the HR as performance measure when validating the RAINFOREST model in cross validation.

Problematically, estimating a Cox model is too computationally expensive to be used in a splitting criterion when training thousands of decision trees. We therefore propose RAINFOREST, a random forest approach in which we introduce a novel splitting criterion that can be optimized to directly predict treatment benefit. For each sample, RAINFOREST requires treatment arm data, survival data and SNP data. Each decision tree should define a class 'benefit' and 'no benefit' which maximizes the difference between treatment effect. We define this difference through the splitting criterion *SurvDiff:*

$$SurvDiff = \left| \frac{\overline{survA_L} - \overline{survB_L}}{\sqrt{\frac{\sigma(survA_L)^2}{n_{A_L}} + \frac{\sigma(survB_L)^2}{n_{B_L}}}} - \frac{\overline{survA_R} - \overline{survB_R}}{\sqrt{\frac{\sigma(survA_R)^2}{n_{A_R}} + \frac{\sigma(survB_R)^2}{n_{B_R}}}} \right|$$

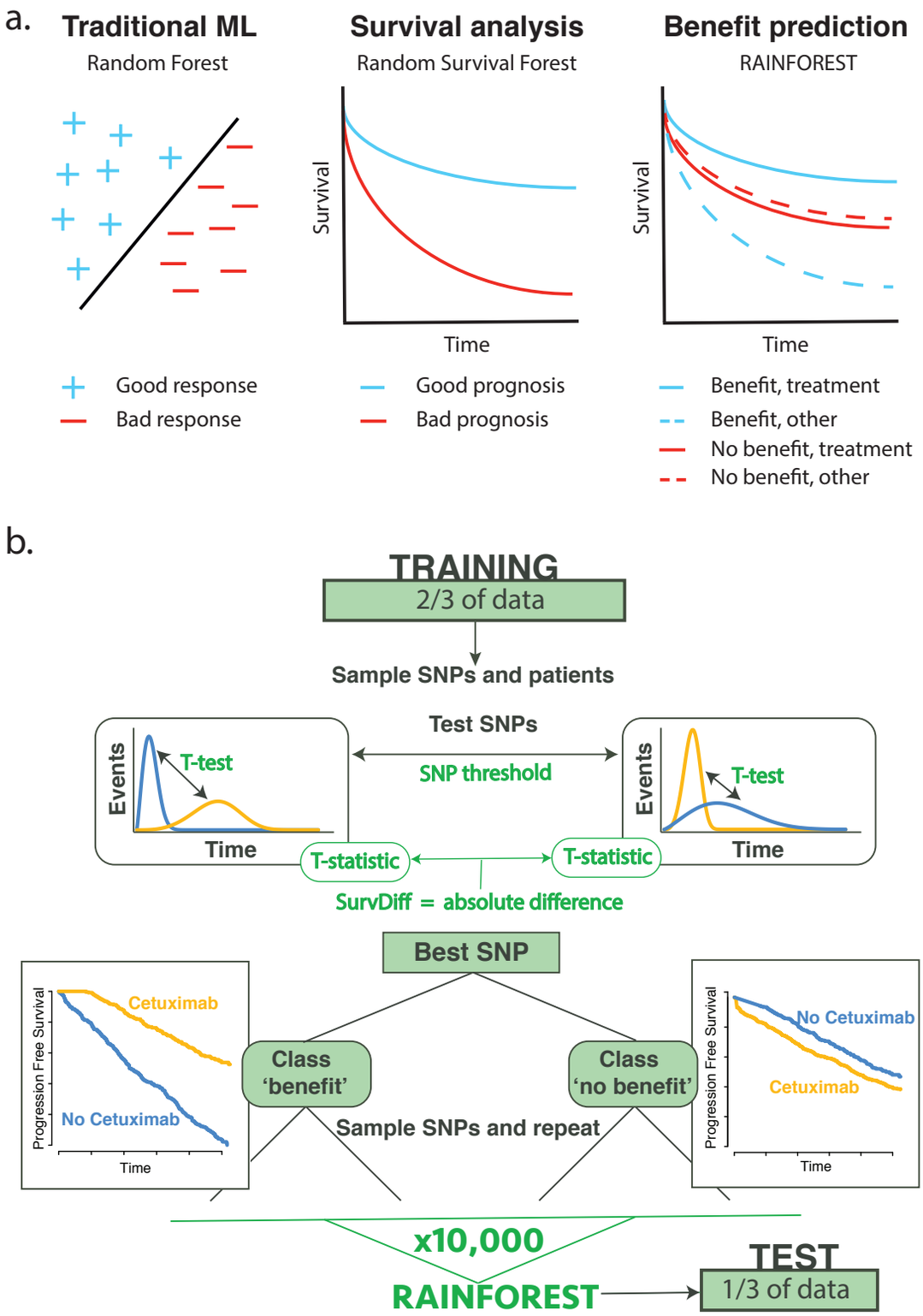


Figure 1. a. An overview of the difference between traditional machine learning, survival analysis and benefit prediction. b. An overview of the RAINFOREST algorithm. The survival curves show examples of what a class 'benefit' and 'no benefit' should look like. We train 10,000 of these individual decision trees to form the RAINFOREST model, which is validated on ⅓ of the data that acts as test data and was not used in training of the model.

where $\overline{survA_L}$ and $\overline{survB_L}$ are the mean survival data for treatment arm A and B in the left node of the split, respectively. Similarly, $\overline{survA_R}$ and $\overline{survB_R}$ are the equivalent in the right node. Moreover, $n_A$and n_B denote the number of samples included in the node in treatment arm A and B, respectively. For each SNP under consideration we test two thresholds (SNP value >0 or >1) to define the left and right node. $SurvDiff$thus corresponds to calculating the absolute difference between the Welch's T-test statistics found in the left and right node. The best SNP is the one resulting in the maximum value of $SurvDiff$.

Using this criterion we train 10,000 decision trees. We further prevent overtraining by restricting every tree to a depth of two. This restricts the tree to a maximum number of four leaves (nodes without a child node) and means every tree uses at most three SNPs. When building a tree using SNP data, the RF can be biased towards choosing non-informative SNPs with a high minor allele frequency over informative SNPs with a lower minor allele frequency (Boulesteix et al. 2012). This bias is not very pronounced in the beginning of a tree, but can dramatically influence SNP selection lower in the tree, when smaller sample sizes are present. We therefore also only split a node further when it contains at least 50 patients. These restrictions also reduce computational cost. An overview of the construction of the RAINFOREST model is given in **Figure 1**.

Survival analysis and event imputation

Survival data is right censored, which means that all patients who did not experience progression of disease by the end of follow-up are censored, i.e. no event is recorded. Cox models can handle censored data by maximizing the partial log likelihood over coefficient β through:

$$\ell(\beta) = \sum_{i:C_i=1} X_i * \beta - \log \sum_{j:Y_j \geq Y_i} \theta_j$$

where $\theta_j = exp(X_j * \beta)$ and X represents the explanatory variable, i.e. the treatment arm in this situation. When estimating the likelihood of an event occurring for subject i at a certain time t the $\theta_j$is summed for every subject j that has not yet experienced an event at t. In this way censored patients can be included and used for optimization up

to the time of censoring, instead of being excluded from the dataset all together. The HR is defined as the exponent of β.

The *SurvDiff* measure does not rely on Cox models. Instead, RAINFOREST deals with the censoring problem by imputation. More specifically, for all censored patients an event time is imputed based on all patients for whom an event was observed as reference. To achieve this, a Weibull distribution is fitted to all uncensored patients through maximum likelihood estimation. The Weibull distribution can be used to adequately parametrize a survival distribution and can also - akin Cox regression - model proportional hazards (Carroll 2003). The cumulative distribution function of a Weibull distribution is described by

$$F(x; k, \lambda) = 1 - e^{-(x/\lambda^k)}$$

where x is the time to event, k is a shape parameter and λ is the scale parameter. In our dataset we find the maximum likelihood is reached with a value of 11.91 for λ and 1.65 for k. Importantly, we find very similar parameters for the distribution when we perform a maximum likelihood estimation for each treatment arm separately, justifying an estimation over the whole dataset. This is in line with the observation in the original trial that there is no significant survival difference between the two treatment arms. For each censored patient we now sample an event time greater than the time of censoring from the estimated Weibull distribution.

Data

In this work the survival and genome wide genotype data from patients enrolled in the CAIRO2 trial are used, which included patients in 79 Dutch centers to test the addition of cetuximab for the treatment of metastatic colorectal cancers. The data generation and processing has been previously described in detail (Pander et al. 2015). Briefly, we use survival data and germline DNA of 553 patients who received treatment regimen CAPOX-B (capecitabine, oxaliplatin and bevacizumab) with cetuximab (n = 274) or without cetuximab (n = 279).

DNA was isolated from peripheral leukocytes and genome wide genotyping was performed with Illumina beadchip arrays. Of all measured variants 647,550 passed all quality checks and we performed no imputation of additional variants. We also exclude

SNPs with a minor allele frequency <5% and SNPs with any missing data, after which 257,008 SNPs remain. Each SNP is coded as 0,1, or 2, corresponding to the number of copies of the alternative allele. We use progression free survival (PFS) as the end point in all analyses.

Univariate SNP analysis

To evaluate the ability of individual SNPs to predict cetuximab benefit, we compute two Cox proportional hazard models per SNP. First, we compute an additive model which contains the SNP and treatment arm as separate variables. The second model also includes an interaction term between the SNP and treatment arm (i.e. treatment arm*SNP). For a SNP that influences treatment benefit, this second model should provide a better fit. We test whether there is a significant difference in model fit using a likelihood ratio test. We rank SNPs on most significant contribution of the interaction term to the model, as measured by the likelihood ratio test p-value. With the best SNPs we define a benefit score by:

$$benefitScore = \sum_{i=1}^{n} X_i \beta_i$$

Where X is the alternative allele count for a certain SNP i and β the Cox regression coefficient associated with the interaction term. We perform forward feature selection to determine the best SNP combination by ranking the SNPs on p-value and adding the top 250 in order. The SNP combination resulting in the lowest HR in class 'benefit' is chosen. We validate this model in a three-fold cross validation.

Random forest using survival-derived labels

We compare the performance of RAINFOREST to the results obtained by a regular RF trained on the survival labels directly (which, as discussed previously, is not necessarily the best measure for treatment benefit). To obtain a labeled dataset, required for training a regular RF, we define the class 'benefit' as the patients with the 25% best progression free survival from the cetuximab arm combined with the patients with the 25% worst progression free survival from the other arm. The other 75% of patients comprise class 'no benefit'. With these labels we define a class benefit that has a

significantly better survival on cetuximab than the rest of the population, satisfying our definition of treatment benefit.

Cross validation fold construction

To evaluate the performance of univariate SNP selection, the regular RF and the RAINFOREST models, we employ 3-fold cross validation. To ensure the results are directly comparable, we use the same folds for all analyses. To obtain a fair estimation of the performance, it is important that the different folds are stratified, i.e. contain a similar and representative part of the whole dataset. Here we cannot balance the folds using training labels, as these are not available. To ensure the cross validation folds are representative, we therefore balance on treatment arm. Furthermore, we require that the HR found between the treatment arms does not differ more than 0.05 between all three folds.

Optimization of mtry parameter

RFs often use an out-of-bag (OOB) error to optimize model parameters. Since in an RF model each tree samples a different subset of the patients, each training sample is not used in a number of trees. The OOB error is determined by classifying each training sample, using only the trees in which a particular sample was not included. However, the OOB error can severely underestimate performance when random sampling is performed from unbalanced classes (Mitchell 2011). As we do not know the labels here, representative sampling is impossible. Using random sampling we indeed see that the OOB performance, defined as the HR between treatment arms in class 'benefit', is close to random (HR class 'benefit' in OOB sample is 1.45 (95% CI 0.94 - 2.26, p =0.10)).

As we cannot obtain a realistic estimation of the performance from the OOB sample in RAINFOREST, we cannot optimize the *mtry* parameter which defines how many features are sampled at every split. However, previous work suggests that the best *mtry* is linked to dataset dimensionality (Goldstein et al. 2010). The RF trained on survival labels uses the same features as RAINFOREST. In training this RF we try several settings for *mtry* ($\sqrt{p}$, $2\sqrt{p}$, 0.1p and 0.2p). For training RAINFOREST we use the same *mtry* setting as in the best performing RF trained on survival based labels ($\sqrt{p}$) and train 10,000 trees.

Results

T-test in SurvDiff criterion captures survival difference

We first assess whether the T-test on the imputed survival data, which is used in the *SurvDiff* splitting criterion, captures the same signal as Cox regression would capture, to ensure this is a suitable measure to use during training of the RAINFOREST model. For each SNP we perform a T-test for both the reference and alternative allele, contrasting the difference in imputed survival between the two treatment arms. We compare the resulting T-test statistic to the equivalent Cox regression β (**Figure 2a**). We find these measures to be highly correlated for both the reference allele (Spearman correlation coefficient = 0.95, $p < 2*10^{-16}$) and the alternative allele (Spearman correlation coefficient = 0.94, $p < 2*10^{-16}$). Importantly, this approach reduces compute time by one order of magnitude (34.41 minutes for the Cox regression computation versus 1.89 for the T-test on a single core). Thus, the T-test approach captures a similar signal as a full survival analysis while keeping it computationally feasible to train a model with thousands of trees.

RAINFOREST can identify patients benefiting from cetuximab

We next trained RAINFOREST to predict cetuximab benefit on the CAIRO2 trial data and validate its performance in a three-fold cross validation. **Figure 2b** shows the survival curves in the dataset as a whole, without any classification. Here we find an HR of 1.11 (95%CI 0.93 – 133, p = 0.25) for cetuximab treatment. **Figure 2c** shows the different HRs found in class 'benefit' when using different operating points of the classifier (i.e. different thresholds on the number of trees classifying a sample as 'benefit'). This curve indicates a direct relationship between the operating point and the HR found in class 'benefit' - we find a lower HR when the threshold is set higher. As no sample has a posterior probability higher than 0.5, we cannot use a majority vote to assign a sample to class 'benefit' or 'no benefit'. The threshold set provides a trade-off between the size of class 'benefit' and the HR found. **Figure 2d** shows the Kaplan Meier plot when the classification threshold that results in the lowest p-value in class 'benefit' is used. Importantly, all thresholds classifying 50% or less of the patients as 'benefit' result in an HR below 1 and would thus provide benefit. We show the combined results across the three cross validation folds, i.e. the predictions for each patient is based on the two folds in which this patient was not present. In class 'benefit' (n = 153) we find a

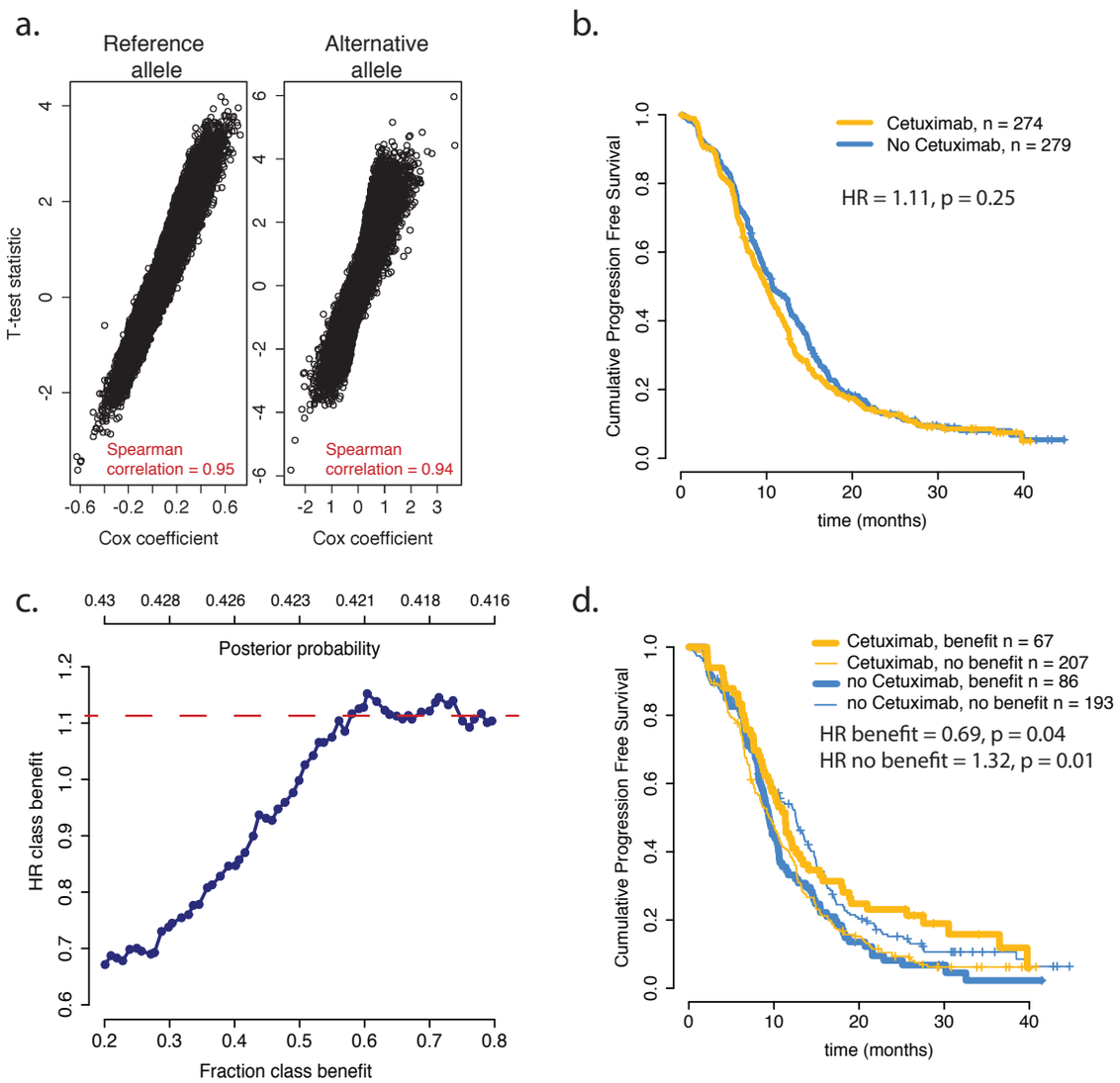


Figure 2. a. Scatterplot of the T-test statistic and Cox regression coefficient per SNP. We perform this analysis once using the reference allele to define class 'benefit' and once using the alternative allele. b. Kaplan Meier of the CAIRO2 survival data used, showing no survival benefit for the patients who received cetuximab. c. The HR found in class 'benefit' when using different threshold on the posterior probability to define benefit. The red dashed line shows the HR between treatments found in the dataset as a whole, without any classification. d. Kaplan Meier of the classification in class 'benefit' and 'no benefit' using the posterior probability threshold associated with the lowest Cox regression p-value in class 'benefit'.

significant HR of 0.69 (95% CI 0.49 - 0.98, p = 0.04) whereas in class 'no benefit' (n = 400) an HR of 1.32 (95% CI 1.07 - 1.62, p = 0.01) is found. This performance is relatively stable in all cross validation folds. More specifically, we find an HR of 0.66 (0.33 - 1.03, p = 0.23) in class 'benefit' in Fold 1, an HR of 0.72 (0.40 - 1.31, p = 0.28) in Fold 2, and an HR of 0.61 (0.44 - 1.09, p = 0.10) in Fold 3. While the original trial concluded addition of cetuximab to the standard regimen has no benefit, this result shows RAINFOREST can successfully identify a subset of patients, comprising 27.7% of the population, that do benefit from cetuximab.

Known and new SNPs are identified in frequently chosen SNPs

Over the three cross validation folds in total 51,154 unique SNPs are used (19,918, 19,982, and 19,810 in the models validated on Fold 1, 2 and 3 respectively). **Figure 3b** shows the number of SNPs overlapping between the three different models. We obtain an empirical p-value for this overlap by randomly sampling 10,000 trees for each fold and computing the overlap. We find the overlap of 781 SNPs between the three folds to be significant ($p < 1*10^{-4}$). We also train a RAINFOREST model using shuffled treatment labels with the same cross validation folds. With shuffled labels the association between genomic data and treatment specific outcome is removed and these models can indeed not predict benefit in hold-out data (HR class benefit = 0.95, 95% CI 0.64 - 1.41, p = 0.8). Between the models trained on shuffled labels only 356 SNPs overlap, which is similar to mean overlap found in random sampling (mean overlap = 344.7). The overlap found in the RAINFOREST model is thus clearly non-random.

Figure 3a shows the number of times each individual SNP is selected across the three cross validation folds. Interestingly, the SNP selected most often, rs885036, has been reported before to predict cetuximab benefit in a univariate analysis of the CAIRO2 trial (Pander et al. 2015). This shows that when univariate signals are present in the data, RAINFOREST will also capture these. In addition to rs885036, we also find a cluster of frequent SNPs on chromosome 5 which have not been reported before. Four of these variants (rs2549782, rs2287988, rs1056893 and rs2255546) are intronic variants within the ERAP1 gene. A fifth SNP (rs10069361) is annotated to LNPEP, a paralog of ERAP1. These SNPs are in high linkage disequilibrium (coefficient of linkage disequilibrium >0.9), where linkage disequilibrium is defined as the squared Pearson correlation coefficient. Both ERAP1 and LNPEP code for aminopeptidases. ERAP1 plays an important role in cleaving proteins into peptides that can be presented by MHC class 1

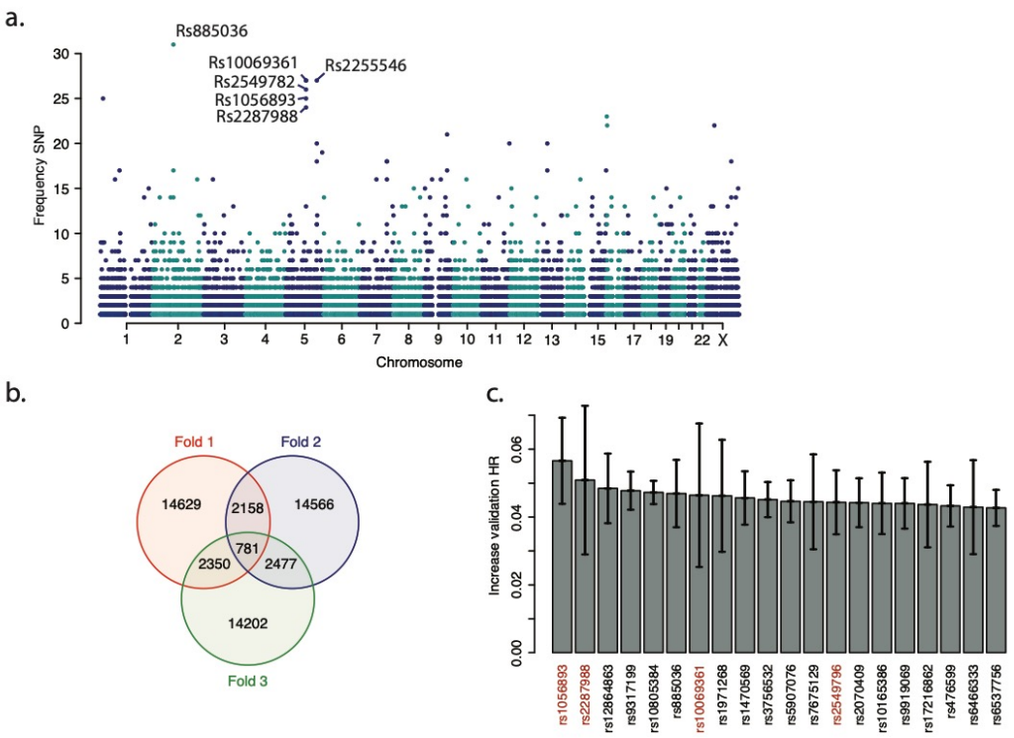


Figure 3. a. Manhattan plot showing the number of times individual SNPs were used in a decision tree across all three cross validation folds. b. Venn diagram showing the overlap in SNPs used in the three models for the three different cross validation folds. c. Barplot showing the 20 SNPs with the greatest influence on validation HR when the data is shuffled. Error bars indicate standard deviation. The SNPs indicated in red text are in LD > 0.9 with each other and all lie in the same region of chromosome 5. SNPs in black are not in high LD with any other SNP in the plot.

proteins to immune cells (Falk and Rötzschke 2002). Cetuximab is a monoclonal antibody and it has been shown that activation of the adaptive of the immune system and presence of cytotoxic T-cells are essential for its antitumor effect (Holubec et al. 2016; Yang et al. 2013). A potential explanation of these observations is that these SNPs represent genetic variation in the T-cell response that influence cetuximab response.

For all 781 SNPs that are present in all three models we also assessed feature importance by shuffling the genotype of the individual SNP and predicting the class labels on the validation again. This eliminates the association between the genetic data and treatment effect, so we can estimate the importance of each SNP. Without exception, shuffling SNPs increases the HR, which means the model performs worse. **Figure 3c** shows the difference in HR for the 20 SNPs with the largest effect. Note that since many

SNPs are only present in a few trees (i.e. the most frequent SNP is only present 31 times), the effect of shuffling is limited. We thus also do not see large changes in validation HR. Despite this limitation, 4 out of 5 SNPs from the chromosome 5 cluster as well as rs885036 are present in the top 20, strengthening their putative role in predicting cetuximab benefit.

Lactate dehydrogenase and age do not determine benefit prediction

High baseline lactate dehydrogenase (LDH) is a known prognostic factor in colorectal cancer (Li et al. 2016), but it does not have a significant interaction with treatment effect in survival analysis in our data (HR = 0.81, 95% CI 0.57 – 1.17, p = 0.26). Patients with high LDH are fairly evenly spread between class 'benefit' (45.6%) and class 'no benefit' (42.0%) and in neither class there is a significant interaction between treatment and high LDH (HR 'benefit' = 0.92, 95% CI 0.46 – 1.84, p = 0.30 and HR 'no benefit' = 0.80, 95% CI 0.52 – 1.22. p = 0.30).

There is also no significant difference in the mean age between the two classes (p = 0.66). The difference in treatment benefit is thus not explained by LDH and age, which are two common patient characteristics used in clinical decision making (van Eeghen et al. 2015; Li et al. 2016).

Sex influences treatment benefit

In the original trial the authors reported that women have a significantly better survival when not treated with cetuximab. Indeed, when considering the patients classified as 'benefit', we find an HR of 0.61 (95%CI 0.40 - 0.94, p = 0.02) for men and an HR of 1.04 (95% CI 0.56 - 1.94, p = 0.90) for women. While for women the HR in class benefit is lower than the overall HR (1.51, 95% CI 1.14 - 2.00, p = 0.003), it is not below 1 and therefore does not signify benefit. Moreover, more men are classified as benefit (31.9%) than women (22.0%). In our dataset we find an HR of 1.73 (p = 0.003, 95% CI 1.20 - 2.49) for the interaction term treatment*sex. We therefore investigate whether the interaction between treatment effect and chromosomal sex could also partly explain the performance of our model. The interaction term for sex*treatment was similar in both classes, giving an HR of 1.71 (p = 0.17, 95% CI 0.80 - 3.63) in class 'benefit' and an HR of 1.67 (p = 0.02, 95% CI 1.09 - 2.56) in class 'no benefit'. Together, this indicates RAINFOREST discovered a signal independent of the sex effect.

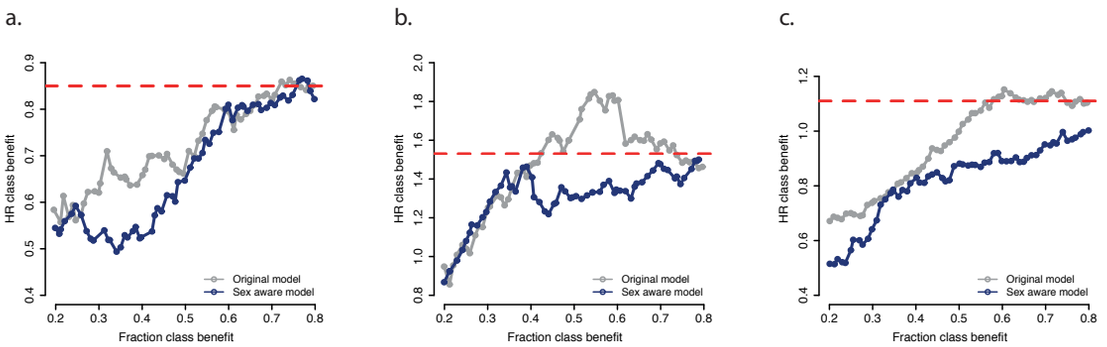


Figure 4. Performance for different sized class benefit (as determined with different thresholds on the posterior probability) for men and women, and the whole dataset. The red dashed line represent the HR found in the population as a whole.

Model incorporating chromosomal sex predicts benefit for men

Since sex is known to influence the outcome of cetuximab treatment and we see a different HR for men and women in our class 'benefit', we also train a RAINFOREST model that incorporates the sex variable, which we call the sex aware model in the rest of this text. The training procedure is the same as before, but in the construction of a tree, in addition to a sample of the SNPs, chromosomal sex can be selected as a splitting variable. We also construct new cross validation folds, in which the stratification is chosen such that, in addition to the overall treatment HR, the interaction term sex*treatment is similar in all folds.

For each fold we train 10,000 trees. On average 1109 trees use the sex variable for a split (1232 for Fold 1, 1263 for Fold 2 and 831 for Fold 3). The optimal HR found in class 'benefit' (n = 131) is 0.52 (95% CI 0.35 - 0.76, p = 0.0007), while the HR in class 'no benefit' (n = 422) is 1.35 (95% CI 1.11 - 1.67, p = 0.004). The sex aware model thus provides a better performance than the original model that did not include the sex variable. However, it should be noted that in this case class 'benefit' consists almost entirely of men (95.4%). We therefore evaluate the optimal threshold for men and women separately, as well as for the whole dataset (**Figure 4**). It follows that the sex-aware model works better than the original model for men for a class 'benefit' below 50%, but not for women. While the sex aware model has a better performance for women in a larger class 'benefit', it should be noted that all these HRs are well above 1 and thus do not represent true benefit.

When considering the selected SNP-variables, 110 SNPs are shared between all three folds of the original model and all three folds of this new model. This includes rs885036 and all the SNPs in the cluster on chromosome 5 described above, underscoring their importance in determining benefit to cetuximab.

When we train a RAINFOREST model in only women, we do find an optimal HR of 0.76 (p = 0.39), suggesting a model can be obtained with a true predictive performance. However, the performance curve (Supplemental Figure 1) does not show the linear relationship between the size of class benefit (as determined by the threshold on the posterior probability) and HR in class benefit. This indicates that a well-defined class 'benefit' cannot be identified by RAINFOREST in this dataset. The sex aware model reflects this fact by not including women in class 'benefit' when given access to this information. This shows RAINFOREST can accommodate this type of known effect and fit a model on the rest of the variables, improving the performance of the model.

Univariate SNP selection does not validate in cross validation

We compare the performance of RAINFOREST to the univariate selection of SNPs (see Methods). This analysis reveals no SNPs that are significant at a multiple testing corrected p-value less than 0.05. We perform forward feature selection by ranking the SNPs on likelihood ratio test p-value to find the optimal SNP combination. With this approach, the models for fold 1, 2 and 3 contain 101, 197 and 190 SNPs respectively. In line with the earlier univariate study (Pander et al, 2015), Rs885036 (the most frequently selected SNP in the RAINFOREST model) is selected in all three folds. With the exception of one other SNP (rs10165386) no other SNPs overlap. Moreover, the model does not result in a significant HR, as we find an HR of 1.00 (95% CI = 0.70 - 1.44, p = 1) in class 'benefit' (n = 138) and an HR of 1.15 (95% CI 0.93 - 1.41, p = 0.19) in class 'no benefit' (n = 415). Univariate selection of the SNPs thus does not lead to a model that validates on unseen patient data.

Random forest on survival based labels does not validate

We also train a classical random forest model on the benefit labels derived from the survival data (see **Methods**). The cross validation is performed using the same folds as in the univariate and RAINFOREST analysis. Since we do have training labels in this case, mtry can be optimized using the OOB error. The default setting often used is the

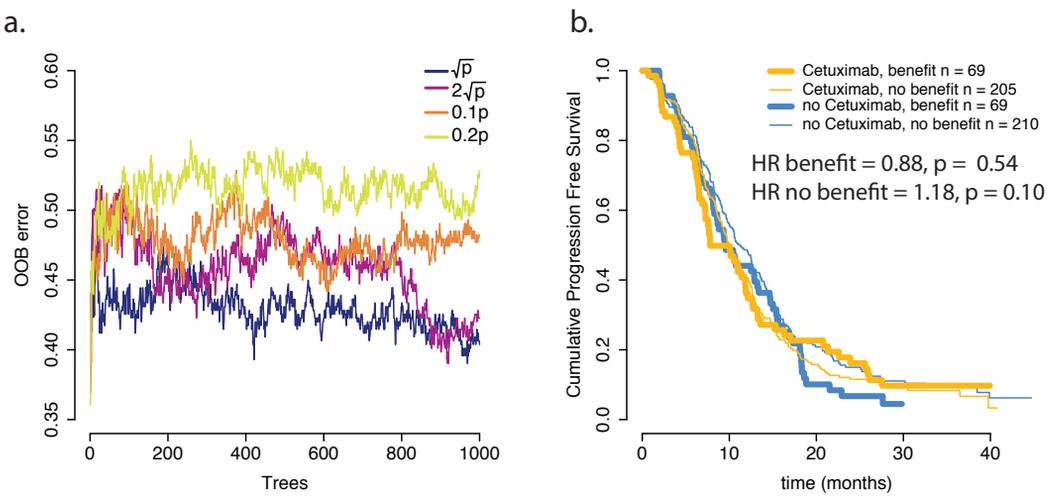


Figure 5 a. The OOB error found for the survival based levels when using different values for *mtry*. b. Kaplan Meier of the classification in class 'benefit' and 'no benefit', using the threshold that defines the class 'benefit' with the lowest Cox regression p-value.

square root of all features available, but it has been suggested that in high dimensional datasets a higher mtry leads to a better performance (Goldstein et al. 2010). We therefore try several values for mtry and evaluate the OOB error. Figure 5a shows that the default $\sqrt{p}$, where p is the total number of features, leads to the lowest error (Figure 5a).

Using the optimal model we find that no patients are classified into the 'benefit' class when using majority vote, despite the fact that both classes are sampled equally in the training data. We therefore classify a sample with where more than 30% of trees indicate benefit as benefiting, as this leads to a class benefit of approximately 25%. Using these settings we train a random forest with 10,000 trees and validate it on the test set. In the test set we set a threshold on the posterior probability that results in the lowest p-value in class 'benefit'. We then find an HR of 0.88 (95% CI 0.59 - 1.32, p = 0.54) in class 'benefit' (n = 138) and an HR of 1.18 (95% CI 0.97 - 1.44, p = 0.10) in class 'no benefit' (n = 415). The Kaplan Meier curve is shown in Figure 5b. While the RF can identify a class 'benefit' with an HR below 1, this is not statistically significant at $p < 0.05$. Similar results are obtained when defining benefit as the top 50% and bottom 50% of the treatment arms (HR benefit = 0.97, 95% CI 0.70 - 1.36, p = 0.88) or when

restricting the RF to a depth of two (HR benefit = 0.92, 95% CI 0.50 - 1.67, p = 0.77). We conclude that predefined benefit labels based on survival outcome are not suitable as training labels for training an RF classifier for treatment benefit.

Discussion

We here demonstrate RAINFOREST, a new approach to predict treatment benefit from patient germline variation data. The RAINFOREST model successfully identifies a subset of patients that benefits from cetuximab treatment in the CAIRO2 trial. It outperforms univariate analysis and traditional random forest models. We demonstrate its performance through cross validation, as the best estimate of the performance on independent validation data. Further validation in a truly independent patient cohort should further establish clinical utility of our approach. Moreover, in this model we have only considered the influence of germline variation on cetuximab benefit. Several tumor characteristics, like KRAS and BRAF mutation status and molecular subtype, have also been shown to correlate with cetuximab response (Salvatore et al., 2010, Trinh et al, 2017). A further analysis could take both tumor and germline variation into account to identify benefiting patients even more comprehensively.

The CAIRO2 trial represents a good test case for RAINFOREST as previous univariate analysis has shown a relation between germline variation and treatment specific survival. Reassuringly, we identify rs885036, the variant identified previously, among the most frequently used SNPs in the RAINFOREST model. Importantly, RAINFOREST identifies a number of previously unknown SNPs, which are not found with a univariate approach, that suggest a role for genetic variation in the immune response in determining cetuximab benefit.

With the sex aware model we show RAINFOREST can be adapted to incorporate characteristics known to be important, such as chromosomal sex. However, as the overlap in important SNPs show, the same signal can still be identified, underscoring the stability of the method.

The authors of the CAIRO2 trial concluded that there was a slight detrimental effect of the addition of cetuximab to the CAPOX-B treatment regimen. This is a clear example

for how RAINFOREST can be applied, as roughly half of all phase 3 clinical trials fail to reach their predefined endpoints and most fail due to insufficient efficacy of the drug (Hwang et al. 2016). As a result, these drugs do not enter the clinic, while it is very possible that a subset of the patient population experiences benefit. RAINFOREST can identify patients that do benefit from drugs which failed to show significant benefit in the patient population as a whole, and thus play an important role in leveraging valuable patient data and find an application for drugs that otherwise would not be introduced to the clinic.

Funding

J.U. is supported by a PhD fellowship from Van Herk Charity.

Data availability

Due to restrictions based on privacy regulations and informed consent of participants, phenotype and genotype data of the CAIRO2 trial cannot be made freely available in a public repository. Data from this study can be obtained upon request. Requests should be directed towards Prof. dr. H.J. Guchelaar (h.j.guchelaar@lumc.nl).

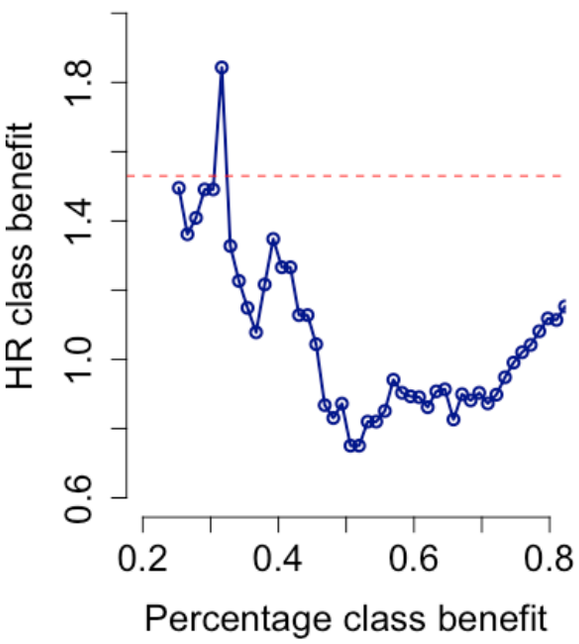
Code availability

The R code used to produce the results in this paper is available at github.com/UMCUGenetics/RAINFOREST. A more configurable, user-friendly Python implementation of RAINFOREST is also provided.

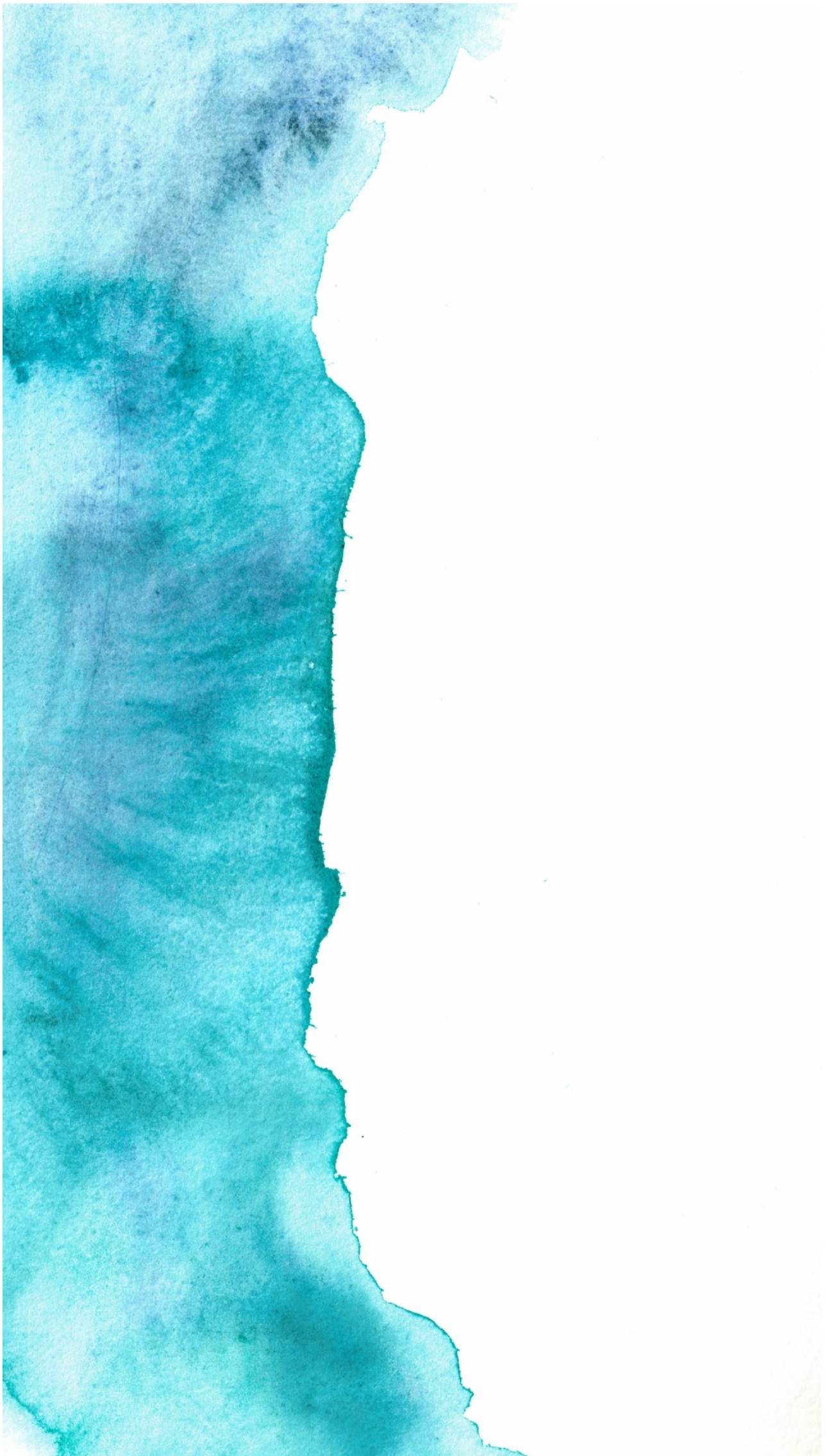
5

Supplementary material

5



Supplementary figure 1. Performance when RAINFOREST is trained and validated on only women.



Chapter 6

Discussion

Personalized medicine has been discussed as the future of cancer treatment for over three decades now, since the discovery and potential targeting of mutations in the RAS signalling pathway (Reddy et al. 1982; Downward 2003). There has been tremendous progress, particularly in matching targeted treatment with specific mutations or cell surface markers. However, despite the fact that we have known for a long time that both germline variation and tumor characteristics influence disease progression and treatment response, there are not many DNA or RNA signatures in use in the clinic (Fröhlich et al. 2018). There is in fact a great gap between the great number of papers reporting gene expression signatures and the ones that have an impact on clinical care (Koscielny 2010). In this thesis several different approaches for predicting treatment benefit are presented. While multiple counterfactual approaches exist, they have so far been used mostly in low dimensional settings for causal inference. Here we present multiple ways of using this kind of reasoning in high dimensional settings to build clinically useful models. These approaches could play an important role in a further realization of personalized medicine - the tailoring of treatment to a patient based on individual characteristics - in cancer treatment. However, there are still various challenges to be faced. We will discuss here what the work in this thesis can contribute and which challenges still have to be addressed.

Reproducibility of signatures and different populations

A major concern and hindrance in clinical adaptation is the lack of reproducibility for many classifiers (Subramanian and Simon 2010). For prognostic signatures it has been shown that many classifiers in fact do not outperform random classifiers when tested on external data (i.e. data the classifier was not trained on) (Tang et al. 2017). Moreover, it was also shown a gene expression classifier with satisfactory internal performance could be trained on completely random data. Proper validation is thus crucial.

In absence of truly independent data, many studies use cross validation to estimate the expected performance, as we also do in **Chapter 2** and **Chapter 5**. However, it is known that cross validation can overestimate performance (Castaldi, Dahabreh, and Ioannidis 2011). In some cases, the cross validation may not have been properly performed (for example, when multiple models are validated and the best is presented), but there are also signatures that were correctly evaluated and yet do not show a satisfactory

performance in independent data. This could be due to the fact that the classifier has fitted a signal specific to the population in the original data (i.e. the data the cross validation was performed on), with the difference in signal between datasets influenced by a true biological difference between the patients or differences in lab procedure or clinical practice.

This thesis contains some examples of this: in **Chapter 2** we see that when we train the model solely on the Total Therapy dataset the classifier does not validate on the Hovon65 dataset, while we can perform a successful cross validation when the two datasets are mixed. With mixed dataset GESTURE has the opportunity to fit the mixed signal, where the model is most likely too specific when trained solely on the Total Therapy dataset. In **Chapter 3** of this thesis cross validation was quite predictive of the performance in independent data, while this was not the case in **Chapter 4**, even though the set-up of the cross validation was nearly identical. Most likely the difference between the breast cancer datasets used in **Chapter 4** is far greater than the multiple myeloma datasets in **Chapter 3**, as the clinical reality is very different for both diseases. Treatment is less guided by patient characteristics in multiple myeloma than in breast cancer; patients in an observational trial are probably more likely to match the population from a randomized clinical trial. Training on randomized data is more suitable for simulated treatment learning, as there are similar patients in both treatment arms by definition. While the strategy of matching patients in the breast cancer dataset - to simulate a clinical trial like setting - improved performance on hold out data from the same population, this classifier still did not validate on external data. There could be trade off in data selection: training within one population leads to better results in that specific population, but is less generalizable to a wider population. This conflict extends to the follow-up time: longer follow up is often beneficial for the training procedure, especially for cancer types with a long median survival. When using a 20-year-old dataset most relevant events will have been recorded, but the setting in which these women were treated is no longer relevant. This meant that the older METABRIC dataset in **Chapter 4** was less useful for training, even though it included far more events than the SCAN-B dataset. There is no clear solution to this problem, though potentially subsetting older datasets to more closely match new datasets could be a strategy (to for example conform to current treatment guidelines). There are limits to this; in METABRIC we could not match patients in a manner that resulted in a dataset

with a hazard ratio in favor of chemotherapy. We should always carefully evaluate whether a dataset can still be relevant. The training data used has important implications for clinical deployment; the intended use population should match the training and validation population.

Finally, for identification of cancer subtypes tumor purity is an important factor. When a biopsy of a solid tumor is taken it will always contain both cancer cells and other cells (for example cells from the immune system). When gene expression is measured on this mixture the outcome will also be influenced by non-cancer cells. It has been shown that the variability in tumor purity biases subtype classification and estimating tumor purity can improve classification results (Aran, Sirota, and Butte 2015; Zhang et al. 2017). Multiple myeloma is a non-solid tumor and cells are sorted to a purity of at least 80% before gene expression is measured. This could lead to a more consistent measurement, less bias and thus a higher chance of successful external validation.

Lack of available data

Absolutely crucial for the training and proper validation of these classifiers is the availability of data. Especially for diseases that are not very prevalent, data available within one institution will not be sufficient. Moreover, as discussed, validation within one dataset or population is no guarantee for predictive availability in another population. All considerations about matching populations are only relevant if enough data is available. Open science and the sharing of data has received a lot of attention in the past few years, but many scientists are still worried that sharing their data will be to their disadvantage (Gewin 2016). While many journals now require a statement on data availability and the data needs to be publicly available (Naughton and Kernohan 2016), many publicly available gene expression datasets (for example in the Gene Expression Omnibus) do not offer enough patient information to enable the training of predictive or even prognostic classifiers on this data. We need systems in place that encourage sharing of all useful data, while of course keeping an eye on privacy concerns. Journal simply requiring data to be available seems to be insufficient. For example, when the British Medical Journal randomly audited 157 research articles in their journal, they found data was available (either publicly or upon request) for only 4.5% of those articles

(Rowhani-Farid and Barnett 2016). However, as more and more journals adopt a data sharing policy and open data is normalized, more data will hopefully be shared. Increasingly, funders also require a data sharing plan and publishers start encouraging data sharing more actively, with Springer Nature starting a research data helpdesk that can facilitate the sharing process (Jones, Grant, and Hrynaszkiewicz 2019). Data can also be assigned a digital object identifier (DOI), so it can be cited and researchers receive credit for the data they made available. Increasingly, researchers are aware of the FAIR data principles: data should be Findable, Accessible, Interoperable, Reusable (Wilkinson et al. 2016). This means it should be clear where data is located, how to gain access, and it should be in a format that can be read and manipulated by commonly used programs. It should be clear which data is included in the file and how it was produced. Importantly, accessible does not mean *freely* accessible. FAIR data can still safeguard privacy. For further model development and validation, wide availability of data is crucial and the research community should take all possible steps to encourage FAIR data sharing.

Integration of different data types

An approach not employed in this thesis is the integration of different data types (i.e. DNA and RNA data). In **Chapter 2, 3 and 4** we use tumor gene expression and in **Chapter 5** we use germline DNA variation. The truth is that the benefit for each treatment is probably influenced both by factors specific to the cancer cells and specific to the individual patient. An important distinction to be made here is the integration of different data types from the same cell (type) and data representing different systems in the body. For the prediction of prognosis in breast cancer it has been shown that tumor gene expression captures most of the information and adding different data types does not improve performance (Aben et al. 2018). However, the data considered there was all taken from the tumor and thus represented the same system. When we would for example combine tumor gene expression and germline DNA data, we are taking data from different systems; the tumor cell and the body surrounding it. The impact of a drug not being metabolized in the liver could never be captured by tumor gene expression for example. In **Chapter 5** we identify SNPs that are predictive of cetuximab benefit. Previously, tumor gene expression profiles and tumor specific mutations that are

predictive of cetuximab response have been identified (Salvatore et al. 2010; Baker et al. 2011). Since information is most likely present in all these data types, a logical next step would be to analyze them together to form a more complete picture of which patients benefit. There are several ways of integrating data; you can pool all data and train a single classifier (early integration) or train separate classifiers and then combine the classifications (late integration) and forms in between. RAINFOREST in **Chapter 5** could easily be adapted to also take tumor gene expression into account, with the values discretized to match the SNP format. However, in early integration dimensionality of datasets matters a lot; the higher dimensional data type can dominate the signal and seem the most important, even though this is not biologically true. Late integration, on the other hand, does not offer opportunities to model interactions between the germline data and tumor gene expression. Early integration may then be preferable, but steps should be taken to bring different data types in the same (dimensional) space.

Interpretability of the classifiers

When a predictive classifier is able to identify which patients benefit from a treatment, the logical next step is to investigate why these patients benefit and how the genes included in the classifier fit in. It has been shown for prognostic classifiers that many classifiers with a similar performance, and yet using completely different genes, can be found (Ein-Dor et al. 2005). Since genes function in pathways and expression is often very correlated, many genes can encode the same signal and simply interpreting the genes included in a classifier may not be useful. In **Chapter 2** and **Chapter 4** we attempt to encode biological information using gene ontology (GO) annotations and these gene sets do indeed perform better than random sets at predicting treatment benefit. However, the GO sets used in different classifiers predicting benefit for the same drug show very little overlap and no (obvious) interpretation of these genes could be formulated. An additional concern is that especially when a classifier is trained in a non-linear way like GESTURE is, it is possible the class 'benefit' is actually composed of multiple subsets; not all benefiting patients benefit for the same reason. We also measure gene expression in bulk, while each multiple myeloma patient probably harbors multiple different tumor clones (Keats et al. 2012). We could be measuring an average of the clones, or just a signal dominated by the largest clone. While there is

clearly enough information present to predict treatment benefit in a meaningful way, it could be a barrier to interpretation.

Gene sets defined by biological knowledge may thus not be useful for interpretation. Interestingly, when we formed entirely data driven gene sets in **Chapter 3**, the individual genes were more crucial for performance than when we used gene sets informed by biological annotation. Without the 14 genes included in the original signature, no signature with a similar performance could be found. The strategy followed in **Chapter 3** (i.e. only selecting genes that the algorithm always ranks highly over different repeats) could be beneficial in finding these crucial genes. However, it should be noted we could not describe a mechanism that links the 14 genes in this classifier and individually they were not differentially expressed between class 'benefit' and 'no benefit'. This approach also does not address the concern of measuring several clones at the same time.

Once we have an interpretation of the genes, a next step could be to functionally validate the findings. The fact that the 14 genes in **Chapter 3** can only be identified together and do not show differential expression by themselves could also be a barrier to proving their role in a functional assay. We would have to under- or overexpress a combination of 14 genes, without the model itself providing a hypothesis on how benefit could be achieved (i.e. which genes should be over- or under-expressed). For even more complicated models, like GESTURE produces, this would be impossible. It is also important to consider what the goal of interpretation could be, beyond providing further insight into the disease. When a clear mechanism can be identified that causes a patient not to benefit from a drug, this could be used for rational design of a drug that could overcome this. It seems clear the models presented in this thesis are far away from playing a role in this.

Clinical practice and clinical utility

Finally, the most important part of work like this is the clinical utility: even if the classifier is completely accurate, would clinical care be changed based on its prediction? In some cases this may be obvious. With the Mammaprint, which can predict which breast cancer patients can safely forgo chemotherapy, it seems the decision is clear. It

should however be noted, that even here the case is not clear cut. Patients can be reluctant to forgo available therapy based on a risk assessment, even if statistically we would not expect benefit. In the case of the Mammaprint, the Dutch Healthcare Institute declined to mandate insurance companies to reimburse the test, citing a possible 2.4% increase in distant metastases if chemotherapy was not given (Zorginstituut Nederland 2018). This test failed to become the standard, even though a prospective clinical trial proved its accuracy. With algorithms and artificial intelligence playing a larger role in society, there has been a lot of public debate on when algorithms can be trusted to make decisions that will impact lives. Which decisions can be taken by non-human systems and where lies the responsibility for the outcome of such a decision? Explainability of the decision plays a large role here (Abdollahi and Nasraoui 2018). It could very well be unreasonable to expect physicians and patients to stake lives on a model for which it cannot be explained why it works. For this purpose identifying which genes are crucial - as discussed above - can be already useful, even if it does not lead to a new treatment; it can aid in the explainability of the treatment decision. Smaller, clearer signatures like the one presented in **Chapter 3** will then be preferable over the large, complicated models built by GESTURE.

Of course, what is clinically useful is not static. When a treatment is standard and given to all patients, it may make more sense to attempt finding a group that does not benefit: for those patients treatment should be changed. However, without a convincing alternative treatment, such a classifier does not have a high probability of being adapted. The discovery of a new drug could render a classifier useless or useful; for example by establishing a new standard treatment or by providing an alternative treatment for a no benefit group.

In light of changing clinical practice, it is crucial to shorten the time between biomarker discovery and introduction in the clinic. As mentioned before, there are limitations here: sufficient follow up is needed. However, adaptive clinical trials could play a role here. This is a trial that changes design based on data gathered during the trial (Barker et al. 2009; Gallo et al. 2006). The I-SPY trial in breast cancer is an example, where inclusion criteria for treatment arms are adapted as the trial goes on to incorporate effects discovered during the trial (Barker et al. 2009). The I-SPY trial is mostly designed around known patient characteristics. There are also strategies that facilitate biomarker

discovery and validation when no obvious candidates are known. Here patients are randomized as normal between two treatment arms and then split in a training and validation cohort during the trial. The training cohort can be used to continuously build a predictive model, while the validation cohort can be used as a prospective trial at the same time (Scher et al. 2011). There are obvious ethical considerations here and in practice this design will be followed only when there is no evidence for superiority of the treatment under investigation in the population as a whole.

There is also the wider context of health care to consider. In most developed countries the cost of healthcare is on the rise and discussions on when treatment is no longer affordable need to be had (Baltagi et al. 2017). Personalized medicine can play an important role in this problem and reduce overall health care expenditure (Jakka and Rossbach 2013). When we can predict who will benefit from more generic treatments, we do not only spare patients who do not benefit unnecessary side effects, we can also reduce the cost of treatment. However, in incurable forms of cancer like Multiple Myeloma, where a patient will always receive a form of treatment and often will be treated until their death, it may be hard to quantify the amount of money saved.

This also relates to the importance of finding a subset of the population that does benefit from drugs that fail to show a significant effect in the population as a whole. Pharmaceutical companies claim high prices for drugs are needed to offset all the costs made in developing drugs that do not reach the market. The more efficient drug development is and the more drugs can be used, the cheaper drugs can (theoretically) be.

Conclusion

Personalized medicine and predictive biomarkers will play an important role in the health care of the future. However, it is also clear that there are different challenges for different diseases and there is not one model to be applied here. Algorithms should be combined with clinical trial design and an awareness of clinical reality. For adaptation in the clinic, simpler models may be better.
In this thesis we present three different algorithms to train a model capable of predicting treatment benefit: GESTURE, STLsig, and RAINFOREST. Considering the

topics discussed, they all have different strengths. For the purpose of training an interpretable model, STLsig seems to produce the best classifiers; GESTURE models are too complicated. However, STLsig is much more sensitive to high censoring rates. When a patient does not have suitable neighbours (i.e. no neighbours who experienced an event), our main measure for benefit (zPFS) cannot be calculated and this patient then drops out of the analysis. This obviously happens more often when there are fewer events recorded in the dataset. STLsig uses the whole distribution of zPFS and its performance is more impacted by patients dropping out than GESTURE, which just uses patients with a high zPFS. For data with fewer events GESTURE(-BC) is thus more suitable. Both GESTURE and STLsig need continuous data like gene expression to calculate distances between patients. RAINFOREST is more versatile; it can handle the discrete values of SNP data, but could also easily be adapted to categorical data like sex. It could potentially also handle gene expression data and a mix of different data types. It would thus be most suitable to be used for integration of patient characteristics and different data types. Together they can hopefully be used to make personalized medicine a reality in cancer treatment.

Addendum

A

References

Abdollahi B and Nasraoui O. 2018. Transparency in Fair Machine Learning: the Case of Explainable Recommender Systems. 10.1007/978-3-319-90403-0_2.

Aben N, Westerhuis JA, Song Y, et al. iTOP: inferring the topology of omics data. Bioinformatics. 2018;34(17):i988-i996. doi:10.1093/bioinformatics/bty636

Akai, H., Yasaka, K., Kunimatsu A., et al. 2018. "Predicting Prognosis of Resected Hepatocellular Carcinoma by Radiomics Analysis with Random Survival Forest." Diagnostic and Interventional Imaging 99 (10): 643–51.

Amin, Mohammed Abdullahel, Go Itoh, Kenji Iemura, Masanori Ikeda, and Kozo Tanaka. 2014. "CLIP-170 Recruits PLK1 to Kinetochores during Early Mitosis for Chromosome Alignment." Journal of Cell Science 127 (Pt 13): 2818–24.

Aran, Dvir, Marina Sirota, and Atul J. Butte. 2015. "Systematic Pan-Cancer Analysis of Tumour Purity." Nature Communications 6 (December): 8971.

Arana, Mercedes E., and Thomas A. Kunkel. 2010. "Mutator Phenotypes due to DNA Replication Infidelity." Seminars in Cancer Biology 20 (5): 304–11.

Athreya, A.P., et al. 2019. "Pharmacogenomics-Driven Prediction of Antidepressant Treatment Outcomes: A Machine-Learning Approach With Multi-trial Replication." Clinical Pharmacology & Therapeutics. https://doi.org/10.1002/cpt.1482.

Audeh, William, Lisa Blumencranz, Heather Kling, Harsha Trivedi, and Gordan Srkalovic. 2019. "Prospective Validation of a Genomic Assay in Breast Cancer: The 70-Gene MammaPrint Assay and the MINDACT Trial." Acta Medica Academica 48 (1): 18–34.

Avet-Loiseau, H., P. Moreau, C. Mathiot, C. Charbonnel, T. Facon, M. Attal, C. Hulin, et al. 2010. "Use of Bortezomib to Overcome the Poor Prognosis of t(4;14), but Not del(17p), in Young Patients with Newly Diagnosed Multiple Myeloma." Journal of Clinical Oncology. https://doi.org/10.1200/jco.2010.28.15_suppl.8113.

Baker, J. B., D. Dutta, D. Watson, T. Maddala, B. M. Munneke, S. Shak, E. K. Rowinsky, et al. 2011. "Tumour Gene Expression Predicts Response to Cetuximab in Patients with KRAS Wild-Type Metastatic Colorectal Cancer." British Journal of Cancer 104 (3): 488–95.

Baltagi, Badi H., Raffaele Lagravinese, Francesco Moscone, and Elisa Tosetti. 2017. "Health Care Expenditure and Income: A Global Perspective." Health Economics 26 (7): 863–74.

Barker, A. D., C. C. Sigman, G. J. Kelloff, N. M. Hylton, D. A. Berry, and L. J. Esserman. 2009. "I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy." Clinical Pharmacology and Therapeutics 86 (1): 97–100.

Barrier, Alain, Pierre-Yves Boelle, François Roser, Jennifer Gregg, Chantal Tse, Didier Brault, François Lacaine, et al. 2006. "Stage II Colon Cancer Prognosis Prediction by Tumor Gene Expression Profiling." Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology 24 (29): 4685–91.

Bennett, Brian D., and Pierre R. Bushel. 2017. "goSTAG: Gene Ontology Subtrees to Tag and Annotate Genes within a Set." Source Code for Biology and Medicine 12 (April): 6.

Bernard, P. S., et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of Clinical Oncology, 27, 1160–1167. (2009).

Bernau, Christoph, Markus Riester, Anne-Laure Boulesteix, Giovanni Parmigiani, Curtis Huttenhower, Levi Waldron, and Lorenzo Trippa. 2014. "Cross-Study Validation for the Assessment of Prediction Algorithms." Bioinformatics 30 (12): i105–12.

Bhutani, M., et al. Investigation of a gene signature to predict response to immunomodulatory derivatives for patients with multiple myeloma: an exploratory, retrospective study using microarray datasets from prospective clinical trials. The Lancet Haematology, 4, e443–e451. (2017).

A

Bianconi, Eva, Allison Piovesan, Federica Facchin, Alina Beraudi, Raffaella Casadei, Flavia Frabetti, Lorenza Vitale, et al. 2013. "An Estimation of the Number of Cells in the Human Body." Annals of Human Biology 40 (6): 463–71.

Block, K. I., et al. Designing a broad-spectrum integrative approach for cancer prevention and treatment. Seminars in Cancer Biology, 35, S276–S304. (2015)

Blotta, S., et al. Canonical and noncanonical hedgehog pathway in the pathogenesis of multiple myeloma. Blood, 120(25), 5002–5013. (2012).

Boulesteix, A. et al. 2012. "Random Forest Gini Importance Favours SNPs with Large Minor Allele Frequency: Impact, Sources and Recommendations." Briefings in Bioinformatics 13 (3): 292–304.

Breiman, Leo. 2001. "Machine Learning." https://doi.org/10.1023/a:1017934522171.

Bullinger, Lars, and Peter J. M. Valk. 2005. "Gene Expression Profiling in Acute Myeloid Leukemia." Journal of Clinical Oncology. https://doi.org/10.1200/jco.2005.05.020.

Burrell, R. A., McGranahan N., Bartek, J. and Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature, 501, 338–345. (2013).

Cancer Stat Facts: Female breast cancer. https://seer.cancer.gov/statfacts/html/breast.html

Cardoso, Fatima, Laura J. van't Veer, Jan Bogaerts, Leen Slaets, Giuseppe Viale, Suzette Delaloge, Jean-Yves Pierga, et al. 2016. "70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer." The New England Journal of Medicine 375 (8): 717–29.

Carlson M. hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2). R package version 3.0.0. (2016).

Carroll, K.J. 2003. "On the Use and Utility of the Weibull Model in the Analysis of Survival Data." Controlled Clinical Trials. https://doi.org/10.1016/s0197-2456(03)00072-2.

Carter, Jessica A., Dariusz C. Górecki, Charles A. Mein, Börje Ljungberg, and Sassan Hafizi. 2013. "CpG Dinucleotide-Specific Hypermethylation of theTNS3gene Promoter in Human Renal Cell Carcinoma." Epigenetics. https://doi.org/10.4161/epi.25075.

Castaldi, Peter J., Issa J. Dahabreh, and John P. A. Ioannidis. 2011. "An Empirical Assessment of Validation Practices for Molecular Classifiers." Briefings in Bioinformatics 12 (3): 189–202.

Chang, H., et al. Analysis of PTEN deletions and mutations in multiple myeloma. Leukemia Research, 30(3), 262–265. (2006).

Chapman, Michael A., Jonathan Sive, John Ambrose, Claire Roddie, Nicholas Counsell, Anna Lach, Mahnaz Abbasian, et al. 2018. "RNA-Seq of Newly Diagnosed Patients in the PADIMAC Study Leads to a Bortezomib/lenalidomide Decision Signature." Blood 132 (20): 2154–65.

Chapman, P. B., A. Hauschild, C. Robert, J. M. G. Larkin, J B A, A. Ribas, D. Hogg, et al. 2011. "Phase III Randomized, Open-Label, Multicenter Trial (BRIM3) Comparing BRAF Inhibitor Vemurafenib with Dacarbazine (DTIC) in Patients with V600EBRAF-Mutated Melanoma." Journal of Clinical Oncology. https://doi.org/10.1200/jco.2011.29.18_suppl.lba4.

Chen, Yuyan, Alvin Kamili, Jayne R. Hardy, Guy E. Groblewski, Kum Kum Khanna, and Jennifer A. Byrne. 2013. "Tumor Protein D52 Represents a Negative Regulator of ATM Protein Levels." Cell Cycle 12 (18): 3083–97.

Cosgun, E et al. 2011. "High-Dimensional Pharmacogenetic Prediction of a Continuous Trait Using Machine Learning Techniques with Application to Warfarin Dose Prediction in African Americans." Bioinformatics 27 (10): 1384–89.

Cox, D. R. 1972. "Regression Models and Life-Tables." Journal of the Royal Statistical Society: Series B (Methodological). https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

Curtis, Christina, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, et al. 2012. "The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups." Nature 486 (7403): 346–52.

Damiano, J. S., A. E. Cress, L. A. Hazlehurst, A. A. Shtil, and W. S. Dalton. 1999. "Cell Adhesion Mediated Drug Resistance (CAM-DR): Role of Integrins and Resistance to Apoptosis in Human Myeloma Cell Lines." Blood 93 (5): 1658–67.

Damiano, J. S., and W. S. Dalton. 2000. "Integrin-Mediated Drug Resistance in Multiple Myeloma." Leukemia & Lymphoma 38 (1-2): 71–81.

De Pas M. D. et al. 2013. "Crizotinib versus Chemotherapy in Advanced ALK-Positive Lung Cancer." New England Journal of Medicine 368: 2385–94.

Dispenzieri, A., et al. Immunoglobulin free light chain ratio is an independent risk factor for progression of smoldering (asymptomatic) multiple myeloma. Blood, 111, 785–789. (2008).

Dong, Hongjuan, Liang Chen, Xiequn Chen, Hongtao Gu, Guangxun Gao, Ying Gao, and Baoxia Dong. 2009. "Dysregulation of Unfolded Protein Response Partially Underlies Proapoptotic Activity of Bortezomib in Multiple Myeloma Cells." Leukemia & Lymphoma 50 (6): 974–84.

Dong, Min Jun, Wang Lin Bo, Zhi Nong Jiang, Mei Jin, Wen Xian Hu, and Shen Jian Guo. 2014. "The Transcription Factor KLF4 as an Independent Predictive Marker for Pathologic Complete Remission in Breast Cancer Neoadjuvant Chemotherapy: A Case–control Study." OncoTargets and Therapy. https://doi.org/10.2147/ott.s68340.

Downward, Julian. 2003. "Targeting RAS Signalling Pathways in Cancer Therapy." Nature Reviews Cancer. https://doi.org/10.1038/nrc969.

Druker, B. J., M. Talpaz, D. J. Resta, B. Peng, E. Buchdunger, J. M. Ford, N. B. Lydon, et al. 2001. "Efficacy and Safety of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in Chronic Myeloid Leukemia." The New England Journal of Medicine 344 (14): 1031–37.

Durinck S, Spellman P, Birney E and Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols, 4, pp. 1184–1191. (2009).

Eales, K. L., Hollinshead, K. E. R., and Tennant, D. A. Hypoxia and metabolic adaptation of cancer cells. Oncogenesis, 5, e190. (2016).

Eeghen, E. E. van, et al. 2015. "Impact of Age and Comorbidity on Survival in Colorectal Cancer." Journal of Gastrointestinal Oncology 6 (6): 605–12.

Ein-Dor, Liat, Itai Kela, Gad Getz, David Givol, and Eytan Domany. 2005. "Outcome Signature Genes in Breast Cancer: Is There a Unique Set?" Bioinformatics 21 (2): 171–78.

Falk, K, and Rötzschke, O. 2002. "The Final Cut: How ERAP1 Trims MHC Ligands to Size." Nature Immunology. https://doi.org/10.1038/ni1202-1121.

Foster, Jared C., Jeremy M. G. Taylor, and Stephen J. Ruberg. 2011. "Subgroup Identification from Randomized Clinical Trial Data." Statistics in Medicine. https://doi.org/10.1002/sim.4322.

Fröhlich, Holger, Rudi Balling, Niko Beerenwinkel, Oliver Kohlbacher, Santosh Kumar, Thomas Lengauer, Marloes H. Maathuis, et al. 2018. "From Hype to Reality: Data Science Enabling Personalized Medicine." BMC Medicine 16 (1): 150.

Gallo, Paul, Christy Chuang-Stein, Vladimir Dragalin, Brenda Gaydos, Michael Krams, and José Pinheiro. 2006. "Adaptive Designs in Clinical Drug Development—An Executive Summary of the PhRMA Working Group." Journal of Biopharmaceutical Statistics. https://doi.org/10.1080/10543400600614742.

García-Laorden, M. Isabel, Ingrid Stroo, Sanne Terpstra, Sandrine Florquin, Jan Paul Medema, Cornelis van T Veer, Alex F. de Vos, and Tom van der Poll. 2017. "Expression and Function of Granzymes A and B in Peritonitis and Sepsis." Mediators of Inflammation 2017 (June): 4137563.
Gewin, Virginia. 2016. "Data Sharing: An Open Mind on Open Data." Nature. https://doi.org/10.1038/nj7584-117a.

Goldstein, B. A., et al. 2010. "An Application of Random Forests to a Genome-Wide Association Dataset: Methodological Considerations & New Findings." BMC Genetics 11 (June): 49.

Hanahan, Douglas, and Robert A. Weinberg. 2000. "The Hallmarks of Cancer." Cell. https://doi.org/10.1016/s0092-8674(00)81683-9.

Hanahan, D., & Weinberg, R. A. Review Hallmarks of Cancer : The Next Generation. Cell, 144(5), 646–674. (2011).

Harrell F.E. et al., Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors Statistics in Medicine, 15, 361-387 (1996).

Hatano, K., J. Kikuchi, M. Takatoku, R. Shimizu, T. Wada, M. Ueda, M. Nobuyoshi, et al. 2009. "Bortezomib Overcomes Cell Adhesion-Mediated Drug Resistance through Downregulation of VLA-4 Expression in Multiple Myeloma." Oncogene. https://doi.org/10.1038/onc.2008.385.

Haynes, Winston A., Aurelie Tomczak, and Purvesh Khatri. 2018. "Gene Annotation Bias Impedes Biomedical Research." Scientific Reports 8 (1): 1362.

Hideshima, Teru, Constantine Mitsiades, Masaharu Akiyama, Toshiaki Hayashi, Dharminder Chauhan, Paul Richardson, Robert Schlossman, et al. 2003. "Molecular Mechanisms Mediating Antimyeloma Activity of Proteasome Inhibitor PS-341." Blood 101 (4): 1530–34.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." Political Analysis. https://doi.org/10.1093/pan/mpl013.

A

Ho, D., Imai, K., King, G., & Stuart, E. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software, 42(8), 1 - 28. doi:http://dx.doi.org/10.18637/jss.v042.i08

Hofman, I. J. F., et al. RPL5 on 1p22.1 is recurrently deleted in multiple myeloma and its expression is linked to bortezomib response. Leukemia, 31(8), 1706–1714. (2017).

Holman, L., Head, M.L., Lanfear, R., and Jennions, M.D. Evidence of experimental bias in the life sciences: Why we need blind data recording. PLoS Biology, 13. doi:10.1371/journal.pbio.1002190 (2015).

Holubec, L et al. 2016. "The Role of Cetuximab in the Induction of Anticancer Immune Response in Colorectal Cancer Treatment." Anticancer Research. https://doi.org/10.21873/anticanres.10985.

Howlader N, et al. SEER Cancer Statistics Review, 1975-2013. In National Cancer Institute. Bethesda, MD. Retrieved from http://seer.cancer.gov/csr/1975_2013/ (2016).

Hwang, T.J. et al. 2016. “Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results.” JAMA Internal Medicine 176 (12): 1826–33.

Ishwaran, Hemant, and Min Lu. 2019. “Random Survival Forests.” Wiley StatsRef: Statistics Reference Online. https://doi.org/10.1002/9781118445112.stat08188.

Jakka, Sairamesh, and Michael Rossbach. 2013. “An Economic Perspective on Personalized Medicine.” The HUGO Journal. https://doi.org/10.1186/1877-6566-7-1.

Jardim, D.L. et al. 2017. “Factors Associated with Failure of Oncology Drugs in Late-Stage Clinical Development: A Systematic Review.” Cancer Treatment Reviews 52 (January): 12–21.

Johnson, W. E., Li, C., and Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics (Oxford, England), 8, 118–27. (2007).

Jones, Allison, Andrew E. Teschendorff, Quanxi Li, Jane D. Hayward, Athilakshmi Kannan, Tim Mould, James West, et al. 2013. “Role of DNA Methylation and Epigenetic Silencing of HAND2 in Endometrial Cancer Development.” PLoS Medicine 10 (11): e1001551.

Jones, Leila, Rebecca Grant, and Iain Hrynaszkiewicz. 2019. “Implementing Publisher Policies That Inform, Support and Encourage Authors to Share Data: Two Case Studies.” Insights the UKSG Journal. https://doi.org/10.1629/uksg.463.

Keats, Jonathan J., Marta Chesi, Jan B. Egan, Victoria M. Garbitt, Stephen E. Palmer, Esteban Braggio, Scott Van Wier, et al. 2012. “Clonal Competition with Alternating Dominance in Multiple Myeloma.” Blood 120 (5): 1067–76.

Kaplan, Henry G., and Judith A. Malmgren. 2008. “Impact of Triple Negative Phenotype on Breast Cancer Prognosis.” The Breast Journal. https://doi.org/10.1111/j.1524-4741.2008.00622.x.

Karadag, Abdullah, Min Zhou, and Peter I. Croucher. 2006. “ADAM-9 (MDC-9/meltrin-γ), a Member of the Adisintegrin and Metalloproteinase Family, Regulates Myeloma-Cell–induced Interleukin-6 Production in Osteoblasts by Direct Interaction with the αvβ5 Integrin.” Blood. https://doi.org/10.1182/blood-2005-09-3830.

Keats, J. J.,et al.. Clonal competition with alternating dominance in multiple myeloma. Blood, 120, 1067–1076. (2012).

Khan, S.A. et al. 2017. "EGFR Gene Amplification and KRAS Mutation Predict Response to Combination Targeted Therapy in Metastatic Colorectal Cancer." Pathology Oncology Research: POR 23 (3): 673–77.

Knudson, A. G. 1971. "Mutation and Cancer: Statistical Study of Retinoblastoma." Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.68.4.820.

Koscielny, S. 2010. "Why Most Gene Expression Signatures of Tumors Have Not Been Useful in the Clinic." Science Translational Medicine. https://doi.org/10.1126/scitranslmed.3000313.

Kuiper, R., A. Broyl, Y. de Knegt, M. H. van Vliet, E. H. van Beers, B. van der Holt, L. el Jarari, et al. 2012. "A Gene Expression Signature for High-Risk Multiple Myeloma." Leukemia 26 (11): 2406–13.

Kumar, S. K., et al. Improved survival in multiple myeloma and the impact of novel therapies. Blood, 111, 2516–2520. (2008)

Kyle, Robert A., Ellen D. Remstein, Terry M. Therneau, Angela Dispenzieri, Paul J. Kurtin, Janice M. Hodnefield, Dirk R. Larson, et al. 2007. "Clinical Course and Prognosis of Smoldering (asymptomatic) Multiple Myeloma." The New England Journal of Medicine 356 (25): 2582–90.

Landowski, Terry H., Nancy E. Olashaw, Deepak Agrawal, and William S. Dalton. 2003. "Cell Adhesion-Mediated Drug Resistance (CAM-DR) Is Associated with Activation of NF-Kappa B (RelB/p50) in Myeloma Cells." Oncogene 22 (16): 2417–21.

Laubach, Jacob, Paul Richardson, and Kenneth Anderson. 2011. "Multiple Myeloma." Annual Review of Medicine 62: 249–64.

Li, G et al. 2016. "The Prognostic Value of Lactate Dehydrogenase Levels in Colorectal Cancer: A Meta-Analysis." BMC Cancer 16 (March): 249.

Le, Dung T., Jennifer N. Durham, Kellie N. Smith, Hao Wang, Bjarne R. Bartlett, Laveet K. Aulakh, Steve Lu, et al. 2017. "Mismatch Repair Deficiency Predicts Response of Solid Tumors to PD-1 Blockade." Science 357 (6349): 409–13.

Lee, Stephanie J., Paul G. Richardson, Pieter Sonneveld, Michael W. Schuster, David Irwin, Jesús-F San Miguel, Bruce Crawford, et al. 2008. "Bortezomib Is Associated with Better Health-Related Quality of Life than High-Dose Dexamethasone in Patients with Relapsed Multiple Myeloma: Results from the APEX Study." British Journal of Haematology 143 (4): 511–19.

Liaw, A. and Wiener, M. Classification and Regression by randomForest. R News 2(3), 18 - 22. (2002).

Lièvre, A., et al. KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. Cancer Research, 66, 3992–3995. (2006).

Lohr, J. G., et al. Widespread genetic heterogeneity in multiple myeloma: Implications for targeted therapy. Cancer Cell, 25, 91–101. (2014).

Lu, Min, Saad Sadiq, Daniel J. Feaster, and Hemant Ishwaran. 2018. "Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods." Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America 27 (1): 209–19.

Luo, Dianhong, Yun He, Haifeng Zhang, Luyang Yu, Hong Chen, Zhe Xu, Shibo Tang, Fumihiko Urano, and Wang Min. 2008. "AIP1 Is Critical in Transducing IRE1-Mediated Endoplasmic Reticulum Stress Response." The Journal of Biological Chemistry 283 (18): 11905–12.

Manilich, E. A., Kiran R.P., Radivoyevitch, T. et al 2011. "A Novel Data-Driven Prognostic Model for Staging of Colorectal Cancer." Journal of the American College of Surgeons 213 (5): 579–88, 588.e1–2.

Meyer, D. et al. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-7. http://CRAN.R-project.org/package=e1071 (2015).

Mitchell, Matthew W. 2011. "Bias of the Random Forest Out-of-Bag (OOB) Error for Certain Input Parameters." Open Journal of Statistics. https://doi.org/10.4236/ojs.2011.13024.

Moreau, Philippe, Paul G. Richardson, Michele Cavo, Robert Z. Orlowski, Jesús F. San Miguel, Antonio Palumbo, and Jean-Luc Harousseau. 2012. "Proteasome Inhibitors in Multiple Myeloma: 10 Years Later." Blood 120 (5): 947–59.

Moreau P, Attal M, Hulin C, Arnulf B, Belhadj K, Benboubker L et al. Bortezomib, thalidomide, and dexamethasone with or without daratumumab before and after autologous stem-cell transplantation for newly diagnosed multiple myeloma (CASSIOPEIA): a randomised, open-label, phase 3 study. The Lancet 2019. 394:29 – 38

Mouneimne, Ghassan, and Joan S. Brugge. 2007. “Tensins: A New Switch in Cell Migration.” Developmental Cell 13 (3): 317–19.

Mothilal, Ramaravind K., Amit Sharma, and Chenhao Tan. 2020. “Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations.” Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3351095.3372850.

Munshi, N. C., & Anderson, K. C. New strategies in the treatment of multiple myeloma. Clinical Cancer Research, 19, 3337–44. (2013).

Mygind, Kasper J., Theresa Störiko, Marie L. Freiberg, Jacob Samsøe-Petersen, Jeanette Schwarz, Olav M. Andersen, and Marie Kveiborg. 2018. “Sorting Nexin 9 (SNX9) Regulates Levels of the Transmembrane ADAM9 at the Cell Surface.” The Journal of Biological Chemistry 293 (21): 8077–88.

Nahi, H., Sutlu, T., Jansson, M., Alici, E. and Gahrton, G. (2011), Clinical impact of chromosomal aberrations in multiple myeloma. Journal of Internal Medicine, 269: 137-147. doi:10.1111/j.1365-2796.2010.02324.x

Narita, T., M. Ri, A. Masaki, F. Mori, A. Ito, S. Kusumoto, T. Ishida, H. Komatsu, and S. Iida. 2015. “Lower Expression of Activating Transcription Factors 3 and 4 Correlates with Shorter Progression-Free Survival in Multiple Myeloma Patients Receiving Bortezomib plus Dexamethasone Therapy.” Blood Cancer Journal 5 (December): e373.

Naughton, Linda, and David Kernohan. 2016. “Making Sense of Journal Research Data Policies”. Insights 29 (1): 84–89. DOI: http://doi.org/10.1629/uksg.284

Neben, K., et al. Administration of bortezomib before and after autologous stem cell transplantation improves outcome in multiple myeloma patients with deletion 17p. Blood, 119, 940–948. (2012).

O’Connell, M.J, , Ian Lavery, Greg Yothers, Soonmyung Paik, Kim M. Clark-Langone, Margarita Lopatin, Drew Watson, Frederick L. Baehner, Steven Shak, Joffre Baker, J. Wayne Cowens, and Norman Wolmark. 2010. “Relationship Between Tumor Gene Expression and Recurrence in Four Independent Studies of Patients With Stage II/III Colon Cancer Treated With Surgery Alone or Surgery Plus Adjuvant Fluorouracil Plus Leucovorin.” Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology 28 (25): 3937–44.

Panczyk, Mariusz. 2014. “Pharmacogenetics Research on Chemotherapy Resistance in Colorectal Cancer over the Last 20 Years.” World Journal of Gastroenterology: WJG 20 (29): 9775–9827.

Pander, J. et al. 2015. “Genome Wide Association Study for Predictors of Progression Free Survival in Patients on Capecitabine, Oxaliplatin, Bevacizumab and Cetuximab in First-Line Therapy of Metastatic Colorectal Cancer.” PloS One 10 (7): e0131091.

Perou, C. M., T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, et al. 2000. “Molecular Portraits of Human Breast Tumours.” Nature 406 (6797): 747–52.

Priestley, Peter, Jonathan Baber, Martijn P. Lolkema, Neeltje Steeghs, Ewart de Bruijn, Charles Shale, Korneel Duyvesteyn, et al. 2019. “Pan-Cancer Whole-Genome Analyses of Metastatic Solid Tumours.” Nature 575 (7781): 210–16.

Rajkumar S. V. (2018). Multiple myeloma: 2018 update on diagnosis, risk-stratification, and management. American journal of hematology, 93(8), 981–1114. https://doi.org/10.1002/ajh.25117

Raponi, Mitch, Yi Zhang, Jack Yu, Guoan Chen, Grace Lee, Jeremy M. G. Taylor, James Macdonald, et al. 2006. “Gene Expression Signatures for Predicting Prognosis of Squamous Cell and Adenocarcinomas of the Lung.” Cancer Research 66 (15): 7466–72.

Reddy, E. P., R. K. Reynolds, E. Santos, and M. Barbacid. 1982. “A Point Mutation Is Responsible for the Acquisition of Transforming Properties by the T24 Human Bladder Carcinoma Oncogene.” Nature 300 (5888): 149–52.

Roccaro, Aldo Maria, Teru Hideshima, Noopur Raje, Shaji Kumar, Kenji Ishitsuka, Hiroshi Yasui, Norihiko Shiraishi, et al. 2006. “Bortezomib Mediates Antiangiogenesis in Multiple Myeloma via Direct and Indirect Effects on Endothelial Cells.” Cancer Research 66 (1): 184–91.

Rowhani-Farid, Anisa, and Adrian G. Barnett. 2016. “Has Open Data Arrived at the British Medical Journal (BMJ)? An Observational Study.” BMJ Open 6 (10): e011784.

Saal, Lao H., Johan Vallon-Christersson, Jari Häkkinen, Cecilia Hegardt, Dorthe Grabau, Christof Winter, Christian Brueffer, et al. 2015. “The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: A Large-Scale Multicenter Infrastructure towards Implementation of Breast Cancer Genomic Analyses in the Clinical Routine.” Genome Medicine 7 (1): 20.

Salvatore, M. Di et al. 2010. “KRAS and BRAF Mutational Status and PTEN, cMET, and IGF1R Expression as Predictive Markers of Response to Cetuximab plus Chemotherapy

in Metastatic Colorectal Cancer (mCRC)." Journal of Clinical Oncology. https://doi.org/10.1200/jco.2010.28.15_suppl.e14065.

Samuel, A. L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." IBM Journal of Research and Development. https://doi.org/10.1147/rd.33.0210.

Santos, C., et al. Intrinsic cancer subtypes-next steps into personalized medicine. Cellular Oncology, 38, 3–16. (2015).

Schaeffer, Jonathan. 2006. "Samuel's Checkers Player." Encyclopedia of Cognitive Science. https://doi.org/10.1002/0470018860.s00001.

Schapire, R. E. A brief introduction to boosting. IJCAI International Joint Conference on Artificial Intelligence, 2, 1401–1406. (1999).

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science, 270(5235), 467 LP – 470. https://doi.org/10.1126/science.270.5235.467

Scher, Howard I., Shelley Fuld Nasso, Eric H. Rubin, and Richard Simon. 2011. "Adaptive Clinical Trial Designs for Simultaneous Testing of Matched Diagnostics and Therapeutics." Clinical Cancer Research: An Official Journal of the American Association for Cancer Research 17 (21): 6634–40.

Sharma, Rajesh. 2019. "Breast Cancer Incidence, Mortality and Mortality-to-Incidence Ratio (MIR) Are Associated with Human Development, 1990–2016: Evidence from Global Burden of Disease Study 2016." Breast Cancer. https://doi.org/10.1007/s12282-018-00941-4.

Shaughnessy, John D., Jr, Fenghuang Zhan, Bart E. Burington, Yongsheng Huang, Simona Colla, Ichiro Hanamura, James P. Stewart, et al. 2007. "A Validated Gene Expression Model of High-Risk Multiple Myeloma Is Defined by Deregulated Expression of Genes Mapping to Chromosome 1." Blood 109 (6): 2276–84.

Shepherd JH, Uray IP, Mazumdar A, et al. The SOX11 transcription factor is a critical regulator of basal-like breast cancer growth, invasion, and basal-like gene expression. Oncotarget. 2016;7(11):13106-13121. doi:10.18632/oncotarget.7437

Simon, Noah, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2011. "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent." Journal of Statistical Software. https://doi.org/10.18637/jss.v039.i05.

Smetana, Jan, Kristina Berankova, Romana Zaoralova, Pavel Nemec, Henrieta Greslikova, Renata Kupska, Aneta Mikulasova, et al. 2013. “Gain(1)(q21) Is an Unfavorable Genetic Prognostic Factor for Patients with Relapsed Multiple Myeloma Treated with Thalidomide but Not for Those Treated with Bortezomib.” Clinical Lymphoma, Myeloma & Leukemia 13 (2): 123–30.

Socinki et al. CheckMate 026: A phase 3 trial of nivolumab vs investigator's choice (IC) of platinum-based doublet chemotherapy (PT-DC) as first-line therapy for stage iv/recurrent programmed death ligand 1 (PD-L1)–positive NSCLC. Ann Oncology, 27, Suppl_6:LBA7_PR (2016).

Soriano, G. P., L. Besse, N. Li, M. Kraus, A. Besse, N. Meeuwenoord, J. Bader, et al. 2016. “Proteasome Inhibitor-Adapted Myeloma Cells Are Largely Independent from Proteasome Activity and Show Complex Proteomic Changes, in Particular in Redox and Energy Metabolism.” Leukemia 30 (11): 2198–2207.

Spear, Brian B., Margo Heath-Chiozzi, and Jeffrey Huff. 2001. “Clinical Application of Pharmacogenetics.” Trends in Molecular Medicine. https://doi.org/10.1016/s1471-4914(01)01986-4.

Staiger, Christine, Sidney Cadot, Raul Kooter, Marcus Dittrich, Tobias Müller, Gunnar W. Klau, and Lodewyk F. A. Wessels. 2012. “A Critical Evaluation of Network and Pathway-Based Classifiers for Outcome Prediction in Breast Cancer.” PloS One 7 (4): e34796.

Stimpson, SE, Coorssen JR, Myers SJ. Isolation and identification of ER associated proteins with unique expression changes specific to the V144D SPTLC1 mutations in HSN-I. Biochemistry & Analytical Biochemistry, 5(1), (2015).

Subramanian, Jyothi, and Richard Simon. 2010. “Gene Expression-Based Prognostic Signatures in Lung Cancer: Ready for Clinical Use?” Journal of the National Cancer Institute 102 (7): 464–74.

Sullivan, I. et al.. 2014. “Pharmacogenetics of the DNA Repair Pathways in Advanced Non-Small Cell Lung Cancer Patients Treated with Platinum-Based Chemotherapy.” Cancer Letters 353 (2): 160–66.

Sun, Xiaodong, Dengwen Li, Yunfan Yang, Yuan Ren, Jingyu Li, Zaizhu Wang, Bin Dong, Min Liu, and Jun Zhou. 2012. “Microtubule-Binding Protein CLIP-170 Is a Mediator of Paclitaxel Sensitivity.” The Journal of Pathology 226 (4): 666–73.

Svachova, Hana, Fedor Kryukov, Elena Kryukova, Sabina Sevcikova, Pavel Nemec, Henrieta Greslikova, Lucie Rihova, Lenka Kubiczkova, and Roman Hajek. 2014. “Nestin Expression

throughout Multistep Pathogenesis of Multiple Myeloma." British Journal of Haematology 164 (5): 701–9.

Syn, Nicholas Li-Xun, Wei-Peng Yong, Boon-Cher Goh, and Soo-Chin Lee. 2016. "Evolving Landscape of Tumor Molecular Profiling for Personalized Cancer Therapy: A Comprehensive Review." Expert Opinion on Drug Metabolism & Toxicology 12 (8): 911–22.

Szalat, Raphael, Herve Avet-Loiseau, and Nikhil C. Munshi. 2016. "Gene Expression Profiles in Myeloma: Ready for the Real World?" Clinical Cancer Research: An Official Journal of the American Association for Cancer Research 22 (22): 5434–42.

Szymczak, S. et al 2009. "Machine Learning in Genome-Wide Association Studies." Genetic Epidemiology. https://doi.org/10.1002/gepi.20473.

Ting, K. R., et al. Novel panel of protein biomarkers to predict response to bortezomib-containing induction regimens in multiple myeloma patients. BBA Clinical, 8, 28–34. (2017).

Tang, H., S. Wang, G. Xiao, J. Schiller, V. Papadimitrakopoulou, J. Minna, I. I. Wistuba, and Y. Xie. 2017. "Comprehensive Evaluation of Published Gene Expression Prognostic Signatures for Biomarker-Based Lung Cancer Clinical Studies." Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO 28 (4): 733–40.

Tanoue, L. T. 2012. "Prognostic and Predictive Gene Signature for Adjuvant Chemotherapy in Resected Non–Small-Cell Lung Cancer." Yearbook of Pulmonary Disease. https://doi.org/10.1016/j.ypdi.2012.01.076.

Therneau T. A Package for Survival Analysis in S. version 2.38, https://CRAN.R-project.org/package=survival. (2015).

Tol, J., et al. (2009). Chemotherapy, Bevacizumab, and Cetuximab in Metastatic Colorectal Cancer. New England Journal of Medicine, 360(6), 563–572. https://doi.org/10.1056/NEJMoa0808268

Trinh, A. et al. (2017) Practical and Robust Identification of Molecular Subtypes in Colorectal Cancer by Immunohistochemistry. Clinical Cancer Research 23 (2), 387 - 398

Tusher, V. G., R. Tibshirani, and G. Chu. 2001. "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response." Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.091062498.

Ubels, Joske, Pieter Sonneveld, Erik H. van Beers, Annemiek Broijl, Martin H. van Vliet, and Jeroen de Ridder. 2018. "Predicting Treatment Benefit in Multiple Myeloma through Simulation of Alternative Treatment Effects." Nature Communications 9 (1): 2943.

Vangsted et al., Drug response prediction in high-risk multiple myeloma. Gene 644 80 - 86 (2018).

Van Vliet, M.H. et al. , An Assay for Simultaneous Diagnosis of t(4;14), t(11;14), t(14;16)/t(14;20), del1p, add1q, del13q, del17p, MS/MF Expression Clusters, and the SKY-92 High Risk Signature in Multiple Myeloma Patients Haematologica ; 98(s1):101. abstract n. P234 (2013).

Veer, Laura J. van 't, Hongyue Dai, Marc J. van de Vijver, Yudong D. He, Augustinus A. M. Hart, Mao Mao, et al. 2002. "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer." Nature. https://doi.org/10.1038/415530a.

Venet, D., Dumont, J. E., and Detours, V. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Computational Biology, 7(10). (2011).

Vittinghoff, Eric, Charles E. McCulloch, Claudine Woo, and Steven R. Cummings. 2010. "Estimating Long-Term Effects of Treatment from Placebo-Controlled Trials with an Extension Period, Using Virtual Twins." Statistics in Medicine 29 (10): 1127–36.

Waks, Adrienne G., and Eric P. Winer. 2019. "Breast Cancer Treatment: A Review." JAMA: The Journal of the American Medical Association 321 (3): 288–300.

Walker, Brian A., Konstantinos Mavrommatis, Christopher P. Wardell, T. Cody Ashby, Michael Bauer, Faith E. Davies, Adam Rosenthal, et al. 2018. "Identification of Novel Mutational Drivers Reveals Oncogene Dependencies in Multiple Myeloma." Blood 132 (6): 587–97.

Walther, A. Genetic prognostic and predictive markers in colorectal cancer. Nature Reviews Cancer, 9, 489–499. (2009).

Wang, Cun, Haojie Jin, Ning Wang, Shaohua Fan, Yanyan Wang, Yurong Zhang, Lin Wei, et al. 2016. "Gas6/Axl Axis Contributes to Chemoresistance and Metastasis in Breast Cancer through Akt/GSK-3β/β-Catenin Signaling." Theranostics 6 (8): 1205–19.

Warde-Farley, et al. The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. Nucleic Acids Research, 38,(SUPPL2) doi:10.1093/nar/gkq537 (2010).

Wilkinson, Mark D., Michel Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." Scientific Data 3 (March): 160018.

Wouden, Cathelijne H. van der, Mandy H. van Rhenen, Wafa O. M. Jama, Magnus Ingelman-Sundberg, Volker M. Lauschke, Lidija Konta, Matthias Schwab, Jesse J. Swen, and Henk-Jan Guchelaar. 2019. "Development of the PGx-Passport: A Panel of Actionable Germline Genetic Variants for Pre-Emptive Pharmacogenetic Testing." Clinical Pharmacology and Therapeutics 106 (4): 866–73.

Wu, De-Wei, Ya-Wen Cheng, John Wang, Chih-Yi Chen, and Huei Lee. 2010. "Paxillin Predicts Survival and Relapse in Non-Small Cell Lung Cancer by microRNA-218 Targeting." Cancer Research 70 (24): 10392–401.

Yamada, K. M., & Araki, M. Tumor suppressor PTEN: modulator of cell signaling, growth, migration and apoptosis. Journal of Cell Science, 114(Pt 13), 2375–2382. (2001).

Yanamandra, N. 2006. "Tipifarnib and Bortezomib Are Synergistic and Overcome Cell Adhesion-Mediated Drug Resistance in Multiple Myeloma and Acute Myeloid Leukemia." Clinical Cancer Research. https://doi.org/10.1158/1078-0432.ccr-05-1792.

Yang, X et al. 2013. "Cetuximab-Mediated Tumor Regression Depends on Innate and Adaptive Immune Responses." Molecular Therapy: The Journal of the American Society of Gene Therapy 21 (1): 91–100.

Yao, Fang, Chi Zhang, Wei Du, Chao Liu, and Ying Xu. 2015. "Identification of Gene-Expression Signatures and Protein Markers for Breast Cancer Grading and Staging." PloS One 10 (9): e0138213.

Yeh, Hsi-Wen, En-Chi Hsu, Szu-Shuo Lee, Yaw-Dong Lang, Yuh-Charn Lin, Chieh-Yu Chang, Suz-Yi Lee, et al. 2018. "PSPC1 Mediates TGF-β1 Autocrine Signalling and Smad2/3 Target Switching to Promote EMT, Stemness and Metastasis." Nature Cell Biology 20 (4): 479–91.

Yin, J et al.2012. "Meta-Analysis on Pharmacogenetics of Platinum-Based Chemotherapy in Non Small Cell Lung Cancer (NSCLC) Patients." PloS One 7 (6): e38150.

Yoshida, Takashi, Masaki Ri, Shiori Kinoshita, Tomoko Narita, Haruhito Totani, Reham Ashour, Asahi Ito, et al. 2018. "Low Expression of Neural Cell Adhesion Molecule, CD56, Is Associated with Low Efficacy of Bortezomib plus Dexamethasone Therapy in Multiple Myeloma." PloS One 13 (5): e0196780.

Zehir, Ahmet, Ryma Benayed, Ronak H. Shah, Aijazuddin Syed, Sumit Middha, Hyunjae R. Kim, Preethi Srinivasan, et al. 2017. “Mutational Landscape of Metastatic Cancer Revealed from Prospective Clinical Sequencing of 10,000 Patients.” Nature Medicine 23 (6): 703–13.

Zhan F, et al. The molecular classification of multiple myeloma. Blood 108(6), 2020-2028. (2006).

Zhang, J., Q. Liang, Y. Lei, M. Yao, L. Li, X. Gao, J. Feng, et al. 2012. “SOX4 Induces Epithelial-Mesenchymal Transition and Contributes to Breast Cancer Progression.” Cancer Research. https://doi.org/10.1158/0008-5472.can-12-1045.

Zheng, Zhihong, Tingbo Liu, Jing Zheng, and Jianda Hu. 2017. “Clarifying the Molecular Mechanism Associated with Carfilzomib Resistance in Human Multiple Myeloma Using Microarray Gene Expression Profile and Genetic Interaction Network.” OncoTargets and Therapy 10 (March): 1327–34.

Zorginstituut Nederland 2018. MammaPrint® niet in basispakket. https://www.zorginstituutnederland.nl/actueel/nieuws/2018/09/28/mammaprint-niet-in-basispakket

A

Summary

Many cancer treatments are associated with serious side effects, while it is known not all patients who receive them see benefit from the treatment. It has become clear both patient and tumor characteristics can influence the response of a cancer patient to a specific treatment. There is therefore great interest in personalized medicine: matching the right drug with the right patient, based on certain predictive features that can be measured. Certain drugs are designed to target a specific mutation in the tumor DNA; this drug is only beneficial for patients whose tumor harbors this alteration. But personalized medicine can also play a role in more generic treatments. Machine learning approaches have been employed to separate poor and good responders on the basis of tumor gene expression, among other things.

However, often there is more than one drug available and a choice has to be made between them, which is a more challenging problem. Most machine learning approaches employed in predicting benefit for a single treatment require labels to train a model. Patients are be labeled as poor or good responders, and the model is optimized to distinguish these two classes. These cannot be employed when predicting whether a patient will benefit more from a certain treatment than from an alternative. We can only observe the response to a treatment the patient actually receives; we cannot know if they would have responded more or less favourably to an alternative treatment. A patient can thus not be labeled as benefiting or not. New methods need to be developed to deal with this problem.

This thesis presents several different algorithms that can train a model capable of identifying patients that will benefit more from the treatment of interest than an alternative treatment. In **Chapter 2,3 and 4** we use the concept of Simulated Treatment Learning (STL). STL relies on the idea that genetically similar patients who received different treatments can be used to model the response to an alternative treatment. Similarity between patients should be defined by genes relevant to treatment benefit. As we do not know beforehand which genes are relevant, the algorithms we build to implement STL need to both select relevant genes and use these to build a model that can classify new patients.

In **Chapter 2** we present GESTURE (Gene Expression-based Simulated Treatment Using similaRity between patiEnts) and demonstrate its utility in Multiple Myeloma, a plasma cell cancer. GESTURE uses predefined gene sets, informed by biological annotation, to define similarity between patients. It then tests which of these gene sets can be used to identify a class 'benefit', i.e. patients who benefit more from the treatment than the population as a whole. We show it can do so successfully for two major treatments in Multiple Myeloma: bortezomib and lenalidomide.

In **Chapter 3** we implement the concept of STL in the algorithm STLsig, which does not need predefined gene sets. While GESTURE could predict in unseen data which patients would benefit from bortezomib or lenalidomide, it produced models that contain hundreds of gene sets and thousands of genes. These models are complicated to interpret. Instead, STLsig builds gene networks specific to the disease and treatment by connecting pairs of genes that are synergistic in their ability to predict benefit. With STLsig we define a 14-gene model that can predict benefit to proteasome inhibitors (like bortezomib) in Multiple Myeloma. These 14 genes present a much simpler model and they are moreover unique: a model with similar performance cannot be found when they are removed from the dataset.

In **Chapter 4** we adapt GESTURE to predict chemotherapy benefit in breast cancer. Breast cancer patients have on average a much better survival than Multiple Myeloma patients. This poses a statistical challenge as the majority of the patients included in the dataset are still alive at the end of follow-up. When two similar patients from different treatment arms are both still alive, we cannot define who benefited more. The adapted version, GESTURE-BC, uses a different criterion to define the best classifier better suited to a dataset with few recorded deaths. We show that GESTURE-BC can identify which patients see benefit from chemotherapy treatment and which patients do not benefit. However, this model did not show performance on older data where patients were treated along different guidelines. This highlights the importance of matching the patient populations in which a model is trained and in which its performance is evaluated.

In **Chapters 2, 3 and 4** we use tumor gene expression to predict treatment benefit. However, this is not the only factor influencing response. In **Chapter 5** we introduce

RAINFOREST (tReAtment benefIt prediction using raNdom FOREST), which predicts treatment benefit using germline DNA variation, which is the inherited genetic variation and not specific to the tumor. We use RAINFOREST to predict cetuximab benefit in metastatic colorectal cancer.

The algorithms presented have different strengths and weaknesses. STLsig provides simpler models, but is less adept at dealing with low event rates, which GESTURE-BC can deal with. Neither can deal with non-continuous data, which RAINFOREST can do. Together, GESTURE, STLsig and RAINFOREST provide a versatile toolbox to predict treatment benefit in different settings and using different data types.

Samenvatting

Veel kankerbehandelingen zijn geassocieerd met ernstige bijwerkingen, terwijl het bekend is dat niet alle patiënten die er mee behandeld worden baat hebben bij het medicijn. Het is bekend dat zowel patiënt- als tumorkenmerken de respons van een kankerpatiënt op een specifieke behandeling kunnen beïnvloeden. Er is dan ook grote belangstelling voor gepersonaliseerde geneeskunde: het matchen van het juiste medicijn met de juiste patiënt, op basis van bepaalde voorspellende kenmerken die kunnen worden gemeten. Sommigemedicijnen zijn gericht op een specifieke mutatie in het DNA van de tumor; dit medicijn is alleen nuttig voor patiënten wiens tumor deze mutatie herbergt. Maar gepersonaliseerde geneeskunde kan ook een rol spelen bij meer generieke behandelingen. In het verleden is machinaal leren ("machine learning") toegepast om patiënten met een slechte en goede respons op een bepaald medicijn van elkaar te onderscheiden. Dit is bijvoorbeeld gedaan op basis van genexpressie in de tumor.

Vaak is er echter meer dan één medicijn beschikbaar en moet er een keuze worden gemaakt welk medicijn het beste is voor de patiënt. Dit is een moeilijker probleem dan respons voor één medicijn voorspellen. De meeste methodes voor het voorspellen van een goede of slechte respons hebben labels nodig; patiënten worden gelabeld als goede of slechte responder en het model wordt geoptimaliseerd om deze groepen van elkaar te onderscheiden. Deze methodes kunnen niet worden gebruikt om te voorspellen of een patiënt meer baat zal hebben bij een bepaalde behandeling dan bij een alternatief. We kunnen alleen kijken naar de respons op een behandeling die de patiënt daadwerkelijk krijgt; we kunnen niet weten of die beter of slechter gereageerd zou hebben op een alternatieve behandeling. Een patiënt kan dus niet worden gelabeld als wel of geen baat hebben. Er moeten nieuwe methoden worden ontwikkeld om dit probleem aan te pakken.

Dit proefschrift presenteert verschillende algoritmen die een model kunnen trainen dat in staat is om patiënten te identificeren die meer baat hebben bij een bepaalde behandeling dan bij een alternatief. In **Hoofdstuk 2, 3 en 4** gebruiken we het concept van Simulated Treatment Learning (STL). STL is gebaseerd op het idee dat genetisch vergelijkbare patiënten die verschillende behandelingen hebben gekregen, kunnen

worden gebruikt om de respons op een alternatieve behandeling te modelleren. Genetische gelijkenis tussen patiënten moet worden bepaald door genen die relevant zijn voor baat bij de behandeling. We weten niet op voorhand welke genen relevant zijn. Daarom moeten de algoritmes zowel relevante genen kunnen selecteren, alsook nieuwe patiënten kunnen classificeren met behulp van deze genen.

In **Hoofdstuk 2** presenteren we GESTURE (Gene Expression-based Simulated Treatment Using similaRity between patiEnts) en demonstreren we het nut ervan in multipel myeloom, een plasmacelkanker. GESTURE maakt gebruik van vooraf gedefinieerde verzamelingen van genen ("gene sets"), gevormd aan de hand van biologische functie, om de gelijkenis tussen patiënten te definiëren. Vervolgens wordt getest welke van deze *gene sets* kunnen worden gebruikt om een 'baat'-groep te identificeren, d.w.z. patiënten die meer baat hebben bij de behandeling dan de rest van de patiëntenpopulatie gemiddeld heeft. We laten zien dat GESTURE in staat is dit te doen voor twee veel gebruikte medicijnen in multipel myeloom: bortezomib en lenalidomide.

In **Hoofdstuk 3** implementeren we het concept van STL in STLsig, een algoritme waarbij het niet nodig is van te voren *gene sets* te definiëren. Hoewel GESTURE in staat is te voorspellen welke patiënten baat zouden hebben bij bortezomib of lenalidomide, gebruikte het hiervoor modellen met honderden *gene sets* en duizenden genen. Het is lastig dit soort modellen te interpreteren. In plaats van vooraf gedefinieerde *gene sets* te gebruiken, maakt STLsig netwerken van genen die specifiek relevant zijn voor de ziekte en de behandeling. Om deze netwerken te maken, verbinden we genen die samen beter in staat zijn om baat te voorspellen, dan met een ander gen. Met STLsig trainen we een model dat baat kan voorspellen voor proteasoomremmers (zoals bortezomib) in multipel myeloom. Het model gebruikt slechts 14 genen en vormt Hiermee een veel simpeler model. Bovendien zijn deze genen uniek in hun voorspellende waarde: als we deze uit de dataset verwijderen kunnen we geen model vinden dat even goed kan voorspellen welke patiënten baat hebben.

In **Hoofdstuk 4** passen we GESTURE aan om te voorspellen welke borstkankerpatiënten baat hebben bij chemotherapie. Borstkankerpatiënten overleven gemiddeld veel langer na de diagnose dan patiënten die lijden aan multipel myeloom.

Daardoor was de meerderheid van de patiënten in de dataset aan het einde van de follow-up periode nog in leven. Dit maakt het statistisch gezien lastiger om een model te trainen. Wanneer twee vergelijkbare patiënten die verschillende medicijnen hebben gekregen allebei nog in leven zijn, kunnen we niet bepalen of de ene patiënt meer baat heeft gehad dan de andere. De aangepaste versie van GESTURE, GESTURE-BC, definieert de beste classificatie met een ander criterium. Dit criterium train op zowel baat als geen baat en rangschikt de *gene sets* op een andere manier, waardoor we een beter model kunnen trainen op data met goede overleving. We demonstreren dat GESTURE-BC kan voorspellen welke patiënten baat hebben bij behandeling met chemotherapie en welke niet. Dit model werkt echter niet goed op een andere, oudere dataset, waar de patiënten volgens andere richtlijnen werden behandeld. Dit laat zien dat het erg belangrijk is om de patiëntenpopulatie waar het model op getraind wordt, te matchen met de populatie waar het in getest wordt.

In **Hoofdstuk 2, 3 en 4** gebruiken we genexpressie van de tumor om baat bij een behandeling te voorspellen. Dit is echter niet de enige factor die de respons beïnvloedt. In **Hoofdstuk 5** introduceren we RAINFOREST (tReAtment benefIt prediction using raNdom FOREST), dat gebruik maakt van verschillen in kiemlijn DNA om baat bij behandeling te voorspellen. Kiemlijn DNA is DNA dat overgeërfd kan worden, dit is dus anders dan het (gemuteerde) DNA van de tumor. We gebruiken RAINFOREST om te voorspellen welke patiënten baat hebben bij behandeling met cetuximab bij uitgezaaide darmkanker.

Al deze algoritmen hebben zwakke en sterke kanten. STLsig kan simpelere, beter te interpreteren, modellen trainen, maar kan minder goed omgaan met data waar de meeste patiënten nog in leven zijn; daar kan GESTURE-BC beter mee omgaan. Allebei de algoritmes hebben continue data zoals genexpressie nodig om gelijkenis tussen patiënten te definiëren, terwijl RAINFOREST ook om kan gaan met andere soorten data. GESTURE, STLsig en RAINFOREST vormen samen een toolbox om baat bij behandeling te voorspellen voor verschillende soorten kanker en verschillende soorten data.

List of publications

Ubels, J., Sonneveld, P., van Beers, E.H., van Vliet, M.H and de Ridder, J. Predicting treatment benefit in multiple myeloma through simulation of alternative treatment effects. *Nat Commun* **9,** 2943 (2018). https://doi.org/10.1038/s41467-018-05348-5

Allahyar A., **Ubels J**. and de Ridder J. A data-driven interactome of synergistic genes improves network-based cancer outcome prediction. *PLoS Comput Biol* **15**(2): e1006657 (2019). https://doi.org/10.1371/journal.pcbi.1006657

Ubels J., Sonneveld P., van Vliet, M.H. and de Ridder, J. Gene networks constructed through simulated treatment learning can predict proteasome inhibitor benefit in Multiple Myeloma. *Clin Cancer Res* September 10 2020 DOI: 10.1158/1078-0432.CCR-20-0742

In press

Ubels J., Schaefers T., Punt C., Guchelaar H.J., and de Ridder J. RAINFOREST: A random forest approach to predict treatment benefit in data from (failed) clinical drug trials. *Bioinformatics* In press (2020)

Acknowledgements

And here we are. While it's still hard to believe that I have now actually made it to the end of my PhD, I know for sure I would not have gotten here without the support of a lot of people. So I'm going to try (and probably inevitably fail) to thank them all.

Ik wil eerst graag **Van Herk Charity** bedanken voor het ondersteunen en mogelijk maken van dit PhD project. Ik wil ook graag **prof. Ivo Touw**, **prof. Marcel Reijnders** en **prof. Peter van der Spek** bedanken voor het plaatsnemen in de beoordelingscommissie van dit proefschrift.

Jeroen, bedankt voor al het advies, support en peptalks van de afgelopen 5 jaar. Het proces is zeker niet zonder hobbels geweest, maar in alle lastige periodes kon ik op je steun rekenen en dat heeft heel veel voor me betekend. Je enthousiasme is erg aanstekelijk; ondanks mijn neiging resultaten nogal eens negatiever te interpreteren dan ze zijn, was ik na een research meeting met jou altijd weer enthousiast over mijn project. Ik heb ook veel plezier gehad aan al onze niet-wetenschappelijke discussies, al ben ik op politiek vlak wellicht wat minder makkelijk te overtuigen. Ik ben heel blij dat ik destijds besloten heb met je mee te gaan naar Utrecht. Bedankt voor alles!

Pieter, bedankt voor al je input en feedback op mijn werk. Jouw klinische blik was onmisbaar voor de multiple myeloma hoofdstukken, die er zeker een stuk beter door zijn geworden. Ik wil je ook bedanken voor de vrijheid die je me hebt gegeven om mijn eigen projecten te ontwikkelen, ook toen dat betekende dat mijn proefschrift wat minder MM gefocust werd.

Martin, je bent mijn supervisor geweest sinds mijn allereerste regel code. Dankzij jouw begeleiding tijdens mijn masterstage had ik direct een goede introductie in de bioinformatica en kon ik de switch vanuit het wetlab maken in mijn PhD. Ook heb je het mogelijk gemaakt dat ik op deze bijzondere manier, tussen bedrijf en universiteit in, mijn PhD kon doen. Bedankt voor alles wat je me geleerd hebt en alle support de afgelopen 7 jaar!

Dharminder, ontzettend bedankt voor dit alles mogelijk maken. Ik vind de vrijheid die ik heb gekregen om mijn eigen onderzoek te doen en tegelijkertijd een kijkje te nemen bij het bedrijfsleven heel bijzonder. En wat voor bedrijfsleven! Ik heb menigeen jaloers

gemaakt met de verhalen over Skyline's borrels en retreats. Het was zeker "our party" and ik heb heel veel lol gehad! Heel veel succes met "the next big thing" en ik reken op die alumni party ;)

And so a big thank you to all my colleagues at **SkylineDx** that have helped and supported me over the years. **Erik**, alle discussies met jou hebben me veel nieuwe inzichten gebracht. Bedankt voor alle gezelligheid en alle taxiritjes van en naar Utrecht! **Kim,** via jou ben ik bij Skyline terecht gekomen en is het allemaal begonnen. Bedankt voor al het advies in het begin van mijn carrière en, misschien nog wel belangrijker, de onvergetelijke bedrijfsuitjes en borrels! **Fanni,** ik heb met niemand zoveel gelachen op werk als met jou. Je maakte het altijd leuk om naar kantoor te komen en een kopje koffie op de 11e was altijd een goede start van de dag! **Domenico,** I am always impressed by your enormous knowledge about everything in the field (and quite a few topics outside our field). No matter what the question, you always have something to contribute or a paper to suggest. My research is certainly better for all the discussion we've had, thank you! **Arjan**, ik ga de ritjes naar Rotterdam missen! Zeker in de laatste fase van mijn PhD was het goed om even stoom af te kunnen blazen. **Rowan,** bedankt voor al het advies rondom de MM data en het samenstellen van alle datasets. Je hebt me een hoop werk bespaard! **Anton,** without you and all your efforts to get the data, the RAINFOREST chapter would not be there. I know the whole GESTURE project could be quite annoying, but you made it happen and some great results came out of it!

Deniz, I cannot thank you enough for the amazing design of my thesis, it makes it so much better! I'm very happy to have met you and look forward to many more game nights and city tours.

While I was only there for a short time, I want to thank all the colleagues at the Bioinformatics lab at the **TU Delft** for getting my PhD off to such a good start. I learned a lot about machine learning and programming, which was invaluable coming from a biology background.

A big thank you to the Hematology department at **Erasmus MC**. Even though my PhD moved into a different direction when I moved to Utrecht, I was always welcome to present in lab meetings and at the Friday morning meeting whenever I needed input. **Mark,** bedankt voor al je hulp met de Hovon65 data en alle input die je hebt geleverd

bij de lab meetings. **Annemiek,** bedankt voor al de goede feedback op mijn eerste paper.

Elisa, Gosia, Kyoko and Mirella, I really enjoyed being on the RSG board with you! You guys certainly improved my organisational skills.

Bedankt **prof. Henk Jan Guchelaar**, voor de samenwerking op het RAINFOREST project en het mogelijk maken om aan de CAIRO2 data te werken.

Thanks to everyone in the **CMM** for the invaluable scientific discussions and for making my PhD so much fun! I will not attempt to name everyone I enjoyed the Friday borrel with over the past 4 years, but thank you for always providing a fun start to the weekend. I have had many enlightening discussions, honed my darts and beer pong skills, learned how to open a beer bottle on the crate, and learned how to tie my shoelaces in a manner that saves 2 seconds, what more can you want?

Mircea, I've missed having you in the department in last year of my PhD, I'm sure I could have used your relaxing influence. Thank you for all the amazing, albeit frequently quite pointless, discussions and the reminder that not everything has to be serious all the time and there's always some time to have fun. **Glen,** I think you are just about the nicest person I have ever met. Chatting with you has cheered me up countless times! **Jose,** thanks for making the department so much more fun. I've very much enjoyed all the beers we've shared over the years. Good luck with finishing up! **Chris** and **Christina,** I'm happy we had our little end-of-PhD support group, it made the whole frustrating and difficult process a bit easier and less lonely. And we all got there! **Luan,** giver of invaluable advice like 'just don't be scared'. Thanks for some of the most random discussions I've ever had and of course, countless bouldering tips. **Sakshi,** I am so inspired by your perserverance and optimism, even in the face of great challenges. Thank you for all of the discussions and life advice!

And of course, a massive thank you to all former and current members of the Riddertjes. I couldn't have wished for a better or more supportive group of people to do my PhD with and I have had the time of my life with you guys. **Amin,** man, thank you for - well, everything. You taught me a lot about machine learning and computer science, which has surely helped my PhD a lot. I really enjoyed working with you on SyNet. I am so

impressed with your ability to always ask the most important questions and identify where we can improve. But besides being a great scientist, you are also a great friend! You made me feel welcome from my very first day in Delft and have been a great support to me during the entire process. I will never forget that you always made time to have a cup of coffee with me and reassure me about the state of my PhD, no matter how busy you were yourself. I am so very happy you decided to stick around! **Sara,** thanks for all the advice, mentorship and friendship. And of course for the epic trip to Oxford! Your amazing passion for science has inspired me a lot (and look, I now even have a GWAS chapter in my thesis!). **Joep,** wat een verlies voor het lab toen jij naar Nieuw-Zeeland vertrok! Je was een oneindige bron van wijsheden en interessante feiten en de eerste hulp voor gestreste PhD studenten. Gelukkig onthoud je je zelfs vanaf de andere kant van de wereld niet van advies! **Roy,** ik ken niemand die zo matter-of-fact met de slimste oplossingen kan komen. Je was dan ook een enorme (en meestal geduldige ;)) bron van hulp voor de non-computer scientist. **Myrthe,** mijn fellow east side kletser. Het was heel gezellig dat je erbij kwam, ook al was het misschien niet altijd het beste voor de productiviteit. Bedankt voor alle check-ins en moral support tijdens de laatste loodjes! **Adrien,** I'm so happy you joined our group, you bring great spirit to everything you do. And of course, thank you for the best game of Clue ever! To my fellow PhD students, **Joanna B, Joanna W, Marleen, Alexandra, Liting, Emmy** and **Marc,** thanks for always being there to give advice, commiserate and have nice masterclasses and retreats with! **Joanna,** my work spouse. Ik ben heel blij dat je bij ons in het lab gekomen bent en dat nu we geen collega's meer zijn, we wel vrienden zijn - zoals officieel established in Oxford. Het was altijd gezellig (misschien soms iets te gezellig), maar ik heb ook heel veel aan je steun gehad in de moeilijke tijden. Heel veel succes met je laatste jaar! **Joanna W,** bedankt voor alle prachtige designs. Ik heb veel geleerd van al je vormgeving tips en mijn papers zijn er zeker mooier van geworden! **Marleen**, ik ben altijd onder de indruk van hoe hard en gestructureerd jij kan werken. Ons project samen is misschien niet echt van de grond gekomen, maar het heeft me wel geïnspireerd tot betere organisatie. Heel veel succes met laatste loodjes, ik weet zeker dat er straks een prachtig proefschrift ligt! **Liting,** you've been a great addition to east side office and thank you for taking such good care of Boete during the lockdown! **Emmy,** mijn corona office buddy! Het is een lastige tijd om een PhD te starten, maar je bent ondanks dat voortvarend van start gegaan. Ik weet zeker dat het project – welke richting het ook krijgt – bij jou in goede handen is. Succes en – minstens zo belangrijk – veel plezier!

Tilman, thanks for all the help on the RAINFOREST project and forcing me to finally use GitHub properly! **Jasmin,** our Shut Up And Write! adminstrator and focus hour monitor. I'm sure some parts of this thesis only got written on time because of your killer combination of discipline and cookies. **Flip,** oplosser van elk mogelijk probleem. Ik denk niet dat ik ooit tevergeefs iets aan je hebt gevraagd. **Monique,** bedankt voor alles wat je geregeld hebt – te veel om op te noemen! - en uiteraard voor pas weggaan nadat ik klaar was met mijn PhD.

Wouter, van de therapie-geitjes voor mijn first year evaluation tot aan het diner om mijn thesis submission te vieren, je bent er mijn hele PhD geweest. Bedankt (en sorry) voor alle avonden waar je mijn gestress aan moest horen en je weer moest zeggen dat het wel goed kwam en ik heus mijn PhD wel zou halen. Kijk eens aan, je had gelijk. Bedankt!

En tot slot, de grootste support van allemaal, mijn familie. Jullie hebben allemaal geweldig met me meegeleefd (en geleden, sorry) de afgelopen jaren en zijn de basis voor alles. **Pap,** op deze mag je trots zijn, want het is zeker ook dankzij jou. Bedankt voor de eindeloze steun en interesse de afgelopen jaren, ik vond het altijd erg leuk om samen over mijn onderzoek te praten. **Mam,** bedankt voor elk succes, groot en klein, met me vieren. En als de successen even uitbleven en het moeilijk werd, stond je altijd klaar met een maaltijd en zo nodig een borrel. **Richtje,** bedankt voor al het no-nonsense advies als ik weer eens een veel te gestrest was over de kleine dingen en had en zo nu en dan het nodige duwtje om het gewoon te doen. **Myrthe en Robbert,** ik denk niet dat dit proefschrift er nu was geweest zonder jullie opvang tijdens de lockdown. Zo ontzettend bedankt. Ook bedankt aan kleine **Kees**, die een welkome afleiding en dagelijkse portie vrolijkheid was in zware tijden. **Myrth,** je hebt zonder twijfel dit proefschrift gered in de laatste fase, maar je was natuurlijk al ver voor corona een grote steun en altijd geïnteresseerd in mijn werk. En laten we niet vergeten dat dit alles begonnen is met jouw suggestie Kim eens te bellen! En als laatste **Sam,** mijn eeuwige steun en toeverlaat. Ik denk niet dat ik jou ooit genoeg kan bedanken voor alles wat je voor me hebt gedaan de afgelopen jaren. Zoals mijn collega's al constateerden nadat je weer eens voor me gekookt had, je bent de beste broer ooit.

Joske

Curriculum Vitae

Joske Ubels was born on December 12th 1992 in Amstelveen, the Netherlands. She grew up in Amersfoort, where she attended Corderius college for her pre-university education. In 2007/2008 she attended the Advanced Academy of the University of West Georgia in Carrollton, Georgia, USA. After returning to the Netherlands she obtained her high school diploma and started the bachelor Biomedical Science at the Vrije Universiteit in Amsterdam in 2009. In 2011 she attended Lunds Universitet in Lund, Sweden for her minor in Neurobiology. After completing an internship in the Oncogenomics lab of the Cancer Center Amsterdam she became interested in oncology research and started the research master Oncology at the Vrije Universiteit in Amsterdam. She first became acquainted with bioinformatics research during an internship at SkylineDx in 2013 and became so enthusiastic she decided to pursue a PhD in the topic. She started her PhD in September 2015 under the supervision of prof. dr. Pieter Sonneveld at Erasmus MC and dr. Jeroen de Ridder, first at the TU Delft, and later at UMC Utrecht. She currently works as a postdoctoral researcher in the lab of dr. Ruben van Boxtel at the Princess Máxima Center for Pediatric Oncology.

PhD portfolio

Name PhD student:	Joske Ubels	PhD period:	Sept 2015 – July 2020
Erasmus MC department:	Hematology	Promotor:	Prof. dr. P. Sonneveld
Research school:	Molecular Medicine	Co-promotor:	Dr. J. de Ridder

Training

Courses	**Year**	**ECTS**
BioSB course Pattern recognition	2015	3
BioSB course Algorithms for Biological Networks	2016	3
Presenting – Breaking Science	2018	2
Research planning and Time management	2018	0.5
Psychological Flexibility	2018	1
Scientific Artwork with Photoshop and Illustrator	2019	0.6
Interpersonal Communication	2019	0.5
This thing called science	2019	2
Total		**12.6**

Cancer, Stem cells & Developmental biology PhD Program	**Year**	**ECTS**
CSND PhD masterclass 2017	2017	1
CSND PhD retreat 2018	2018	1
GSLS PhD Day – Talking science	2018	0.3
CSND PhD masterclass	2018	1
GSLS PhD Day – Transparent Science	2019	0.3
Total		**3.6**

Conferences	**Year**	**ECTS**
Bioinformatics & Systems Biology meeting	2016	1
European Conference on Computational Biology (poster)	2016	1
Bioinformatics & Systems Biology meeting (talk)	2017	1
European Hematology Association meeting (E-poster)	2017	1
Intelligent Systems for Molecular Biology (poster)	2017	1
Bioinformatics & Systems Biology meeting (poster)	2018	1
Utrecht Bioinformatics Center symposium (talk)	2018	1
Bioinformatics & Systems Biology meeting (poster)	2019	1
Intelligent Systems for Molecular Biology (poster)	2019	1
Total		**9**

Teaching

	Year	**ECTS**
Daily supervisor literature study MSc	2017	1
Daily supervisor mini-project MSc	2017	2
Daily supervisor literature study MSc	2017	1
Daily supervisor minor internship MSc	2018	3
Supervision paper discussion CSND introductory course	2018	0.1
Lecturer CSND course Analytics & Algorithms for Omics Data	2019	2
Total		**9.1**

Total ECTS	**34.3**