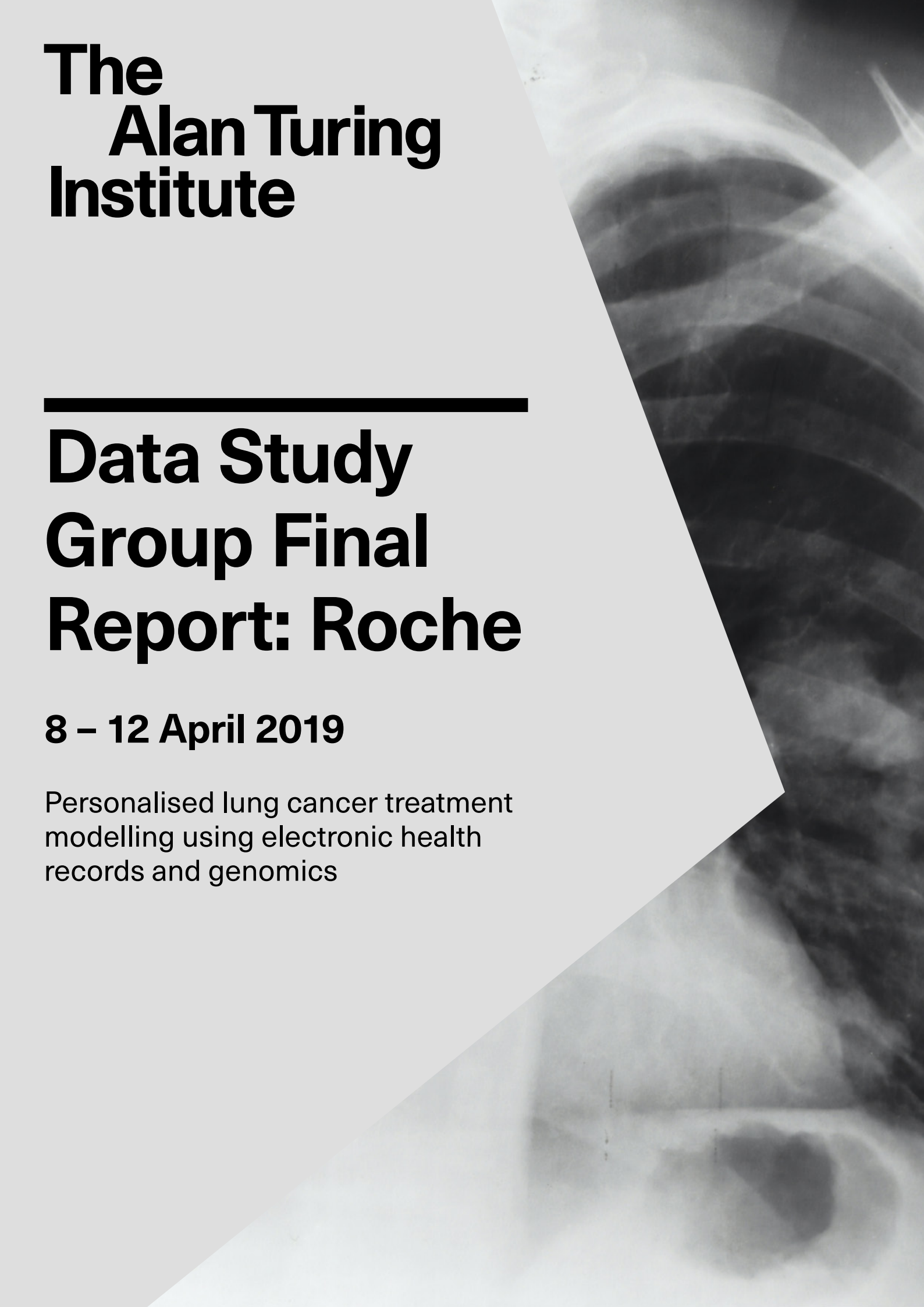
The Alan Turing Institute

Data Study Group Final Report: Roche

8 – 12 April 2019

Personalised lung cancer treatment modelling using electronic health records and genomics



https://doi.org/10.5281/zenodo.3876989

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1

1 Executive Summary

1.1 Challenge Overview

Cancer immunotherapy (CIT) is a promising new type of cancer treatment that uses the patient's own immune system to fight cancer cells. CIT drugs work to stop the cancer cells from turning off the immune system's T-cells by inhibiting the PD-L1 produced by the tumour cells (PD-L1 is a protein that binds to PD-1 receptors on T-cells and prevents the immune system from attacking the cancer cells).

CIT is currently being used to treat patients with non-small cell lung cancer (NSCLC) for whom chemotherapy or other drugs have failed. CIT is also being used as part of the first-line treatment in patients with advanced NSCLC (aNSCLC - stage III and higher). Theoretically, patients with high PD-L1 expression levels are more likely to respond well to CIT; however, in practice, patient outcomes vary considerably.

In this data study group, we investigated different approaches for predicting survival time for patients treated with CIT as first line of treatment, using both electronic health records and tumour genomic data. We also investigated the causal effects of CIT vs other oncology treatments, and studied treatment heterogeneity. The results contribute to identifying patients who are most likely to benefit from CIT.

1.2 Data Overview

Two main datasets, provided by the Roche, were used in this data study:

1. Electronic health records of 51,884 aNSCLC patients, collected by Flatiron Health, Inc. from over 280 clinics across the United States. This dataset contains data on patients' demographics, diagnosis, line(s) of therapy received (including medications administered), lab test results, progression and mortality. We will refer to this dataset as the 'EDM' dataset.
2. Electronic health records and genomic data by Foundation Medicine, Inc. for a subset of 5,866 aNSCLC patients in the EDM dataset. This dataset, in addition to the tables present in the EDM dataset, contains data on patients whose clinicians requested Foundation Medicine biomarker testing, and includes all classes of mutation in approximately 400 genes, computationally determined ancestry, therapy details and tumour response. We will refer to this dataset as the 'FMI' dataset.

Note that these datasets had to be analysed separately, as even when they contain the same patients, identifiers in each of the two datasets are different (by design).

1.3 Main Objectives

The overall aim of this data study was to explore various methods for predicting patient survival after CIT using electronic health records and genomic data. Specifically, we sought to answer the following questions:

1. Which model most accurately predicts patient survival after CIT initiation? and does adding genomic features to demographics and clinical variables improve the accuracy of the models?
2. Which treatment option (CIT or chemotherapy) is likely to be more effective in aNSCLC patients?
3. Are there subgroups of patients who benefit more from CIT?

1.4 Approach

To answer the first question (see Section 1.3), we performed:

- Modelling of the relationship between variables using probabilistic classification models;
- Binary prediction of patient survival using random forests;
- Estimation of patient survival using the Kaplan-Meier estimator;
- Statistical analysis of baseline and time-varying factors using the Cox regression model and dynamic prediction of patient survival using joint longitudinal-survival models.

To answer the second question, we compared different treatment options (CIT, chemotherapy, combination) using a formal potential outcomes framework, to obtain the average causal treatment effect. Assuming we have enough variables to control for the confounding between CIT-treated and not treated, we estimated the causal effects of CIT on 6-month survival, by targeted maximum likelihood estimation, combined with SuperLearner, an ensemble method, that included random forests and gradient boosting amongst the candidate learners.

Finally, to answer the third question, we performed:

- Dimensionality reduction and cluster analysis using spectral clustering and the PAM algorithm;
- Data visualisation using the t-SNE algorithm;
- Cluster analysis using an integrative approach;
- causal forests, to characterise the factors most strongly associated with heterogeneity of the causal treatment effect comparing CIT to other treatments.

1.5 Main Conclusions

The results of binary prediction and estimation of patient survival revealed the lab tests (including Albumin, Calcium, Creatinine and Haemoglobin levels) done prior to the initiation of CIT as first-line treatment to be highly predictive of survival, while demographics (e.g., smoking history) were found to be not of predictive relevance in the fitted models. These results agree with those of dynamic prediction of patient survival.

Furthermore, adding the genomic data did not improve the quality of binary predictions. Similarly, analysis of the data using unsupervised methods (cluster analysis and data visualisation) with and without the genomic data showed that without the genomic features, better clusterings are obtained.

The results of comparing different treatment options showed that having immunotherapy as part of the treatment (vs chemotherapy alone) increases the patient survival time and hence, treatment options that include immunotherapy are likely to be more effective.

1.6 Limitations

We encountered the following limitations regarding the data and the methods that were applied:

- Missing data;
- The EDM and FMI datasets are not linkable and hence were processed separately;
- The bias introduced in the data as a result of genetic data being available for only a subset of patients;
- The data are observational and hence, causal inferences are subject to strong unverifiable assumptions holding;
- Lack of a 'control group' and no validation dataset that we could use to benchmark the methods.

1.7 Recommendations and Future Work

Below is a summary of the future research avenues proposed:

- Additional data pre-processing to address data quality issues;
- Including additional variables in the models;
- Optimising the methods that were applied to the data;
- Extending the analyses to other patient cohorts;
- Validating the results in intervention trials.

2 Quantitative Problem Formulation

Our main task in this data study was to develop accurate models for predicting patient survival time from the initiation of treatment, as well as to identify factors that are predictors of survival time. Given that the outcome variable is time from treatment initiation until the patient's death, this problem falls into the domain of survival analysis. Moreover, survival analysis methods can handle censored (or incomplete) observations, one of the limitations of this dataset.

To exploit the large collection of existing machine learning tools, we considered a simplified version of this task by focusing on predicting binary survival (alive or dead) at 6 and 12 months for a cohort of patents who had received CIT as first-line treatment. Two categories of models commonly used for this task are: probabilistic classification models (e.g., regression models to model the survival time as a function of a set of predictor variables) and decision trees (e.g., random forests).

We then considered the more complex task of dynamically predicting survival time. Three easy-to-interpret models widely used for this task are: the Kaplan-Meier estimator, which can be used to estimate survival probabilities as a function of time and compare survival curves for two or more groups of patients; the Cox proportional hazards regression model, which allows testing for differences in survival times of two or more patient groups while allowing to adjust for covariates of interest; and joint longitudinal-survival models for studying the association between time-updated variables (e.g., monthly lab test results) and survival time.

In addition to supervised methods, we also considered the application of unsupervised methods such as cluster analysis and data visualisation. We began with clustering and dimensionality reduction techniques as the data are high-dimensional. We then employed an integrative clustering approach. In integrative cluster analysis, we assume the data arise from multiple sources (e.g., if we aim to infer groups of patients, data for each patient are available across a number of datasets). Such approach is commonly used in the analysis of genomic datasets wherein we observe multiple measurements for patients (e.g., mRNA, microRNA, reverse phase protein arrays, DNA methylation, somatic copy number alterations, etc.). It may not be suitable to model such datasets as arising from a common family of statistical distributions and so we cannot simply concatenate these data and analyse them using standard clustering methods such as k-means or hierarchical cluster analysis.

Our other task in this data study was to compare the effect of CIT to that of other treatment options on patients' survival. For this task, it is important to recognise that the data arose from routine clinical practice, so the patients receiving CIT as first-line, are very likely to be systematically different in their baseline characteristics to those not receiving it in terms of, for example, their disease severity, age, lab results and even socioeconomic status. These baseline characteristics which are simultaneously associated with the treatment received and survival time are referred to as *confounders*. Assuming we control for all the confounders, we performed causal inference to estimate the average causal treatment effect of CIT vs the rest (adjusting for the confounding by using inverse

probability of treatment weights). We then studied treatment effect heterogeneity, and described the covariates which are most associated with heterogeneous causal effects. These covariates are potential effect modifiers (subgroup effects that should be investigated in future studies).

3 Dataset Overview

3.1 Data Preparation

The datasets provided by Roche contain clinical data for over 50,000 aNSCLC patients, of whom around 6,000 have had genomic profiling. For this subset, clinical and genomic data were linked at the individual patient level, resulting in a rich clinico-genomic database.

The data for this data study were prepared as follows: First, patient identifiers and demographics for qualifying patients (e.g., patients who received CIT as first-line treatment) were extracted. Next, relevant index dates (e.g., start date of CIT) were created, and for each patient, *baseline lab and clinical data* were extracted, defined as the last measurement recorded prior to the index date. These variables were merged with the patients' demographic data. The following prognostic factors that characterise patients with lung cancer, identified from the literature [1], were considered:

- Age
- Gender (higher treatment success rate in females)
- Histological type of cancer
- Co-morbidities
- Smoking history
- Additional malignancies (in particular neuroendocrine tumours)
- Insurance category (a good proxy for socioeconomic status)

The following lab variables were also chosen according to the literature [1]:

- Albumin levels (kidney/liver function)
- Creatinine levels (kidney/liver function)
- Haemoglobin levels (blood disorders)
- Calcium levels (thyroid tumour indicator)
- Lactate Dehydrogenase (cell damage indicator)
- Neutrophil count (immune system response)
- Leukocyte count (immune system response)
- Lymphocytes (immune system response).

These factors are re-measured over time, so baseline and time-updated versions were extracted and used in accordance with the goal of the corresponding analyses.

Genomic data preparation

The PD-L1 expression level, a key indicator of immune response, was processed as follows: we mapped the 'PercentStaining' and 'ExpressionLevel' variables to the same categories: <1% to PD-L1-NEGATIVE, 2%-49% to PD-L1-LOW and 50%-100% to PD-L1-HIGH. Of the two variables, we chose 'ExpressionLevel' as the PD-L1 expression level if it existed, and chose 'PercentStaining' otherwise. This is due to the staining being less precise.

Missing values, resulting from the cases where the patient never had a test or the date of the test is unknown, were coded as 'Unknown'. Unfortunately, baseline PD-L1 expression level was missing for many patients, so this variable was ultimately discarded.

The genomic data from the FMI dataset were prepared from data generated by six enrichment capture kits: CF2 (n=616), D2-R2 (n=2), DX1 (n=208), T4b (n=35), T5a (n=494) and T7 (n=4511). While CF2 and D2-R2 captured free circulating DNA and RNA from peripheral blood, all other kits were applied to solid tumour samples. Since different gene panels covered different genes, in order to exclude spurious correlations, analyses using these data were restricted to patients whose DNA was sequenced using the T7 kit. The following variants were included:

- Single nucleotide variants (SNVs)
- Copy number variants (CNVs)
- Non-human variants (NHs)
- Large rearrangements (REs)

These variants were transformed into a gene burden score (sum of variants identified within each gene) for each variant type.

Note that when preparing the genomic data, we assumed equal burden and directionality of the variants. An optimal modelling of the genomic data would include biological variables such as variant zygosity (heterozygous or homozygous) and rarity, tumour cellularity and whether a variant is germ-line or somatic. Similarly, when modelling CNVs, we only reported if an abnormal number of copy numbers had been identified. An improved version of this modelling would include also the extent of duplications and their rarity.

In the last step of data preparation, we added the outcome (e.g., survival, 6 and 12-month mortality) variables.

3.2 Data Quality Issues

- Some patients were lost to follow-up;

- The lab test data for a large portion of patients were incomplete. The analyses that required these data were performed on complete cases (also known as list-wise deletion, i.e., patients were removed from the analyses sets). This could potentially introduce biases to our results;
- For some patients, ethnicity data are missing;
- Using different tools for the genetic tests could potentially have introduced bias into the genomic data.

4 Binary Prediction and Estimation of Patient Survival

4.1 Task Description

In this section, we predicted survival for patients who received CIT as first-line treatment and identified the variables associated with survival. To do this, we:

- Modelled the relationship between variables using probabilistic classification models;
- Predicted binary patient 6-month survival using random forests;
- Estimated patient survival using the Kaplan-Meier estimator.

4.2 Experimental Setup

Data

For probabilistic classification models, lab test data and demographics from the EDM dataset were used. As previously described (see Section 3.1), only patients with all the 8 baseline lab variables and demographic data were included.

For binary predictions, clinical and genomic data from the EDM and FMI datasets were used.

For estimating patient survival, a subset of observations for 7,700 patients from the EDM dataset was used, including the follow-up time (in days) since the date of the first round of CIT until the date of death, when it was present, or the date of the last visit recorded otherwise. In the latter case, the observations correspond to censored patients. The event indicator is thus set to 0, if patient is censored and 1 otherwise. These data also included age, gender, smoking status, histology and the results of any lab test done immediately before the initiation of CIT.

Methods

Three probabilistic classification models with different penalisation schemes were applied to the EDM dataset:

Method	Accuracy (t=0.3)	Accuracy (t=0.8)
Binary GLM	0.887	0.843
RLR	0.887	0.877
LASSO	0.887	0.847

Table 1: Accuracies achieved by three probabilistic classification models on the EDM dataset using two different thresholds

- A binary generalised linear model (GLM, logistic regression) with no penalisation;
- A ridge logistic regression (RLR) model that penalises large coefficients and shrinks them towards zero to reduce overfitting;
- A LASSO logistic regression, which penalises the coefficients in such a way that many of them are identically zero.

The penalty (for each penalised logistic regression) was chosen by 10-fold cross-validation.

Random forest classifiers were applied to the EDM and FMI datasets. In addition, a random forest regressor was trained to predict patient survival time (in days) for those who died within 6 months. The Kaplan-Meier estimator was used to estimate patient survival time (in months) for a subset of patients from the EDM dataset.

4.3 Results

Demographics were consistently considered unimportant by all probabilistic classification models. The binary GLM model showed statistically significant coefficients for Albumin (-ve), Haemoglobin (+ve), Calcium (+ve) and LDH (+ve). The RLR model showed similar results, with additional strong positive associations with Creatinine, Leukocytes and Neutrophils. Finally, LASSO showed similar strong positive associations with Haemoglobin and Calcium, with other effects reduced. Coefficients for Lymphocytes, histology and smoking status were set to 0. Positive values suggest an increased chance of survival at 6 months, whilst negative values suggest a decreased probability of survival. Table 1 shows the accuracies of predictions for these three models using threshold cutoffs of 0.3 and 0.8 (prediction values below the thresholds are set to 0 and 1 otherwise).

The random forest classifier applied to the EDM dataset achieved an out of bag (OOB) score (1-error, where error is the average error cross-validated in a manner similar to k-folds, averaged across the 'bags' of data on which each tree was trained) of 83%. This random forest's relative feature importance (variable importance) revealed that lab test factors are the most predictive of survival at 6 months. Age at treatment initiation is also important. Demographics are relatively less important, including smoking status. While the performance of this model is promising, the data included approximately 85% of individuals in

Method	Data	Performance
RFC (6-month binary outcome)	EMD	OOB = 83%
RFC (12-month binary outcome)	EMD	OOB = 65%
RFR (6-month survival time)	EMD	$R^2 = 0.854$
RFC (6-month binary outcome)	FMI (demographics only)	OOB = 63%
RFC (6-month binary outcome)	FMI (genomic data only)	OOB = 50%
RFC (6-month binary outcome)	FMI (all, except lab data)	OOB = 63%

Table 2: Performances of a random forest regressor (RFR) and five random forest classifiers (RFC) on the EDM and FMI datasets

one class and thus, the model did not improve on a naive prediction strategy based on class weighting.

Random forests were also used for 12 months survival outcome (defined as being alive 12 months after treatment initiation). Variable importance did not change substantially relative to the rankings found for 6 months, however the OOB score dropped to 65%. This is concerning given that a naive prediction strategy based on class weighting would achieve a score of 75%.

To confirm the validity of the lab test factors observed when predicting 6 and 12 months survival, a random regression forest was used, including the same predictor variables as the binary outcome models. This model's relative variable importance confirmed that lab tests done immediately before CIT initiation are key prognostic factors for short-term mortality. This model achieved an R^2 value of 0.854, indicating that 85% of the variance in observed survival time within-6-months is explained. This has to be interpreted in light of potential systematic differences in those who are censored within the first 6 months and those who survive longer than 6 months (the complement of the set studied in this analysis).

As for the FMI dataset, a random forest classifier was applied to 6-month binary survival outcome, each of the demographic data only, genomic data only and the combination of both, but crucially not lab data. The best performance with an OOB score of 63% was observed when modelling demographics. We did not observe any improvement by adding the genomic data to this model. However, these models did not include lab test variables, which were found to be highly predictive of survival by the models applied to the EDM dataset. Moreover, the genomic data were modelled without considering some important biological variables which might have increased their discriminatory power, and the results were not corrected for factors such as ancestry. Also, it is worth mentioning that in this subset of the dataset, the number of features was approximately three times larger than the number of samples. Table 2 summarises the performances of these methods.

Figure 1 shows the survival curve for all the patients in the EDM dataset, generated by the Kaplan-Meier estimator. The vertical and horizontal lines correspond to the median survival time (see vertical axis), which is 18 months (see horizontal axis). Figure 2 shows the survival time estimates stratified by the three histology groups. A logrank test showed significant difference in sur-

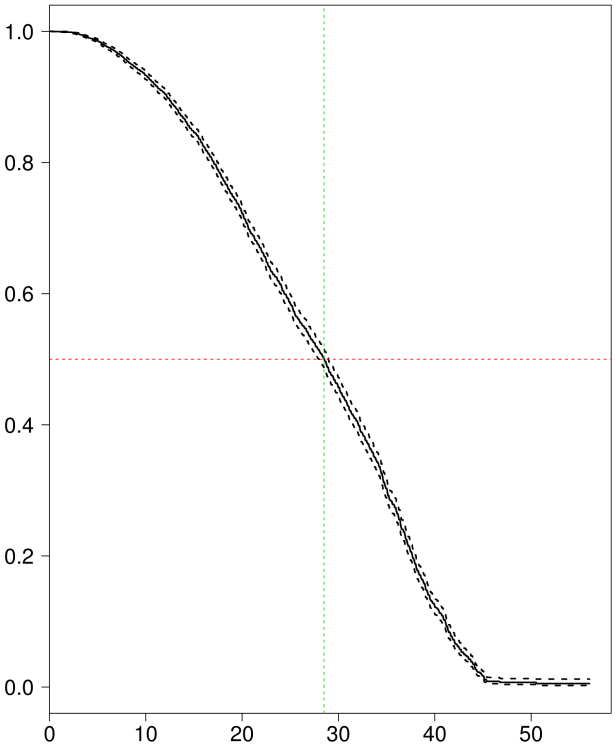


Figure 1: Survival curve (Kaplan-Meier) for patients in the EDM dataset. The vertical and horizontal axes represent the survival rate and time (months), respectively.

vival times across these histology groups. Figure 3 shows the survival estimates stratified by quantiles of the Creatinine level. A logrank test showed significant difference in survival times in the four groups: patients with a Creatinine level between 0.87 and 1.1 units survived the longest (on average 30.2 months), and those with the minimum Creatinine level survived the shortest (on average 26.2 months).

A similar analysis showed significant difference in survival times depending on Albumin, Calcium or Neutrophils levels with p-values (computed by a logrank test) of <1e-6. The higher the Albumin level, the longer the survival time: the average survival times are 22.9 months for patients with Albumin levels in the lowest quartile of the distribution (6.7 to 32 units) and 35.1 months for patients with Albumin levels in the highest quartile of the distribution (40 to 298 units). The average difference in survival times with respect to Calcium levels range from 25.3 months for patients in the lowest quartile (1.08 to 8.7 units) to 30.6 months for patients with a Calcium level between the median and 3rd quartile of the distribution (9.1 to 9.5 units). The lower the Neutrophils level, the longer the survival time: on average, 30.3 months for patients with Neutrophils levels between 0 to 3.6 units and 25.1 months for patients with Neutrophils levels of 8.9 to 5700 units.

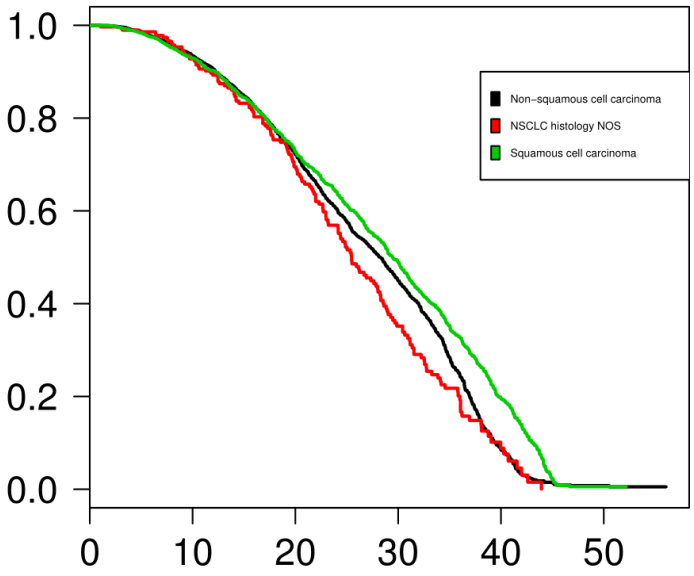


Figure 2: Survival time estimates for patients in the EDM dataset stratified by three histology groups. The vertical and horizontal axes represent the survival rate and time (months), respectively.

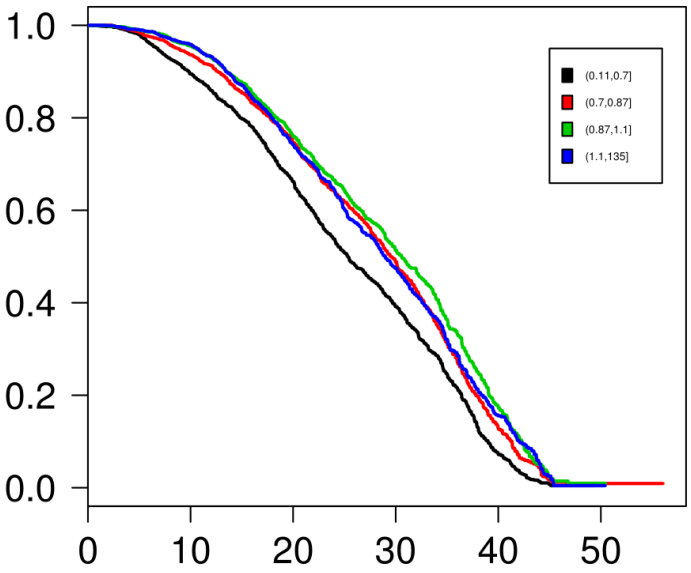


Figure 3: Survival time estimates for patients in the EDM dataset stratified by quantiles of the Creatinine level. The vertical and horizontal axes represent the survival rate and time (months), respectively.

5 Dynamic Prediction of Patient Survival

5.1 Task Description

In this section, we performed survival analysis using dynamic prediction models for patients who received CIT as first-line treatment, to be able to model the relationship between time-varying clinical measures and mortality in these patients.

5.2 Experimental Setup

Data

- Population: a subset of 11,142 patients from the EDM dataset who had started CIT as first-line treatment were considered; 6,116 (55%) patients died over the study period. Mean follow-up was 1.6 years (SD of 1 year and maximum follow-up of 4.6 years);
- Design: time of CIT initiation was used as entry date. Patients were followed up to the earliest of the date of death, the study end date (December 2018) or the last recorded follow-up date, irrespective of any changes in their treatment or whether immunotherapy was discontinued;
- Baseline measures: the most recent measures prior to the entry date were extracted;
- Time-varying measures: lab test measures were extracted for each patient from the study start date to the end of follow-up. We looked at Albumin (as serum Albumin has been demonstrated to be a strong predictor of survival in patients with lung cancer, with low serum Albumin being associated with reduced survival. 9,421 (85%) patients with a baseline Albumin measure prior to the index date were eligible for inclusion in the analysis. After the index date, there were on average 14 time-varying Albumin measures per patient) as well as other prognostic markers including Haemoglobin, Lactate Dehydrogenase and Neutrophils to Lymphocytes ratio.

Methods

For survival analysis, we used joint longitudinal-survival models, which allow investigation of the association between a biomarker's evolution over time and survival time. We fitted joint models for Albumin and Haemoglobin, incorporating all post-baseline measures for each patient (using Weibull baseline hazard function and shared parameter joint model linked through subject specific random effects).

For modelling the relationship between time-varying measures and survival time, with time horizon of 6 months, we used Cox regression models.

5.3 Results

Association between Albumin and survival: Time-varying Albumin was found to be predictive of survival time, according to a Cox regression model (adjusted for baseline Albumin) reporting a hazard ratio (HR) of 0.96 (95%CI, lower=0.95, upper=0.97).

A strong association between time-varying Albumin and survival time was also found by the joint model, reporting a HR of 0.91 (95%CI, lower=0.90, upper=0.92).

Association between Haemoglobin and survival: Baseline Haemoglobin was found to be predictive of survival time, according to a Cox regression model reporting a HR of 0.94 (95%CI, lower=0.92, upper=0.96). Time-varying Haemoglobin was also found to be predictive of survival time according to the Cox regression model (adjusted for baseline Haemoglobin) reporting a HR of 0.90 (95%CI, lower=0.88, upper=0.91).

Similarly, a strong association between time-varying Haemoglobin and survival time was found by the joint model, reporting a HR of 0.96 (95%CI, lower=0.95, upper=0.97).

Overall, time-varying Albumin and Haemoglobin levels measured after the initiation of immunotherapy may be good predictors of future mortality risk.

6 Comparative effectiveness of CIT

6.1 Task Description

In this section, we studied the causal effects of three treatment options: (1) chemotherapy only, (2) CIT only or (3) combined therapy. We examined the impact of each type of treatment on patient survival at 6 months after treatment initiation (first-line), while adjusting for confounding by including all baseline variables as potential confounders in analyses. In addition, we obtained counterfactual predictions of 6-month survival for each treatment option (this is the predicted survival had a patient been treated, possibly contrary to fact, with each of the treatment options). Counterfactual predictions can be used to inform decision making when treating new patients.

6.2 Experimental Setup

Data

A subset of 1,887 patients from the FMI dataset were considered (602 patients were exposed to CIT while 1,285 patients had mono-therapy; 878 patients were exposed to chemotherapy while 1,009 patients had another type of treatment involving immunotherapy). The data contain 1,592 distinct variables, including age, gender, race, histology, smoking status plus variables indicating disease

progression, additional malignancies, tumour type, tissue of origin and purity (both pathology and computational estimates), PD-L1 status, ECOG value and tumour mutational burden (TMB) score. Genomic data were also included.

For some patients, the clinical variables mentioned above were missing. We used the *missingness indicator approach*. That is, for missing categorical variables, an extra category was created to indicate that the value of that variable is unknown. For missing continuous variables, we performed mean imputation and created a separate binary missing indicator variable. We the added all the missing indicator variables to the set of variables available to control for confounding. This method is valid if the variables with missing data are only confounders when observed, and not when missing. The genetic features are available for all the patients.

The treatment options (in terms of combinations of drugs) were defined as follows:

1. Chemotherapy:
    - Carboplatin, Pemetrexed
    - Carboplatin, Paclitaxel
2. Immunotherapy:
    - Nivolumab
    - Pembrolizumab
3. Combined chemotherapy and immunotherapy:
    - Bevacizumab, Carboplatin, Pemetrexed
    - Carboplatin, Pembrolizumab, Pemetrexed
    - Bevacizumab, Carboplatin, Paclitaxel

While there exist many other treatment options, the ones chosen were the most frequent.

The outcome was 6 months survival, i.e., being alive 6 months after treatment initiation (binary).

Methods

We estimated the average causal treatment effect for each treatment vs the others. Due to the presence of confounding by indication (where typically sicker patients are prescribed the latest treatment), we used causal inference techniques to obtain unbiased treatment effect estimates, assuming we have adjusted for a sufficient set of variables to achieve conditional exchangeability (referred to as no unobserved confounding) in our models. In addition, we made the following assumptions:

- Positivity: all patients have non-zero probability of being assigned each treatment option;

- No unobserved confounders: all variables affecting both the treatment assignment and outcomes are measured and adjusted for;
- There is no interference between patients.

First, for each treatment vs the union of the other two, we generated a propensity score model, adding all baseline variables as potential confounders. This propensity score model was fitted using the SuperLearner (SL) algorithm [4], to avoid unrealistic parametric modelling assumptions. SL is an ensemble of algorithms (e.g., random forests, logistic regression, etc.) which chooses a convex combination of the trained learners that yields the best performance. We used 5-fold cross validation. The trained SL propensity score model was used to predict individual propensity scores, to obtain inverse propensity score weighting and to re-weight the observed data (known as inverse probability weighting - IPW).

Second, for each treatment option, we also generated data-adaptive (binary) outcome models, controlling for all baseline variables as confounders, using the SL algorithm (same library as before). To do this, we split the data randomly into training (80%) and test (20%) subsets, and ran the SL algorithm with 5-fold cross-validation. The generated models were then used to obtain counterfactual outcome predictions for each treatment option (employing a one-vs-rest strategy). This aims to capture the 6-month survival that each individual would have had, if possibly contrary to fact, they had been treated with each of the treatments, in turn. We also ran outcome models using weighted loss functions, with the weights being the inverse of the estimated propensity scores, wherever possible (for example, for random forests). This was done so that the outcome model is "tuned" better in the regions of poor overlap, where the PS weights are larger. An outcome model was then fitted using the weights in the loss functions. The parameters for this model were tuned using cross-validation.

6.3 Results

When computing the average treatment effects, the results showed that having immunotherapy as part of the treatment increases the probability of survival at 6 months. Figure 4 shows the probability of survival at 6 months for CIT plus immunotherapy vs chemotherapy treatment options, when using IPW loss functions. On the other hand, the results showed that having CIT alone does not improve the outcome. Figure 5 shows the probability of survival at 6 months for CIT vs chemotherapy plus immunotherapy treatment options. These can be interpreted as counterfactual predictions of survival, had the patient followed the corresponding treatment (possibly contrary to fact).

The SL algorithm selected the Ranger algorithm [5] (a fast implementation of random forests for high-dimensional data) as the best performing method. Figure 6 shows the estimated risks for each method. For a single SL run, we obtained the predicted probability of survival at 6 months after CIT vs either monotherapy. Figure 7 shows the density of probability of survival.

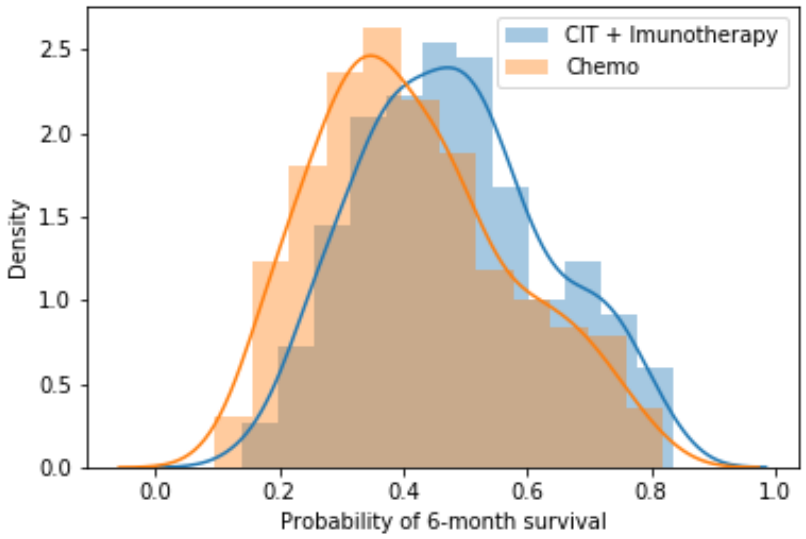


Figure 4: Probability of survival at 6 months in patients treated with CIT plus immunotherapy compared to that of patients treated with chemotherapy

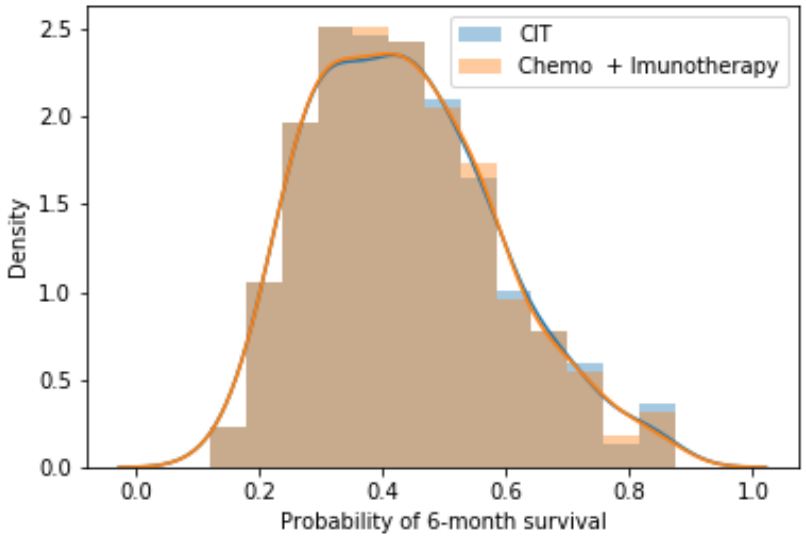


Figure 5: Probability of survival at 6 months in patients treated with CIT compared to that of patients treated with chemotherapy plus immunotherapy

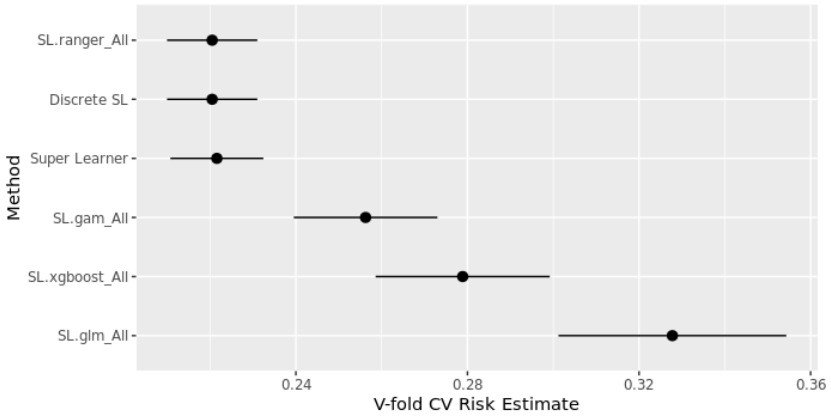


Figure 6: Estimated risks for each method used in SuperLearner

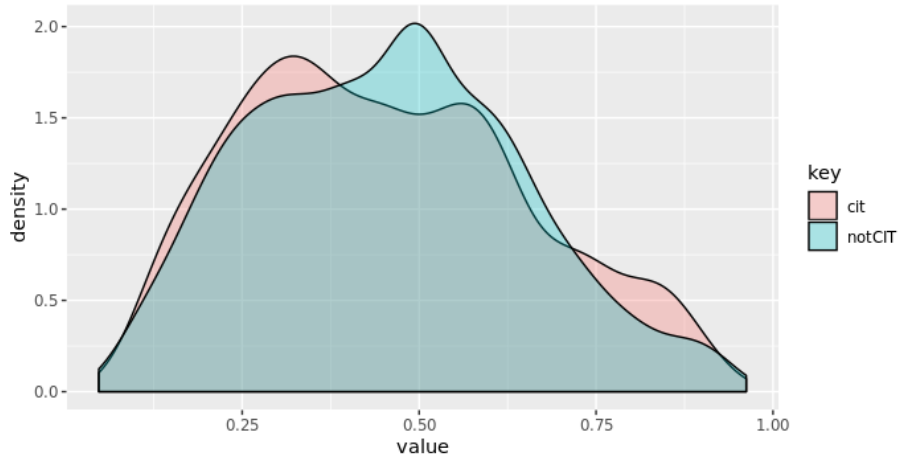


Figure 7: Density of probability of survival at 6 months in patients treated with CIT compared to that of patients not treated with CIT

7 Treatment heterogeneity

We explore treatment heterogeneity, aimed at explaining why some patients respond better to treatment, and characterising which features seem to contribute to patient response. This section reports our efforts to answer the third question set in Section 1.3.

7.1 Causal forests

For this, we used the same data as the causal inference analyses above.

Methods

Recent causal machine learning methods to test for treatment effect heterogeneity in a data-adaptive manner and avoid overfitting have been proposed. Here, we focused on the generalisation of *causal forests* presented by [6]. Causal forests adapt the random forests algorithm to train in a way that maximises heterogeneity in the estimated (or predicted) treatment effect, as opposed to minimising RMSE in outcome prediction. Causal forests also incorporate sample splitting so that the trees are honest, meaning we have valid confidence intervals. The generalised random forests framework improves on the original causal forests by also incorporating an *orthogonalisation* step to deal with confounding.

Briefly, separate random forests are built for treatment and outcome on all the potential baseline confounders. Then, predicted treatment assignment (from the propensity score random forest predictions) and predicted outcome are used to obtain residualised "outcomes" for our next models. These residuals are the difference between the observed quantity and its corresponding prediction. These residuals are used as outcomes in a second step, for building causal forests. These residuals can be thought of as a de-confounded version of treatment and outcome. The use of such residuals as outcomes is common in the doubly robust literature. A test for treatment heterogeneity is also built into the procedure. Causal forests perform sample splitting and, therefore, the trees are 'honest', i.e., they are trained on one part and evaluated on another. In addition, we used the variable importance feature of the causal forests to explore the variables most associated with treatment heterogeneity.

Results

We repeated the causal analysis using causal forests for chemotherapy vs any other treatment option involving immunotherapy. As seen in the plot of individual treatment effects shown in Figure 8, most patients would have had negative treatment effects had they been given chemotherapy compared to any other treatment option involving immunotherapy. As for the heterogeneous effects, the results showed that after adding the genomic features, treatment effects are much smaller. We explored the variables associated with treatment effect heterogeneity, and produced a variable importance plot shown in Figure 9. As

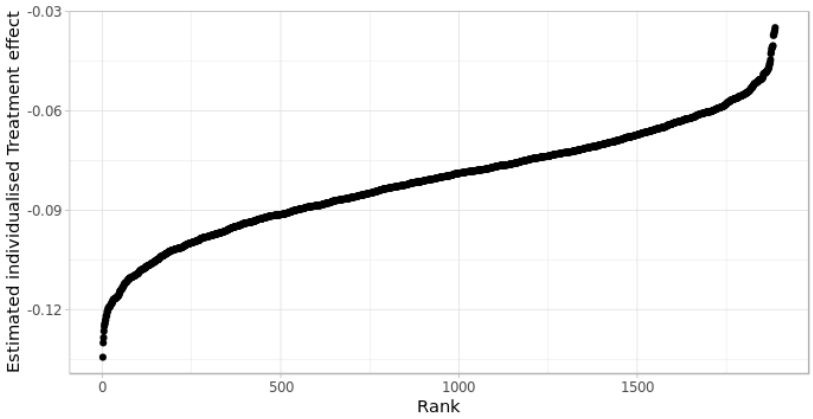


Figure 8: Individualised treatment effect for chemotherapy compared to other treatment options involving immunotherapy

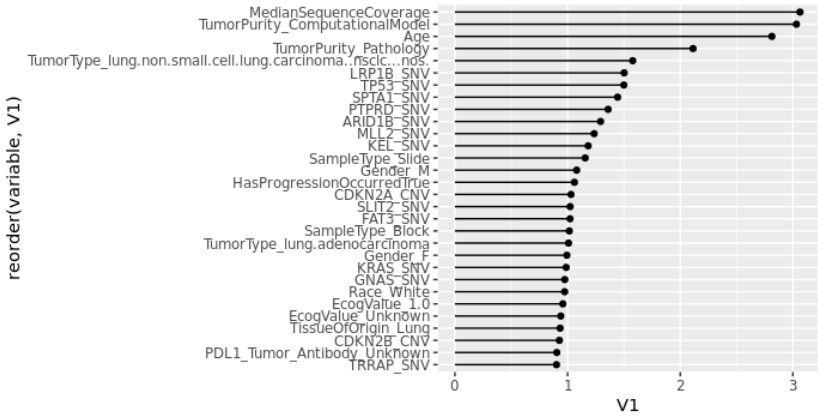


Figure 9: Variable importance plot for variables associated with treatment effect heterogeneity

seen in the plot, age, tumour purity and certain genetic mutations are the most important variables explaining the treatment effect heterogeneity. Due to time constraints, we did not apply sample splitting.

We produced plots of the individualised treatment effects on age and top three genetic mutations (LRP1B, TP53 and PTPRD), shown in Figure 10. The effect of age is quite flat. As seen in the plot, not having mutations in selected genes results in negative treatment effects with CIT, meaning that it would be beneficial to not give combination therapy to patients with these mutations.

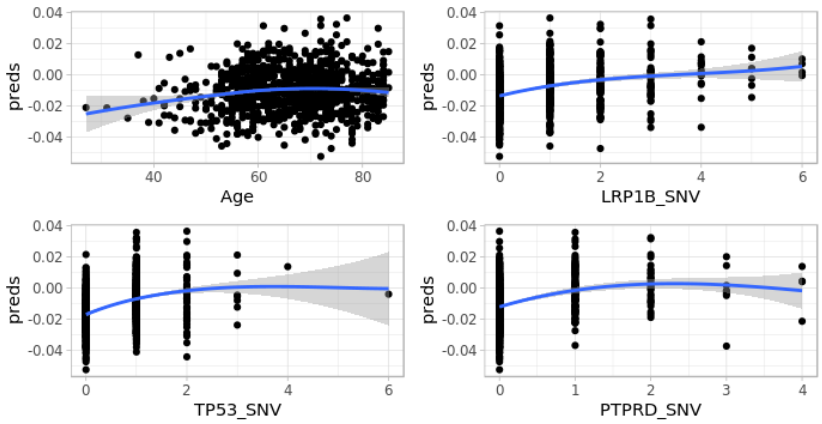


Figure 10: Individualised treatment effects on age and top three genetic mutations

7.2 Unsupervised Analysis of Clinical and Genomic Data

Task Description

In this section, we analysed both the clinical and genomic data using unsupervised methods (cluster analysis and data visualisation) to reveal the underlying structure in the data (e.g., subgroups of patients who benefit more from CIT).

Data

For cluster analysis using hierarchical and spectral clustering and the PAM (or k-medoids) algorithm as well as for data visualisation, we focused on patients in the EDM and FMI datasets who received CIT as first-line treatment.

For integrative cluster analysis, we used the data on copy number variants (CNVs) and single nucleotide variants (SNVs), filtered to select patients who received CIT but not until after their first-line treatment. Ad-hoc feature selection was performed by selecting the 100 most variant genes across the EDM and FMI datasets. Figure 11 displays a heatmap showing the SNV data following feature selection to select the 100 most variant genes. Similarly, Figure 12 displays a heatmap showing the CNV data following feature selection to select the 100 most variant genes.

Methods

We used a number of unsupervised clustering and dimensionality reduction techniques including hierarchical and spectral clustering, the PAM algorithm [2] and the t-SNE algorithm [3]. These methods were applied to two subsets of the data: the clinical data (the EDM dataset minus the lab test data and the FMI dataset minus the genomic data) and the clinical data plus the genomic features (from the FMI dataset).

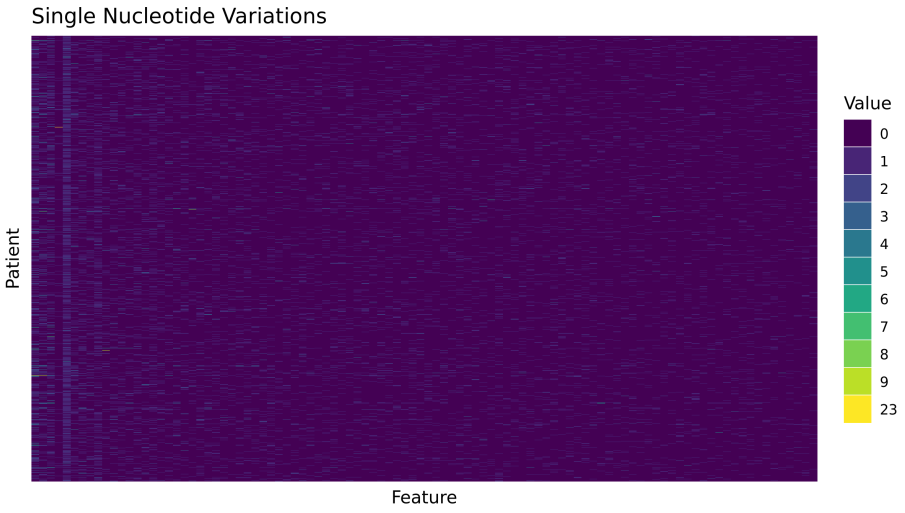


Figure 11: Heatmap of the SNV data

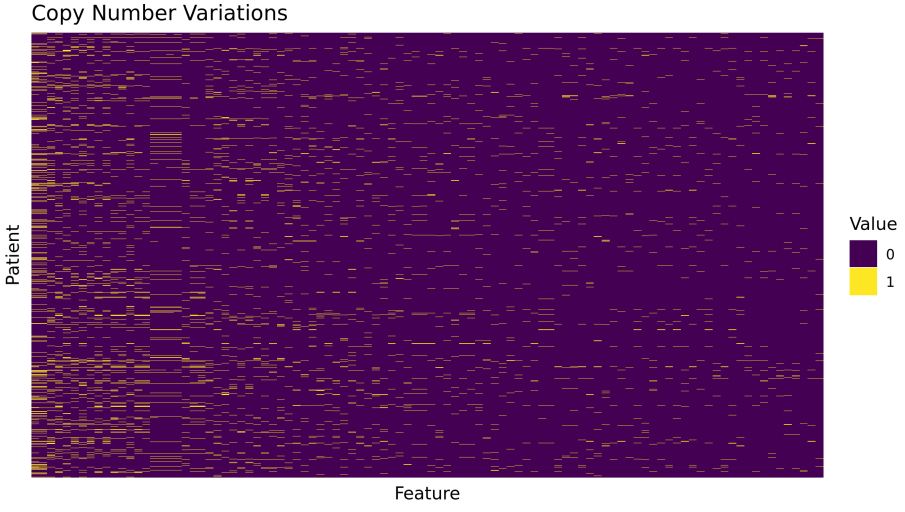


Figure 12: Heatmap of the CNV data

The parameters for these methods were set as follows: for hierarchical and spectral clustering, Euclidean distance and the RBF kernel were used to compute the distances and similarities, respectively. Perplexity and number of iterations for the t-SNE algorithm were set to 30 and 500, respectively.

In the case of hierarchical and spectral clustering and the t-SNE algorithm applied to the clinical data plus the genomic features, columns containing missing values were omitted while non-numeric values were converted to numeric ones. An approach designed to overcome the limitations of clustering mixed data types was employed using the Gower distance, partitioning around medoids and silhouette width (an internal validation metric used as an aggregated measure of how similar an observation is to its own cluster compared to its closest neighbouring cluster). To compute the Gower distance, for each variable type an appropriate distance metric was used and scaled to fall between 0 and 1. Then, A linear combination employing user-specific weights was used to compute the final distance matrix. The matrices used for each variable type were as follows:

- quantitative (interval): range-normalised Manhattan distance;
- ordinal: variable is first ranked and the Manhattan distance is then used with an adjustment for ties;
- nominal: first, variables of n categories are converted to n binary columns and then the Dice coefficient is used.

Due to the existence of custom distance matrices, partitioning around medoids was chosen as the clustering algorithm. The PAM algorithm is more robust to noise and outliers compared to the k-means, and offers the added benefit of having an observation serve as the exemplar for a cluster. To determine the number of clusters (k=5), we used silhouette width.

Integrative cluster analysis was performed using particleMDI [7], an algorithm that uses particle Monte Carlo methods to infer a shared cluster structure across multiple datasets. In contrast to many other methods, particleMDI does not enforce strict agreement in cluster allocations across all datasets, instead allowing cluster structure in one dataset to inform the cluster structure in the others. The maximum number of clusters, the number of particles, the proportion of allocations assumed known in each iteration and the number of iterations were set to 10, 32, 0.33 and 1000, respectively. The output of particleMDI can be visualised via a posterior similarity matrix (PSM), an $n \times n$ matrix with entry (i, j) indicating how frequently patients i and j (of n total patients) are assigned to the same cluster. A consensus cluster allocation across datasets can then be reached by taking the average of individual PSM values across datasets and performing hierarchical cluster analysis on the corresponding similarity matrix.

Results

Hierarchical and spectral clustering did not produce good clusterings (low silhouette coefficients) and hence, their outputs were not considered for further analysis. The t-SNE algorithm also did not produce distinct clusters. Note that when visualising the Euclidean distances prior to running t-SNE, we observed a high correlation between a subset of features.

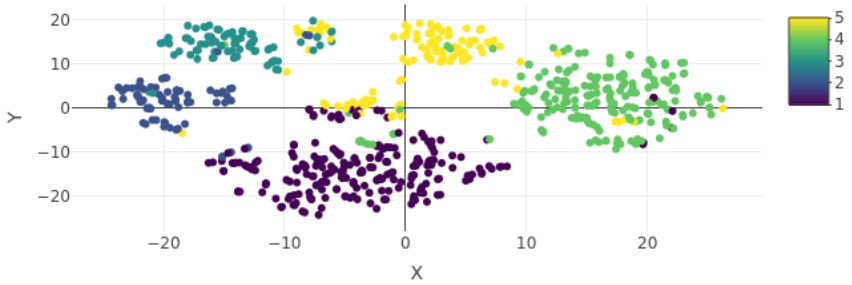


Figure 13: Result of clustering the clinical data using the PAM algorithm

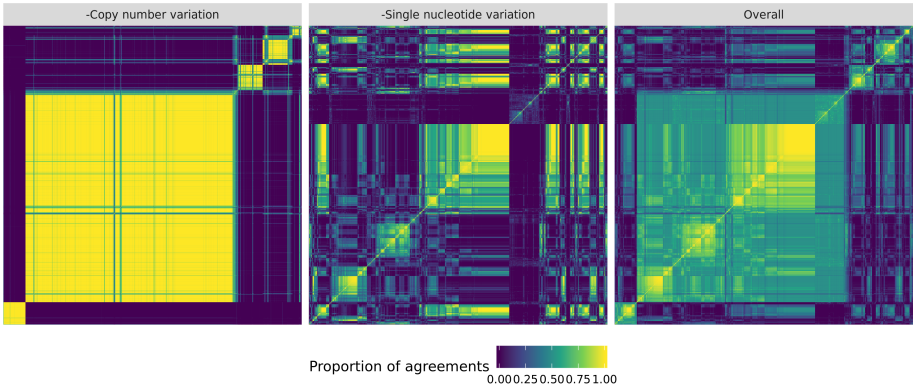


Figure 14: Heatmaps of the PSMs for the SNV and CNV data generated by the particleMDI algorithm. Patients are presented along both axes.

The PAM algorithm applied to the clinical data allowed for clustering of patients in nearly five distinct clusters. The result is shown in Figure 13. We extracted a summary of each cluster: for instance, cluster one comprises of majority white females with history of smoking and non-squamous cell carcinoma and a life span of 303 days (median), while cluster five comprises of older white males with history of smoking and squamous cell carcinoma whose life span is nearly 6 months longer (465 days median). The PAM algorithm applied to the clinical data plus the genomic features, however, did not produce any clustering.

As for the integrative cluster analysis, the consensus cluster allocation suggests the likely existence of five clusters across the datasets. Interestingly, there appears to be reasonable agreement in the structure across datasets, with the SNV data providing some additional specificity on the clusters inferred from the CNV data. Figure 14 displays a heatmap showing the frequency with which pairs of patients are assigned to the same cluster. The clusters inferred, however, do not appear to offer any clinical insights in relation to the relative survival times across patient groups.

Due to time constraints, we did not investigate the inferred clusters further.

However, evaluating the differences in survival curves for these clusters can lead to useful inputs for deciding upon patients who may particularly benefit from being treated with CIT.

8 Future Work and Research Avenues

The following research avenues were proposed by the data study participants:

- Additional data pre-processing including improved preparation of the genomic data;
- Inclusion of socioeconomic features: insurance data were available in both datasets, but were not used in any of the models. However, insurance category is a strong indicator of socioeconomic status which may have prognostic relevance (it was observed that many patients changed their insurance provider during the course of treatment);
- Extension of the analyses to other patient cohorts: for instance, can we predict survival for patients who have not received CIT at all? or can we predict patient response to CIT if the patient has been given other treatments previously? Moreover, analyses using causal inference techniques can be repeated for smaller time windows, i.e., discretising the survival time and obtaining predictions for the binary outcomes, which will allow using time-varying confounders. Such analysis can also be repeated for other lines of treatment (known as treatment intensification), exploiting the longitudinal aspect of the data to find optimal treatments (known as optimal dynamic treatment regimes or optimal policy learning);
- Optimisation of methods: due to time constraints, parameters for some of the methods were not tuned. Tuning the parameters for various methods and repeating the analyses using the features selected by feature selection methods can be performed. For dynamic prediction of patient survival, more complex joint models to evaluate other aspects of biomarker trajectory (e.g., slope or AUC, and adjustment for other baseline measures) and landmarking models that incorporate multiple time-varying measures can be employed;
- Application of other methods (e.g., Bayesian methods).

9 References

[1] Paesmans, M., 2012. Prognostic and predictive factors for lung cancer. *Breathe*, 9(2), pp.112–121.

[2] Kaufman, L. and Rousseeuw, P.J., 1987. Clustering by means of medoids. *Statistical Data Analysis Based on the L1-Norm and Related Methods*. Y. Dodge, Ed, pp.405–416.

[3] Maaten, L.V.D. and Hinton, G., 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov), pp.2579–2605.

[4] Van der Laan, M.J., Polley, E.C. and Hubbard, A.E., 2007. Super learner. *Statistical applications in genetics and molecular biology*, 6(1).

[5] Wright, M.N. and Ziegler, A., 2015. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *arXiv preprint arXiv:1508.04409*.

[6] Athey, S., Tibshirani, J., and Wager, S., 2019. Generalized random forests. *The Annals of Statistics*, 47(2), pp.1148–1178.

[7] Cunningham, N., Griffin, J.E., Wild, D.L. and Lee, A., 2018. particleMDI: A Julia Package for the Integrative Cluster Analysis of Multiple Datasets. In *International Conference on Bayesian Statistics in Action*, pp.65–74.

10 Team Members

- **Maryam Abdollahyan** is currently a Digital Fellow at the Barts Health NHS Trust. Her background is in Machine Learning with applications in Bioinformatics and Health Informatics. She was involved in preparing the data for the binary prediction of patient survival and applied the unsupervised methods (cluster analysis and data visualisation) to the clinical and genomic data. She took primary responsibility for editing the final report.
- **Ioana Bica** is a PhD student at the University of Oxford and the Alan Turing Institute. Her research focuses on developing machine learning models with applications in healthcare, in particular for estimating individualised treatment effects. She contributed to the causal predictions by performing data processing, implementing models and analysing the results.
- **Alexander Buchholz** is a postdoctoral researcher in the MRC Biostatistics Unit at the University of Cambridge working on methodologies for genomics. Before joining, he completed a PhD in Paris, specialising in Bayesian computation. He contributed to the construction of patient cohorts and pre-processing of the data. Moreover, he participated in the dynamic prediction of patient survival.
- **Susana Conde** is a research fellow in Statistics at the University of Warwick based at the Alan Turing Institute. She is currently interested in functional data analysis and methodology development with applications in biology, weather and genetics. She was involved in modelling and producing descriptive statistics of 6-month survival rates for patients in the EDM dataset.
- **Nathan Cunningham** is a PhD student in the Oxford-Warwick Statistics Programme based at the University of Warwick. His thesis focuses on the development of particle Monte Carlo methods for integrative cluster analysis of multiple genomic datasets. He worked on the preparation of datasets and the application of integrative cluster analysis to the genomic data.

- **John Dennis** is a research fellow in Medical Statistics at the University of Exeter. His research focuses on statistical methods for precision medicine in type 2 diabetes. He has a particular interest in methods for treatment selection and evaluating competing risks such as the benefits and risks of medications. He worked on data preparation for survival analysis and on dynamic prediction of mortality using joint-longitudinal survival models.
- **Karla DiazOrdaz** is an Associate Professor of Biostatistics at the London School of Hygiene and Tropical Medicine. She specialises in machine learning-enhanced causal inference methodology, with a focus on applications to electronic health records. She led the causal analyses (TMLE with Super Learner) and also those for heterogeneous treatment effects (causal forests). She edited and approved the final version of this report.
- **Fiona Grimm** is a senior data analyst at the Health Foundation. Her work focuses on using data and innovative methodologies to produce analysis on the big issues affecting health and care in the UK. She was the facilitator of the team and coordinated the project. She also contributed to study design and report writing.
- **Snezhana Ilieva** is a data scientist in PwC's Financial Services Data and Analytics practice. She holds an MSc in Statistics from the London School of Economics and Political Science. She was involved in the data preparation (including both genomic and clinical data) and application of unsupervised methods for the binary prediction of patient survival.
- **Franz Kiraly** is a Lecturer in Statistics at University College London. He was the Turing DSG Principal Investigator, helping shape this data study questions and over-seeing the work, and editing earlier versions of this report.
- **Chen Li** is a PhD student at University of Cambridge. She conducted exploratory single gene associational analysis based on the FMI data.
- **Weiqi Liao** is an ESRC-funded PhD student based at the University of Southampton. His role is to explore the use of different statistical models for investigating diagnostic pathways of lung cancer in primary care using electronic health records.
- **Enrico Mossotto** is a BRC postdoctoral researcher in Medical Machine Learning at the University of Southampton. His role consists of mod-elling next generation sequencing data from patients affected by common complex diseases and applying machine learning methodologies in order to integrate multi-omics data alongside longitudinal data extracted from electronic health records. Enrico worked on the extraction, transforma-tion and manipulation of the genomic data. Moreover, he investigated the application of random forest classifiers on a subset of individuals for which demographics and genomic data were available. He also helped with the study design for the binary prediction of patient survival.
- **Hector Page** is currently a data scientist at Privitar, a London-based software company specialising in data privacy. He holds a PhD in Computational Neuroscience from the University of Oxford, and previously held

a postdoctoral position at UCL's Institute of Behavioural Neuroscience. His group work focused on 6-months survival rates for patients who received CIT as the first-line treatment, data cleaning/wrangling, feature engineering, and random forest modelling. He also helped with the study design for the binary prediction of patient survival.

- **Mario Parreno-Centeno** is a Research Fellow in the Secrier Lab at the UCL Genetics Institute. His research involves modelling and integrating data from multi-omics sources to address various questions around systemic changes in cancer cells that may lead to cancer progression and resistance to therapies. To this end, he employs genomic, transcriptomic, methylation and digital pathology image datasets to investigate how mutational processes, tumour evolution and immunity drivers may in combination determine distinct cancer outcomes.
- **Ben Swallow** is a Lecturer in Statistics at the University of Glasgow with research interests in Bayesian inference, model selection and sensitivity in complex biotic systems. His contributions to the project consisted of extracting and cleaning lab and demographics data from the EDM dataset. He also fitted a variety of regression models to the binary survival data and assisted with the study design for the analysis of patient survival.

turing.ac.uk
@turinginst