APPENDIX

METHODS

Initial Data Processing and Handling of Missing Data

The distributions of the continuous predictor variables were assessed. Pleural fluid LDH, NLR and C-reactive protein (CRP) were log-transformed to improve assumptions of normality. The outcome variable Survival (Days) was combined with censoring information to create a single survival variable for use in Cox-proportional hazards models. Multi-level factors (e.g. symptoms, histological sub-type) were converted to numeric values to be included in subsequent models. This was achieved by creating a set of Boolean variables that recorded whether each level was observed in each sample. Any Boolean variables defined from these multi-level factors that were too sparse (defined as <5 samples being labelled with the least frequent alternative) were excluded. For example, only one patient had the symptom code for cough, therefore this was dropped from downstream analysis. Missing values in all predictor and outcome variables were imputed using the mean value for the variable of interest.

Associations between Predictor and Outcome Variables

The factors in the data collected were assessed to determine if any were significantly associated. For two categorical factors, a chi-squared test was used. For an association between a categorical factor and a continuous variable, ANOVA was used. For an association between two continuous variables, a Spearman correlation test was used.

Definition and Balancing of Training and Validation Sets

The 269 patients were divided into balanced training (n=169) and validation sets (n=100) for subsequent assessment of model performance. Median values for each variable (or level of multi-level factors) were calculated and tests of association performed between each variable and its allocated set. For categorical factors a chi-squared test was used and for continuous variables an analysis of variance was used. The training and validation sets were balanced by maximising the minimum p-value observed from all tests for association over 100,000 random allocations of samples.

Signature Generation Details

Cross-validation

Cross-validation was used to measure performance of the potential models being generated. To implement this, the training samples were subdivided into five subsets randomly and models of interest trained on 4/5 of the samples and predictions obtained for the mutually exclusive 1/5 of samples. This was repeated five times, selecting all possible combinations of 4 out of 5 subsets in order to generate predictions for all samples using models trained on mutually exclusive sets of samples. The process was repeated 10 times using different random allocations of samples to five independent subsets.

Data Normalisation

The data were normalised to ensure that the features examined had equivalent magnitude and variance. All independent variables were scaled such that they had a mean of 0 and standard deviation of 1 across samples. However, since this

normalisation procedure shares information across samples, samples were normalised within each round of cross-validation to avoid overestimating model performance.

Permutation Testing

To test whether the generated signatures performed better than would be expected by chance, permutation tests were performed. To do this, labels (i.e. measures of response) were permuted with respect to the data for each sample and the signature generation process was repeated to calculate performance measures.

Model Finalisation

Once signatures were generated for a variety of models and input data types, models were finalised where the full training set was used to train the model (rather than using subsets as is the case under cross-validation). This finalised model could then be applied to the reserved validation subset to obtain a measure of performance for the final trained model.