OraMod

VPH based predictive model for oral cancer reoccurrence in the clinical practice

TITLE	**D3.3 The validated OraMod predictive model**		
Deliverable No.	D3.3		
EDITOR	VUmc- D.E. te Beest		
Contributors	D.E. te Beest, M.A. van de Wiel, Steven Mes, R.H.Brakenhof (VUmc)		
WorkPackage No.	WP3	**WorkPackage Title**	Oral Cancer Evolution Model
Status[1]	Final	**Version No.**	2
Dissemination level	PU		
DOCUMENT ID			
FILE ID			
Related documents	DoW version 2013-09-23, OraMod_results report092015.xlsx		

Distribution List

Organization	Name of recipients
UNIPR	T. Poli, D. Lanfranco, A. Ferri, E. Sesenna, E.M. Silini, R. Perris, G. Chiari, M. Silva, N. Sverzellati, N. Bertani, S. Rossi, C. Azzoni, L. Bottarelli
VUMC	R. H. Brakenhoff, M van de Wiel, D. Te Beest, S. Mes, P. de Graaf, R. Leemans
Fraunhofer	F. Jung, S. Steger, S. Wesarg
ST- Italy	Sarah Burgarella, Marco Cereda
VCI	G. Aristomenopoulos
OneToNet	A. Ruggeri, F. Dezza, A. Turetta, G. Scoponi
UDUS	K. Scheckenbach
VTT	M. Kulju, Y. Ranta
European Commission	EC Officers and experts

[1] Status values: TOC, DRAFT, FINAL

Revision History

Revision no.	**Date of Issue**	**Author(s)**	**Brief Description of Change**
1	26-10-2015	D.E. te Beest, Mark van de Wiel	First draft
2	18-02-2016	D.E. te Beest, Mark van de Wiel	Revision

Addressees of this document

This document should be distributed as guidance to all the personnel of OraMod Consortium partners involved in the project execution.

This document is Public, therefore it will be published on OraMod web site.

Executive Summary

In this document we describe (1) how we can deliver patient-centered feedback from the predictive model, (2) by comparing the accuracy of the predictive model with the results obtained from a random forest we demonstrate that the applied regression techniques have a good performance, (3) by internal validation of the models we demonstrate their validity.

Table of Contents

List of tables and figures

1 About this document

In this document we describe how we provide patient centered feedback and how this help the clinician improve their decision. In the next section, we compare the performance of the predictive model based on regression techniques with the performance of predictive models based the random forest. The aim of this comparison is to widen the scope of methods that where considered, and to confirm that regression techniques have a good performance compared to a random forest. Finally we present the result of the internal validation of the predictive models through cross-validation.

2 Patient centered feedback

It is both crucial for the acceptance of the predictive model and for the interpretation of the predictions by the clinicians, that the predictions are transparent. Some aspect of the tumor can be very important for individual patients, but if they occur infrequent across the population level, they may not be included in the predictive model. Also, if a patient has a bad prognosis according to the predictive model, a clinicians needs to know what variables contributed to this prediction. To facilitate this we visualize the contribution of the various factors that are incorporated in the predictive model.

The predictive models used in the OraMod system are based on regression techniques. The advantage of these regression techniques is their transparency. In a regression framework, for each variable that is included in the model we estimate a coefficient. Each variable is then multiplied by its coefficient, and the final prediction is obtained by summing up all the different products (e.g. one product is, for example, coefficient*unityears). To visualize the contributions of the various variables, we can plot the product belonging to the different variables (hence the contribution of that variable to the final prediction). Within the OraMod project we can use this to visualize what variables are taken into account, and how much these variables contribute to the risk prediction. We can use this transparency to deliver patient centered feedback. Figure 1 exemplifies how these contributions can be visualized.

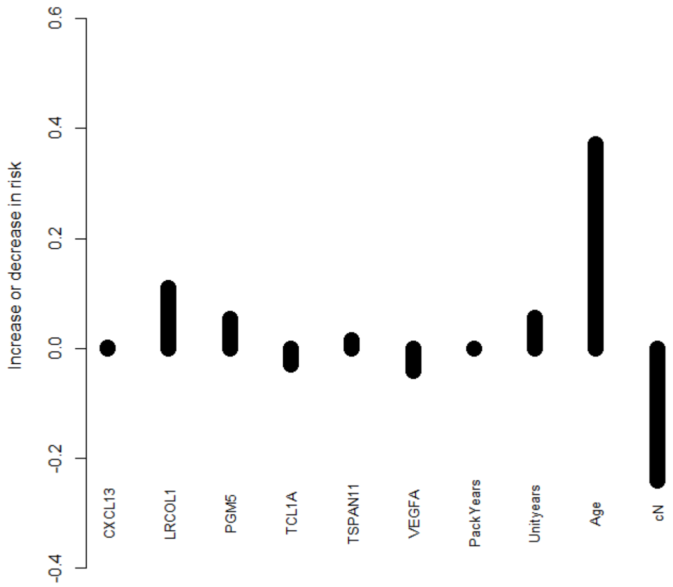

Figure 1: Example of patient centered output of the prediction model

The patient centered feedback will be completed with i) A graph that displays the patient's prediction in the context of (a selected) set of patients (Figure 2) and ii) A number that represents the uncertainty of the patient's prediction. For the first, code is available, which will be aligned with WP5, in particular for the patient selection. The latter will be computed on the prospective data, because fixing the selected variables allows the use of standard methodology to produce uncertainty estimates.

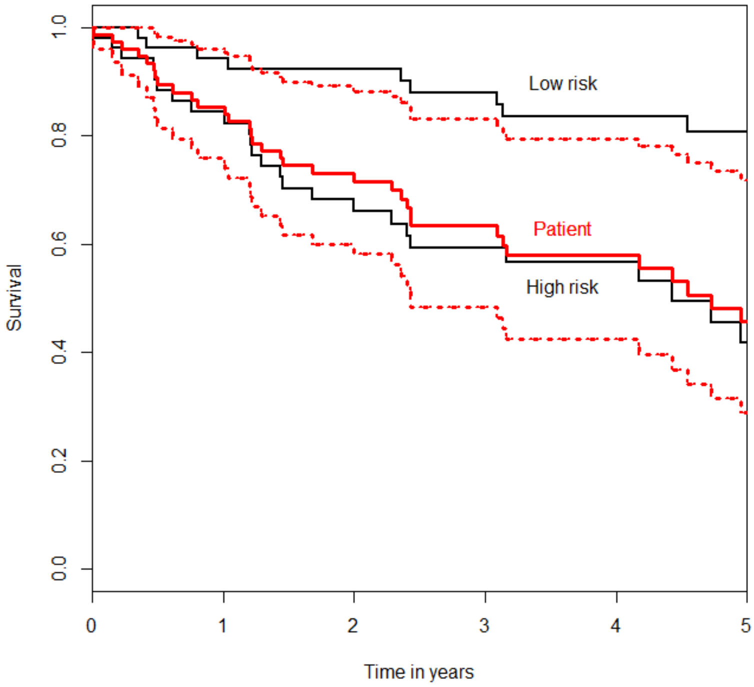

Figure 2: Example of Kaplan–Meier with in red a patient's prediction, the black lines indicate how the patient is positioned in the population

3 Comparison with Random Forest

The models described in D3.2 are built using regression techniques. To assess whether an alternative method provides a similar performance, we compare results of the regression models with those of a random forest (Breiman, 2001). A random forest is machine learning method that builds a large number (e.g. a 1000) decision trees as uses these decision trees collectively to make a prediction. The advantage of the decision trees and therefore also of a random forest is that they do not assume linearity between the predictor variables and the outcome variable. A random forest is able to deal with non-linear relationships and it is also scale invariant. Another advantage of the random forest is that it is, similar as regression methods, able to handle both classification and OS data (Ishwaran et al. 2009). One disadvantage of a random forest is that it is more difficult to combine, for example, clinical and genomics data. In a regression setting it is possible to give a penalty to the genomics data (due its higher dimension) and leave the clinical data unpenalized. In a random forest it is standard to that all variables are treated equal (which would actually put clinical data at a disadvantage). To compare both methods as fair as possible, we only compare their performance for genomics-only models. Predictions with cox and logistic regression are the result of a leave-one-out cross-validation (LOOCV). For random forest we use the out-of-box predictions (OOB). For both overall survival (OS) and Lymph node metastasis (LNM) we use in this analysis the full data set, n=125, and all 60 genes.

3.1 Overall survival

One convenient way of comparing the two methods for their OS score is through the Brier error (a standard evaluation measure). Both the random forest and cox regression have brier score of 0.224. If we look at the predicted 5 year OS probabilities resulting from both methods (Figure 2), we also see a strong agreement in the predictions.

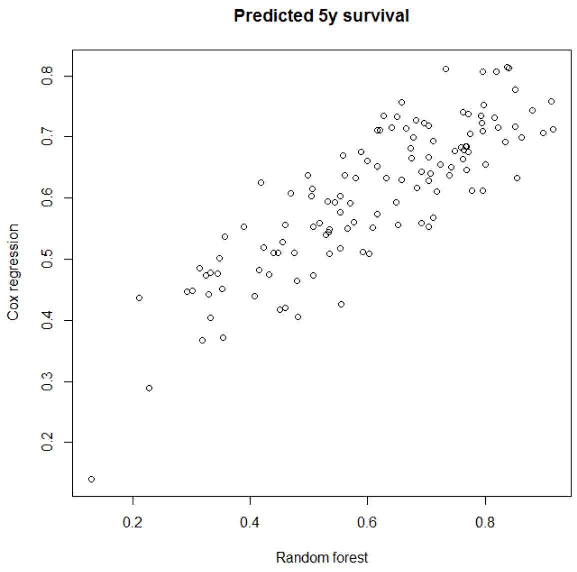

Figure 3: 5-year OS probabilties of a logistic regression versus those of of a randomFores

3.2 Lymph node metastasis

For LNM prediction, tables 1-4 demonstrate that logistic regression and random forest have comparable performance. When we compare the predictions on both methods at a patient level, we also see that two methods are in good agreement (Figure 3).

			Prediction LNM	
			Absence	Present
Randomforest	LNM	Absent	40	20
		Present	21	44
Logistic regression	LNM	Absent	42	19
		Present	19	45

Table 1: LOOCV predictions logistic regression versus OOB predictions random forest

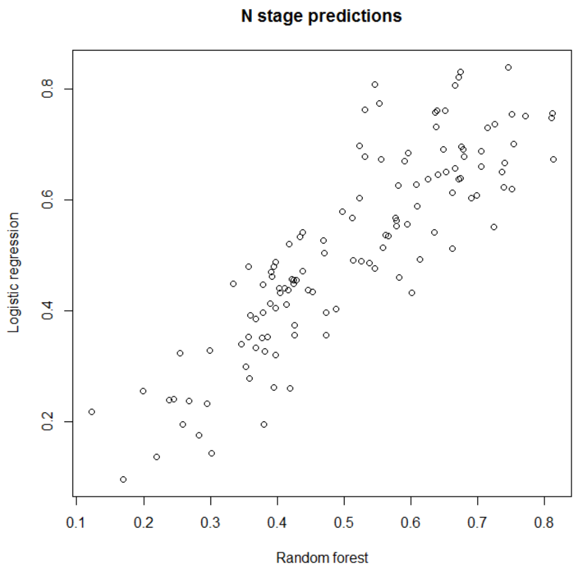

Figure 4: Predictions of a logistic regression versus the predictions of a randomForest

3.3 Conclusion

These results demonstrate that the regression techniques give a better or equal performance compare to a random forest on both binary (LNM) and OS data. The additional advantages of regression techniques are there transparency (as discussed in D3.1), and the more natural way of combining different types of data. The OraMod predictive models will thus be based on regression techniques.

4 Internal validation with cross-validation

The internal validation was conducted with a leave-one-out cross-validation. This means we leave one observation out, refit the model, and predict the observation that was left out in the first step. This procedure is then repeated for each observation. The result of this procedure is that we obtain, per observation, a prediction that is independent of the model the prediction was made with. This method gives an indication of the quality of the predictions made by the model and asses how the models will generalize to a new independent data set. For the genomics data (the selected OS genes) we take into account parameter uncertainty (by repeating the gene selection procedure each cross-validation fold). For clinical, pathology and imaging data we only asses the parameter uncertainty. For these three types of data the dimensions of the data are relatively small (e.g. low number of candidate variables), hence the model uncertainty is low. Note that in principle we could also use the LNM gene set for prediction of OS, however these genes do not contribute to the predictive power of the OS models.

For LNM classification the AUC was used as a metric for predictive performance (random classification results in AUC ≈ 0.50). For OS, the integrated AUC (iAUC; Heagerty et al, 2000) was used for evaluation of predictive performance were the we integrated over a 5-years period (to avoid a too large impact of very long survivors). For visualizing performances we use ROC curves for binary classification (LNM) and Kaplan-Meier curves for OS, where the individuals are grouped according to their cross-validated predictions. The genomic models have been fit with penalized regression. We tried both a lasso and ridge regression for all model (Hastie et al., 2001), the Kaplan-Meier plots display the fit of the ridge regression.

4.1 Clinical/Pathology/Genomics

For the model based on only the clinical data we find an iAUC of 0.692 (Figure 4). Including genomics data on top of the clinical data increases the iAUC to 0.720. The model based on clinical and pathological data has an iAUC of 0.763. Including genomics on top of clinical and pathological does not improve the iAUC. The lasso model gave similar results for the Clinical+Genomics (iAUC 0.719) and the Clinical+Pathology+Genomics models (iAUC 0.763).

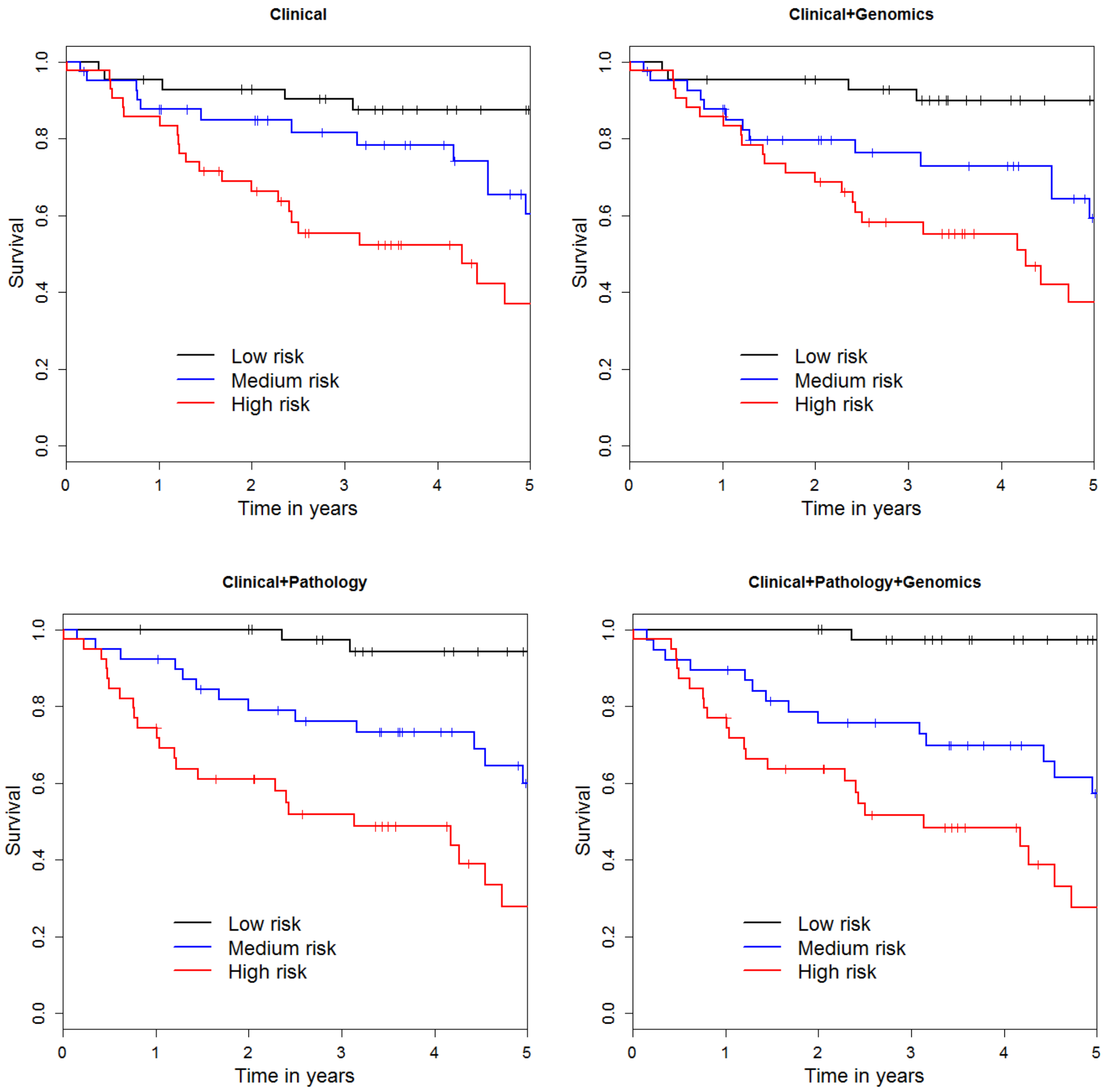

Figure 5: Kaplan-Meier curves for OraMod predictive models

4.2 Models with Imaging data

Models with tumor imaging data can be only evaluated on the subset of 70 patients for whom the imaging data is available. Calculated over these 70 patients, the clinical model has an AUC of 0.620, lower than the AUC across the whole data set. To evaluate the Clinical+Imaging model we follow the same procedure as used for variable selection of the imaging variables. We first we train the clinical model across all patients. We then use the fit of the clinical model as offset in the Clinical+Imaging model. The training of the clinical model was included in the cross-validation loop. This resulted in the iAUC of 0.628, which is marginally higher than the iAUC of the clinical model alone. A model trained in a similar fashion with additionally genomics data (with ridge regression) had an iAUC of 0.652 (Figure 5). The equivalent lasso model had an iAUC of 0.636. We should emphasize that results may improve somewhat when the prospective data are available due to the larger number of samples.

For the models with lymph node imaging data, which are Clinical + Pathology +Imaging and Clinical + Pathology + Imaging + Genomics, we only have 32 patients available, which is insufficient to do a cross-validation.

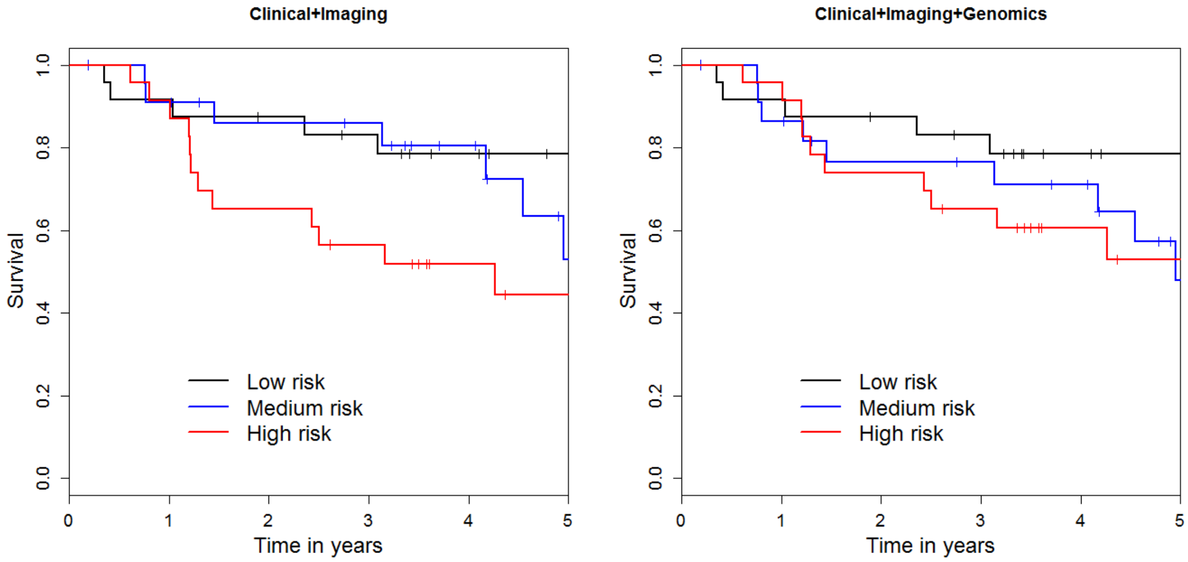

Figure 6: Kaplan-Meier curves for models with imaging data

4.3 Lymph node metastasis models

The LNM models are only based on genomics data. For training the model we have two options, (1) either we only use the LNM gene set, or (2) we use both the LNM and OS genomics gene sets. There is small difference in AUC between these models (0.663 versus 0.709 with ridge regression), meaning that the OS genes give a small improvement in prediction (Figure 6). With a lasso regression, the iAUC for the 6 and 12 gene models are 0.643 and 0.662, respectively.

The LNM predictions are particular clinically relevant for a subgroup of the patients. Patients that are clinical found to be N0, and that additionally have a small tumor (T1 or T2) could potentially avoid an operation if they are found negative for the LNM prediction, as illustrated in Figure 6.

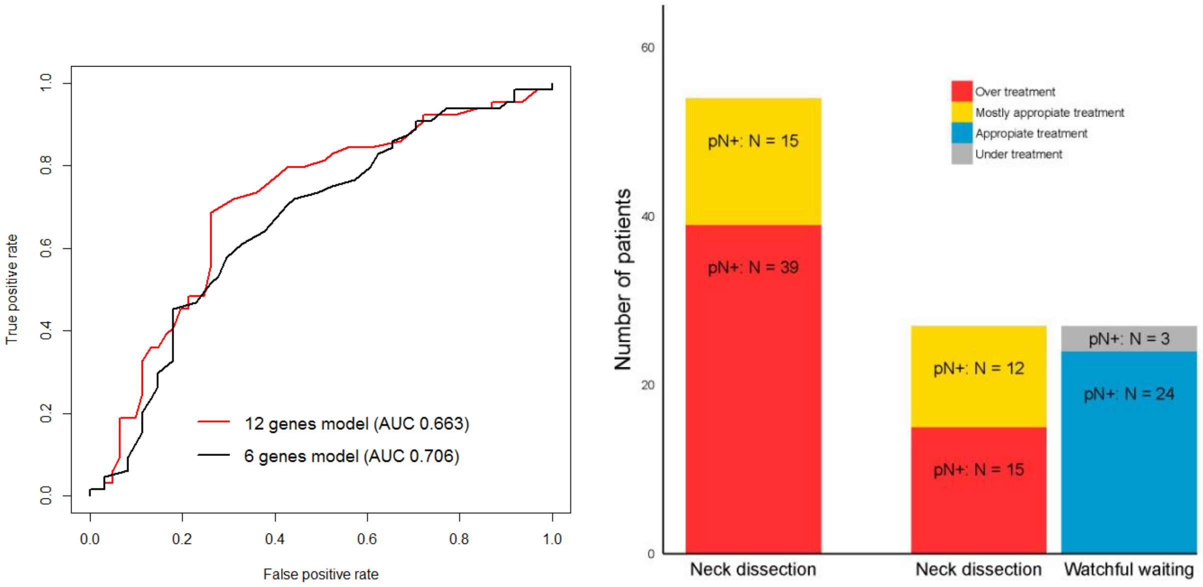

Figure 7: ROC of the LNM genomic models, and a boxplot indicating the number of patients that would receive a more appropriate treatment

4.4 Recurrence

For recurrence we identified a set of pathological variables, and one significant gene (PGM5). We evaluate two models, one based on the pathological variables and one based on the pathological variables with the genomic variable PGM5. We do not include a penalty term for PGM5 although the gene PGM5 was selected on the same data we use for evaluation and the estimated coefficient is therefore not entirely unbiased (e.g. over-optimistic). After excluding 9 patients because of missing values in the pathology data 116 patients remained to evaluate the model, of which only 14 experienced a recurrence. Note that for any outcome variable with such few occurrences, it is difficult to train a robust model, and accurately predict such an outcome.

A model based on the pathological variables had an AUC of 0.678, the model that contained the pathological variables and the genomic variables had an AUC of 0.742. Although these AUCs seem reasonable, one should mind the disbalance between cases and controls. With a standard classification threshold of 0.5 most cases are predicted to have no recurrence due to the low number of occurrences (and the associated low intercept). If we classify every patients with a predicted probability of higher than 0.2 as a recurrence, we obtain results as displayed in Table 2. Although it is clear that probability of recurrence is higher if a recurrence is predicted, the differences are fairly small.

	No recurrence	Recurrence
Absence recurrence predicted	74	8
Recurrence predicted	28	6

Table 2: LOOCV predictions recurrence with the pathology + PGM5 model

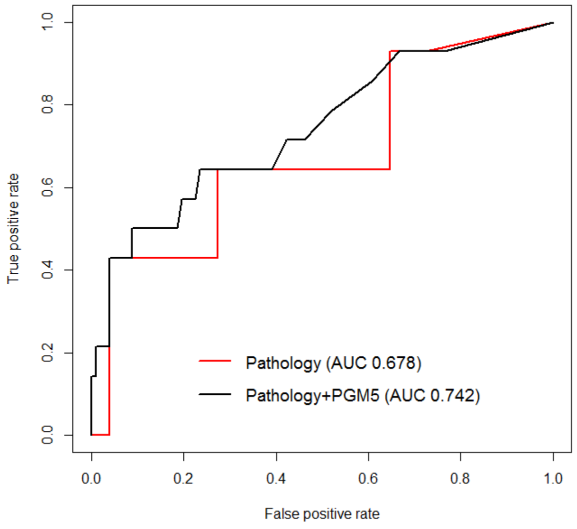

Figure 8: ROC of the recurrence models

5 References

Breiman, L. Random forests. Machine Learning, 45(1): 5–32, 2001

Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York (2001)

Heagerty, P. J., Lumley, T., & Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 337-344.

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat 2008;2(3):841-860.