OraMod

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

TITLE	**D3.2 Prediction simulation tool**		
Deliverable No.	D3.2		
EDITOR	VUmc- M.A. van de Wiel		
Contributors	M.A. van de Wiel, D.E. te Beest, 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	OraMod_ D3.2_Prediction simulation tool.pdf		
FILE ID	OraMod D3.2		
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	21-9-2015	D.E. te Beest, Mark van de Wiel	First draft
2	17-02-2016	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

We distinguish four types of prognostic factors (clinical, pathological, genomic & imaging). We decided to build stepwise models that include an increasing number of factors. These models are static, but the clinician still has the flexibility to assess and compare the predictions from the various models. The consortium decided to not allow for complete flexibility in terms of prioritization of the factors for predictions for two major reasons: 1) It would require retraining and re-evaluating the model for each new setting the clinician enters, which is time-consuming and technically challenging because it requires a link between the clinician's terminal and the model building server; 2) It may facilitate too subjective predictions, making it hard to standardize those between clinicians.

In this document we describe how we selected the variables for the various OraMod predictive models:

Overall Survival (OS):

- Model 1: OS ~ Clinical
- Model 2: OS ~ Clinical + Genomic
- Model 3: OS ~ Clinical + Pathological
- Model 4: OS ~ Clinical + Pathological + Genomic
- Model 5: OS ~ Clinical + Imaging
- Model 6: OS ~ Clinical + Imaging + Genomic
- Model 7: OS ~ Clinical + Pathological + Imaging
- Model 8: OS ~ Clinical + Pathological + Imaging + Genomic.

In addition, we built the following model for Lymph Node Metastasis (LNM).

- Model 9: LNM ~ Genomic

Table of Contents

Executive Summary 3
1 About this document 6
2 Data 7
2.1 Data used 7
3 Variable selection for the predictive models 8
3.1 Variable selection 8
3.2 Clinical data 8
3.3 Pathological data 12
3.4 Tumor imaging data 14
3.5 Lymph node imaging data 17
3.6 Genomic data 19
3.7 Multitype analyses 23
4 Model fitting 26
5 Software 27
6 References 28

List of tables and figures

Table 1. Univariate p-values and coefficients of clinical variables for OS 9
Table 2. Univariate p-values and coefficients of clinical variables for LNM 10
Table 3. Univariate p-values and coefficients of clinical variables for recurrence 11
Table 4. Univariate p-values and coefficients of pathological variables for OS 12
Table 5. Univariate p-values and coefficients of pathological variables for recurrence 13
Table 6. Univariate p-values and coefficients of the tumor imaging data for OS 14
Table 7. Univariate p-values and coefficients of the tumor imaging data for LNM. 15
Table 8. Univariate p-values and coefficients of the tumor imaging data for LNM. 16
Table 9. Univariate results with the lymph node imaging data with OS 17
Table 10. Univariate results with the lymph node imaging data with LNM 18
Table 11. Univariate results with the lymph node imaging data with recurrence 18
Table 12. Summaries on potential housekeeping genes 19
Table 13. Overview of the selected genes 20
Table 14. Results of the stability selection across the set of 102 patients. The genes selected for LNM and OS for OraMod (e.g. the most selected genes) are marked in blue. 21
Table 15. Overview of the clinical, pathological, imaging and genomics variables selected 25

Figure 1: Input and output for OraMod prediction 27
Figure 2: Screenshot predictions with R. Note that this screenshot is only relevant for developers; users will not need to access R directly (only via the OraMod user interface). 27

Abbreviations and definitions

OS	Overall survival
LNM	Lymph node metastasis
OSCC	Oral Squamous Cell Carcinoma
AIC	Akaike Information Criterion
ROC	Receiver Operating Characteristic
LOOCV	Leave-one-out-cross-validation
AUC	Area Under the ROC Curve

1 About this document

This document describes (1) how clinical, pathological, imaging, and genomics variables were selected; (2) how we combined the clinical, pathological, imaging, and genomic data into predictive models; (3) how predictive models are trained for OraMod; (4) how the models will integrate with the OraMod platform.

2 Data

2.1 Data used

Variable selection and model training were conducted with 125 samples from three different hospitals: VUmc, Amsterdam; UNIPR, Parma; and UDUS, Dusseldorf. Note that these samples are independent of those used for the initial gene selection (which was based on an independent set of microarray data). The genomic data used here concerns qPCR data for 60 genes that were preselected from microarray experiments using independent data. For the selection of final gene set we used a subset of 102 of the 125 cases. The remaining 23 cases were not available at the deadline for the gene selection, but they can be used to train the model. The imaging data are available for 70 people, and the lymph node imaging data are available for 32 people. The clinical and genomic data have no missing values. The pathology and imaging data contain some missing values. Samples with missing values are omitted.

3 Variable selection for the predictive models

3.1 Variable selection

We first describe how we selected the variables that are used in the various predictive models for OraMod. OraMod considers four data types: Clinical, Pathological, Imaging and Genomics. We first consider separate predictive models; multitype predictive models are considered further on. Endpoints are overall survival (OS), recurrence and lymph node metastasis (LNM). For OS all analysis are done with cox regression. For LNM and recurrence, the analyses are done with a logistic regression.

For each outcome variable we first screened all variables univariately based on their p-values, and consider whether the estimated coefficients are biologically plausible (higher risk with a higher disease stage). If the coefficients are not biologically plausible, the variable is either omitted or the number of categories is reduced. Variables with p-value lower than 0.15 were considered as candidates for a multivariate model. Next, we used a stepwise regression to identify a multivariate model (with procedure *step* in R). We performed both forward and backward selection and each step a variable can be added or dropped, and then select the best model identified. Stepwise regression aims to balance fit of the model and the number of parameters included, in our case by maximizing the Akaike Information Criterion (AIC). If a variable is included in the multivariate model with coefficients that are not biologically plausible, this variable is then subsequently removed from the set of candidates.

3.2 Clinical data

Overall survival

After univariate screening (Table 1) the following variables were used as input for stepwise selection: PackYears, UnitYears, age, cN, ACE27, Stage, Sex , ECOG, and PriorMal. From these variables, the stepwise regression selects: Unityears, Age, cN, and PriorMal.

For both smoking and drinking, there are four variables available: yes/no/stopped, the number of years, the quantity, and the time multiplied by the quantity (denoted unityears/packyears). For both smoking and drinking unityears and packyears are the most significant variables, and they also give the most complete indication of the amount alcohol/smoking. In the multivariate modelling, we thus only considered unityears and packyears to represent smoking and drinking. We also omit HB, because it contains too many missing variables (29 out of 125).

There is a (known, but also evident in the data) correlation between drinking and smoking: if we would omit unityears, packyears is selected instead. Due to the correlation between unityears and packyears, we additionally included packyears in the clinical model (note that our aim is not to retrieve the most parsimonious model). Including packyears give a small increase in area-under-the-curve (AUC).

PriorMal is selected in the stepwise regression, but will be omitted from the clinical model, because there are only 16 out of 109 patients with a prior malice. This low number, together with its weak effect (see univariate p-value) renders its corresponding parameter difficult to estimate accurately. In fact, removal of PriorMal, renders a small increase in AUC.

Table 1. Univariate p-values and coefficients of clinical variables for OS

	p-value	**Coefficient**	**2nd Coefficient**[1)]
PackYears	4.9E-03	0.02	
Unityears	2.0E-04	0.00	
Age	1.1E-03	0.04	
cN	0.028	0.66	
ACE27	0.019	0.93	1.16
CTNMT (T1/T2,T3,T4)[2)]	0.314	0.03	0.56
CTNMT (T1/T2,T3/T4)	0.386	0.25	
Stage (1/2,3,4)	0.013	0.09	0.97
Stage (1/2,3/4)	0.098	0.50	
Sex	0.100	0.51	
HbDiag	0.156	-0.23	
ECOG	0.136	0.59	
PriorMal	0.142	0.52	
TOBACCO	0.159	0.30	
SigQuanDay	6.5E-03	0.04	
NoSmokYrs	0.039	0.02	
ALCOHOL	0.201	0.39	
AlcUnDay	1.4E-03	0.10	
NoDrinkYrs	0.022	0.02	

[1)] In a univariate regression model variables with three levels have two coefficients. One level is represented by the intercept, two additional levels both have a coefficient.
[2)] Denotes groups used in the analysis

Lymph node metastasis

For LNM the univariate analysis revealed only one variable that could potentially be used in prediction (Table 2), CTNMT (cT-stage I&II versus II&IV). The variable CTNM by itself is not good enough for prediction, but possibly this variable can be used on top of a genomic model. Note that cN and stage are very significant, however these variables code for whether or not a clinical N stage was identified and overlap with the LNM outcome variable. They do not contain information about identifying undiscovered LNM cases, which is our aim. The variables NoDrinkYrs and alcohol are borderline significant, but without a relationship with unityears, it unlikely these are genuine effects that can be used in prediction.

Table 2. Univariate p-values and coefficients of clinical variables for LNM

	p-value	**Coefficient**	**2nd Coefficient[1)]**
PackYears	0.54	4.8E-03	
Unityears	0.50	7.9E-04	
Age	0.34	-0.014	
cN	2.2E-07	3.3	
ACE27	0.13	-0.90	-0.28
CTNMT (T1/T2,T3,T4)[2)]	0.07	0.91	0.75
CTNMT (T1/T2,T3/T4)	0.02	0.84	
Stage (1/2,3,4)	1.1E-06	1.90	2.15
Stage (1/2,3/4)	1.0E-06	2.01	
Sex	0.44	0.28	
HbDiag	0.13	0.21	
ECOG	0.36	-0.51	
PriorMal	0.03	-1.30	
TOBACCO	0.91	0.19	0.07
SigQuanDay	0.21	0.021	
NoSmokYrs	0.95	5.2E-04	
ALCOHOL	0.05	1.06	0.74
AlcUnDay	0.35	4.7E-02	
NoDrinkYrs	0.02	2.0E-02	

[1)] In a univariate regression model variables with three levels have two coefficients. One level is represented by the intercept, two additional levels both have a coefficient.

[2)] Denotes groups used in the analysis

Recurrence

For recurrence we find no variables that could be used in a clinical model (Table 3). Two variables have p-value below 0.10, but closer examination of the coefficients reveals that these coefficients are not biologically plausible (disease stage III has a lower risk than disease stage I/II).

Table 3. Univariate p-values and coefficients of clinical variables for recurrence

	p-value	**Coefficient**	**2nd Coefficient[1)]**
PackYears	0.55	0.01	
Unityears	0.70	0.001	
Age	0.84	0.005	
cN	0.15	0.82	
ACE27	0.64	0.32	-0.32
CTNMT (T1/T2,T3,T4)[2)]	0.07	-0.29	1.34
CTNMT (T1/T2,T3/T4)	0.27	0.63	
Stage (1/2,3,4)	0.02	-1.35	1.08
Stage (1/2,3/4)	0.59	0.32	
Sex	0.27	0.68	
HbDiag	0.34	-0.28	
ECOG	0.56	-0.63	
PriorMal	0.51	-0.71	
TOBACCO	0.42	0.44	1.08
SigQuanDay	0.38	0.02	
NoSmokYrs	0.73	0.01	
ALCOHOL	0.57	0.73	1.06
AlcUnDay	0.59	0.04	
NoDrinkYrs	0.33	0.01	

[1)] In a univariate regression model variables with three levels have two coefficients. One level is represented by the intercept, two additional levels both have a coefficient.
[2)] Denotes groups used in the analysis

3.3 Pathological data

Variable Selection

For the pathological data we followed the same procedure as for the clinical data. We first screened the variables by univariate analysis. Variables with a p-value lower than 0.15 and with estimated coefficients that are biologically plausible are candidates for the stepwise regression. LNM results with pathological data are omitted, as the prediction for LNM is made before the pathological data are available. For 109 cases the pathological data are complete, for the remaining 16 patients some variables are missing.

Overall Survival

After univariate screening (Table 4) nine pathological variables were selected for the multivariate analysis, (VasInfil,BoneInfil,Inflam,Margins,ExCapsSpr,PTNMT,PTNMN,InvaDep,NoMLN). In the stepwise selection, four variables were selected: BoneInfil, Inflam, NoMLN, and Margins. Due to missing values, the multivariate selection was conducted with N = 109. Note that some pathological variables with low p-values (notably ExCapsSpr) were not selected by due strong collinearity with NoMLN.

Although Margins is selected by the stepwise regression, closer examination shows that only 12 out of 109 cases have close margins. This means the effect of close margins is difficult to estimate accurately. As a result, removing Margins from the model does not decrease the AUC of the model. As our aim is to obtain a robust pathological model, we therefore do not include Margins in the pathological model.

Table 4. Univariate p-values and coefficients of pathological variables for OS

	p-value	**Coefficient**	**2nd Coefficient[1)]**	**3rd Coefficient[1)]**
PatInva	0.97	0.01		
VasInfil	0.03	0.67		
PNInfil	0.37	0.27		
BoneInfil	9.4E-04	1.00		
Inflam	8.2E-04	-1.02		
Margins	3.1E-05	1.51		
ExCapsSpr	3.0E-04	1.11		
PTNMT (T1,T2,T3,T4)[2)]	0.02	0.18	-0.24	1.04
PTNMT (T1/T2,T3/T4)	0.08	0.52		
PTNMN (N0,N1,N2)	4.0E-03	-0.37	0.91	
PTNMN (N0,N1/N2)	0.05	0.57		
pStage3C	4.9E-05	-1.61	0.73	
pStage	0.44	0.24		
Diamtr	0.22	0.01		
InvaDep	4.8E-03	0.05		
NoMLN	1.4E-07	0.30		

[1)] In a univariate regression model variables with three levels have two coefficients. One level is represented by the intercept, two additional levels both have a coefficient. Variables with four levels have three coefficients.

[2)] Denotes groups used in the analysis

Recurrence

For recurrence 8 pathological variables where retained after univariate screening (Table 5): PatInva, BoneInfil, Margins, ExCapsSpr, PTNMT, PTNMN, pStage, InvaDep. Stepwise regression selected ExCapsSpr, BoneInfil, and PatInva. Due to missing values, the multivariate selection was conducted with N = 110.

Table 5. Univariate p-values and coefficients of pathological variables for recurrence

	p-value	**Coefficient**	**2nd Coefficient[1)]**	**3rd Coefficient[1)]**
PatInva	0.053	2.05		
VasInfil	0.34	0.57		
PNInfil	0.37	0.52		
BoneInfil	6.0E-03	1.62		
Inflam	0.27	0.69		
Margins	0.14	1.09		
ExCapsSpr	3.0E-03	1.79		
PTNMT (T1,T2,T3,T4)[2)]	0.056	0.02	0.88	1.75
PTNMT (T1/T2,T3/T4)	0.017	1.49		
PTNMN (N0,N1,N2)	0.038	1.66	1.52	
PTNMN (N0,N1/N2)	0.021	1.57		
pStage (1/2,3,4)	0.089	0.54	1.49	
pStage (1/2,3/4)	0.11	1.26		
Diamtr	0.36	0.02		
InvaDep	0.027	0.07		
NoMLN	0.51	0.08		

[1)] In a univariate regression model variables with three levels have two coefficients. One level is represented by the intercept, two additional levels both have a coefficient. Variables with four levels have three coefficients.

[2)] Denotes groups used in the analysis

3.4 Tumor imaging data

Overall Survival

A number of variables have a p-value lower than 0.15 (Table 6) used in univariate screening. We additionally only select variables that occur in at least 8 patients. We additionally omitted ExtrinsicTongueMusculatureInvasion because the estimated coefficient is biologically implausible (due to the small size of the data set this can happen by chance). We furthermore omit Axis3, as it is unlikely that this variable has a genuine effect, without volume also having an effect. The following features where thus retained as candidate variables for the multivariate model: BuccalSpaceInvasion, FloorOfTheMouthInvasion, MasticatorSpaceInvasion, and MaxillarySinusInvasion.

Table 6. Univariate p-values and coefficients of the tumor imaging data for OS

	p-value	**Coefficient**	**2nd Coefficient[1)]**
Axis1	0.71	-4.0E-03	
Axis2	0.32	0.012	
Axis3	0.072	0.019	
BoneInfiltration	0.18	0.27	
BuccalSpaceInvasion	0.081	0.41	
CarotidInvasion			
CutaneousInvasion			
ExtrinsicTongueMusculatureInvasion	0.047	-1.02	
FloorOfTheMouthInvasion	0.069	-0.38	
MasticatorSpaceInvasion	0.018	0.64	
MaxillarySinusInvasion	0.051	0.61	
ParapharingealSpaceInvasion			
PerineuralInfiltration			
PterigoPalatineInvasion	0.22	0.51	
PterygoidInvasion	0.015	0.77	
Side			
SkullBaseInvasion			
SpreadAcrossMidline	0.93	0.022	
TNecrosis	0.76	0.15	
Tstaging (T1/T2,T3/T4)[2)]	0.32	-0.37	
Volume	0.77	3.6E-03	
Nstaging (N0,N1,N2)[2)]g	0.90	0.023	0.23
NumberOfLymphNodes	0.22	0.29	

[1)] In a univariate regression model variables with three levels have two coefficients. One level is represented by the intercept, two additional levels both have a coefficient.
[2)] Denotes groups used in the analysis

Lymph node metastasis

The variables Nstaging, NumberOfLymphNodes are significant (Table 7), however these variables code for whether or not a positive lymph node was identified and overlap with the LNM outcome variable. They do not contain information about identifying undiscovered LNM cases, which is our aim. The variable MaxillarySinusInvasion occurs only in 8 cases, has only a weak effect, and is thus not suitable for prediction (its coefficient indicates that MaxillarySinusInvasion reduces the probability that a LNM occurs). For LNM, no variables were retained after the univariate screening.

Table 7. Univariate p-values and coefficients of the tumor imaging data for LNM.

	p-value	**Coefficient**	**2nd Coefficient[1)]**
Axis1	0.40	0.01	
Axis2	0.28	0.02	
Axis3	0.36	-0.01	
BoneInfiltration	0.37	-0.24	
BuccalSpaceInvasion	0.99	-0.01	
CarotidInvasion			
CutaneousInvasion			
ExtrinsicTongueMusculatureInvasion	0.29	0.40	
FloorOfTheMouthInvasion	0.61	0.13	
MasticatorSpaceInvasion	0.38	0.38	
MaxillarySinusInvasion	0.13	-0.74	
ParapharingealSpaceInvasion			
PerineuralInfiltration			
PterigoPalatineInvasion	0.58	0.39	
PterygoidInvasion	0.69	-0.20	
Side			
SkullBaseInvasion			
SpreadAcrossMidline	0.48	0.23	
TNecrosis	0.99	-0.01	
Tstaging (T1/T2,T3/T4)[2)]	0.94	0.04	
Volume	0.30	0.02	
Nstaging (N0,N1,N2)[2)]g	1.7E-03	2.17	1.25
NumberOfLymphNodes	8.8E-03	1.02	

[1)] In a univariate regression model variables with three levels have two coefficients. One level is represented by the intercept, two additional levels both have a coefficient.
[2)] Denotes groups used in the analysis

Recurrence

For recurrence we did not find imaging variables with p-value lower than 0.15 (Table 8) and no variables were retained after the univariate screening.

Table 8. Univariate p-values and coefficients of the tumor imaging data for LNM.

	p-value	**Coefficient**	**2nd Coefficient[1)]**
Axis1	0.98	-5.4E-04	
Axis2	0.90	2.6E-03	
Axis3	0.97	8.2E-04	
BoneInfiltration	0.44	0.29	
BuccalSpaceInvasion	0.67	0.19	
CarotidInvasion			
CutaneousInvasion			
ExtrinsicTongueMusculatureInvasion	0.68	-0.26	
FloorOfTheMouthInvasion	0.59	-0.21	
MasticatorSpaceInvasion	0.59	-0.41	
MaxillarySinusInvasion	0.39	0.46	
ParapharingealSpaceInvasion			
PerineuralInfiltration			
PterigoPalatineInvasion	0.99	-15.7	
PterygoidInvasion	0.99	-7.9	
Side			
SkullBaseInvasion			
SpreadAcrossMidline	0.61	-0.27	
TNecrosis	0.18	1.1	
Tstaging (T1/T2,T3/T4)[2)]	0.48	0.53	
Volume	0.47	0.02	
Nstaging (N0,N1,N2)[2)]g	0.50	-0.92	0.11
NumberOfLymphNodes	0.91	0.05	

[1)] In a univariate regression model variables with three levels have two coefficients. One level is represented by the intercept, two additional levels both have a coefficient.
[2)] Denotes groups used in the analysis

3.5 Lymph node imaging data

We calculated the p-value with the lymph node imaging data for OS, n-stage and recurrence. The lymph node imaging data is available for only a small subset of the data (n= 32 people). If multiple lymph nodes are available for a particular patient, we selected the largest lymph node (by volume) for analysis. The variables Axis1, Axis2, and Axis3 are analyzed univariately, but in the multivariate analysis we only consider volume (which is derived from the three axes).

Overall Survival

For OS, we retained shapedeviation and volume as candidate variables for the multivariate stepwise regression (Table 9).

Table 9. Univariate results with the lymph node imaging data with OS

	p-value	**Coefficient**
Axis1	0.479	0.019
Axis2	0.094	0.068
Axis3	0.070	0.067
Cluster	0.750	-0.152
ExtraNodalSpreading	0.966	-0.017
Necrosis	0.733	0.108
ShapeDeviation	0.051	-1.191
Volume	0.086	0.074

Lymph node metastasis & recurrence

For N-stage and recurrence we do not find variables with a p-value below 0.15 (Table 10-11) and thus no variables were retained after the univariate screening.

Table 10. Univariate results with the lymph node imaging data with LNM

	p-value	**Coefficient**
Axis1	0.417	0.037
Axis2	0.832	0.013
Axis3	0.613	0.033
Cluster	0.757	0.206
ExtraNodalSpreading	0.948	0.035
Necrosis	0.863	-0.078
ShapeDeviation	0.473	0.559
Volume	0.402	0.099

Table 11. Univariate results with the lymph node imaging data with recurrence

	p-value	**Coefficient**
Axis1	0.95	3.4E-03
Axis2	0.97	0.003
Axis3	0.73	0.027
Cluster	1.00	-16.16
ExtraNodalSpreading	0.85	-0.15
Necrosis	0.87	-0.11
ShapeDeviation	0.61	-0.56
Volume	0.99	-8.6E-04

3.6 Genomic data

More details of the initial genes selection where we select 60 genes using the microarray data are described in deliverable D3.1. Of these 60 genes, 20 genes where selected primarily for on LNM, another 20 where primarily selected for OS, and 20 genes where first pre-selected on OS, and then selected on recurrence. We now describe the further analysis of these data, and how we selected the final gene sets.

Housekeeping genes

Since the genomic data concerns qPCR data we needed to choose a suitable housekeeping gene from four candidates. We did so based on Spearman correlation with the mean expression of the non-housekeeping genes and the variability across samples. Out of 4 candidates to be used as housekeeping gene, we chose GUSB, because it showed a relative small variability across samples and the best correlation with the mean of the 60 genes, as shown in the Table 12.

Table 12. Summaries on potential housekeeping genes

	Variance across all samples
GAPDH	1,89
GUSB	1,15
RPL4	1,28
RPLP0	1,35

	Correlation with the mean of 60 genes
GAPDH	0,76
GUSB	0,82
RPL4	0,75
RPLP0	0,73

Gene selection

Selecting genes on microarray data and testing these genes on independent set of patients on PCR data comes with some complications. Firstly, the dimension we selected from is large (≈ 20.000 genes), there is thus always a possibility that the some selected genes where selected by chance and did not have a genuine effect. Secondly, we do a platform transition so the genes are measured differently which could implicate that some effects are less strong. The second point was partly dealt with in the technical validation. Despite these difficulties, we find that the selected genes validate well. Calculated across all 60 genes, we find that 19 genes have a FDR lower than 0.20. On top of a clinical model, there are for OS 7 genes with a FDR < 0.20. For LNM we find 29 genes with a FDR < 0.20. Further testing with the global test (Goeman et al., 2005) , which renders one p-value for the set of genes as a whole, revealed that the selected genes are significant for both OS and LNM (p-values 0.003 and 0.0002, respectively), and marginally significant for recurrence (p-value 0.09). We also assessed the additive value of the genes with the global test Global Test. For OS, the genomic data on top of the clinical model gives a p-value of 0.06. On top of the clinical and pathological data, the genomic data have a p-value of 0.30. It is thus unlikely that the genes will have additional predictive value on top of a pathological data.

While for other types of variables we mostly perform variable selection for the purpose of improving predictive performance, there is an additional motivation for gene selection: the envisioned, final

platform for measuring gene expression is the ST electronics gene card which contains multiples of six genes. We decided to select 12 genes in total which should have predictive value for both LNM and OS. For both selection of LNM, OS and recurrence, we used all 60 genes to select from. Primary reason is that 60 genes is a relatively small set of genes to select from, there is little difference in selecting from 20 or from 60 genes. Also, preliminary analyses across the 60 genes showed that OS genes could potentially improve prediction of LNM on top of the LNM genes. Selection of genes was based on set of n=102 persons. Note that the data set consists of 125 people, but for 23 people data were not available at the moment the selections where make.

Hence, we take this practical constraint into account. Genes were selected using stability selection (Meinshausen and Bühlmann, 2010). Stability selection involves repeatedly fitting a lasso model to bootstrapped samples of the data set. A lasso is a type of penalized regression technique that sets the coefficients of those genes that contribute least to the predictive value of the model to zero, all remaining genes with a non-zero coefficient are thus selected. Across all bootstrapped models we then count how often each genes was selected, and the six genes with highest selection frequency are selected. To specifically select genes that predict well in a model of six genes, each individual lasso model consists of six genes (selected according to the lasso path). For LNM we selected genes with a genomics-only lasso model. For OS we selected using a genomics-only model and a clinical & genomics model (see description further on). In the latter case, we need to take multi-type models into account, because for the gene cards all models need to be based on the same genes. We did not take into account clinical, genomics & pathology model as the global test indicated that the genomic data do not contain added predictive value on top of such a model (p-value 0.34). The selected genes are displayed in Table 13, and the results of the selection procedure are presented in Table 14.

Table 13. Overview of the selected genes

OS		**LNM**	
Gene	**Assay**	**Gene**	**Assay**
PGM5	Hs00222671_m1	CLEC3B	Hs00162844_m1
VEGFA	Hs00900055_m1	KRT23	Hs00210096_m1
TCL1A	Hs00951350_m1	KLRC1	Hs00970274_m1
CXCL13	Hs00757930_m1	EXPH5	Hs00323579_m1
LRCOL1	Hs01113075_m1	TPM1	Hs00165966_m1
TSPAN11	Hs01391666_m1	FN1	Hs01113075_m1

Table 14. Results of the stability selection across the set of 102 patients. The genes selected for LNM and OS for OraMod (e.g. the most selected genes) are marked in blue.

Genes	LNM	OS	OS (clinical Null)	Genes+Clinical/Genes
ADAM12	0	1	9	10
ADCY4	24	107	6	113
AMPD1	11	29	27	56
AREG	179	14	34	48
ARG1	64	70	106	176
ATP6V0A1	35	3	13	16
C9orf116	19	229	274	503
CALD1	39	7	8	15
CBFA2T3	137	103	154	257
CCDC88B	49	96	14	110
CCND1	136	18	39	57
CDKN2A	14	59	127	186
CLEC3B	667	15	8	23
COL11A1	27	13	10	23
COL4A5	5	66	16	82
COL5A1	6	0	11	11
COL6A1	1	3	10	13
CTTN	21	35	49	84
CXCL13	8	211	297	508
DEFB103	89	66	99	165
EIF5	6	41	25	66
EXPH5	688	165	128	293
FN1	414	78	33	111
IL8	2	105	314	419
INHBB	25	171	26	197
IRX5	2	70	42	112
KIAA1551	66	145	171	316
KLRC1	357	176	111	287
KMT2A	3	11	12	23
KRT23	326	28	140	168
LRCOL1	192	330	261	591

Genes	LNM	OS	OS (clinical Null)	Genes+Clinical/Genes
LUM	12	13	12	25
MPZL2	61	151	26	177
MYBPH	53	8	33	41
NDRG1	24	11	13	24
P4HA1	110	12	52	64
PDHX	14	246	208	454
PGM5	3	778	782	1560
POU2AF1	13	71	100	171
PRAME	81	47	68	115
PTPRB	2	9	57	66
RGS5	168	201	47	248
SCG5	7	11	34	45
SELE	96	107	54	161
SELP	79	206	132	338
SERPINB2	136	16	158	174
SERPINH1	258	112	45	157
SLC5A12	47	79	23	102
SPANXA	109	83	220	303
SPOCK1	72	15	10	25
SPRR2G	228	29	10	39
SYNPO2	2	4	13	17
TANC2	65	28	41	69
TCL1A	136	463	396	859
TIMM8B	7	17	30	47
TNNC1	22	10	31	41
TNXB	24	13	30	43
TPM1	407	67	39	106
TSPAN11	107	362	251	613
VEGFA	9	366	517	883

3.7 Multitype analyses

In the previous section the various data sets were analyzed separately. In the next section we combine the different data sets in multitype models.

Clinical & pathological

For OS we have good clinical and pathological predictor variables. Hence, we developed a combined model. The candidate variables we used are those variables that where retained after the univariate screening. We dropped cN, as this variable is superseded by the pathology data. We also dropped PriorMal and ECOG, as both have only a low number of occurrences. The set of clinical variables then becomes: PackYears, Unityears ,Age, ACE27, Stage, and Sex. The set of candidate pathological variables is: VasInfil, BoneInfil, Inflam, Margins, ExCapsSpr, PTNMT, PTNMN, InvaDep, and NoMLN. A stepwise regression across this set clinical and pathological variables, selects two clinical variables (Age and PackYears), and three pathological variables (NoMLN, Inflam, BoneInfil). Due to the correlation between PackYears and UnitYears (see description clinical model), we also include Unityears in this model.

Clinical & tumor imaging

As the imaging data are only available for subset of the data, we first train a clinical model on all 125 cases. In the next step we then conduct a stepwise regression with the following candidate imaging variables: BuccalSpaceInvasion, FloorOfTheMouthInvasion, MasticatorSpaceInvasion, MaxillarySinusInvasion, using the linear predictor of the clinical model as offset. Hence, we assess the additional value of the imaging-derived variables for those samples for which these are available. The stepwise regression selects the variable MaxillarySinusInvasion.

Clinical, pathological, tumor imaging

For the combined clinical, pathological, and imaging model we follow the same procedure as with the clinical & imaging model. We first fit the clinical & pathology model(with N = 117), and include the linear predictors as offset in the stepwise regression. We again use same set of candidate (BuccalSpaceInvasion, FloorOfTheMouthInvasion, MasticatorSpaceInvasion, MaxillarySinusInvasion). None of these variables were selected by the stepwise regression.

Clinical & Lymph node imaging

For survival we analyzed the possible models with lymph node imaging data in similar fashion as the clinical/tumor imaging model. We first train a clinical model. The offsets of this model is included as offset in the stepwise regression with the lymph node imaging variables. The candidate variables in used the stepwise regression are ShapeDeviation, Volume. Neither of these variables where selected in the stepwise regression.

Clinical, pathological & Lymph node imaging

In this analysis we include the fit of the clinical/pathological as offset, and then conduct a stepwise regression with the lymph node imaging variables. The candidate variables used in the stepwise regression are ShapeDeviation and Volume. Of these variables, volume was selected.

Clinical - lymph node & Tumor imaging

When analyzing the lymph node imaging variables separately, no variables were selected on top of the clinical model. In the analysis were we included the tumor imaging variables on top of the clinical variables, we selected MaxillarySinusInvasion. For the model based on lymph node & tumor imaging data, on top of the clinical data, we thus select MaxillarySinusInvasion.

Clinical, pathological, lymph node & Tumor imaging

In the analysis that considered the tumor imaging data separately, on top of a clinical/pathological model, no variables were selected. When analyzing the lymph node imaging separately on top of a clinical/pathological model, we selected volume. The model with combined lymph node & tumor imaging, on top of the clinical/pathology, thus includes lymph node volume.

Lymph node metastasis & recurrence models for imaging

For LNM we did not find tumor or lymph node imaging variables or that can be used for prediction and pathology is not relevant for prediction (because this prediction is only relevant before surgery). For recurrence we found only found pathology variables. The fitted model based on clinical, pathology, and imaging is thus the same as the model based on pathology alone.

Models with Genomics

The LNM genomics model consists of the six selected genes. The clinical & genomics model consists of the variables of the clinical model and the six selected genes. The clinical, pathology & genomics model consists of the variables selected for the clinical & pathology model, together with the set of 6 genes. By extension, we also combine the clinical, pathology, imaging, and genomics.

Table 15. Overview of the clinical, pathological, imaging and genomics variables selected

Survival models	Clinical				Pathology			Imaging	Genomics	
Clinical	Age	Packyears	Unityears	cN						
Clinical+Genomics	Age	Packyears	Unityears	cN					Survival gene set	
Clinical+Pathology	Age	Packyears	Unityears		Bone infiltration	NoMLN	Inflammation			
Clinical+Pathology+Genomics	Age	Packyears	Unityears		Bone infiltration	NoMLN	Inflammation		Survival gene set	
Clinical+Imaging	Age	Packyears	Unityears	cN				MaxillarySinusInvasion (Tumor)		
Clinical+Imaging+Genomics	Age	Packyears	Unityears	cN				MaxillarySinusInvasion (Tumor)	Survival gene set	
Clinical+Pathology+Imaging	Age	Packyears	Unityears		Bone infiltration	NoMLN	Inflammation	Volume (Lymph node)		
Clinical+Pathology+Imaging+Genomics	Age	Packyears	Unityears		Bone infiltration	NoMLN	Inflammation	Volume (Lymph node)	Survival gene set	
LNM models										
Genomics									LNM gene set	
Genomics2 (optional)									LNM gene set	Survival gene set

4 Model fitting

All models were trained with the complete data set (n=124). For OS, we a standard used cox regression for the models that do not contain genomic data. We use penalized cox regression using the R-package *penalized* (Goeman, 2010) for the model that do contain genomic data. In these models the clinical and pathological variables where left unpenalized and only the genes where penalized. For LNM, we use penalized logistic regression. We have also developed a new algorithm, called GRridge (Van de Wiel et al., 2015). This algorithm can include co-data (auxiliary data) to boost the prediction for a given data set. For the OraMod data, we have used the independent microarray data p-values as co-data. However, the predictions did not improve here with respect to the basic algorithms, most likely because the dimension of the gene space is fairly small (only 60) and the genes were already selected for their associations with the responses. We did acknowledge the OraMod project in the methodological paper (Van de Wiel et al., 2015), because the research questions in that paper were strongly motivated those in the OraMod project. Further (internal) model validation is described in D3.3

5 Software

All models are implemented in the statistical software R. The survival models are the result of a Cox regression. The LNM models are built with a logistic regression. The input of the predictive model consists of a simple csv file containing the patient data. The output will consists of a prediction per patient, an indication of the patient in a low/medium or high risk and how the various variables included in the model contribute to the risk. Note that output of R also consists of a simple csv file (see Figures 5 and 6). This file contains all the information needed by OraMod partner VCI for the relevant visualizations in the OraMod system. It can be called from the DOS prompt within the OraMod system. We have already proposed some visualizations such as those in D3.3. The explicit link with R will also allow to compute patient-specific prediction uncertainties (D3.3).

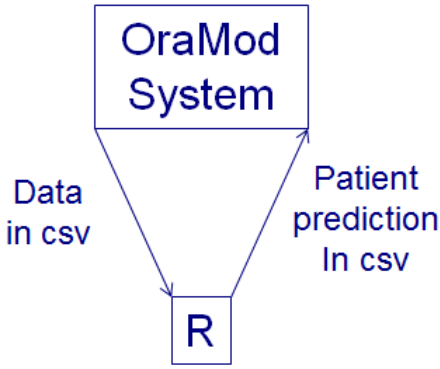

Figure 1: Input and output for OraMod prediction

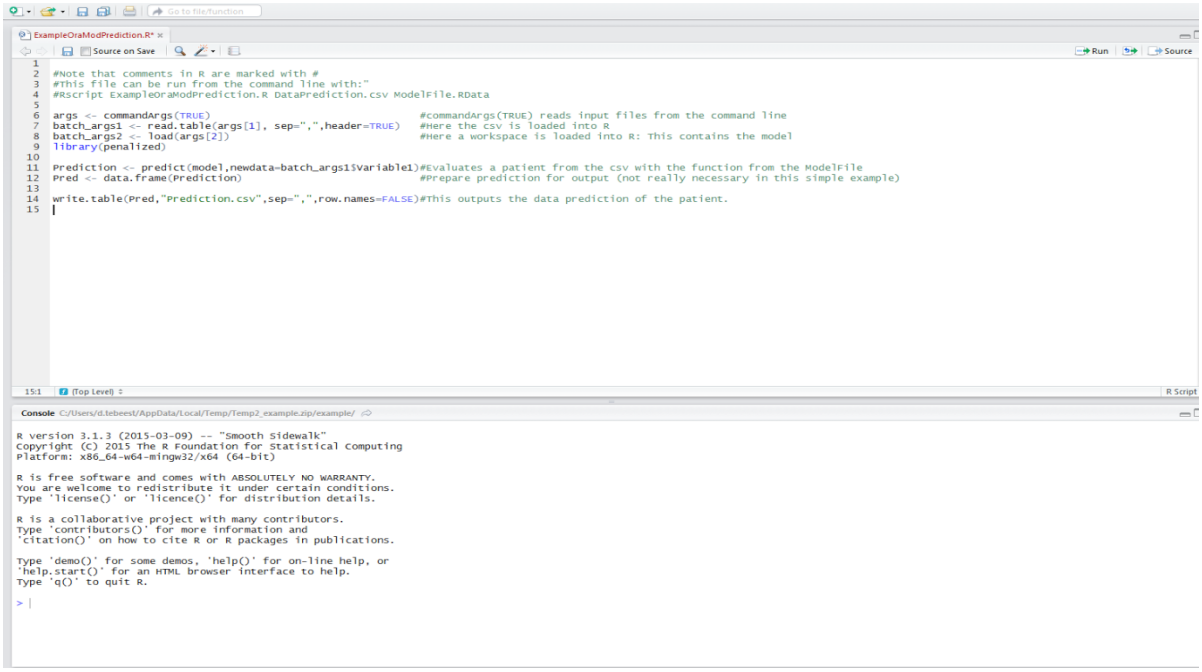

Figure 2: Screenshot predictions with R. Note that this screenshot is only relevant for developers; users will not need to access R directly (only via the OraMod user interface).

6 References

Goeman, J. J., Oosting, J., Cleton-Jansen, A. M., Anninga, J. K., & Van Houwelingen, H. C. (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics*, *21*(9), 1950-1957.

Goeman, J. J. (2010). L1 penalized estimation in the cox proportional hazards model. *Biometrical Journal*, 52(1), 70-84.

Meinshausen, N., and Bühlmann, P.. "Stability Selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72, no. 4 (2010): 417–73. doi:10.1111/j.1467-9868.2010.00740.x.

Van de Wiel, M.A., Lien, T.G., Verlaat,W., van Wieringen, W.N., Wilting, S.M. (2016). Better prediction by use of co-data: Adaptive group-regularized ridge regression. *Statistics in Medicine*, **35**, 368-381