OraMod

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

TITLE	**D3.1 Oral Cancer Predictive Model**		
Deliverable No.	D3.1		
EDITOR	VUmc- M.A. van de Wiel		
Contributors	M.A. van de Wiel,,D.E. te Beest, Steven Mes, R.H.Brakenhof (VUmc) S. Rossi, N. Bertani, R. Perris (UNIPR)		
WorkPackage No.	WP3	**WorkPackage Title**	Oral Cancer Evolution Model
Status[1]	Final	**Version No.**	2
Dissemination level	PU		
DOCUMENT ID	OraMod D3.1 Oral Cancer Model		
FILE ID	OraMod D3.1		
Related documents	DoW version 2013-09-23		

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	10-11-2013	D.E. te Beest, Mark van de Wiel, Steven Mes	First draft
2	12-11-2014	D.E. te Beest, Mark van de Wiel, Steven Mes	Revision and approval by project manager and Coordinator

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

The oral cancer evolution model is the core of the OraMod project and platform. It will be built starting from the original retrospective data from NeoMark (FP7-ICT-VPH-224483 project).

The model building starts from the preliminary assessment of the genomic fingerprint for Oral Squamous Cell Carcinoma (OSCC) reoccurrence, identified by NeoMark using RNA microarrays.

This assessment is performed in a step-by-step process where the genomic fingerprint from NeoMark is validated on a larger retrospective cohort (including VUmc retrospective cases) in order to identify a list of most relevant genes, and to confirm the predictive relevance of NeoMark genomic fingerprint.

As already done in NeoMark, genes from most recent literature have also been considered in the genomic fingerprint analysis and identification. To this aim a precise literature survey has been conducted by UNIPR biomolecular team (see Appendix) and by VUmc.

The so selected genes are then validated by means of RT-PCR in order to define a final list of most relevant genes (genomic fingerprint) to be used for the model building and for the acquisition of genomic data in OraMod system, by means of the lab-on-chip and RT-PCR developed by ST-Italy.

Following the selection of genes, the model building (model training) is performed, by means of biostatistical algorithms, which will consider all the relevant data (clinical, risk factors, imaging, from histology and the selected genes). as described in section 3 of the present document. The model will be developed using the statistical software package R.

In order to allow simulations, different models will be built (stepwise approach), starting from clinical data and gradually adding in different combinations the other data types (imaging, pathology, genomics), in order to allow simulations (see task 3.2 and 3.3):

1. clinical data only
2. clinical data + pathology data
3. clinical data + genomic data
4. clinical data + imaging data
5. clinical data + pathology data + genomic data
6. clinical data + pathology data + imaging data
7. clinical data + pathology data + imaging data + genomic data (i.e. complete oral cancer evolution model).

The model endpoints have also been defined: (1) loco-regional reoccurrence; (2) lymph-nodes metastasis and (3) overall survival.

The so developed model will be integrated into the OraMod platform and will be accessible from the Virtual Patient user interfaces developed in WP5.

Table of Contents

List of tables and figures

Abbreviations and definitions

FDR	False discovery rate
PCR	Polymerase Chain Reaction
OSCC	Oral Squamous Cell Carcinoma

1 About this document

This document describes (1) how the OraMod genes were selected and (2) how the predictive model for OraMod will be built on (1) clinical, (2) pathological, (3) Imaging, and (3) genomics data. The appendix includes the results of the analyses that were conducted to select 60 genes from the 30000+ genes of the micro array data.

2 Selection of the OraMod genes

One of the first aims of the OraMod project is to select 60 genes from 30.000+ probes tested on microarray. These genes will be further tested on PCR on a new set of patients. The selection of 60 genes was based on two data sets: the Neomark data, and the VUmc data. The Neomark data came from the original Neomark project, supplemented with additional patients and longer follow-up times. We included a second data set from the VUmc to ensure that the selected genes are relevant for multiple populations. The VUmc data set contains patients enrolled at VUmc and the Utrecht Medical Centre. In addition, we use information on published gene signatures to increase the robustness of our gene selection.

2.1 Outcome measures used

Primary analyses indicated that a recurrence is difficult to predict for both the NeoMark and VUmc data. We therefore took a broader perspective and considered recurrence, overall survival, and N stage as outcome variables. We spread the selected genes over these three outcome measures, 20 genes were selected for N stage, 20 genes were selected for recurrence, and 20 genes were selected for overall survival. Note that we expect an overlap in predictive ability, because, for example. gene selected for predicting recurrence will likely also partly predict survival.

2.2 Data sets used

The NeoMark microarray data were preprocessed with package Limma in R using loess. Probes with more than 20% missing values were deleted. The NeoMark data consists of a total of 106 samples. Of these samples, seven were not included after checking the quality of the microarrays with MA plots, leaving 99 samples for downstream analyses. For the analysis that used recurrence as the outcome variable, we excluded another five samples, because they either had a metastasis or the medical information was insufficient to identify whether they had a recurrence, another five samples were excluded because the follow-up time was shorter than 24 months. After careful inspection of the medical records, three samples were reclassified from a recurrence to an absence of a recurrence. For the recurrence analysis, in total 89 individuals of which 25 individuals had a recurrence. For the analysis that used N stage as outcome variable, two individuals were excluded, because information on the N stage was missing. Hence, in in total 97 were included of which 49 had an N stage greater than 0. For overall survival analysis no additional exclusions were needed and the data consisted of 99 individuals, of which 31 people did not survive the follow-up

time. Individuals that did not survive the follow up time had a median survival time of 21 months. Individuals that did survived the follow up time were had a median follow up time of 39 months.

The VUMC data were also preprocessed in limma by Wessel van Wieringen using similar quality control criteria . For N stage and survival we, selected cases on (1) oral cavity, (2) negative for HPV, resulting in 150 individuals. Of these, 90 had an N stage greater than 0, 80 individuals did not survive the follow-up time. Individuals that did not survive the follow up time had a median survival time of 20 months. Individuals that did survive the follow up time were had a median follow up time of 83 months. For recurrence we additionally selected on (3) no metastasis, and (4) a follow-up of at least 24 months, resulting of 109 individuals with 21 recurrences.

In the VUMC data, 7 out of 8 HPV positives were oropharynx tumors and only 1 HPV positive was oral cavity. As NeoMark data consists only of samples with oral cavity, the number of recurrence caused by HPV should be low and hence no further samples were excluded from those data.

2.3 Methods used for gene selection

The selection of genes is based on two types of analyses, one univariate, and one multivariate. The univariate approach complements the multivariate one, because multivariate selection methods mostly select genes with orthogonal information. However, if such a gene fails in the PCR validation phase, their may be no 'back-up' gene from the multivariate approach. The genes selected by the univariate apporach may fill this gap. For each outcome measure we selected 10 genes univariately and 10 genes multivariately. For recurrence and N stage the univariate p-values were calculated with t-tests. For survival we used cox regression to calculate p-values. The p-values were then adjusted with Benjamini-Hochberg procedure to calculated false discovery rates (FDR).

For the multivariate selection we use a Lasso (as implemented in package glmnet in R). The Lasso is a regression technique that by L1-penalization sets regression coefficients of genes that do not explain the response variable to zero, thus selecting genes that do explain the response variable. This multivariate analysis thus identifies a group of parameters that contribute on each other in explaining the existing variation in the response variable. The lasso is fitted using randomly chosen cross-validation folds and subset of genes selected can vary between runs. To select the best genes, we run the lasso a number of times and select those genes that were selected by the lasso most often. We do so, because selection by lasso is known to be instable and this procedure helps to stabilize the selection (Meinshausen et al. 2010). For recurrence and N stage we used a logistic regression lasso, and for survival we used a Cox Lasso. For

recurrence and survival, we included disease stage (TNM) as un-penalized covariate, to select the genes in the context of this well-known predictor. We also tested models without disease stage as covariate, but this did not change the list of selected genes.

A complete overview of the results can be found in the accompanying appendix, selected genes are listed in blue. Note that many genes are present in multiple tables, hence often more than 5 or 10 genes were selected per table. Also note that not all genes can be tested on qPCR, so some genes that would be candidates on the basis of predictive performance, were not selected.

2.4 Literature genes and the selection of the genes

For N stage we used the list of Van Hooff et al. (2012), and selected those genes that have a FDR below 0.01 on their data. The 298 probes that remained were used in the univariate and multivariate analyses on the VUMC and NeoMark data. Analyses indicated that both data sets have a strong signal for N stage. Multivariately, we selected those genes that were selected most often by the lasso across both data sets. Univariately, we combined p-values of both data sets using Fisher's combined probability test and calculated FDR's based on the combined p-value.

For survival we combined several gene signatures found in literature (1395 probes, Chen et al. 2008, Chung et al. 2006, De Cecco 2014, Jung et al. 2013, Lohavanichbutr et al. 2013, Onken et al. 2014, Rickman et al. 2008, Thurlow et al. 2010, Winter et al. 2007), and one signature based on integration of the VUmc RNA expression data with matched DNA copy number as performed by Wessel van Wieringen. This "copy number signature" holds the 343 probes for which copy number alterations and the gene expressions are highly correlated. Half the genes were then selected from the combined survival signature (1726 probes), and the other half were selected from the analyses that included all probes (37622 probes for VUmc data, and 40736 for NeoMark data). We thus conducted both the univariate and multivariate analysis twice on each data set. The VUmc data proved to have a stronger signal than the NeoMark data for survival (lower FDR, and more genes selected by the lasso) and the majority of genes for survival were selected on basis of the VUmc data. Possibly, this can be explained by the relatively short median follow-up time of NeoMark cases compared to VUmc cases (39 months versus 83 months, respectively).

In both data sets, we were not able to directly identify genes of enough significance to attempt validation. For this reason, the recurrence genes were selected by using the relationship that exist between overall survival and recurrence. We first preselected genes that correlate with survival, and from this subset we selected genes that correlate with recurrence. We followed the same approach as for survival with respect

to the literature genes: half the genes were selected from the survival signatures, and for other half we considered all genes as candidates. In the analyses where we selected from all genes, we first selected genes with a FDR lower than 0.10 for survival on the VUmc data (217 probes). In those analyses that selected from survival signatures, we first selected those genes with a FDR lower than 0.15 for survival on the VUmc data (77 probes).

3 Training the predictive model

3.1 The predictive model

The 60 selected genes will be validated on a new set independent patients by quantitative real-time PCR. These data will then form the input on which the predictive model will be trained. Next to these 60 genes, the predictive model will also incorporate approximately 30 imaging variables, and a number of clinical and pathological variables. These different types of data are intrinsically very different, for example in the number of parameters they contain, the type of parameters, and the signal to noise ratio, these aspects should be taken into account when building the predictive model. The plan is to use penalized regression techniques, where each type of data has a separate penalty. This ensures that each type of data receives weights according to its contribution to the predictive accuracy of the model. One advantage of a linear model is the relative ease with which such a model can be implemented in the OraMod system. Alternative methods will, however, be considered, e.g. a random forest.

3.2 Testing the additional contributions of the various data sets

Although the different types of data are different, they all measure the same patient and tumor, and there will likely be great number of dependencies between these data sets. It is important to properly test whether the various data sets contribute on top of each other, taking into account these dependencies. More specifically, for a model that contains outcome variable Y and covariates X_1, we are interested in how much a second set of covariate, X_2, contributes on top of X_1 in explaining Y. A naive method would be to permute the rows of X_2. This permutation randomizes X_2 and generates a distribution of the null hypothesis that X_2 does not contribute. In a situation where X_1 and X_2 are independent this would be a valid and effective method. Here that is not the case, and the dependency between X_2 and X_1 must be taken into account.

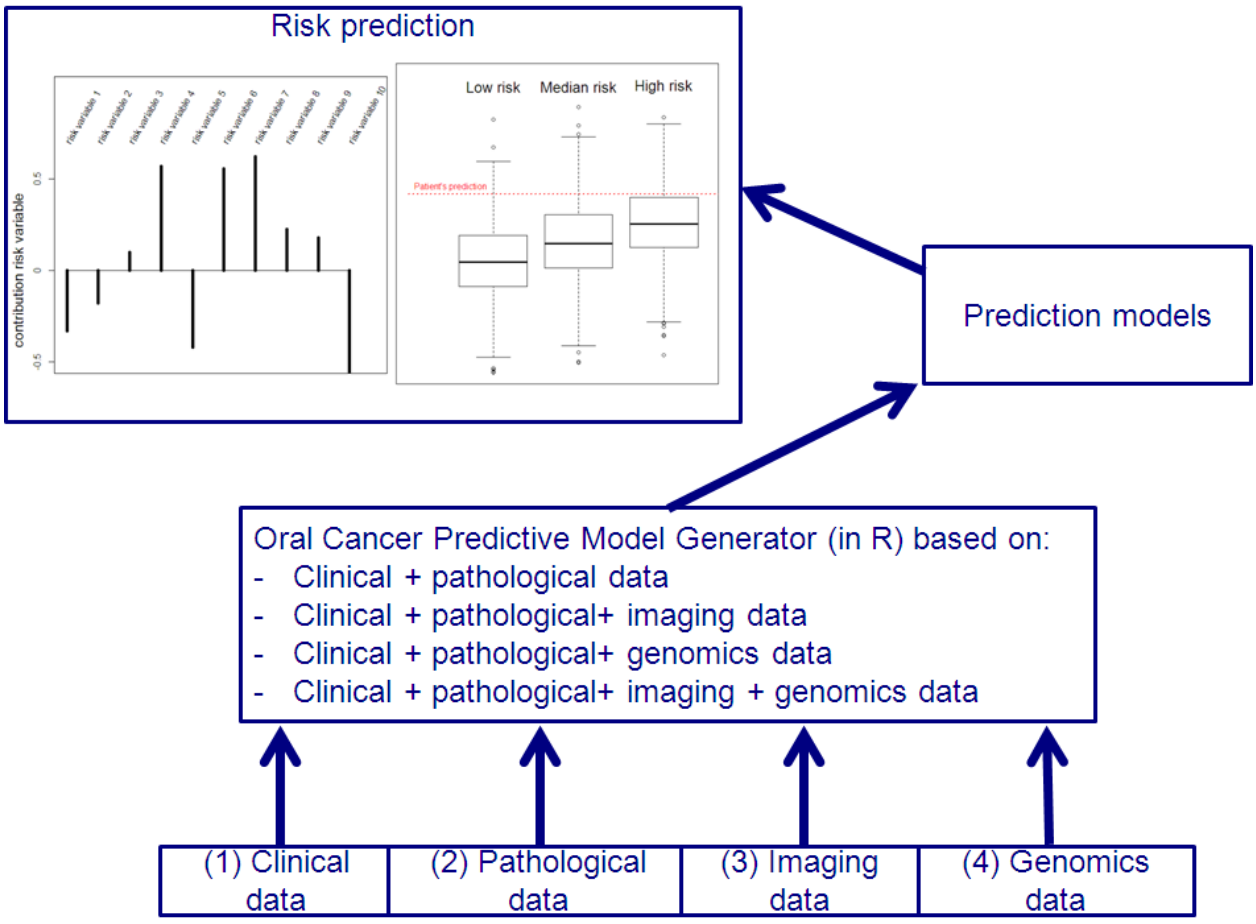
One method we are considering for this problem is the global test (Goeman et al. 2004; Goeman et al. 2005). With the global test it is possible to test how much a given set of covariates (here X_2) contributes to a given null model (here defined on Y given X_1). Currently, the global test only allows a low dimensional x_1, because the null model is fitted with a generalized linear model. In this project we encounter a situation where we have two data sets that have a large number of parameters (p) compared to the number of individuals (n). Although possibly p will be smaller than n, multicollinearity is nevertheless a problem. One methodological aspect we are working on is an extension of the global test that allows for more complex

null models, e.g. a ridge null model. This requires us to generate the distribution for this model under the null hypothesis that X_2 does not contribute to the ability the predict Y, but is conditional on X_1. This can be done by repeatedly drawing $X_2|X_1$ from a multivariate normal distributions. This requires specifying a multivariate normal distribution for X and estimate its covariance matrix. For a high dimensional X the covariance matrix can estimated by ridge shrinkage as implemented in R package rags2ridges (van Wieringen & Peeters, 2014).

4 Embedding the predictive model in the OraMod platform

We will make four different types of predictive models based on different combinations of data sets (See Figure 1). Models will be trained in R, the resulting predictive model will be used to make predictions for new patients. The predictions, including their uncertainty, will be visualized in the OraMod system and placed in to context of the available data.

Fig. 1. Predictive models in the OraMod system



5 References

Chen, Chu, Eduardo Méndez, John Houck, Wenhong Fan, Pawadee Lohavanichbutr, Dave Doody, Bevan Yueh, et al. "Gene Expression Profiling Identifies Genes Predictive of Oral Squamous Cell Carcinoma." *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 17, no. 8 (2008): 2152–62. doi:10.1158/1055-9965.EPI-07-2893.

Chung, Christine H., Joel S. Parker, Kim Ely, Jesse Carter, Yajun Yi, Barbara A. Murphy, K. Kian Ang, et al. "Gene Expression Profiles Identify Epithelial-to-Mesenchymal Transition and Activation of Nuclear Factor-kappaB Signaling as Characteristics of a High-Risk Head and Neck Squamous Cell Carcinoma." *Cancer Research* 66 (2006): 8210–18. doi:10.1158/0008-5472.CAN-06-1213.

De Cecco, L., P. Bossi, L. Locati, S. Canevari, and L. Licitra. "Comprehensive Gene Expression Meta-Analysis of Head and Neck Squamous Cell Carcinoma Microarray Data Defines a Robust Survival Predictor." *Annals of Oncology : Official Journal of the European Society for Medical Oncology / ESMO* 25, no. 8 (2014): 1628–35. doi:10.1093/annonc/mdu173.

Goeman, Jelle J., Jan Oosting, Anne Marie Cleton-Jansen, Jakob K. Anninga, and Hans C. van Houwelingen. "Testing Association of a Pathway with Survival Using Gene Expression Data." *Bioinformatics* 21 (2005): 1950–57. doi:10.1093/bioinformatics/bti267.

Goeman, Jelle J., Sara Van de Geer, Floor De Kort, and Hans C. van Houwellingen. "A Global Test for Groups Fo Genes: Testing Association with a Clinical Outcome." *Bioinformatics* 20 (2004): 93–99. doi:10.1093/bioinformatics/btg382.

Jung, Alain C., Sylvie Job, Sonia Ledrappier, Christine Macabre, Joseph Abecassis, Aurélien de Reyniès, and Bohdan Wasylyk. "A Poor Prognosis Subtype of HNSCC Is Consistently Observed across Methylome, Transcriptome, and miRNome Analysis." *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research* 19, no. 15 (2013): 4174–84. doi:10.1158/1078-0432.CCR-12-3690.

Lohavanichbutr, Pawadee, Eduardo Méndez, F. Christopher Holsinger, Tessa C. Rue, Yuzheng Zhang, John Houck, Melissa P. Upton, et al. "A 13-Gene Signature Prognostic of HPV-Negative OSCC: Discovery and External Validation." *Clinical Cancer Research* 19 (2013): 1197–1203. doi:10.1158/1078-0432.CCR-12-2647.

Meinshausen, Nicolai, and Peter Bühlmann. "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.

Onken, Michael D., Ashley E. Winkler, Krishna-Latha Kanchi, Varun Chalivendra, Jonathan H. Law, Charles G. Rickert, Dorina Kallogjeri, et al. "A Surprising Cross-Species Conservation in the Genomic Landscape of Mouse and Human Oral Cancer Identifies a Transcriptional Signature Predicting Metastatic Disease." *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research* 20, no. 11 (2014): 2873–84. doi:10.1158/1078-0432.CCR-14-0205.

Rickman, D. S., R. Millon, A. De Reynies, E. Thomas, C. Wasylyk, D. Muller, J. Abecassis, and B. Wasylyk. "Prediction of Future Metastasis and Molecular Characterization of Head and Neck Squamous-Cell Carcinoma Based on Transcriptome and Genome Analysis by Microarrays." *Oncogene* 27 (2008): 6607–22. doi:10.1038/onc.2008.251.

Thurlow, Johanna K., Claudia L. Peña Murillo, Keith D. Hunter, Francesca M. Buffa, Shalini Patiar, Guy Betts, Catharine M. L. West, et al. "Spectral Clustering of Microarray Data Elucidates the Roles of Microenvironment Remodeling and Immune Responses in Survival of Head and Neck Squamous Cell Carcinoma." *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology* 28, no. 17 (2010): 2881–88. doi:10.1200/JCO.2009.24.8724.

Van Hooff, Sander R., Frank K. J. Leusink, Paul Roepman, Robert J. Baatenburg de Jong, Ernst-Jan M. Speel, Michiel W. M. van den Brekel, Marie-Louise F. van Velthuysen, et al. "Validation of a Gene Expression Signature for Assessment of Lymph Node Metastasis in Oral Squamous Cell Carcinoma." *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology* 30 (2012): 4104–10. doi:10.1200/JCO.2011.40.4509.

Wieringen, WN van, and C. F. W. Peeters. "Ridge Estimation of Inverse Covariance Matrices from High-Dimensional Data." *arXiv Preprint arXiv:1403.0904* 269553 (2014): 2007–13.

Winter, Stuart C., Francesca M. Buffa, Priyamal Silva, Crispin Miller, Helen R. Valentine, Helen Turley, Ketan A. Shah, et al. "Relation of a Hypoxia Metagene Derived from Head and Neck Cancer to Prognosis of Multiple Cancers." *Cancer Research* 67, no. 7 (2007): 3441–49. doi:10.1158/0008-5472.CAN-06-3322.

Appendix I-III. Tables for the gene selection

These tables will be made available after acceptance of the manuscript that describes the selection and validation of the genes.

Appendix IV Survey of literature for oral cancer predictive genes

This table will be made available after acceptance of the manuscript that describes the selection and validation of the genes.