Genetic classification of non-HPV related head and neck cancer identifies subgroups associated with clinical parameters

Priya A. Kanhai

Vrije Universiteit
Faculty of Sciences
Business Mathematics and Informatics
De Boelelaan 1081a
1081 HV Amsterdam

Organisation:
VUmc Cancer Center
Postbus 7057
1007 MB Amsterdam

08 September 2008

Preface

The BMI paper is written at the end of all the courses done during the Master and before the start of the internship. The purpose of this paper is to do research on a topic that is BMI related in a proper academic way. This research should contain at least two of the following three aspects: business, mathematics, and computer science.

This paper provides the results of statistical testing done on Head and Neck squamous cell carcinomas (HNSCC). During this project the main question was to find out whether there exists a relationship between one or more clinical variables and subgroups of HNSCC. Different statistical methods / tests are taken into consideration.

I would like to thank my supervisor, Mark van de Wiel and also Serge Smeets for all the guidance and support during my research on this topic.

Contents

1 Introduction

In Chapter 1 an introduction on Head and Neck squamous cell carcinoma, microarrays, and WECCA is given. Chapter 2 discusses the genomic and clinical data that was used. Chapter 3 gives an explanation of the method that was used. In Chapter 4 the results are given. The last Chapter provides a conclusion.

1.1 Introduction HNSCC and genetic damage

Head and Neck squamous cell carcinoma (HNSCC) is 5th among the most common cancers in the Western World. There are about 780.000 persons worldwide who suffer from this disease per year **[5]**. HNSCC develops in the mucosal linings of the upper respiratory and digestive tract.

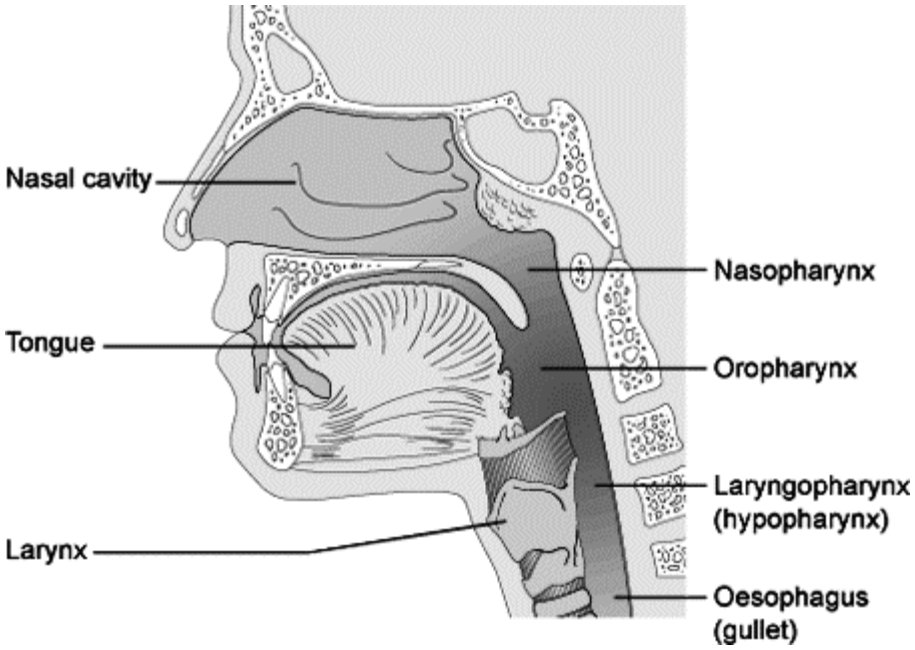

Figure 1: The upper aerodigestive tract

Despite significant advances in local control, long-term survival of cancer patients has only moderately improved during the last 20 years. It stayed around 50% for patients who had an advanced stage of HNSCC **[6]**.The identification of the cancer genes causally involved in carcinogenesis will be essential to make headway in detecting this malignancy at an earlier stage and developing novel therapies.
It is widely accepted that an accumulation of genetic and epigenetic alterations in oncogenes[1] and tumor suppressor genes[2] forms the basis for the progression of a normal cell to a cancer cell, referred to as multi-step carcinogenesis. It is assumed that 4 to 6 genetic hits are necessary to generate a malignant phenotype **[12]**. HNSCC cells often display extensive chromosomal changes, including high level amplifications, and large deletions and gains, as well as translocations.

[1] Definition: Oncogenes are mutated genes that are resident in cellular chromosomes **[20]**
[2] Definition: Tumor suppressor genes are normal genes that slow down cell division, repair DNA mistakes, and tell cells when to die (a process known as apoptosis or programmed cell death) **[21]**. Uncontrolled cell growth happens when these genes don't work correctly, which leads to cancer **[21]**.

HNSCC arises by a chemical etiology[1] encompassing well-established causative life style related agents, like tobacco smoking and alcohol abuse. Recently, the role of the human papillomavirus (HPV) in head and neck carcinogenesis as a separate etiologic factor has been firmly established **[13]**. Knowing that HPV infected tumors are genetically distinct from the other HNSCC **[19]** has shown that HNSCC is a genetically heterogeneous disease. To test this hypothesis we compared CGH-profiles of a set of unselected HNSCC without HPV involvement. Sophisticated methods for analysis of the ordinal data have only recently become available and made it possible to distinguish subgroups **[2]**.

1.2 Introduction microarrays and WECCA

Microarray comparative genomic hybridization (maCGH) is a method that measures chromosomal alterations, gains and losses on a chromosome. This is done by scanning DNA spots on a microarray and sending information concerning their intensity to a computer for analysis. A more elaborate description is given in Figure 2.

[1] Definition: The study of the causes. For example, study of a disorder **[22]**.

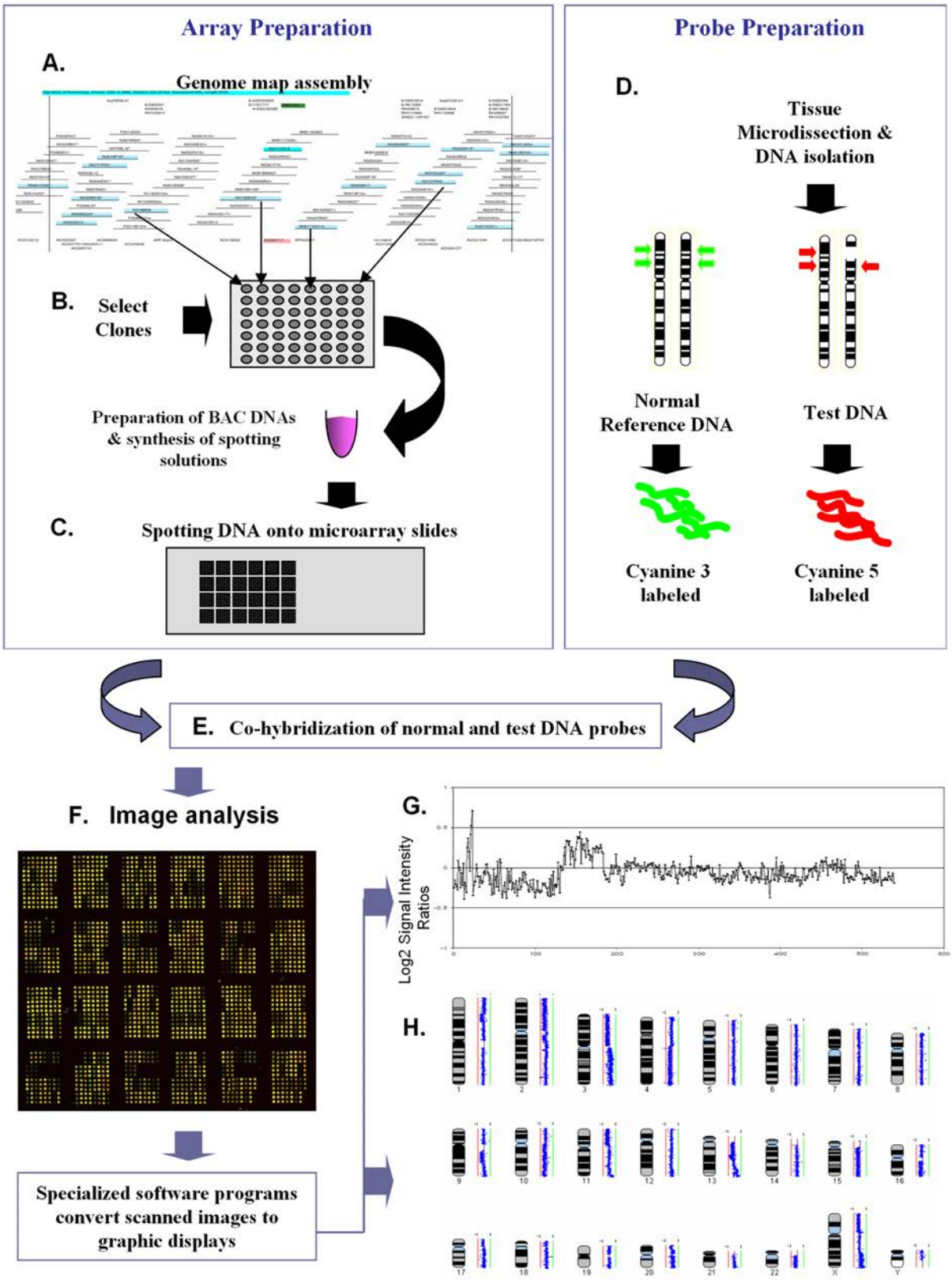

Figure 2: Principle of array CGH

Figure 2 shows the steps in bacterial artificial chromosome (BAC) array CGH.

(A) BAC clones are selected from a physical map of the genome. (B) DNA samples are extracted from selected BAC clones and their identity is confirmed by DNA fingerprinting or sequence analysis. (C) A multi-step amplification process generates sufficient material from each clone for array spotting. Each clone is spotted in replicate onto a solid support. (D) Reference DNA and test DNA are differentially labelled with cyanine 3 and cyanine 5 respectively. (E) The two labelled products are combined and hybridized onto the spotted slide. (F) Images from hybridized slides are obtained by scanning in two channels. Signal intensity ratios from individual spots can be displayed as a simple plot (G) or by using more complex software such as Imagene that can display copy number alterations throughout the whole genome.

"Weighted clustering of called aCGH data (WECCA) is a method for weighted clustering of the ordinal array comparative genomic hybridization (aCGH) data"**[2]**. Two types of similarity measures are introduced in **[2]**, i.e. agreement and concordance. The definition of these similarity measures as provided in **[2]** is as follows:
Agreement:
"A clone of two samples agrees if they are identical".
Concordance:
"Two samples are concordant if their DNA copy numbers of clone A are larger than that of clone B, or both DNA copy numbers of clone A are smaller than that of clone B, or clones A and B have the same DNA copy numbers for both samples."
Clustering can be done per clone or per region (for example, one clone with a small amplification or a complete chromosome arm). In the first case the weights are specified per clone, where in the second case per region.
In **[2]** a new type of linkage is introduced, 'Total linkage'. It is argued that this type of linkage is best suited for ordinal data.

The steps of WECCA are given as follows in **[2]**:

1. Assign weights to each clone, or construct the regions and assign weights to these regions
2. Form the initial clusters, each containing one sample
3. Calculate the similarity between all cluster pairs
4. Merge the two clusters with the highest similarity
5. Iterate between step 3 and 4 until one final cluster remains

This type of clustering can be seen as agglomerative hierarchical clustering.

1.3 Aim of the study

The aim of this study is to evaluate if clinical variables as, gender, smoking, alcohol, lymph node status, age, etc. are relates to the subgroups of HNSCC. This evaluation was done using statistical methods. These methods are available in R **[7]**. R is a free software and can be used for statistical computations **[7]**.

2 Data

2.1 Genomic data

To explore whether we could identify subgroups of tumors with a certain level of similarity with respect to genetic alterations, we analyzed the maCGH data of 39 HNSCC. Unsupervised clustering with WECCA **[2]** was performed and enabled the discovery of three distinct groups. The genomic data consisted of three categories: 'loss', 'normal' and 'gain'. The clustering produced a heat map (see Figure 3).

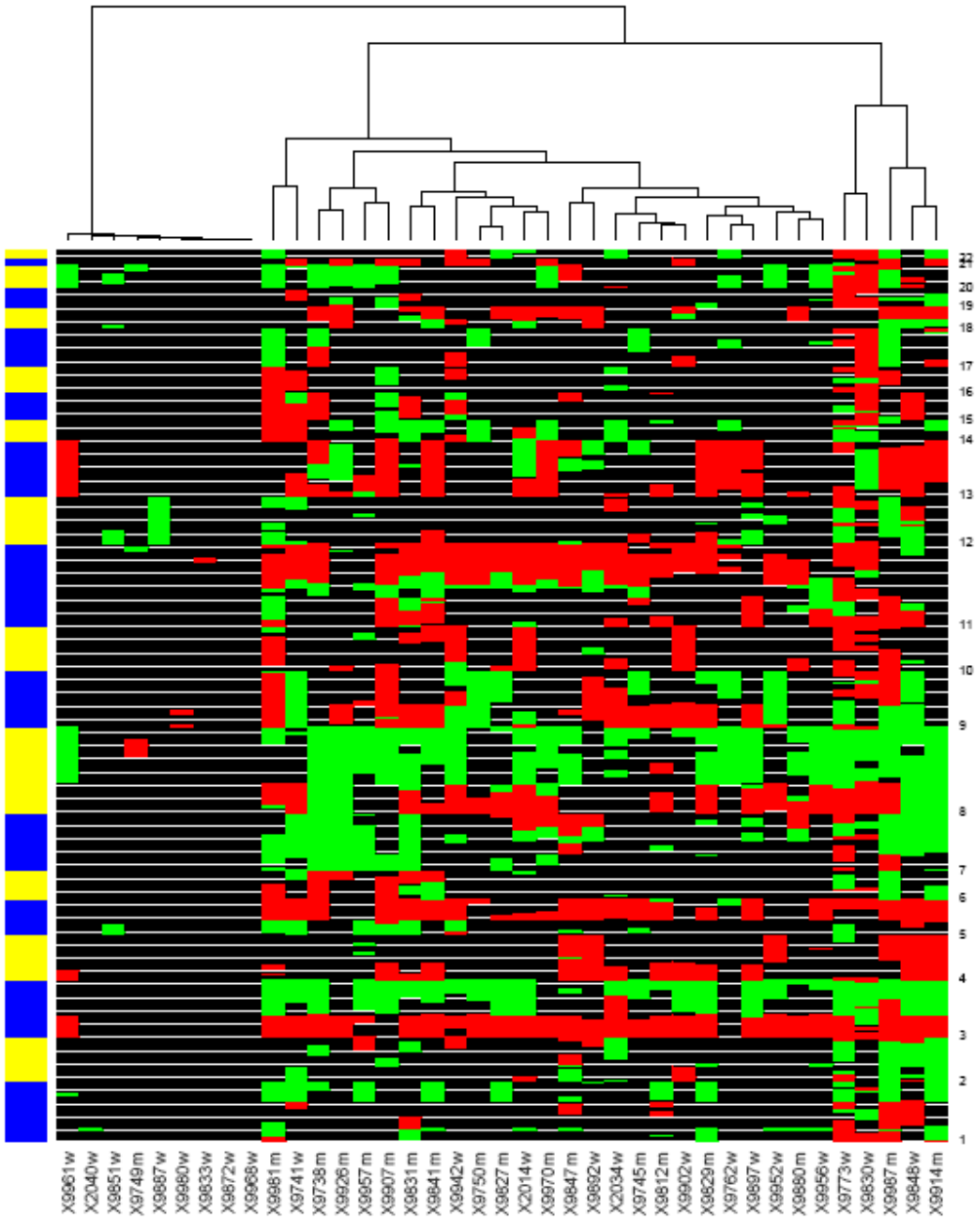

Figure 3: Heat map of genomic data

On the east-side the chromosome group number is given. All tumor DNAs are hybridized to DNA of normal individuals of the opposite gender and therefore the sex-chromosomes are not included in the analysis. The north-side of Figure 3 provides the clustering and the south-side of the picture the patient's code. The colours red, green and black stand for gain, loss and normal respectively. Based on the clustering one can make a cut-off at a level (see Figure 4).

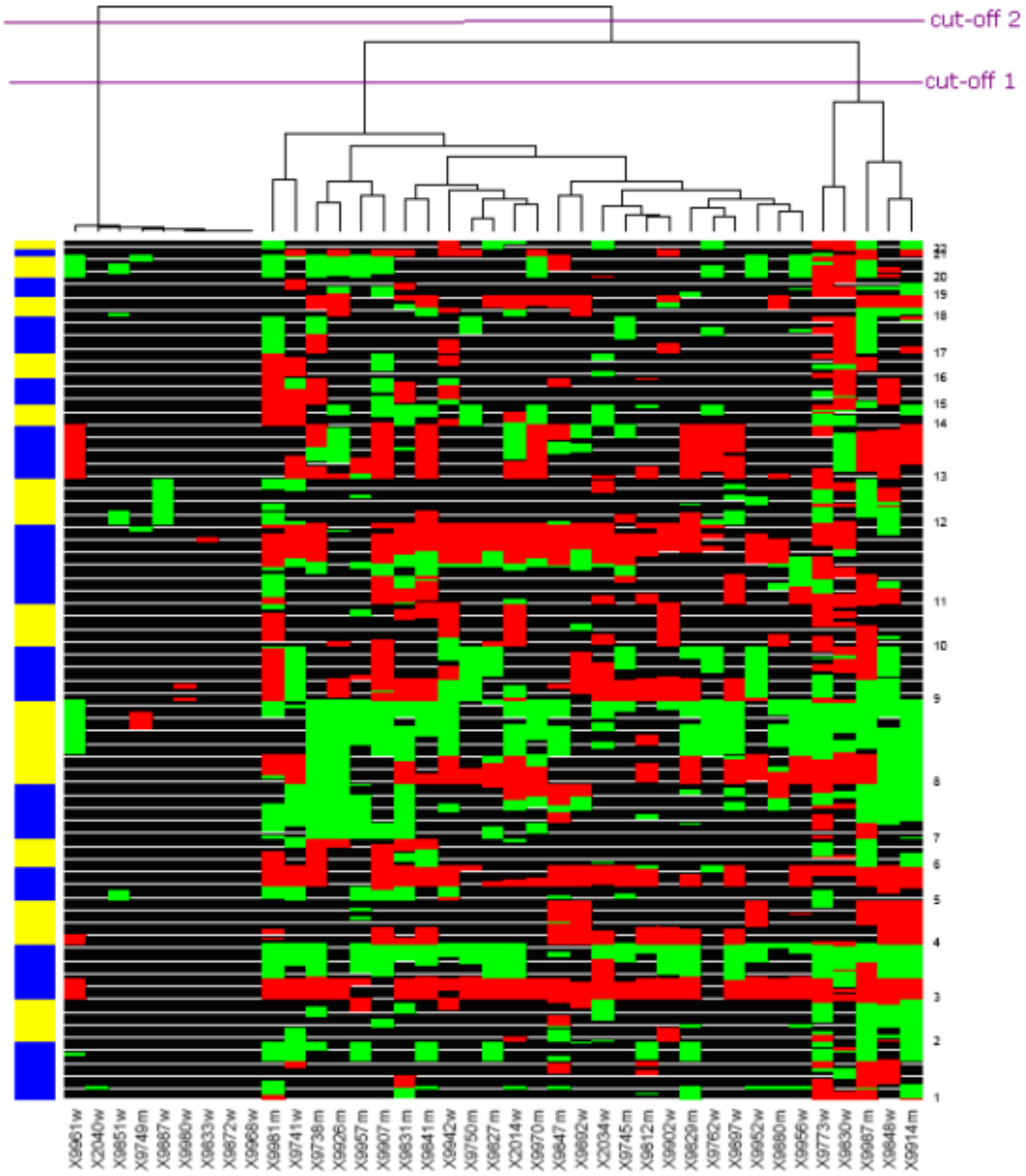

Figure 4: Heat map of genomic data with cut-offs

Three clusters are formed when 'cut-off 1' is applied and two clusters are formed in case 'cut-off 2' is applied. Both cut-offs are considered for this study.

We get the following data when 'cut-off 1' is applied:
Reading from left to right: patient number X9961W until patient number X9968W belong to cluster 2 (9 patients), patient number X9981M until patient number

X9956W belong to cluster 1 (25 patients), patient number X9773W until patient number X9914M belong to cluster 2 (5 patients)
The following data is obtained when ‘cut-off 2’ is applied:
Reading from left to right: patient number X9961W until patient number X9968W belong to cluster 2 (9 patients), patient number X9981M until patient number X9914M belong to cluster 1 (30 patients)

2.2 Clinical data

This Section discusses which kind of clinical variables are used and gives a description of these variables.
The data (see Table 1) which was provided by the department of Otolaryngology/Head-Neck Surgery of the VU Medical Center and consisted of 13 clinical variables s: ‘Mut1’, ‘Age’, ‘PTNMT’, ‘PTNMN’, ‘Smoking’,’Packyears’, ‘Alcohol’, ‘Unitsyear’, ‘Poetacode’, ‘Gender’, ‘Codeloctum’, ‘Stage’, and ‘Rec1’.

Pat_Code	Mut1	Gender	Age	Codeloctur	PTNMT	PTNMN	Stage	Smoking	Packyears	Alcohol	Unitsyear	Rec1	Poetacode
9872	0	F	69	OC	1	0	II	F	84	N	0	2P	ND
9833	0	F	50	OC	2	0	II	C	33	C	320	2P	ND
9887	0	F	43	OC	4	0	IVA	C	5	N	0	DF	ND
9968	0	F	71	OC	1	0	I	C	4	N	0	DF	ND
9961	0	F	73	OC	2	1	III	C	55	N	0	DF	ND
9851	0	M	44	OC	1	0	I	N	0	N	0	DF	ND
2040	0	M	55	OC	1	0	I	NA	NA	NA	NA	DF	ND
9980	0	F	81	OC	1	0	I	N	0	N	0	DF	ND
9897	0	M	72	OC	4	0	IVA	C	57	C	150	DF	ND
9952	0	F	61	OC	3	2B	IVA	C	21	C	NA	REG	ND
9892	0	M	67	OC	2	0	II	F	34	C	14	DF	ND
9902	0	F	53	OC	2	0	II	C	35	C	140	DF	ND
9848	0	M	59	OP	3	0	III	C	40	C	120	LOC	ND
9749	1	F	56	OP	3	1	III	C	80	C	405	LOC	ND
9812	1	F	65	OC	4	2B	IVA	F	8	N	0	DF	ND
9880	1	F	71	OC	NA	NA	NA	C	40	C	90	REG	D
9745	1	M	52	OP	3	0	III	C	34	C	170	DM	ND
9827	1	M	57	OP	1	0	I	C	40	C	117	DF	D
9750	1	M	65	OC	3	2B	IVA	C	47	C	94	DF	ND
9762	0	F	76	OC	1	0	I	N	0	N	0	LOC	ND
9956	0	M	38	OC	2	0	II	N	0	N	0	DF	ND
9907	1	M	46	OC	2	2B	IVA	C	21	C	140	LOC	D
9957	1	M	51	OC	4	0	IVA	C	33	C	450	DF	D
9926	1	M	60	OP	2	0	II	C	60	C	240	DF	D
9738	1	M	60	OC	2	1	III	C	20	C	300	DM	D
9914	1	M	78	OP	3	0	III	F	45	C	180	DF	D
9987	1	M	79	OC	1	0	I	C	61	C	244	DF	D
2034	0	F	45	OC	2	1	III	NA	NA	NA	NA	DF	ND
2014	0	F	68	OC	3	0	III	C	49	C	98	DF	ND
9741	0	F	53	OP	3	2B	IVA	F	35	F	280	DM	ND
9830	0	M	42	OC	2	2B	IVA	C	27	C	216	LOC	ND
9942	0	F	55	OC	2	2B	IVA	C	36	C	NA	DM	ND
9773	0	M	49	OC	2	2B	IVA	C	23	C	140	DF	ND
9829	1	F	74	OC	1	0	I	N	0	N	0	DF	D
9831	1	M	55	OC	3	2C	IVA	F	19	C	75	DF	D
9970	1	F	49	OC	1	0	I	N	0	N	0	NA	D
9847	1	M	57	OP	3	0	III	C	37	C	312	DF	D
9841	1	F	55	OP	4	0	IVA	C	65	C	160	DF	D
9981	1	F	55	OP	3	1	III	C	37	C	148	DF	D

Table 1: Clinical data

The description of the clinical variables can be found in the Appendix.

The number of patients was 39. Of these 39 patients there were 6 patients that contained some missing information.

3 Method

The goal is to investigate if there exists a relationship between the clustering and the clinical variables. In order to do this, first a statistical model had to be chosen. The model used for this investigation was the logistic regression model.
In this Chapter we first introduce a Linear Model. After that the Generalized Linear Model is discussed. In Section three the basic concept behind analysis of deviance is introduced. In the fourth Section two different ways to handle data is provided. Section five discusses clinical variables that were converted. The last Section provides research questions that are formulated.

3.1 Linear Models

Linear Models assume that there exists a linear relationship between the variables you give as input and the output you expect.
The definition of a linear model as given in the Lecture Notes in Chapter 1 **[1]**:
A linear model is a statistical model where random observations $Y_1, \ldots, Y_n$, also known as independent random response variables, are described by a linear combination of p+1 unknown parameters $\beta_0, \ldots, \beta_p$ plus unobservable random errors.

$$\Omega: \begin{cases} Y_i = \beta_0 + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + e_i \\ Ee_i = 0 \\ Cov(e_i, e_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases} \end{cases} \qquad \textbf{(3-1)}$$

for $i, j = 1,\ldots,n$ and where the $\{x_{ij}\}$ are known constant coefficients.

In the Lecture Notes in Chapter 3 **[1]**, this linear model is rewritten as follows:

$$\begin{aligned} &(i) \quad Y_i \sim N(\mu_i, \sigma^2) \\ &(ii) \quad \eta_i = x_i^T \beta_i \\ &(iii) \quad \eta_i = g(\mu_i) = \mu_i \end{aligned} \qquad \textbf{(3-2)}$$

for $i = 1,\ldots,n$, where x_i is a vector of explanatory variables, $\beta = (\beta_0,\ldots,\beta_p)^T$ is a vector of p+1 unknown constants, β_0 belongs to the explanatory variable and is called the intercept, and η_i is a vector of predictors. Part (i) is called the random component and part (ii) the systematic component. In the random component the distribution of Y_i is specified. In part (iii) the link function g comes in and gives the relation between part (i) and part (ii).

In general, the vector $\beta = (\beta_0,\ldots,\beta_p)^T$ can be estimated by the least squares estimator $\hat{\beta} = (\hat{\beta}_0,\ldots,\hat{\beta}_p)^T$ which is the value where $S(\beta)$ is minimized.

$$S(\beta) = \sum_{i=1}^{n} (Y_i - EY_i)^2 \qquad \textbf{(3-3)}$$

The least square estimator $\hat{\beta}$ is found by differentiating $S(\beta)$ with respect to, and then setting this derivative to zero. However, the solution $\hat{\beta}$ is not unique. If we want $\hat{\beta}$ to be unique then we need to specify some additional conditions / restrictions.

3.2 Generalized Linear Models

In Linear Models **(3-2**) the link function is equal to the identity function. When using Generalized Linear Models we are allowed to use any other function. The assumption made when using Linear Models is that the response variable is normally distributed. For Generalized Linear Models any other distributions are allowed.

In order to investigate the relationship between the different clusters and the clinical variables we used the binomial distribution and thus therefore the logistic regression model.

In the Lecture Notes in Chapter 3 **[1]**, this model is given as follows:

$$\begin{aligned} &(i) \quad \mathrm{n}_i Y_i \sim Bin(\mathrm{n}_i, \mu_i) \\ &(ii) \quad \eta_i = x_i^T \beta_i \\ &(iii) \quad \eta_i = g(\mu_i) = \log\left[\frac{\mu_i}{(1-\mu_i)}\right] \end{aligned} \qquad \textbf{(3-4)}$$

for $i = 1,...,n$

This model is chosen, because the clinical data consisted of count data and the number of patient was known. However, we cannot apply this data blindly to our model. This is because we have three clusters when applying ‘cut-off 1’and when using the binomial distribution only two clusters, i.e. 0-1 data is allowed. Despite this, we want to use this distribution and therefore we do the following: First take a look at patients who only belong to clusters one and two, then take a look at patients belonging to clusters one and three, and in the end consider patients belonging to only clusters two and three. However, when this was applied, we saw that there were 25 patients assigned to cluster one, 9 to cluster two, and 5 to cluster three. Basically, we could not compare cluster three to cluster one and two, because of the limited number of patients that are assigned. There was no problem in applying the model in **(3-4)** for ‘cut-off 2’, because the patient data was divided in two clusters.

In the end the following was done:

1. Investigated whether there exists a relationship between clinical variables and patients assigned to cluster one and two for ‘cut-off 1’.
2. Investigated whether there exists a relationship between clinical variables and patients assigned to cluster one and two for ‘cut-off 2’.

3.3 Analysis of Deviance

If we want to know which (explanatory) variable to include in the final model, then we need to perform an analysis of deviance. The analysis of deviance is a goodness-of-fit measure i.e. it measures the quality of the model. The deviance in Generalized Linear Models is comparable to the residual sum of squares in Linear Models.

The deviance statistic is defined as:

$$D = 2\phi(l(\tilde{\theta}) - l(\hat{\theta}))$$

where

- ϕ is the dispersion parameter. When applying the Generalized Linear Models overdispersion can occur when the observed variance of the data is larger than the predicted variance. The dispersion parameter, ϕ, is introduced to the model to lower this overdispersion effect.

- $l(\tilde{\theta})$ is the maximum likelihood estimate in the saturated model (the full / big model) and $l(\hat{\theta})$ is the log-likelihood estimate using the current model (the reduced / small model). The smaller the deviance, the closer the fitted model is to the saturated model. If the deviance is large, the fit is poor. The difference in deviance between the saturated model and the small model has a χ^2 distribution with difference in number of parameters as degrees of freedom if the smaller model suffices. Thus, if the current model is adequate then a comparison with the saturated model can be done using the Chi-squared statistic. The Pearson Chi-squared statistic is defined as:

$$\chi^2 = \phi \sum_{i=1}^{n} \frac{(Y_i - E_\beta Y_i)^2}{Var_\beta Y_i}$$

The Chi-squared statistic will give the difference in deviances, but since the full model always has zero deviance this is just the residual deviance. Hence for models we can test the residual deviance as a Chi-squared statistic to check whether the current model is adequate.

3.4 Two approaches of handling data

There are two approaches available to investigate the relationship between the response variable and the explanatory variables: the univariate approach and the multivariate approach.

The univariate approach looks at one variable at the time, and the multivariate approach considers two or more variables together at the time. Both approaches are applied to the clinical data.

For the univariate approach the ‘anova’ function will be used. If more than one glm object is specified in this function, the table has a row for the residual degrees of freedom and deviance for each model. For all but the first model, the change in

degrees of freedom and deviance is also given. It is conventional to list the models from smallest to largest, because we are testing the relevance of inclusion of the new variable.

For the multivariate approach the 'step' function in R is used. This 'step' function selects the final model based on the Akaike Information Criterion (AIC).
$AIC(k) = -2\log likelihood + 2k$
where k is the number of parameters that is minimized with respect to k.
This criterion gives a penalty if too many parameters are used. The 'step' function starts with an full model and adds or drops variables based on the AIC. The 'step' function in R, can select variables in three directions "forward", "backward", and "both". If one uses the "forward" direction, the 'step' function will only add variables one by one. If one uses the "backward" direction, the 'step' function will only drop variables one by one. However, if one uses "both" as direction, then the 'step' function will drop and add variables. The inclusion of the next variable depends on the presence of other variables that are already in the model. Suppose you have one model with 5 independent variables that are not correlated and another model with 4 dependent variables that are correlated. If both have the same Mean Square Error then the AIC would pick the second model as the 'best' model and not the first one. However, using simple logic, we would say that the first model is the 'best'. The AIC used in the 'step' function doesn't take into account correlated variables.

For the multivariate analysis all the patients who had at least one missing input for a variable were removed and for the univariate analysis only the patients that had for that specific variable a missing input were deleted.

3.5 Converted clinical variables

There were dummy variables introduced for the following clinical variables: 'Smoking', 'Alcohol', 'PTNMN'. Dummy variables are introduced when for instance the numbers 1,2 and 3 do not represent numeric-but categorical-variables. Example: 1-> Good, 2->Bad, 3->None. In this case "Bad" is not twice as big as "Good" thus "2" is not twice as big as "1".
The variables 'Gender', 'Poetacode' and 'Codeloctum' are also converted to '0'and '1'.
The variable 'Stage' is converted from 'I', 'II', 'III', 'IVA' to 1, 2, 3, and 4. No dummy variables were introduced for this variable, because 'II' is considered twice as bad as 'I'.

3.6 Research questions

There was one dataset with all the clinical information. From this dataset two types were made based on the cut-offs. The first type contained patients that were clustered using 'cut-off 1' and the second type contained patients that were clustered using 'cut-off 2'.

Using both types the following questions were investigated:
Which clinical variables are related to the subgroups identifying HNSCC using the univariate approach?

An analysis of deviance was performed given a model with one clinical variable and the empty model (with no clinical variables). From the results, the clinical variables that had a p-value of smaller than 0.15 were selected and added into a model. A higher level can be used than 0.05 or 0.10 **[19]**.
The hypothesis that can be considered was:

H_0: The saturated model is not better than the reduced model

H_1: The saturated model is better than the reduced model

Using formula **(3-4)** this means:

$H_0 : \beta = 0$

$H_1 : \beta \neq 0$

The null-hypothesis is rejected if the p-value is smaller than 0.15.

In the end, the final model with selected clinical variables (p-value < 0.15) was achieved by repeating the following procedure:

1. Start with an empty model
2. Add the clinical variable which contained the smallest p-value in the univariate analysis and has not already been added to the model
3. Perform analysis of deviance
4. If the model becomes better (p-value < 0.15) go to Step 2. If the model does not become better than drop the last selected clinical variable and go to Step 2.
5. Repeat Step 4 until all the selected clinical variables has been considered.

There was an additional question:
Which clinical variables are related to the subgroups identifying HNSCC using the mulitvariate approach?

Based on the results, an analysis of deviance was performed similar to the one with the univariate approach (Step 1 until 5) to select the final model.

4 Results

This Chapter provides the results of the univariate and multivariate approaches on dataset type 1 and dataset type 2. First the results of the univariate approach are given using both dataset types and then that the results of the multivariate approach is provided. In the last Section a summary of the results is given.

4.1 Results univariate approach dataset type 1

In Table 2 the p-values are given for performing an analysis of deviance on an empty model and a model with one clinical variable on dataset type 1.

Clinical variable	p-value	significant
'Mut1'	0.008	*
'Age'	0.654	
'PTNMT'	0.063	*
'PTNMN'	0.157	
'Smoking'	0.863	
'Packyears'	0.811	
'Alcohol'	0.020	*
'Unitsyear'	0.395	
'Poetacode'	0.002	*
'Gender'	0.166	
'Codeloctum'	0.279	
'Stage'	0.236	
'Rec1'	0.059	*

Table 2: p-values of the univariate analysis on dataset type 1

From Table 2 we can see that the clinical variables 'Mut1', 'PTNMT' , 'Alcohol', 'Poetacode', and 'Rec1' have a p-value smaller than 0.15.

In order to select the final model we will repeat the procedure described in Section 3.6. The clinical variables ''Mut1', 'PTNMT' , 'Alcohol', 'Poetacode', and 'Rec1' will be added in the following order 'Poetacode', 'Mut1', 'Alcohol', 'Rec1', 'PTNMT' (from smallest p-value to largest p-value, where all p-values are smaller than 0.15). Also, for the selection of the final model, we will use the dataset with removing all patients who had at least one missing input.

First the clinical variable 'Poetacode' will be added to the empty model.

```
Model 1: cluster ~ 1
Model 2: cluster ~ Poetacode_c
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        27     33.503
2        26     24.731  1    8.772     0.003
```

Table 3: Analysis of Deviance on dataset type 1, model with Poetacode versus empty model

From Table 3 we see that we can reject the null-hypothesis under the significance level of 0.15. Therefore, we may assume that the model with 'Poetacode' is better than the empty model.
Now we add to the model with 'Poetacode' the clinical variable 'Mut1' and perform an analysis of deviance on this model and the model with only 'Poetacode'.

```
Model 1: cluster ~ Poetacode_c
Model 2: cluster ~ Mut1 + Poetacode_c
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        26    24.7306
2        25    23.9068  1   0.8238    0.3641
```

Table 4: Analysis of Deviance on dataset type 1, model with Poetacode versus model with Poetacode and Mut1

From Table 4 we see that we cannot reject the null-hypothesis under the significance level of 0.15. Therefore, we may assume that the model with 'Poetacode' and 'Mut1' is not better than the model with only 'Poetacode'.
Now we will drop the clinical variable 'Mut1' and add to the model with 'Poetacode' the clinical variable 'Alcohol' and perform an analysis of deviance on this model and the model with only 'Poetacode'.

```
Model 1: cluster ~ Poetacode_c
Model 2: cluster ~ Poetacode_c + Alcohol_c1 + Alcohol_c2
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        26    24.7306
2        24    20.4546  2   4.2760    0.1179
```

Table 5: Analysis of Deviance on dataset type 1, model with Poetacode versus model with Poetacode and Alcohol

From Table 5 we see that we can reject the null-hypothesis under the significance level of 0.15. Therefore, we may assume that the model with 'Poetacode' and 'Alcohol' is better than the model with only 'Poetacode'.
Now we will add to the model with 'Poetacode' and 'Alcohol' the clinical variable 'Rec1' and perform an analysis of deviance on this model and the model with only 'Poetacode' and 'Alcohol'.

```
Model 1: cluster ~ Poetacode_c + Alcohol_c1 + Alcohol_c2
Model 2: cluster ~ Poetacode_c + Alcohol_c1 + Alcohol_c2 + Rec1
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        24    20.4546
2        21    15.8566  3   4.5980    0.2037
```

Table 6: Analysis of Deviance on dataset type 1, model with Poetacode and Alcohol versus model with Poetacode, Alcohol, and Rec1

From Table 6 we see that we cannot reject the null-hypothesis under the significance level of 0.15. Therefore, we may assume that the model with 'Poetacode', 'Alcohol', and 'Rec1' is not better than the model with only 'Poetacode' and 'Alcohol'.
Now we will drop the clinical variable 'Rec1' and add to the model with 'Poetacode' and 'Alcohol' the clinical variable 'PTNMT' and perform an analysis of deviance on this model and the model with only 'Poetacode' and 'Alcohol'.

```
Model 1: cluster ~ Poetacode_c + Alcohol_c1 + Alcohol_c2
Model 2: cluster ~ Poetacode_c + Alcohol_c1 + Alcohol_c2 + PTNMT
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        24    20.4546
2        23    19.5190  1   0.9356    0.3334
```

Table 7: Analysis of Deviance on dataset type 1, model with Poetacode and Alcohol versus model with Poetacode, Alcohol, and PTNMT

From Table 7 we see that we cannot reject the null-hypothesis under the significance level of 0.15. Therefore, we may assume that the model with 'Poetacode', 'Alcohol', and 'PTNMT' is not better than the model with only 'Poetacode' and 'Alcohol'.

The final model is the model with only 'Poetacode' and 'Alcohol'.

4.2 Results univariate approach dataset type 2

In Table 8 the p-values are given for doing an analysis of deviance on an empty model and a model with one clinical variable on dataset type 2.

Clinical variable	p-value	significant
'Mut1'	0.011	*
'Age'	0.756	
'PTNMT'	0.069	*
'PTNMN'	0.117	*
'Smoking'	0.753	
'Packyears'	0.932	
'Alcohol'	0.008	*
'Unitsyear'	0.271	
'Poetacode'	0.002	*
'Gender'	0.063	*
'Codeloctum'	0.225	
'Stage'	0.209	
'Rec1'	0.060	*

Table 8: p-values of the univariate analysis on dataset type 2

From Table 8 we can see that the clinical variables 'Mut1', 'PTNMT', 'PTNMN' 'Alcohol', 'Poetacode', 'Gender', and 'Rec1' have a p-value smaller than 0.15.

In order to select the final model we will repeat the procedure described in Section 3.6. The clinical variables 'Mut1', 'PTNMT', 'PTNMN' 'Alcohol', 'Poetacode', 'Gender', and 'Rec1' will be added in the following order 'Poetacode', 'Alcohol', 'Mut1', 'Rec1', 'Gender', 'PTNMT', and 'PTNMN' (from smallest p-value to largest p-value, where all p-values are smaller than 0.15). Also, for the selection of the final model, we will use the dataset with removing all patients who had at least one missing input.

First the clinical variable 'Poetacode' will be added to the empty model.

```
Model 1: cluster ~ 1
```

```
Model 2: cluster ~ Poetacode_c
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        32     36.555
2        31     27.910  1    8.644     0.003
```

Table 9: Analysis of Deviance on dataset type 2, model with Poetacode versus empty model

From Table 9 we see that we can reject the null-hypothesis under the significance level of 0.15. Therefore, we may assume that the model with 'Poetacode' is better than the empty model.
Now we add to the model with 'Poetacode' the clinical variable 'Alcohol' and perform an analysis of deviance on this model and the model with only 'Poetacode'.

```
Model 1: cluster ~ Poetacode_c
Model 2: cluster ~ Alcohol_c1 + Alcohol_c2 + Poetacode_c
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        31    27.9102
2        29    21.8883  2   6.0219    0.0492
```

Table 10: Analysis of Deviance on dataset type 2, model with Poetacode versus model with Poetacode and Alcohol

From Table 10 we see that we can reject the null-hypothesis under the significance level of 0.15. Therefore, we may assume that the model with 'Poetacode' and 'Alcohol' is better than the model with only 'Poetacode'.
Now we add to the model with 'Poetacode' and 'Alcohol' the clinical variable 'Mut1' and perform an analysis of deviance on this model and the model with only 'Poetacode' and 'Alcohol'.

```
Model 1: cluster ~ Alcohol_c1 + Alcohol_c2 + Poetacode_c
Model 2: cluster ~ Alcohol_c1 + Alcohol_c2 + Poetacode_c + Mut1
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        29    21.8883
2        28    21.8009  1   0.0874    0.7675
```

Table 11: Analysis of Deviance on dataset type 2, model with Poetacode and Alcohol versus model with Poetacode, Alcohol and Mut1

From Table 11 we see that we cannot reject the null-hypothesis under the significance level of 0.15. Therefore, we may assume that the model with 'Poetacode' and 'Alcohol' is better than the model with 'Poetacode', 'Alcohol' and 'Mut1'.
We will now drop the variable 'Mut1'and add to the model with 'Poetacode' and 'Alcohol' the clinical variable 'Rec1' and perform an analysis of deviance on this model and the model with only 'Poetacode' and 'Alcohol'.

```
Model 1: cluster ~ Alcohol_c1 + Alcohol_c2 + Poetacode_c
Model 2: cluster ~ Alcohol_c1 + Alcohol_c2 + Poetacode_c + Rec1
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        29    21.8883
2        26    16.8597  3   5.0286    0.1697
```

Table 12: Analysis of Deviance on dataset type 2, model with Poetacode and Alcohol versus model with Poetacode, Alcohol and Rec1

From Table 12 we see that we cannot reject the null-hypothesis under the significance level of 0.15. Therefore, we may assume that the model with 'Poetacode' and 'Alcohol' is better than the model with 'Poetacode', 'Alcohol' and 'Rec1'.
We will now drop the variable 'Rec1'and add to the model with 'Poetacode' and 'Alcohol' the clinical variable 'Gender' and perform an analysis of deviance on this model and the model with only 'Poetacode' and 'Alcohol'.

```
Model 1: cluster ~ Alcohol_c1 + Alcohol_c2 + Poetacode_c
Model 2: cluster ~ Alcohol_c1 + Alcohol_c2 + Poetacode_c + Gender_c
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        29    21.8883
2        28    18.4360  1   3.4524    0.0632
```

Table 13: Analysis of Deviance on dataset type 2, model with Poetacode and Alcohol versus model with Poetacode, Alcohol, and Gender

From Table 13 we see that we can reject the null-hypothesis under the significance level of 0.15. Therefore, we may assume that the model with 'Poetacode', 'Alcohol', and 'Gender' is better than the model with 'Poetacode' and 'Alcohol'.
We will add to the model with 'Poetacode', 'Alcohol', and 'Gender' the clinical variable 'PTNMT' and perform an analysis of deviance on this model and the model with only 'Poetacode', 'Alcohol', and 'Gender'.

```
Model 1: cluster ~ Alcohol_c1 + Alcohol_c2 + Poetacode_c + Gender_c
Model 2: cluster ~ Alcohol_c1 + Alcohol_c2 + Poetacode_c + Gender_c +
    PTNMT
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        28    18.4360
2        27    17.2665  1   1.1695    0.2795
```

Table 14: Analysis of Deviance on dataset type 2, model with Poetacode, Alcohol, and Gender versus model with Poetacode, Alcohol, Gender, and PTNMT

From Table 14 we see that we cannot reject the null-hypothesis under the significance level of 0.15. Therefore, we may assume that the model with 'Poetacode', 'Alcohol', and 'Gender' is better than the model with 'Poetacode', 'Alcohol', 'Gender', and 'PTNMT'.
Now we will drop the clinical variable 'PTNMT' and add to the model with 'Poetacode', 'Alcohol', and 'Gender' the clinical variable 'PTNMN' and perform an analysis of deviance on this model and the model with only 'Poetacode', 'Alcohol', and 'Gender'.

```
Model 1: cluster ~ Alcohol_c1 + Alcohol_c2 + Poetacode_c + Gender_c
Model 2: cluster ~ Alcohol_c1 + Alcohol_c2 + Poetacode_c + Gender_c +
    (PTNMN_c1 + PTNMN_c2 + PTNMN_c3)
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        28    18.4360
2        25    12.1580  3   6.2780    0.0988
```

Table 15: Analysis of Deviance on dataset type 2, model with Poetacode, Alcohol, and Gender versus model with Poetacode, Alcohol, Gender, and PTNMN

From Table 15 we see that we can reject the null-hypothesis under the significance level of 0.15. Therefore, we may assume that the model with 'Poetacode',

‘Alcohol’,‘Gender’, and ‘PTNMN’ is better than the model with ‘Poetacode’, ‘Alcohol’, and ‘Gender’.

The final model is the model with only ‘Poetacode’, ‘Alcohol’, and ‘Gender’.

4.3 Results multivariate approach dataset type 1

Table 16 provides the result of the multivariate analysis done on dataset type 1.

```
Step:  AIC=20
cluster ~ Mut1 + Age + Unitsyear + Gender_c + Codeloctum_c +
    Smoking_c1 + Smoking_c2 + Alcohol_c1 + Alcohol_c2

               Df  Deviance    AIC    LRT   Pr(Chi)
<none>            1.774e-08 20.000
- Gender_c      1     5.461 23.461  5.461 0.0194496 *
- Age           1     5.545 23.545  5.545 0.0185317 *
- Smoking_c2    1     6.326 24.326  6.326 0.0118990 *
- Smoking_c1    1     7.640 25.640  7.640 0.0057093 **
- Codeloctum_c  1     8.300 26.300  8.300 0.0039650 **
- Unitsyear     1    13.508 31.508 13.508 0.0002375 ***
- Mut1          1    13.772 31.772 13.772 0.0002063 ***
- Alcohol_c2    1    17.283 35.283 17.283 3.220e-05 ***
- Alcohol_c1    1    19.046 37.046 19.046 1.276e-05 ***
```

Table 16: Result on multivariate analysis on dataset type 1, in backward direction

From Table 16 we can see that the clinical variables ‘Gender’, ‘Age’, ‘Smoking’, ‘Codeloctum’, ‘Unitsyear’, ‘Mut1’, and ‘Alcohol’ are selected as final model by the ‘step’ function in backward direction.

```
Step:  AIC=16
cluster ~ Poetacode_c + PTNMN_c3 + Gender_c + Alcohol_c1 + Unitsyear
+     Packyears + Age

               Df  Deviance    AIC       LRT Pr(Chi)
<none>            3.136e-07  16.00
+ Smoking_c1    1 3.252e-08  18.00 2.811e-07  0.9996
+ Alcohol_c2    1 1.764e-07  18.00 1.372e-07  0.9997
+ Smoking_c2    1 2.624e-07  18.00 5.120e-08  0.9998
+ Codeloctum_c  1 2.731e-07  18.00 4.048e-08  0.9998
+ Mut1          1 2.923e-07  18.00 2.137e-08  0.9999
+ Mut1_1        1 2.923e-07  18.00 2.137e-08  0.9999
+ PTNMN_c2      1 3.194e-07  18.00      0.00  1.0000
+ PTNMN_c1      1 3.195e-07  18.00      0.00  1.0000
+ Stage_c       1 8.383e-06  18.00      0.00  1.0000
+ Rec1          3 9.829e-09  22.00 3.038e-07  1.0000
+ PTNMT         1    144.18 162.18      0.00  1.0000
```

Table 17: Result on multivariate analysis on dataset type 1, in forward direction

From Table 17 we can see that the clinical variables ‘Poetacode, ‘PTNMN_c3’, ‘Gender’, ‘Alcohol_c1’, ‘Unitsyear’, ‘Packyears, and ‘Age’ are selected as final model by the ‘step’ function in forward direction.

```
Step:  AIC=17.55
cluster ~ Poetacode_c + PTNMN_c3 + Alcohol_c1 + Unitsyear + Packyears
```

```
               Df Deviance     AIC       LRT   Pr(Chi)
<none>              5.5452 17.5452
+ PTNMT         1   3.8191 17.8191    1.7261 0.1889107
+ Stage_c       1   3.8191 17.8191    1.7261 0.1889107
- Packyears     1   8.9974 18.9974    3.4522 0.0631682 .
+ Age           1   5.4606 19.4606    0.0845 0.7712357
+ Alcohol_c2    1   5.5452 19.5452 1.243e-08 0.9999111
+ Smoking_c2    1   5.5452 19.5452 5.064e-09 0.9999432
+ Smoking_c1    1   5.5452 19.5452 4.462e-09 0.9999467
+ Mut1          1   5.5452 19.5452 2.471e-09 0.9999603
+ Mut1_1        1   5.5452 19.5452 2.471e-09 0.9999603
+ Codeloctum_c  1   5.5452 19.5452 2.406e-09 0.9999609
+ PTNMN_c2      1   5.5452 19.5452    0.0000 1.0000000
+ PTNMN_c1      1   5.5452 19.5452    0.0000 1.0000000
+ Gender_c      1   5.5452 19.5452    0.0000 1.0000000
+ Rec1          3   3.8191 21.8191    1.7261 0.6311485
- Unitsyear     1  14.6224 24.6224    9.0772 0.0025882 **
- Poetacode_c   1  16.1201 26.1201   10.5749 0.0011463 **
- PTNMN_c3      1  20.4606 30.4606   14.9155 0.0001124 ***
- Alcohol_c1    1  20.6897 30.6897   15.1445 9.959e-05 ***
```

Table 18: Result on multivariate analysis on dataset type 1, in both directions

From Table 18 we can see that the clinical variables 'Poetacode', 'PTNMN_c3', 'Alcohol_c1', 'Unitsyear', and 'Packyears' are selected as final model by the 'step' function in both directions.

4.4 Results multivariate approach dataset type 2

Table 19 provides the result of the multivariate analysis done on dataset type 2.

```
Step:  AIC=18
cluster ~ Age + Unitsyear + Gender_c + Poetacode_c + PTNMN_c3 +
    Smoking_c1 + Smoking_c2 + Alcohol_c1

              Df  Deviance    AIC    LRT   Pr(Chi)
<none>           3.691e-08 18.000
- Gender_c     1     5.461 21.461  5.461 0.0194496 *
- Age          1     5.545 21.545  5.545 0.0185317 *
- Smoking_c2   1     6.326 22.326  6.326 0.0118990 *
- Smoking_c1   1     7.640 23.640  7.640 0.0057093 **
- Poetacode_c  1    10.458 26.458 10.458 0.0012210 **
- Unitsyear    1    11.303 27.303 11.303 0.0007739 ***
- PTNMN_c3     1    15.719 31.719 15.719 7.348e-05 ***
- Alcohol_c1   1    16.322 32.322 16.322 5.345e-05 ***
```

Table 19: Result on multivariate analysis on dataset type 2, backward direction

From Table 19 we can see that the clinical variables 'Age', 'Unitsyear', 'Gender', 'Poetacode', 'PTNMN_c3', 'Smoking', and 'Alcohol_c1' are selected as final model by the 'step' function in backward direction.

```
Step:  AIC=16
cluster ~ Poetacode_c + PTNMN_c3 + Gender_c + Alcohol_c1 + Unitsyear
+     Packyears + Age
```

```
                Df  Deviance    AIC       LRT Pr(Chi)
<none>             4.037e-07  16.00
+ Smoking_c1     1 3.978e-08  18.00 3.639e-07  0.9995
+ Alcohol_c2     1 2.031e-07  18.00 2.007e-07  0.9996
+ Smoking_c2     1 3.194e-07  18.00 8.436e-08  0.9998
+ Codeloctum_c   1 3.493e-07  18.00 5.445e-08  0.9998
+ Mut1           1 3.836e-07  18.00 2.010e-08  0.9999
+ Mut1_1         1 3.836e-07  18.00 2.010e-08  0.9999
+ PTNMN_c2       1 4.138e-07  18.00      0.00  1.0000
+ PTNMN_c1       1 4.164e-07  18.00      0.00  1.0000
+ PTNMT          1 7.021e-06  18.00      0.00  1.0000
+ Rec1           3 1.269e-08  22.00 3.910e-07  1.0000
+ Stage_c        1    288.35 306.35      0.00  1.0000
```

Table 20: Result on multivariate analysis on dataset type 2, forward direction

From Table 20 we can see that the clinical variables 'Poetacode', 'PTNMN_c3', 'Gender', 'Alcohol_c1', 'Unitsyear', 'Packyears', and 'Age' are selected as final model by the 'step' function in forward direction.

```
Step:  AIC=17.55
cluster ~ Poetacode_c + PTNMN_c3 + Alcohol_c1 + Unitsyear + Packyears

                Df Deviance    AIC       LRT   Pr(Chi)
<none>               5.545 17.545
+ PTNMT          1    3.819 17.819     1.726  0.188911
+ Stage_c        1    3.819 17.819     1.726  0.188911
- Packyears      1    8.997 18.997     3.452  0.063168 .
+ Age            1    5.461 19.461     0.085  0.771236
+ Codeloctum_c   1    5.545 19.545 1.159e-09  0.999973
+ Mut1           1    5.545 19.545 7.227e-10  0.999979
+ Mut1_1         1    5.545 19.545 7.227e-10  0.999979
+ PTNMN_c2       1    5.545 19.545 1.589e-10  0.999990
+ PTNMN_c1       1    5.545 19.545 8.638e-11  0.999993
+ Gender_c       1    5.545 19.545     0.000  1.000000
+ Alcohol_c2     1    5.545 19.545     0.000  1.000000
+ Smoking_c2     1    5.545 19.545     0.000  1.000000
+ Smoking_c1     1    5.545 19.545     0.000  1.000000
+ Rec1           3    3.819 21.819     1.726  0.631149
- Unitsyear      1   15.001 25.001     9.456  0.002104 **
- Poetacode_c    1   16.570 26.570    11.025  0.000899 ***
- PTNMN_c3       1   21.365 31.365    15.819 6.969e-05 ***
- Alcohol_c1     1   22.169 32.169    16.623 4.558e-05 ***
```

Table 21: Result on multivariate analysis on dataset type 2, both directions

From Table 21 we can see that the clinical variables 'Poetacode', 'PTNMN_c3', 'Alcohol_c1', 'Unitsyear', and 'Packyears' are selected as final model by the 'step' function in forward direction.

4.5 Summary of all the results

For the univariate approach on dataset type 1, we can conclude that there is a actual final model, namely the model with only 'Poetacode', and 'Alcohol'.

For the univariate approach on dataset type 2, we can conclude that there is a actual final model, namely the model with only ‘Poetacode’, ‘Alcohol’, and ‘Gender’.

For the multivariate approach in backward direction on dataset type 1, we can conclude that the final model is the model with the clinical variables ‘Gender’, ‘Age’, ‘Smoking’, ‘Codeloctum’, ‘Unitsyear’, ‘Mut1’, and ‘Alcohol’.

For the multivariate approach in forward direction on dataset type 1, we can conclude that the final model is the model with the clinical variables ‘Poetacode, ‘PTNMN_c3’, ‘Gender’, ‘Alcohol_c1’, ‘Unitsyear’, ‘Packyears, and ‘Age’.

For the multivariate approach in both directions on dataset type 1, we can conclude that the final model is the model with the clinical variables ‘Poetacode’, ‘PTNMN_c3’, ‘Alcohol_c1’, ‘Unitsyear’, and ‘Packyears’.

For the multivariate approach in backward direction on dataset type 2, we can conclude that the final model is the model with the clinical variables ‘Age’, ‘Unitsyear’, ‘Gender’, ‘Poetacode’, ‘PTNMN_c3’, ‘Smoking’, and ‘Alcohol_c1’.

For the multivariate approach in forward direction on dataset type 2, we can conclude that the final model is the model with the clinical variables ‘Poetacode’, ‘PTNMN_c3’, ‘Gender’, ‘Alcohol_c1’, ‘Unitsyear’, ‘Packyears’, and ‘Age’.

For the multivariate approach in both directions on dataset type 2, we can conclude that the final model is the model with the clinical variables ‘Poetacode’, ‘PTNMN_c3’, ‘Alcohol_c1’, ‘Unitsyear’, and ‘Packyears’.

All the above described results are visualized in Table 22. A ‘x’ means that the clinical variable was in the output.

Clinical variable	Data set type 1				Data set type 2			
	Uni - variate	Multivariate			Uni- variate	Multivariate		
		Back - ward	For- ward	Both		Back- ward	For- ward	Both
Poetacode	x		x	x	x	x	x	x
Alcohol	x	x			x			
Unitsyear		x	x	x		x	x	x
Gender		x	x		x	x	x	
Age		x	x			x	x	
Smoking		x				x		
Packyears			x	x		x		x
Codeloctum		x						
Mut1		x						
PTNMN_c3			x	x		x	x	x
Alcohol_c1			x	x		x	x	x

Table 22: Results for both dataset types on univariate and multivariate analysis

5 Conclusion and discussion

The results from the univariate and multivariate analysis are different. First of all, because the selection for the univariate analysis is based on the deviance ($D = 2\phi(l(\tilde{\theta}) - l(\hat{\theta}))$) and the selection of the multivariate analysis is based on the AIC ($AIC(k) = -2\log likelihood + 2k$). The deviance is calculated as two times the difference between the log likelihood of the full model and the small model, where the AIC only takes into account two times the log likelihood of the estimated model plus 2 times the numbers of parameters. These two measures are different: the last one gives a penalty to the number of parameters, where the first one does not. Therefore, the results of the univariate and multivariate analysis are a bit different.
For both data types the results of the univariate analysis are a bit different. For the first dataset type the clinical variables 'Alcohol' and 'Poetacode' are included in the final model, where for the second dataset type not only the clinical variables 'Alcohol' and 'Poetacode', but also the clinical variable 'Gender' are included related to the genetic subgroups of HNSCC. That the results are not the same could be easily explained by the fact that the composition of the patients for the two dataset types is are not the same. The patients in the first dataset type are a subset of the ones in the second dataset type, and it is therefore not surprising that the clinical variables delivered as output for the first dataset type are also a subset of the output from the second dataset type.
Considering the results of the multivariate analysis for both dataset types, we see that for the direction "both" the results are surprisingly the same, where the results for backward and forward directions are a bit different. The results for the multivariate analysis for backward and forward direction are not the same, and could be explained by that for the backward direction the elimination of a clinical variable depends on all the variables already in the model, where for the forward direction the 'step' function adds clinical variables to an in the beginning empty model. For the forward direction one starts with an empty model, where for the backward direction one starts with a full model. In order to do both add and drop clinical variables, one should consider the multivariate analysis for the selection of variables in both directions. These results are more consistent. So, considering the results achieved with multivariate analysis in both directions, we see that the clinical variables 'Poetacode', 'Unitsyear', 'Packyears', 'PTNMN_c3', and 'Alcohol_c1' related to the genetic subgroups of HNSCC for both dataset types. However, the clinical variables 'PTNMN' and 'Alcohol' are not fully included and therefore we need to revise our previous conclusion, by saying that the clinical variables 'Poetacode', 'Unitsyear', and 'Packyears' are related to the genetic subgroups of HNSCC for both dataset types. As the clinical variables 'Unitsyear' and 'Alcohol' are highly correlated and in fact more or less describe the same thing, namely drinking alcohol, we can use these as one variable. Also, the clinical variables 'Packyears' and 'Smoking' are highly correlated and can thus therefore also be used as one variable, namely smoking cigarettes. In **[14]** it was found that not only the use of alcohol was linked to HNSCC, but also the use of tobacco products. For the multivariate analysis during this study it was found that 'Smoking' and/or 'Packyears' were related to the subgroups of HNSCC. The results for the univariate analysis do not indicate that the use of tobacco products is related to the subgroups of HNSCC.

Overall we can conclude that which clinical variables are related to the subgroups of HNSCC does not depend on where the cut-off is made for the multivariate analysis done in both directions. However, for the univariate analysis it does depend where the cut-off is made. And also, independently of which dataset types is used and which analysis is performed, the use of alcohol is strongly related to the subgroups of HNSCC. Next to this, the clinical variable ‘Poetacode’ is also related to the subgroups of HNSCC. Only for one analysis, namely the multivariate analysis in the backward direction for dataset type 1, this clinical variable was not in the output of the final model.

List of Figures

List of Tables

References

[1] De Gunst, M.C.M. (najaar 2006), "Statistical Models", Lecture notes VU Amsterdam.

[2] Van Wieringen, W.N., M.A. van de Wiel, B. Ylstra (…….), "Weighted clustering of called array CGH data".

[3] http://www.kwfkankerbestrijding.nl/index.jsp?objectid=15033

[4] http://www.cancer.org/docroot/CRI/content/CRI_2_4_1x_What_Is_Cancer.asp

[5] Lemaire, F., R. Millon, J. Young,, A. Cromer, C. Wasylyk, I. Schultz, D. Muller, P. Marchal, C. Zhao, D. Melle, L. Bracco, J. Abecassis, B. Wasylyk (2003), "Differential expression profiling of head and neck squamous cell carcinoma (HNSCC)" , British Journal of Cancer; 89: 1940 – 1949.

[6] Karahatay, S., K . Thomas , S . Koybasi , C . Senkal , S . ElOjeimy , X . Liu , J . Bielawski , T . Day , M . Gillespie , D . Sinha (2007), "Clinical Relevance of Ceramide Metabolism in the Pathogenesis of Human Head and Neck Squamous Cell Carcinoma (HNSCC): Attenuation of C_{18}-ceramide in HNSCC Tumors Correlates with Lymphovascular Invasion and Nodal Metastasis", Cancer Letters; 256(1): 101–111.

[7] http://www.r-project.org/

[8] http://learn.genomics.utah.edu/units/basics/tour/chromosome.swf

[9] http://www.enotalone.com/article/5356.html

[10] Jiang, W., M. Zahurak, Z. Zhou, H.L. Park, Z. Guo, G. Wu, D. Sidransky, B. Trink, J.A. Califano (2007), "Alterations of GPI transamidase subunits in head and neck squamous carcinoma", PubMed.

[11] http://www.watisgenomics.nl/genomics/genomics/i000502.html

[12] Hanahan, D., R. A.Weinberg RA (2000) "The hallmarks of cancer", Cell; Vol. 100, 57–70.

[13] Braakhuis, B.J.M., P.J.F. Snijders, W.J.H. Keune, C.J.L.M Meijer, H.J. Ruijter-Scheppers, C.R. Leemans, et al.(2004) "Genetic patterns in head and neck cancers that contain or lack transcriptionally active human papillomavirus", J Natl Cancer Inst;96:998–1006

[14] Wiseman, S.M., H. Swede , D.L. Stoler, G.R. Anderson, N.R. Rigual , W.L. Jr Hicks, W.G. Douglas, D. Tan , T.R. Loree (2003), "Squamous cell carcinoma of the head and neck in nonsmokers and nondrinkers: an analysis of clinicopathologic characteristics and treatment outcomes", Annals of Surgical Oncology;10(5):551-7.

[15] Poeta, M.L., J. Manola, M. A. Goldwasser, A. Forastiere, N. Benoit, J.A. Califano, J.A. Ridge, J. Goodwin, D. Kenady, J. Saunders, W. Westra, D. Sidransky, and W.M. Koch, (2007) "*TP53* Mutations and Survival in Squamous-Cell Carcinoma of the Head and Neck", N Engl J Med.; 357(25): 2552–2561.

[16] http://www.unsolvedmysteries.oregonstate.edu/

[17] http://www.bio.davidson.edu/Courses/genomics/chip/chip.html

[18]Smeets SJ, Braakhuis BJM, Abbas S, et al. Genome-wide DNA copy number alterations in head and neck squamous cell carcinomas with or without oncogene-expressing human papillomavirus. *Oncogene* 2006; **25**: 2558-2564.
[19] http://www.few.vu.nl/~mavdwiel/StatMod/ch3.pdf
[20] http://ceh.uconn.edu/classroom/defonco.html
[21] http://www.cancer.org/docroot/ETO/content/ETO_1_4x_oncogenes_and_tumor_suppressor_genes.asp
[22] http://www.medterms.com/script/main/art.asp?articlekey=3334

Appendix

A description of the clinical variables is given in the table below.

Clinical variable	**Description**	**Input**
'Mut1'	Describes mutation status of the TP53 gene.	0 or 1
'Age'	Says how old a person is.	in years
'PTNMT'	Pathological tumor stage: based on tumor size and histopathological differentiation.	1, 2, 3 , or 4
'PTNMN'	Pathological nodal stage: describes the number of tumor-positive lymphnodes.	0, 1, 2B , and 2C
'Smoking'	Contains information about the smoking status of a person: if the person is smoking at the moment (C), if he or she has smoked in the past, but stopped now (F), and the last one if the person has never smoked (N).	C, F, or N
'Packyears'	Pack-years were taken as a measure of cumulative tobacco consumption. It is calculated by multiplying the number of packs of cigarettes smoked per day by the number of years the person has smoked.	number of pack years
'Alcohol'	Contains information about the alcohol status of a person: if the person is drinking at the moment (F), if he or she has drank in the past, but stopped now (F), and the last one if the person has never drank (N).	C, F, or N
'Unit year'	Unit-years were taken as a measure of cumulative alcohol consumption and were calculated as the number of years drinking multiplied by the number of units per day. A unit is defined as one alcoholic beverage (equivalent to approximately 15 mL of alcohol).	number of Unit Years
'Poetacode'	Describes if the mutation leads to inactivation of the protein **[15]**.	Disruptive (D) and non-disruptive (ND)
'Gender'	Tells you if a person is a male or a female.	M or F
'Stage'	Describes how aggressive the tumor was.	I, II, III, IVA.
'Codeloctum'	Describes the location of the tumor; OC = Oral cavity, OP = Oropharynx.	OC or OP
'Rec1'	Provides the information if a tumor has reoccurred and on what location and	2P, DF, REG, LOC, or DM.

	timeframe compared to the primary tumor. 2P = second primary, DF = desease free, REG = regional recurrence, LOC, Local recurrence, DM = distant metastasis.	

Table A1: Description of the clinical variables