Principal component analysis

# 1 Summary Statistics

## 1.1 Demographics

The study involved 77 participants with the majority using Medtronic780 (40.3%) and Omnipod (29.9%) insulin pumps, while others used Medtronic640 (16.9%) and tslim (13.0%). Most participants utilized an open insulin delivery system (58.4%), and nearly all had Type I diabetes (93.5%). The gender distribution showed 62.3% females and 37.7% males, with an average age of 49.4 years. The average duration since diabetes diagnosis was 24.3 years, and the average duration of being diabetic was 23.6 years.

Table 1.1  
Demographic caracteristics

|  |  |  | Statistic (N=77) |
| --- | --- | --- | --- |
| Type of insulin pump | medtronic640 |  | 13 (16.9%) |
| medtronic780 |  | 31 (40.3%) |
| omnipod |  | 23 (29.9%) |
| tslim |  | 10 (13.0%) |
| Boucle | Open |  | 45 (58.4%) |
| Semi-closed |  | 32 (41.6%) |
| Diabetes type | Other |  | 5 (6.5%) |
| Type I |  | 72 (93.5%) |
| Sex | Female |  | 48 (62.3%) |
| Male |  | 29 (37.7%) |
| Age | Mean (SD) |  | 49.4 (14.9) |
| Median (IQR) |  | 49.0 (24.0) |
| Range |  | 22.0 - 76.0 |
| Duration after diabetes test | Mean (SD) |  | 24.3 (13.1) |
| Median (IQR) |  | 23.0 (19.0) |
| Range |  | 3.0 - 51.0 |
| Duration of being diabetic | Mean (SD) |  | 23.6 (13.1) |
| Median (IQR) |  | 23.0 (19.0) |
| Range |  | 2.0 - 50.0 |
| Source: Questionaire Baseline | | | |

## 1.2 Summary of Scores

The study assessed various health indices among participants. The World Health Organization Five Well-Being Index had a mean score of 58.6 with a range from 8.0 to 88.0, indicating varying levels of well-being. The Patient Health Questionnaire 9 showed a mean score of 6.9, reflecting mild depressive symptoms, with scores ranging from 0.0 to 26.0. The assessment of precariousness and health inequalities had a mean score of 19.6, highlighting moderate levels of socio-economic difficulties. Other metrics such as compliance (mean: 1.0), fragility (mean: 2.0), and various diabetes distress scales provided insights into the participants’ compliance, fragility, and distress levels, with glycated hemoglobin levels averaging 7.3, indicating overall blood glucose control. The Time in Range (TIR) metric, which measures the percentage of time participants’ blood glucose levels were within the target range, had a mean of 62.2%, suggesting moderate glycemic control. Additionally, the incidence of severe hypoglycemia in the past six months was relatively low, with a mean of 0.6 episodes, indicating that most participants experienced few severe hypoglycemic events.

Table 1.2  
Evaluation of various health indices

|  |  |  | Statistic (N=77) |
| --- | --- | --- | --- |
| World health organization five well being index | Mean (SD) |  | 58.65 (21.96) |
| Median (IQR) |  | 64.00 (32.00) |
| Range |  | 8.00 - 88.00 |
| Patient Health Questionnaire 9 | Mean (SD) |  | 6.88 (5.91) |
| Median (IQR) |  | 6.00 (8.00) |
| Range |  | 0.00 - 26.00 |
| Assessment of precariousness and health inequalities in health examination centers | Mean (SD) |  | 19.57 (19.43) |
| Median (IQR) |  | 15.38 (21.29) |
| Range |  | 0.00 - 81.07 |
| Gired : compliance | Mean (SD) |  | 0.99 (0.91) |
| Median (IQR) |  | 1.00 (1.00) |
| Range |  | 0.00 - 3.00 |
| Fried : fragility | Mean (SD) |  | 1.97 (1.39) |
| Median (IQR) |  | 2.00 (2.00) |
| Range |  | 0.00 - 4.00 |
| Diabetes distress scale | Mean (SD) |  | 2.10 (0.78) |
| Median (IQR) |  | 1.94 (1.06) |
| Range |  | 1.00 - 5.35 |
| Diabetes distress scale emotional | Mean (SD) |  | 2.76 (1.17) |
| Median (IQR) |  | 2.60 (1.60) |
| Range |  | 1.00 - 5.80 |
| Diabetes distress scale healthcare providers | Mean (SD) |  | 1.55 (0.74) |
| Median (IQR) |  | 1.25 (0.75) |
| Range |  | 1.00 - 4.50 |
| Diabetes distress scale lifestyle | Mean (SD) |  | 1.89 (0.84) |
| Median (IQR) |  | 1.60 (1.20) |
| Range |  | 1.00 - 5.60 |
| Diabetes family distress scale | Mean (SD) |  | 2.15 (1.37) |
| Median (IQR) |  | 1.67 (1.34) |
| Range |  | 1.00 - 6.00 |
| Glycated hemoglobin | Mean (SD) |  | 7.35 (0.83) |
| Median (IQR) |  | 7.20 (1.20) |
| Range |  | 5.88 - 10.30 |
| Time in range = time on target | Mean (SD) |  | 62.24 (16.65) |
| Median (IQR) |  | 65.00 (24.20) |
| Range |  | 23.90 - 89.00 |
| Severe hypoglycemia in the past 6 months | Mean (SD) |  | 0.57 (1.33) |
| Median (IQR) |  | 0.00 (0.00) |
| Range |  | 0.00 - 6.00 |
| Source: Questionnaire Baseline | | | |

## 1.3 Summary of EORTCQLQBR23

The scoring method for the EORTC QLQ-BR23 involves two main steps: calculating the raw score and performing a linear transformation. The raw score for each multi-item scale is the average of the corresponding item scores, while single-item measures use the item’s score as the raw score. Certain items require reverse scoring before statistical analysis. To standardize raw scores to a 0-100 range, a linear transformation is applied: for functional scales, the formula is , and for symptom scales, the formula is . High scores on functional scales indicate better functioning, while high scores on symptom scales indicate higher symptom severity.

Table 1.3  
Summary of EORTC QLQ-BR23 Scores

|  |  |  | Statistic (N=25) |
| --- | --- | --- | --- |
| Body Image | Mean (SD) |  | 57.33 (31.02) |
| Median (IQR) |  | 58.33 (50.00) |
| Range |  | 0.00 - 100.00 |
| Sexual Function | Mean (SD) |  | 24.67 (21.58) |
| Median (IQR) |  | 33.33 (33.33) |
| Range |  | 0.00 - 83.33 |
| Sexual Enjoyment | Mean (SD) |  | 47.22 (22.29) |
| Median (IQR) |  | 50.00 (33.33) |
| Range |  | 0.00 - 66.67 |
| Missing |  | 13 (52.0%) |
| Future Perspective | Mean (SD) |  | 40.00 (34.69) |
| Median (IQR) |  | 33.33 (66.67) |
| Range |  | 0.00 - 100.00 |
| Systemic Therapy Side Effects | Mean (SD) |  | 37.94 (21.28) |
| Median (IQR) |  | 42.86 (42.86) |
| Range |  | 9.52 - 72.22 |
| Breast Symptoms | Mean (SD) |  | 19.67 (17.66) |
| Median (IQR) |  | 16.67 (25.00) |
| Range |  | 0.00 - 66.67 |
| Arm Symptoms | Mean (SD) |  | 16.00 (20.81) |
| Median (IQR) |  | 0.00 (33.33) |
| Range |  | 0.00 - 55.56 |
| Upset by Hair Loss | Mean (SD) |  | 60.78 (39.50) |
| Median (IQR) |  | 66.67 (66.67) |
| Range |  | 0.00 - 100.00 |
| Missing |  | 8 (32.0%) |
| Source: Questionnaire Baseline | | | |

## 1.4 Summary of SATISFACTION WITH BREASTS

Table 1.4  
Summary of Satisfaction with Breasts Scores

|  |  |  | Statistic (N=25) |
| --- | --- | --- | --- |
| Satisfaction with Breasts | Mean (SD) |  | 60.92 (28.12) |
| Median (IQR) |  | 68.75 (25.00) |
| Range |  | 0.00 - 100.00 |
| Source: Questionnaire Baseline | | | |

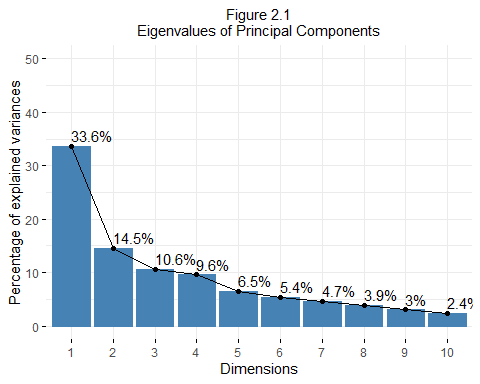
# 2 Principal Component Analysis (with encoded categories, typepompe excluded)

## 2.1 Eigenvalues of Principal Components

**Practical Implications** Based on the scree plot:

The scree plot shown in Figure 2.1 displays the eigenvalues of the principal components derived from a Principal Component Analysis (PCA). Each bar represents the percentage of the total variance explained by each principal component. Here’s how to interpret this plot:

1. **X-axis (Dimensions)**:
   * This axis represents the principal components, numbered from 1 to 10 in this plot.
2. **Y-axis (Percentage of Explained Variance)**:
   * This axis shows the percentage of the total variance in the dataset explained by each principal component.
3. **Bars (Percentage of Explained Variance by Each Component)**:
   * Each bar’s height indicates the proportion of the total variance explained by the corresponding principal component. For example, the first principal component explains approximately 33.6% of the total variance, while the second explains about 14.5%.
4. **Line Plot (Cumulative Explained Variance)**:
   * The black line connecting the points at the top of each bar represents the cumulative explained variance. It helps to visualize how much total variance is explained as more principal components are included.



## 2.2 Individual Plot

**Figure 2.2.1: Individuals Factor Map (PC1 vs PC2)**

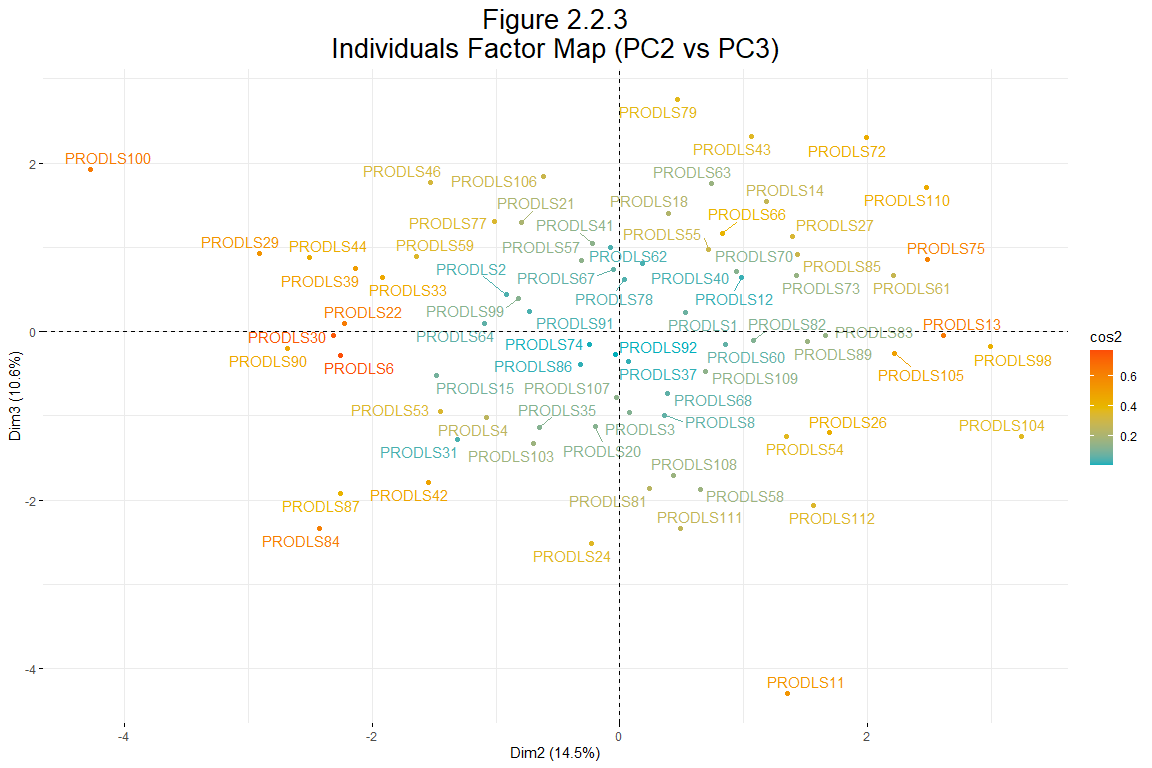
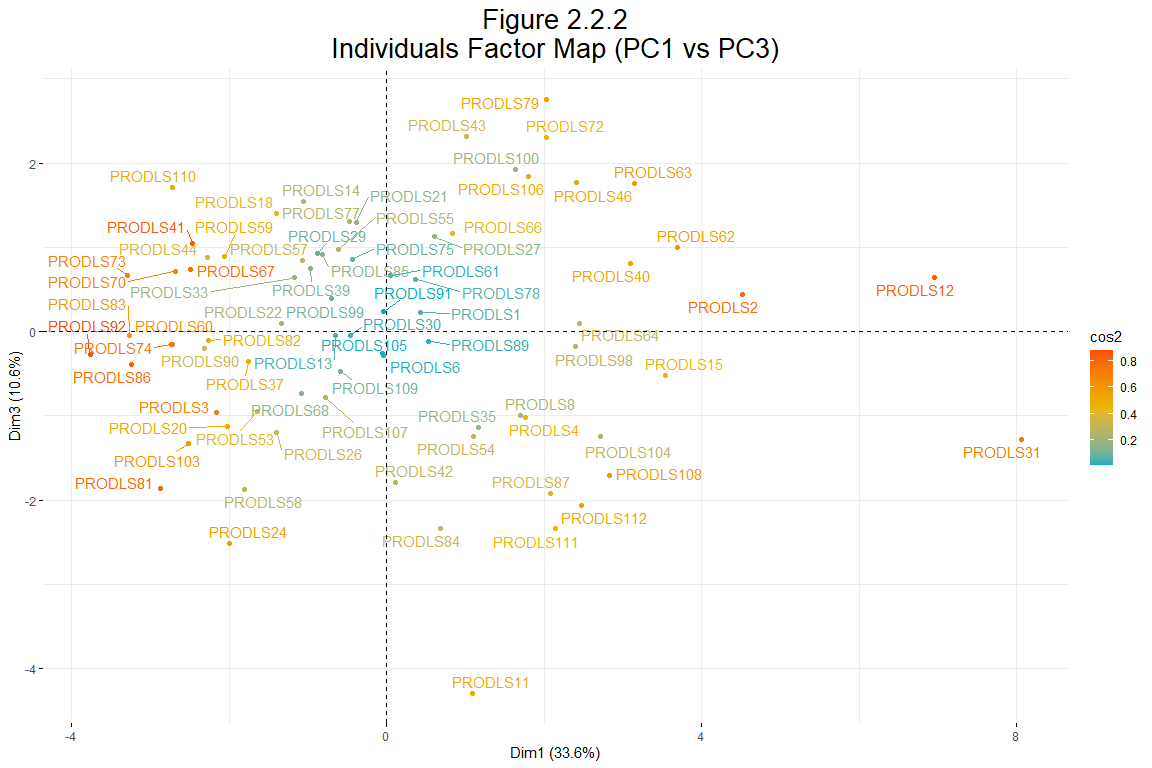
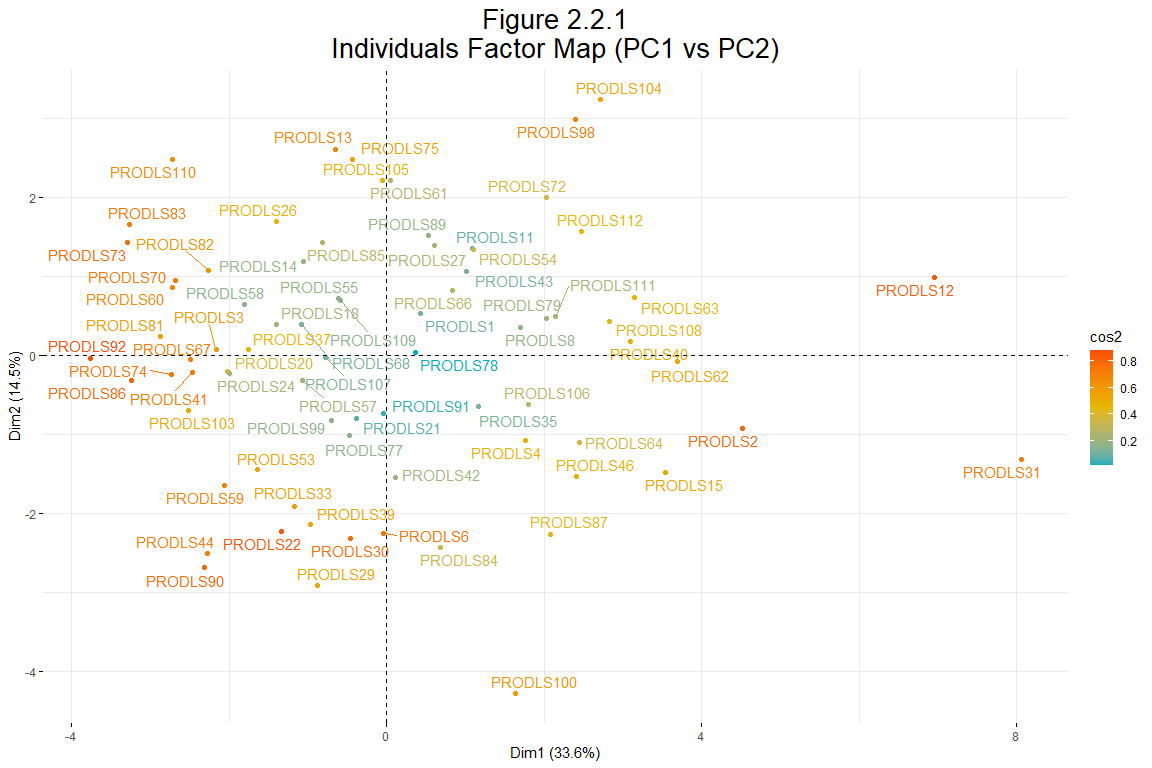
* **Axes**:
  + **PC1**: Explains 33.6% of the variance.
  + **PC2**: Explains 14.5% of the variance.
* **Color**: Points are colored based on cos2 values, indicating representation quality.
* **Interpretation**: Points farther from the origin indicate individuals with extreme values on PC1 and PC2. Clustering suggests similar profiles.

**Figure 2.2.2: Individuals Factor Map (PC1 vs PC3)**

* **Axes**:
  + **PC1**: Explains 33.6% of the variance.
  + **PC3**: Explains 10.6% of the variance.
* **Color**: Points colored by cos2 values.
* **Interpretation**: Shows additional variance captured by PC3. Identifies unique characteristics.

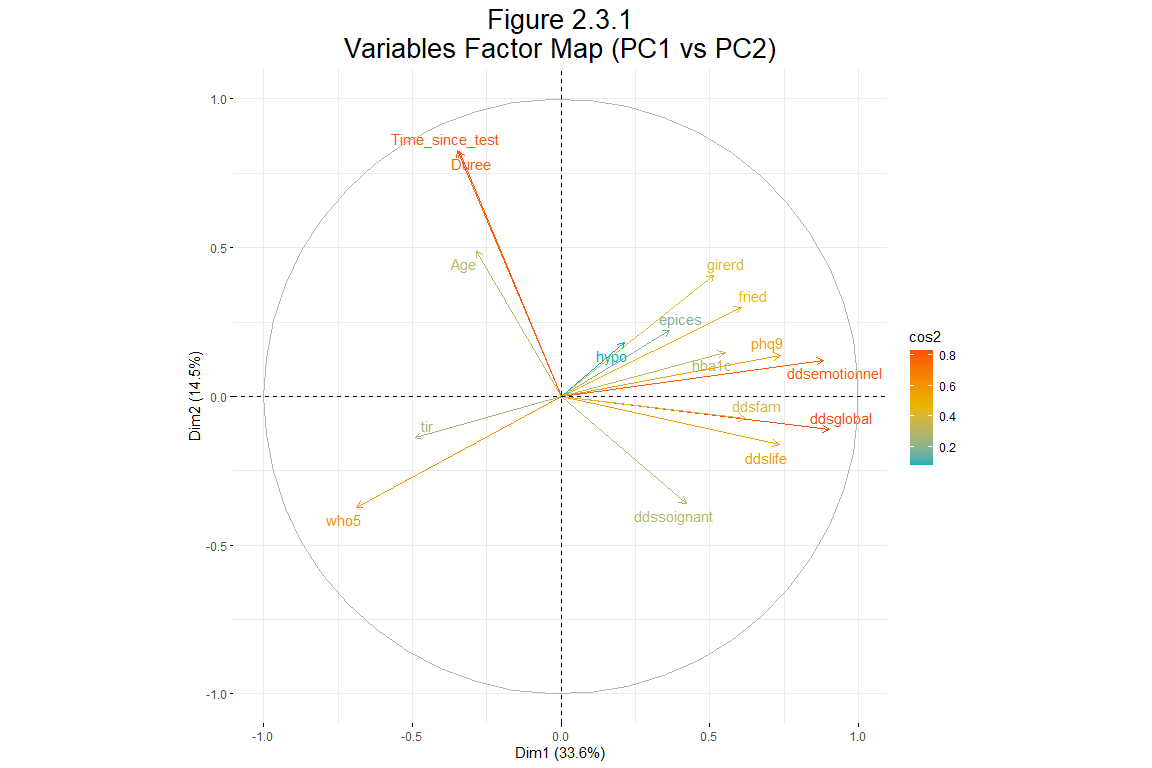
**Figure 2.2.3: Individuals Factor Map (PC2 vs PC3)**

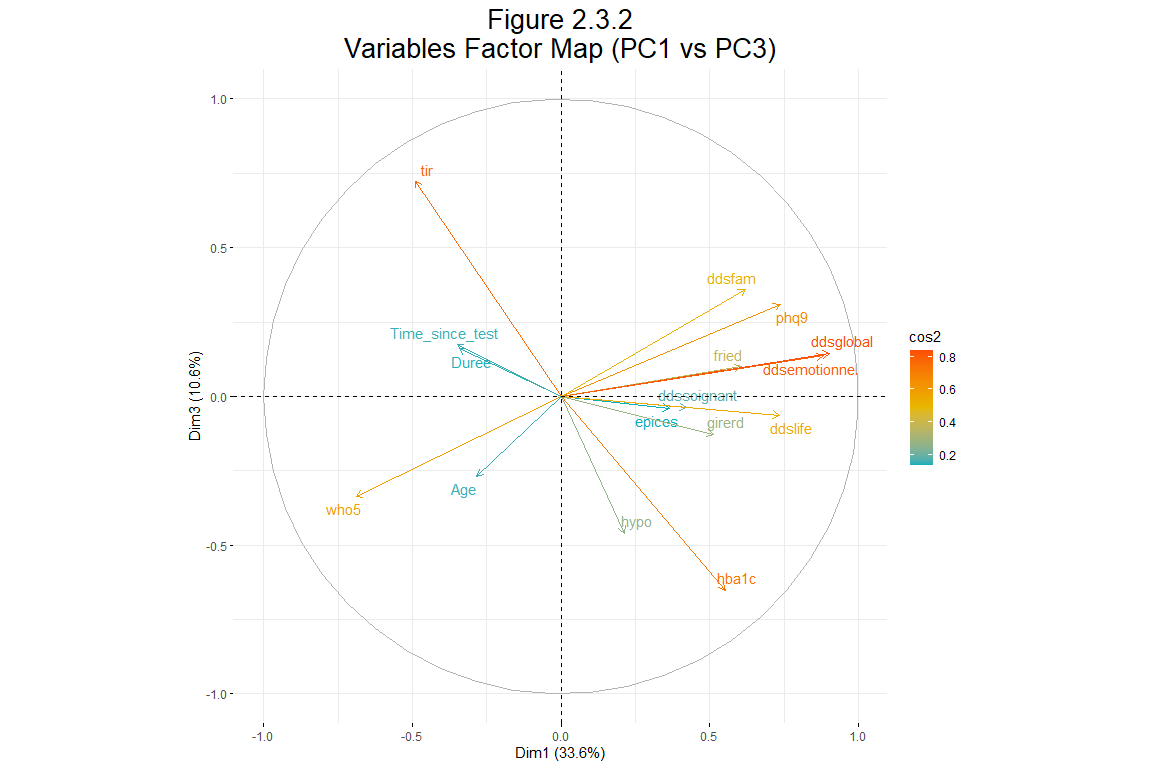
* **Axes**:
  + **PC2**: Explains 14.5% of the variance.
  + **PC3**: Explains 10.6% of the variance.
* **Color**: Points colored by cos2 values.
* **Interpretation**: Highlights variance captured by PC2 and PC3. Useful for examining unique relationships.

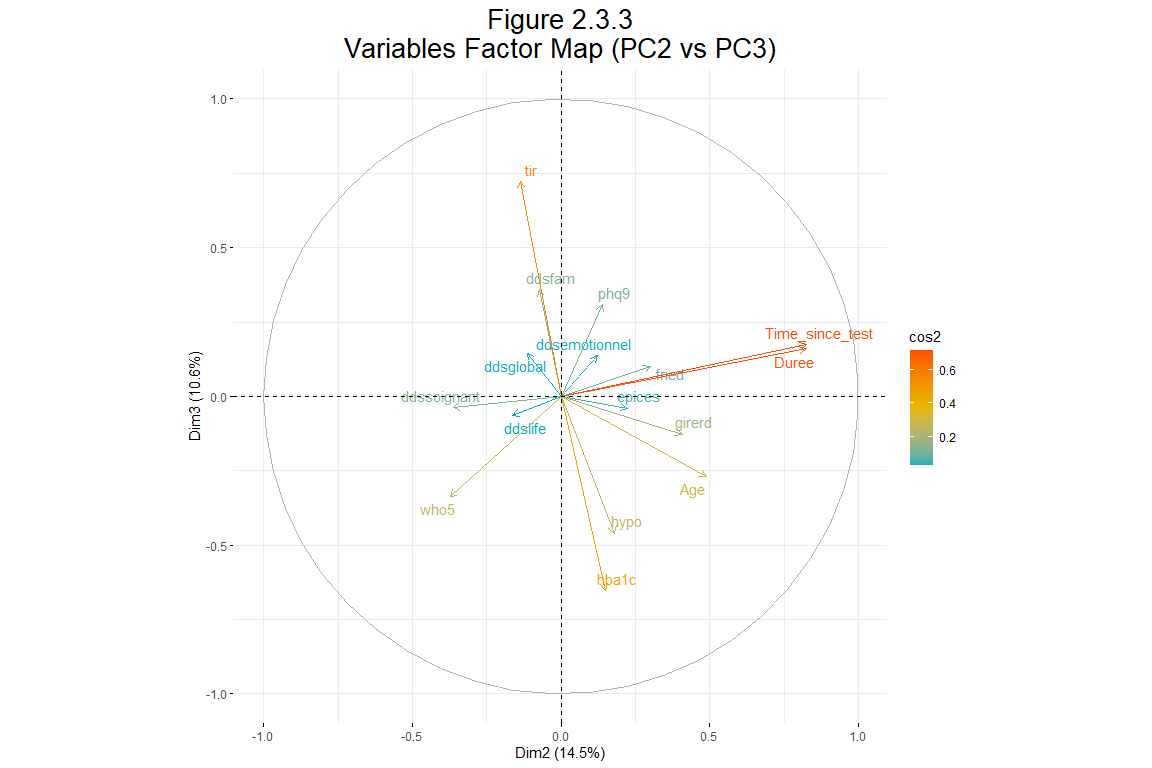


## 2.3 Variable Plot

* **Interpretation**:
  + **Vectors**: Each vector represents a variable. The length and direction indicate the contribution and correlation of the variable with the principal components.
  + **Clustering**: Variables that cluster together are correlated.
  + **Proximity to Circle**: Variables close to the circle’s edge are well represented by PC1 and PC2, while those closer to the center are less well represented.

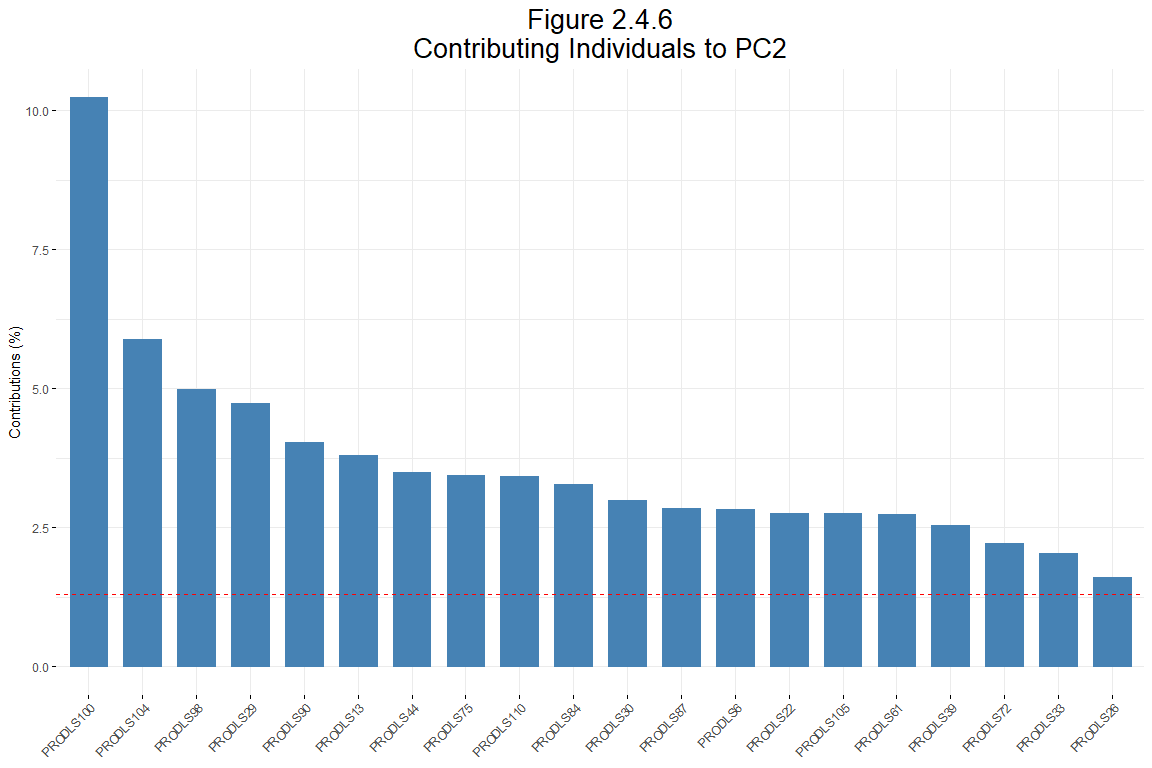
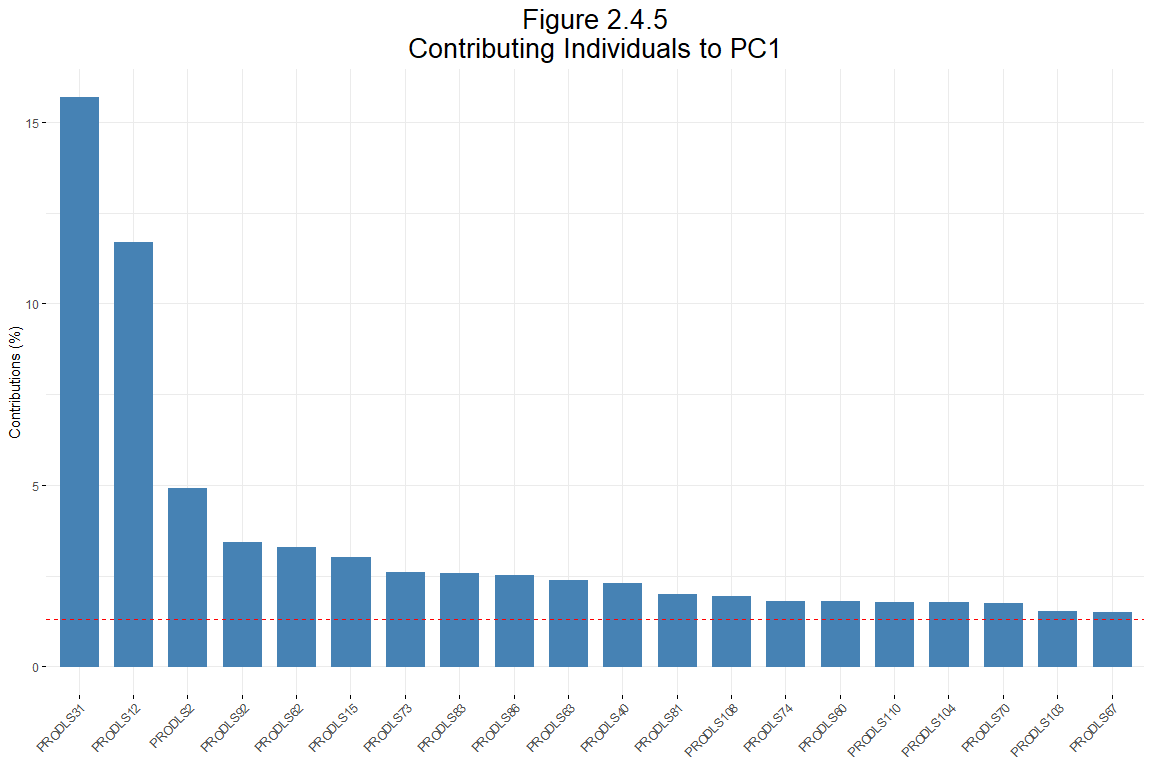
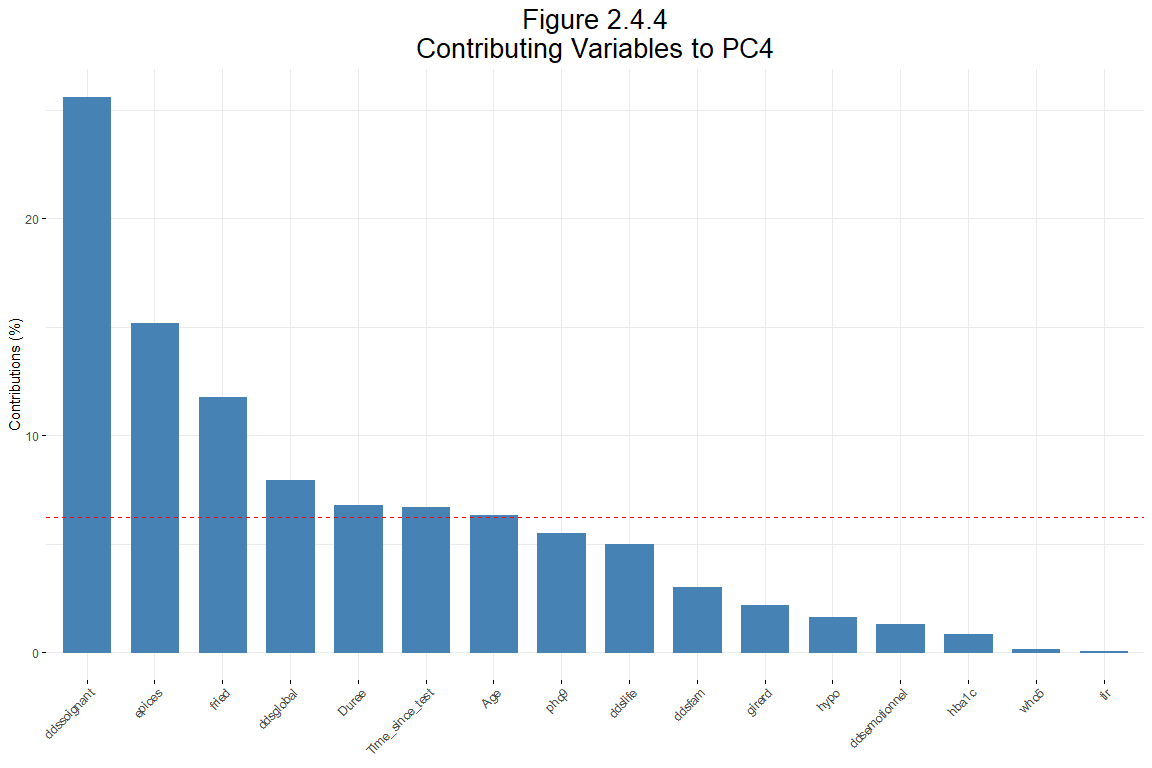
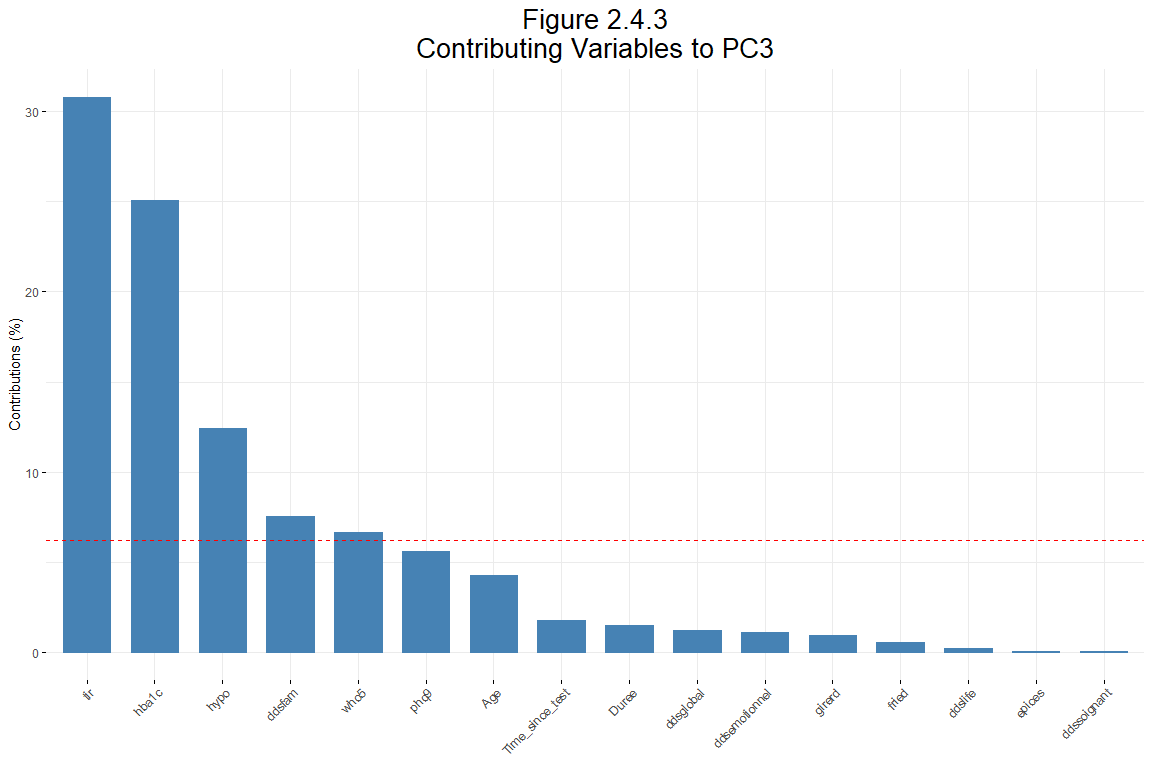
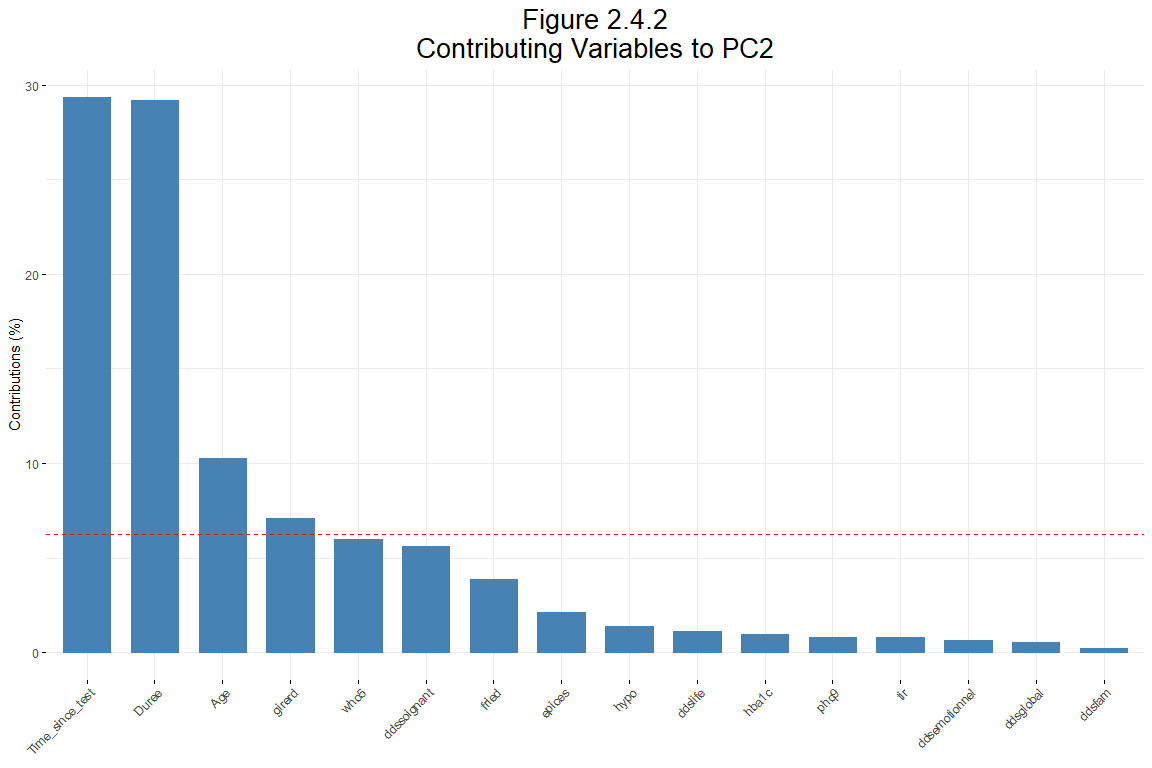
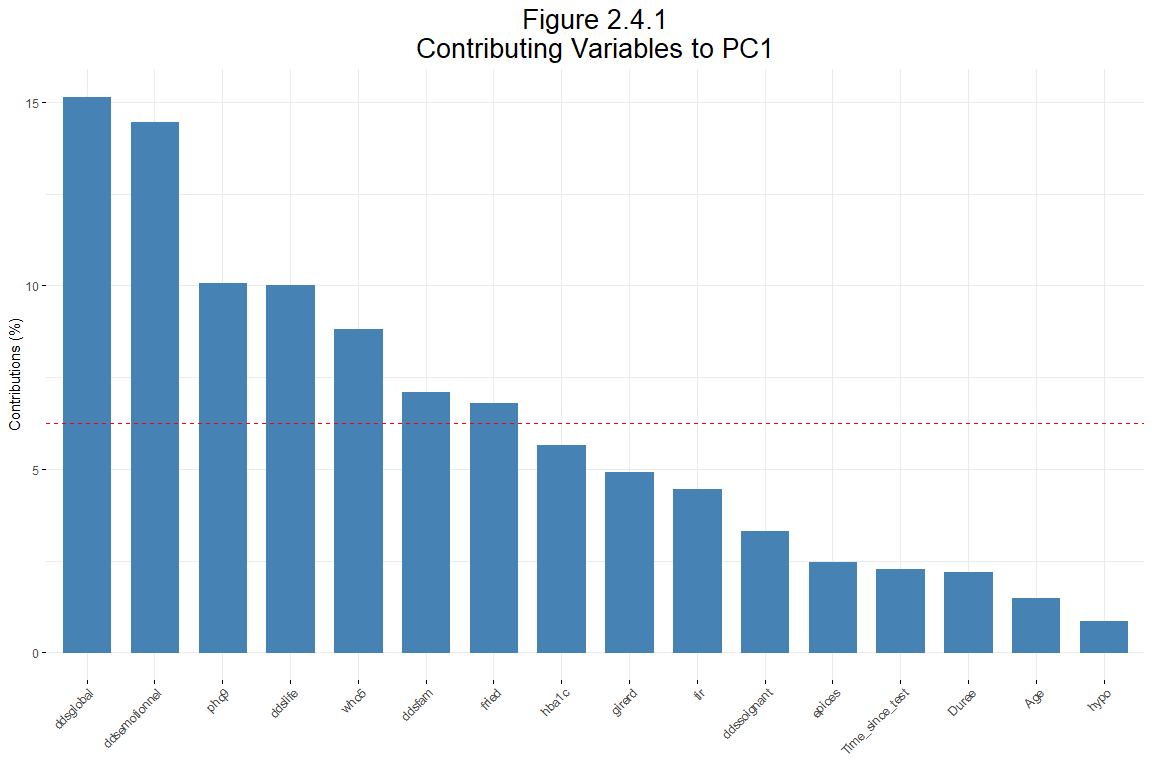




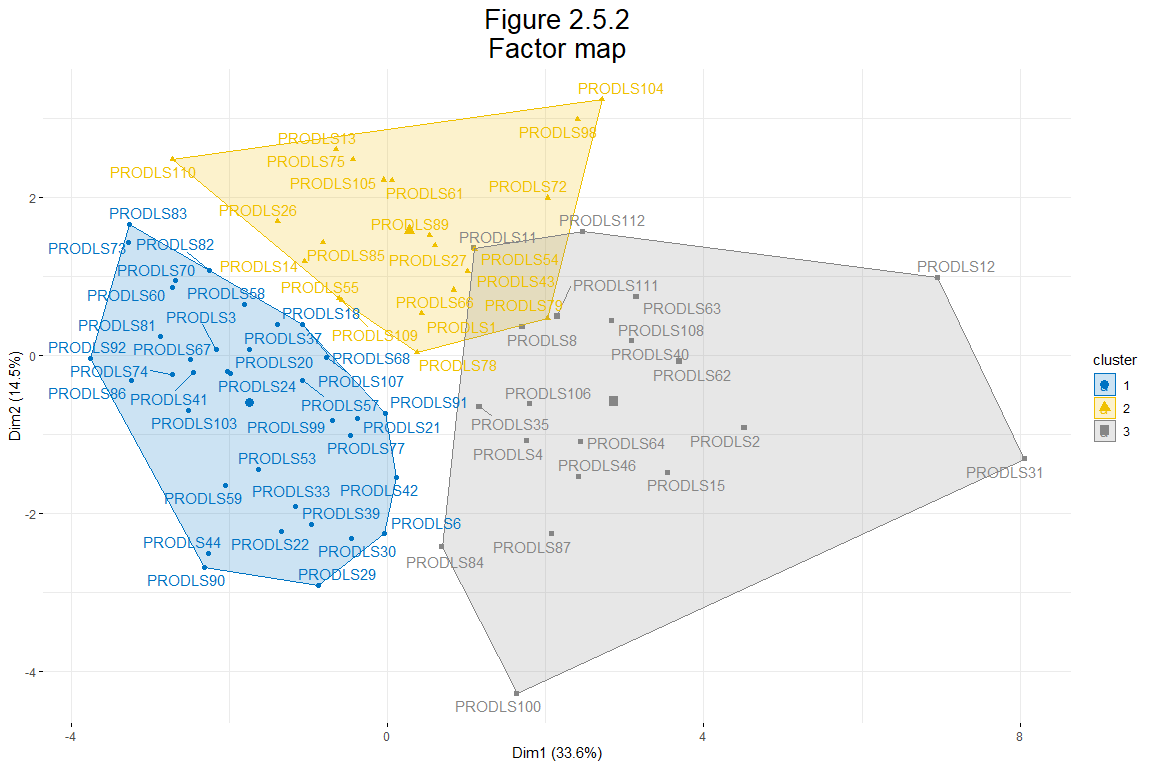
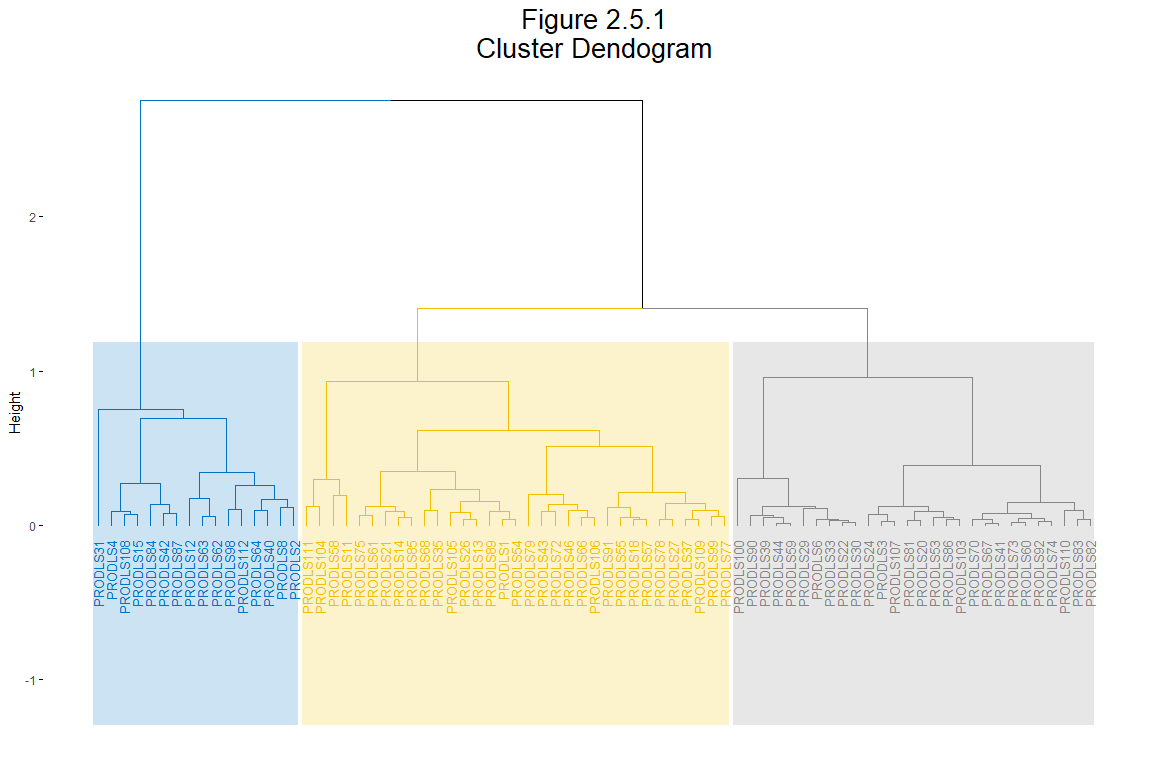


## 2.4 Contributing variables analysis

The bar plots illustrate the contributions of variables to the first 4 principal components.



## 2.5 Clustering analysis



### 2.5.1 Summary of Hierarchical Clustering Analysis after PCA

**Objective:** To group variables into clusters based on their similarity after performing Principal Component Analysis (PCA).

**Variables Analyzed:** A variety of health and behavioral metrics, including scores on psychological assessments, types of insulin pumps used, and blood glucose measures.

**Clusters Identified:** Three distinct clusters were identified based on the hierarchical clustering.

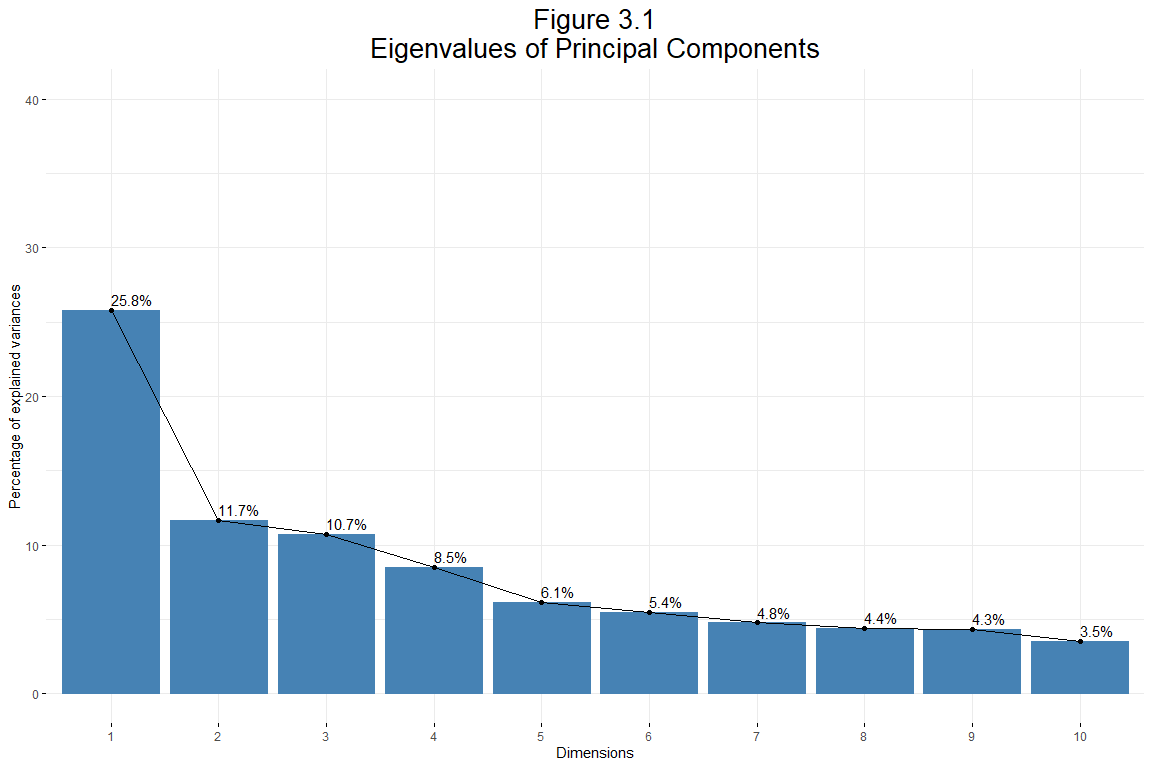
Table 2.1  
Description of each cluster by quantitative variables

| Variable | v.test | Mean in category | Overall mean | sd in category | Overall sd | p.value | Cluster |
| --- | --- | --- | --- | --- | --- | --- | --- |
| character | numeric | numeric | numeric | numeric | numeric | character | numeric |
| who5 | 5.79 | 74.11 | 58.65 | 11.53 | 21.82 | <0.005 | 1 |
| tir | 3.71 | 69.75 | 62.24 | 12.34 | 16.54 | <0.005 | 1 |
| epices | -2.37 | 13.97 | 19.57 | 15.09 | 19.31 | 0.018 | 1 |
| hba1c | -3.34 | 7.01 | 7.35 | 0.56 | 0.82 | <0.005 | 1 |
| ddslife | -3.73 | 1.51 | 1.89 | 0.50 | 0.83 | <0.005 | 1 |
| ddsfam | -3.98 | 1.49 | 2.15 | 0.52 | 1.36 | <0.005 | 1 |
| fried | -4.46 | 1.22 | 1.97 | 1.11 | 1.38 | <0.005 | 1 |
| girerd | -4.65 | 0.47 | 0.99 | 0.64 | 0.90 | <0.005 | 1 |
| phq9 | -5.02 | 3.28 | 6.88 | 3.38 | 5.87 | <0.005 | 1 |
| ddsglobal | -5.30 | 1.60 | 2.10 | 0.38 | 0.78 | <0.005 | 1 |
| ddsemotionnel | -5.95 | 1.92 | 2.76 | 0.60 | 1.16 | <0.005 | 1 |
| Time\_since\_test | 5.27 | 37.24 | 24.34 | 7.87 | 13.06 | <0.005 | 2 |
| Duree | 5.21 | 36.24 | 23.57 | 7.87 | 12.98 | <0.005 | 2 |
| who5 | -3.96 | 42.48 | 58.65 | 17.61 | 21.82 | <0.005 | 2 |
| ddsglobal | 5.48 | 2.93 | 2.10 | 0.80 | 0.78 | <0.005 | 3 |
| ddsemotionnel | 5.00 | 3.89 | 2.76 | 1.03 | 1.16 | <0.005 | 3 |
| ddslife | 4.40 | 2.60 | 1.89 | 0.99 | 0.83 | <0.005 | 3 |
| phq9 | 4.19 | 11.65 | 6.88 | 5.62 | 5.87 | <0.005 | 3 |
| girerd | 3.79 | 1.65 | 0.99 | 0.91 | 0.90 | <0.005 | 3 |
| ddsfam | 3.61 | 3.10 | 2.15 | 1.69 | 1.36 | <0.005 | 3 |
| hba1c | 3.51 | 7.91 | 7.35 | 1.02 | 0.82 | <0.005 | 3 |
| fried | 3.29 | 2.85 | 1.97 | 1.24 | 1.38 | <0.005 | 3 |
| ddssoignant | 3.21 | 2.01 | 1.55 | 0.98 | 0.74 | <0.005 | 3 |
| hypo | 2.45 | 1.20 | 0.57 | 1.81 | 1.32 | 0.014 | 3 |
| epices | 2.39 | 28.49 | 19.57 | 25.23 | 19.31 | 0.017 | 3 |
| Age | -2.54 | 42.10 | 49.39 | 14.90 | 14.83 | 0.011 | 3 |
| who5 | -2.57 | 47.80 | 58.65 | 21.11 | 21.82 | 0.010 | 3 |
| Duree | -4.34 | 12.65 | 23.57 | 6.92 | 12.98 | <0.005 | 3 |
| tir | -4.44 | 48.02 | 62.24 | 16.85 | 16.54 | <0.005 | 3 |
| Time\_since\_test | -4.58 | 12.75 | 24.34 | 5.97 | 13.06 | <0.005 | 3 |
| n: 30 | | | | | | | |
| Source: Questionaire Baseline | | | | | | | |

# 3 Factor Analysis of Mixed Data (typepompe excluded)

## 3.1 Dimensions overview

This plot represents the results of a Factor Analysis of Mixed Data (FAMD), which is a technique used to analyze datasets containing both categorical and continuous variables. The goal of FAMD is to reduce the dimensionality of the data while preserving as much variability as possible.

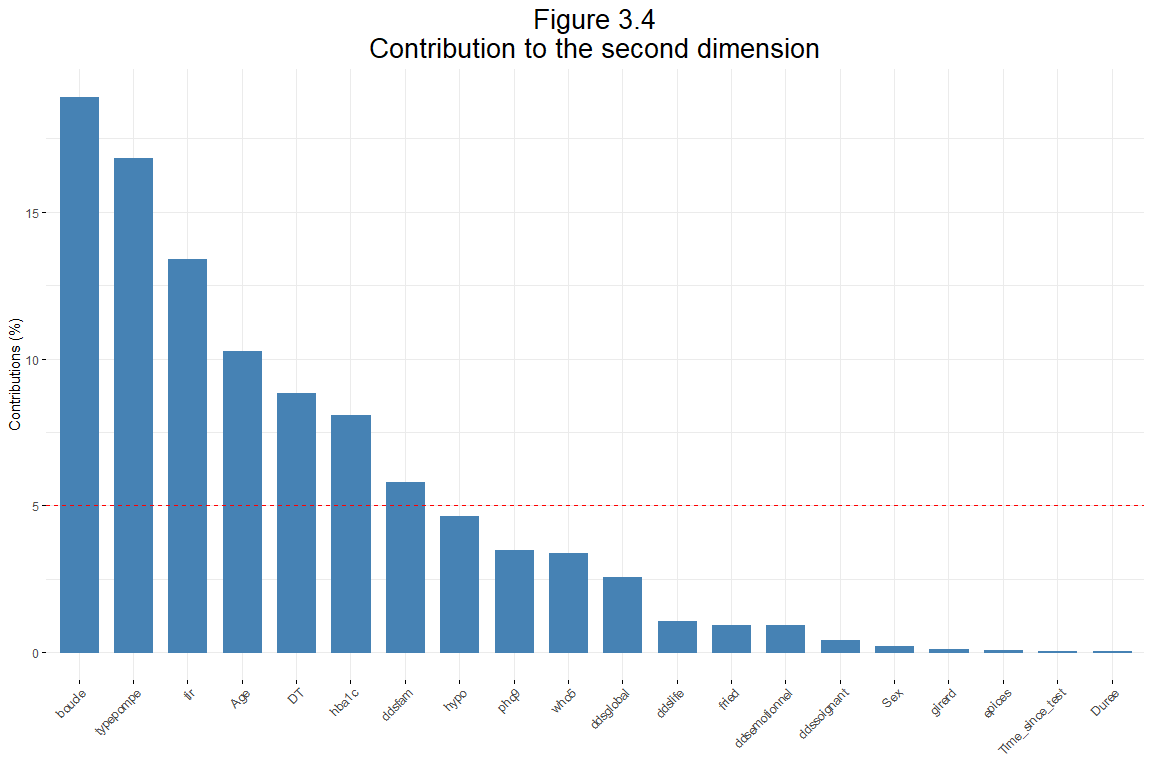
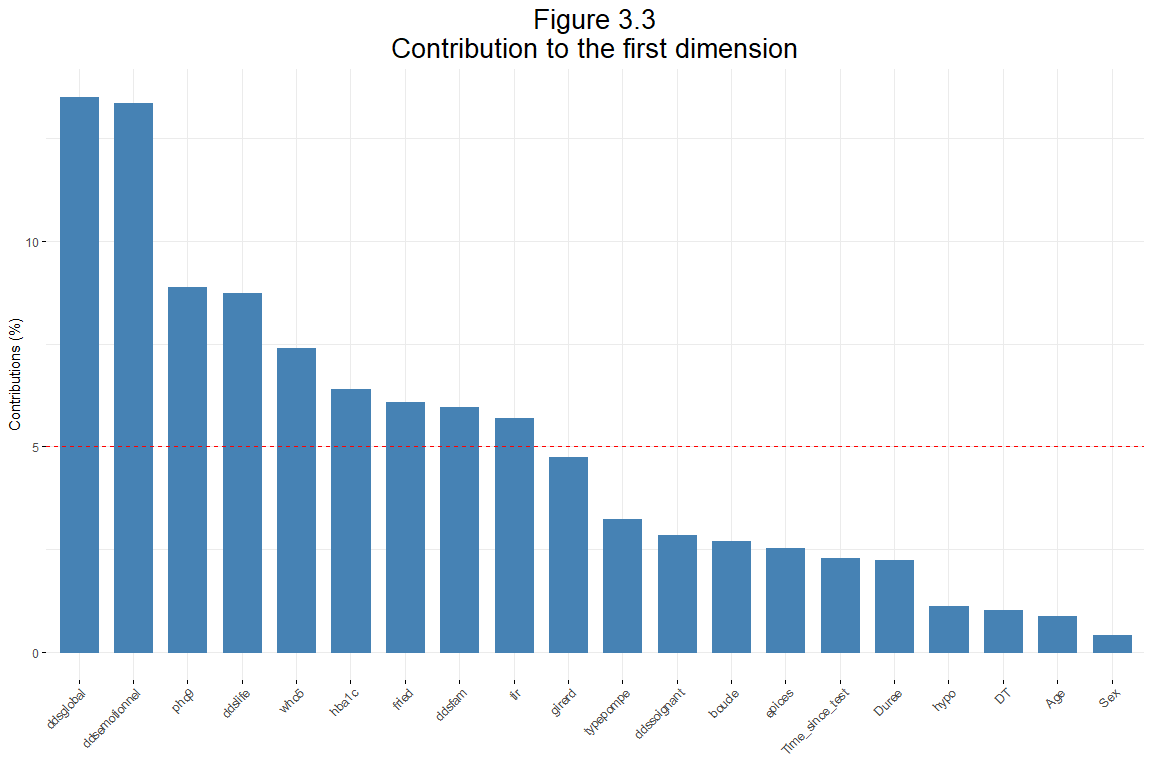
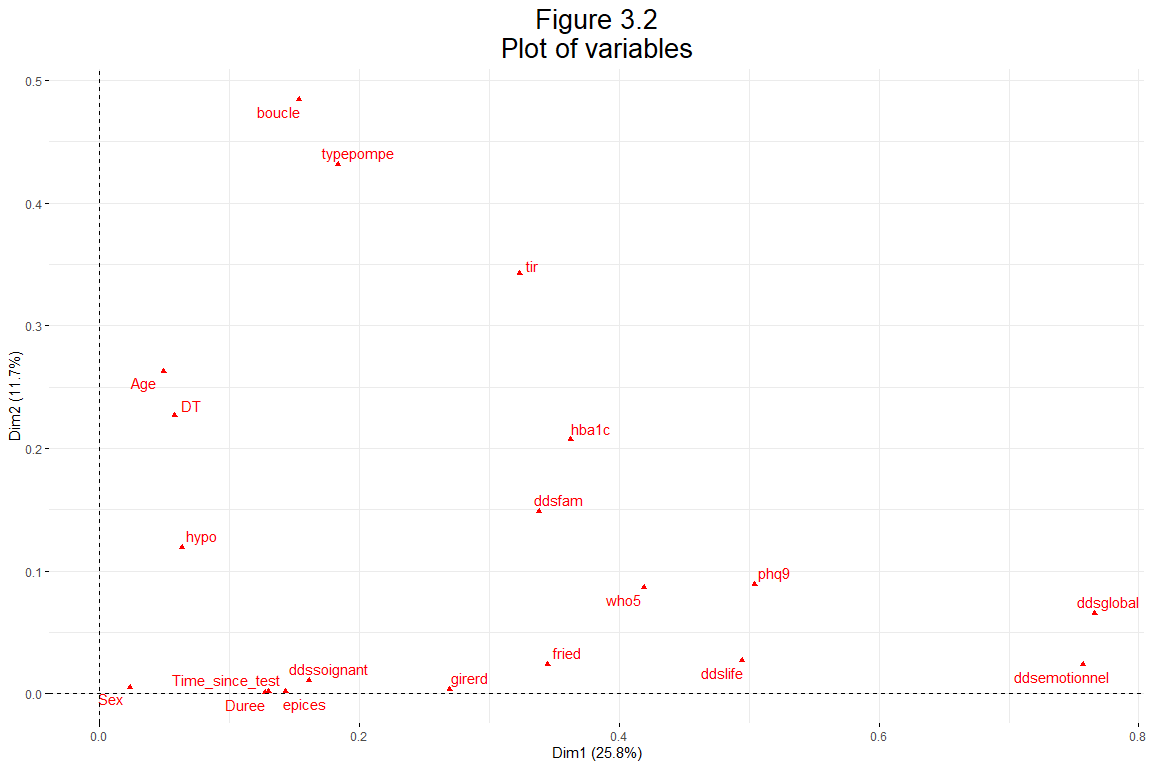


**Practical Implications**

Based on the FAMD analysis:

* **Key Variables:** The most contributing variables to Dim1 (ddsglobal, ddsemotionnel, phq9) and Dim2 (Time\_since\_test, Duree) are crucial for understanding the primary sources of variability in the data.
* **Focus Areas:** Variables with high contributions should be the focus of further analysis and interpretation to understand their impact and relationships within the dataset.
* **Dimensionality Reduction:** Variables closer to the origin with lower contributions can be considered for removal or further investigation if dimensionality reduction is needed, simplifying the analysis without significant loss of information.

This analysis helps to identify the most influential variables driving variability in the dataset, guiding targeted and informed decision-making.

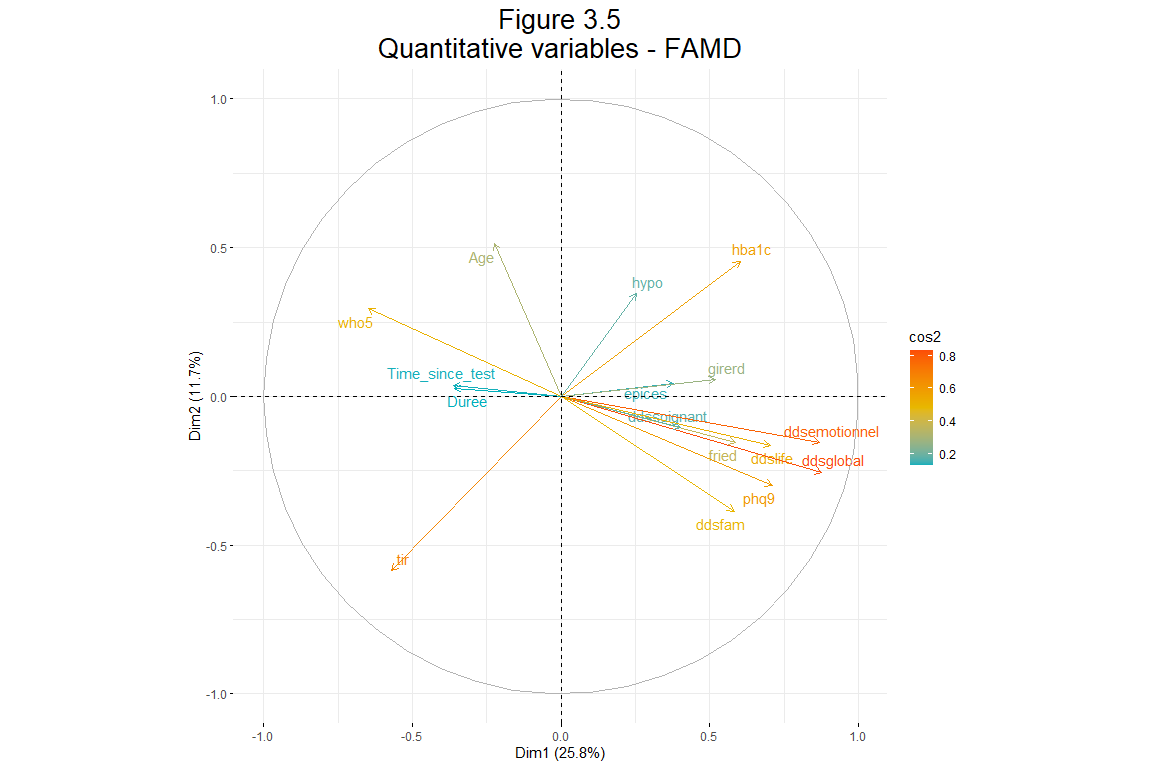


## 3.2 Quantitative variables visualization

**Quantitative Variables - FAMD (Figure 3.5)** The plot shows the distribution of quantitative variables across the first two principal components (Dim1 and Dim2). Variables such as ddsemotionnel, ddsglobal, and ddslife have strong contributions to Dim1, as indicated by their long vectors along the horizontal axis. Variables positioned closer to the origin have lower contributions to both dimensions, indicating less influence on the variability captured by these components.

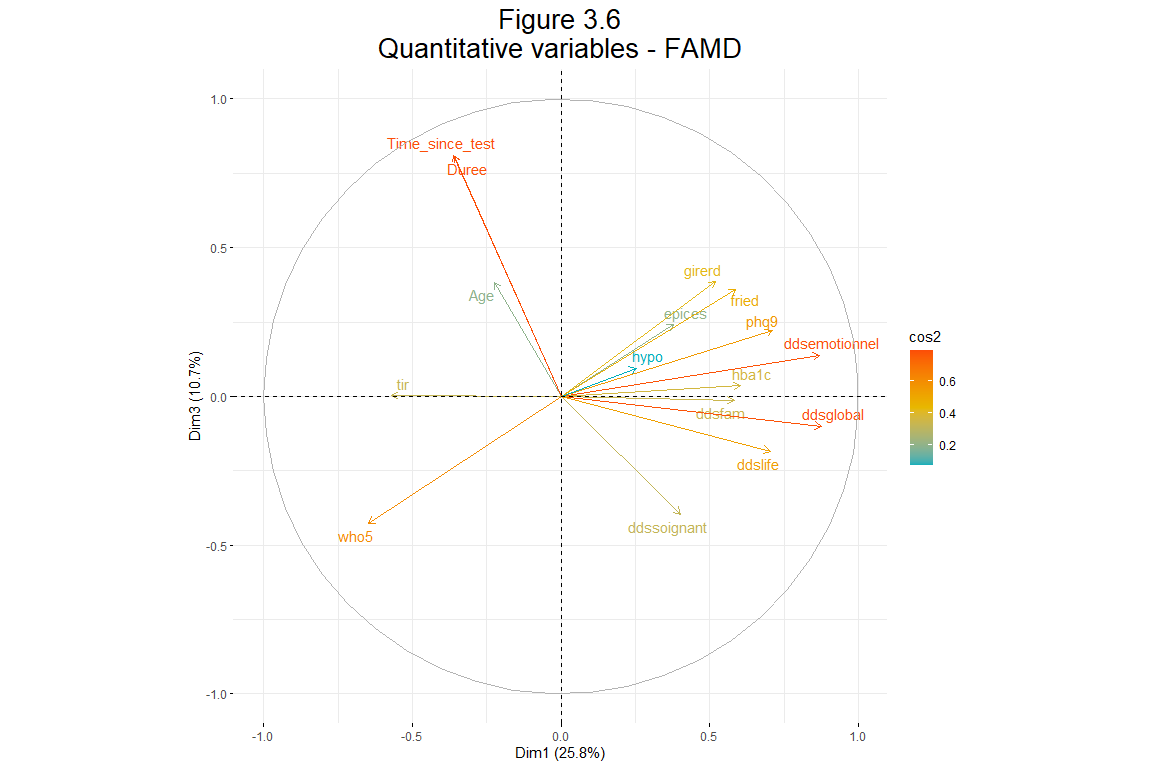
**Key Observations:**

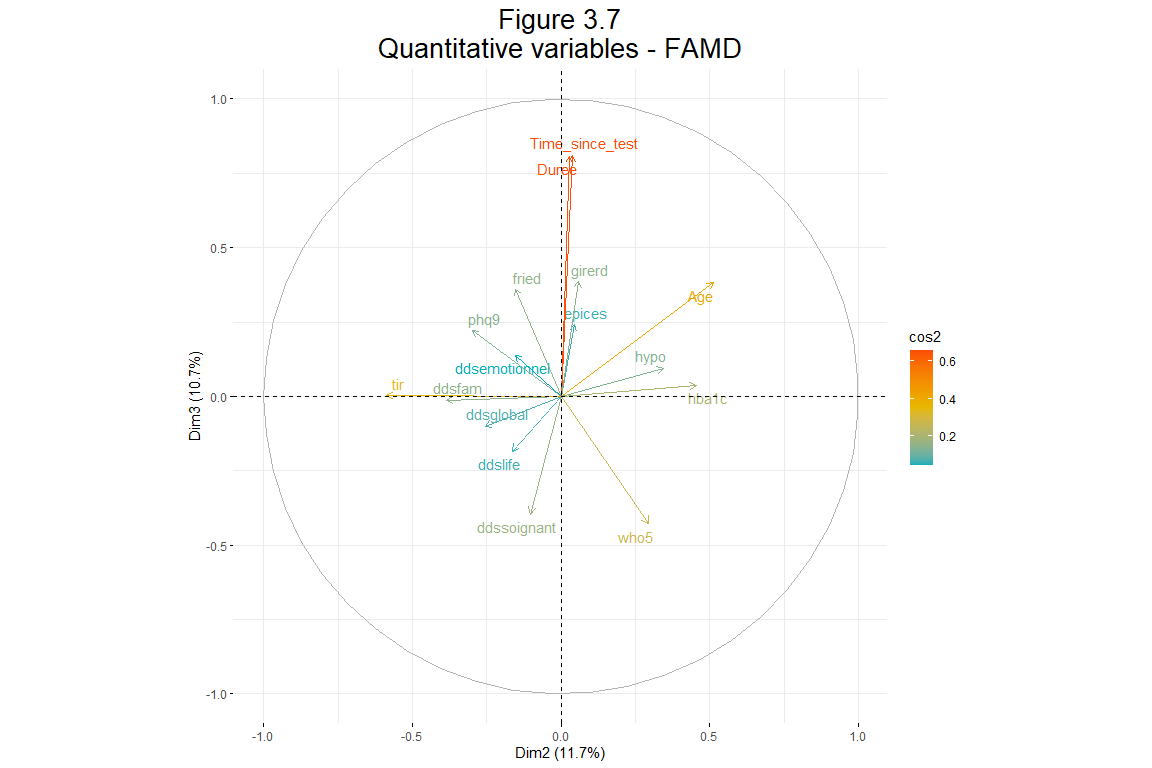
* **Dim1 Contributions:** Variables like ddsemotionnel, ddsglobal, and ddslife contribute significantly to Dim1, indicating these variables are major drivers of the variance in the first principal component.
* **Variable Relationships:** Variables pointing in similar directions are positively correlated. For example, ddsglobal and ddsemotionnel point in nearly the same direction, indicating a strong positive correlation.

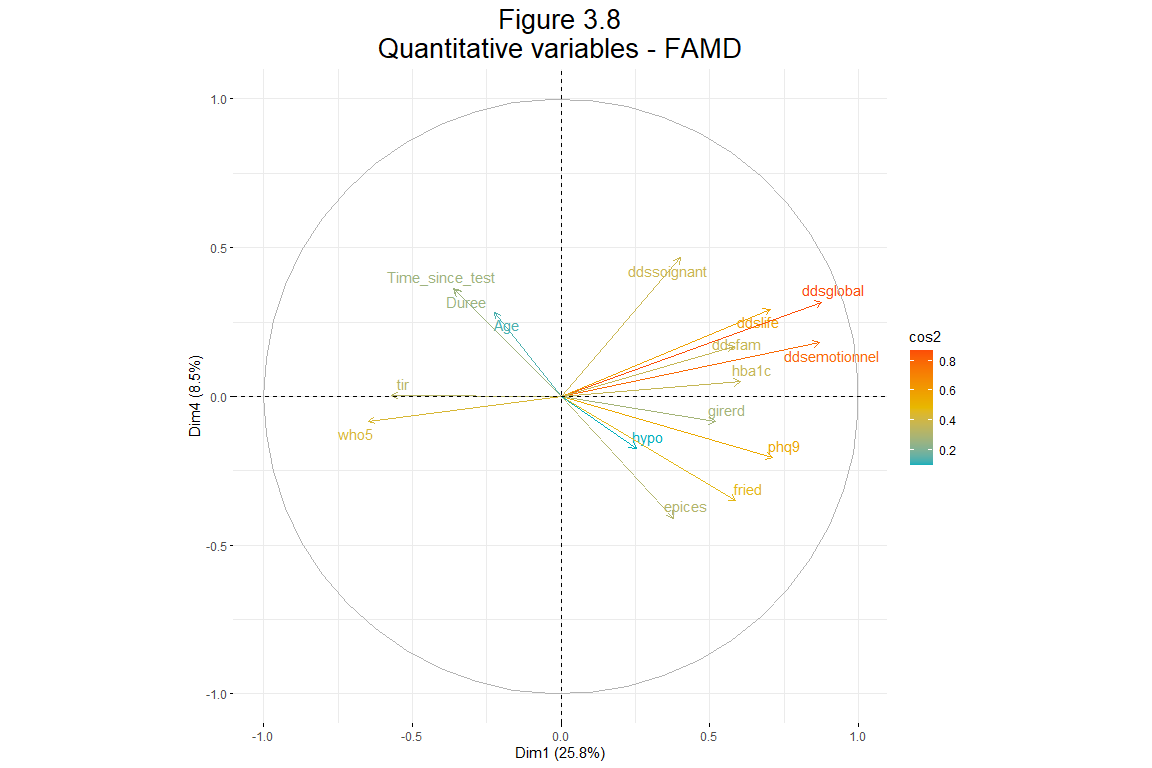


**Specific Observations**

* **Health Metrics**: Variables like hba1c, ddsemotionnel, and ddsglobal have high contributions to Dim1.
* **Well-being and Psychological Measures**: Variables such as phq9, ddsiflife, fried are grouped closely, indicating a relationship in this dimension.



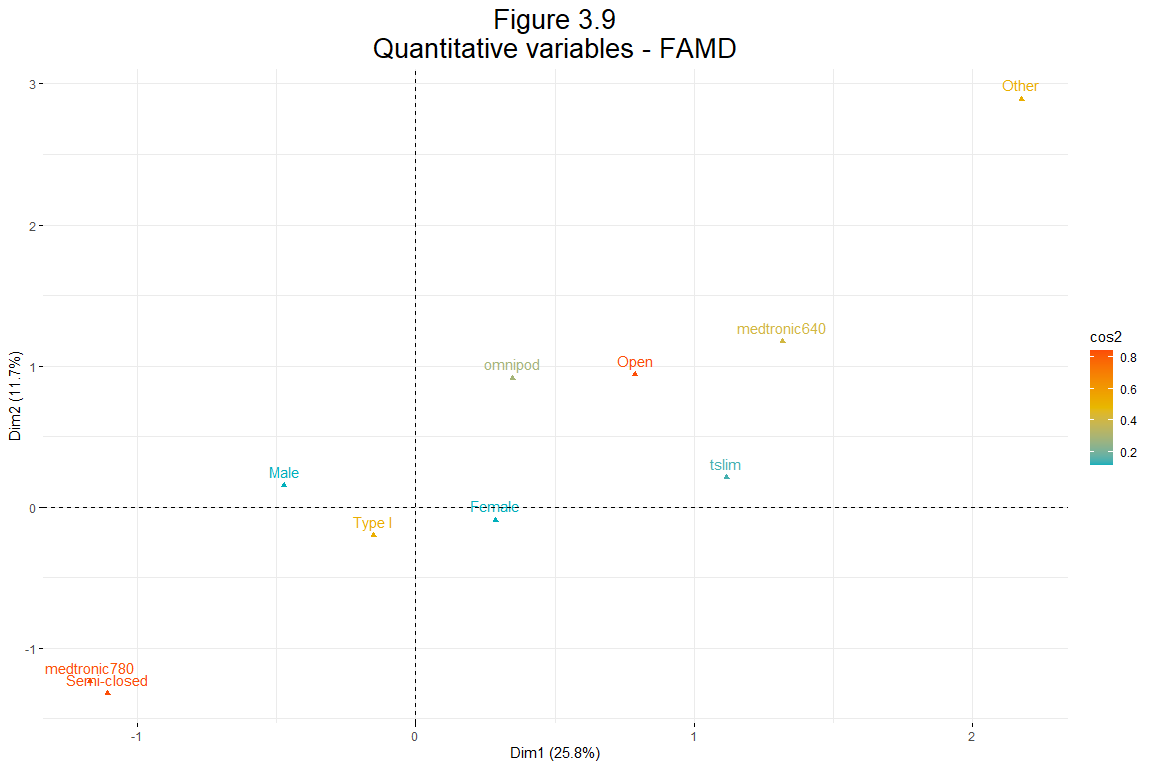




## 3.3 Quantitative variables visualization

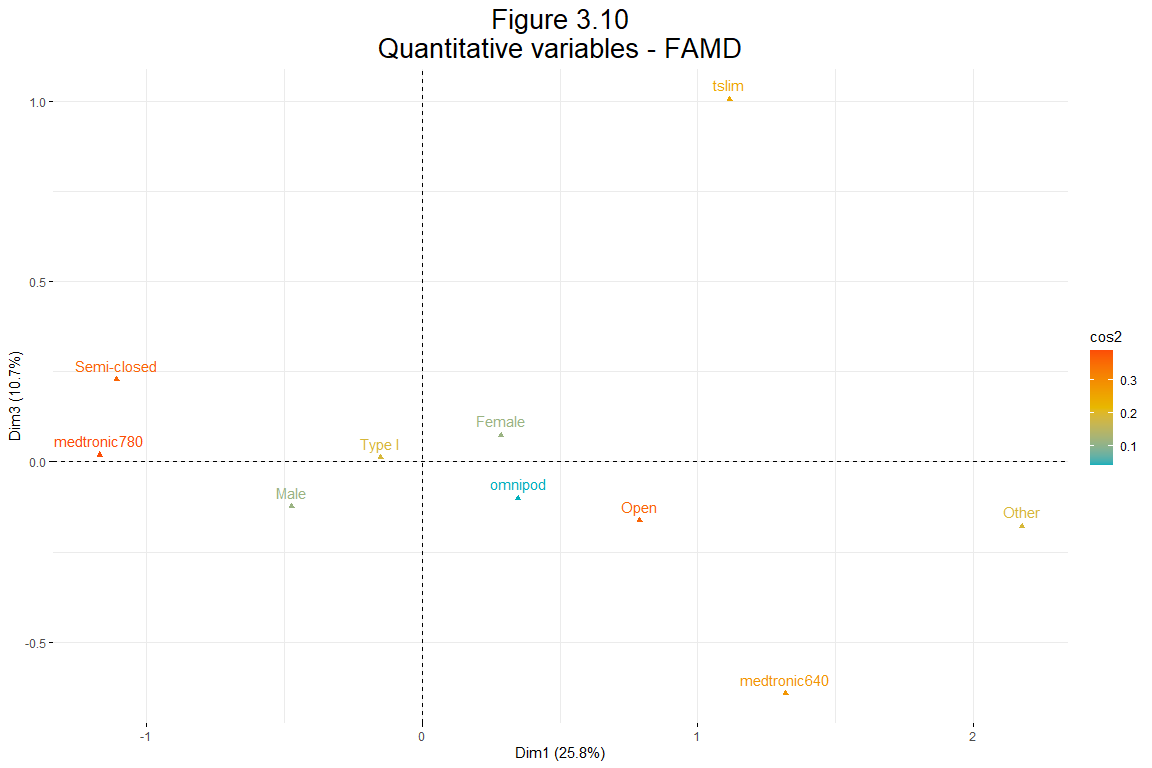
**Specific Observations**

* **Gender and Device Type**: Male and Female are well separated along Dim1, with different insulin pump types clustered around each gender.
* **Type I Diabetes**: The variable Type I is close to the center, suggesting it does not vary significantly along these dimensions.
* **Other**: This variable is positioned distinctly, indicating a specific influence on Dim2.



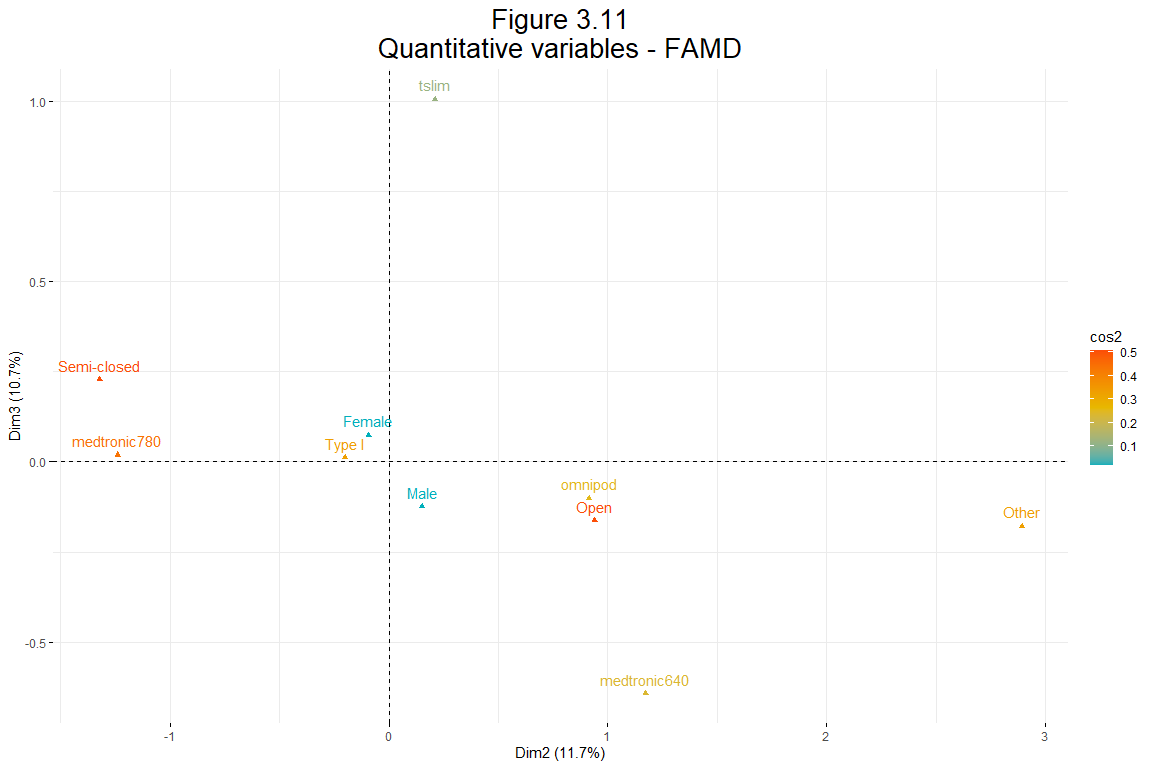
**Specific Observations**

* **Gender Distribution**: Male and Female are separated along Dim1, showing distinct patterns.
* **Open Systems**: The Open variable has a moderate contribution, positioned between other pump types.

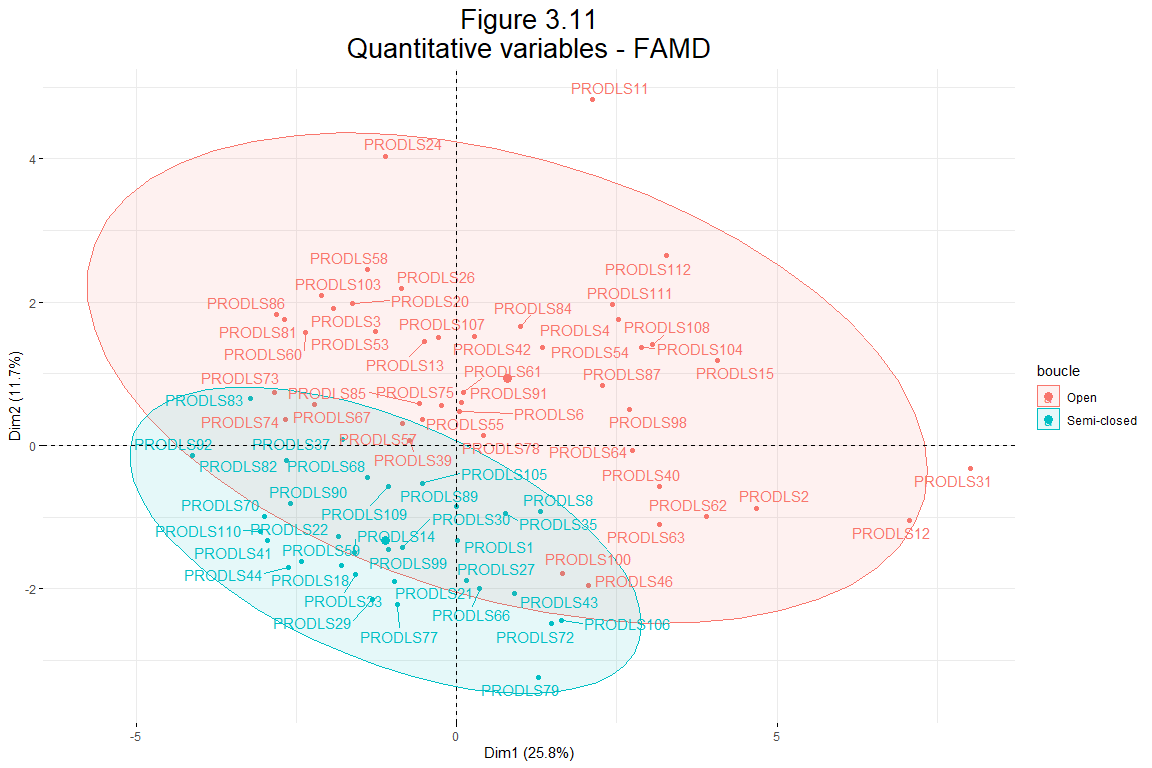


**Specific Observations**

* **Type I Diabetes**: This variable remains central, suggesting consistent representation.
* **Other Insulin Pumps**: The Other category shows a specific pattern along Dim3.



## 3.4 Plotting individuals



# 4 Building Statistical Model to fit our data

## 4.1 Binomial Logistic Regression Results (for some variables, boucle as dependent variabele)

**Interpretation of Coefficients**

The plot displays the log odds ratios (log(OR)) for different variables (hba1c, tir, and hypo) in relation to the type of insulin pump used. Here is a practical summary of the results:

**hba1c**

* The log(OR) is close to zero with a wide confidence interval that crosses zero, indicating no significant association between hba1c and the type of insulin pump used.

**tir**

* The log(OR) is negative with a narrow confidence interval, suggesting that higher tir values are associated with lower odds of using a specific insulin pump compared to the reference category. The confidence interval not crossing zero indicates that this result is statistically significant.

**hypo**

* The log(OR) is close to zero with a wide confidence interval that crosses zero, indicating no significant association between hypo and the type of insulin pump used.

**Practical Implications**

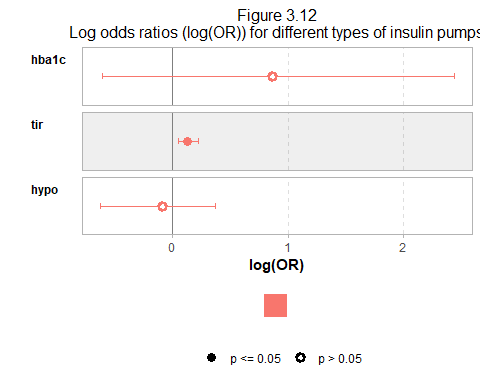
The results suggest that there is statistically significant associations tir with the type of insulin pump used when compared to the reference category. This implies that other variables may not be strong predictors of insulin pump type choice in this dataset.

**Notes**

* **Statistical Significance**: The p-values for all coefficients indicate that none of the associations are statistically significant at the 0.05 level.
* **Confidence Intervals**: Wide confidence intervals suggest variability in the estimates, and intervals crossing zero indicate non-significant effects.

Further research with a larger sample size or additional predictors may be necessary to identify significant predictors of insulin pump type.

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2



**Model Fit Statistics**

* Null deviance: 104.539 on 76 degrees of freedom
* Residual deviance: 77.929 on 73 degrees of freedom
* AIC: 85.929
* Number of Fisher Scoring iterations: 5

**Practical Conclusions**

1. **Intercept**:
   * The intercept estimate is -15.00690 with a p-value of 0.0634. This indicates that when all predictors are at zero, the log odds of the outcome occurring is -15.00690. However, this is not statistically significant at the 0.05 level.
2. **hba1c**:
   * The estimate for hba1c is 0.86449 with a standard error of 0.77200. The p-value is 0.2628, indicating that hba1c is not a statistically significant predictor of the outcome in this model at the 0.05 level. Thus, there is insufficient evidence to suggest that hba1c is associated with the outcome.
3. **tir**:
   * The estimate for tir is 0.12997 with a standard error of 0.04300. The p-value is 0.0025, which is highly significant (p < 0.01). This indicates that tir is a statistically significant predictor of the outcome. For every unit increase in tir, the log odds of the outcome “Semi-closed” occurring increases by 0.12997.
4. **hypo**:
   * The estimate for hypo is -0.08698 with a standard error of 0.24233. The p-value is 0.7196, indicating that hypo is not a statistically significant predictor of the outcome in this model. There is no sufficient evidence to suggest that hypo is associated with the outcome.

**Interpretation**

* The model suggests that among the predictors, only tir is a significant predictor of the outcome. This means that as tir increases, the likelihood of the “Semi-closed” outcome occurring increases.
* The non-significance of hba1c and hypo indicates that these variables do not have a statistically significant effect on the outcome in this dataset.
* The model fit statistics (deviance and AIC) provide a measure of how well the model fits the data, with lower values indicating a better fit.

Table 4.1  
Summary of Logistic regression

| Term | Estimate | Std. Error | z value | Pr(>|z|) |
| --- | --- | --- | --- | --- |
| character | numeric | numeric | numeric | character |
| (Intercept) | -15.01 | 8.08 | -1.86 | 0.0634 |
| hba1c | 0.86 | 0.77 | 1.12 | 0.2628 |
| tir | 0.13 | 0.04 | 3.02 | 0.0025 |
| hypo | -0.09 | 0.24 | -0.36 | 0.7196 |
| n: 4 | | | | |

## 4.2 Binomial Logistic Regression Results (for most of variables, boucle as dependent variabele)

**Interpretation:**

* Several predictors are significant, indicating they have a meaningful impact on the outcome variable.
* The model has reduced deviance, suggesting a good fit.

**Logistic Regression Model Adjustments**

* **Time\_since\_test**: The model performs better without this variable.
* **DT**: Removed due to high Variance Inflation Factor (VIF), indicating multicollinearity.
* **ddsemotionnel**, **ddssoignant**, **ddslife**, **ddsfam**: Eliminated due to its high correlation with **dsglobal**.
* **Sex**: Excluded from the model as it was not a significant predictor.

Table 4.1  
Summary of Logistic regression

| Term | Estimate | Std. Error | z value | Pr(>|z|) |
| --- | --- | --- | --- | --- |
| character | numeric | numeric | numeric | character |
| (Intercept) | -1.1780 | 0.4828 | -2.4402 | 0.0147 |
| Age | -1.5483 | 0.5091 | -3.0412 | 0.0024 |
| Duree | 1.0247 | 0.4416 | 2.3205 | 0.0203 |
| who5 | -0.1079 | 0.6388 | -0.1689 | 0.8659 |
| phq9 | -0.7359 | 0.6424 | -1.1456 | 0.2520 |
| epices | -0.5328 | 0.4376 | -1.2173 | 0.2235 |
| girerd | 0.4502 | 0.4113 | 1.0945 | 0.2738 |
| fried | 1.0457 | 0.5275 | 1.9824 | 0.0474 |
| ddsglobal | -1.1839 | 0.8352 | -1.4174 | 0.1564 |
| hba1c | 2.3153 | 1.0887 | 2.1266 | 0.0335 |
| tir | 4.3423 | 1.3256 | 3.2756 | 0.0011 |
| hypo | 0.2639 | 0.4349 | 0.6067 | 0.5441 |
| n: 12 | | | | |

