Scientific Visualization (2025/26)
Assignment 2

PDAC DATA INTEGRATION USING MISS-SNF

integrating multi-omics datasets for pancreatic ductal adenocarcinoma

SARA MANCINI
66458A

BETÜL GÜL
V13000

Index

What is Miss-SNF?

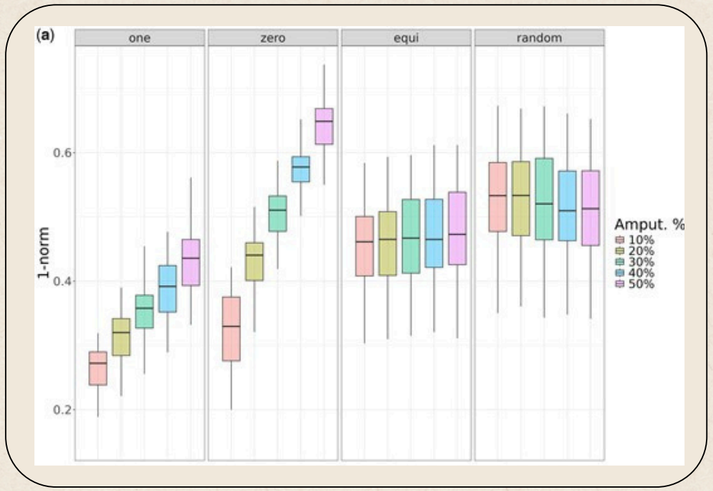
https://github.com/AnacletoLAB/missSNF

Miss-SNF extends the original Similarity Network Fusion (SNF; Wang et al., 2014), so instead of discarding patients missing large fractions of features or performing global imputation, patients with too many missing values are identified and treated as **missing samples**, that are handled during similarity-network construction using one of several strategies.

[1]

Example of comparison of miss-SNF strategies (one, zero, equidistant, and random);

/https://doi.org/10.1093/bioinformatics/btaf150



MAIN FUNCTION (miss.snf)

- Takes in input a list of matrices (which may contain NA values), one per omics layer. Rows = patients, Columns = features.
- Imputes partially missing patients (optional).
- Detects missing patients in each matrix based on a per-patient missingness threshold (*perc.na*) and removes patients that are missing in all matrices.
- Constructs Per-Matrix Affinity Matrices: for each matrix, computes similarity using *sims[i]* (Typical values*: "scaled.exp.euclidean"* – continuous data; *"scaled.exp.chi2"* – binary/categorical data), converts similarity to affinity using the K-nearest-neighbor graph. Modifies affinities for missing patients according to mode:

Reconstruct (One)	Reconstruction of missing patients during fusion
Ignore (Zero)	Missing patients ignored for that layer
Equidistant	Missing patients assigned uniform similarity to all others
Random	Assign random similarities

- Performs an enhanced SNF cross-diffusion that preserves missing-data structure.

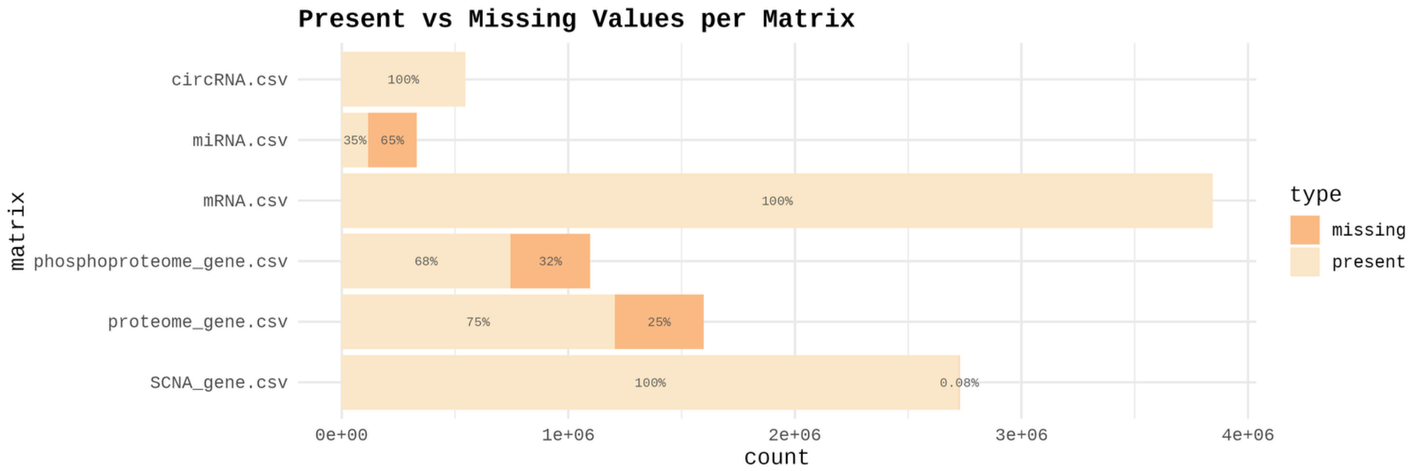
Chosen Dataset

PDAC_data.csv

Amount of Patients: **137** Total Matrices: 7 Target Matrix: clinical_data.csv

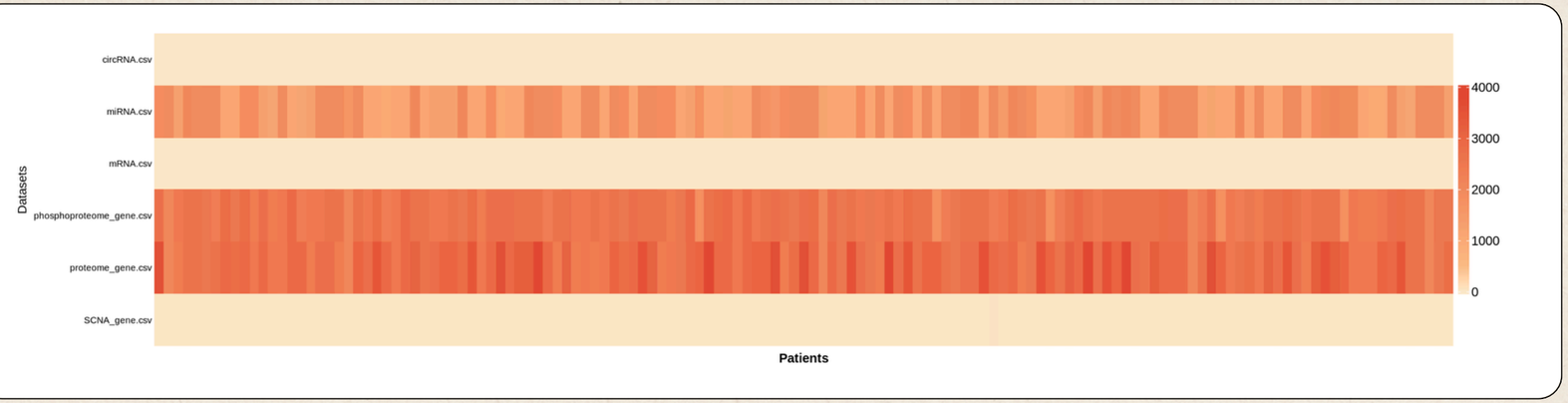
Omics matrices:

[2]



Patient missing values amount per matrix:

[3]



HELPER FUNCTIONS

get.miss.pts

Identifies which patients are considered “missing” for each matrix, based on the threshold *perc.na*.

perc_na = **0.2**

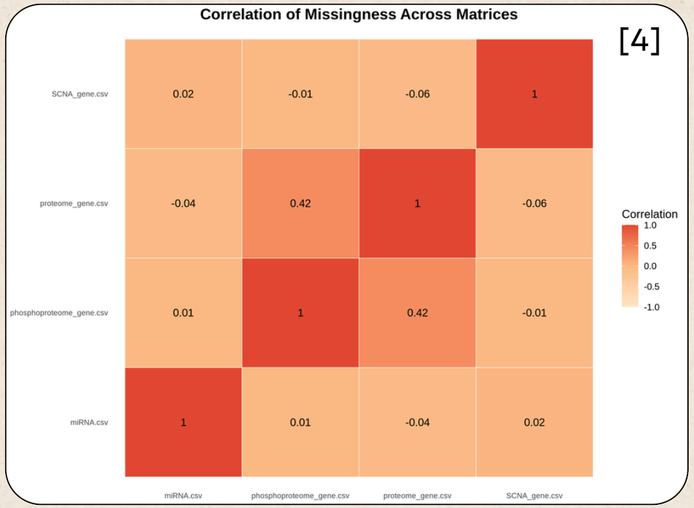
Amount of Missing Patients per Matrix:

circRNA.csv	0
miRNA.csv	137
mRNA.csv	0
phosphoproteome_gene.csv	137
proteome_gene.csv	132
SCNA_gene.csv	0

↓

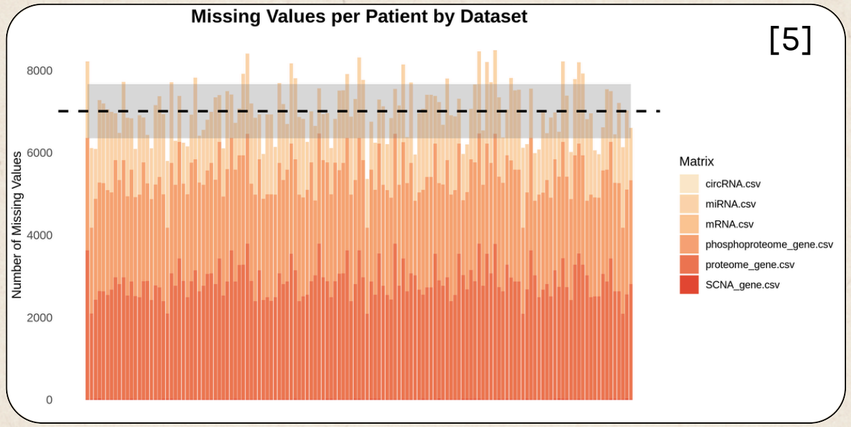
Most patients are considered missing in 3/6 datsets

Why are patients missing?



[4]

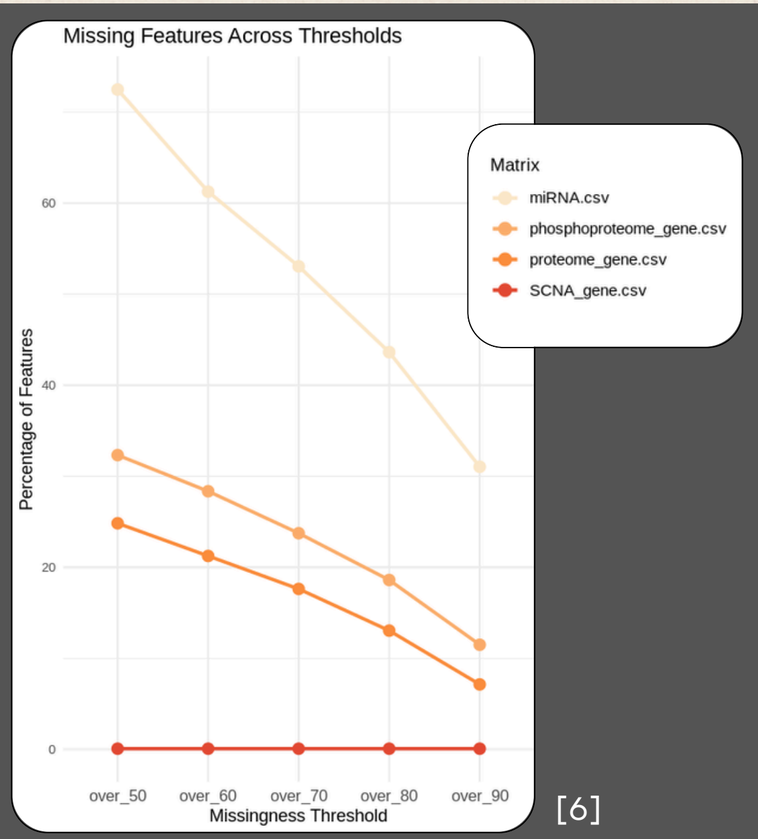
- miRNA.csv and SCNA_gene.csv: **very low correlations** with all other matrices, this suggests that missing values are mostly **independent** from missingness in the other datasets.
- phosphoproteome_gene.csv and proteome_gene.csv show a **moderate correlation** (0.416)



[5]

Mean fraction missing per patient	7010.467
Stdev fraction missing per patient	658.056
Variability % of missing values across patients	**9.39 %**

INTERPRETATION Missingness is fairly consistent across patients, with mostly very low correlation across datasets, suggesting that values are largely missing at random: the main issue lies in **features** - most missingness is concentrated in a limited subset, rather than being widespread.



[6]

Approach used: Feature Removal

Chosen threshold: **> 80%** of NAs

↓

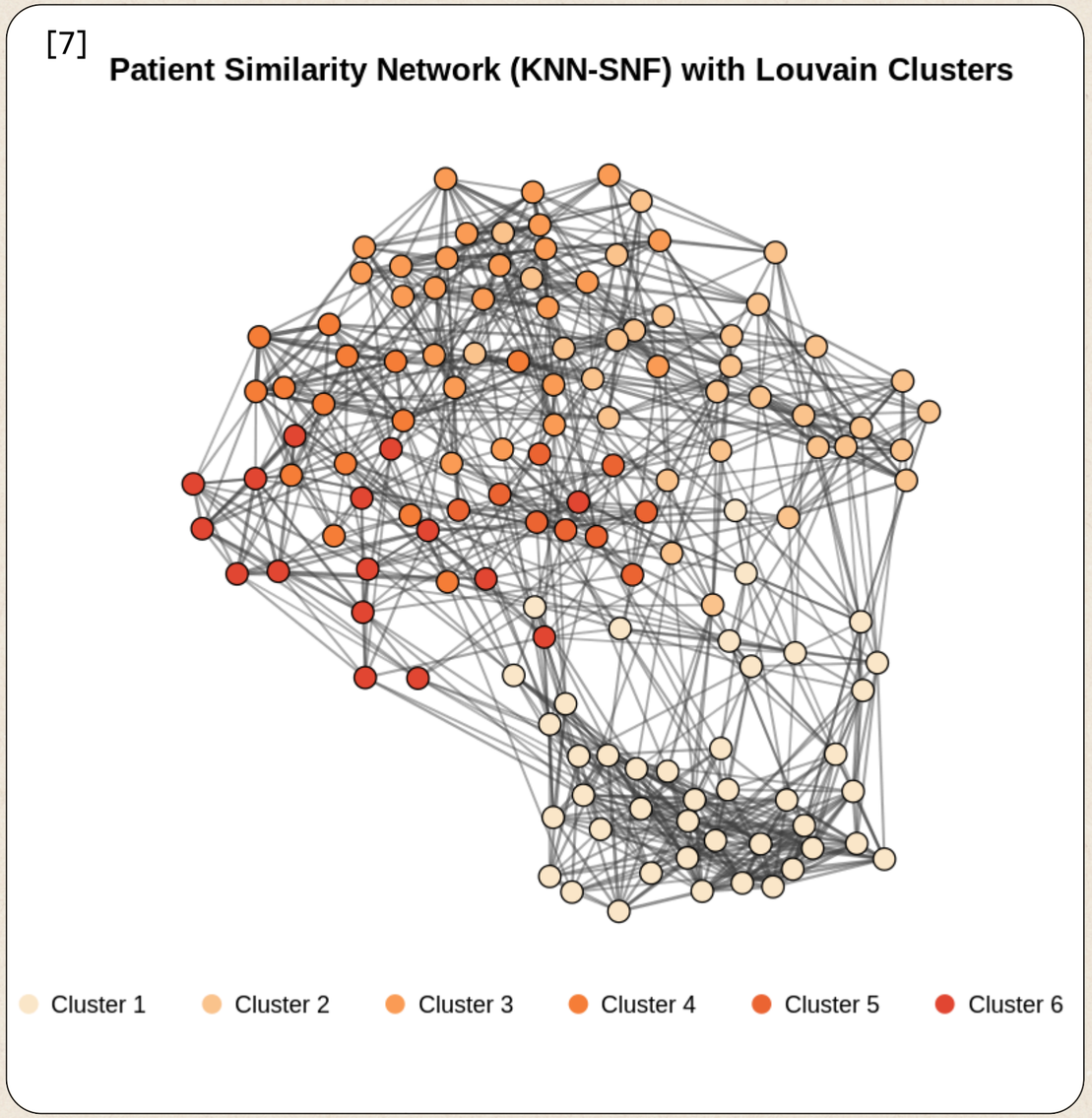
New get.miss.pts results:

miRNA.csv	80
phosphoproteome_gene.csv	34
proteome_gene.csv	16

Miss-SNF Results

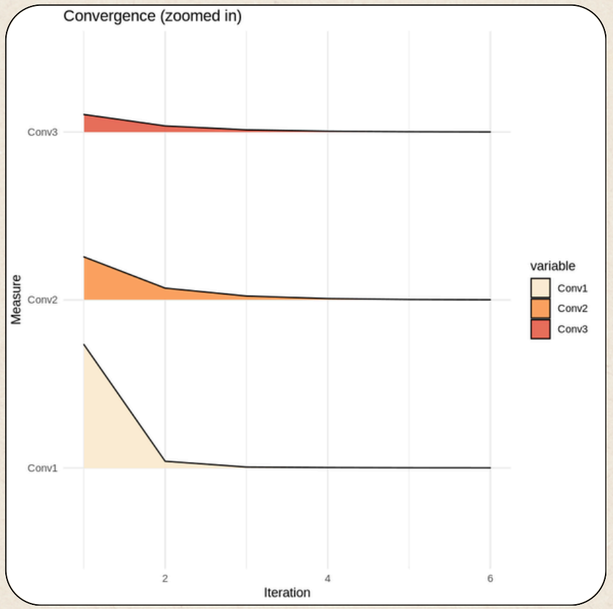
mode: reconstruct, **perc.na** = 0.2, **impute**: median

[7]



Number of patients per cluster:

Cluster 1	43	Cluster 4	14
Cluster 2	31	Cluster 5	9
Cluster 3	24	Cluster 6	16



[8] **CONVERGENCE CURVES**

The model stabilized quickly, reaching convergence only after 4 iterations.

- **K-nearest-neighbors** (KNN) graph was constructed by keeping, for each patient, only the top-K (k=10) strongest similarity connections and setting all weaker links to zero.
- The matrix was then symmetrized and the **Louvain** community detection algorithm was applied for clustering.

Modularity: 0.5587

Mean silhouette: 0.0047

Numeric Variables Validation

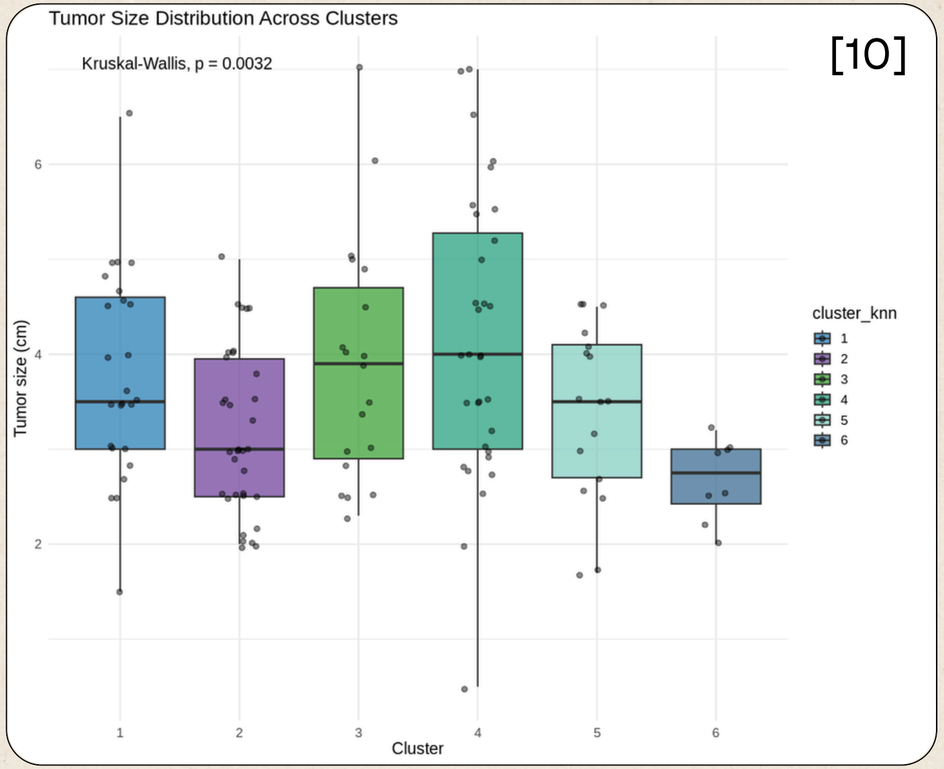
variable	KW H statistic	df	p_value	p_fdr	significant_0_05	significant_0_05_fdr
tumor_size_cm	17.801413	5	0.003205869	0.01282347	YES	YES
bmi	7.982472	5	0.157204295	0.23174317	NO	NO
follow_up_days	7.695955	5	0.173807375	0.23174317	NO	NO
age	5.089488	5	0.405056963	0.40505696	NO	NO

Upon calculating CHI-squared metrics tumor size was found to be the most significant: pairwise similarities [9], distributions across cluster [10] were calculated.

[9]

cluster 1	cluster 2	p_raw	p_adj	sig_0_05_raw	sig_0_05_fdr
1	6	0.0053	0.0264	YES	YES
2	4	0.0032	0.0264	YES	YES
4	6	0.0037	0.0264	YES	YES
3	6	0.0150	0.0564	YES	NO
1	2	0.0207	0.0620	YES	NO
5	6	0.0377	0.0943	YES	NO
2	3	0.0525	0.1124	NO	NO
4	5	0.0957	0.1794	NO	NO
2	6	0.1798	0.2997	NO	NO
1	5	0.2735	0.4103	NO	NO
2	5	0.3150	0.4296	NO	NO
1	4	0.4012	0.4774	NO	NO
3	4	0.4232	0.4774	NO	NO
3	5	0.4456	0.4774	NO	NO
1	3	0.9621	0.9621	NO	NO

[10]



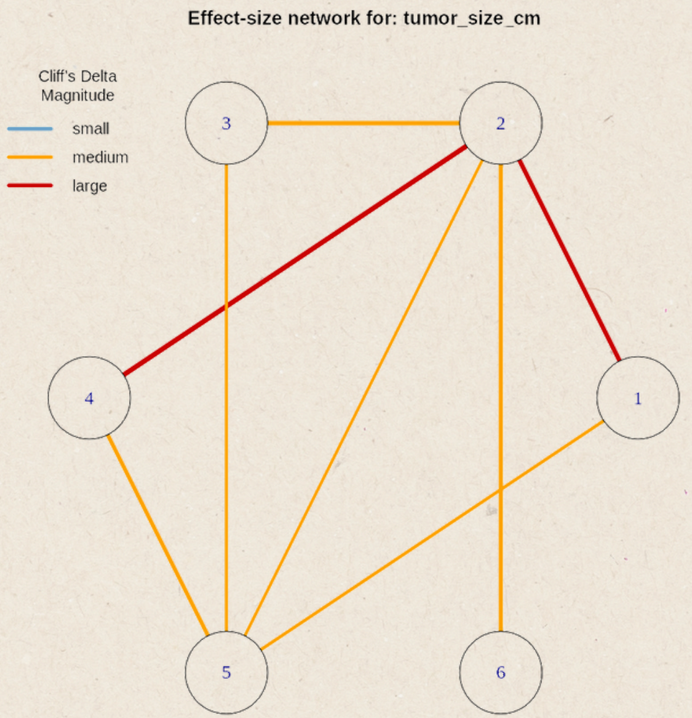
Numeric Variables Validation

Cliff's delta was also calculated for magnitudes, scores for the age variable are illustrated in table [12] and plotted in effect-size networks for each numeric variable.

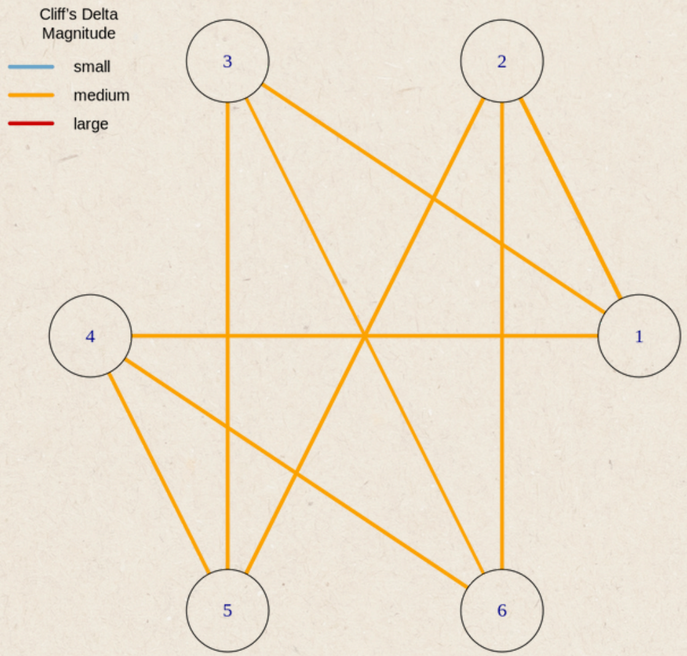
[12]

variable	cluster1	cluster2	delta	magnitude
age	1	2	-0.17764706	small
age	1	3	-0.03809524	negligible
age	1	4	-0.18000000	small
age	1	5	-0.34352941	medium
age	1	6	-0.36000000	medium
age	2	3	0.10224090	negligible

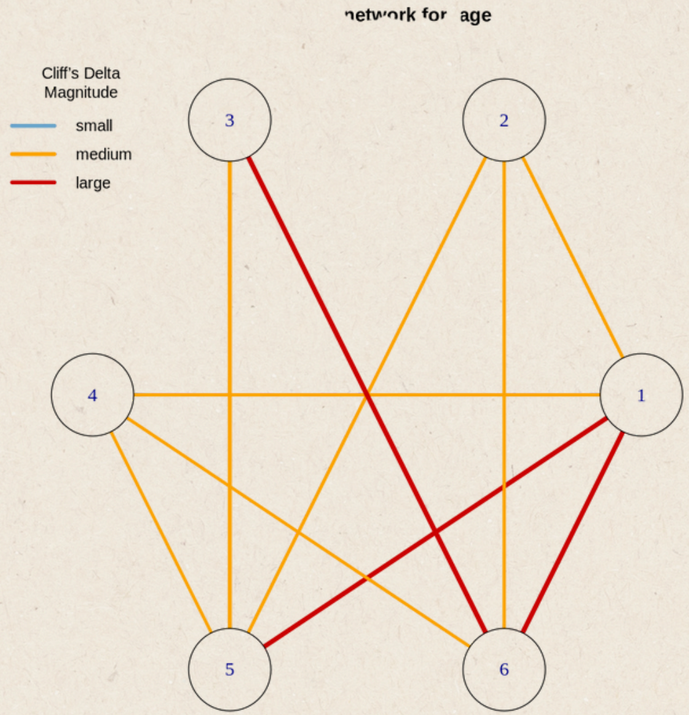
tumor_size_cm



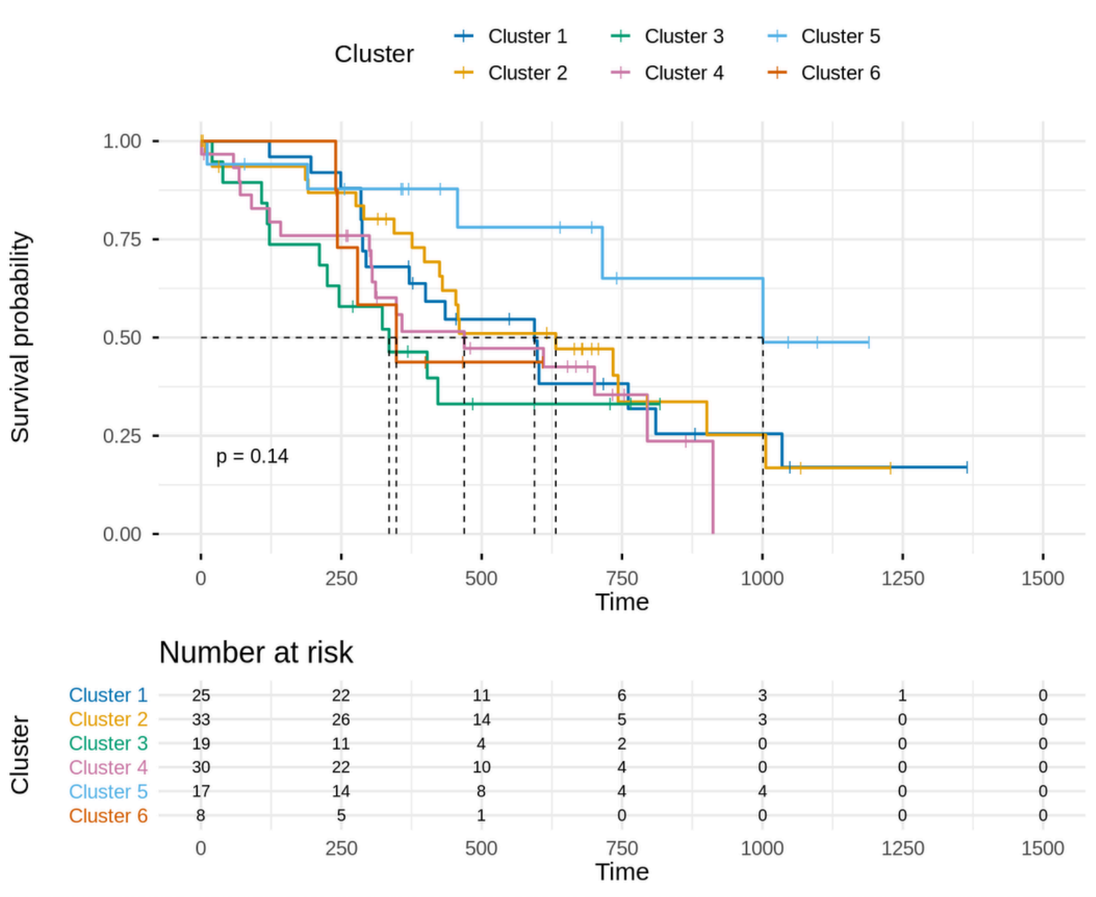
BMI

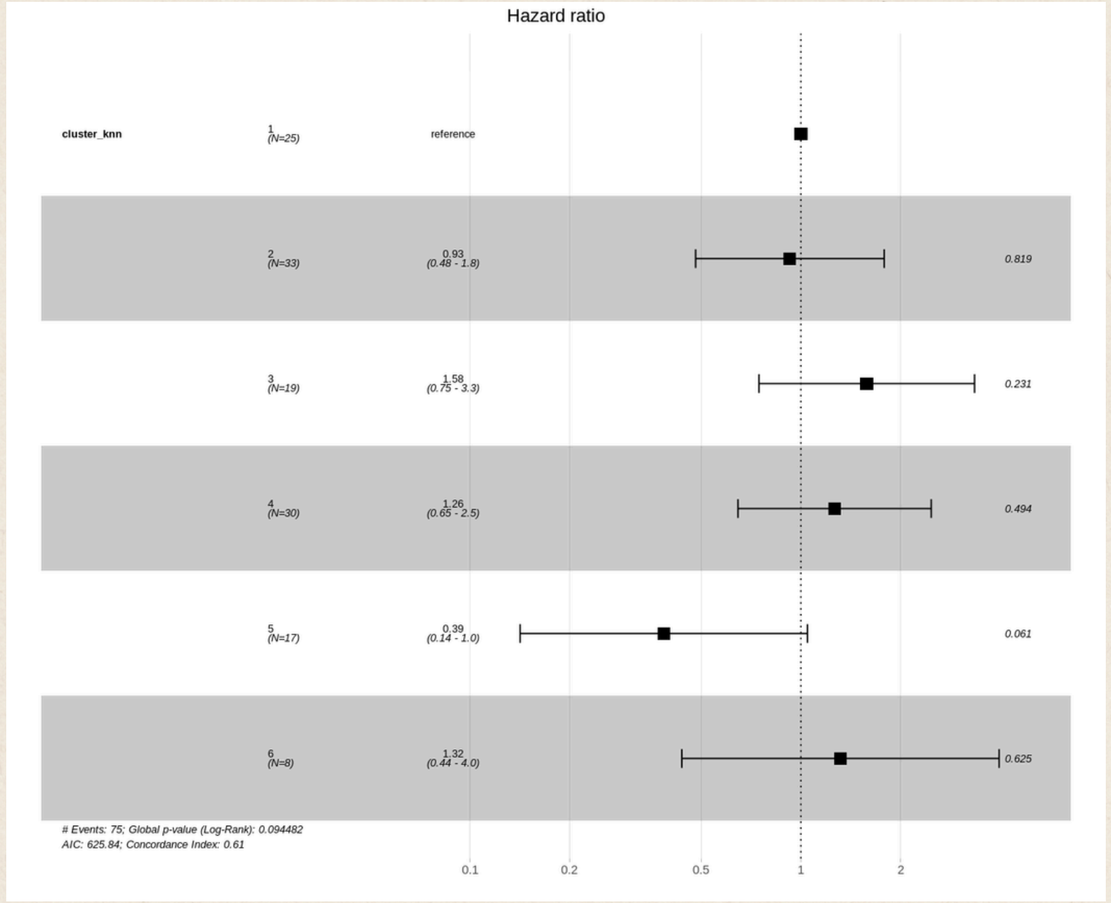


age



Survival Analysis





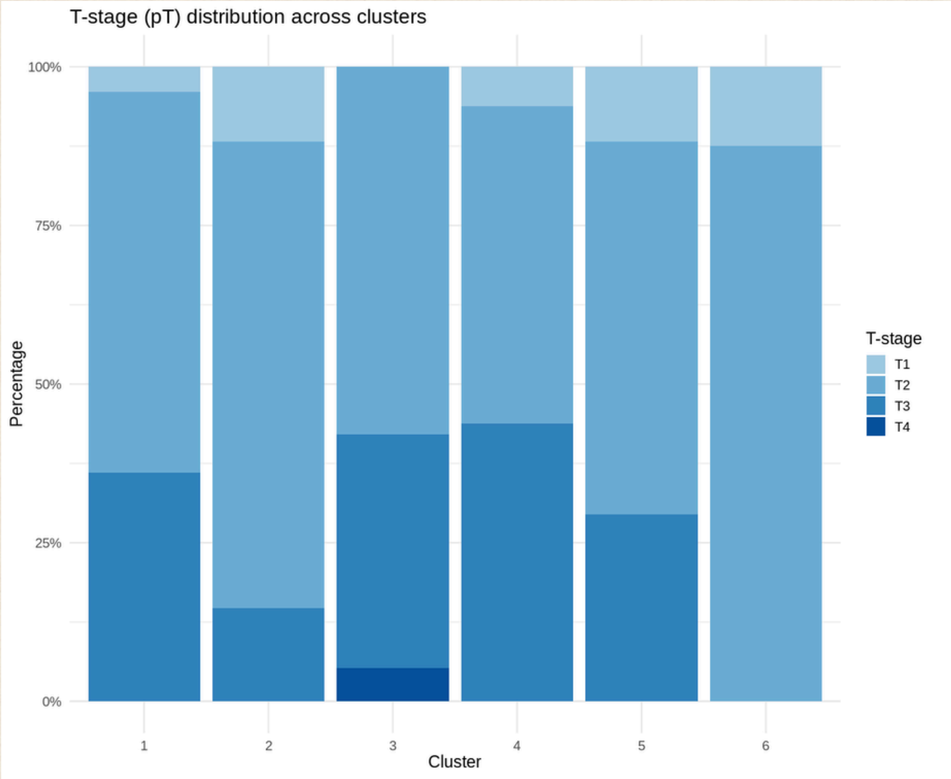
Ordinal Variables Validation

variable	chi_squared	df	p_value	p_fdr	significant_0_05_raw	significant_0_05_fdr
pn_factor	4.248242	5	0.5142565	0.5142565	NO	NO
pt_factor	12.694393	5	0.0264173	0.1320865	YES	NO
pm_factor	8.035216	5	0.1543055	0.2915123	NO	NO
cm_factor	7.677835	5	0.1749074	0.2915123	NO	NO
residual_tumor_ord	5.873171	5	0.3187563	0.3984453	NO	NO

Only pT stage shows a noticeable cluster difference (raw p = 0.026), but it's not significant after FDR, so the signal is weak.

Other ordinal variables (pN, pM, cM, residual tumor) show no significant differences across clusters.

Ordinal clinical variables provide limited cluster separation, with pT being the only mild signal.



Nominal Variables Validation

variable	test type	statistic	p value	significant_0_05
sex	Chi-square (simulated)	5.024842	0.43175682	NO
race	Chi-square (simulated)	NaN	NaN	NA
tumor_site	Chi-square (simulated)	15.435412	0.78322168	NO
tumor_necrosis	Chi-square (simulated)	10.060786	0.07019298	NO
lymph_vascular_invasion	Chi-square (simulated)	6.810897	0.75572443	NO
perineural_invasion	Chi-square (simulated)	26.735504	0.00459954	YES
vital_status	Chi-square (simulated)	7.209204	0.21627837	NO
cause_of_death	Chi-square (simulated)	49.110468	0.20067993	NO
tumor_stage_pathological	Chi-square (simulated)	30.962613	0.19148085	NO
alcohol_consumption	Chi-square (simulated)	23.894518	0.23807619	NO
tobacco_smoking_history	Chi-square (simulated)	35.663814	0.07039296	NO
is_this_patient_lost_to_follow_up	Chi-square (simulated)	6.362956	0.27447255	NO

We tested whether nominal clinical variables differed across clusters.
Some variables had too many small categories, so first we simplified them into clinically meaningful binary groups – for example, alcohol use, tumor site and smoking status.

Nominal Variables Validation

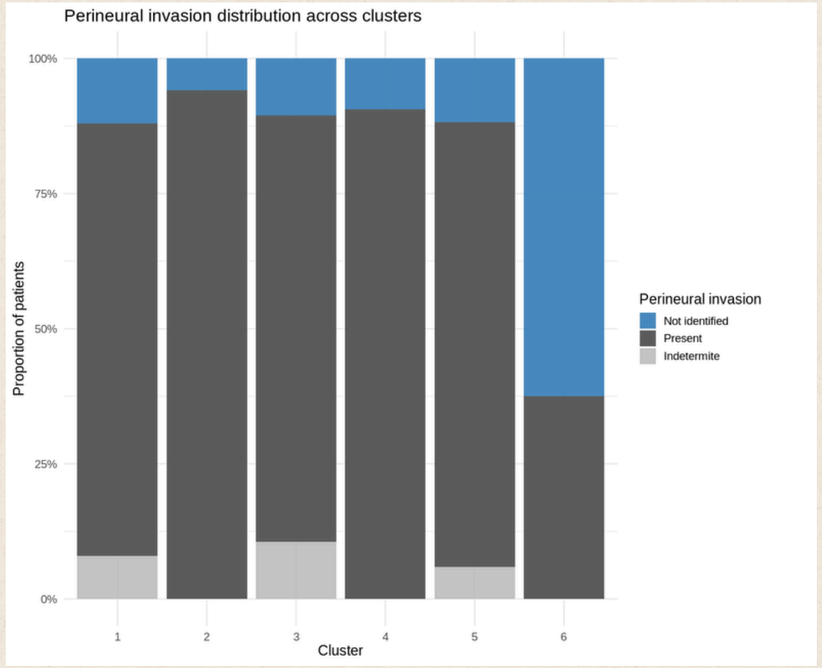
variable	cluster1	cluster2	p_value
perineural_invasion	1	2	0.193707281
perineural_invasion	1	3	1
perineural_invasion	1	4	0.230363369
perineural_invasion	1	5	1
perineural_invasion	1	6	0.018314905
perineural_invasion	2	3	0.085632535
perineural_invasion	2	4	0.667887668
perineural_invasion	2	5	0.266602982
perineural_invasion	2	6	0.001200074
perineural_invasion	3	4	0.257639337
perineural_invasion	3	5	1
perineural_invasion	3	6	0.013803588
perineural_invasion	4	5	0.562837497
perineural_invasion	4	6	0.003795673
perineural_invasion	5	6	0.016850426

variable	test_type	statistic	p_value	significant_0_05
tumor_site_bin	Fisher exact	NA	0.1776142	NO
alcohol_bin	Chi-square (simulated)	15.35874	0.1149885	NO
death_bin	Fisher exact	NA	0.3055991	NO
smoking_bin	Chi-square (simulated)	23.54818	0.0079992	YES

After this recoding, only smoking showed a significant association with the clusters (p = 0.007). Since smoking was the only nominal feature with a meaningful signal, we visualized it separately to show how its distribution varies across clusters.

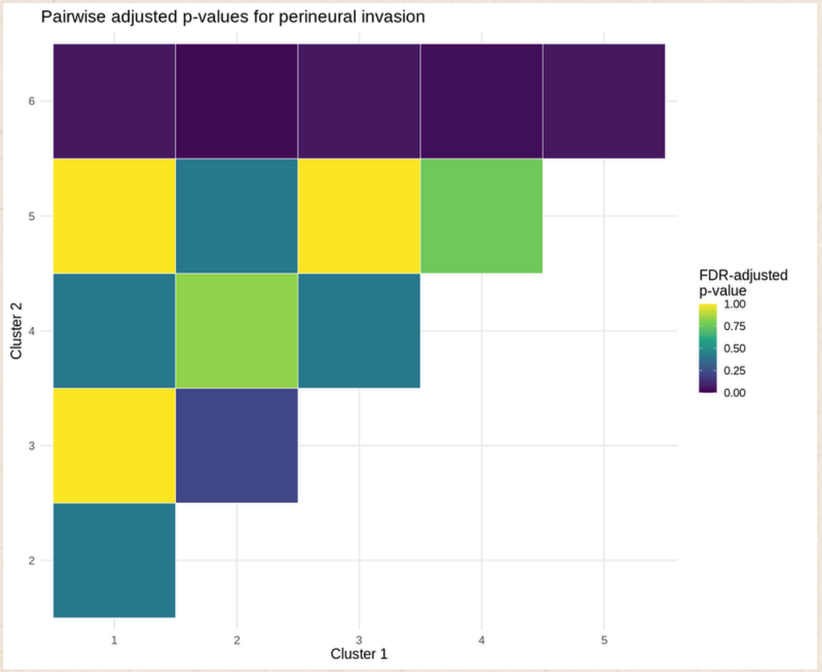
variable	cluster1	cluster2	p_value
smoking_bin	1	2	0.02856001
smoking_bin	1	3	0.05181567
smoking_bin	1	4	0.0629617
smoking_bin	1	5	0.58500067
smoking_bin	1	6	0.32236284
smoking_bin	2	3	0.90697476
smoking_bin	2	4	0.84761744
smoking_bin	2	5	0.01280069
smoking_bin	2	6	0.36222502
smoking_bin	3	4	0.43416263
smoking_bin	3	5	0.04115867
smoking_bin	3	6	0.17300091
smoking_bin	4	5	0.01807009
smoking_bin	4	6	0.64152867
smoking_bin	5	6	0.12868178

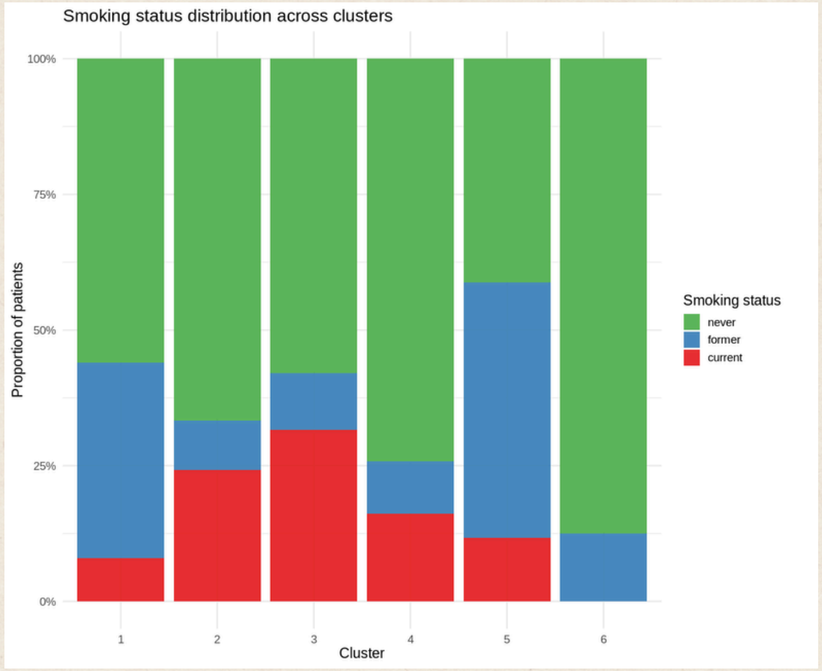
Nominal Variables Validation



Perineural invasion rates differ across clusters, with Cluster 6 showing noticeably higher ‘Not identified’ cases. This aligns with the global and pairwise results, where Cluster 6 significantly diverges from several other clusters.

Pairwise tests highlight that Cluster 6 differs from multiple clusters (FDR-adjusted p < 0.05), supporting the idea that this cluster may represent a clinically distinct subtype.

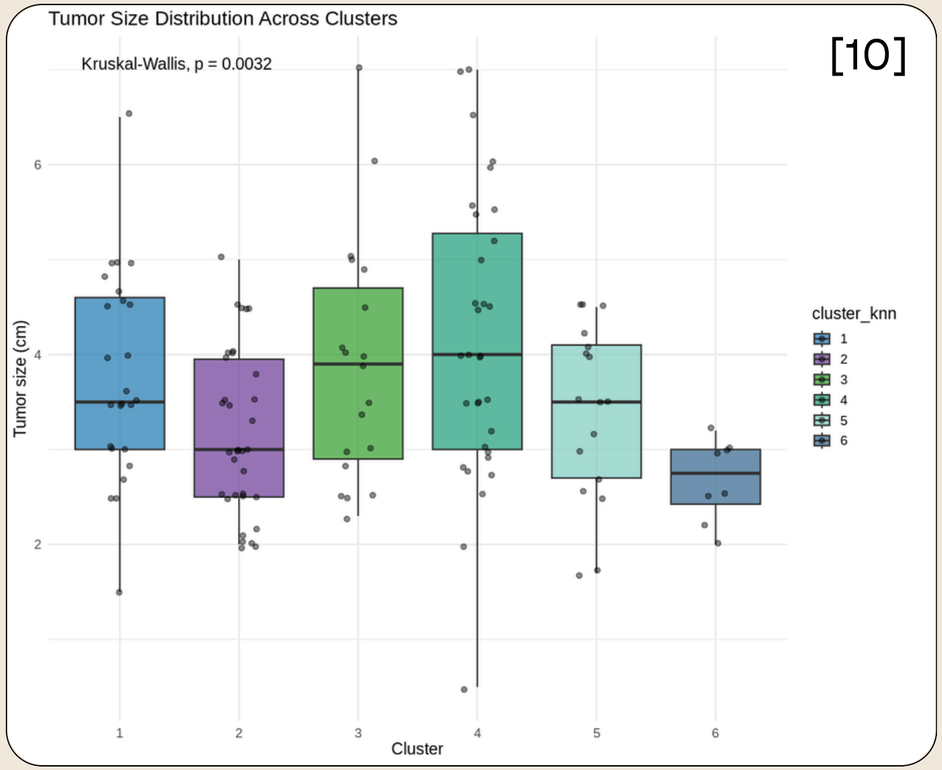


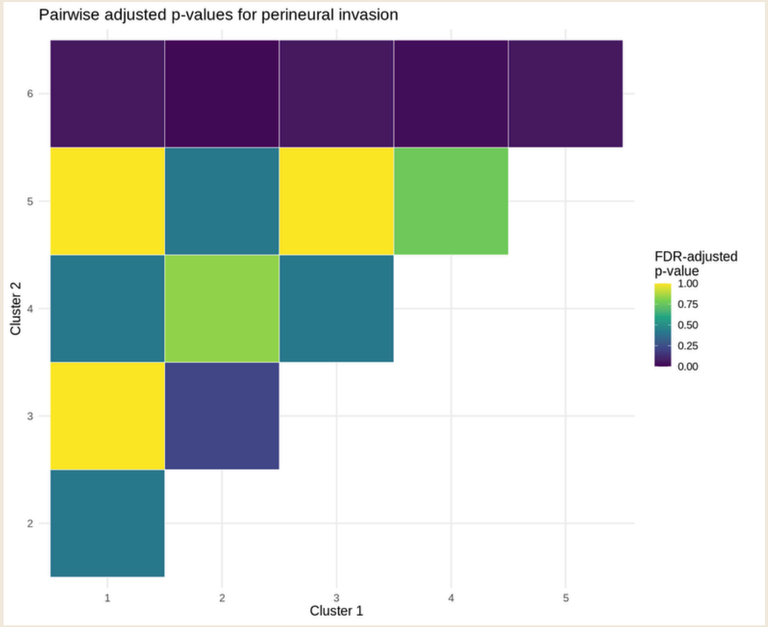


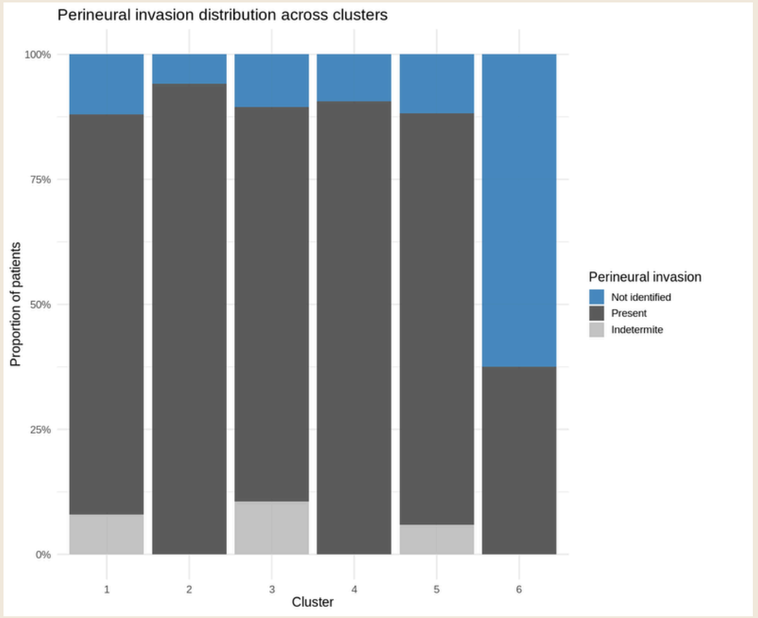
Smoking status also varies across clusters, especially between Cluster 1–2 and 2–5, though differences are smaller and not concentrated in a single cluster. This suggests a moderate but not dominant cluster-level effect

CONCLUSION

Clustering is not driven by demographic variables but is enriched for tumor-related features, particularly tumor size.



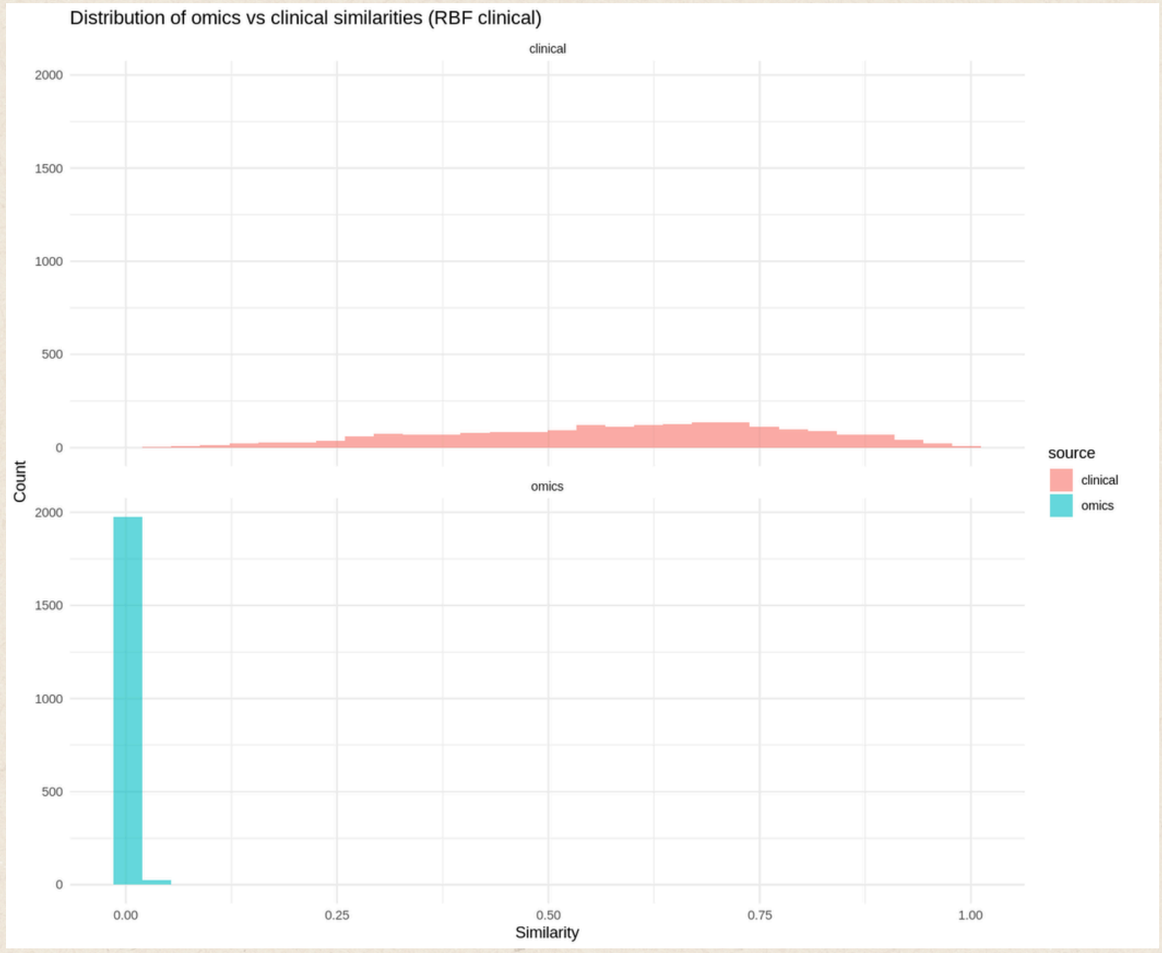




Across multiple analyses, Cluster 6 consistently emerges as distinct, suggesting a biologically and clinically meaningful subgroup.

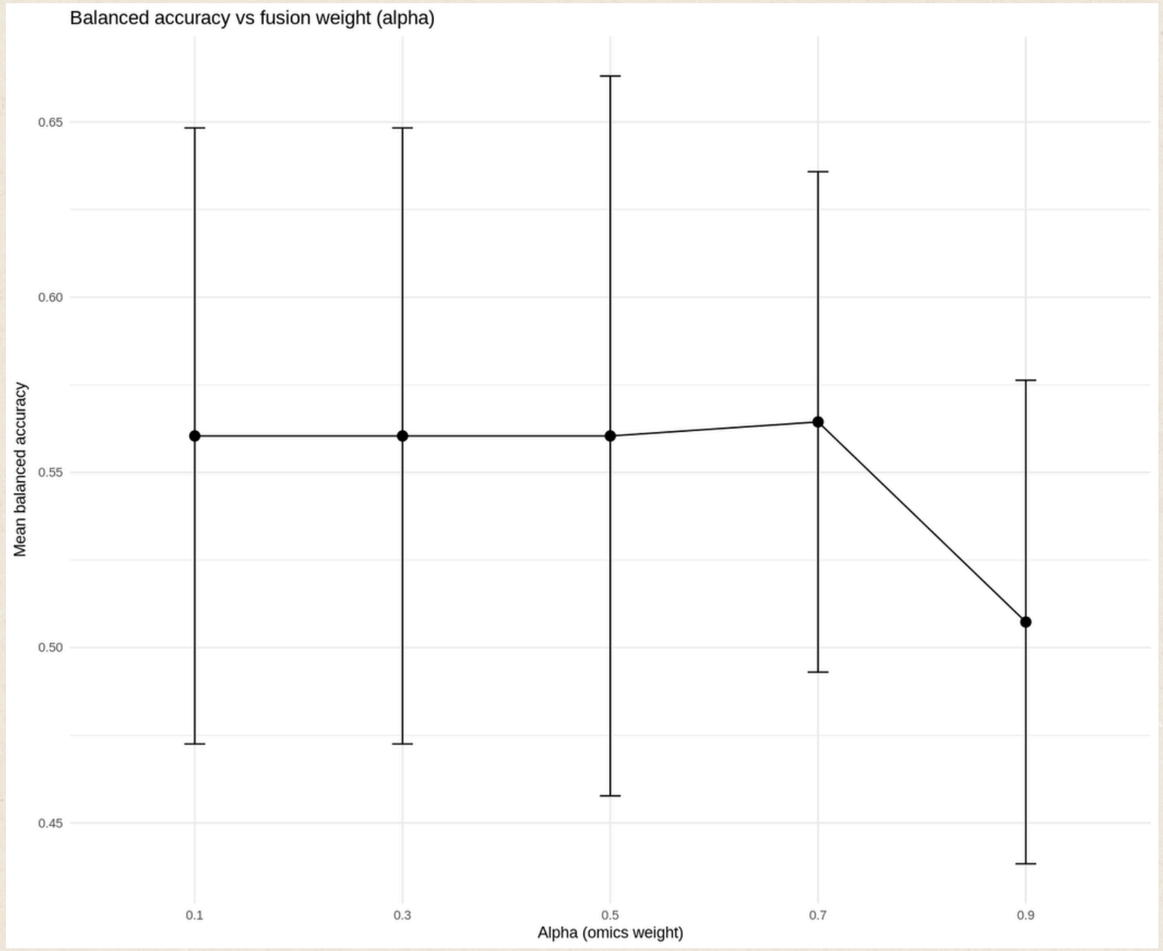
Overall, these results support that the clustering captures clinically interpretable structure, not just technical variation.

PREDICTION



Omics similarities are concentrated near zero, while clinical similarities show a wider spread, indicating that the two data sources carry complementary information.

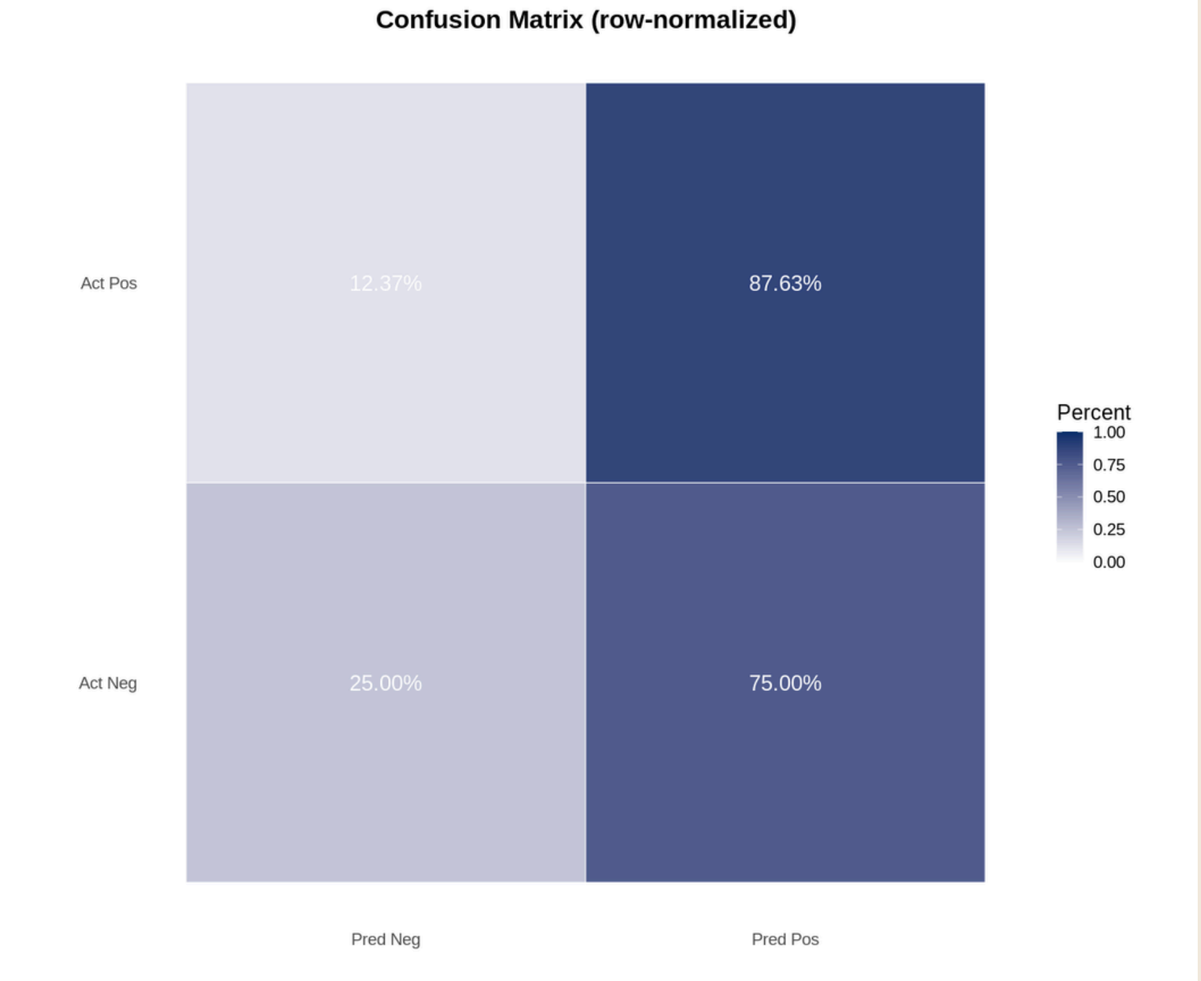
CURSE OF DIMENSIONALITY



Balanced accuracy across different fusion weights (alpha). Performance peaks around α = 0.7, suggesting that combining omics and clinical similarities improves prediction.

Weighted K-NN Classification

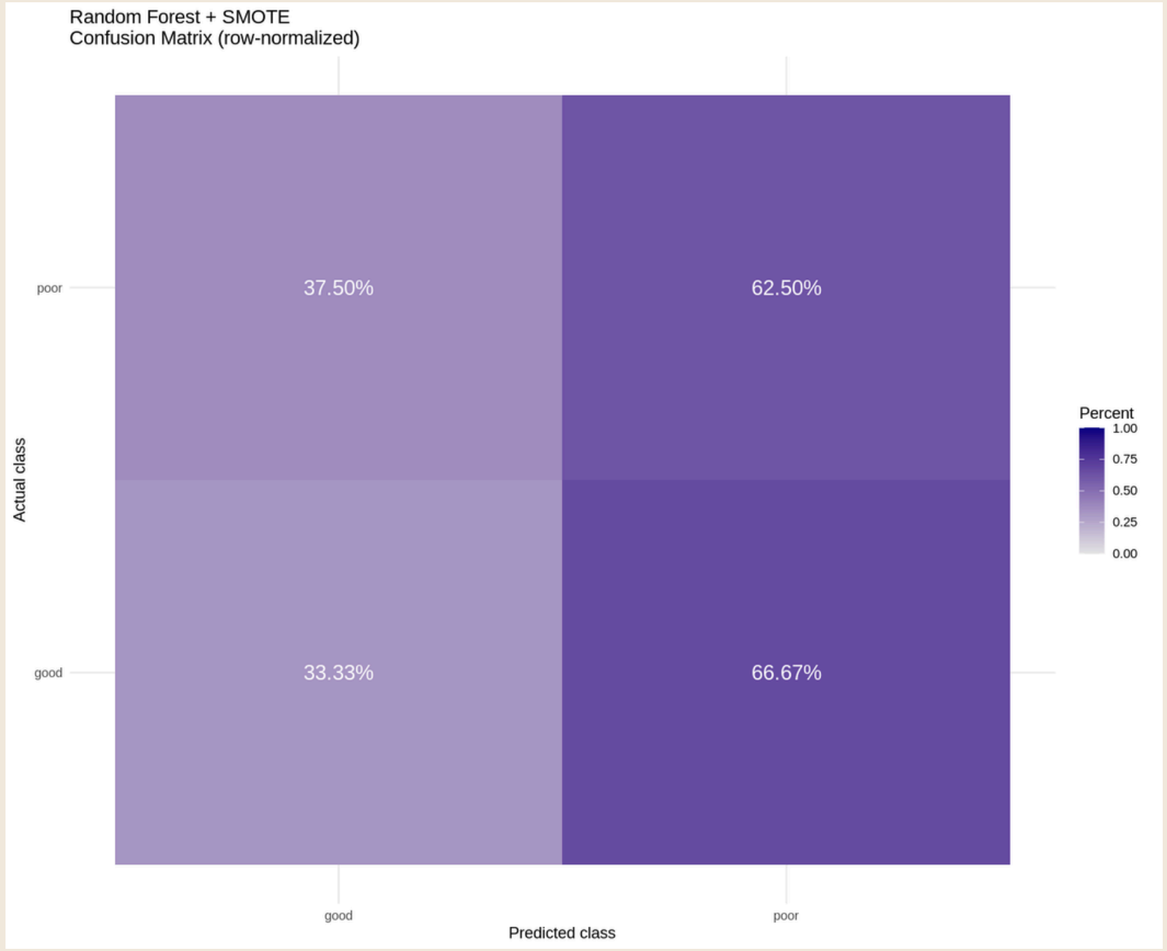
metric	value
Sensitivity	0.88
Specificity	0.25
Pos Pred Value	0.76
Neg Pred Value	0.43
Precision	0.76
Recall	0.88
F1	0.81
Prevalence	0.73
Detection Rate	0.64
Detection Prevalence	0.84
Balanced Accuracy	0.56



metric	value
Accuracy	0.71
Kappa	0.15
AccuracyLower	0.62
AccuracyUpper	0.78
AccuracyNull	0.73
AccuracyPValue	0.76
McnemarPValue	0.02

Random Forest + SMOTE (Supervised Learning)

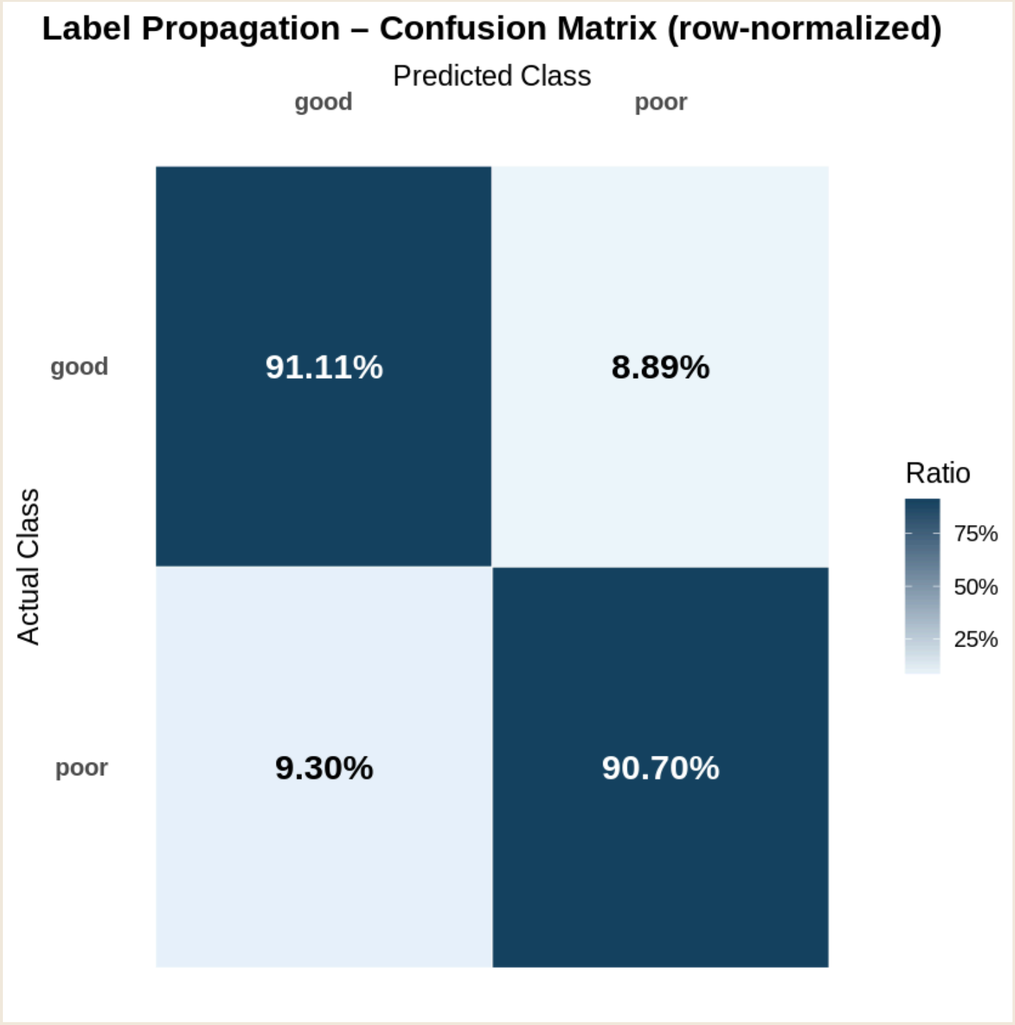
metric	value
Sensitivity	0.625
Specificity	0.333
Pos Pred Value	0.556
Neg Pred Value	0.4
Precision	0.556
Recall	0.625
F1	0.588
Prevalence	0.571
Detection Rate	0.357
Detection Prevalence	0.643
Balanced Accuracy	0.479



metric	value
Accuracy	0.5
Kappa	-0.043
AccuracyLower	0.23
AccuracyUpper	0.77
AccuracyNull	0.571
AccuracyPValue	0.792
McnemarPValue	1

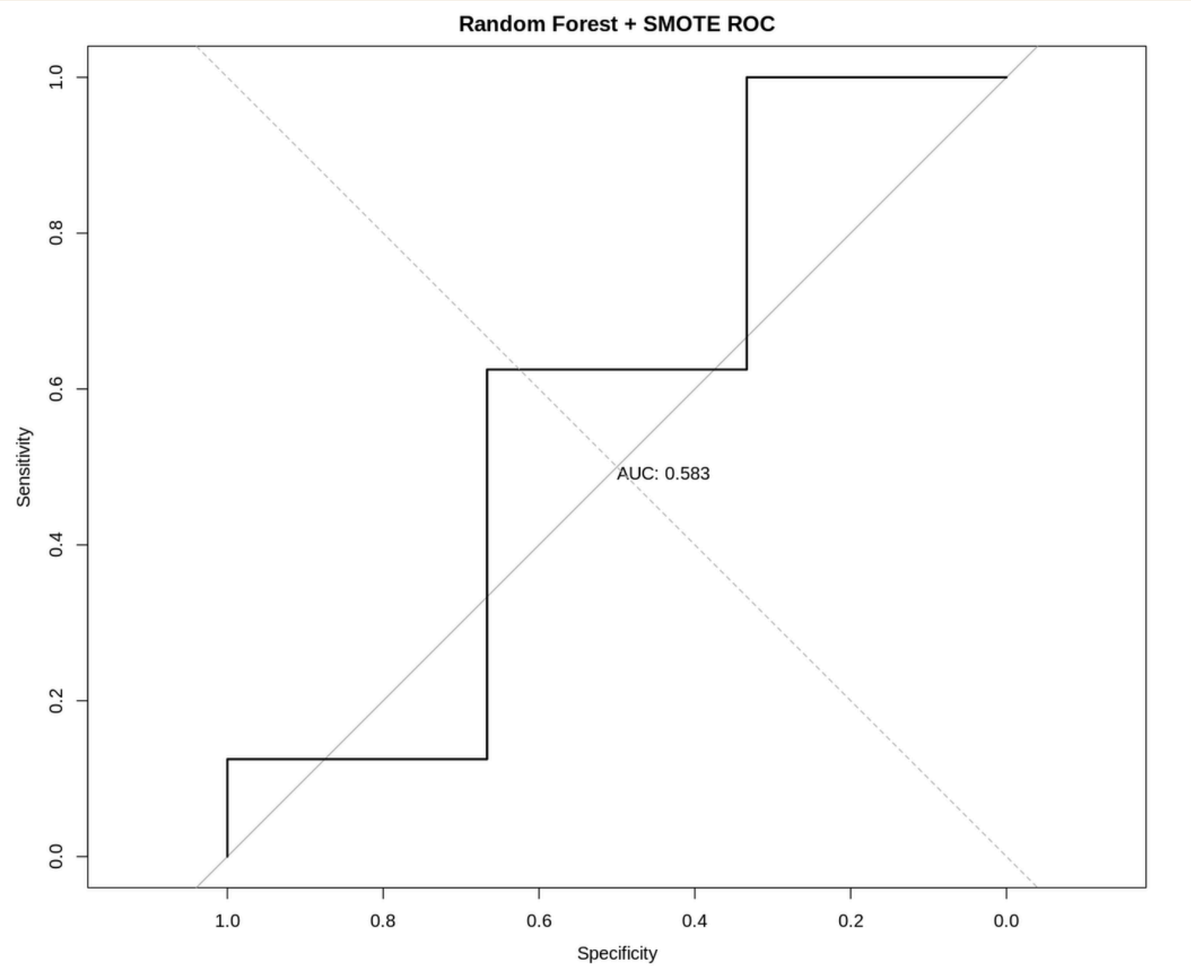
Label Propagation

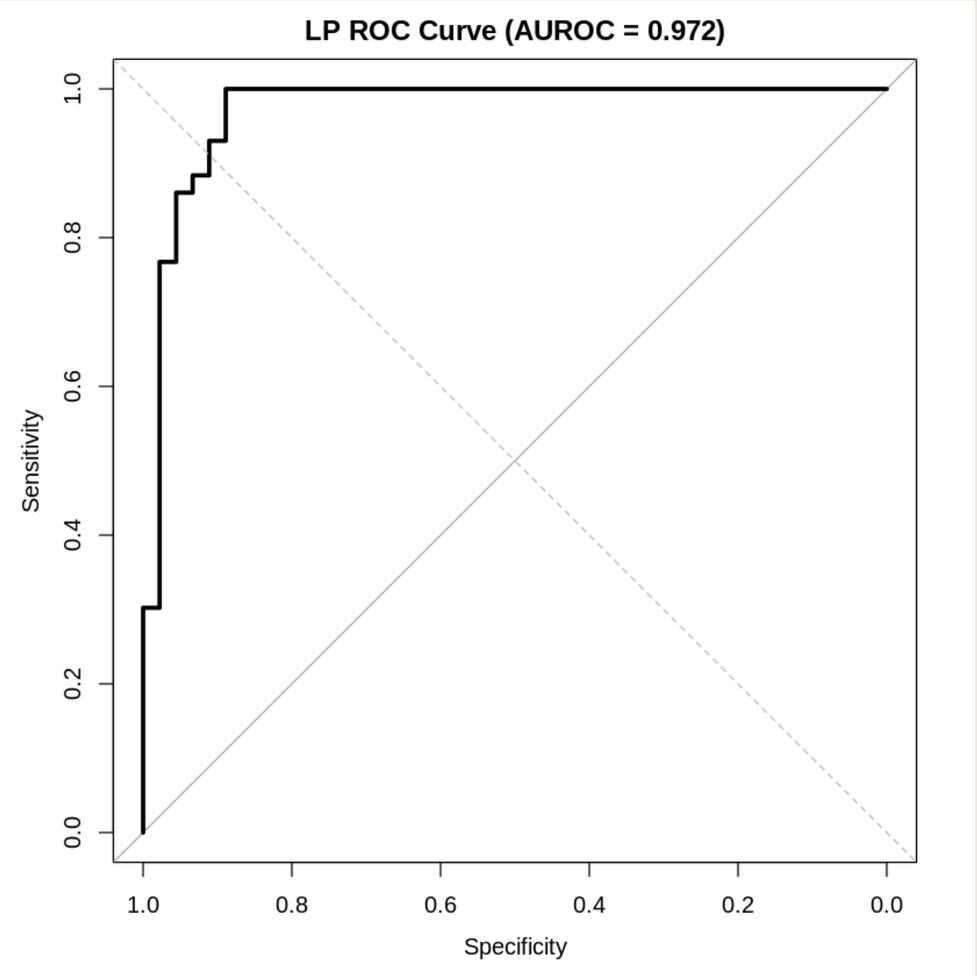
metric	value
Sensitivity	0.907
Specificity	0.911
Pos Pred Value	0.907
Neg Pred Value	0.911
Precision	0.907
Recall	0.907
F1	0.907
Prevalence	0.489
Detection Rate	0.443
Detection Prevalence	0.489
Balanced Accuracy	0.909



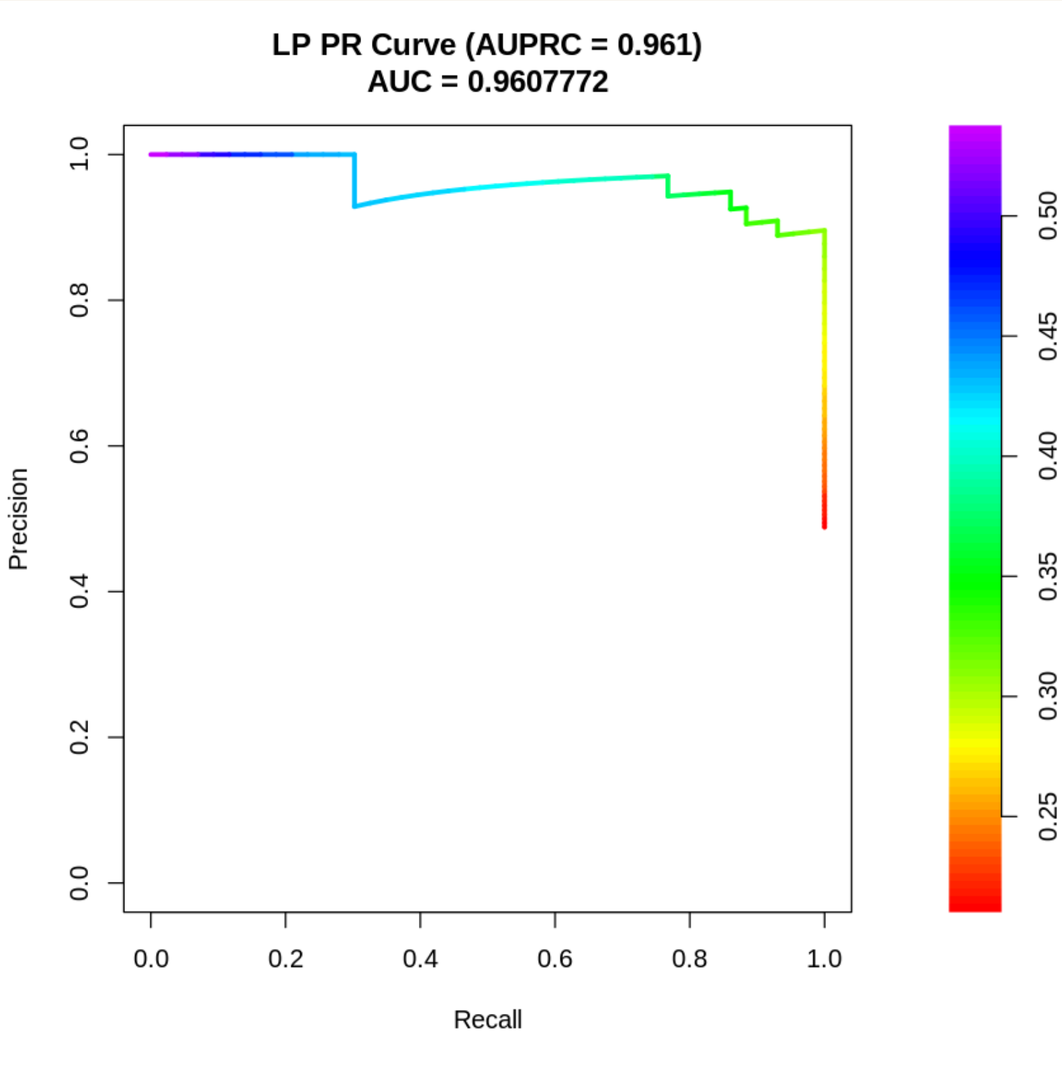
metric	value
Accuracy	0.909
Kappa	0.818
AccuracyLower	0.829
AccuracyUpper	0.96
AccuracyNull	0.511
AccuracyPValue	< 1e-14
McnemarPValue	1

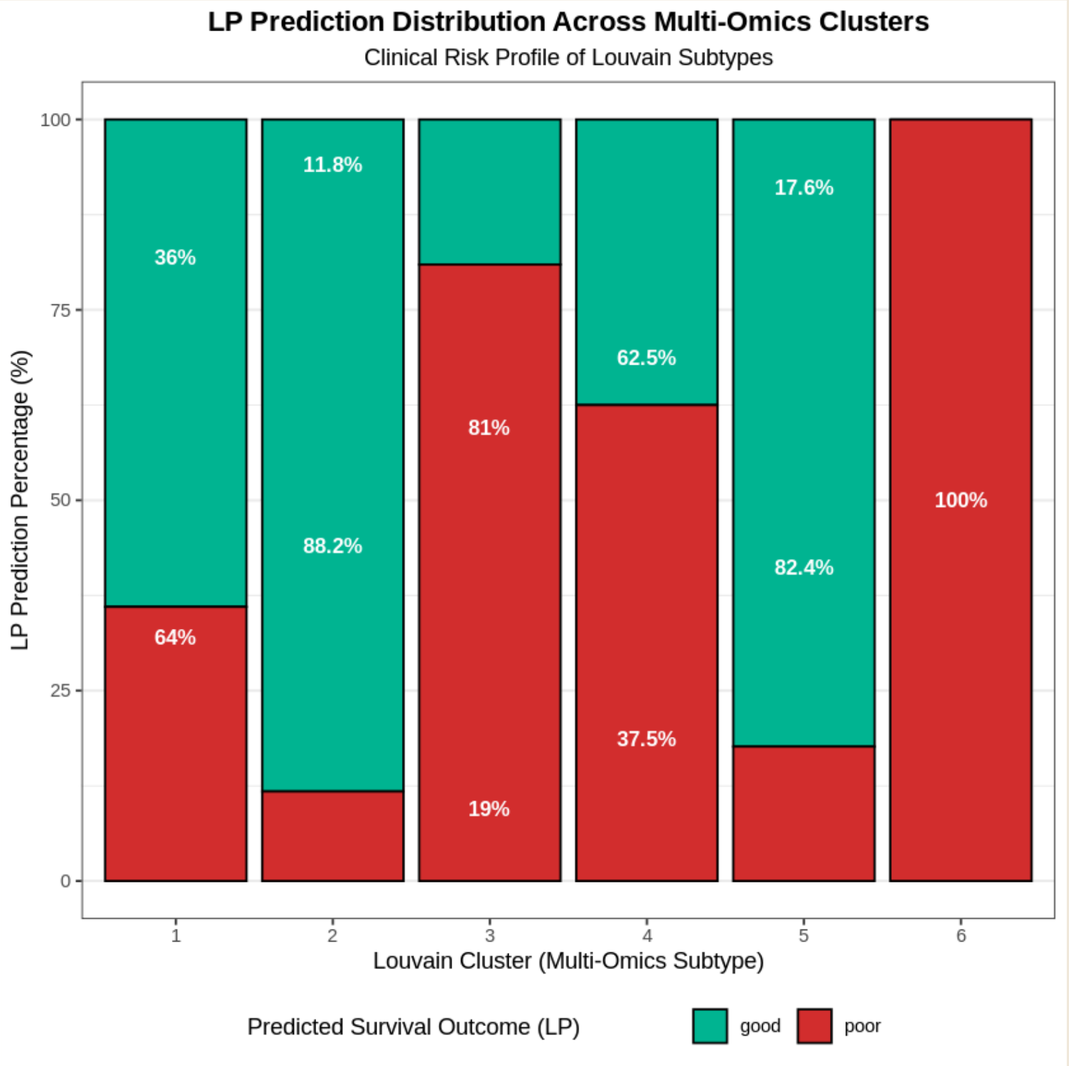
Random Forest + SMOTE vs. Label Propagation





Label Propagation





THANK YOU!