Endotype discovery in a severe TBI cohort with cerebral microdialysis using dimensionality reduction and clustering techniques

Tomasz Kulinski[1,2], Elham Rostami[1,2,3]

[1] Department of Medical Sciences, Uppsala University, [2] Akademiska Sjukhuset, Uppsala, [3] Karolinska Institutet

Background: Severe TBI patient populations are expected to show significant heterogeneity (Maas et al., 2017). Identifying the underlying endotypes from physiological data with dimensionality reduction and clustering is a first step towards a precision medicine approach to TBI.

Materials and methods: 48 variables from <72 h since admission including vital records, bedside physiological monitoring, CT scores and common blood-derived clinical chemistry from 195 patients hospitalized at Akademiska Sjukhuset in Uppsala between 2008 to 2020 for severe TBI with cerebral microdialysis. The median value of time series variables was used. PCA and UMAP dimensionality reduction was performed on the data, with imputation of missingness using EASE (Steck, 2016). Finite mixture modelling (FMM) of the data was done on the unimputed dataset using code by Åkerlund (2023).

Results: PCA showed that the majority of variance between patients was due to tissue trauma and inflammation blood biomarkers, along with cerebral glycerol (table 1); excluding blood variables from consideration exposed the importance of other cerebral microdialysis and physiological monitoring variables (table 2). UMAP-based visual clusterings in 2D were heavily affected by data missingness. FMM with 3, 4 and 5 clusters produced internally consistent patient assignments to clusters that were divergent from results attained with UMAP (figure 1).

Conclusion: Latent structures are present in the analyzed severe TBI cohort, albeit further work is necessary to elucidate their importance, interpretation and comparability to other works in the field. Severe missingness necessitates careful consideration of methods. PCA explained variance recapitulates knowledge about the importance of specific biomarkers.

References

Åkerlund CAI, Holst A, Bhattacharyay S, et al. Lancet Neurol. 2024;23(1):71-80. doi:10.1016/S1474-4422(23)00358-7

Maas AIR, Menon DK, Adelson PD, et al. Lancet Neurol. 2017;16(12):987-1048. doi:10.1016/S1474-4422(17)30371-X

Steck H. In: The World Wide Web Conference. ACM; 2019:3251-3257. doi:10.1145/3308558.3313710

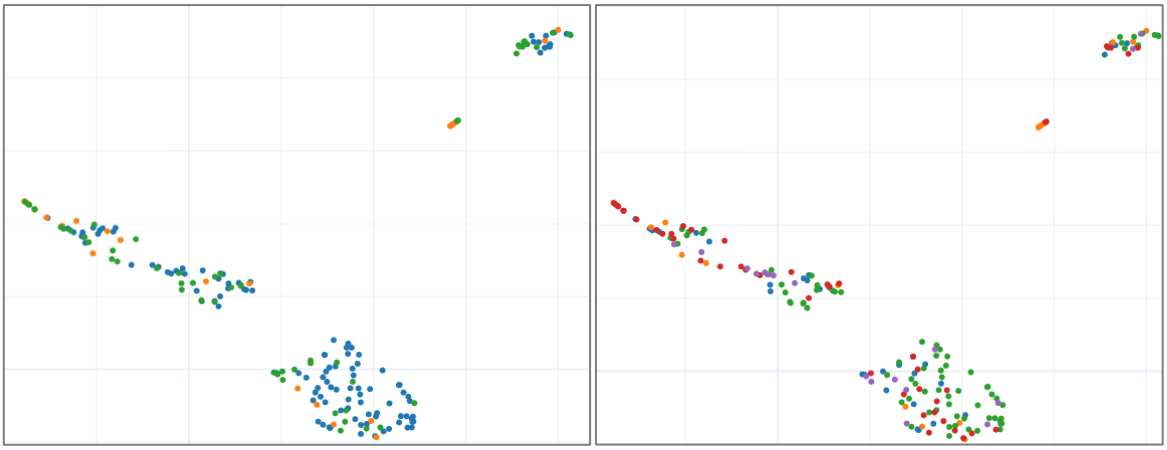

Figure 1. UMAP with cosine metric was used to generate the embeddings, determining the locations of points, which were colored according to cluster labels determined by the finite mixture model (color palette is arbitrary). Left: an example FMM cluster labelling with 3 groups, right: 5 groups.

Table 2. Most and second most significant variables for the first 6 principal components (PC1…6) of PCA of the full dataset. Note that, for any given component, the squared variable projections onto it sum to 1. This includes those not shown in the table, since each component has as many coordinates as there are variables (48), though most are very close to 0. (*) are time series variables, meaning that each patient had (or at the very least was expected to have) more than one value measured at different moments, which was summarized by the median over time.

PC*n*	% explained variance	1st most influential variable	Proj.	2nd most influential variable	Proj.
1	75	Plasma myoglobin *	0.998	Cerebral glycerol *	0.052
2	20	Cerebral glycerol *	0.998	Plasma CRP *	0.026
3	2.5	Blood erythrocyte MCHC *	0.743	Blood thrombocytes *	0.415
4	0.52	Plasma CRP *	0.797	Cerebral pyruvate *	0.329
5	0.49	Cerebral pyruvate *	0.875	Cerebral glutamate *	0.216
6	0.33	Systolic blood pressure *	0.514	Blood oxygen saturation *	0.348

Table 2. Most and second most significant variables for the first 6 principal components (PC1…6) of PCA, excluding the blood clinical chemistry variables. Note that, for any given component, the squared variable projections onto it sum to 1. This includes those not shown in the table, since each component has as many coordinates as there are variables (48), though most are very close to 0. (*) are time series variables, see table 1 caption for more information.

PC*n*	% explained variance	1st most influential variable	Proj.	2nd most influential variable	Proj.
1	30	Mean blood pressure *	0.350	Blood oxygen saturation (SaO_2)	0.350
2	12	Cerebral lactate *	0.517	Cerebral pyruvate *	0.413
3	8.6	Mean intracranial pressure *	0.422	Cerebral lactate/pyruvate ratio *	0.315
4	7.4	Age (years)	0.531	Worst CT Marshall score	0.464
5	5.3	Cerebral glycerol *	0.503	Anticoagulants	0.439
6	5.1	GCS eye score	0.635	Cerebral glycerol *	0.318