Clinical Dataset Exploration Explanation Notebook

narrative of `clinica/clinica_data_exploration.ipynb`.

1. Objectives and Dataset

- Goal: characterize PD vs MSA (MSA-P, MSA-C) clinically; surface discriminative variables; prepare for integration with imaging models.
- Classes considered: MSA-P, MSA-C,PD (MSA-P and MSA-C are sometimes merged as MSA).

1.1 Legend

Column	Type	Short explanation
anni_dalla_diagnosi	numeric	Years elapsed since formal diagnosis.
anni_dopaminoaagonisti	numeric	Duration of dopamine-agonist therapy (years).
anni_l_dopa	numeric	Duration of levodopa therapy (years).
anno_diagnosi	numeric	Calendar year of diagnosis.
anno_esordio_disautonomia	numeric	Year of first autonomic symptom.
anno_esordio_sintomi_motori	numeric	Year of first motor symptom.
anno_esordio_sintomi_non_motori	numeric	Year of first non-motor symptom.
anno_nascita	numeric	Year of birth.
compass_gi	numeric	COMPASS-31 gastrointestinal sub-score.
compass_oh	numeric	COMPASS-31 orthostatic hypotension sub-score.
compass_pupil	numeric	COMPASS-31 pupillomotor sub-score.
compass_sudor	numeric	COMPASS-31 sudomotor (sweating) sub-score.
compass_totale	numeric	Total COMPASS-31 autonomic dysfunction score.
compass_uin	numeric	COMPASS-31 urinary sub-score.
compass_vasc	numeric	COMPASS-31 vasomotor sub-score.

Column	Type	Short explanation
delta_off_on	numeric	Difference between UPDRS_OFF and UPDRS_ON (treatment effect).
durata_malattia	numeric	Disease duration from onset (years).
eta_attuale	numeric	Current patient age.
eta_diagnosi	numeric	Age at diagnosis.
eta_esordio	numeric	Age at first motor symptom.
h_and_y	numeric	Hoehn & Yahr stage (1–5).
ledd	numeric	Levodopa equivalent daily dose (mg/day).
ledd_per_anno	numeric	LEDD normalized per year of disease.
n_anomalie_mri	numeric	Number of abnormal MRI findings.
n_red_flags_msa	numeric	Count of MSA “red-flag” features (per MDS).
n_red_flags_msa_clinic_certified	numeric	Clinically certified number of red flags.
parkinsonism	numeric	Severity/composite score of parkinsonian signs (rigidity, bradykinesia, tremor).
percentuale_risposta_ldopa	numeric	% improvement after acute L-Dopa test.
progression_rate	numeric	Calculated progression speed (e.g. H&Y / disease years).
ritardo_diagnostico	numeric	Diagnostic delay (years from symptom onset to diagnosis).
updrs_off	numeric	UPDRS-III motor score in OFF-medication state.

Column	Type	Short explanation
updrs_on	numeric	UPDRS-III motor score in ON-medication state.
atrofia_cervelletto	binary	MRI: cerebellar atrophy.
atrofia_del_putamen	binary	MRI: putaminal atrophy.
atrofia_peduncoli_cerebellari_medi	binary	MRI: middle cerebellar peduncle atrophy.
atrofia_ponte	binary	MRI: pontine atrophy.
behavioural_alteration	binary	Behavioural or personality changes.
caduta_segnale_putamen	binary	MRI: putaminal signal loss on T2*/SWI.
cadute	binary	History of falls.
carrozzina	binary	Wheelchair use.
cerebellar_syndrome	binary	Presence of cerebellar signs (ataxia, dysmetria).
cognitive_decline	binary	Cognitive impairment or dementia.
cold_discolored_hands_and_feet	binary	Peripheral vasomotor disturbance (autonomic).
constipation	binary	Chronic constipation.
craniocervical_dyst_induced_dy_l_dopa	binary	Craniocervical dystonia induced by L-Dopa.
deambulaz_appoggio	binary	Ambulates with support.
deambulaz_autonoma	binary	Ambulates independently.
drooling	binary	Hypersalivation / drooling.
erectile_disfunction	binary	Erectile dysfunction (autonomic

Column	Type	Short explanation
		symptom).
fatigue	binary	Fatigue / lack of energy.
hot_cross_bun_sign	binary	MRI: pontine cruciform hyperintensity typical of MSA-C.
hyposmia	binary	Reduced sense of smell.
inspiratory_sighs	binary	Sighing or irregular breathing pattern.
iperintensita_peduncoli_cerebellari_medi	binary	MRI: MCP hyperintensity.
iperintensita_putamen	binary	MRI: putaminal hyperintensity.
jerky_myoclonic_postural_or_kinetic_tremor	binary	Irregular / jerky tremor type.
moderate_to_severe_postural_instability_w_3_yrs_of_motor_onset	binary	Postural instability within 3 years of motor onset.
normal_rmn	binary	Normal brain MRI.
pain	binary	Chronic or neuropathic pain.
pathologic_laughter_or_crying	binary	Emotional incontinence (pseudobulbar affect).
poor_l_dopa_responsivenes	binary	Poor or absent clinical response to L-Dopa.
postural_deformities	binary	Axial/postural deformities (camptocormia, Pisa).
rapid_progression_w_3_yrs	binary	Rapid disease progression within 3 years.
rbd	binary	REM sleep behavior disorder.
russamento_osas	binary	Snoring / sleep apnea (OSAS).
severe_dysphagia_w_3_yrs	binary	Severe dysphagia within 3 years of

Column	Type	Short explanation
		onset.
severe_speech_impairement_w_3_yrs	binary	Severe dysarthria within 3 years of onset.
sonnolenza_diurna	binary	Excessive daytime sleepiness.
stridor	binary	Laryngeal stridor (inspiratory noise).
unexplained_babinski	binary	Pathological Babinski sign unexplained by stroke.
unexplained_urinary_urge_incontinence	binary	Urinary urge incontinence unexplained by obstruction.
unexplained_voiding_difficulties	binary	Urinary retention / difficulty voiding unexplained by prostate disease.
visual_alteration	binary	Visual disturbances or blurring.
anamnestic_oh	binary	History of orthostatic hypotension from anamnesis.
diagnosi_definita	categorical	Final confirmed diagnosis (PD / MSA-P / MSA-C / Control).
diagnosi_di_invio	categorical	Referral diagnosis at first evaluation.
gruppo_eta	categorical	Age group (e.g. <50, 50-60, 60-70, >70).
sesso	categorical	Sex (M/F).
stadio_malattia	categorical	Disease stage grouping (e.g. early / mid / late).
data_di_nascita	date	Full birth date.

pointed by Grazia

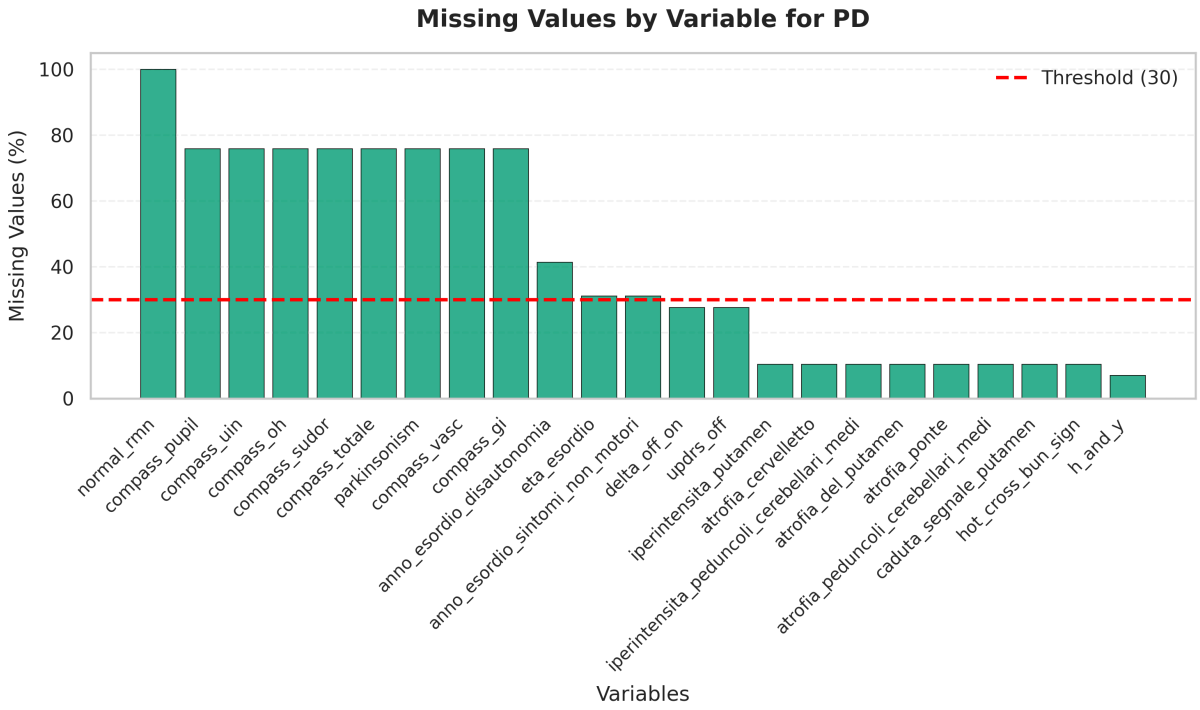
Set	Variables Included
Clinician-Certified MSA Red Flags	`poor_l_dopa_responsivenes`, `rapid_progression_w_3_yrs`, `moderate_to_severe_postural_instability_w_3_yrs_of_motor_onset`, `craniocervical_dyst_induced_dy_l_dopa`, `severe_speech_impairement_w_3_yrs`, `severe_dysphagia_w_3_yrs`, `unexplained_babinski`, `jerky_myoclonic_postural_or_kinetic_tremor`, `postural_deformities`, `unexplained_voiding_difficulties`, `unexplained_urinary_urge_incontinence`, `stridor`, `inspiratory_sighs`, `cold_discolored_hands_and_feet`, `pathologic_laughter_or_crying`

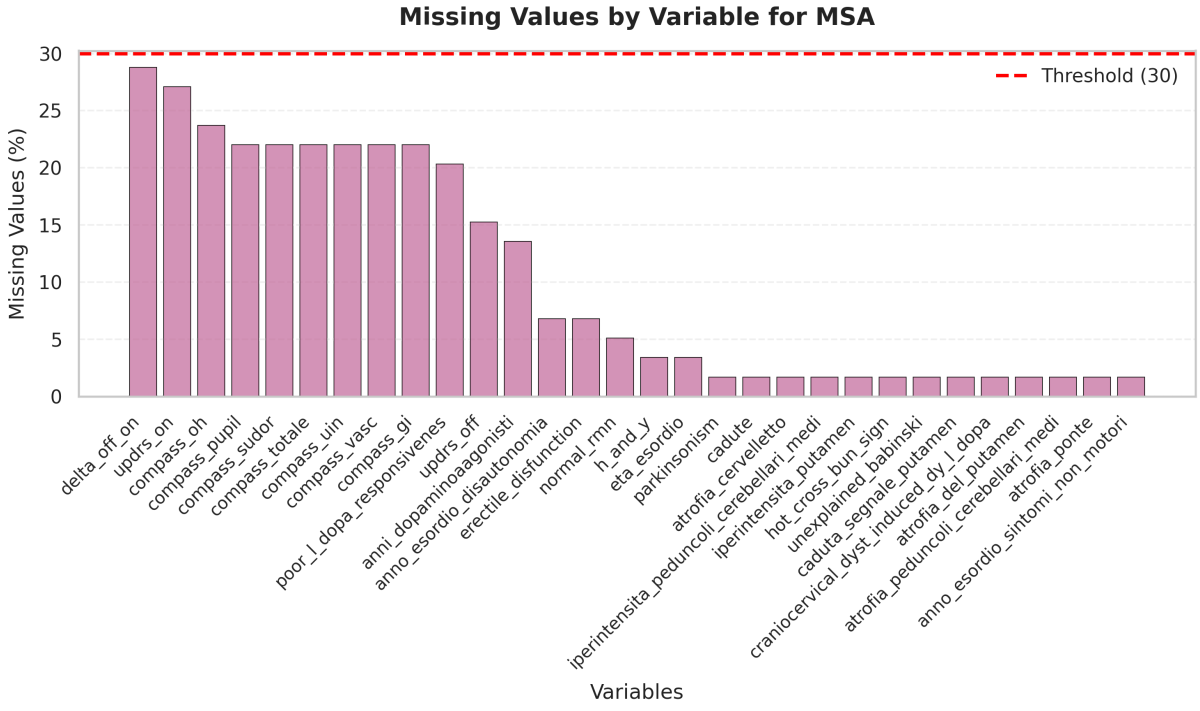
2. Data Ingestion and Harmonization

- Load clinical csv (harmonizing different rapresentations of missing values to Nan)
- Column names normalized: Unicode stripped, whitespace collapsed, then slugified to ASCII snake_case.
- Diagnosis labels trimmed and harmonized (e.g., `MSA-P/C` → `MSA-P`).
- Numeric casting based on patterns; categorical cleaning (uppercasing, NA handling).
- Duplicates check

3. Missingness Profiling

- Quantify missingness per variable per class to guide usable features.

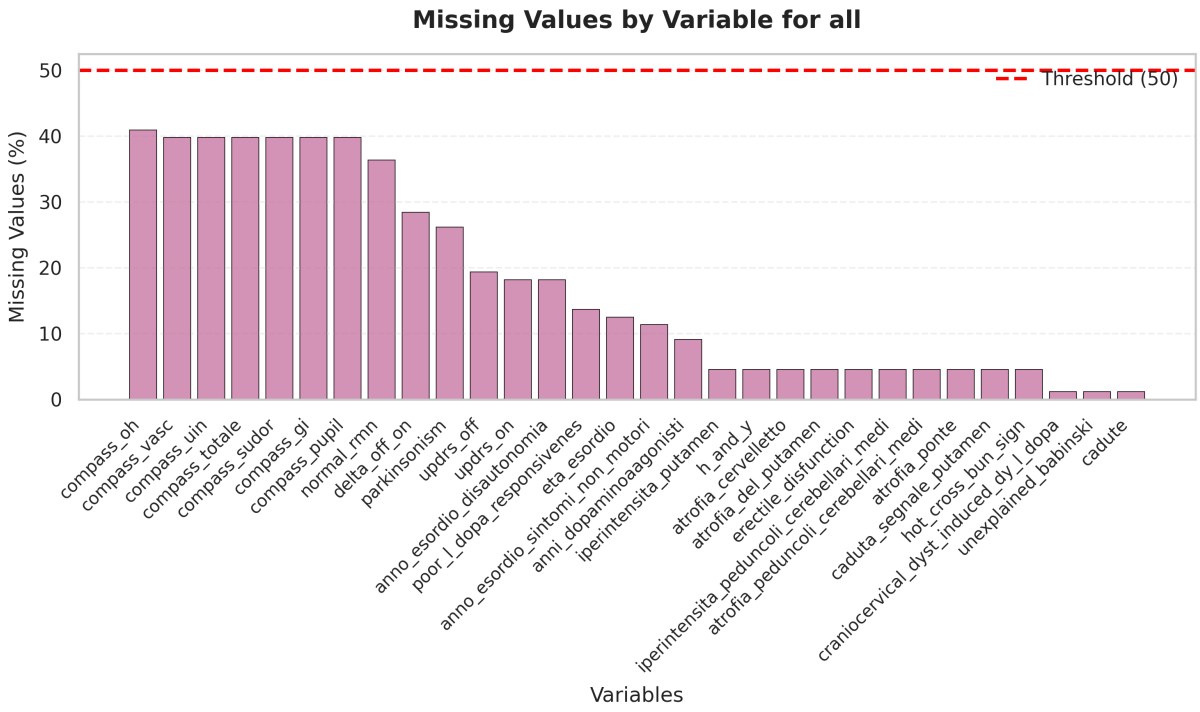




📝 **NOTE:** which thrashold do i keep?

🚨 **IMPORTANT:** Note that PD has much more missings

Remaining columns after removal of high missing columns



4. Derived Variables and Cleaning

- Create analysis-friendly fields (examples):
  - Timing/severity: `eta_attuale`, `eta_esordio`, `durata_malattia`, `percentuale_risposta_ldopa`.
  - Aggregates: `n_red_flags_msa`, `n_anomalie_mri` as the sum of the corresponding binary symptoms columns

Variable	Short description
eta_attuale	Current patient age.
ritardo_diagnostico	Diagnostic delay (years from symptom onset to diagnosis).
anni_dalla_diagnosi	Years elapsed since formal diagnosis.
percentuale_risposta_ldopa	% improvement after acute L-Dopa test.
ledd_per_anno	Levodopa equivalent daily dose normalized per year of disease.
n_red_flags_msa	Count of MSA “red-flag” AI generated.
n_red_flags_msa_clinic_certified	Clinically certified count od MSA red flags.
n_anomalie_mri	Number of abnormal MRI findings.
stadio_malattia	Disease stage grouping (e.g. early / mid / late).
gruppo_eta	Age group (e.g. <50, 50-60, 60-70, >70).
progression_rate	Calculated progression speed (e.g. H&Y / disease years).

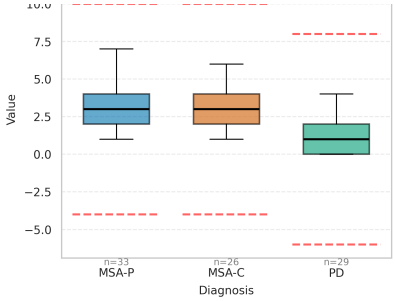
5. Outlier Detection

- Use Tukey's fences on key continuous variable to spot extreme values.
- Provides a record of suspected outliers for clinician review.

Method: Tukey's Fences

- Lower fence = Q1 − 1.5 × IQR
- Upper fence = Q3 + 1.5 × IQR
- Values outside these fences are flagged as outliers


Hoehn & Yahr Stage
(4 outliers total)
compass_gi
(1 outliers total)
compass_oh
(1 outliers total)
compass_totale
(1 outliers total)
N° MRI Abnormalities
(4 outliers total)
compass_vasc
(1 outliers total)
Progression Rate
(2 outliers total)
Years on L-DOPA
(2 outliers total)
Years on DA Agonists
(1 outliers total)
LEDD (mg/day)
(1 outliers total)
LEDD per Year
(1 outliers total)
UPDRS Δ(OFF-ON)
(1 outliers total)
L-DOPA Response (%)
(1 outliers total)
parkinsonism
(1 outliers total)
Disease Duration (years)
(1 outliers total)
Diagnostic Delay (years)
(1 outliers total)
Value
Diagnosis
MSA-P
MSA-C
PD



Consolidated Outlier Details (Tukey k=3)

Variable	Patient ID	Diagnosis	Value	Q1	Q3	Lower Fence	Upper Fence
Diagnostic Delay (years)	6663	MSA-P	12.0	2.0	4.0	-4.0	10.0
Disease Duration (years)	6663	MSA-P	12.0	3.0	5.0	-3.0	11.0
Hoehn & Yahr Stage	6008	PD	1.0	0.0	0.0	0.0	0.0
Hoehn & Yahr Stage	6366	PD	1.0	0.0	0.0	0.0	0.0
Hoehn & Yahr Stage	6383	PD	1.0	0.0	0.0	0.0	0.0
Hoehn & Yahr Stage	6424	PD	1.0	0.0	0.0	0.0	0.0
L-DOPA Response (%)	6308	MSA-P	53.12	12.73	20.31	-10.01	43.05
LEDD (mg/day)	5996	MSA-C	1208.0	0.0	287.5	-862.5	1150.0
LEDD per Year	7461	PD	400.0	80.0	156.25	-148.75	385.0
N° MRI Abnormalities	6008	PD	0.0	1.0	1.0	1.0	1.0
N° MRI Abnormalities	6323	PD	0.0	1.0	1.0	1.0	1.0
N° MRI Abnormalities	6690	PD	0.0	1.0	1.0	1.0	1.0
N° MRI Abnormalities	7787	PD	0.0	1.0	1.0	1.0	1.0
Progression Rate	7179	MSA-C	39.0	5.94	12.75	-14.5	33.19
Progression Rate	7155	PD	40.0	3.78	8.4	-10.09	22.27
UPDRS Δ(OFF-ON)	6308	MSA-P	-53.13	-20.31	-12.72	-43.06	10.02
Years on DA Agonists	5996	MSA-C	6.0	0.0	0.0	0.0	0.0
Years on L-DOPA	5996	MSA-C	6.0	0.0	1.0	-3.0	4.0
Years on L-DOPA	6663	MSA-P	10.0	1.0	3.0	-5.0	9.0
compass_gi	6651	PD	2.0	6.5	7.5	3.5	10.5
compass_oh	6008	PD	8.0	0.0	0.0	0.0	0.0
compass_totale	6008	PD	33.0	12.0	16.5	-1.5	30.0
compass_vasc	5969	MSA-C	1.67	0.0	0.0	0.0	0.0
parkinsonism	7132	MSA-P	2.33	1.0	1.0	1.0	1.0

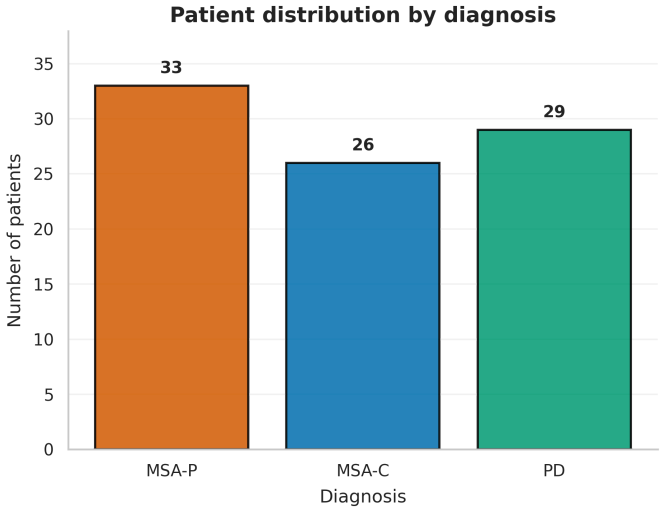
⚠️ **TODO:** Make clinician verify outliers and decide what to do with those

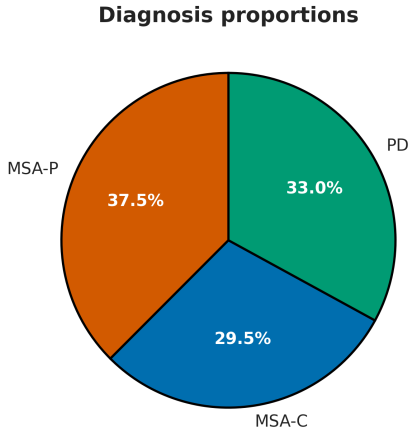
6. Cohort Overview

- Distribution of patients across definitive diagnoses.
- A compact summary table (N, M/F, age at onset, duration, current age).

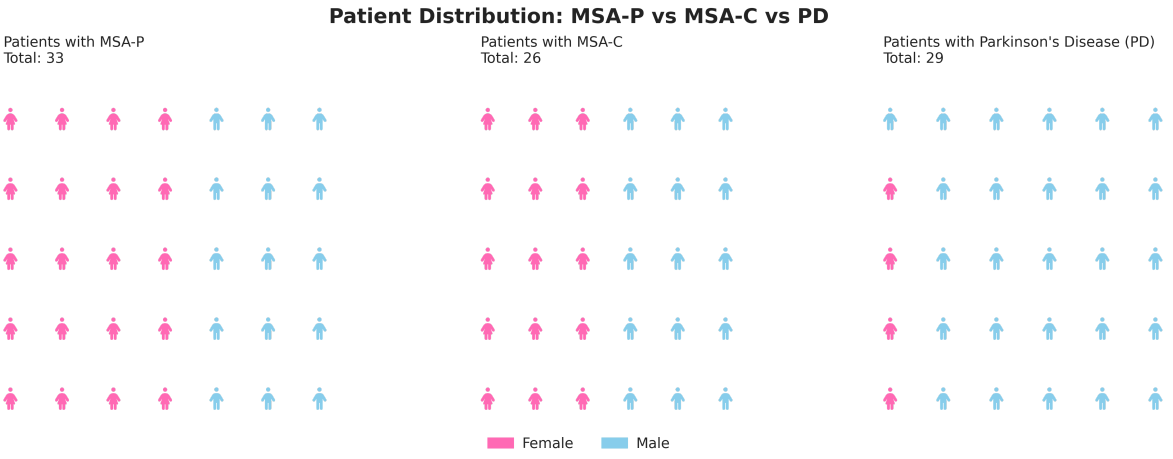
Clinical and Demographic Characteristics by Diagnosis

Diagnosis	N	M/F	Age onset	Duration	Current age
MSA-P	33	14/19	61.0±7.5	4.0±2.5	67.1±7.5
MSA-C	26	13/13	56.8±8.7	3.1±1.7	61.7±8.7
PD	29	25/4	55.1±11.4	6.9±4.6	63.8±9.5





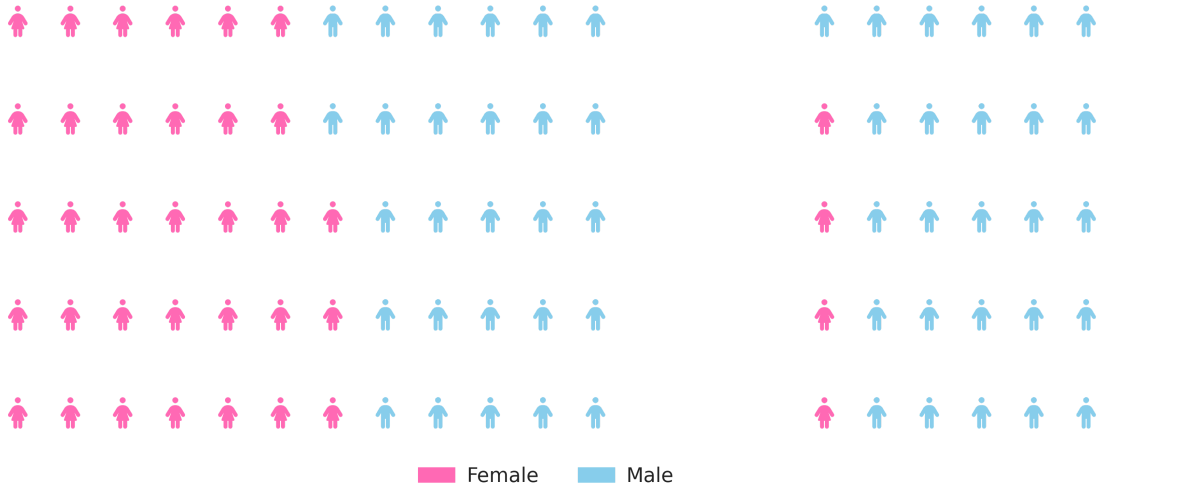
> 🚨 **WARNING:** PD patients are predominantly males

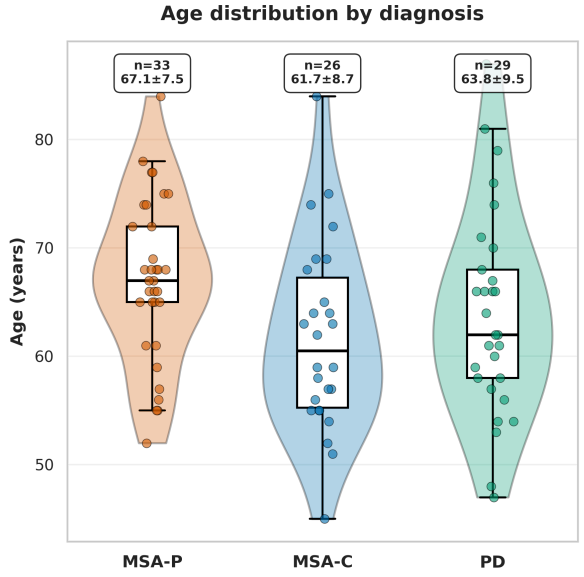


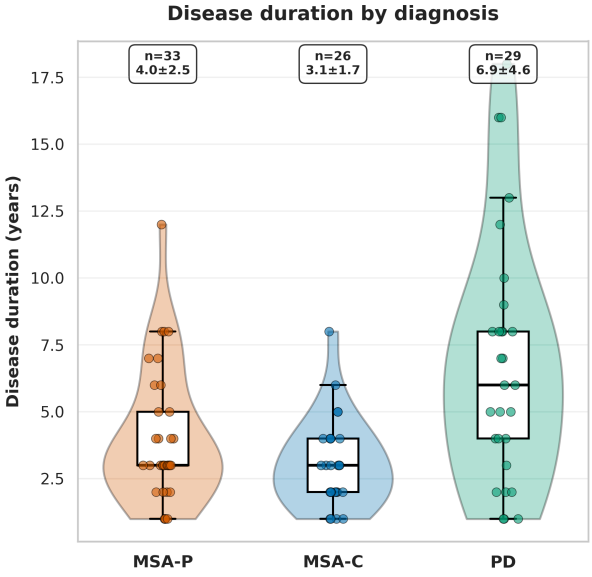
Patient Distribution: MSA vs PD

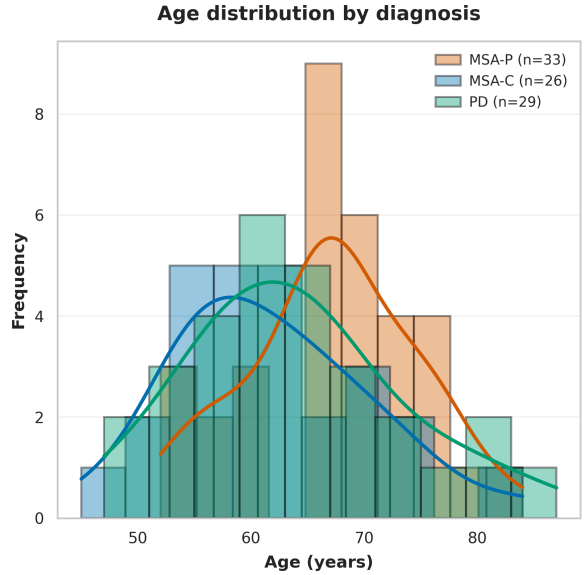
Patients with Multiple System Atrophy (MSA)
Total: 59

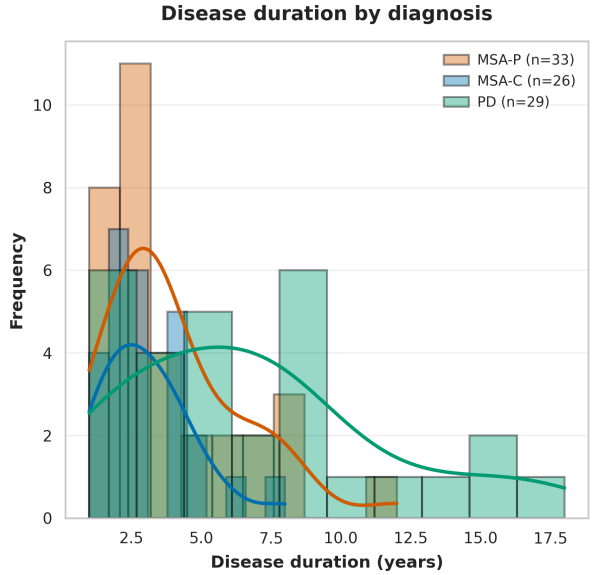
Patients with Parkinson's Disease (PD)
Total: 29







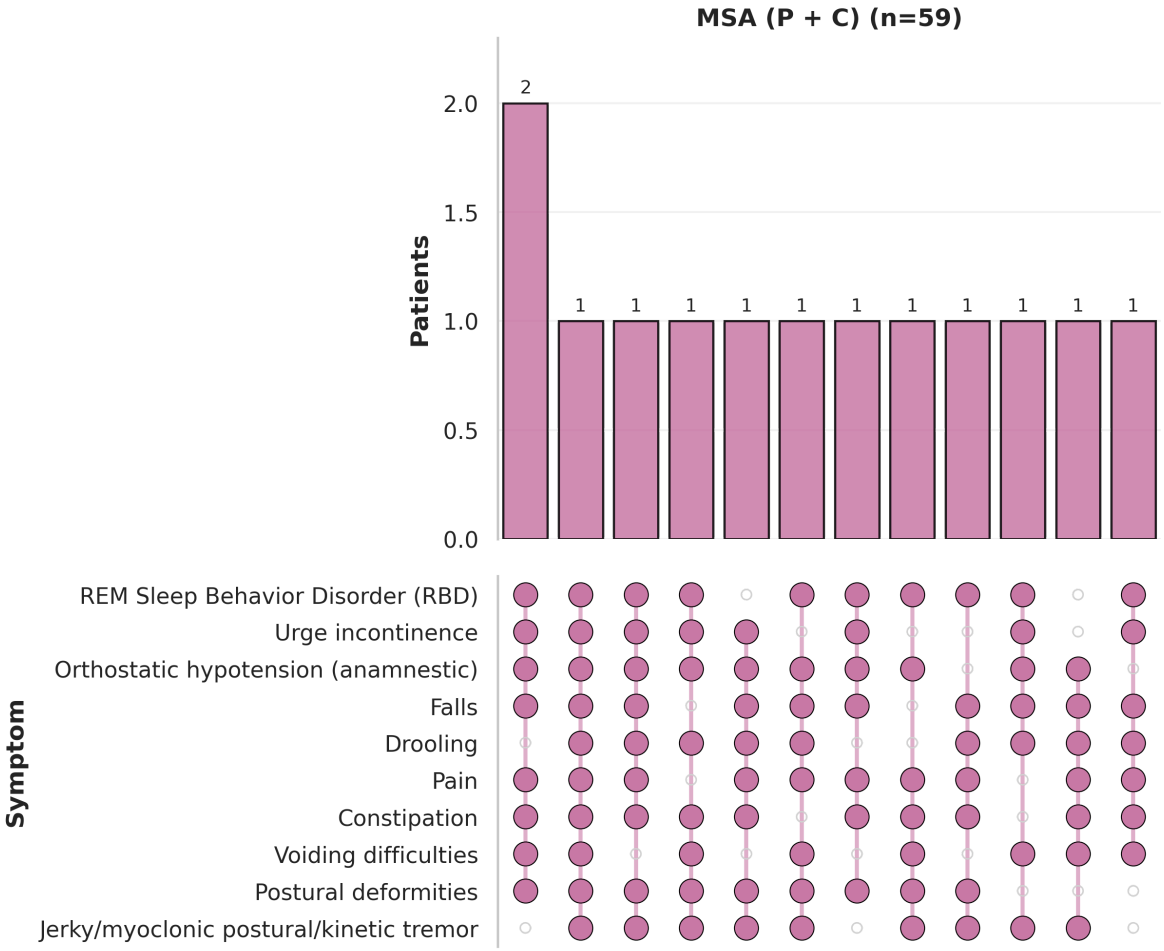


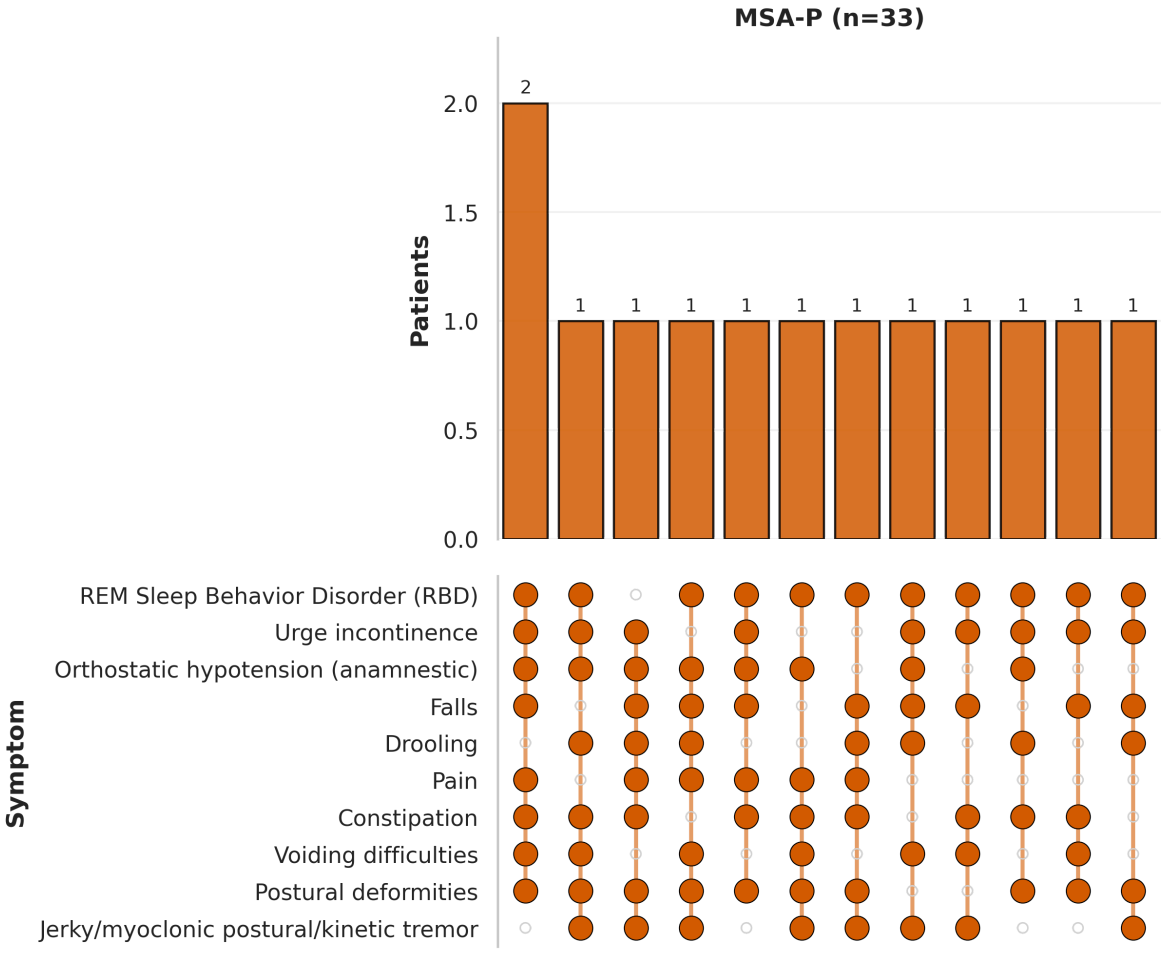


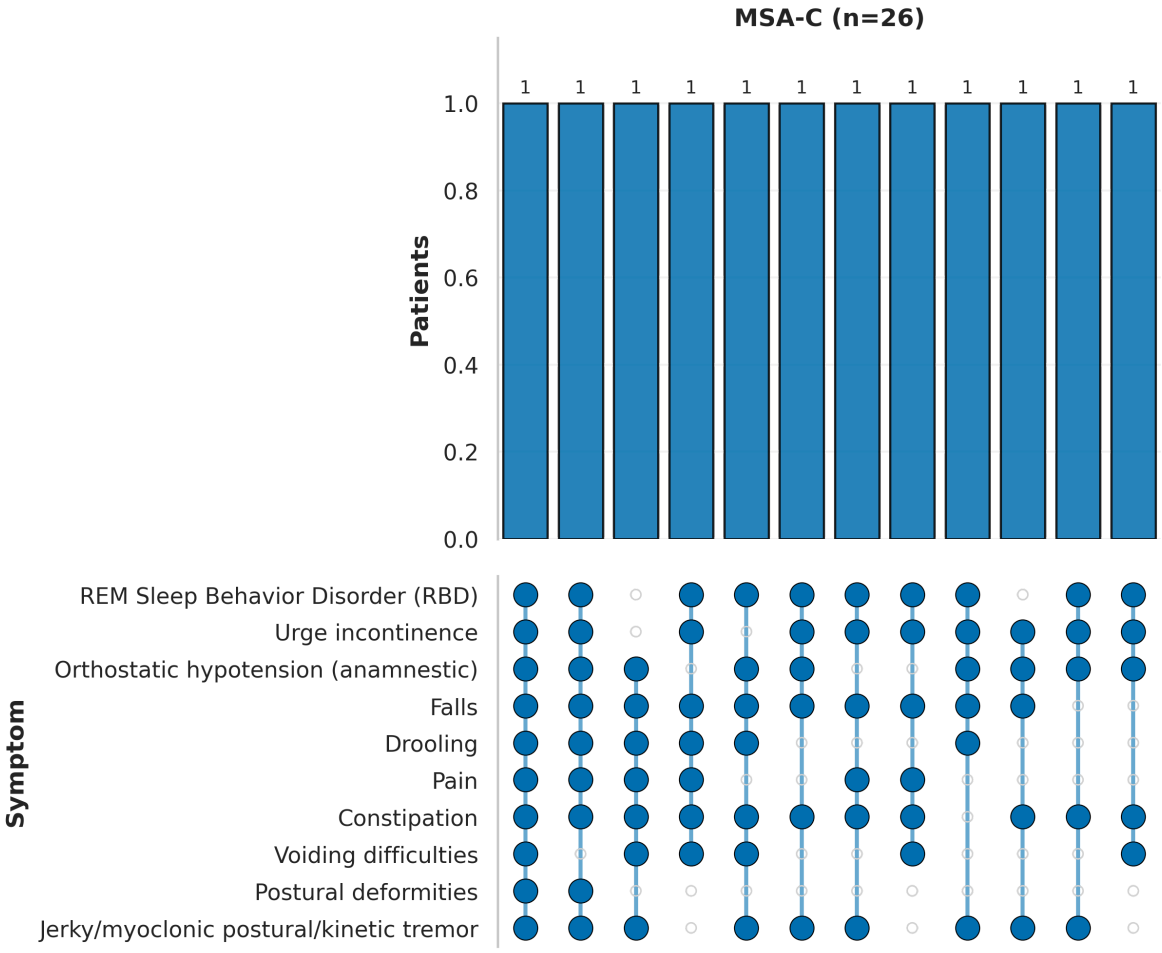
7. Multi-symtoms clusters and Co-occurrence

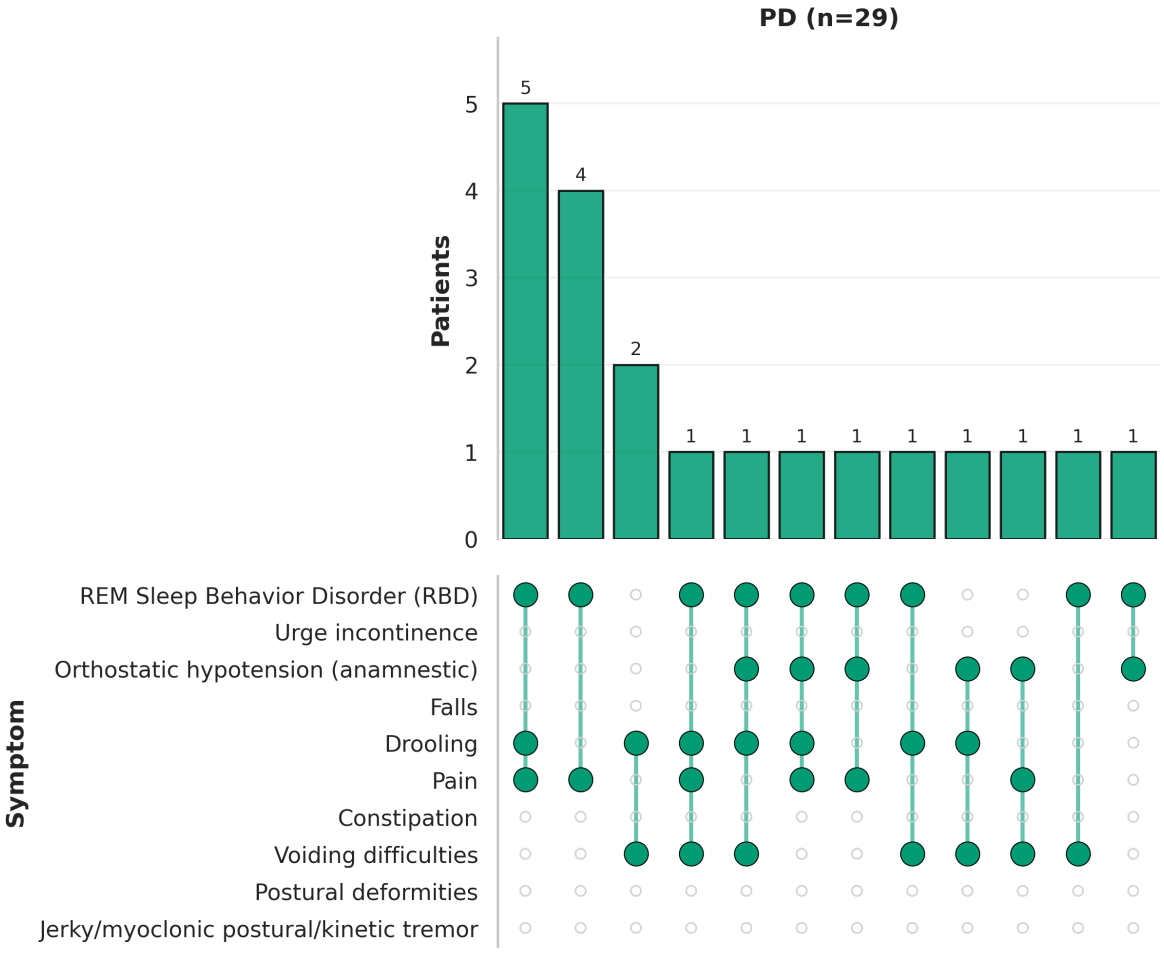
This analysis identifies and visualizes patterns of symptom co-occurrence within each diagnostic group (MSA-P, MSA-C, PD) to reveal characteristic multi-symptom profiles that may aid differential diagnosis.

Rationale: While univariate symptom prevalence identifies individual discriminators, parkinsonian syndromes are clinically characterized by specific constellations of concurrent symptoms.


MSA (P + C) (n=59)
Patients
2.0
1.5
1.0
0.5
0.0
2
1
1
1
1
1
1
1
1
1
1
1
Symptom
REM Sleep Behavior Disorder (RBD)
Urge incontinence
Orthostatic hypotension (anamnestic)
Falls
Drooling
Pain
Constipation
Voiding difficulties
Postural deformities
Jerky/myoclonic postural/kinetic tremor


MSA-P (n=33)
Patients
2.0
1.5
1.0
0.5
0.0
2
1
1
1
1
1
1
1
1
1
1
1
Symptom
REM Sleep Behavior Disorder (RBD)
Urge incontinence
Orthostatic hypotension (anamnestic)
Falls
Drooling
Pain
Constipation
Voiding difficulties
Postural deformities
Jerky/myoclonic postural/kinetic tremor


MSA-C (n=26)
1 1 1 1 1 1 1 1 1 1 1 1
1.0
0.8
0.6
0.4
0.2
0.0
Patients
Symptom
REM Sleep Behavior Disorder (RBD)
Urge incontinence
Orthostatic hypotension (anamnestic)
Falls
Drooling
Pain
Constipation
Voiding difficulties
Postural deformities
Jerky/myoclonic postural/kinetic tremor

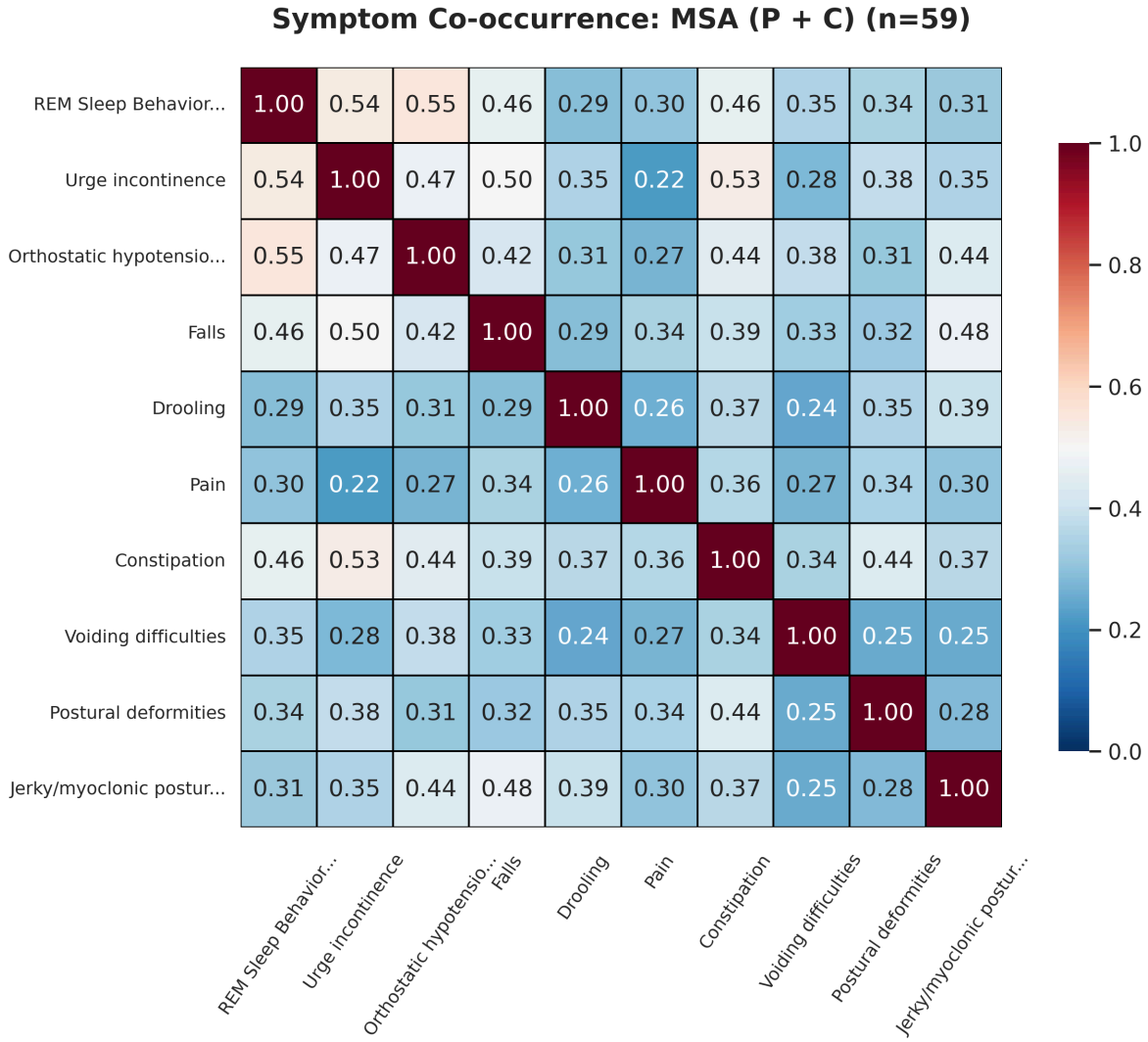


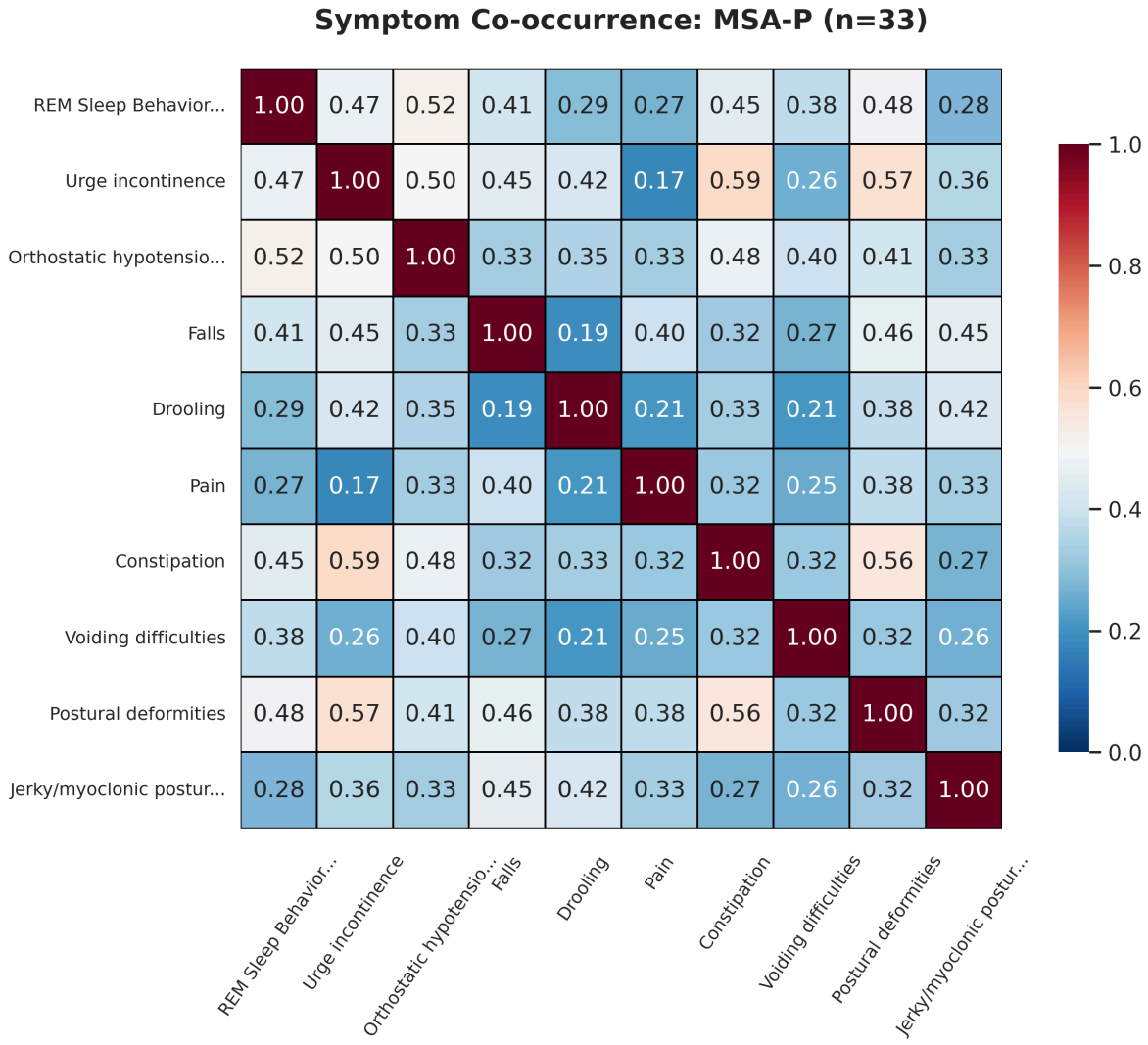
7.1 symptoms co-occurence heatmaps

This analysis quantifies pairwise symptom co-occurrence within each diagnostic group using the Jaccard similarity coefficient, revealing which symptoms tend to appear together in **individual patients**.

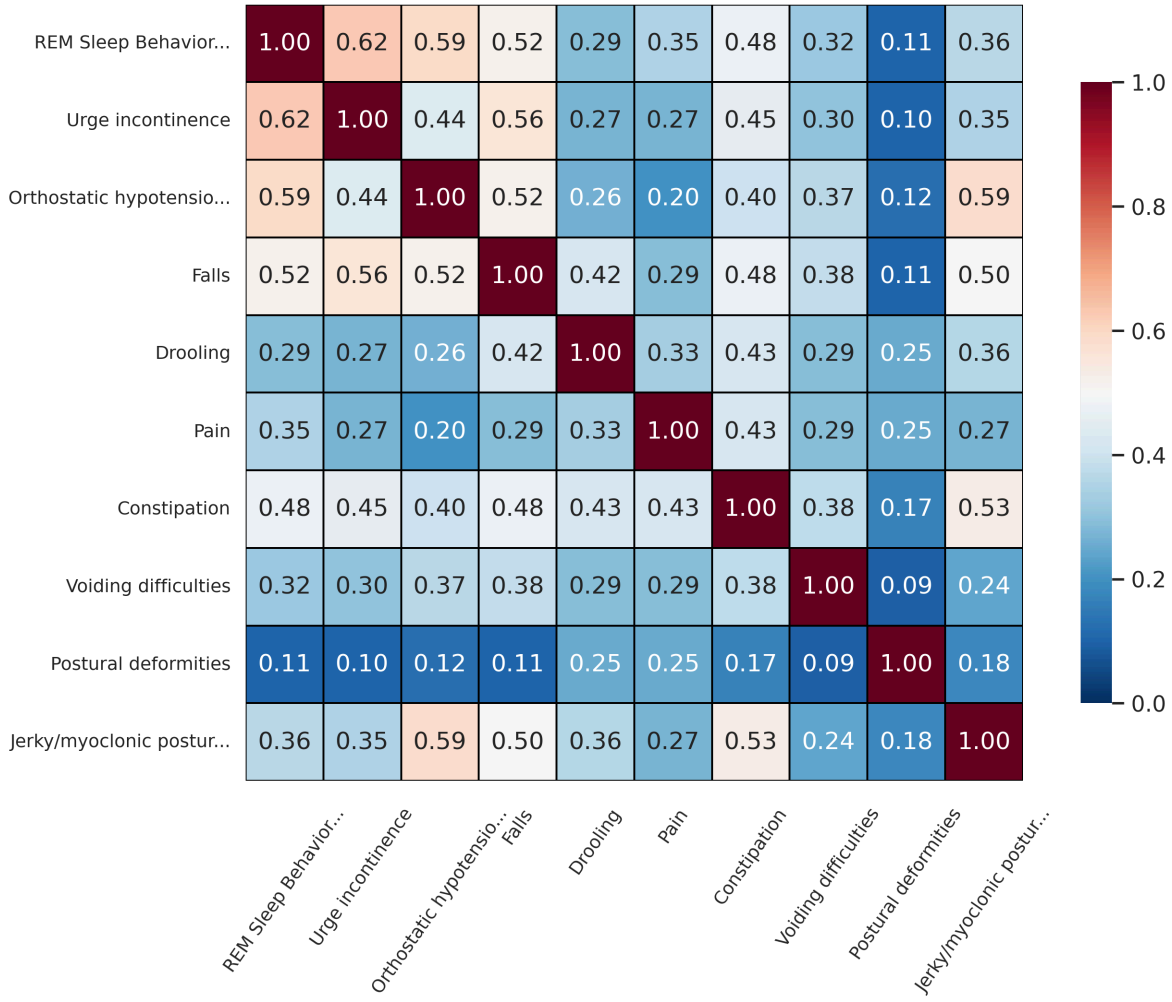
Methods:

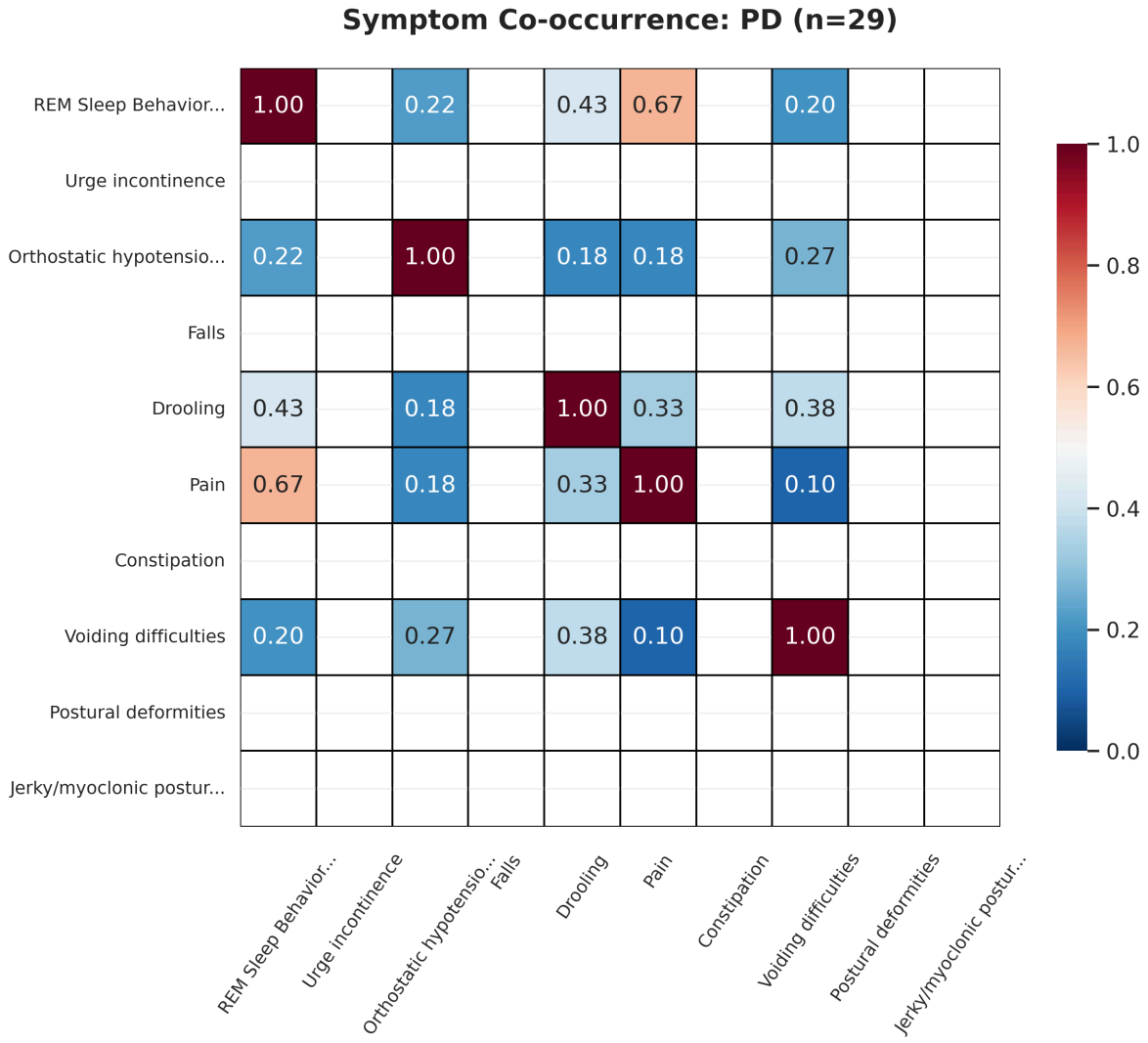
- **Jaccard index**: For each symptom pair, computed as the ratio of co-occurrence (intersection) to total presence (union): $J(A,B) = |A \cap B| / |A \cup B|$
- **Scale**: Ranges from 0 (symptoms never co-occur) to 1 (perfect overlap—every patient with symptom A also has symptom B, and vice versa)


Symptom Co-occurrence: MSA (P + C) (n=59)
REM Sleep Behavior... 1.00 0.54 0.55 0.46 0.29 0.30 0.46 0.35 0.34 0.31
Urge incontinence 0.54 1.00 0.47 0.50 0.35 0.22 0.53 0.28 0.38 0.35
Orthostatic hypotensio... 0.55 0.47 1.00 0.42 0.31 0.27 0.44 0.38 0.31 0.44
Falls 0.46 0.50 0.42 1.00 0.29 0.34 0.39 0.33 0.32 0.48
Drooling 0.29 0.35 0.31 0.29 1.00 0.26 0.37 0.24 0.35 0.39
Pain 0.30 0.22 0.27 0.34 0.26 1.00 0.36 0.27 0.34 0.30
Constipation 0.46 0.53 0.44 0.39 0.37 0.36 1.00 0.34 0.44 0.37
Voiding difficulties 0.35 0.28 0.38 0.33 0.24 0.27 0.34 1.00 0.25 0.25
Postural deformities 0.34 0.38 0.31 0.32 0.35 0.34 0.44 0.25 1.00 0.28
Jerky/myoclonic postur... 0.31 0.35 0.44 0.48 0.39 0.30 0.37 0.25 0.28 1.00
REM Sleep Behavior...
Urge incontinence
Orthostatic hypotensio...
Falls
Drooling
Pain
Constipation
Voiding difficulties
Postural deformities
Jerky/myoclonic postur...
1.0
0.8
0.6
0.4
0.2
0.0


Symptom Co-occurrence: MSA-P (n=33)
REM Sleep Behavior...
Urge incontinence
Orthostatic hypotensio...
Falls
Drooling
Pain
Constipation
Voiding difficulties
Postural deformities
Jerky/myoclonic postur...
1.00 0.47 0.52 0.41 0.29 0.27 0.45 0.38 0.48 0.28
0.47 1.00 0.50 0.45 0.42 0.17 0.59 0.26 0.57 0.36
0.52 0.50 1.00 0.33 0.35 0.33 0.48 0.40 0.41 0.33
0.41 0.45 0.33 1.00 0.19 0.40 0.32 0.27 0.46 0.45
0.29 0.42 0.35 0.19 1.00 0.21 0.33 0.21 0.38 0.42
0.27 0.17 0.33 0.40 0.21 1.00 0.32 0.25 0.38 0.33
0.45 0.59 0.48 0.32 0.33 0.32 1.00 0.32 0.56 0.27
0.38 0.26 0.40 0.27 0.21 0.25 0.32 1.00 0.32 0.26
0.48 0.57 0.41 0.46 0.38 0.38 0.56 0.32 1.00 0.32
0.28 0.36 0.33 0.45 0.42 0.33 0.27 0.26 0.32 1.00
1.0
0.8
0.6
0.4
0.2
0.0

Symptom Co-occurrence: MSA-C (n=26)





📝 **NOTE:** PD heatmaps lacks columns of the symptoms which are always not expressed (ie set to 0)

8. Motor Function and L-Dopa Responsiveness

Objective: Compare motor severity (UPDRS OFF) and treatment response across diagnoses. higher UPDRS values means severe motor impairment

Clinical Relevance:

- **UPDRS OFF score:** Measures motor impairment without medication, reflecting disease severity
- **L-dopa responsiveness: KEY DIAGNOSTIC CRITERION** distinguishing MSA from PD
  - PD: Typically >30% improvement with L-dopa (good response)
  - MSA: Poor L-dopa response (<30% improvement) is a core diagnostic feature

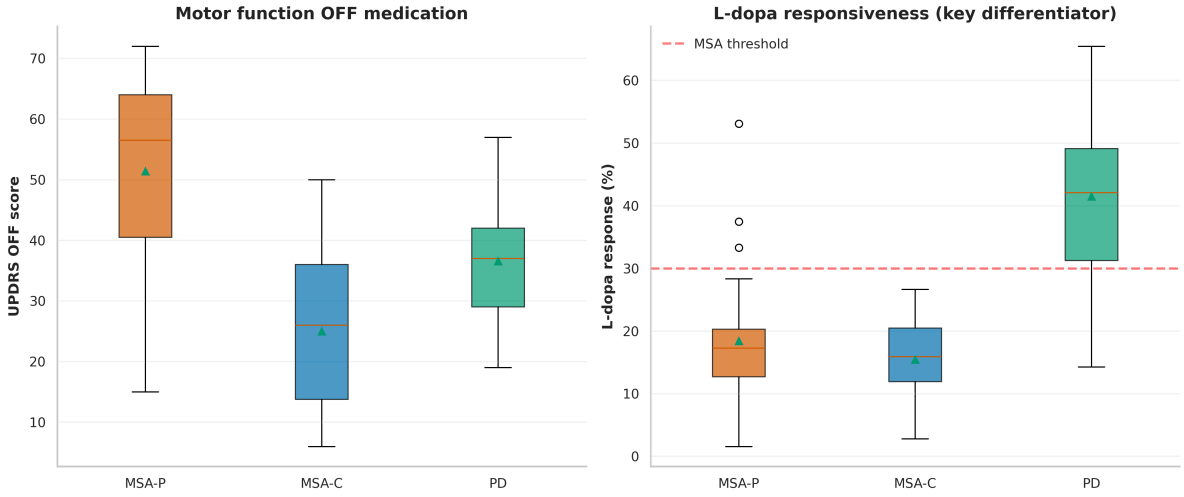
Clinical Importance: This is one of the most critical clinical differentiators between MSA and PD

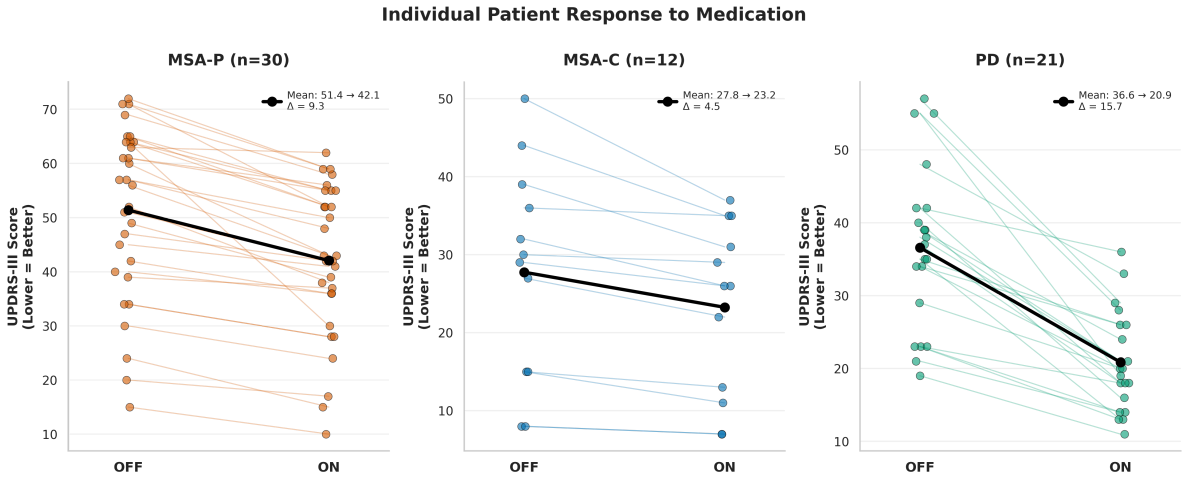
Expected Findings:

- Similar or higher UPDRS OFF scores in MSA (more severe motor impairment)
- Significantly lower L-dopa response in MSA vs. PD
- High inter-individual variability, especially in MSA

Results

- expected finding one and two are respected (can also be seen by the slope of the graph plotting values of UPDRS OFF and ON)
- although MSA-P presents outliers it doesnt show higher variability than PD



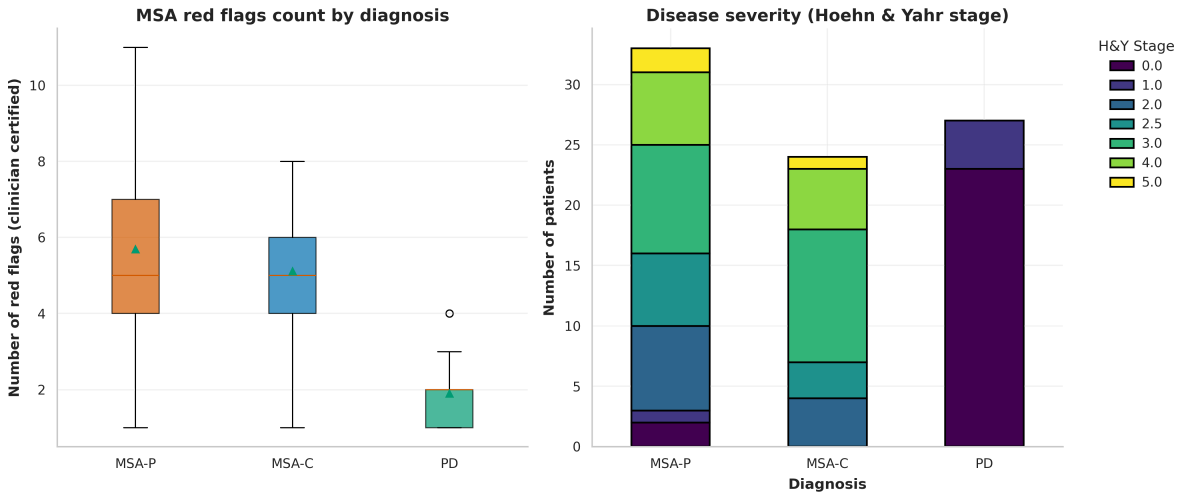


9. MSA Red Flags and Hoehn & Yahr Severity

Objective: Quantify MSA-specific clinical red flags and assess disease staging.

Clinical Relevance:

- **Red flags:** Clinical features suggestive of MSA (e.g., rapid progression, early autonomic failure, poor L-dopa response, cerebellar signs)
- **Hoehn & Yahr (H&Y) staging:** Standard PD staging system (0-5), also applicable to parkinsonian disorders
  - Stage 1-2: Unilateral/bilateral involvement, no balance impairment
  - Stage 3: Balance impairment, physically independent
  - Stage 4-5: Severe disability, wheelchair-bound



🚨 **IMPORTANT:** MSA patients are overwhelmingly in worse disease stages than PD ones

10. Diagnostic Delay and Progression

Objective: Quantify time from symptom onset to diagnosis and calculate progression rates.

Clinical Relevance:

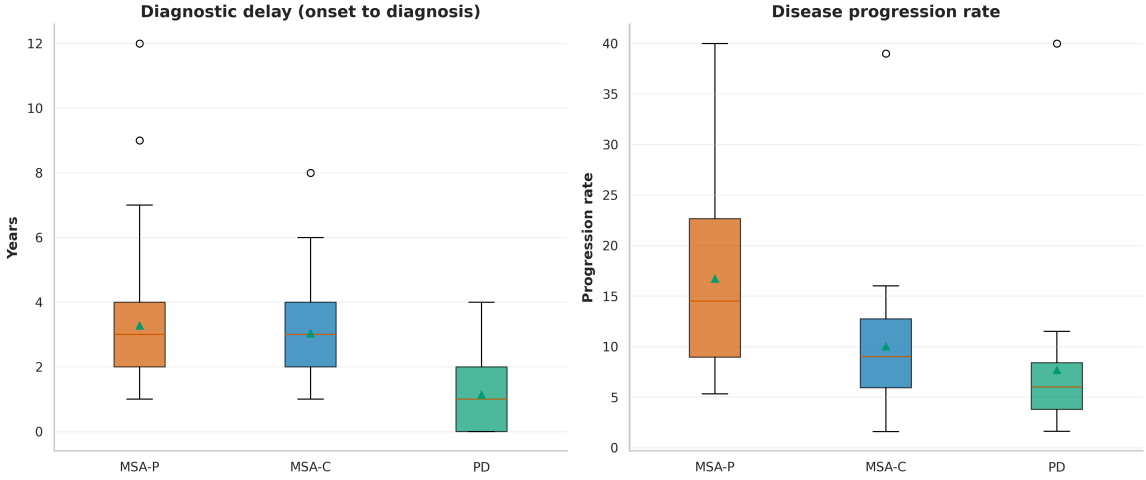
- **Diagnostic delay:** Time between symptom onset and formal diagnosis
  - Reflects diagnostic complexity and symptom overlap between conditions
  - Longer delays may indicate atypical presentations
- **Progression rate:** Speed of clinical decline (e.g., H&Y stage change per year)
  - MSA progresses faster than PD (key differentiator)
  - Rapid progression within 3 years is an MSA red flag

Expected Findings:

- MSA may show longer diagnostic delays due to initial misdiagnosis as PD
- MSA demonstrates faster progression rates than PD
- High variability in both measures reflects diagnostic complexity

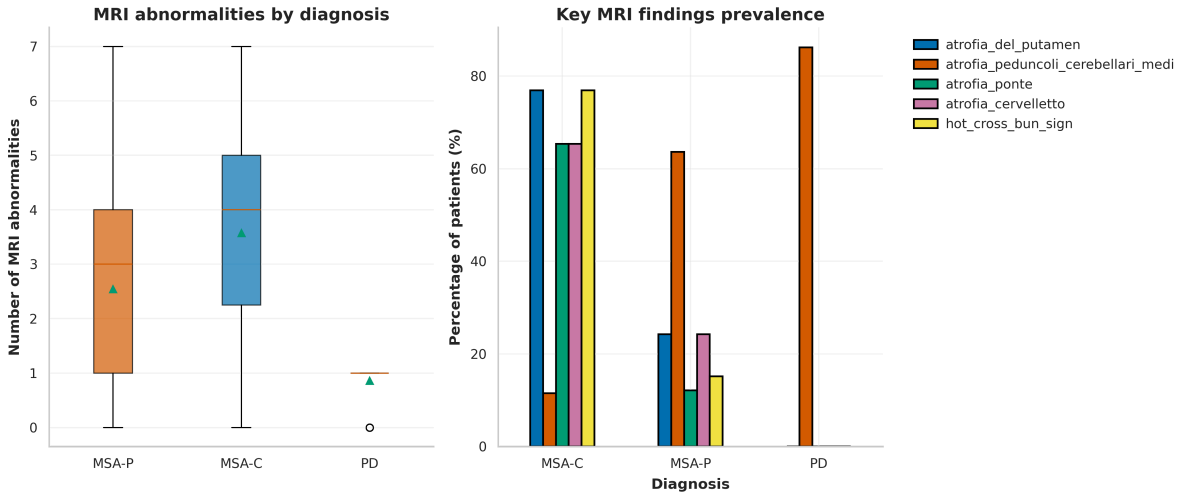
Results Analysis

- all expected findings are respected



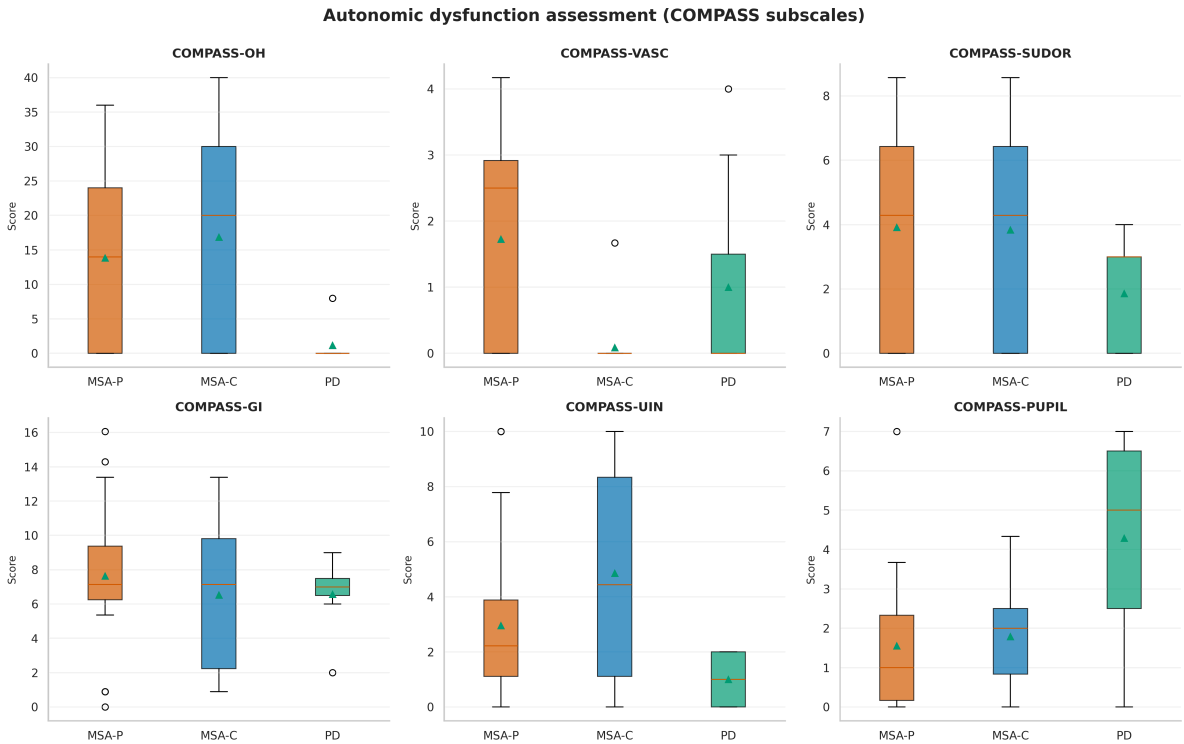
11. MRI Abnormalities (Supportive Features)

- Total abnormality count by diagnosis; specific signs: hot-cross-bun, putamen atrophy/signal changes.
- Clinical reading: MSA-C shows pontocerebellar signs; MSA-P shows putaminal changes; PD often near-normal MRI.



12. Autonomic Dysfunction Profile (COMPASS)

- Compare subscales COMPASS across classes.

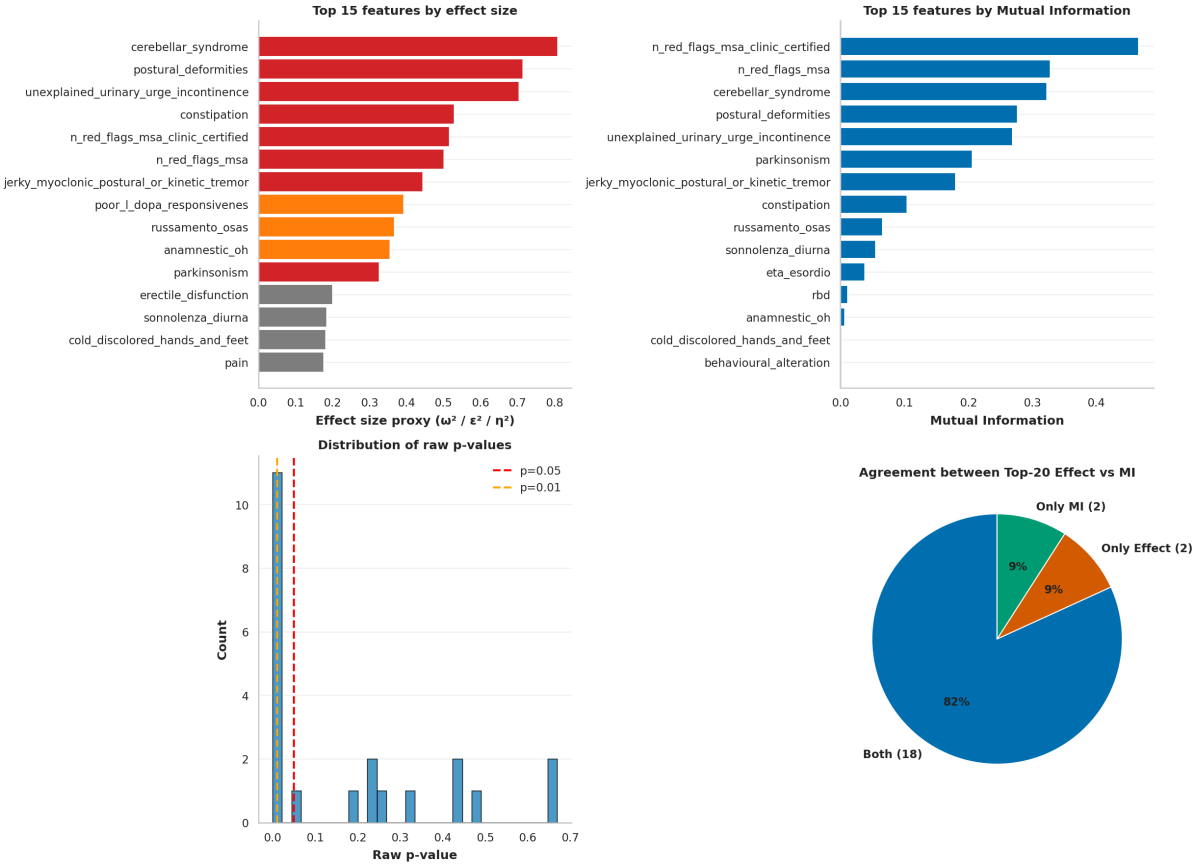


🚨 **IMPORTANT:** COMPASS is a high missing value (almost 40%) expecially for PD patients where almost 80% of them do not have a compass value

13. Univariate Screening and Mutual Information

Since the dataset contains both continouse (with different distribution) and categorical features proper statistical tests on early diagnoses features (ie features that a patient could present before being diagnosed/at first clinical visit) to understand which of those are statistically more relevant to distinguish diagnoses class and could potentially be integrated into a ML model.

- For each feature: appropriate test (ANOVA/Welch/Kruskal for continuous (decision based on feature distribution ie if they respect normality and homoschedacity); Chi-square for binary).
- Effect sizes and p-values aggregated and adjusted using FDR (Benjamini–Hochberg False Discovery Rate)
- **Mutual Information (MI)** is computed for each feature to quantify any **non-linear dependencies** between that variable and the diagnosis, beyond what ANOVA captures.
- A combined visualization summarizes top features by q-value/effect and MI.
- the best top K(=10) features (for both MI and ANOVA/Kruskal) are taken to be used with ML model.



Top 10 features by effect size and MI

Feature	Test	p_value	eta2	omega2	eps2	cramers_v	q_value	Significant	effect_proxy
parkinsonism	Kruskal	1.6422149786502O7e-05	nan	nan	0.32312514637521234	nan	5.161247075757793e-05	***	0.32312514637521234
anamnestic_oh	Chi2	0.004114689655306599	nan	nan	nan	0.35333450l1074575	0.008229379310613199	**	0.35333450l1074575
constipation	Chi2	5.143136601709189e-06	nan	nan	nan	0.52608864252265 62	1.885816753960036 2e-05	***	0.52608864252265 62
unexplained_urinary_urge_incontinence	Chi2	3.847906567135673 4e-10	nan	nan	nan	0.7019181783176 48	1.693078889539696 5e-09	***	0.7019181783176 48
cold_discolored_hands_and_feet	Chi2	0.24434460023318857	nan	nan	nan	0.17896011198558773	0.35837208034200996		0.17896011198558773
sonnolenza_diurna	Chi2	0.23613785519691022	nan	nan	nan	0.18111645627663214	0.35837208034200996		0.18111645627663214
russamento_osas	Chi2	0.00293025460397775 73	nan	nan	nan	0.36408871184758895	0.00686260971687352 45	**	0.36408871184758895
cerebellar_syndrome	Chi2	3.876032807057358e-13	nan	nan	nan	0.8059268245992087	8.527272175526187e-12	***	0.8059268245992087
postural_deformities	Chi2	2.011439274802452 3e-10	nan	nan	nan	0.7123424930690485	1.236326691987281e-09	***	0.7123424930690485
jerky_myoclonic_postural_or_kinetic_tremor	Chi2	0.00018740500100840 78	nan	nan	nan	0.44164564421136887	0.00051536375277312 14	***	0.44164564421136887

14. Differential and Within-Diagnosis Correlations

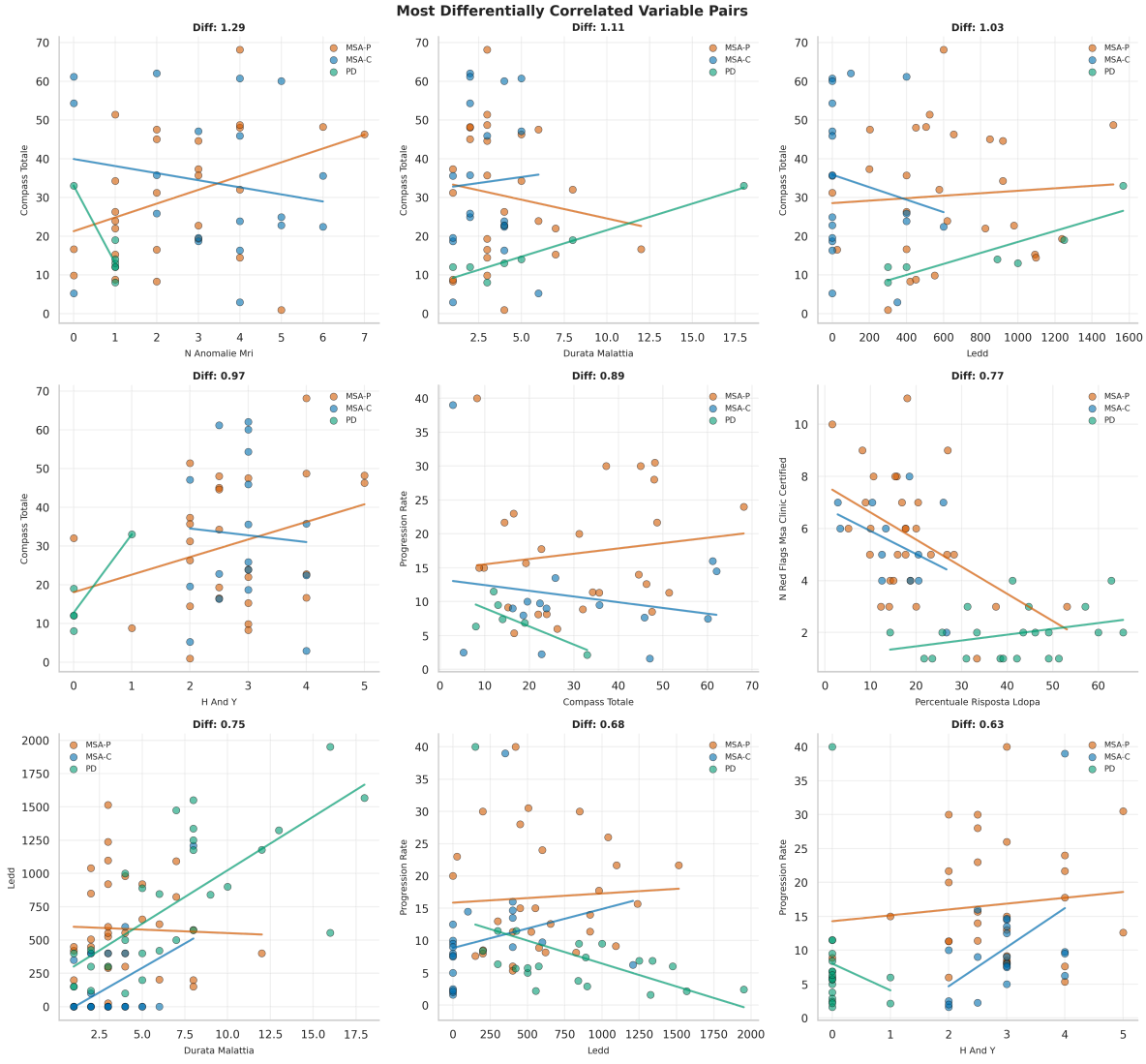
- Differential correlations: identify pairs of continouse variables whose association differs between diagnoses.
- Within-diagnosis correlation heatmaps (MSA-P, MSA-C, PD) for key clinical variables.

Method:

1. Calculate correlation matrices **separately** for each diagnosis (MSA-P, MSA-C, PD)
2. For each variable pair, compute the **maximum absolute difference** in correlation coefficients across groups:

$$\text{Max Diff} = \max\left(|r_{\text{MSA-P}} - r_{\text{PD}}|, \quad |r_{\text{MSA-C}} - r_{\text{PD}}|, \quad |r_{\text{MSA-P}} - r_{\text{MSA-C}}| \right)$$

3. Rank pairs by Max Diff to identify **most differentially correlated** relationships



Main Insights

- **Autonomic–Structural Link**
  Compass Totale increases with *N° Anomalie MRI* only in **MSA-P**, suggesting that structural damage (putamen/cerebellum) parallels autonomic failure — absent in PD.

  strange MSA-C inverse correlation.

- **Disease Duration Effect**
  Compass Totale rises with *Durata Malattia* in in PD and MSA-C indicating **progressive autonomic decline**

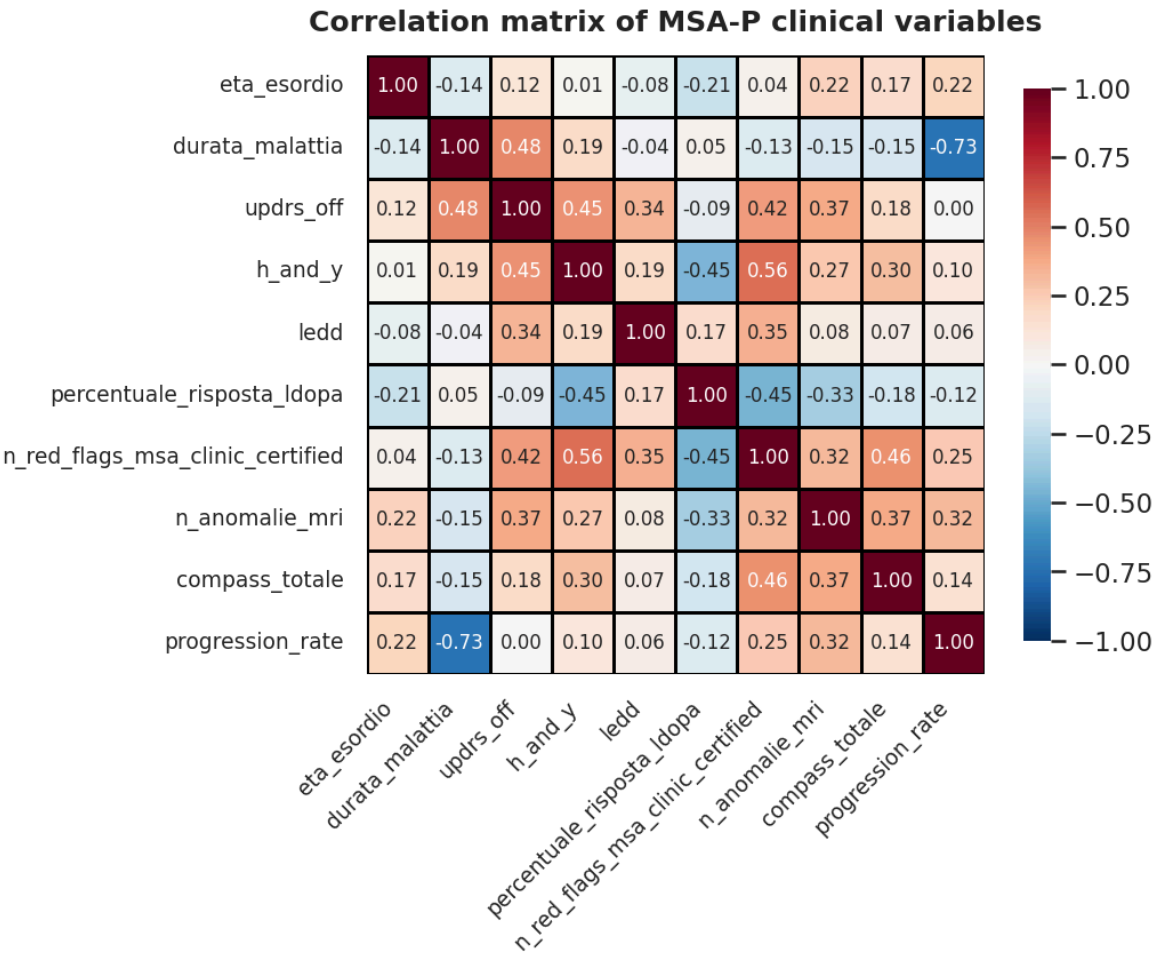
  strange MSA-P inverse relationship

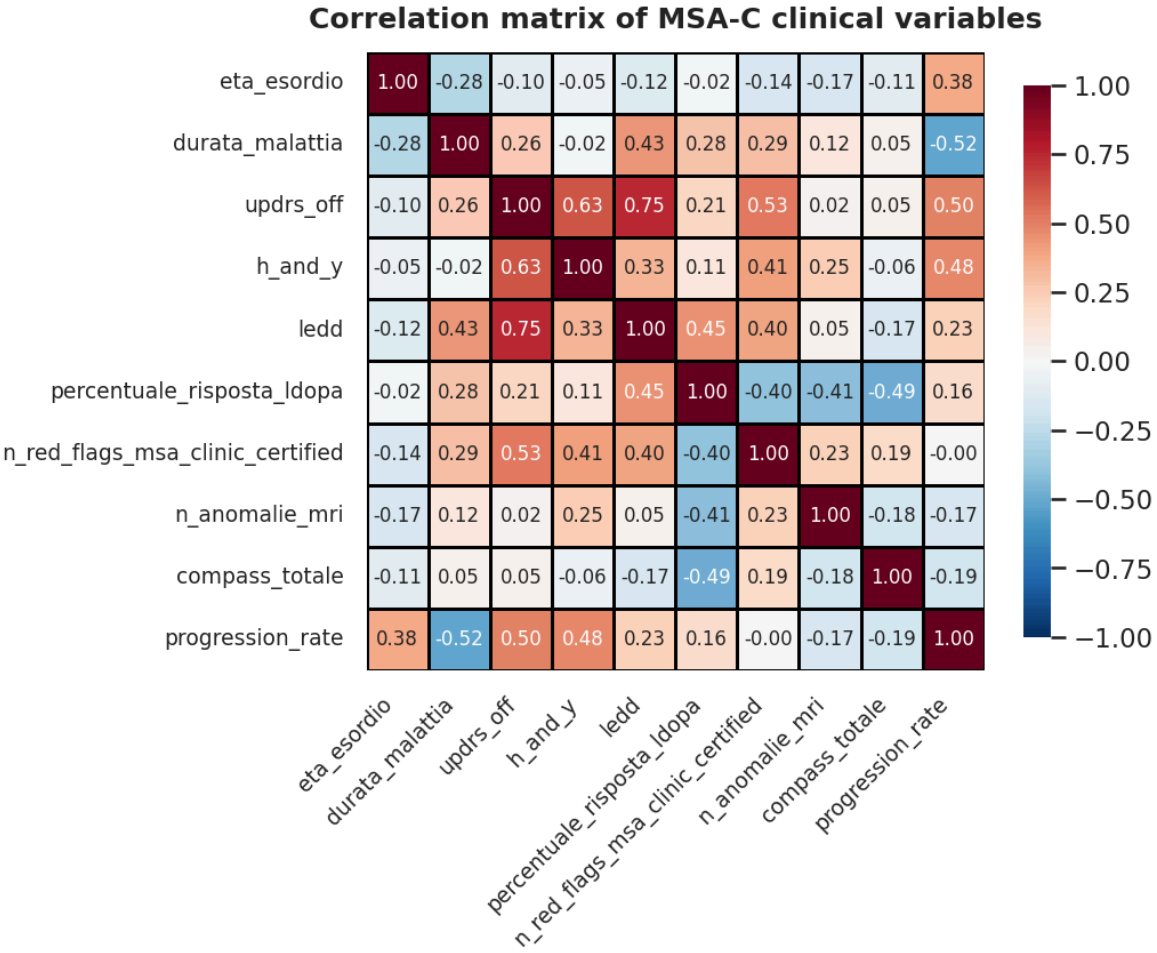
- **Medication Dynamics**
  - In PD, *LEDD* scales with disease duration (normal titration).
  - In MSA, higher *LEDD* does **not** improve *Compass Totale* or *Progression Rate* as much as it does for Pd patients consistent with **poor dopaminergic responsiveness** of MSA.
- **Motor–Autonomic Coupling**
  H&Y correlates positively with *Compass Totale* and *Progression Rate* in MSA, but not in PD supporting **parallel motor and autonomic progression** in MSA.

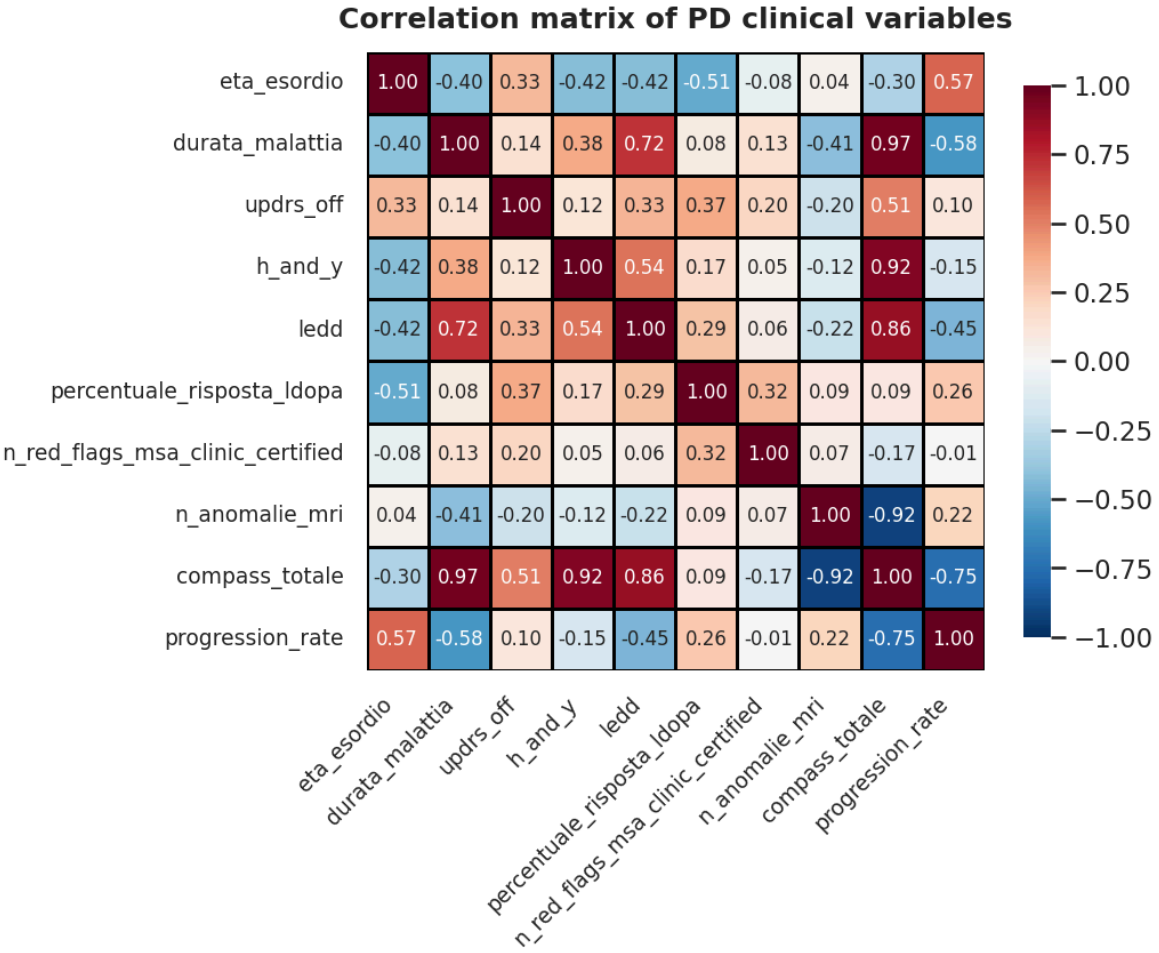
- **Diagnostic Red Flags**
  Patients with low *% L-dopa response* show more *MSA red flags*, consistent with **poor dopaminergic responsiveness** of MSA.

> 🚨 **IMPORTANT:** COMPASS is a high missing value (almost 40%) expecially for PD patients where almost 80% of them do not have a compass value

Within diagnosis symptoms correlation (Pearson)




Correlation matrix of MSA-C clinical variables
eta_esordio 1.00 -0.28 -0.10 -0.05 -0.12 -0.02 -0.14 -0.17 -0.11 0.38
durata_malattia -0.28 1.00 0.26 -0.02 0.43 0.28 0.29 0.12 0.05 -0.52
updrs_off -0.10 0.26 1.00 0.63 0.75 0.21 0.53 0.02 0.05 0.50
h_and_y -0.05 -0.02 0.63 1.00 0.33 0.11 0.41 0.25 -0.06 0.48
ledd -0.12 0.43 0.75 0.33 1.00 0.45 0.40 0.05 -0.17 0.23
percentuale_risposta_ldopa -0.02 0.28 0.21 0.11 0.45 1.00 -0.40 -0.41 -0.49 0.16
n_red_flags_msa_clinic_certified -0.14 0.29 0.53 0.41 0.40 -0.40 1.00 0.23 0.19 -0.00
n_anomalie_mri -0.17 0.12 0.02 0.25 0.05 -0.41 0.23 1.00 -0.18 -0.17
compass_totale -0.11 0.05 0.05 -0.06 -0.17 -0.49 0.19 -0.18 1.00 -0.19
progression_rate 0.38 -0.52 0.50 0.48 0.23 0.16 -0.00 -0.17 -0.19 1.00
eta_esordio
durata_malattia
updrs_off
h_and_y
ledd
percentuale_risposta_ldopa
n_red_flags_msa_clinic_certified
n_anomalie_mri
compass_totale
progression_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Correlation matrix of PD clinical variables
eta_esordio 1.00 -0.40 0.33 -0.42 -0.42 -0.51 -0.08 0.04 -0.30 0.57
durata_malattia -0.40 1.00 0.14 0.38 0.72 0.08 0.13 -0.41 0.97 -0.58
updrs_off 0.33 0.14 1.00 0.12 0.33 0.37 0.20 -0.20 0.51 0.10
h_and_y -0.42 0.38 0.12 1.00 0.54 0.17 0.05 -0.12 0.92 -0.15
ledd -0.42 0.72 0.33 0.54 1.00 0.29 0.06 -0.22 0.86 -0.45
percentuale_risposta_ldopa -0.51 0.08 0.37 0.17 0.29 1.00 0.32 0.09 0.09 0.26
n_red_flags_msa_clinic_certified -0.08 0.13 0.20 0.05 0.06 0.32 1.00 0.07 -0.17 -0.01
n_anomalie_mri 0.04 -0.41 -0.20 -0.12 -0.22 0.09 0.07 1.00 -0.92 0.22
compass_totale -0.30 0.97 0.51 0.92 0.86 0.09 -0.17 -0.92 1.00 -0.75
progression_rate 0.57 -0.58 0.10 -0.15 -0.45 0.26 -0.01 0.22 -0.75 1.00
eta_esordio
durata_malattia
updrs_off
h_and_y
ledd
percentuale_risposta_ldopa
n_red_flags_msa_clinic_certified
n_anomalie_mri
compass_totale
progression_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

15. ML model integration

Highly discriminative easy-to-collect features are selected (selected in step 13):

- 'unexplained_urinary_urge_incontinence'
- 'russamento_osas'
- 'anamnestic_oh'
- 'sonnolenza_diurna'
- 'cold_discolored_hands_and_feet'
- 'cerebellar_syndrome'
- 'postural_deformities'
- 'parkinsonism'
- 'constipation'
- 'jerky_myoclonic_postural_or_kinetic_tremor'

few simple ML/statistical models are used to make predictions.

- dummy = random guess
- logistic regression (logreg)
- random forest (rf)

8 fold cross validation is performed

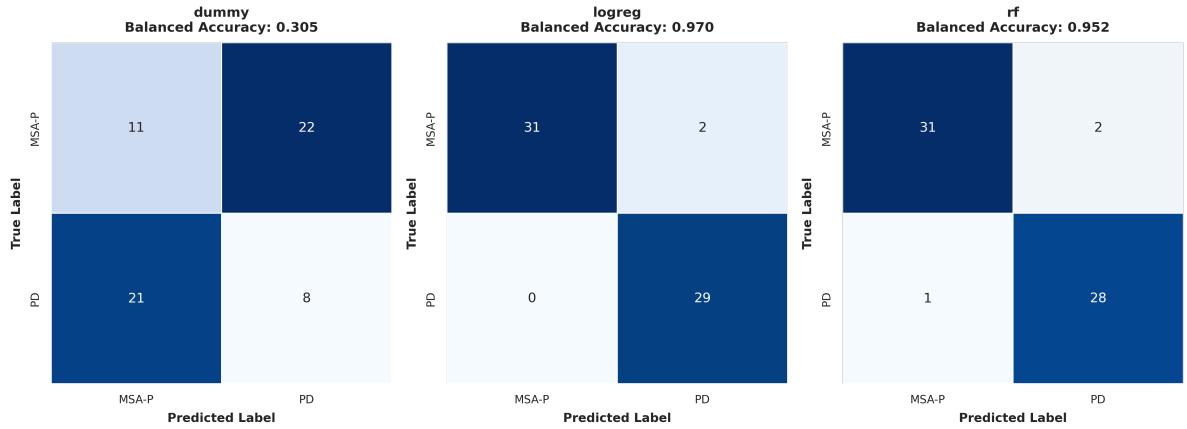
The **Matthews Correlation Coefficient (MCC)** is included as a key metric for model evaluation instead of f1.

- **MCC** is a balanced measure, even for imbalanced classes, giving a value between -1 (inverse prediction) and +1 (perfect prediction); 0 indicates random performance.

Model Comparison Summary

Model	BAC	Macro F1	MCC	ROC-AUC
dummy	0.3010 ± 0.1187	0.3008 ± 0.1186	-0.3969 ± 0.2381	0.3046
logreg	0.9688 ± 0.0579	0.9683 ± 0.0588	0.9436 ± 0.1043	0.9472
rf	0.9531 ± 0.0647	0.9524 ± 0.0657	0.9155 ± 0.1167	0.9404

OUT OF FOLD CONFUSION MATRIX aggregates prediction for each test fold data. ie these are the prediction aggregated over each fold test set



⚠️ **TODO:** Clinician features review