Supplementary material for “Predicting correlated outcomes from molecular data”

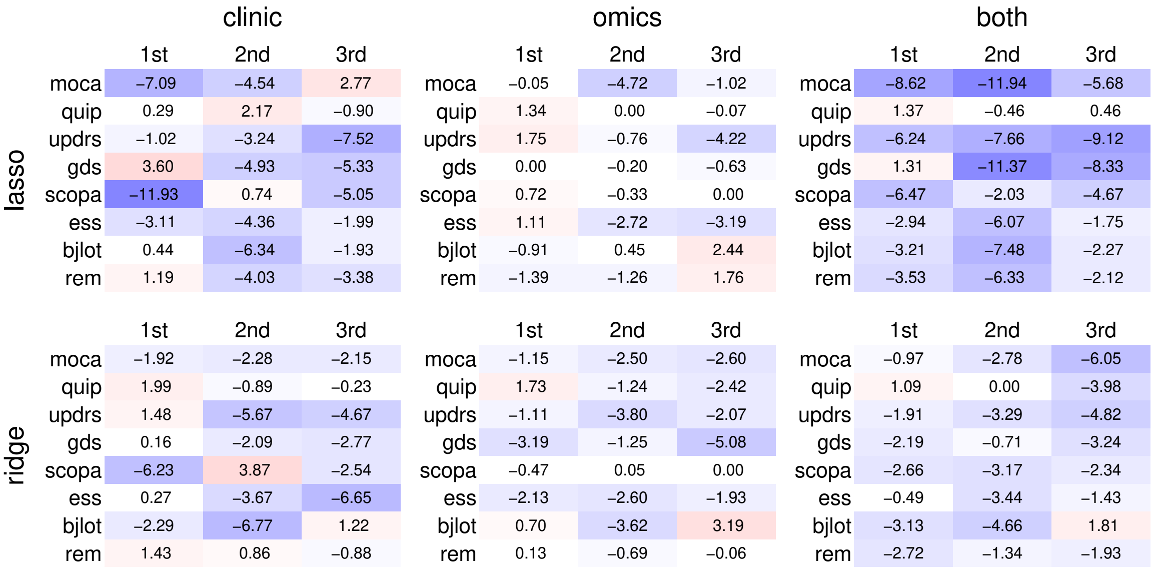
Armin Rauschenberger, Enrico Glaab

Luxembourg Centre for Systems Biomedicine (LCSB),
University of Luxembourg, Esch-sur-Alzette, Luxembourg

This supplementary material includes further information on the application. Table A describes the pre-processing of clinical and genomic data. Figures A and B show the change in predictive performance from univariate to multivariate regression. For reproducibility of the simulation and the application, see the vignettes of the R package `joinet` (https://github.com/rauschenberger/joinet, https://cran.r-project.org/package=joinet).

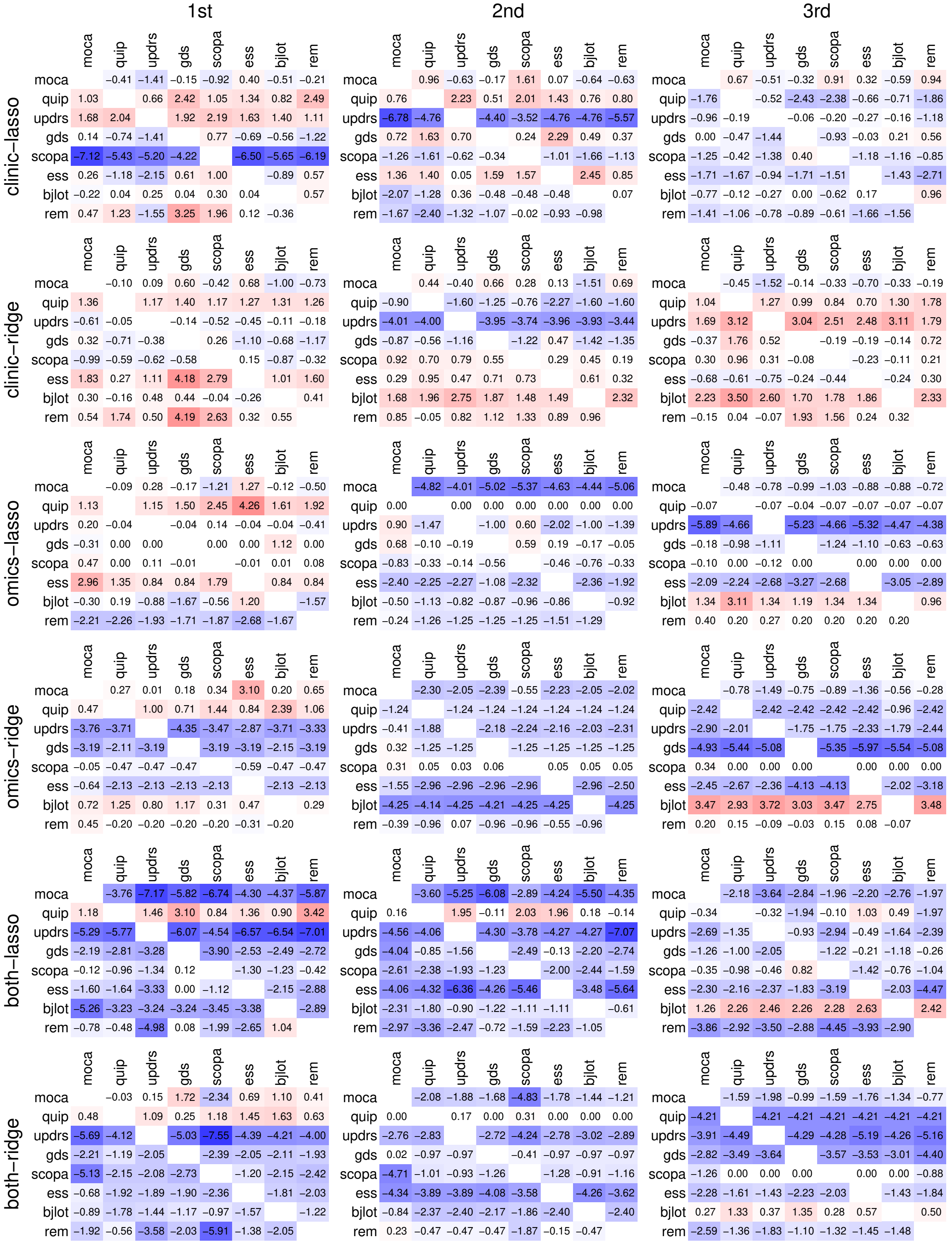
Table A: Pre-processing of clinical and genomic data.

clinical inputs	From the curated baseline clinical data (“PPMI_Baseline_Data_02Jul2018.csv”), a table with 683 rows (samples) and 134 columns (variables), we retain the 423 samples with Parkinson’s disease (i.e. apprdx=1) and the 130 clinical variables (i.e. without “site”, “apprdx”, “event_id” and “symptom5_comment”). After imputing the missing values ($\approx 3\%$) by chained random forests (R package `missRanger`), we transform the categorical variables (i.e. “apoe”, “snca_rs356181”, “snca_rs3910105” and “mapt”) to dummy variables. This leads to a matrix with 423 rows (samples) and 138 columns (variables).
genomic inputs	We use the count matrix generated by Zhi Zhang’s RNA-seq pipeline (Python package `sumrnaseq.main 0.1.0`, https://pypi.org/project/sumrnaseq.main/). From the baseline genomic data (“ppmi_rnaseq_bl_pd_hc-2019-01-11.Rdata”), a table with 363 rows (samples) and 57 820 columns (variables), we remove the transcripts without an HGNC gene symbol (i.e. names starting with ENSG0000 or ENSGR0000). This reduces the dimensionality by approximately 38%. After calculating the library sizes (total count for each sample) and the abundances (total count for each gene), we remove the genes with a low abundance (less than 10 times the sample size), and adjust the counts for different library sizes (R package `edgeR`). We then Anscombe-transform ($x \to 2\sqrt{x+3/8}$) and standardise the counts (mean zero and variance one for each gene). This leads to a matrix with 363 rows (samples) and 17 714 columns (variables).
clinical outputs	From the curated follow-up clinical data (“PPMI_Year_1-3_Data_02Jul2018.csv”), a table with 1 783 rows (samples, visits) and 124 columns (variables), we retain the visits of interest (i.e. “V04”, “V06” and “V08” but not “ST”) and the variables of interest (i.e. “moca”, “quip”, “updrs_totscore”, “gds”, “scopa”, “ess”, “bjlot” and “rem”) together with the sample and visit identifiers (“PATNO” and “EVENT_ID”). We then reshape the data from the long to the wide format. This leads to a matrix with 654 rows (samples) and $3 \times 8 = 24$ columns (variables, visits).



		clinic			omics			both		
		1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
lasso	moca	−7.09	−4.54	2.77	−0.05	−4.72	−1.02	−8.62	−11.94	−5.68
	quip	0.29	2.17	−0.90	1.34	0.00	−0.07	1.37	−0.46	0.46
	updrs	−1.02	−3.24	−7.52	1.75	−0.76	−4.22	−6.24	−7.66	−9.12
	gds	3.60	−4.93	−5.33	0.00	−0.20	−0.63	1.31	−11.37	−8.33
	scopa	−11.93	0.74	−5.05	0.72	−0.33	0.00	−6.47	−2.03	−4.67
	ess	−3.11	−4.36	−1.99	1.11	−2.72	−3.19	−2.94	−6.07	−1.75
	bjlot	0.44	−6.34	−1.93	−0.91	0.45	2.44	−3.21	−7.48	−2.27
	rem	1.19	−4.03	−3.38	−1.39	−1.26	1.76	−3.53	−6.33	−2.12
		1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
ridge	moca	−1.92	−2.28	−2.15	−1.15	−2.50	−2.60	−0.97	−2.78	−6.05
	quip	1.99	−0.89	−0.23	1.73	−1.24	−2.42	1.09	0.00	−3.98
	updrs	1.48	−5.67	−4.67	−1.11	−3.80	−2.07	−1.91	−3.29	−4.82
	gds	0.16	−2.09	−2.77	−3.19	−1.25	−5.08	−2.19	−0.71	−3.24
	scopa	−6.23	3.87	−2.54	−0.47	0.05	0.00	−2.66	−3.17	−2.34
	ess	0.27	−3.67	−6.65	−2.13	−2.60	−1.93	−0.49	−3.44	−1.43
	bjlot	−2.29	−6.77	1.22	0.70	−3.62	3.19	−3.13	−4.66	1.81
	rem	1.43	0.86	−0.88	0.13	−0.69	−0.06	−2.72	−1.34	−1.93

Figure A: Percentage change in cross-validated mean squared error from univariate to multivariate regression, with decreases in blue and increases in red. We consider six different settings, namely lasso or ridge regularisation (outer row) for clinical data, omics data, or both (outer column). Each multivariate model supports the prediction for one tool (inner row) at one visit (inner column) with the same tool at the other visits.



clinic–lasso

1st

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		−0.41	−1.41	−0.15	−0.92	0.40	−0.51	−0.21
quip	1.03		0.66	2.42	1.05	1.34	0.82	2.49
updrs	1.68	2.04		1.92	2.19	1.63	1.40	1.11
gds	0.14	−0.74	−1.41		0.77	−0.69	−0.56	−1.22
scopa	−7.12	−5.43	−5.20	−4.22		−6.50	−5.65	−6.19
ess	0.26	−1.18	−2.15	0.61	1.00		−0.89	0.57
bjlot	−0.22	0.04	0.25	0.04	0.30	0.04		0.57
rem	0.47	1.23	−1.55	3.25	1.96	0.12	−0.36	

2nd

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		0.96	−0.63	−0.17	1.61	0.07	−0.64	−0.63
quip	0.76		2.23	0.51	2.01	1.43	0.76	0.80
updrs	−6.78	−4.76		−4.40	−3.52	−4.76	−4.76	−5.57
gds	0.72	1.63	0.70		0.24	2.29	0.49	0.37
scopa	−1.26	−1.61	−0.62	−0.34		−1.01	−1.66	−1.13
ess	1.36	1.40	0.05	1.59	1.57		2.45	0.85
bjlot	−2.07	−1.28	0.36	−0.48	−0.48	−0.48		0.07
rem	−1.67	−2.40	−1.32	−1.07	−0.02	−0.93	−0.98	

3rd

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		0.67	−0.51	−0.32	0.91	0.32	−0.59	0.94
quip	−1.76		−0.52	−2.43	−2.38	−0.66	−0.71	−1.86
updrs	−0.96	−0.19		−0.06	−0.20	−0.27	−0.16	−1.18
gds	0.00	−0.47	−1.44		−0.93	−0.03	0.21	0.56
scopa	−1.25	−0.42	−1.38	0.40		−1.18	−1.16	−0.85
ess	−1.71	−1.67	−0.94	−1.71	−1.51		−1.43	−2.71
bjlot	−0.77	−0.12	−0.27	0.00	−0.62	0.17		0.96
rem	−1.41	−1.06	−0.78	−0.89	−0.61	−1.66	−1.56	

clinic–ridge

1st

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		−0.10	0.09	0.60	−0.42	0.68	−1.00	−0.73
quip	1.36		1.17	1.40	1.17	1.27	1.31	1.26
updrs	−0.61	−0.05		−0.14	−0.52	−0.45	−0.11	−0.18
gds	0.32	−0.71	−0.38		0.26	−1.10	−0.68	−1.17
scopa	−0.99	−0.59	−0.62	−0.58		0.15	−0.87	−0.32
ess	1.83	0.27	1.11	4.18	2.79		1.01	1.60
bjlot	0.30	−0.16	0.48	0.44	−0.04	−0.26		0.41
rem	0.54	1.74	0.50	4.19	2.63	0.32	0.55	

2nd

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		0.44	−0.40	0.66	0.28	0.13	−1.51	0.69
quip	−0.90		−1.60	−1.25	−0.76	−2.27	−1.60	−1.60
updrs	−4.01	−4.00		−3.95	−3.74	−3.96	−3.93	−3.44
gds	−0.87	−0.56	−1.16		−1.22	0.47	−1.42	−1.35
scopa	0.92	0.70	0.79	0.55		0.29	0.45	0.19
ess	0.29	0.95	0.47	0.71	0.73		0.61	0.32
bjlot	1.68	1.96	2.75	1.87	1.48	1.49		2.32
rem	0.85	−0.05	0.82	1.12	1.33	0.89	0.96	

3rd

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		−0.45	−1.52	−0.14	−0.33	−0.70	−0.33	−0.19
quip	1.04		1.27	0.99	0.84	0.70	1.30	1.78
updrs	1.69	3.12		3.04	2.51	2.48	3.11	1.79
gds	−0.37	1.76	0.52		−0.19	−0.19	−0.14	0.72
scopa	0.30	0.96	0.31	−0.08		−0.23	−0.11	0.21
ess	−0.68	−0.61	−0.75	−0.24	−0.44		−0.24	0.30
bjlot	2.23	3.50	2.60	1.70	1.78	1.86		2.33
rem	−0.15	0.04	−0.07	1.93	1.56	0.24	0.32	

omics–lasso

1st

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		−0.09	0.28	−0.17	−1.21	1.27	−0.12	−0.50
quip	1.13		1.15	1.50	2.45	4.26	1.61	1.92
updrs	0.20	−0.04		−0.04	0.14	−0.04	−0.04	−0.41
gds	−0.31	0.00	0.00		0.00	0.00	1.12	0.00
scopa	0.47	0.00	0.11	−0.01		−0.01	0.01	0.08
ess	2.96	1.35	0.84	0.84	1.79		0.84	0.84
bjlot	−0.30	0.19	−0.88	−1.67	−0.56	1.20		−1.57
rem	−2.21	−2.26	−1.93	−1.71	−1.87	−2.68	−1.67	

2nd

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		−4.82	−4.01	−5.02	−5.37	−4.63	−4.44	−5.06
quip	0.00		0.00	0.00	0.00	0.00	0.00	0.00
updrs	0.90	−1.47		−1.00	0.60	−2.02	−1.00	−1.39
gds	0.68	−0.10	−0.19		0.59	0.19	−0.17	−0.05
scopa	−0.83	−0.33	−0.14	−0.56		−0.46	−0.76	−0.33
ess	−2.40	−2.25	−2.27	−1.08	−2.32		−2.36	−1.92
bjlot	−0.50	−1.13	−0.82	−0.87	−0.96	−0.86		−0.92
rem	−0.24	−1.26	−1.25	−1.25	−1.25	−1.51	−1.29	

3rd

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		−0.48	−0.78	−0.99	−1.03	−0.88	−0.88	−0.72
quip	−0.07		−0.07	−0.04	−0.07	−0.07	−0.07	−0.07
updrs	−5.89	−4.66		−5.23	−4.66	−5.32	−4.47	−4.38
gds	−0.18	−0.98	−1.11		−1.24	−1.10	−0.63	−0.63
scopa	−0.10	0.00	−0.12	0.00		0.00	0.00	0.00
ess	−2.09	−2.24	−2.68	−3.27	−2.68		−3.05	−2.89
bjlot	1.34	3.11	1.34	1.19	1.34	1.34		0.96
rem	0.40	0.20	0.27	0.20	0.20	0.20	0.20	

omics–ridge

1st

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		0.27	0.01	0.18	0.34	3.10	0.20	0.65
quip	0.47		1.00	0.71	1.44	0.84	2.39	1.06
updrs	−3.76	−3.71		−4.35	−3.47	−2.87	−3.71	−3.33
gds	−3.19	−2.11	−3.19		−3.19	−3.19	−2.15	−3.19
scopa	−0.05	−0.47	−0.47	−0.47		−0.59	−0.47	−0.47
ess	−0.64	−2.13	−2.13	−2.13	−2.13		−2.13	−2.13
bjlot	0.72	1.25	0.80	1.17	0.31	0.47		0.29
rem	0.45	−0.20	−0.20	−0.20	−0.20	−0.31	−0.20	

2nd

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		−2.30	−2.05	−2.39	−0.55	−2.23	−2.05	−2.02
quip	−1.24		−1.24	−1.24	−1.24	−1.24	−1.24	−1.24
updrs	−0.41	−1.88		−2.18	−2.24	−2.16	−2.03	−2.31
gds	0.32	−1.25	−1.25		−1.25	−1.25	−1.25	−1.25
scopa	0.31	0.05	0.03	0.06		0.05	0.05	0.05
ess	−1.55	−2.96	−2.96	−2.96	−2.96		−2.96	−2.50
bjlot	−4.25	−4.14	−4.25	−4.21	−4.25	−4.25		−4.25
rem	−0.39	−0.96	0.07	−0.96	−0.96	−0.55	−0.96	

3rd

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		−0.78	−1.49	−0.75	−0.89	−1.36	−0.56	−0.28
quip	−2.42		−2.42	−2.42	−2.42	−2.42	−0.96	−2.42
updrs	−2.90	−2.01		−1.75	−1.75	−2.33	−1.79	−2.44
gds	−4.93	−5.44	−5.08		−5.35	−5.97	−5.54	−5.08
scopa	0.34	0.00	0.00	0.00		0.00	0.00	0.00
ess	−2.45	−2.67	−2.36	−4.13	−4.13		−2.02	−3.18
bjlot	3.47	2.93	3.72	3.03	3.47	2.75		3.48
rem	0.20	0.15	−0.09	−0.03	0.15	0.08	−0.07	

both–lasso

1st

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		−3.76	−7.17	−5.82	−6.74	−4.30	−4.37	−5.87
quip	1.18		1.46	3.10	0.84	1.36	0.90	3.42
updrs	−5.29	−5.77		−6.07	−4.54	−6.57	−6.54	−7.01
gds	−2.19	−2.81	−3.28		−3.90	−2.53	−2.49	−2.72
scopa	−0.12	−0.96	−1.34	0.12		−1.30	−1.23	−0.42
ess	−1.60	−1.64	−3.33	0.00	−1.12		−2.15	−2.88
bjlot	−5.26	−3.23	−3.24	−3.24	−3.45	−3.38		−2.89
rem	−0.78	−0.48	−4.98	0.08	−1.99	−2.65	1.04	

2nd

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		−3.60	−5.25	−6.08	−2.89	−4.24	−5.50	−4.35
quip	0.16		1.95	−0.11	2.03	1.96	0.18	−0.14
updrs	−4.56	−4.06		−4.30	−3.78	−4.27	−4.27	−7.07
gds	−4.04	−0.85	−1.56		−2.49	−0.13	−2.20	−2.74
scopa	−2.61	−2.38	−1.93	−1.23		−2.00	−2.44	−1.59
ess	−4.06	−4.32	−6.36	−4.26	−5.46		−3.48	−5.64
bjlot	−2.31	−1.80	−0.90	−1.22	−1.11	−1.11		−0.61
rem	−2.97	−3.36	−2.47	−0.72	−1.59	−2.23	−1.05	

3rd

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		−2.18	−3.64	−2.84	−1.96	−2.20	−2.76	−1.97
quip	−0.34		−0.32	−1.94	−0.10	1.03	0.49	−1.97
updrs	−2.69	−1.35		−0.93	−2.94	−0.49	−1.64	−2.39
gds	−1.26	−1.00	−2.05		−1.22	−0.21	−1.18	−0.26
scopa	−0.35	−0.98	−0.46	0.82		−1.42	−0.76	−1.04
ess	−2.30	−2.16	−2.37	−1.83	−3.19		−2.03	−4.47
bjlot	1.26	2.26	2.46	2.26	2.28	2.63		2.42
rem	−3.86	−2.92	−3.50	−2.88	−4.45	−3.93	−2.90	

both–ridge

1st

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		−0.03	0.15	1.72	−2.34	0.69	1.10	0.41
quip	0.48		1.09	0.25	1.18	1.45	1.63	0.63
updrs	−5.69	−4.12		−5.03	−7.55	−4.39	−4.21	−4.00
gds	−2.21	−1.19	−2.05		−2.39	−2.05	−2.11	−1.93
scopa	−5.13	−2.15	−2.08	−2.73		−1.20	−2.15	−2.42
ess	−0.68	−1.92	−1.89	−1.90	−2.36		−1.81	−2.03
bjlot	−0.89	−1.78	−1.44	−1.17	−0.97	−1.57		−1.22
rem	−1.92	−0.56	−3.58	−2.03	−5.91	−1.38	−2.05	

2nd

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		−2.08	−1.88	−1.68	−4.83	−1.78	−1.44	−1.21
quip	0.00		0.17	0.00	0.31	0.00	0.00	0.00
updrs	−2.76	−2.83		−2.72	−4.24	−2.78	−3.02	−2.89
gds	0.02	−0.97	−0.97		−0.41	−0.97	−0.97	−0.97
scopa	−4.71	−1.01	−0.93	−1.26		−1.28	−0.91	−1.16
ess	−4.34	−3.89	−3.89	−4.08	−3.58		−4.26	−3.62
bjlot	−0.84	−2.37	−2.40	−2.17	−1.86	−2.40		−2.40
rem	0.23	−0.47	−0.47	−0.47	−1.87	−0.15	−0.47	

3rd

	moca	quip	updrs	gds	scopa	ess	bjlot	rem
moca		−1.59	−1.98	−0.99	−1.59	−1.76	−1.34	−0.77
quip	−4.21		−4.21	−4.21	−4.21	−4.21	−4.21	−4.21
updrs	−3.91	−4.49		−4.29	−4.28	−5.19	−4.26	−5.16
gds	−2.82	−3.49	−3.64		−3.57	−3.53	−3.01	−4.40
scopa	−1.26	0.00	0.00	0.00		0.00	0.00	−0.88
ess	−2.28	−1.61	−1.43	−2.23	−2.03		−1.43	−1.84
bjlot	0.27	1.33	0.37	1.35	0.28	0.57		0.50
rem	−2.59	−1.36	−1.83	−1.10	−1.32	−1.45	−1.48	

Figure B: Percentage change in cross-validated mean squared error from univariate to multivariate regression, with decreases in blue and increases in red. We consider six different settings, namely lasso or ridge regularisation for clinical data, omics data, or both (outer row), and analyse each visit separately (outer column). Each multivariate model supports the prediction for one tool at one visit (inner row) with another tool at the same visit (inner column).