Proceedings of Machine Learning Research 193:88-102, 2022 Machine Learning for Health (ML4H) 2022

Supplementary material for mmVAE: multimorbidity clustering using Relaxed Bernoulli β-Variational Autoencoders

Charles Gadd* CWLGADD@GMAIL.COM
University of Oxford

Krishnarajah Nirantharakumar K.NIRANTHARAN@BHAM.AC.UK
University of Birmingham

Christopher Yau CHRISTOPHER.YAU@WRH.OX.AC.UK
University of Oxford, Alan Turing Institute, Health Data Research UK

* Corresponding author

Primary condition		Prevalence (of $N = 290812$)	Total co-morbidities 1	2	3+
Mental health	Depression	133060 (45.78%)	41929	37533	53628
	Anxiety	97239 (33.45%)	27634	27040	42581
	OthMental	47124 (16.2%)	11746	12153	23225
	eatingdisorderuom	11645 (4.0%)	2795	2816	6034
	substance misuse	11638 (4.0%)	2085	2559	6994
	SMHmm	11108 (3.8%)	1294	2002	7812
	alcoholproblem	5683 (2.0%)	9283626	1129	3626
Cardiovascular	hypertension	6876 (2.4%)	2113	1763	3000
	CongHeart	2770 (1.0%)	994	715	1061
	Valve	1115 (0.4%)	346	291	478
	stroketiaincidprevkoo	1010 (0.3%)	266	263	481
	IHD MI	664 (0.2%)	156	144	364
	AF	219 (0.1%)	40	58	121
	Cardiomyopathy	207 (0.1%)	52	53	102
Dermatology	OthSkin	28085 (9.7%)	9231	7362	11492
	psoriasis mm	18620 (6.4%)	6428	4985	7207
	Atopic eczema	16581 (5.1%)	4430	4264	7887
Gynaecology	Female infertility	24121 (8.3%)	8893	6465	8767
	pcoskoo	21012 (7.2%)	6946	5639	8427
	endometriosis	10507 (3.6%)	3151	2729	4627
	leiomyoma	3338 (1.1%)	1205	846	1287
Haematology	VTEall	3851 (1.3%)	1112	902	1837
	perniciousanaemia	858 (0.3%)	231	237	390
	sickle cell	119 (0.3%)	46	37	36
	haemophilia	63 (0%)	22	16	25
Rheumatology	InflamArth	6609 (2.3%)	2145	1670	2794
	Ehler	1691 (0.6%)	440	416	835
	SpondArth	984 (0.3%)	232	240	512
Orthopaedic	chronicbackpain	3107 (1.1%)	735	756	1616
	scoliosis	3238 (1.1%)	1148	863	1227
	osteoporosis	571 (0.2%)	141	100	330
	Vertebrae	5885 (2.0%)	1672	1432	2781
Neurology	Migraine	64686 (22.2%)	21332	17406	25948
	OthHeadache	23626 (8.1%)	6333	6022	11271
	epilepsy mm	7238 (2.5%)	2346	1872	3020
	MS	900 (0.3%)	300	237	363
	spina bifida	644 (0.2%)	197	170	277
	iih	577 (0.2%)	144	140	293
	periph neuro	2776 (1.0%)	728	672	1376

	Primary condition	Prevalence (of $N = 290812$)	Total co-morbidities 1	2	3+
Respiratory	Asthma	85315 (29.3%)	30375	22920	32100
	PulmHtn	45 (0%)	11	9	25
	interstitiallungdiseasemm	50 (0%)	11	13	26
	bronchiectasisdraftv1	382 (0.1%)	75	93	214
	cf	239 (0.1%)	79	78	82
	osafinal	854 (0.3%)	250	177	427
	sarcoid	292 (0.1%)	92	78	122
	copd	323 (0.1%)	44	53	226
Other	Allergic Rhinoconjunctivitis	92340 (31.8%)	34050	24925	33429
	ibs mm	49439 (17.0%)	14435	13054	21950
	Thyroid	15824 (5.4%)	5327	4101	6396
	AutoSkin	4108 (1.4%)	1357	1121	1630
	InflamBowel	3348 (1.2%)	1157	918	1273
	InflamEye	3270 (1.1%)	1104	852	1314
	cholelithiasis	8682 (3.0%)	2579	2125	3978
	prithrombocytopenia imrd	578 (0.2%)	212	161	205
	Cancer	2890 (1.0%)	957	791	1142
	cataract	861 (0.3%)	269	214	378
	deaf	889 (0.3%)	278	243	368
	CKDall	604 (0.2%)	95	155	354
	ChrLiverAll	2446 (0.8%)	696	531	1219
	ulcer peptic	1562 (0.5%)	430	372	760
	slesystemic2019	612 (0.2%)	154	143	315
	oa	2204 (0.8%)	568	509	1127
	AdrenalAll	62 (0%)	19	9	34
	Pituitary	2010 (0.7%)	678	481	851
	retinal detach	473 (0.2%)	157	119	197
	pth	134 (0%)	37	30	67
	hfincidenceprevkoo	383 (0.1%)	81	91	211
	blindmm	118 (0%)	27	22	69
	solidorgantransplant	181 (0.1%)	26	38	117
	NeuroDev	2542 (0.9%)	724	632	1186
	turnerssyndrome imrd	57 (0%)	24	7	26
	marfansyndrome imrd	125 (0%)	32	35	58
	HIVall	267 (0.1%)	133	63	71
	DiabAll	5071 (1.7%)	1350	1243	2478
	DiabRetino	1720 (0.6%)	328	412	980
	coeliac	1529 (0.5%)	477	406	646
	urolithiasis	1973 (0.7%)	598	480	895
	Somatoform	5060 (1.7%)	269	214	378

Table 1: The total number of patients with each of the 79 health conditions in the full dataset (after removing non-multimorbidity previous pregnancy cases), and the total number of patients with different numbers of co-morbidities.

Conditions	Condition combination	Prevalence
2+	Depression, Anxiety	13329 (4.59%)
	Asthma, AllergicRhinoConj	9967 (3.43%)
	Depression, other mental health	4162 (1.43%)
	Asthma, Depression	3799 (1.3%)
	AllergicRhinoConj, Depression	3703 (1.3%)
	AllergicRhinoConj, Migraine	3432 (1.1%)
	Migraine, Depression	3163 (1.1%)
	Asthma, Migraine	2717 (0.9%)
	Anxiety, Depression, other mental health	2715 (0.9%)
	AllergicRhinConj, Anxiety	2253 (0.8%)
	AllergicRhinConj, Anxiety, Depression	2195 (0.8%)
3+	Anxiety, Depression, other mental health	2715 (0.9%)
	AllergicRhinConj, Anxiety, Depression	2195 (0.8%)
	Depression, Anxiety, Asthma	2113 (0.73%)
	Asthma, Anxiety, Depression	2113 (0.7%)
	Migraine, Anxiety, Depression	1863 (0.6%)
	Anxiety', Depression, ibs	1609 (0.6%)
4+	Asthma, AllergicRhinConj, Anxiety, Depression	931 (0.32%)
	Asthma, Anxiety, Depression, other mental health	510 (0.18%)
	AllergicRhinConj, Migraine, Anxiety, Depression	451 (0.16%)
	AllergicRhinConj, Anxiety, Depression, ibs	422 (0.2%)
	Asthma, Migraine, Anxiety, Depression	411 (0.1%)

Table 2: The top unique combinations of health conditions, and their respective prevalence for different levels of multimorbidity.

1. Experiments

In this section we show additional results which were excluded from the main text due to space constraints. We present the results for multiple choices of β on the reduced (i) asthma (ii) female infertility and (iii) cancer sub-group. In Fig. 1 we plot the factor odds ratios. In Fig. 2-3 we plot the prevalence plots for each of these factor. In Fig. 4-5 we plot the cluster odds ratios. In Fig. 6-7 we plot the prevalence plots for each of these clusters. In Fig. 8 we plot the cluster-factor association matrices. We observe a sharp decrease in factor-cluster sparsity as β increases.

In Fig. 9 we plot the *out-of-distribution* prevalence histograms. We observe that as β increases, and particularly as it shifts into a positive regime, the number of clusters out-of-distribution samples get mapped to overfits.

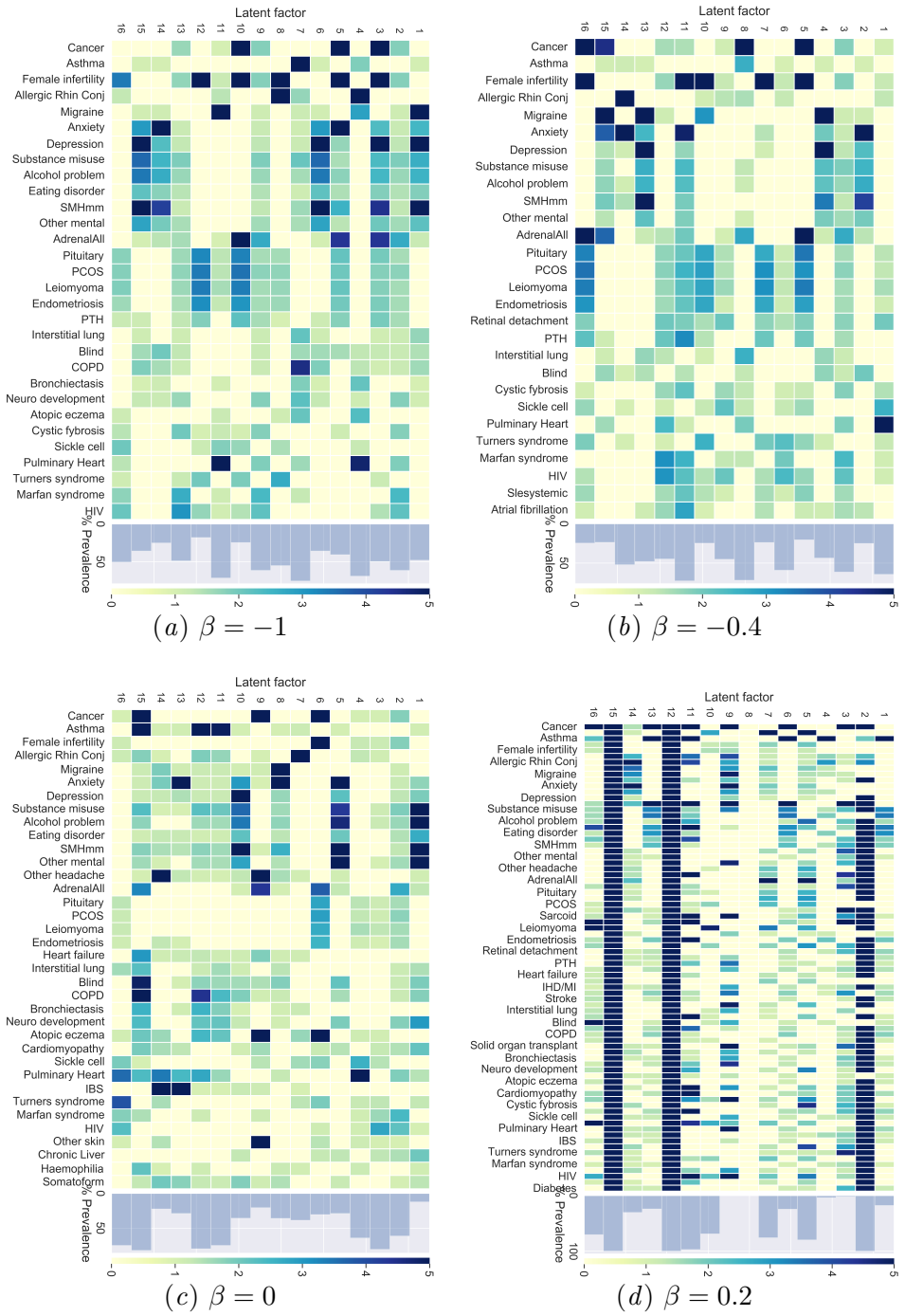


Figure 1: Factor odds ratio on a cohort with (i) asthma (ii) cancer and (iii) female infertility, and the percentage of individuals with each factor. Conditions with no odds ratio above 2 are filtered and odds ratios are truncated.

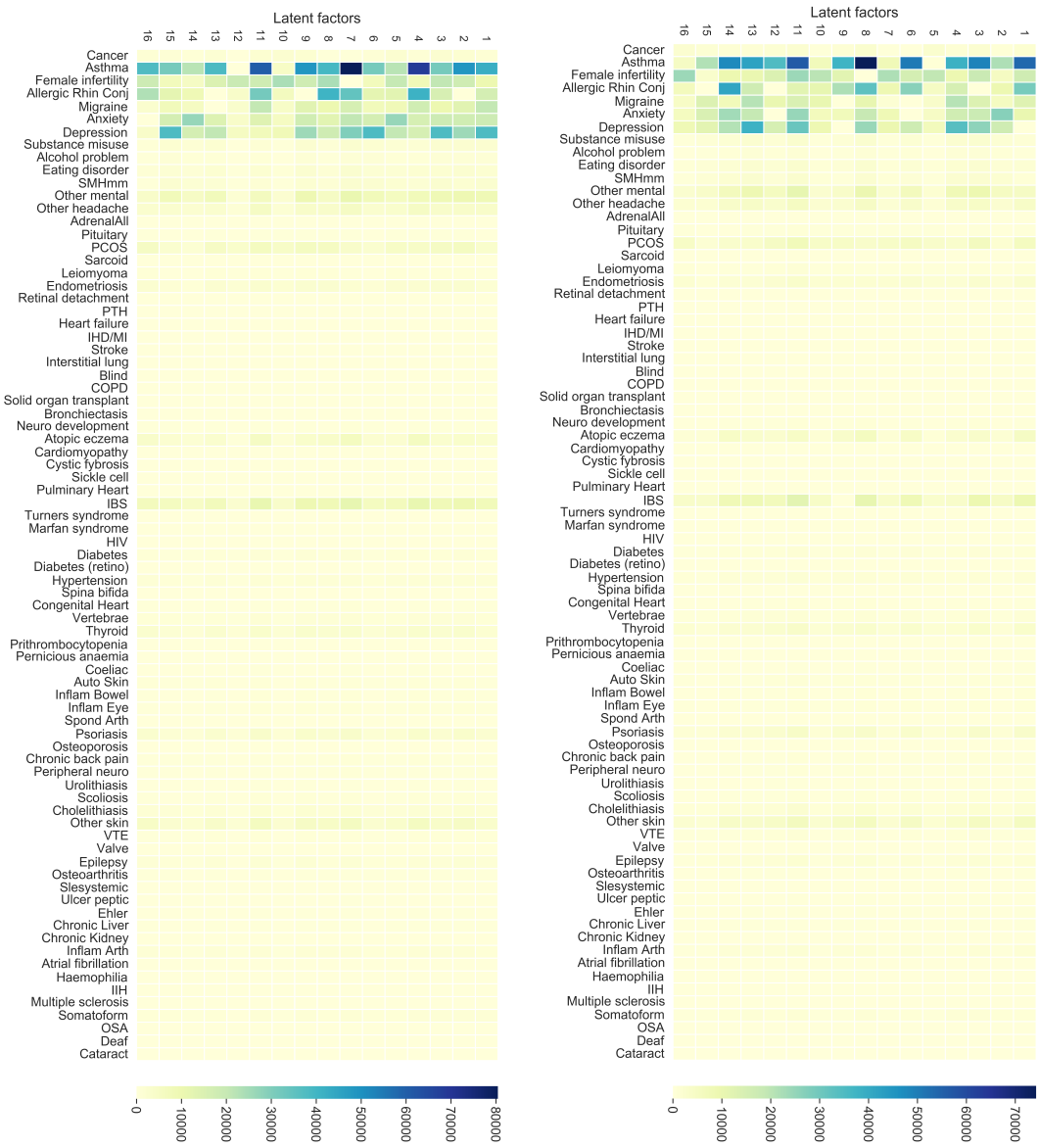
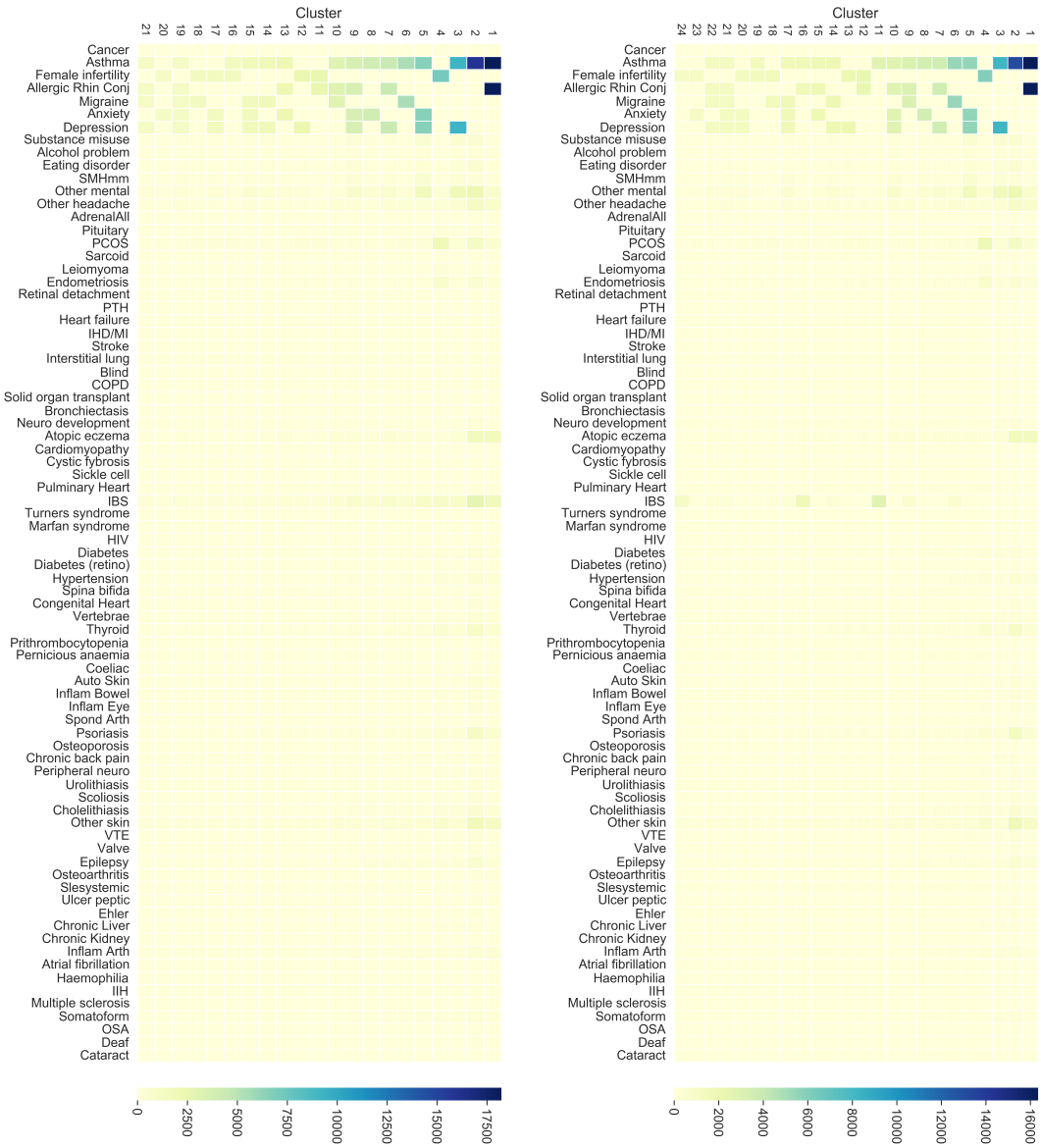


Figure 2: Factor prevalence on a cohort with (i) asthma, (ii) cancer, or (iii) female infertility health conditions.

Figure 3: Factor prevalence on a cohort with (i) asthma, (ii) cancer, or (iii) female infertility health conditions.

Figure 4: Clustering odds ratio on a cohort with (i) asthma, (ii) cancer, or (iii) female infertility health conditions, and the percentage of individuals in each cluster. Clusters with fewer than $< 1\%$ of the samples are not shown and odds ratios are truncated.

Figure 5: Clustering odds ratio on a cohort with (i) asthma, (ii) cancer, or (iii) female infertility health conditions, and the percentage of individuals in each cluster. Clusters with fewer than $< 1\%$ of the samples are not shown and odds ratios are truncated.

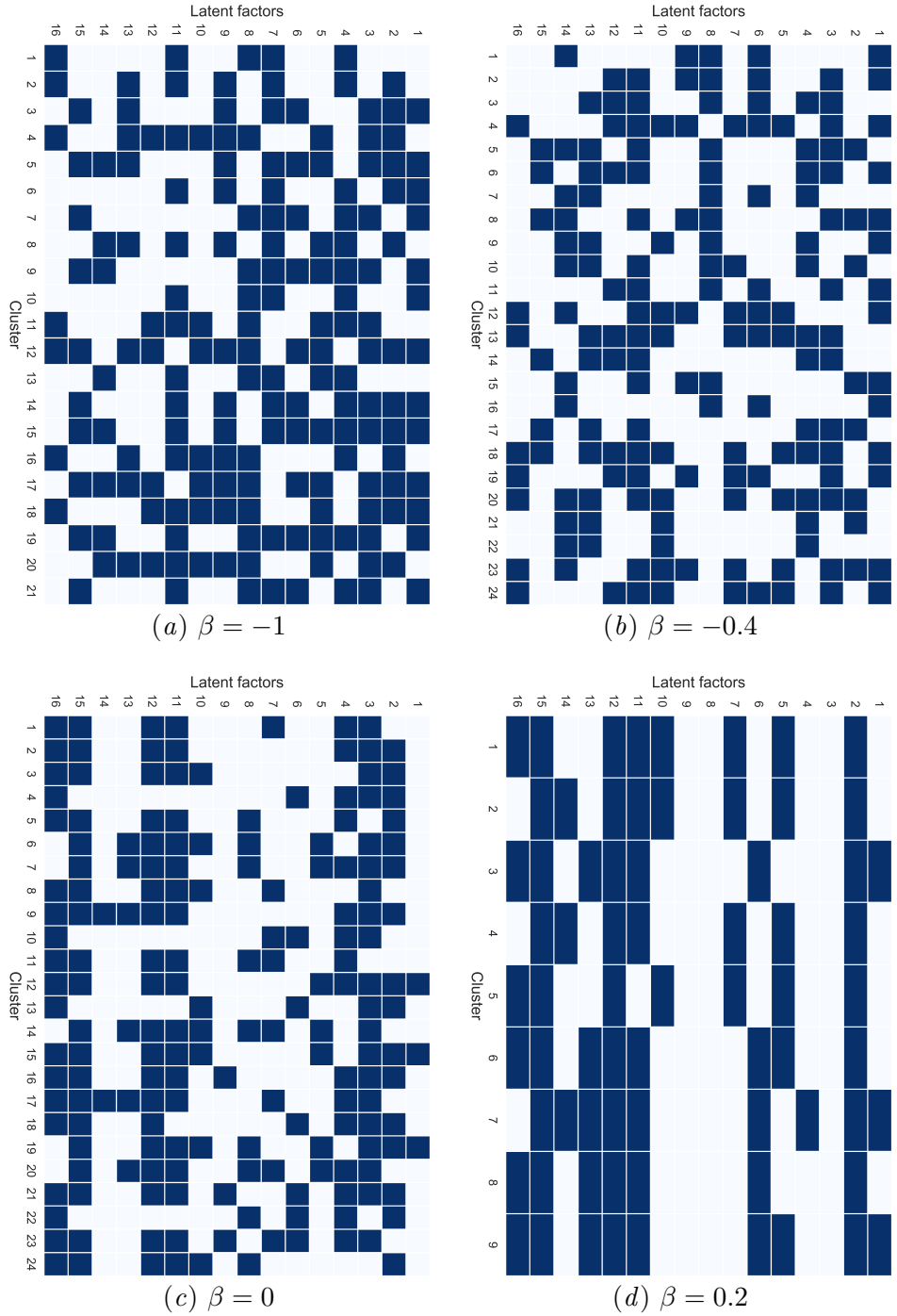


(a) $\beta = -1$ (b) $\beta = -0.4$

Figure 6: Cluster prevalence on a cohort with (i) asthma, (ii) cancer, or (iii) female infertility health conditions.

(a) $\beta = 0$ (b) $\beta = 0.2$

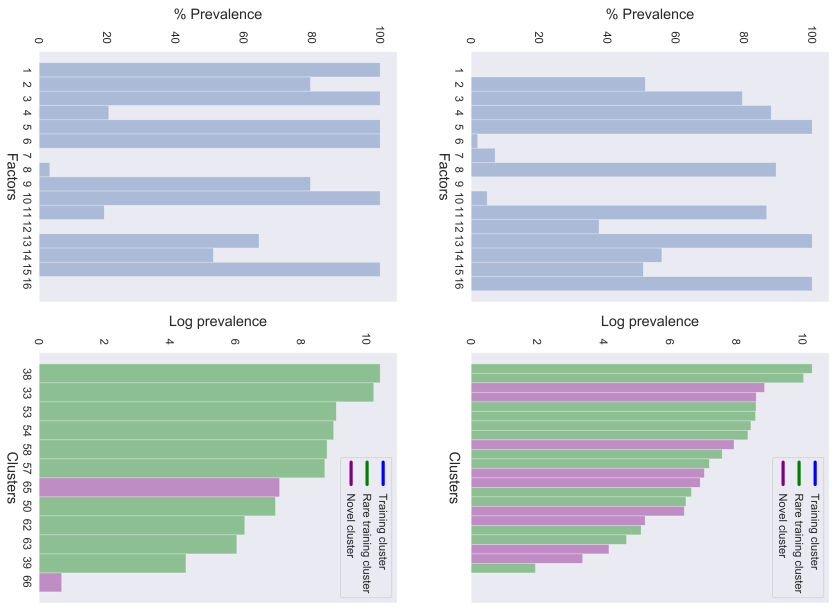
Figure 7: Cluster prevalence on a cohort with (i) asthma, (ii) cancer, or (iii) female infertility health conditions.



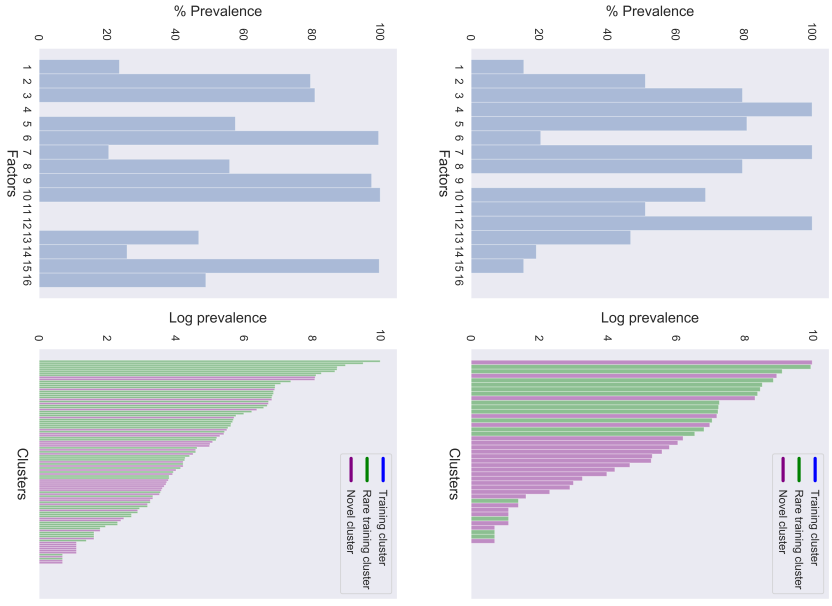
(a) $\beta = -1$ (b) $\beta = -0.4$

(c) $\beta = 0$ (d) $\beta = 0.2$

Figure 8: Cluster factor association matrices, indicating which factors are 'turned on' for each cluster.



(a) $\beta = -1$ (b) $\beta = -0.4$



(c) $\beta = 0$ (d) $\beta = 0.2$

Figure 9: *Out-of-distribution* histogram plots. A test depression cohort is encoded using a `mmVAE` model trained on (i) asthma, (ii) cancer, or (iii) female infertility cases. No test combinations have been seen in training. Above (blue): The prevalence of test cases in each factor. Below (green/pink): The prevalence of test cases in clusters, labelled given their relation to the training clusters.