Prediction of Autism Risk From Family Medical History Data Using Machine Learning: A National Cohort Study From Denmark

Supplemental Information

Contents

Table S1. Disorders and corresponding ICD 8 and 10-codes

Mental			Cardiometabolic			Neurologic			Birth defect			Autoimunne			Other		
Diagnosis	ICD		Diagnosis	ICD		Diagnosis	ICD		Diagnosis	ICD		Diagnosis	ICD		Diagnosis	ICD	
	8	10		8	10		8	10		8	10		8 & 9	10		8 & 9	10
ASD	299.00-299.03	F84.0, F84.1, F84.5 F84.8, F84.9	Any diabetes but type 1 diabetes	250, 761.1	E11-E14, G59.0- G63.2, H28.0, H36.0, M14.2, N08.3, O24.1-O24.4, O24.9	Systemic atrophies	331.09, 332 348.09, 348.20, 348.29, 348.99	G10-G14	CNS	740-743	G00-Q07	Thyrotoxicosis	242.00	E05.0	Asthma	493.00, 793.01, 493.09	J45, J46
Psychoactive substance use	291, 303, 304	F10-F19	Type 1 diabetes	249	E10, O24.0	Extrapyramid	342	G20-G26	Eye	744	Q10-Q15	Thyroiditis	245.03	E06.3	Allergies	493.02, 507, 691.00, 999.49, 708.09,	J45.0, J30, L20, T78.0-T78.4, H10.1
Schizophrenia	295, 297, 298	F20-F29	Obesity	277, 277.99, 278	E66, O99.21, Z68.2-Z68.4	Other degenerative		G30-G32	Ear	745	Q16-Q18	Pri adrenocortical	255.1	E27.1			
Bipolar disorder	296.19, 296.39, 298.19	F30-F31	Hypertension	400-404, 760.2, 637.0, 637.1, 637.9, 762.1, 762.2	I10-I15, O10, O11, O13-O16	Inflammatory of CNS	320-324	G00-G09	Heart	746, 747	Q20-Q28	Rheumatoid arthritis	712.19, 712.29, 712.59	M05, M06			
Depression	296.09, 296.29, 298.09, 300.49	F32-F33				Demyelinating of CNS	340, 341.01	G35-G37	Respiratory	748	Q30-Q34	Juvenile arthritis	712.09	M08			
Neurotic/stress disorder not OCD	300 (excl 300.9), 305	F40, F41, F43-F48				Episodic but not Epilepsy, migraine or sleep disorder	347.00, 347.01, 347.09	G42, G44-G46	Lip	749	Q35-Q37	Dermatopolymyositis	716	M33			
OCD	300.3	F42.0, F42.1, F42.2				Migraine	346	G43	Digestive	750, 751	Q38-Q45	Polymyalgia		M31.5, M31.6, M35.3			
Behavioral syndrome - physiological (not anorexia)	306.49, 306.58, 306.59	F50-F59 (exl F50.0)				Sleep disorders		G47	Genital	752	Q50-Q56	Scleroderma	734.0	M34 (excl: M34.2)			
Anorexia nervosa	306.50	F50.0				Nerve disorder not polyneuropath	350-358 (excl:354)	G50-G59	Urinary tract	753	Q60-Q64	Lupus erythema	734.19	M32.1, M32.8, M32.9			
Adult personality disorder	301,302	F60-F69				Polyneuropath	354	G60-G64	Musculoskeletal	754-756	Q65-Q79	Sjogren	734.90	M350			
Intellectual disability	310-315	F70-F79				Myoneural	330, 733.09	G70-G73	Skin	757	Q80-Q84	Ankylos spondil.	712.49	M45.9			
Psychological dev. disorder - not ASD	306.10, 306.11, 306.12, 306.18, 306.19	F80-89 (Exl: F84.0,F48.1, F84.5, F84.8, F84.9)				Cerebral palsy	343	G80-G83	Other	758, 759	Q85-Q99 (Excl:Q85.0, Q85.1, Q87.1, Q87.8, Q90,Q93.5, Q93.8, Q98.0, Q98.1,Q98.2,Q98.3, Q98.4, Q99.2)	Celiac not irritabl bowel synd	269.00	K90.0			
Emotional not ADHD or Tic	30609, 30679, 30689, 308	DF91-DF98 (Exl: DF951, 952, 988)				Other neurologic	347.93, 347.94, 347.95, 349.00, 349.01, 349.09	G90-G99	ASD specific	759.83, 759.30, 759.51	D82.1, Q71.0, Q85.0, Q85.1, Q87.1, Q87.8, Q90, Q93.5, Q93.8, Q98.0, Q98.1, Q98.2, Q98.3, Q98.4, Q99.2	Irritable bowel syndrome	564.19	K58			
ADHD		F90, F98.8				Epilepsy	345	G40, G41				Crohn	563.01	K50			
Tic disorder	30629	F95.1, F95.2										Ulcerative colitis	563.19	K51			
Mental-unspecified		F99										Pernicious anem	281.0	D51.0			
												Hemolytic anem	283.90, 283.91	D59.1			
												Purpura	287.31, 287.39	D69.3			
												Multple sclerosis	340	G35			
												Guillain-Bar	354	G61.0			
												Myasthen grav.	733.09	G70.0			
												Psoriasis	696.09, 696.10, 696.19	L40 (excl: L40.4)			
												Alopecia areata	704.00	L63			
												Vitiligo	709.01	L80.9			

The disorders should be mutually exclusice meaning that no ICD-codes are used more than one time to create a disorder

Table S2. Detailed analytic description

Stage	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6
Phase	Development	Development	Development	Development	Development	Deployment
Purpose	*Data partitioning*	*Tuning*	*Validation and Retraining*	*Replicate stage 2-3 for different sets of candidate morbidity indicators*	*Out-of-sample testing*	*Application of the FMRS*
Data source	Entire study cohort	Training sub-sample	Training sub-sample	Training sub-sample	Test sub-sample	Entire study cohort
Description of stage	Setting a seed (ensuring that one can replicate the random stochastic process) we divided the study cohort into a test and training sub-sample. 80% of the study cohort were randomly partitioned into the training sub-sample and 20% into the test sub-sample	We tested 10 different algorithms. Each of these where tuned individually selecting the optimal parameters based on which combinations of different tuning paramaters provided the highest F-measure score. For each algorithm, we first used a random search to locate the optimal parameter space. We then used a grid search around that specific parameter space to more precisely identify the optimal parameter value. Below are listed each model with an indication of the paramaters that where tuned for each of them. All random grid searches were evaluated using the same seed (the same stochastic process).	Using the initial tuned parameters, we calculated performance measures for the algorithms and started to compare models. As with the tuning stage, we used the same seed. As described in the main paper, an important issue in this study was the extremely imbalanced class distribution (a ratio of 1 case to 63 non-cases). The initial tuning and performance reflected this and we experimented with both upscaling the cases, downscaling the non-cases and mixing the two approaches. Downscaling the non-cases were vastly superior to the other two approaches in terms of with computational speed and model performance, thus - with a downscaling ratio of 1 to 1 (randomly selected from the non-cases), we again tuned each algorithm parameter listed in stage 2, with first a random search and then a grid search using the same seed for the downscaling across the algorithms to ensure comparability. Three super-learners incorporating the different algorithms in different ways were then estimated using the parameter values identified using the same downscaling process. The average performance measure across the 10-fold validation were than computed, also extracting the standard deviation across the 10 different runs.	Stage 2 and 3 were replicated for differents sets of candidate morbidity indicators. First, a model based on 353 candidate indicators were trained. As described in the paper we then calculated the variable importance using two differet algorithms: Random Forest and Extreme Gradient Boost. Each algorithm calculated the importance for each of the 353 variables mesured by the total amount the Gini index is decreased by splits over each morbidity indicator, averaged over all trees. The resulted in two importance ranking (which were widely similar as shown in Figure SX). As it is not clear which algorithm were superior in terms of estimating the variable importance we choose to include a variable in the best performing variable set if they were either included in the top based on estimations on either the Random Forest or the Gradient Boost algorithm. We cycled through 3 different restrictions of included variabled. First we restricted the included variables to those that were either among the top 30 variables in either the RF or the EGB (41 variables were in either one or the other), then among the 15 best performing (21 varaibles were included) or among the top 3.	Having chosen the best performing algorithm in stage 4, we now applied this algorithm to the test subsample and calculated the F-score, area under the curve (AUC), sensitivity, predictive positive value, deviance, kappa and specificity.	For each study participant a family morbidity risk score (FMRS) was calculated using the estimated predicted probability from the best performing algorithm. This score was then applied to the whole study population. We calculated the distribution of the FMRS and compared the similarity of the distrubtion of ASD cases compared to non-cases. We then used logistic regression with ASD as the outcome estimated unadjusted and adjusted associations with the FMRS. We compared the adjusted estimates to other representations of family morbidity history: parental psychiatric disorders yes/no and sibling ASD history yes/no. Lastly, we assessed potential interaction between the FMRS and gender, birth weight and paprental socioeconomic position.
				We also replicated stage 2 and 3 for a number of sensitivity analyses. First running the main analyses separate for men and women and then included the candidate indicators as counting the number of times, the diagnosis in question had occurred within each family member type.		

Table S3. Included candidate morbidity indicators by family member type after initial removal of 83 indicators due to lower than 40 exposed ASD cases.

Diagnosis	Grandparents	Aunts/uncles	Mother	Father	Cousin	Sibling	
ASD		x	x	x	x	x	Mental
Psychoactive substance use	x	x	x	x	x	x	
Schizophrenia	x	x	x	x	x	x	
Bipolar disorder	x	x	x	x	x	x	
Depression	x	x	x	x	x	x	
Neurotic/stress disorder not OCD	x	x	x	x	x	x	
OCD	x	x	x	x	x	x	
Behavioral syndrome - physiological (not anorexia)	x	x	x	x	x	x	
Anorexia nervosa		x	x		x	x	
Adult personality disorder	x	x	x	x	x	x	
Intellectual disability	x	x	x	x	x	x	
Psychological dev. disorder - not ASD		x	x		x	x	
Emotional not ADHD or Tic	x	x	x	x	x	x	
ADHD		x	x	x	x	x	
Tic disorder					x	x	
Mental-unspecified	x	x	x	x	x	x	
Any diabetes but type 1 diabetes	x	x	x	x	x	x	Cardiometabolic
Type 1 diabetes	x	x	x	x	x	x	
Obesity	x	x	x	x	x	x	
Hypertension	x	x	x	x	x	x	
Systemic atrophies	x						Neurologic
Extrapyramid	x	x	x	x	x	x	
Other degenerative	x	x					
Inflammatory of CNS	x	x	x	x	x	x	
Demyelinating of CNS	x	x	x	x			
Episodic but not Epilepsy, migraine or sleep disorder	x	x	x	x	x	x	
Migraine	x	x	x	x	x		
Sleep disorders	x	x	x	x	x	x	
Nerve disorder not polyneuropath	x	x	x	x	x		
Polyneuropath	x	x	x	x	x		
Myoneural	x	x	x	x	x	x	
Cerebral palsy	x	x	x	x	x		
Other neurologic							
Epilepsy	x	x	x	x	x	x	

CNS	x	x	x		x		Birth defect
Eye	x	x	x	x	x	x	
Ear	x	x	x	x	x	x	
Heart	x	x	x	x	x	x	
Respiratory	x	x	x	x	x	x	
Lip		x	x		x	x	
Digestive	x	x	x	x	x	x	
Genital	x	x	x	x	x	x	
Urinary tract	x	x	x	x	x	x	
Musculoskeletal	x	x	x	x	x		
Skin	x	x	x	x	x	x	
Other	x	x	x		x	x	
ASD specific	x	x	x		x	x	
Thyrotoxicosis	x	x	x	x	x		Autoimmune
Thyroiditis	x	x	x		x		
Pri adrenocortical	x						
Rheumatoid arthritis	x	x	x	x	x	x	
Juvenile arthritis		x			x		
Dermatopolymyositis	x						
Polymyalgia	x	x					
Scleroderma	x	x					
Lupus erythema	x	x	x		x		
Sjogren	x	x	x				
Ankylos spondil.	x	x	x	x	x		
Celiac not irritabl bowel synd	x	x	x		x	x	
Irritable bowel syndrome	x	x	x	x	x	x	
Crohn	x	x	x	x	x	x	
Ulcerative colitis	x	x	x	x	x	x	
Pernicious anem	x						
Hemolytic anem	x						
Purpura	x	x					
Multple sclerosis	x	x	x	x	x		
Guillain-Bar	x	x					
Myasthen grav.	x						
Psoriasis	x	x	x	x	x	x	
Alopecia areata	x	x			x		
Vitiligo							
Asthma	x	x	x	x	x	x	Other
Allergies	x	x	x	x	x	x	

Table S4.10-fold cross-validated performance for each candidate model with percent (%) difference to base model

F-score	score	difference	Kappa	kappa	difference	AUC	SD AUC	difference	TPR	SD TPR	difference	MMCE	GPR	PPV	SD PPV	difference	NPV	Log loss	model	Algorithm
0,047			*0,020*			*0,557*			*0,289*			*0,184*	*0,088*	*0,026*			*0,986*	*0,692*	*Parental psychiatric history*	*Logistic*
0,05051728	0,001095	7%	0,021320108	0,001115	7%	0,642811305	0,00541	15%	0,492832538	0,015265	71%	0,291231152	0,114543806	0,026623369	0,000567	1%	0,988754145	0,659398135	All morbidity indicators	Logistic
0,05469664	**0,001684**	**16%**	**0,025945396**	**0,001745**	**30%**	**0,644034228**	**0,007048**	**16%**	**0,446273076**	**0,017603**	**54%**	**0,24255551**	**0,114015431**	**0,029135178**	**0,000903**	**10%**	**0,988532539**	**0,659375839**	**All morbidity indicators**	**Gradient Boosting**
0,048631425	0,000691	3%	0,019159366	0,000711	-4%	0,646669625	0,004398	16%	0,547170995	0,01107	89%	0,336582987	0,117996978	0,025446655	0,000358	-4%	0,989244785	0,596871566	All morbidity indicators	Random Forest
0,049270365	0,020523	5%	0,020716524	0,023761	4%	0,641146849	0,009829	15%	0,63951026	0,333486	121%	0,518005905	0,116127972	0,027883488	0,015116	6%	0,990070807	0,76701885	All morbidity indicators	Neural Networks
0,052498	0,000922	12%	0,02348545	0,000936	17%	0,647212633	0,004685	16%	0,475875235	0,013174	65%	0,270045161	0,114977927	0,027781738	0,000477	5%	0,988723373	0,656638232	All morbidity indicators	Elastic Net
0,05117208	0,001076	9%	0,022012799	0,001124	10%	0,645398716	0,005191	16%	0,493675805	0,007763	71%	0,287900725	0,115416693	0,026984919	0,000583	2%	0,988823476	0,66047099	All morbidity indicators	Support Vector Machines
0,046541542	0	-1%	0,018683849	0	-7%	0,585245634	0	5%	0,239647549	0	-17%	0,154455333	0,078578157	0,025778832	0	-2%	0,985997224	0,525647069	All morbidity indicators	K nearest Neighbors
0,053344276	0,003028	13%	0,024395138	0,003342	22%	0,651268259	0,009402	17%	0,478128395	0,042804	65%	0,268545683	0,116004949	0,028278299	0,001818	7%	0,988806059	0,630607259	All morbidity indicators	Stacked - Average
0	0	-100%	0	0	-100%	0,651203042	0,00898	17%	0	0	-100%	0,015722678		0,3	0,483046	1036%	0,984277322	0,077695232	All morbidity indicators	Stacked - CV
0,0478279	0,00101	2%	0,009736978	0	-51%	0,6462462	0,009592	16%	0,5522304	0,016147	91%	0,3456861	0,1174863	0,0249966	0,000524	-5%	0,9892148	0,6052082	All morbidity indicators	Stacked - Hill climb
0,052505625	0,000976	12%	0,023539705	0,000995	18%	0,643508241	0,004167	16%	0,463649072	0,013375	60%	0,263071294	0,113586741	0,02782898	0,000508	5%	0,988575351	0,658237149	Top 41	Logistic
0,05486774	**0,001703**	**17%**	**0,026162789**	**0,001777**	**31%**	**0,641960603**	**0,006883**	**15%**	**0,438075038**	**0,011376**	**52%**	**0,23735734**	**0,113226352**	**0,029267254**	**0,000928**	**11%**	**0,988444509**	**0,660349819**	**Top 41**	**Gradient Boosting**
0,049984365	0,001197	6%	0,020807613	0,001238	4%	0,633869897	0,005734	14%	0,476766218	0,01205	65%	0,284965799	0,112133788	0,026375089	0,000638	0%	0,988506699	0,530848469	Top 41	Random Forest
0,054214612	0,022175	15%	0,026323027	0,025963	32%	0,6417159	0,007416	15%	0,554172328	0,279444	92%	0,401073789	0,115203207	0,031690787	0,018049	20%	0,989099084	0,672601584	Top 41	Neural Networks
0,053370112	0,001091	14%	0,024497774	0,001112	22%	0,64329927	0,004409	15%	0,454233608	0,013492	57%	0,253319376	0,113477725	0,028350981	0,00057	7%	0,988530824	0,658522334	Top 41	Elastic Net
0,052306526	0,001366	11%	0,02336008	0,001453	17%	0,638959656	0,00768	15%	0,456157056	0,022091	58%	0,260025919	0,112466272	0,027748575	0,000766	5%	0,98846918	0,664604352	Top 41	Support Vector Machines
0,048810853	0,000832	4%	0,019756498	0,000858	-1%	0,621748653	0,005094	12%	0,425191074	0,010216	47%	0,260553954	0,104919868	0,025892049	0,000444	-2%	0,987817167	0,547578952	Top 41	K nearest Neighbors
0	0	-100%	0	0	-100%	0,645194301	0,007457	16%	0	0	-100%	0,015722678		0,2	0,421637	657%	0,984277322	0,0778661	Top 41	Stacked - Average
0,052925668	0,002299	13%	0,023991516	0,002577	20%	0,642466045	0,007511	15%	0,466885356	0,042144	62%	0,263977189	0,114262287	0,028079096	0,001415	6%	0,988642166	0,596291691	Top 41	Stacked - CV
0,044944574	0,016148	-4%	0,015493444	0,01828	-23%	0,641601024	0,007154	15%	0,676103611	0,268498	134%	0,543273204	0,118076764	0,024194605	0,010881	-8%	0,98993798	0,767686433	Top 41	Stacked - Hill climb
0,053074872	0,000587	13%	0,024214263	0,000601	21%	0,634393864	0,003105	14%	0,446785775	0,011246	55%	0,25065327	0,112268067	0,028213992	0,00031	7%	0,988419256	0,661308299	Top 21	Logistic
0,053813748	**0,001335**	**14%**	**0,025033083**	**0,00141**	**25%**	**0,63484436**	**0,007221**	**14%**	**0,439761682**	**0,01541**	**52%**	**0,243227192**	**0,112251526**	**0,028663035**	**0,000745**	**9%**	**0,988390658**	**0,660558724**	**Top 21**	**Gradient Boosting**
0,051731303	0,001906	10%	0,022881074	0,002031	14%	0,621854071	0,006382	12%	0,423834165	0,008573	47%	0,244602976	0,108034829	0,027549634	0,00108	4%	0,98804465	0,615764604	Top 21	Random Forest
0,061511967	0,019689	31%	0,034586501	0,02304	73%	0,632622329	0,007526	14%	0,412115169	0,250656	43%	0,266715149	0,108808449	0,036294498	0,016225	37%	0,987751066	0,602950374	Top 21	Neural Networks
0,053191053	0,000623	13%	0,024343253	0,000642	22%	0,634455206	0,003296	14%	0,445520896	0,010015	54%	0,24937398	0,112250417	0,028284608	0,000333	7%	0,9884129	0,661256326	Top 21	Elastic Net
0,052564042	0,002425	12%	0,023704234	0,002714	19%	0,629166992	0,007412	13%	0,442806639	0,039976	53%	0,252284645	0,111050023	0,027968195	0,001487	6%	0,988322925	0,665603349	Top 21	Support Vector Machines
0,064066173	0,003296	36%	0,038105834	0,003341	91%	0,598300837	0,004988	7%	0,226202994	0,014387	-22%	0,103859669	0,091875098	0,037319733	0,001857	41%	0,986553339	1,093098566	Top 21	K nearest Neighbors
0	0	-100%	0	0	-100%	0,634624888	0,007126	14%	0	0	-100%	0,015722678		0,3	0,483046	1036%	0,984277322	0,078117355	Top 21	Stacked - Average
0,060875568	0,002624	30%	0,033161952	0,002824	66%	0,634349774	0,007985	14%	0,344716852	0,01538	19%	0,167504539	0,107243737	0,033397963	0,001588	26%	0,987696834	0,519371565	Top 21	Stacked - CV
0,050969545	0,018576	8%	0,022437971	0,021314	12%	0,634081357	0,007144	14%	0,566835159	0,308225	96%	0,440482306	0,113232612	0,028445058	0,013434	8%		0,729853929	Top 21	Stacked - Hill climb
0,068117286	0,002269	45%	0,042254174	0,0023	111%	0,577043	0,004792	4%	0,241239901	0,010022	-17%	0,103755822	0,097809856	0,039658749	0,001278	50%	0,986809108	0,671186999	Top 3	Logistic
0,068129941	**0,002287**	**45%**	**0,042266552**	**0,002348**	**111%**	**0,577099027**	**0,004189**	**4%**	**0,241240867**	**0,008429**	**-17%**	**0,103755827**	**0,0978215**	**0,039666342**	**0,001326**	**50%**	**0,986808974**	**0,673612972**	**Top 3**	**Gradient Boosting**
0,068117286	0,002269	45%	0,042254174	0,0023	111%	0,575919358	0,005128	3%	0,241239901	0,010022	-17%	0,103755822	0,097809856	0,039658749	0,001278	50%	0,986809108	0,53030143	Top 3	Random Forest
0,055256242	0,027823	18%	0,028865784	0,03372	44%	0,577069237	0,004198	4%	0,593676815	0,431155	105%	0,529629083	0,111467181	0,037544652	0,029351	42%		0,728745079	Top 3	Neural Networks
0,068117286	0,002269	45%	0,042254174	0,0023	111%	0,577043	0,004792	4%	0,241239901	0,010022	-17%	0,103755822	0,097809856	0,039658749	0,001278	50%	0,986809108	0,671176527	Top 3	Elastic Net
0,068129941	0,002287	45%	0,042266552	0,002348	111%	0,575033306	0,003753	3%	0,241240867	0,008429	-17%	0,103755827	0,0978215	0,039666342	0,001326	50%	0,986808974	0,675116967	Top 3	Support Vector Machines
0	0	-100%	0	0	-100%	0,5	0	-10%	0	0	-100%	0,015722678		1	0	3686%	0,984277322	0,543042044	Top 3	K nearest Neighbors
0	0	-100%	0	0	-100%	0,577143154	0,004184	4%	0	0	-100%	0,015722678		1	0	3686%	0,984277322	0,078780646	Top 3	Stacked - Average
0,068129941	0,002287	45%	0,042266552	0,002348	111%	0,577089423	0,004187	4%	0,241240867	0,008429	-17%	0,103755827	0,0978215	0,039666342	0,001326	50%	0,986808974	0,469387968	Top 3	Stacked - CV
0	0	-100%	0	0	-100%	0,500720456	0,00152	-10%	0	0	-100%	0,015722678		1	0	3686%	0,984277322	0,542303019	Top 3	Stacked - Hill climb

Table S5. Test sample performance fit with absolute and percentage (%) differences to parental psychiatric history

	Parental	**Top 41**	**Absolut**	**Percentage**
F-score	*0,048*	0,054	0,006	12%
Kappa	*0,020*	0,025	0,005	27%
AUC	*0,560*	0,643	0,084	15%
TPR	*0,294*	0,441	0,147	50%
MMCE	*0,184*	0,244	0,061	33%
GPR	*0,088*	0,113	0,025	28%
PPV	*0,026*	0,029	0,002	9%
NPV	*0,986*	0,988	0,002	0%
Log loss	*0,692*	0,663	-0,030	-4%

Table S6. Performance for the machine learning algorithm with the count indicators for each candidate model with percent (%) difference to base model

Data type	AUC (Diff to base model;ln %)	F-score (Diff to base model;ln %)	Kappa (Diff to base model;ln %)	PPV (Diff to base model;ln %)	TPR (Diff to base model;ln %)	Algorithm
Count Indicator	0.639 (0.08;14.68%)	0.05 (0.003;6%)	0.021 (0.002;10.3%)	0.026 (0.001;2.6%)	0.488 (0.2;69.2%)	Logistic
Count Indicator	0.645 (0.09;15.71%)	0.054 (0.006;13.4%)	0.025 (0.006;30.3%)	0.029 (0.003;10.5%)	0.459 (0.171;59.2%)	Elastic Net
Count Indicator	0.641 (0.08;15.04%)	0.051 (0.004;8.7%)	0.022 (0.003;17.5%)	0.027 (0.001;5.5%)	0.474 (0.185;64.2%)	Support Vector Machines
Count Indicator	0.643 (0.09;15.4%)	0.054 (0.007;14.2%)	0.025 (0.006;32.7%)	0.029 (0.003;11.6%)	0.443 (0.155;53.7%)	Gradient Boosting
Count Indicator	0.642 (0.08;15.25%)	0.044 (-0.003;-7%)	0.014 (-0.005;-25.6%)	0.023 (-0.003;-11.4%)	0.612 (0.324;112.2%)	Random Forest
Count Indicator	0.634 (0.08;13.68%)	0.075 (0.027;57.6%)	0.05 (0.031;160.5%)	0.045 (0.019;72.9%)	0.228 (-0.06;-20.9%)	Neural Networks
Count Indicator	0.555 (0;-0.33%)	0.037 (-0.01;-21.6%)	0.007 (-0.012;-61.5%)	0.019 (-0.006;-24.6%)	0.407 (0.118;41%)	K nearest Neighbors
Count Indicator	0.648 (0.09;16.33%)	0.047 (-0.001;-1.1%)	0.017 (-0.002;-9.8%)	0.024 (-0.001;-5.4%)	0.577 (0.289;100.1%)	Stacked - Average
Count Indicator	0.649 (0.09;16.39%)	0 (-0.047;-100%)	0 (-0.019;-100%)	1 (0.974;3774.1%)	0 (-0.289;-100%)	Stacked - CV
Count Indicator	0.637 (0.08;14.37%)	0.076 (0.029;60.8%)	0.052 (0.033;171.2%)	0.046 (0.021;79.9%)	0.212 (-0.076;-26.5%)	Stacked - Hill climb

Table S7. Levels for the Top 41 FMRS

Category	Values
1 (Lowest)	<=0.41
2	0.41>=0.464
3	0.464>=0.518
4	0.518>=0.572
5	0.572>=0.626
6	0.626>=0.68
7	0.68>=0.734
8	0.734>=0.788
9	0.788>=0.842
1 (Highest)	>=0.842

Table S8. Adjusted* Odds Ratios (OR) with with 95%CI for the FMRS based on all candidate indicators, the top 21 and the top 3, respectively

		353 candidate indicators			Top 21			Top 3			
		ASD cases	**OR**	**95%CI**	**ASD cases (%)**	**OR**	**95%CI**		**ASD cases**	**OR**	**95%CI**
FMRS								FMRS			
	1 (Lowest)	1428 (0.8%)	*ref*		11010 (1.1%)	*ref*		1 (Lowest)	20145 (1.3%)	*ref*	
	2	10066 (1.1%)	1,12	(1,06--1,18)	4459 (1.5%)	1,29	(1,25-1,34)	2	2560 (2.7%)	2,00	(1,92-2,08)
	3	4824 (1.6%)	1,58	(1,49-1,68)	3509 (2.0%)	1,70	(1,63-1,77)	3	1000 (3.2%)	2,21	(2,07-2,35)
	4	3324 (2.1%)	2,05	(1,92-219)	2201 (2.5%)	2,13	(2,04-2,24)	4	1574 (7.5%)	5,21	(4,94-5,5)
	5	2275 (2.9%)	2,82	(2,64-3,02)	1354 (3.2%)	2,75	(2,6-2,92)	5	1096 (11.3%)	7,99	(7,48-8,53)
	6	1193 (4.2%)	3,91	(3,61-4,23)	721 (4%)	3,28	(3,04-3,55)	6 (Highest)	143 (13.7%)	9,63	(8,04-11,55)
	7	583 (5.7%)	5,19	(4,7-5,74)	482 (5%)	4,02	(3,66-4,42)				
	8	811 (6.8%)	6,39	(5,84-6,99)	1071 (7.5%)	6,19	(5,8-6,61)				
	9	1317 (10.1%)	9,49	(8,77-10,27)	1180 (10.4%)	8,70	(8,15-9,28)				
	1 (Highest)	697 (15.7%)	15,52	(14,06-17,14)	531 (15.3%)	13,52	(12,27-14,89)				

Table S9. Adjusted* Odds Ratios (OR) with 95%CI for the FMRS restricting the study population to i) only children, ii) complete linkages, iii) incomplete linkages, iv) test subsample

		Only children			Complete linkages			Incomplete linkages			Test sub sample		
		(%)	OR	95%CI	(%)	OR	95%CI	(%)	OR	95%CI	(%)	OR	95%CI
FMRS													
	1 (Lowest)		**	-	49 (0.7%)	*ref*		39 (0.5%)	*ref*		16 (0.58%)	*ref*	
	2	2341 (1.7%)	*ref*		6983 (1.0%)	0,89	(0,67-1,18)	3563 (1.2%)	1,19	(0,86-1,64)	2126 (1.1%)	1,07	(0,65-1,75)
	3	1480 (2.4%)	1,25	(1,17-1,33)	3588 (1.4%)	1,20	(0,9-1,59)	1922 (1.9%)	1,55	(1,12-2,13)	1106 (1.5%)	1,44	(0,87-2,37)
	4	1026 (2.8%)	1,53	(1,43-1,66)	2267 (1.9%)	1,65	(1,24-2,19)	1249 (2.5%)	1,97	(1,43-2,73)	708 (2.1%)	1,96	(1,19-3,23)
	5	793 (3.5%)	1,97	(1,82-2,15)	1473 (2.5%)	2,17	(1,63-2,9)	971 (3.3%)	2,55	(1,84-3,53)	504 (2.8%)	2,64	(1,59-4,36)
	6	382 (4.3%)	2,28	(2,04-2,56)	726 (3.4%)	2,89	(2,19-3,87)	472 (4.2%)	3,02	(2,17-4,21)	227 (3.4%)	3,14	(1,88-5,25)
	7	121 (5.2%)	2,75	(2,27-3,33)	497 (5.1%)	4,17	(3,1-5,61)	160 (4.8%)	3,38	(2,36-4,83)	125 (4.6%)	3,98	(2,34-6,74)
	8	48 (7.6%)	3,89	(2,87-5,28)	746 (6.9%)	5,81	(4,33-7,78)	154 (7.6%)	6,11	(4,27-8,78)	176 (6.6%)	5,83	(3,47-9,79)
	9	9 (34.6%)	28,49	(12,13-66,9)	1320 (10.5%)	9,08	(6,81-12,1)	159 (10.4%)	8,46	(5,9-12,14)	308 (10.1%)	10,06	(6,04-16,75)
	10 (Highest)		**	-	466 (17.5%)	16,17	(11,97-21,8)	36 (12%)	9,16	(5,67-14,8)	96 (17.0%)	16,69	(9,68-28,79)
							*Adjusted for sex, birth weight, gestational age, birth year, maternal and paternal age						
							** Not enough cases to estimate						

Table S10. Adjusted Odds Ratios (OR) with with 95%CI for the FMRS additionally adjusting for the number of aunts and uncles, cousins, siblings and grandparents

		OR*	95%CI
FMRS			
	1 (Lowest)	*ref*	-
	2	1,29	(1,25-1,34)
	3	1,70	(1,63-1,77)
	4	2,13	(2,04-2,24)
	5	2,75	(2,6-2,92)
	6	3,28	(3,04-3,55)
	7	4,02	(3,66-4,42)
	8	6,19	(5,8-6,61)
	9	8,70	(8,15-9,28)
	10 (Highest)	13,52	(12,27-14,89)

*also adjusted for gender, birth weight, gestational age, birth year, maternal, paternal age and highest parental educational attainment

Figure S1. Illustration of linkages to family members

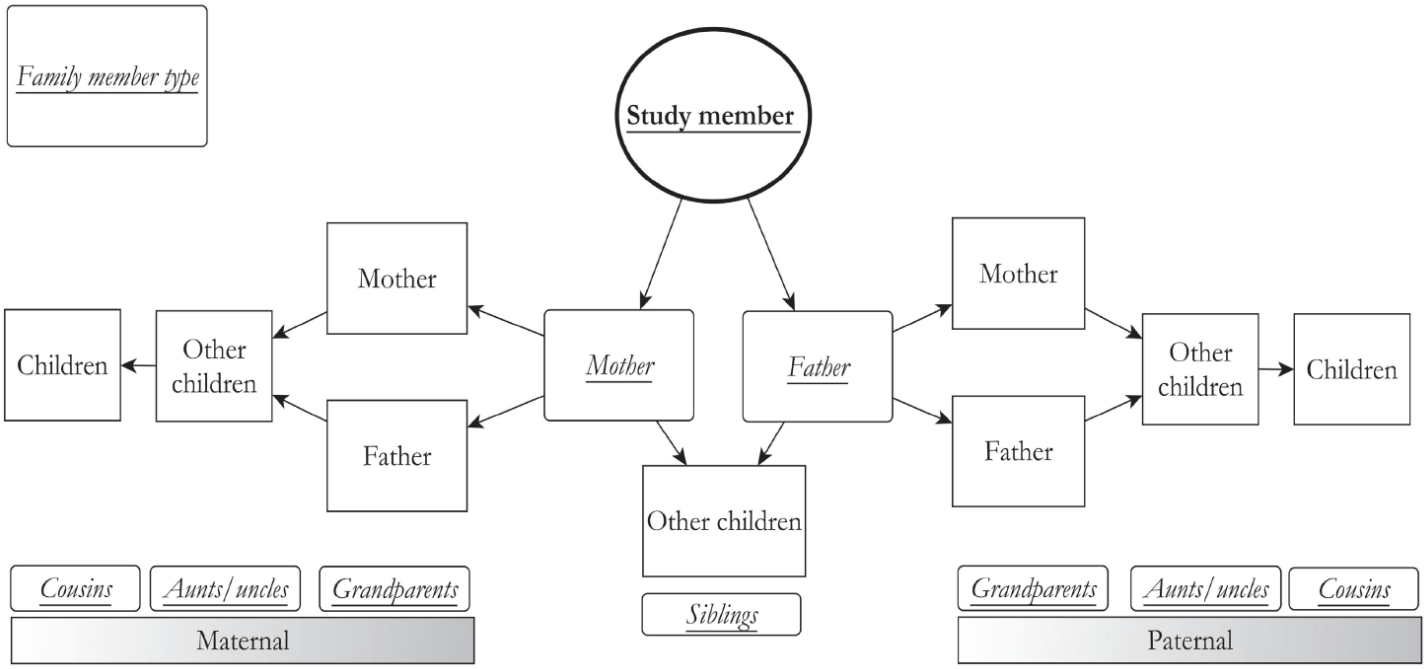


Figure S2. Prevalence and percent (%) for each morbidity indicator

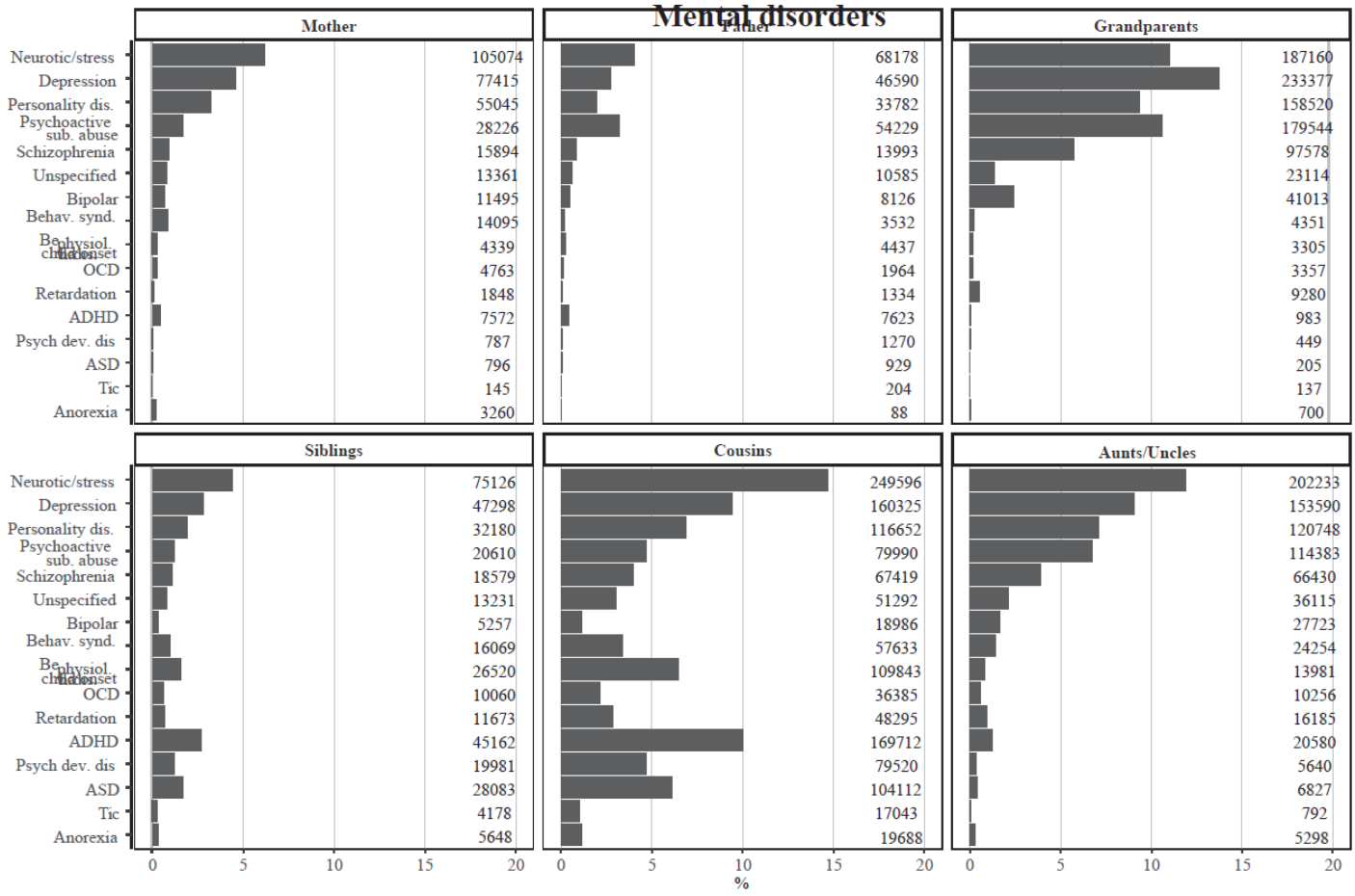


Figure S2. Prevalence and percent (%) for each morbidity indicator, continued

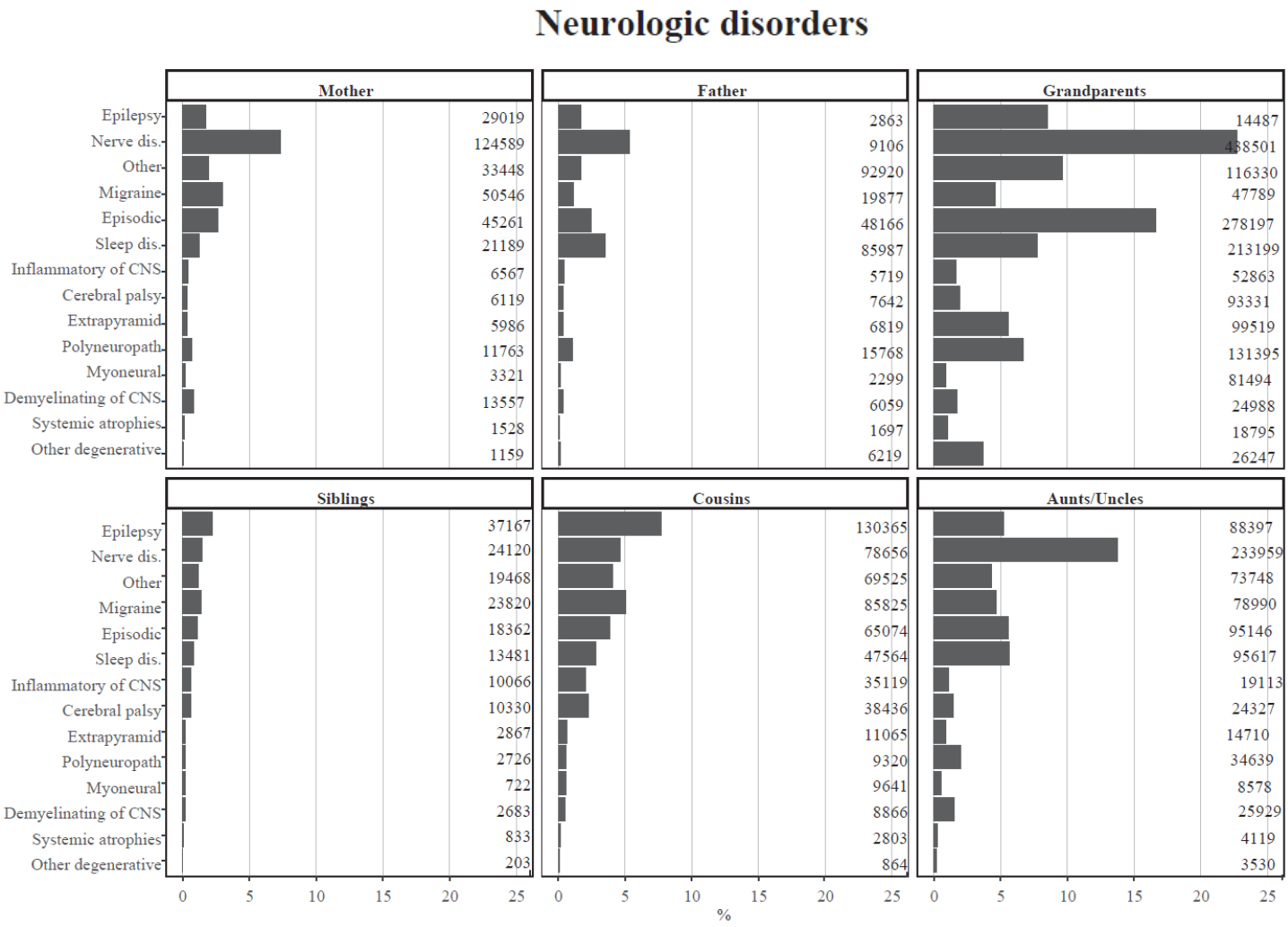


Figure S2. Prevalence and percent (%) for each morbidity indicator, continued

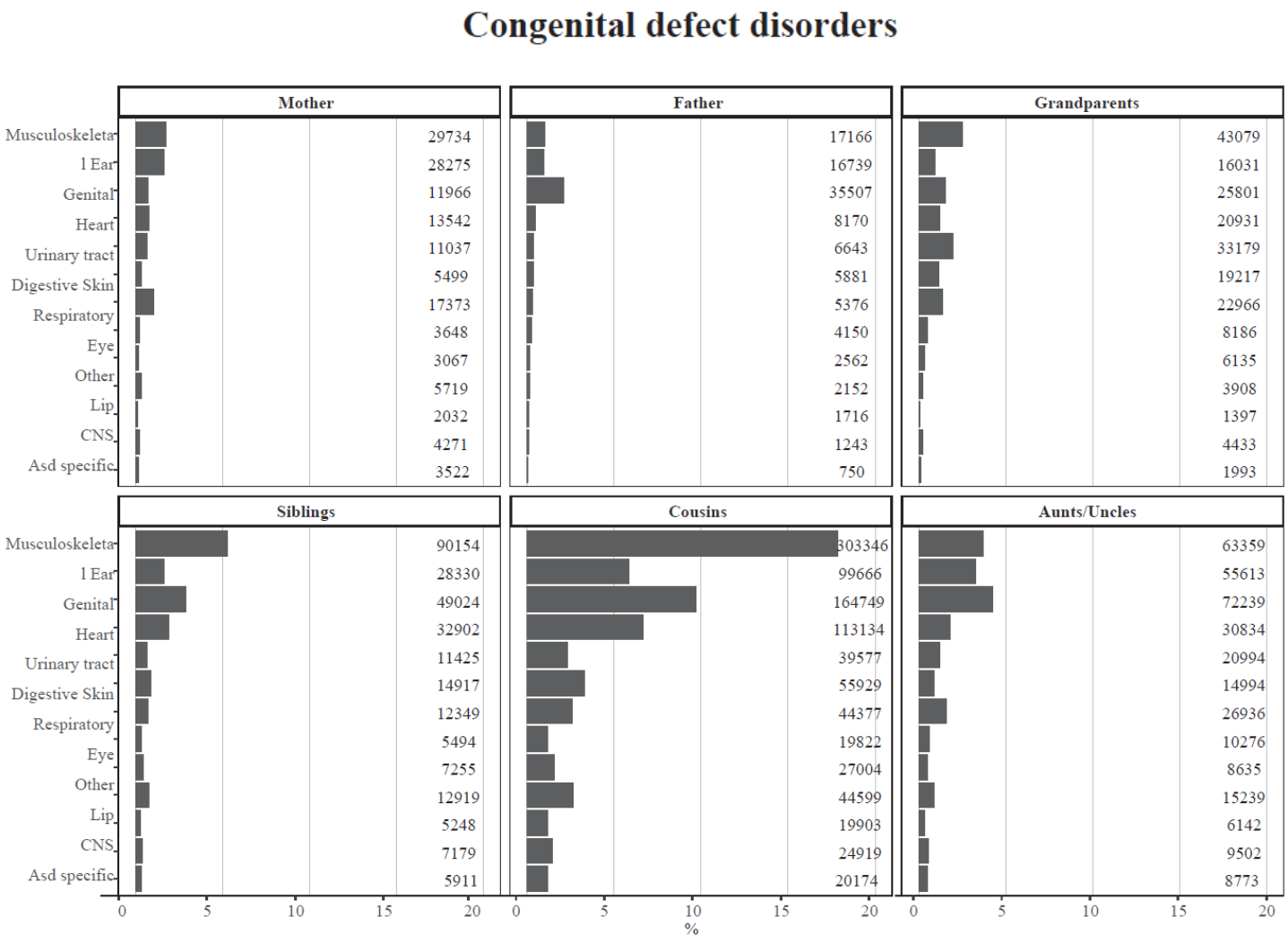


Figure S2. Prevalence and percent (%) for each morbidity indicator, continued

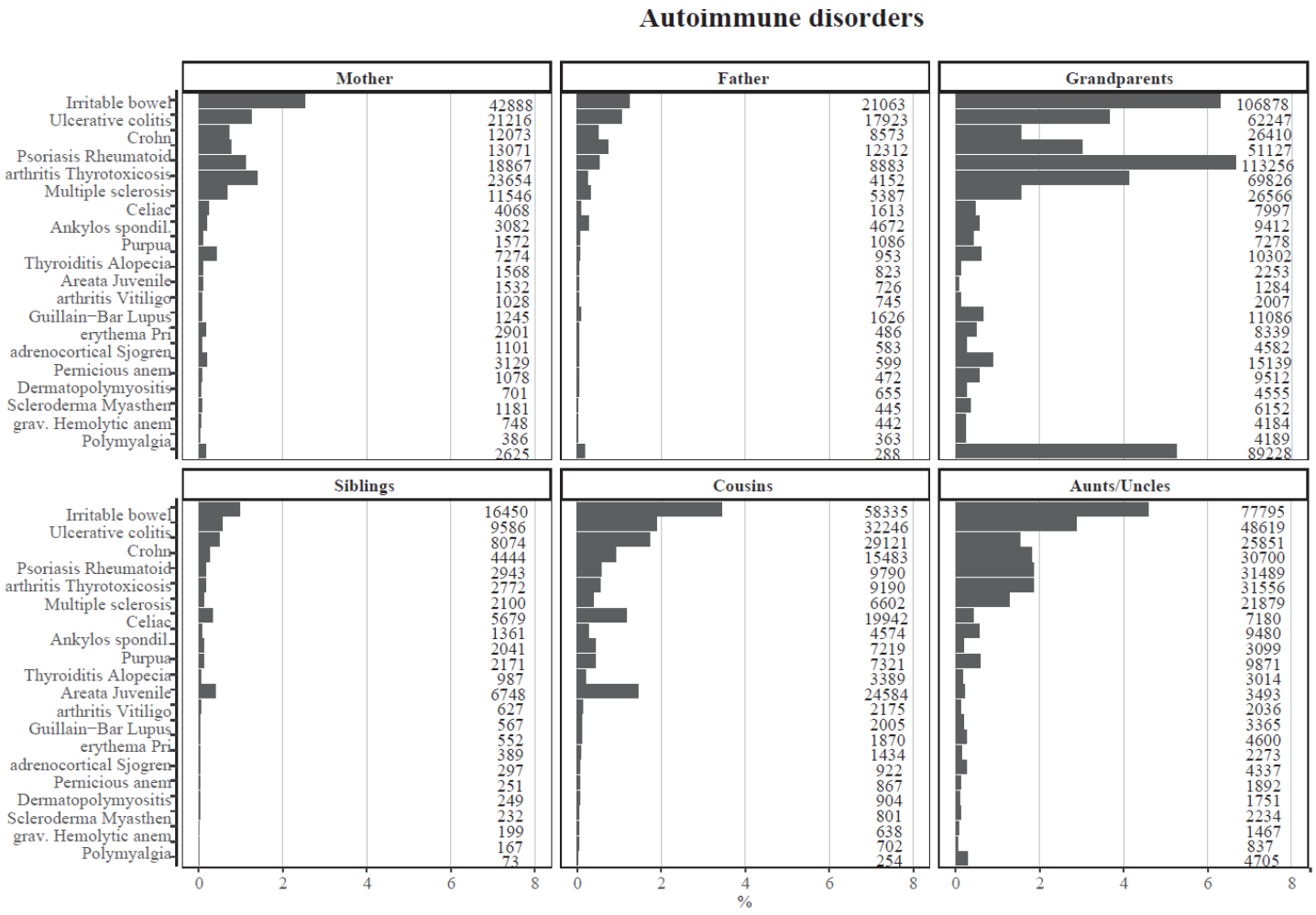


Figure S2. Prevalence and percent (%) for each morbidity indicator, continued

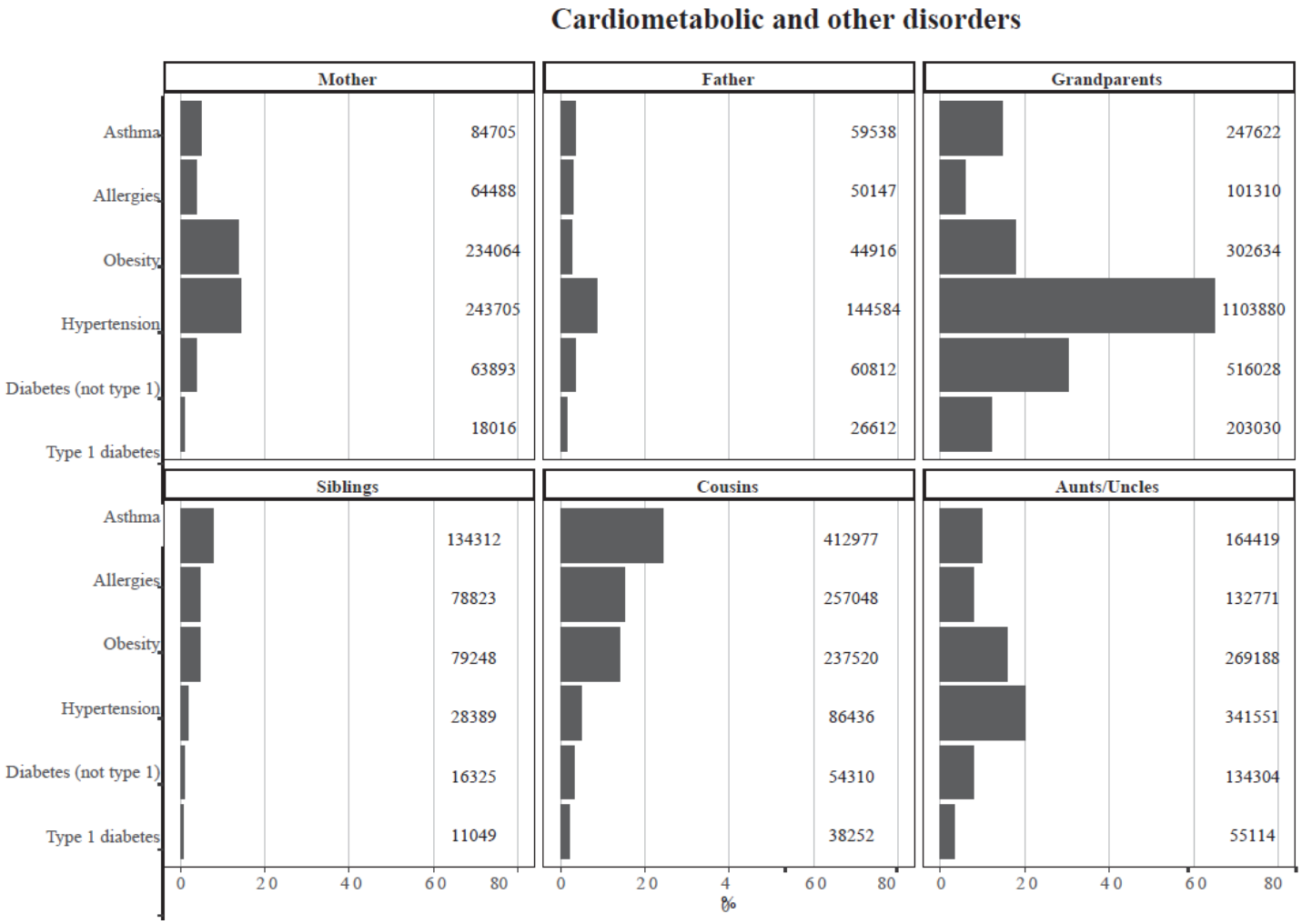
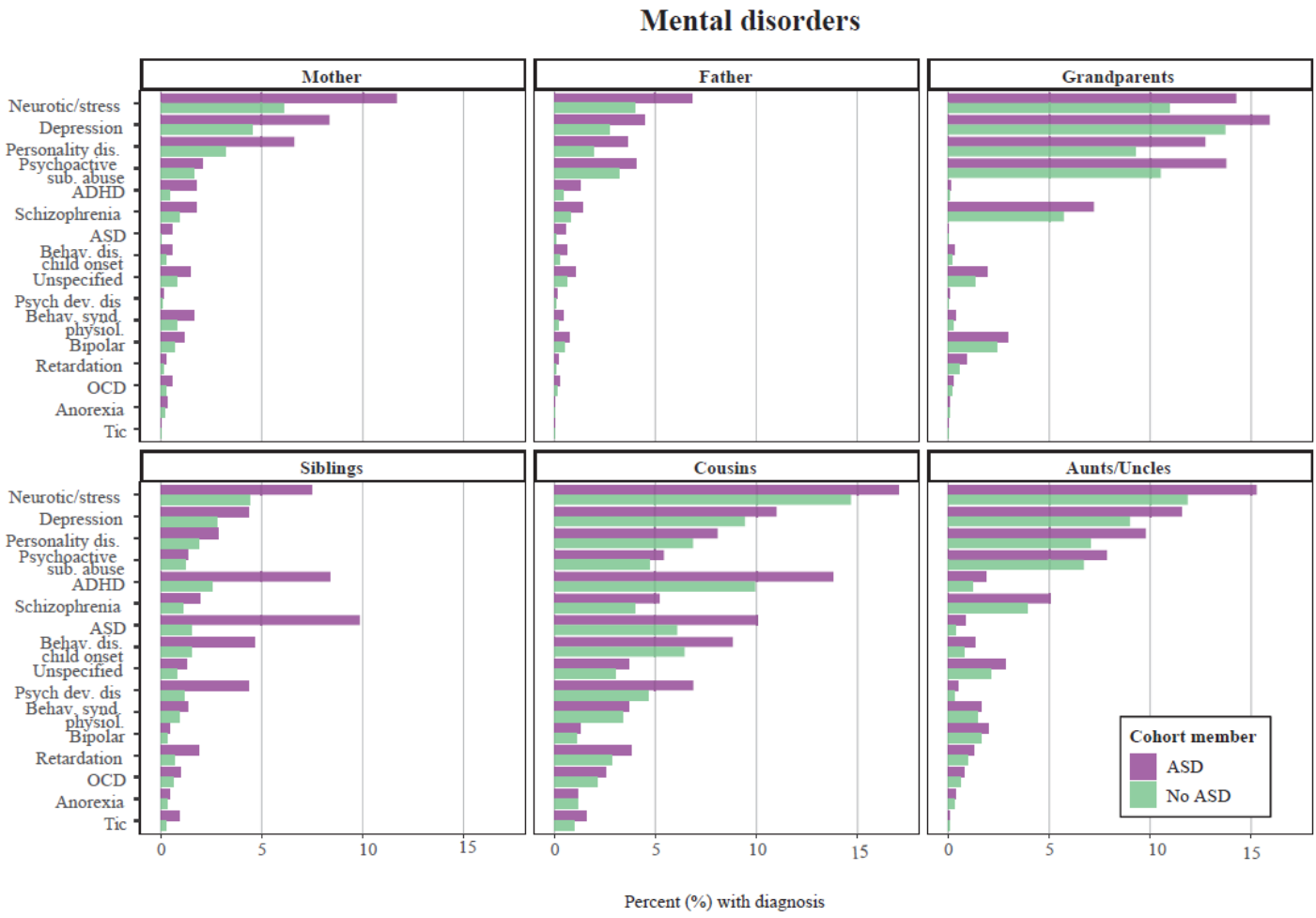


Figure S3. Percent (%) for each morbidity indicator by ASD/no ASD



If there is below 5 cases then the bar is removed

Figure S3. Percent (%) for each morbidity indicator by ASD/no ASD, continued

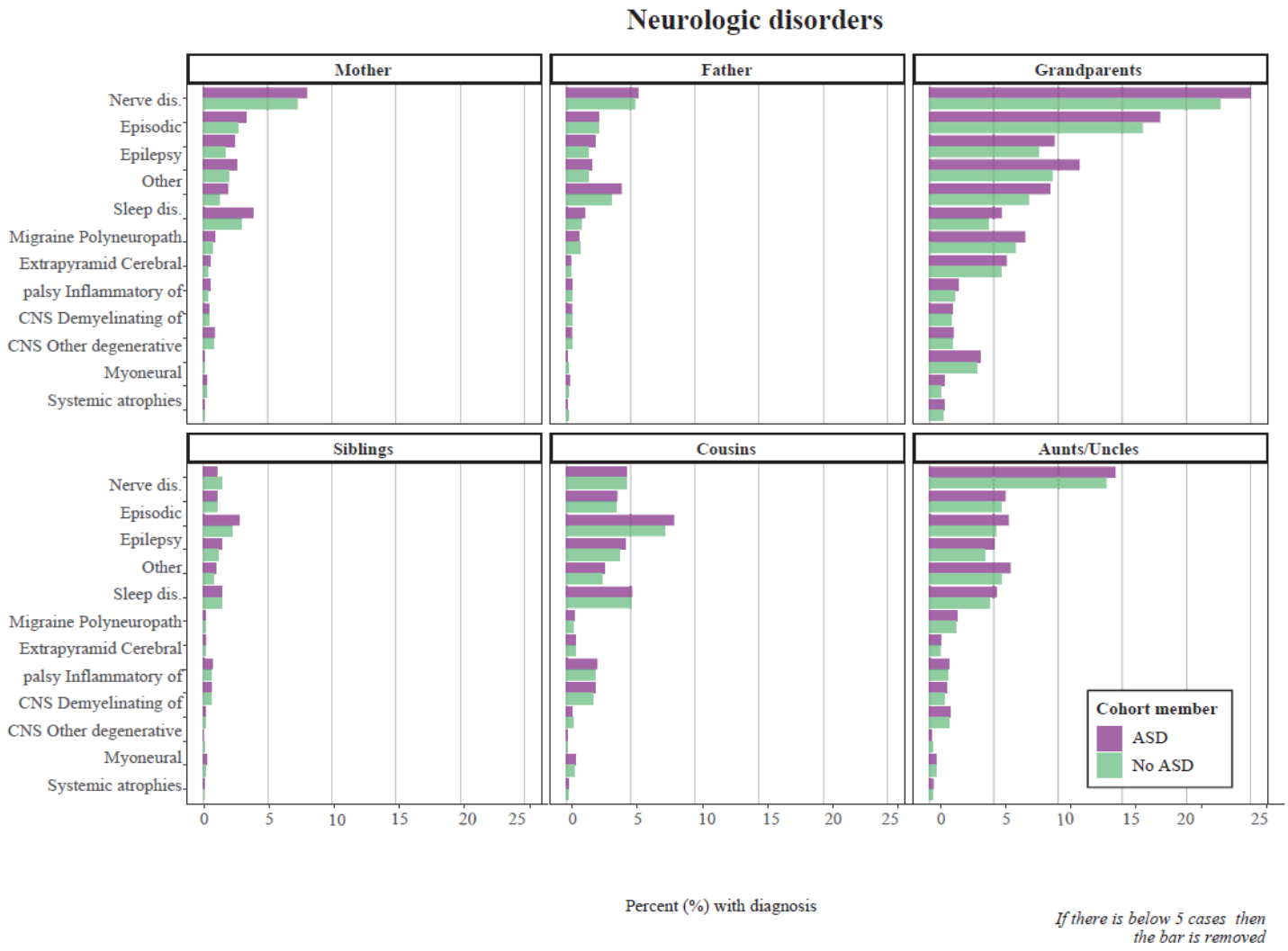
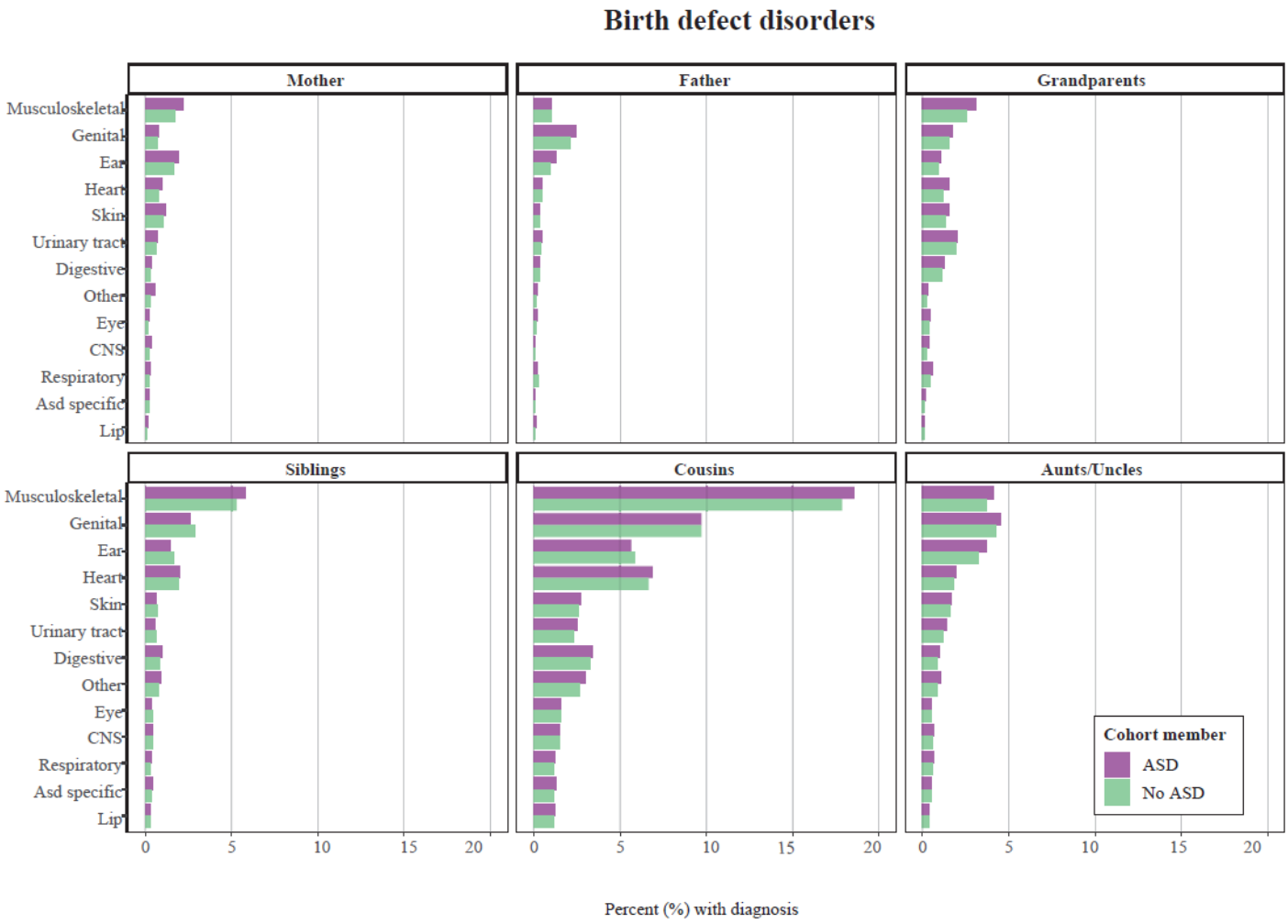
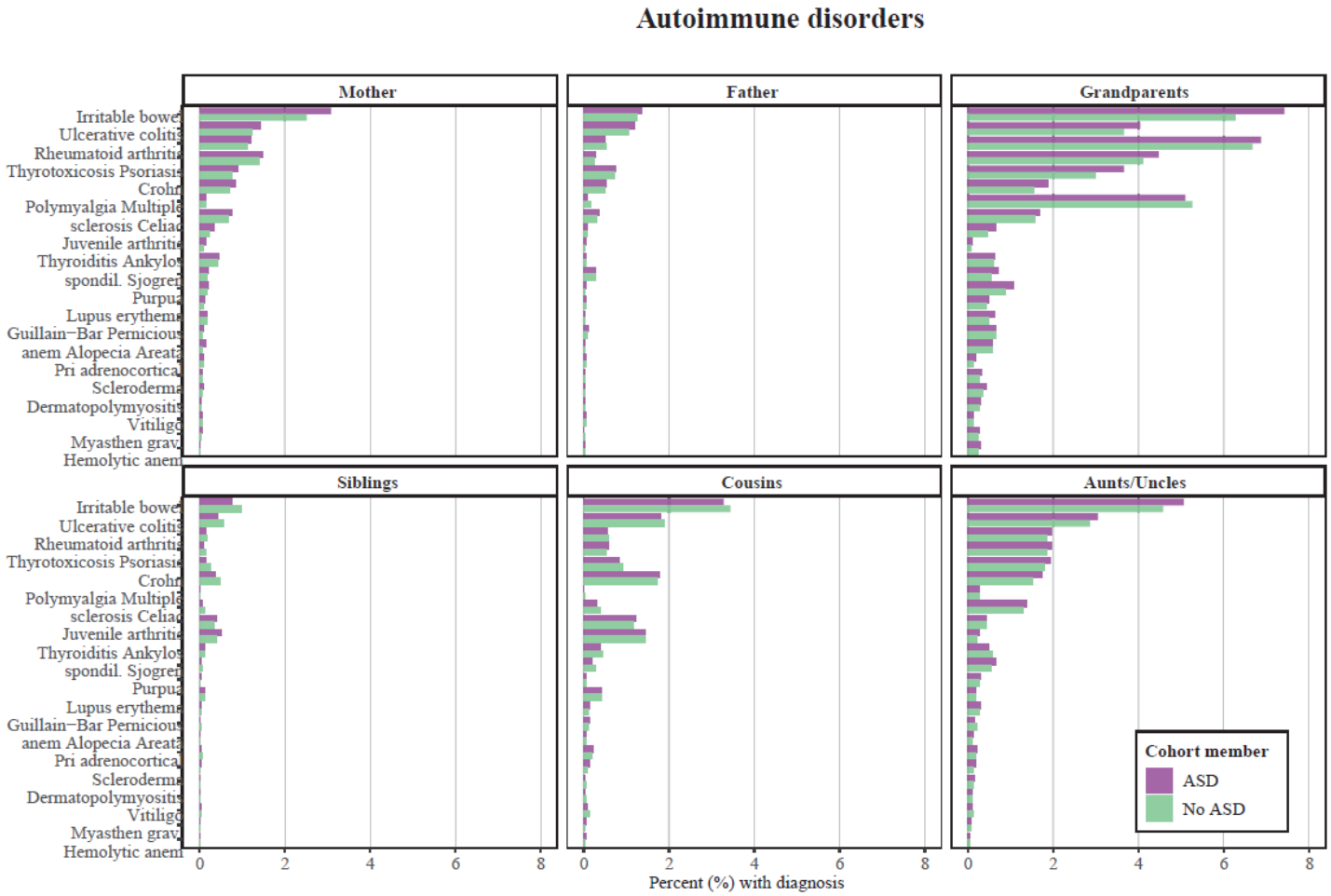


Figure S3. Percent (%) for each morbidity indicator by ASD/no ASD, continued



If there is below 5 cases then the bar is removed

Figure S3. Percent (%) for each morbidity indicator by ASD/no ASD, continued



If there is below 5 cases then the bar is removed

Figure S3. Percent (%) for each morbidity indicator by ASD/no ASD, continued

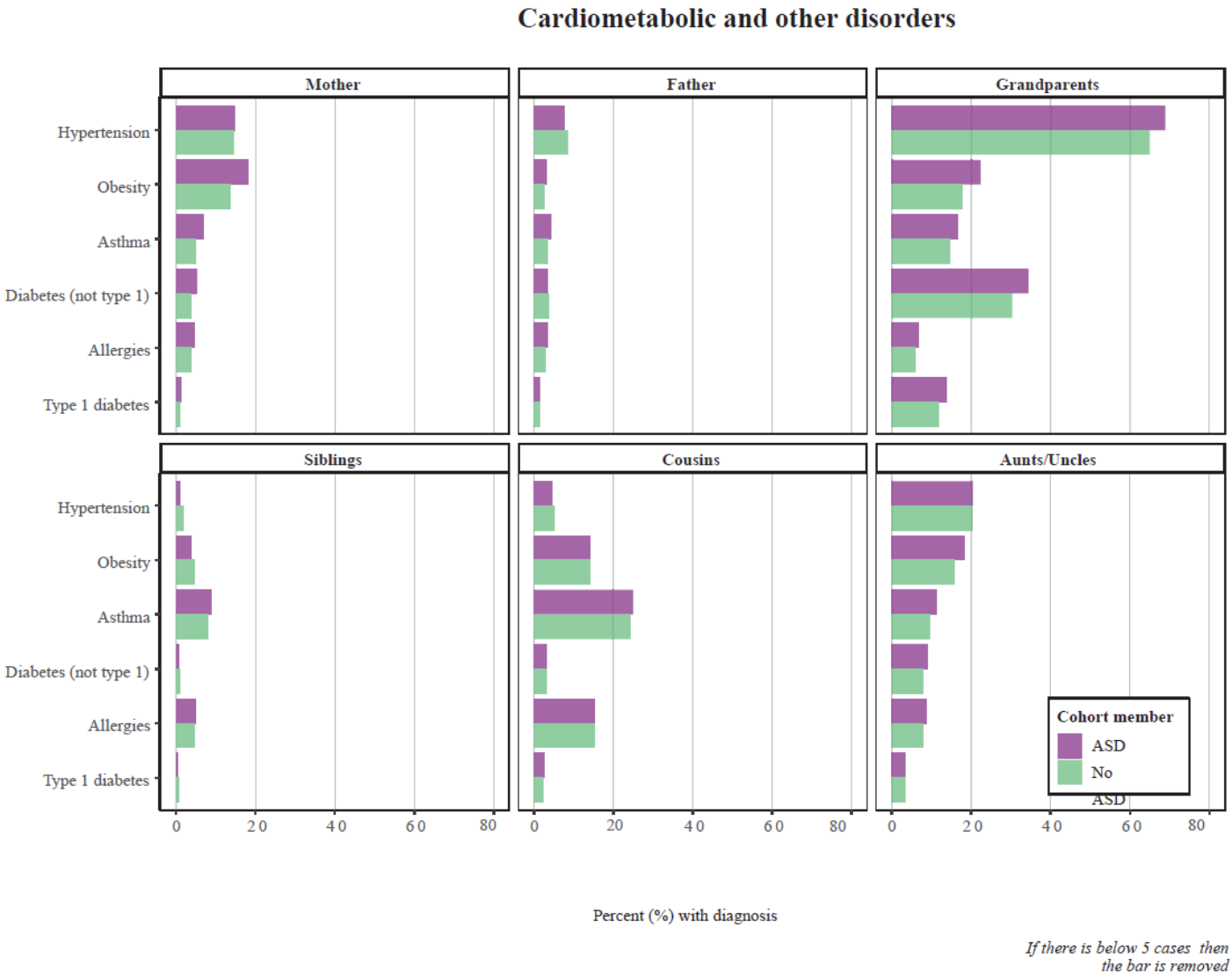


Figure S4. Importance ranking of 30 most important predictors by Random Forest (A) and Extreme Gradient Boost (B)

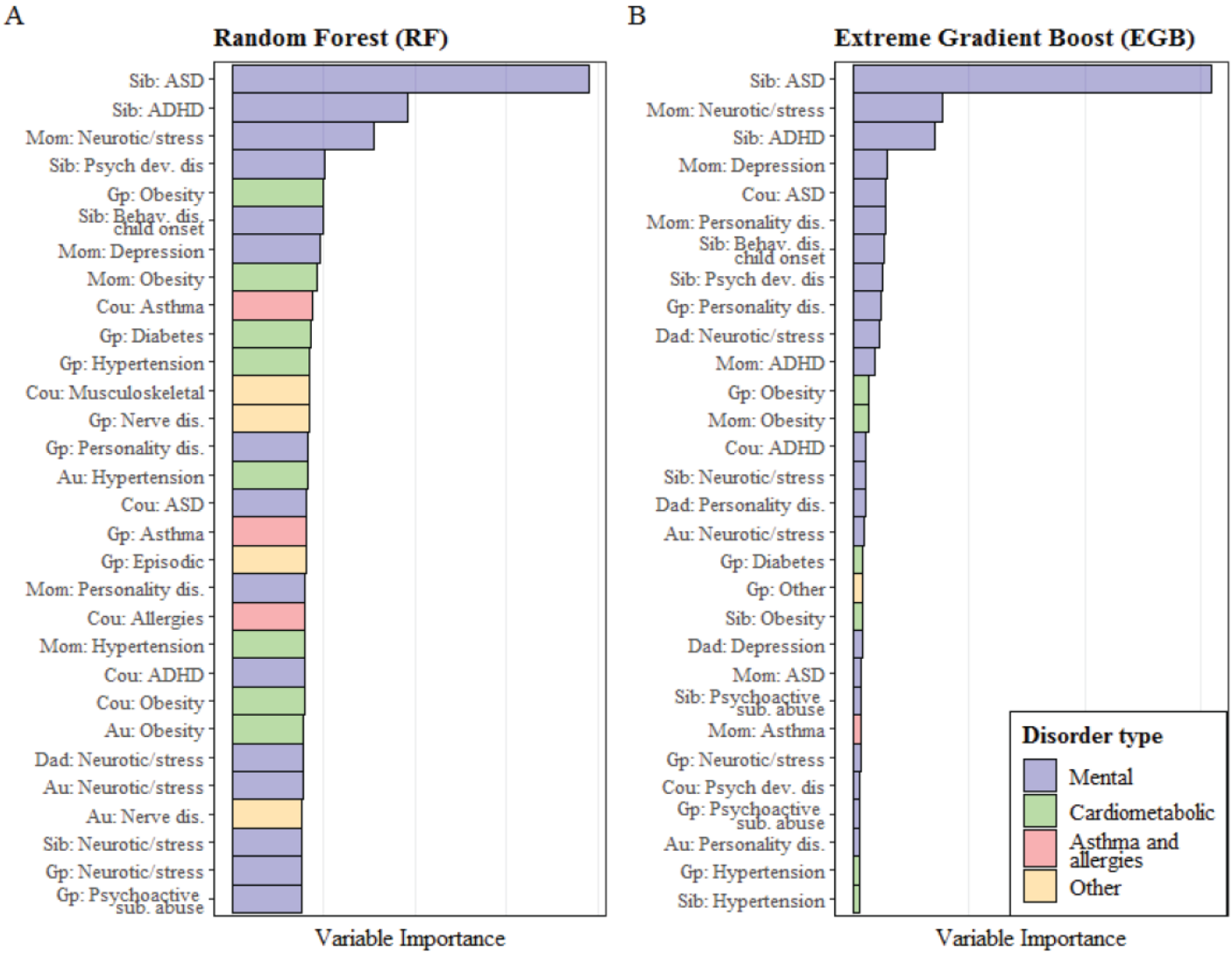


Figure S5. Illustration of performance measures on area-under-the-curve (AUC), F-score and Kappa for all morbidity indicators, top 41, top 21, top 3, respectively

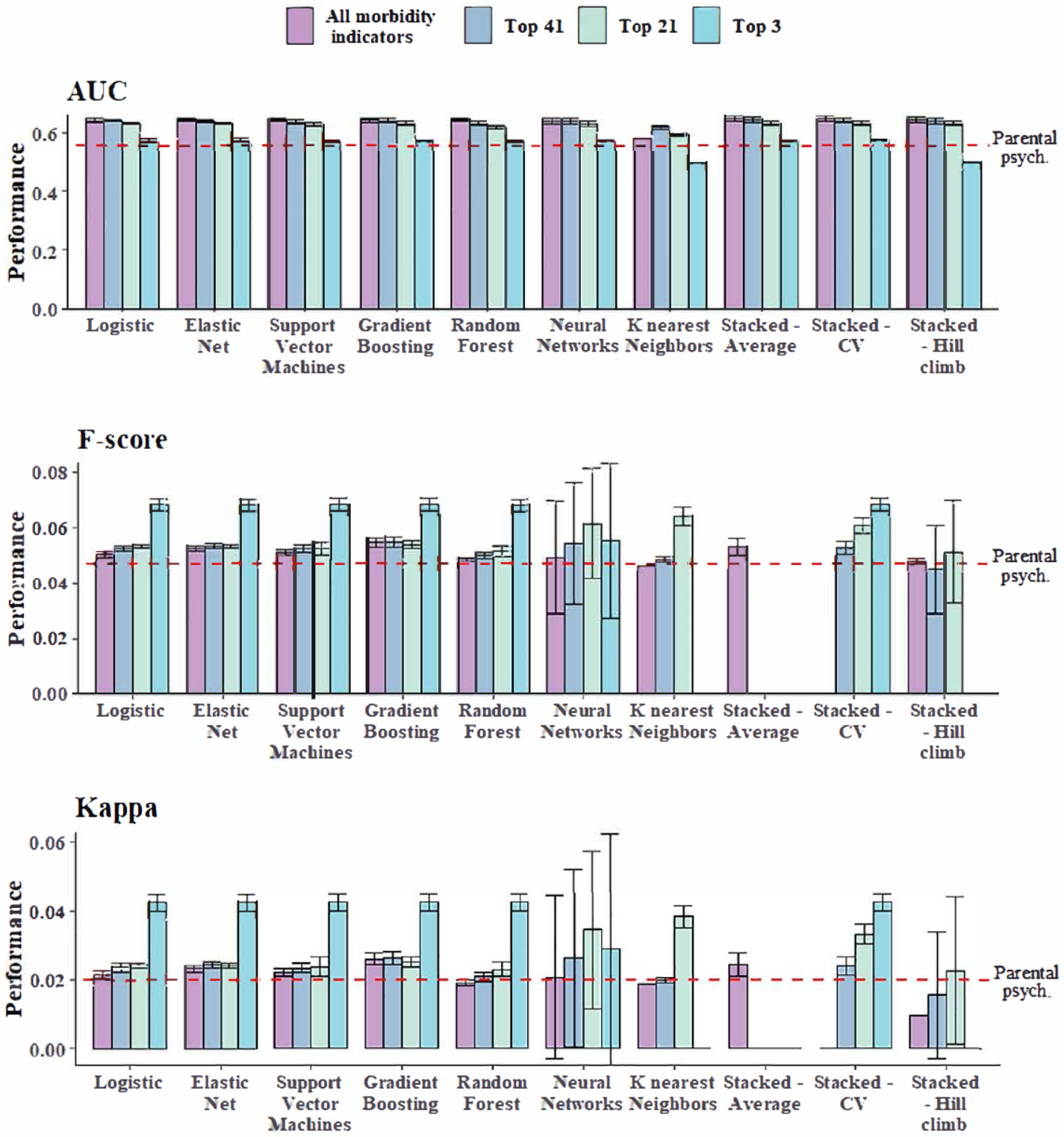


Figure S6. Importance ranking of 30 most important predictors by Random Forest and Extreme Gradient Boost depending on sex of the cohort member

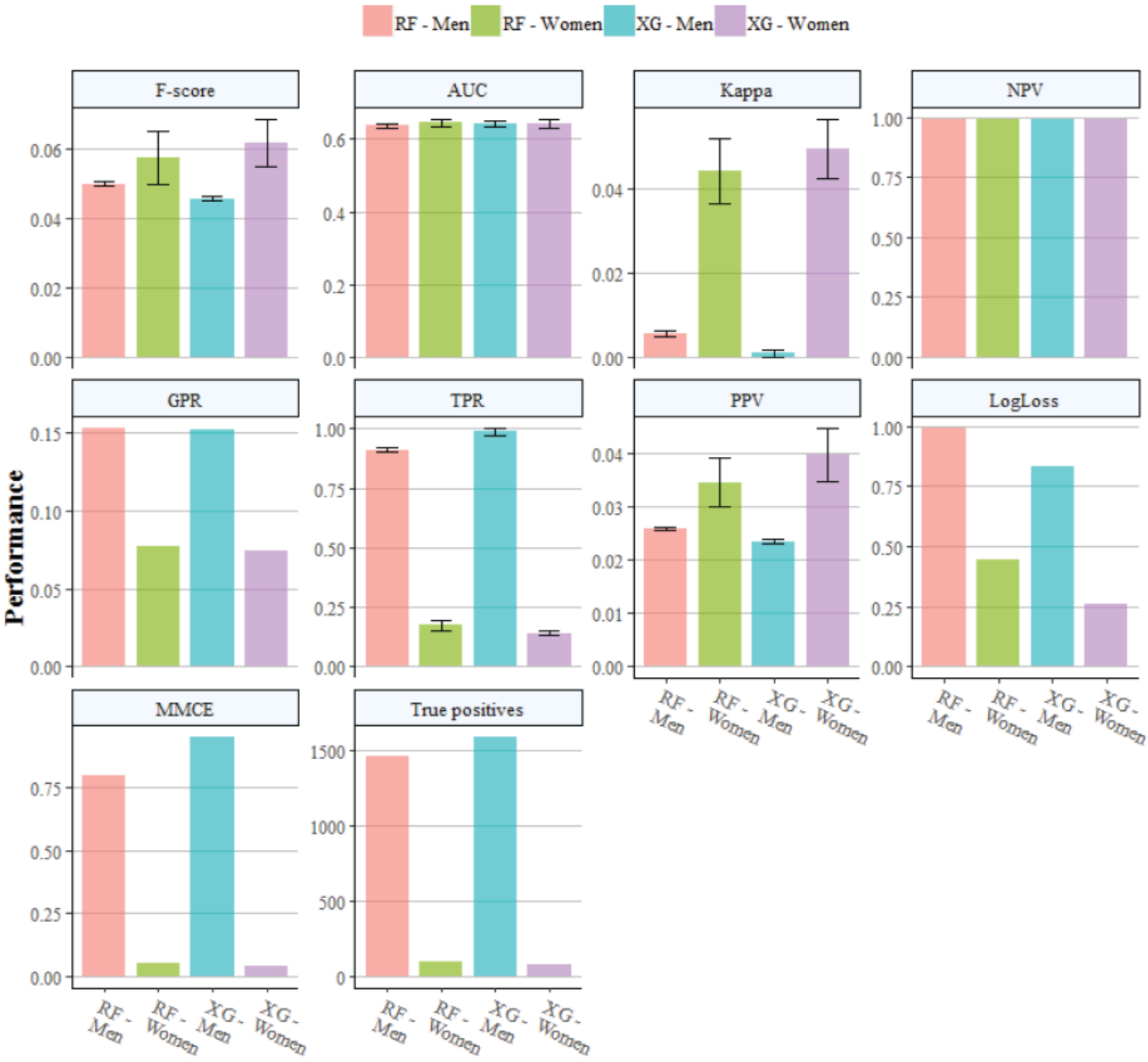
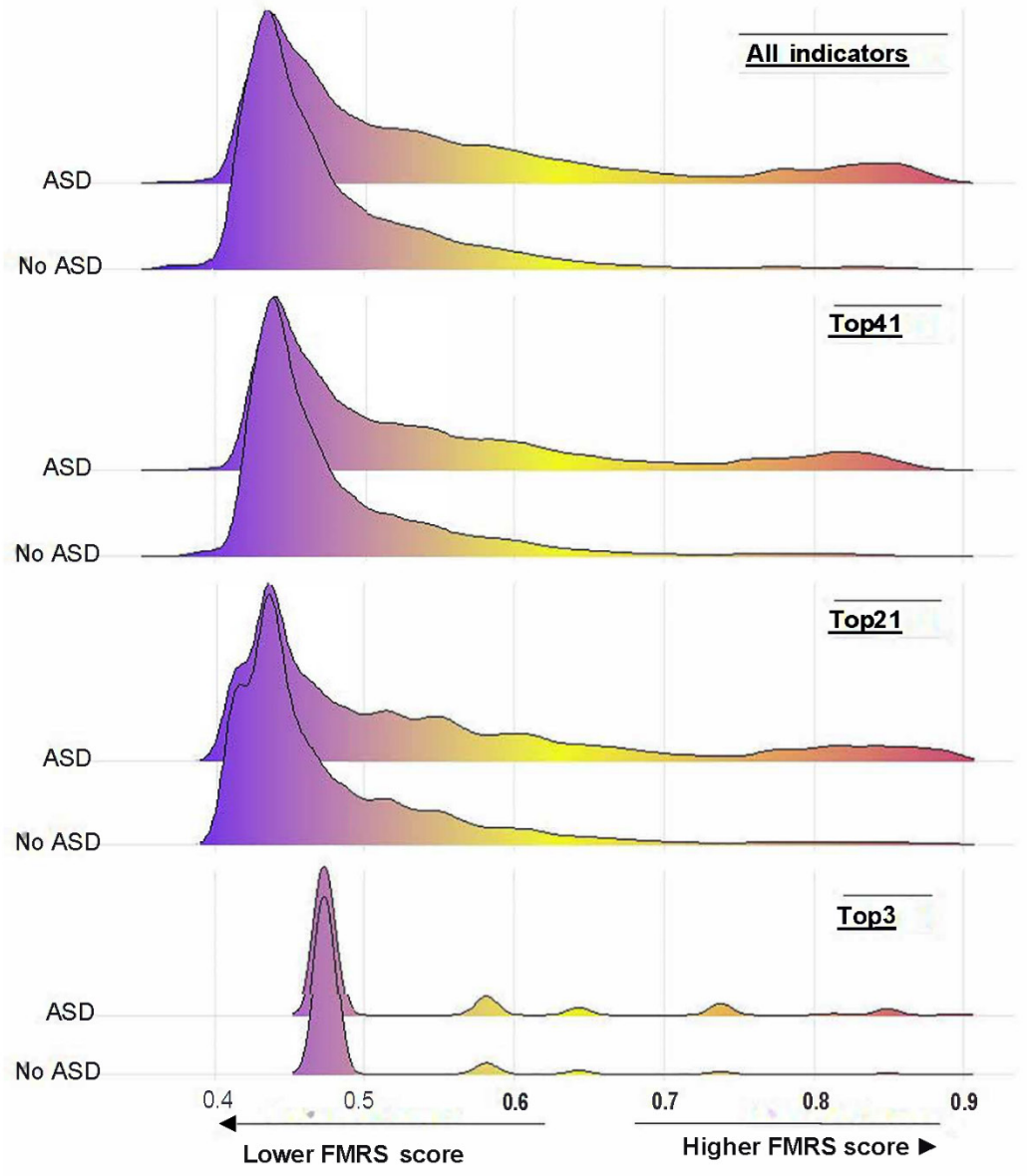


Figure S7. Importance ranking of 30 most important predictors by Random Forest (bottom) and Extreme Gradient Boost (top) for men (A and C) and women (B and D), respectively

Figure S8


All indicators
ASD
No ASD
Top41
ASD
No ASD
Top21
ASD
No ASD
Top3
ASD
No ASD
0.4
0.5
0.6
0.7
0.8
0.9
Lower FMRS score
Higher FMRS score ▶