Weakly Semi-supervised Phenotyping Using Electronic Health Records

Isabelle-Emmanuella Nogues[1], Jun Wen[2], Yucong Lin[2,3], Molei Liu[1], Sara K. Tedeschi[4], Alon Geva[2,5,6], Tianxi Cai[1,2*], Chuan Hong[2*]

[1]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA;

[2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA;

[3]Center for Statistical Science, Tsinghua University, Beijing, China;

[4] Department of Medicine, Division of Rheumatology, Inflammation and Immunity, Brigham and Women's Hospital, Boston, MA, USA;

[5] Department of Anesthesiology, Critical Care, and Pain Medicine, and Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA;

[6] Department of Anesthesia, Harvard Medical School, Boston, MA, USA.

* Cai and Hong contributed equally.

1 Cross-validation

We perform all experiments with results averaged across 100 bootstrap replications of the labeled data, in order to correct for potential randomness in sampling of the labeled observations. Within each bootstrap replication, we apply 5-fold cross-validation.

2 Additional Figures

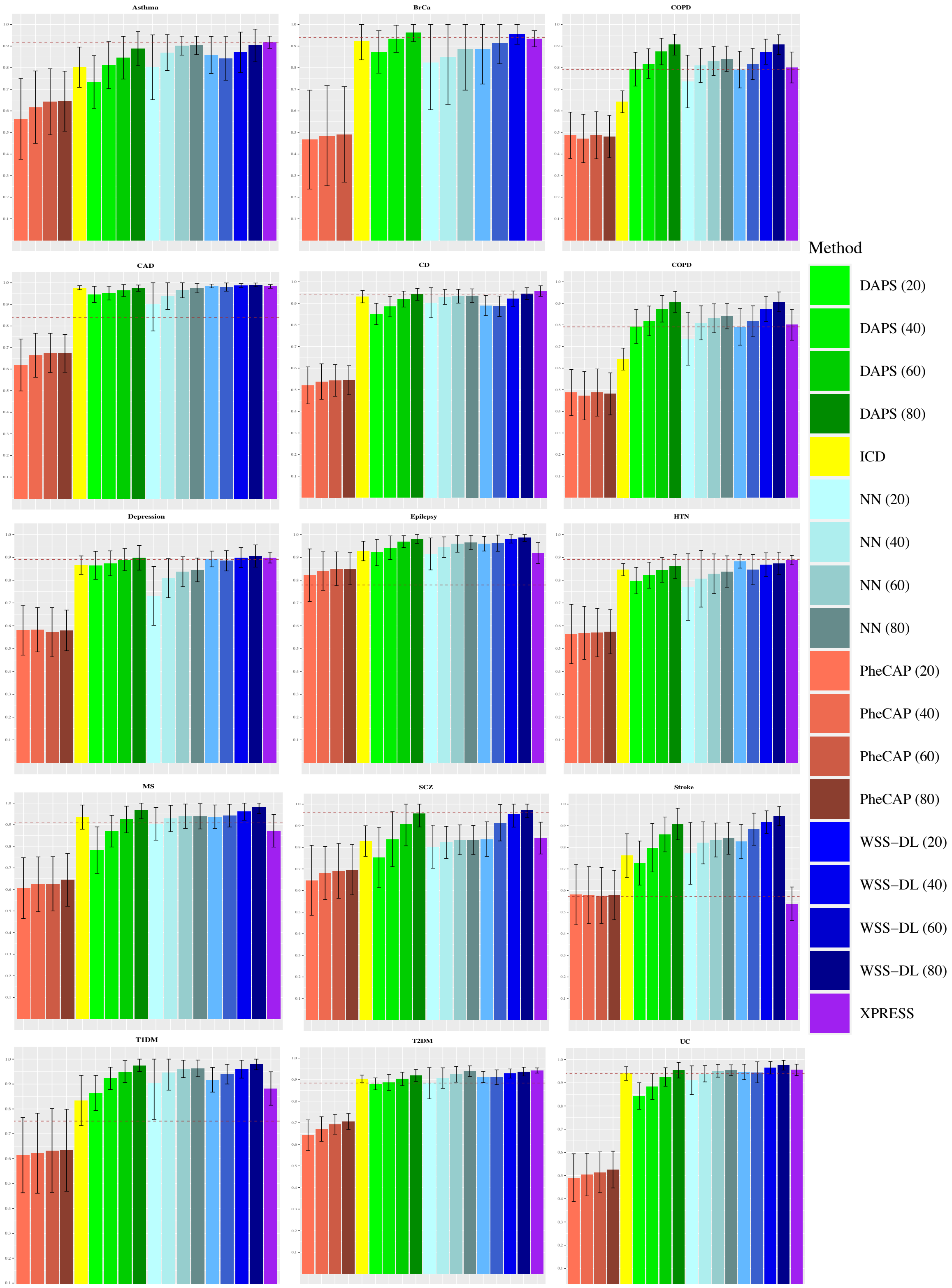

Fig. 1. Comparison of AUCs with gold standard labels for ICD-9 count, MAP, XPRESS, PheCAP (n = 20, 40, 60, and 80), DAPS (n = 20, 40, 60, and 80), NN (n = 20, 40, 60, and 80), and WSS-DL (n = 20, 40, 60, and 80) for 15 disease phenotypes. From left to right, top to bottom: Asthma, Breast Cancer, Chronic Obstructive Pulmonary Disorder, Depression, Epilepsy, Hypertension, Multiple Sclerosis, Rheumatoid Arthritis, Schizophrenia, Stroke, Type 1 Diabetes Mellitus, Type 2 Diabetes Mellitus, and Ulcerative Colitis.

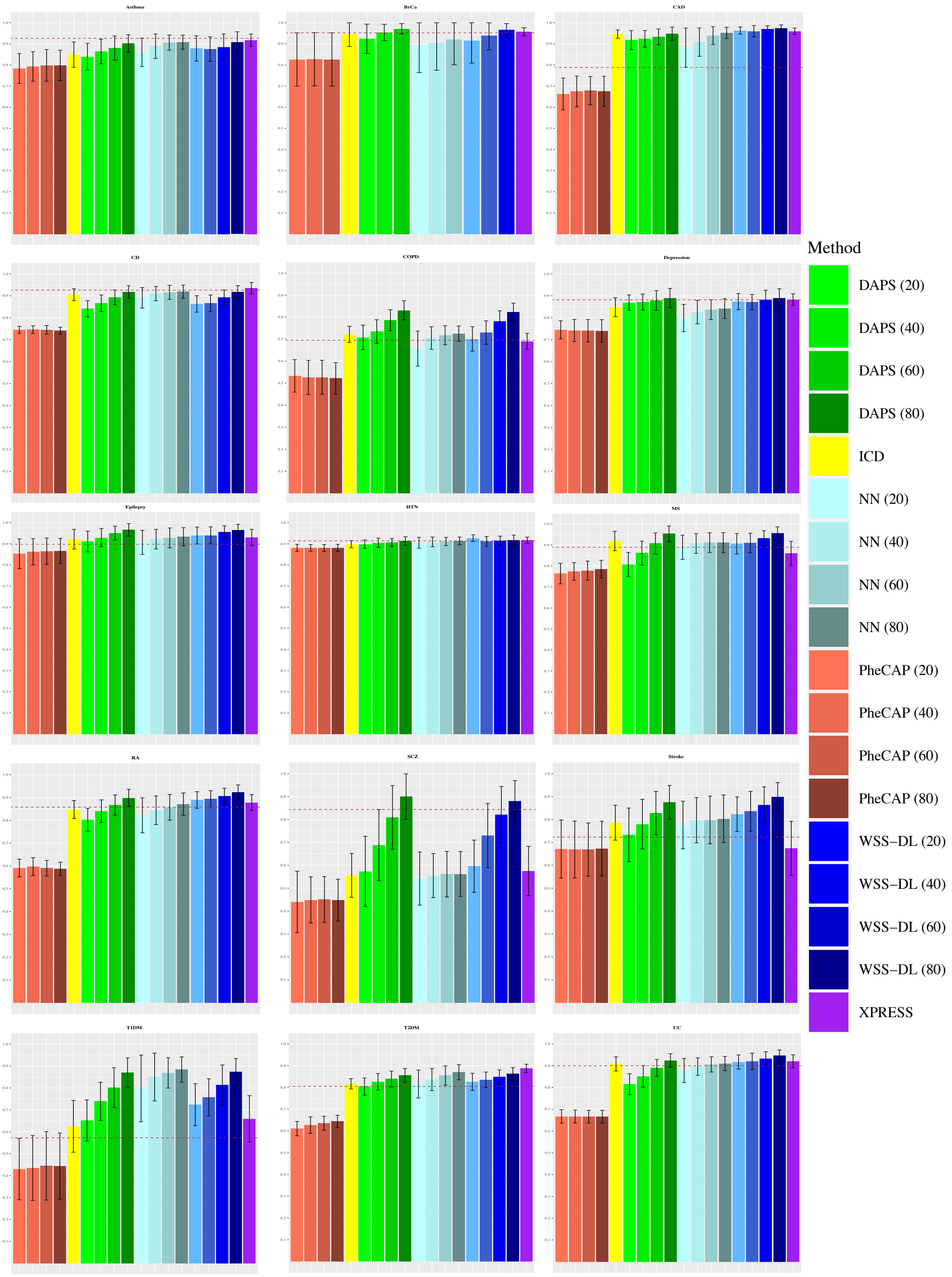

Fig. 2. Comparison of F-scores with gold standard labels for ICD-9 count, MAP, XPRESS, PheCAP (n = 20, 40, 60, and 80), DAPS (n = 20, 40, 60, and 80), NN (n = 20, 40, 60, and 80), and WSS-DL (n = 20, 40, 60, and 80) for 15 disease phenotypes. From left to right, top to bottom: Asthma, Breast Cancer, Chronic Obstructive Pulmonary Disorder, Depression, Epilepsy, Hypertension, Multiple Sclerosis, Rheumatoid Arthritis, Schizophrenia, Stroke, Type 1 Diabetes Mellitus, Type 2 Diabetes Mellitus, and Ulcerative Colitis.

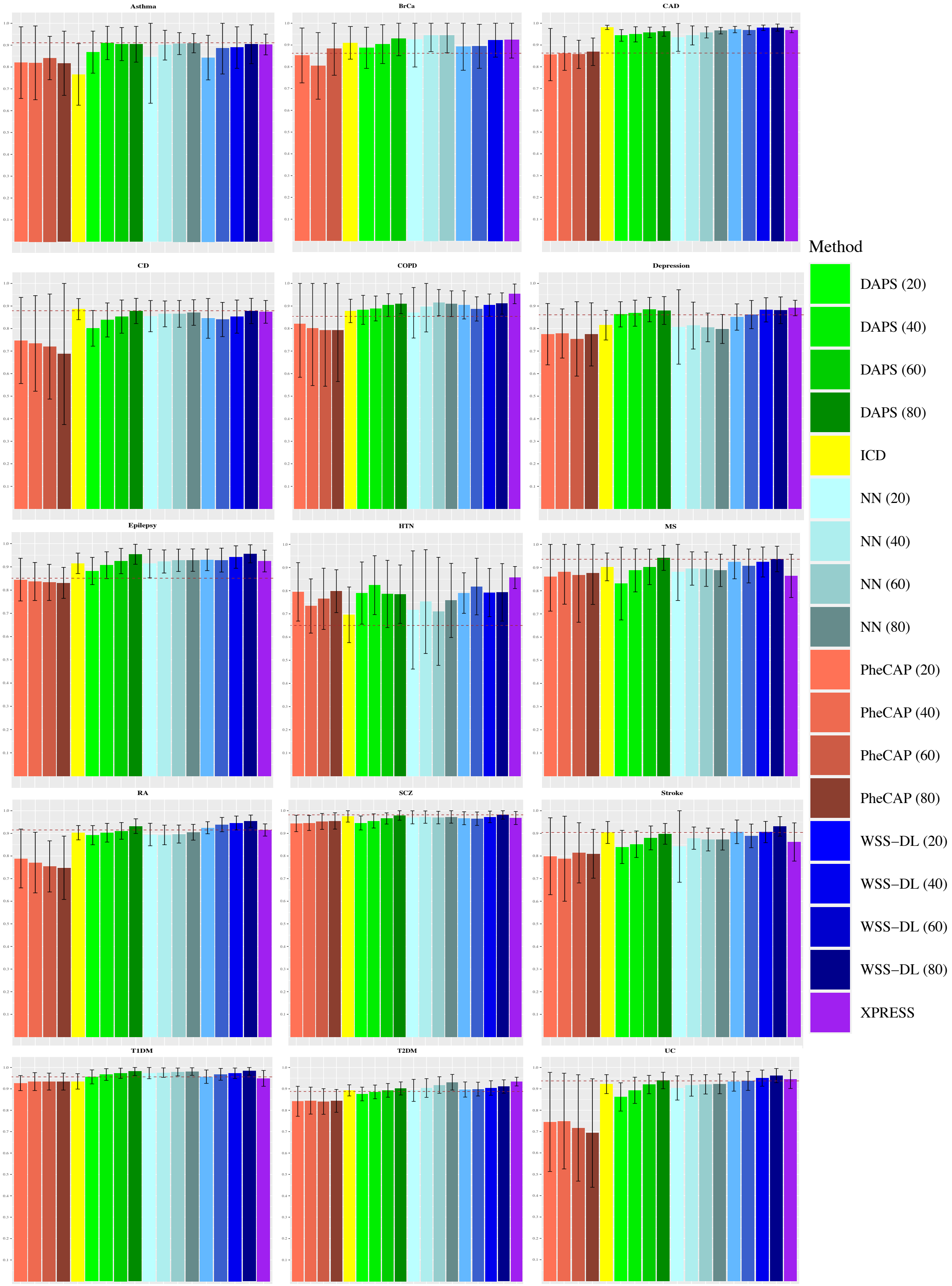

Fig. 3. Comparison of NPVs with gold standard labels for ICD-9 count, MAP, XPRESS, PheCAP (n = 20, 40, 60, and 80), DAPS (n = 20, 40, 60, and 80), NN (n = 20, 40, 60, and 80), and WSS-DL (n = 20, 40, 60, and 80) for 15 disease phenotypes. From left to right, top to bottom: Asthma, Breast Cancer, Chronic Obstructive Pulmonary Disorder, Depression, Epilepsy, Hypertension, Multiple Sclerosis, Rheumatoid Arthritis, Schizophrenia, Stroke, Type 1 Diabetes Mellitus, Type 2 Diabetes Mellitus, and Ulcerative Colitis.

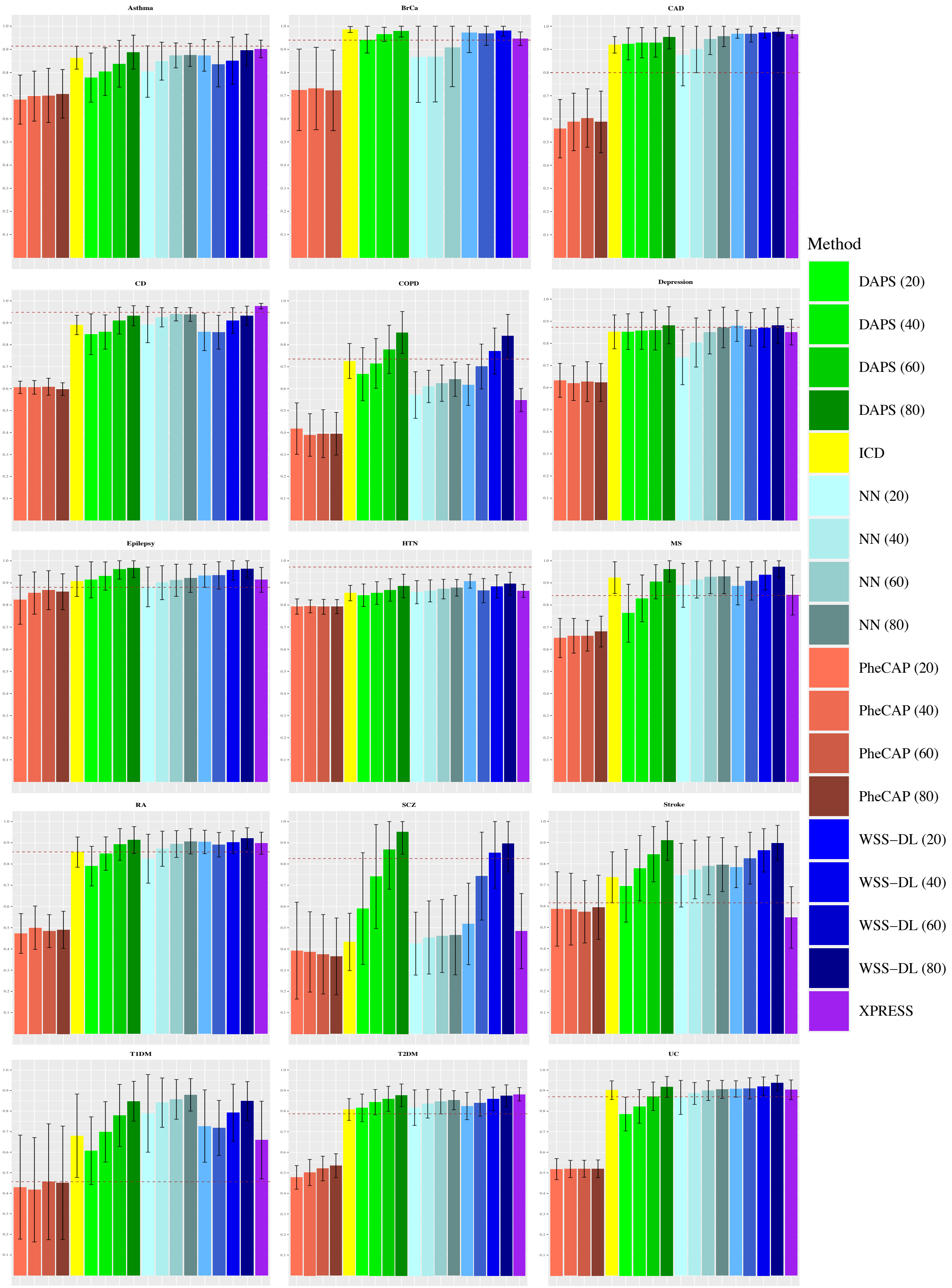

Fig. 4. Comparison of PPVs with gold standard labels for ICD-9 count, MAP, XPRESS, PheCAP (n = 20, 40, 60, and 80), DAPS (n = 20, 40, 60, and 80), NN (n = 20, 40, 60, and 80), and WSS-DL (n = 20, 40, 60, and 80) for 15 disease phenotypes. From left to right, top to bottom: Asthma, Breast Cancer, Chronic Obstructive Pulmonary Disorder, Depression, Epilepsy, Hypertension, Multiple Sclerosis, Rheumatoid Arthritis, Schizophrenia, Stroke, Type 1 Diabetes Mellitus, Type 2 Diabetes Mellitus, and Ulcerative Colitis.

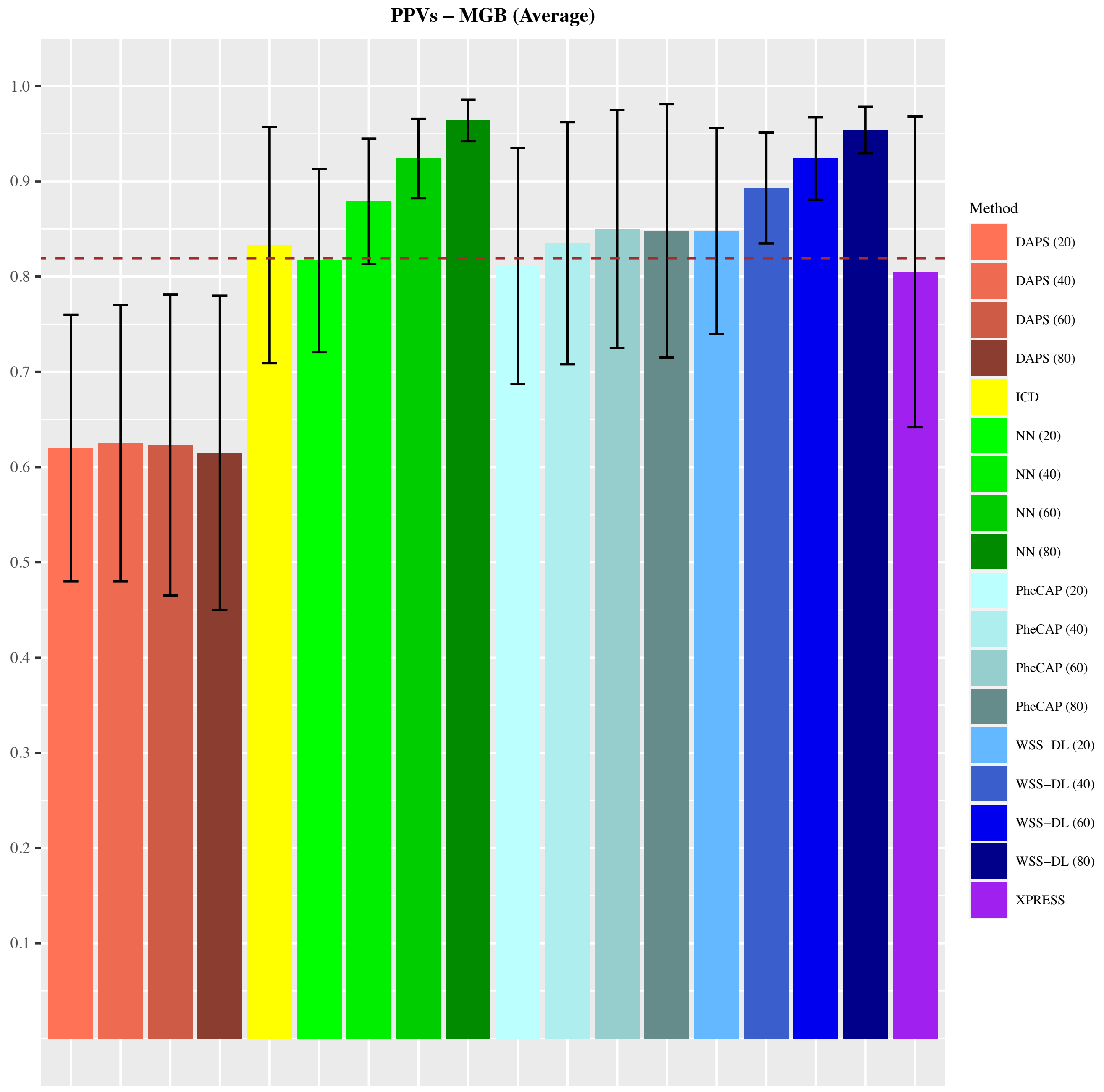

Fig. 5 Comparison of PPVs with gold standard labels for ICD-9 count, MAP, XPRESS, PheCAP (n = 20, 40, 60, and 80), DAPS (n = 20, 40, 60, and 80), NN (n = 20, 40, 60, and 80), and WSS-DL (n = 20, 40, 60, and 80), averaged over 15 phenotypes (MGB).

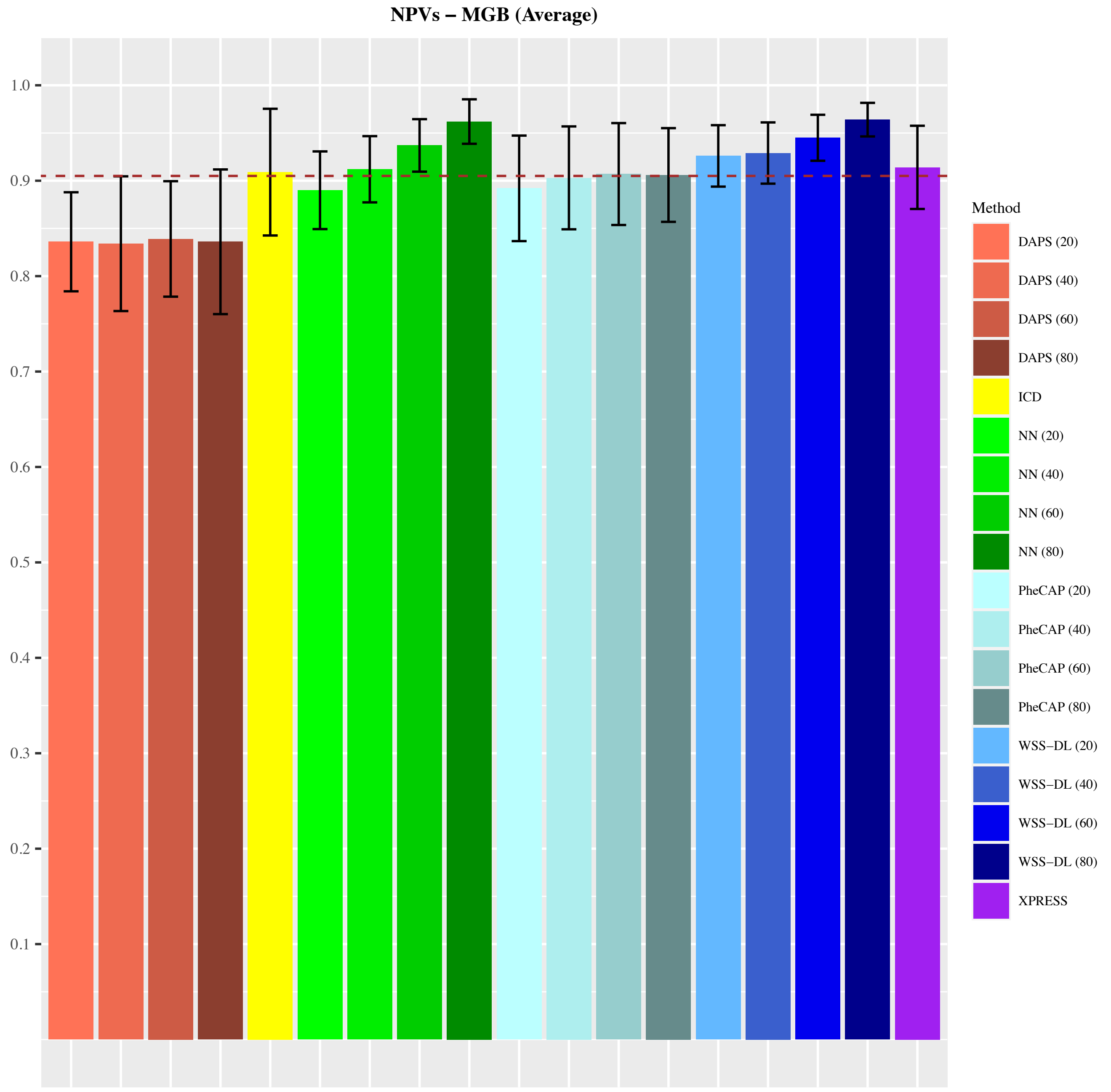

Fig. 6 Comparison of NPVs with gold standard labels for ICD-9 count, MAP, XPRESS, PheCAP (n = 20, 40, 60, and 80), DAPS (n = 20, 40, 60, and 80), NN (n = 20, 40, 60, and 80), and WSS-DL (n = 20, 40, 60, and 80), averaged over 15 disease phenotypes (MGB).

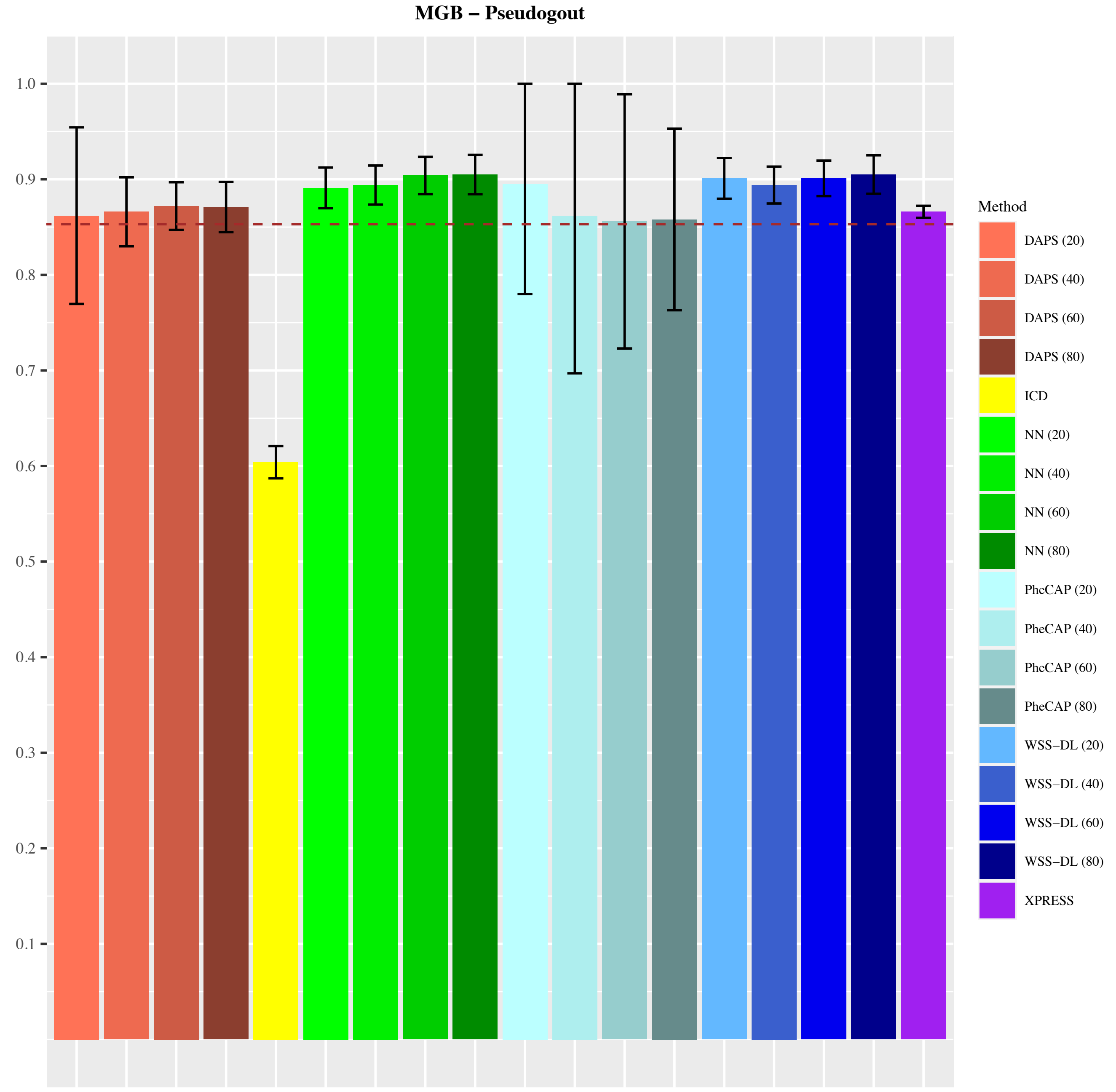

Fig. 7 Comparison of NPVs with gold standard labels for ICD-9 count, MAP, XPRESS, PheCAP (n = 20, 40, 60, and 80), DAPS (n = 20, 40, 60, and 80), NN (n = 20, 40, 60, and 80), and WSS-DL (n = 20, 40, 60, and 80), for MGB – Pseudogout cohort.

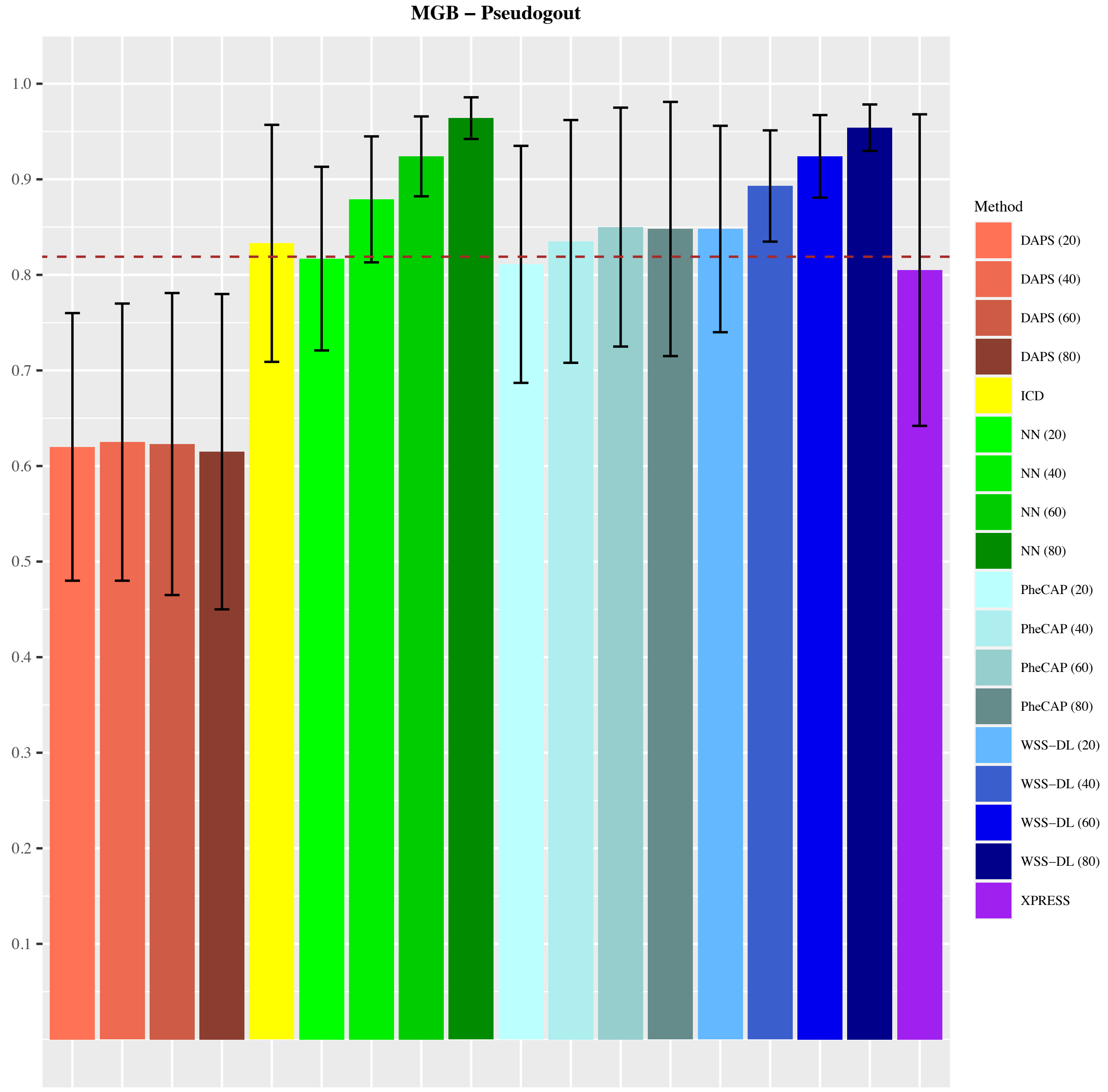

Fig. 8 Comparison of PPVs with gold standard labels for ICD-9 count, MAP, XPRESS, PheCAP (n = 20, 40, 60, and 80), DAPS (n = 20, 40, 60, and 80), NN (n = 20, 40, 60, and 80), and WSS-DL (n = 20, 40, 60, and 80), for MGB-Pseudogout cohort.

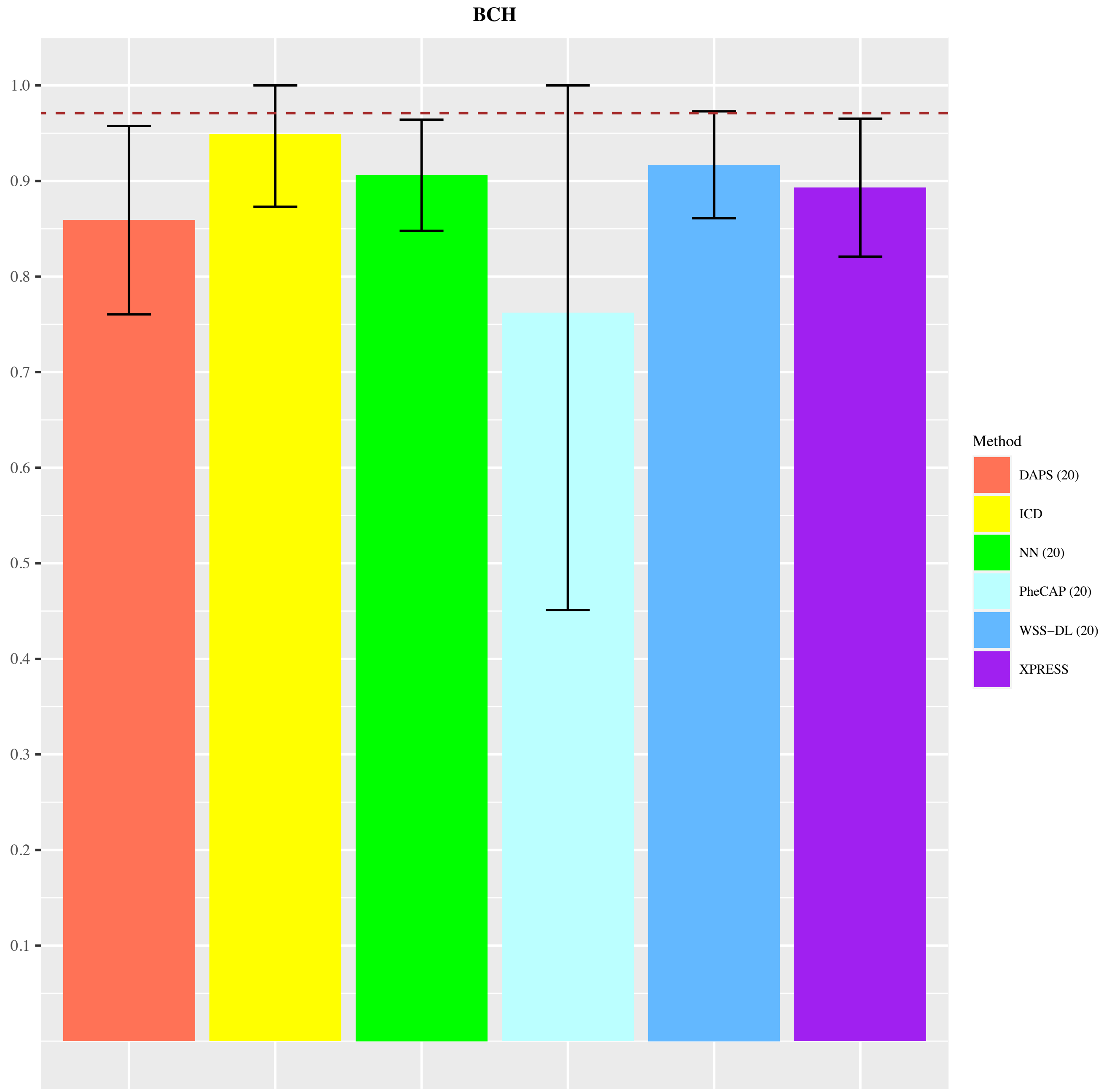

Fig. 9 Comparison of NPVs with gold standard labels for ICD-9 count, MAP, XPRESS, PheCAP (n = 20), DAPS (n = 20), NN (n = 20), and WSS-DL (n = 20) for BCH cohort.

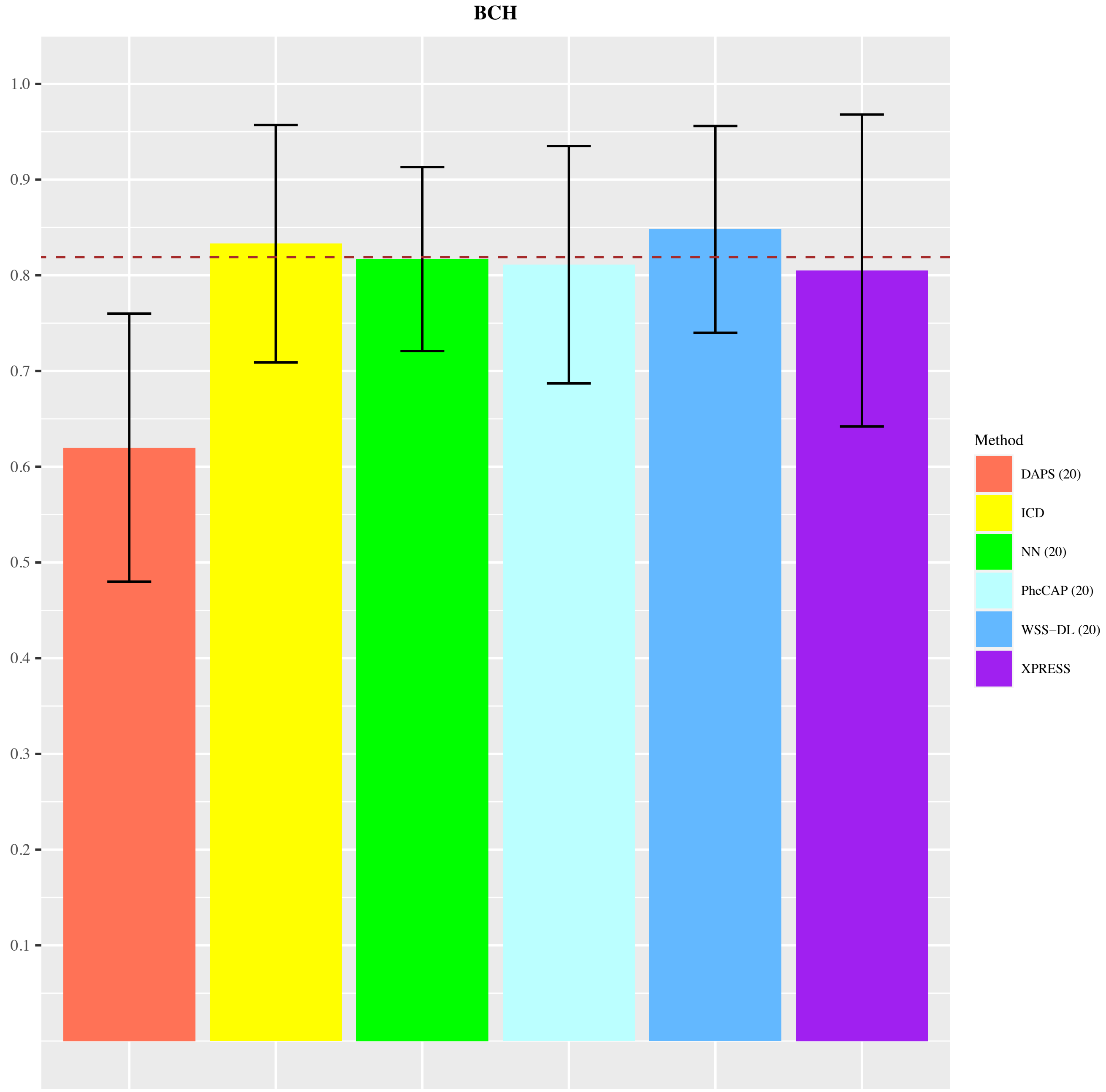

Fig. 10 Comparison of PPVs with gold standard labels for ICD-9 count, MAP, XPRESS, PheCAP (n = 20), DAPS (n = 20), NN (n = 20), and WSS-DL (n = 20) for BCH cohort.

Method	Asthma	PARDS	BrCa	CAD	CD	COPD	Depression	Epilepsy	HTN	MS	pGout	RA	SCZ	Stroke	T1DM	T2DM	UC
WSS-DL (20)	0.859	0.907	0.886	0.985	0.89	0.791	0.893	0.96	0.883	0.937	0.696	0.942	0.838	0.827	0.917	0.912	0.947
WSS-DL (40)	0.843	NA	0.915	0.98	0.887	0.817	0.885	0.962	0.846	0.942	0.777	0.947	0.914	0.884	0.94	0.911	0.945
WSS-DL (60)	0.871	NA	0.957	0.987	0.922	0.874	0.899	0.981	0.868	0.961	0.797	0.957	0.955	0.917	0.96	0.928	0.965
WSS-DL (80)	0.903	NA	NA	0.99	0.945	0.907	0.906	0.986	0.874	0.981	0.817	0.969	0.975	0.945	0.98	0.937	0.976
ICD	0.802	0.521	0.924	0.977	0.931	0.642	0.866	0.928	0.846	0.935	0.604	0.923	0.829	0.762	0.834	0.904	0.94
Silver std.	0.918	0.919	0.94	0.838	0.939	0.791	0.89	0.779	0.89	0.908	0.717	0.926	0.963	0.573	0.751	0.884	0.939
PheCAP (20)	0.802	0.53	0.824	0.9	0.903	0.736	0.731	0.915	0.77	0.904	0.535	0.882	0.803	0.772	0.904	0.883	0.911
PheCAP (40)	0.87	NA	0.851	0.939	0.93	0.81	0.809	0.945	0.806	0.929	0.546	0.895	0.823	0.821	0.947	0.907	0.939
PheCAP (60)	0.902	NA	0.886	0.966	0.933	0.831	0.837	0.959	0.828	0.939	0.561	0.909	0.835	0.834	0.961	0.924	0.952
PheCAP (80)	0.903	NA	NA	0.975	0.936	0.841	0.845	0.965	0.838	0.939	0.572	0.918	0.834	0.842	0.963	0.939	0.954
Xpress	0.918	0.772	0.934	0.983	0.957	0.801	0.899	0.919	0.888	0.872	0.748	0.939	0.843	0.539	0.882	0.942	0.956
DAPS (20)	0.563	0.636	0.467	0.619	0.52	0.487	0.581	0.822	0.564	0.606	0.603	0.577	0.647	0.581	0.614	0.642	0.491
DAPS (40)	0.617	NA	0.485	0.664	0.538	0.472	0.583	0.84	0.569	0.624	0.62	0.597	0.681	0.579	0.622	0.671	0.504
DAPS (60)	0.642	NA	0.491	0.675	0.543	0.487	0.572	0.85	0.57	0.626	0.63	0.594	0.691	0.577	0.633	0.693	0.514
DAPS (80)	0.645	NA	NA	0.673	0.544	0.481	0.58	0.85	0.574	0.644	0.634	0.595	0.697	0.579	0.634	0.706	0.526
NN (20)	0.734	0.912	0.873	0.946	0.851	0.793	0.865	0.921	0.798	0.782	0.742	0.878	0.753	0.727	0.864	0.88	0.843
NN (40)	0.812	NA	0.934	0.952	0.885	0.819	0.874	0.942	0.822	0.87	0.776	0.9	0.838	0.798	0.923	0.887	0.884
NN (60)	0.846	NA	0.963	0.964	0.92	0.875	0.89	0.969	0.845	0.924	0.795	0.925	0.908	0.86	0.95	0.903	0.925
NN (80)	0.888	NA	NA	0.975	0.942	0.907	0.898	0.982	0.86	0.968	0.816	0.948	0.957	0.908	0.975	0.919	0.954

Table 1. AUCs of disease status prediction - WSS-DL (n = 20, 40,60,80), PheCAP, WSS-DL (n = 20, 40,60,80), silver standard labels, raw ICD-9 codes, DAPS (n = 20, 40,60,80), and NN (n = 20, 40,60,80)

Method	Asthma	PARDS	BrCa	CAD	CD	COPD	Depression	Epilepsy	HTN	MS	pGout	RA	SCZ	Stroke	T1DM	T2DM	UC
WSS-DL (20)	0.874	0.837	0.973	0.968	0.859	0.617	0.879	0.932	0.908	0.886	0.436	0.904	0.518	0.784	0.727	0.825	0.908
WSS-DL (40)	0.836	NA	0.97	0.968	0.857	0.701	0.864	0.934	0.865	0.909	0.512	0.891	0.743	0.826	0.718	0.84	0.911
WSS-DL (60)	0.851	NA	0.981	0.972	0.91	0.771	0.87	0.958	0.884	0.936	0.55	0.903	0.854	0.863	0.792	0.86	0.921
WSS-DL (80)	0.897	NA	NA	0.976	0.932	0.84	0.881	0.964	0.896	0.972	0.578	0.922	0.897	0.898	0.849	0.875	0.937
ICD	0.864	0.594	0.986	0.92	0.89	0.726	0.852	0.907	0.854	0.924	0.294	0.856	0.433	0.736	0.68	0.808	0.902
Silver std.	0.914	0.794	0.94	0.8	0.948	0.735	0.873	0.88	0.972	0.843	0.785	0.857	0.826	0.616	0.456	0.787	0.87
PheCAP (20)	0.804	0.504	0.867	0.875	0.892	0.571	0.737	0.882	0.858	0.889	0.252	0.825	0.425	0.746	0.789	0.817	0.867
PheCAP (40)	0.849	NA	0.869	0.901	0.925	0.61	0.804	0.901	0.864	0.914	0.257	0.872	0.454	0.773	0.841	0.836	0.886
PheCAP (60)	0.874	NA	0.908	0.945	0.939	0.625	0.851	0.912	0.872	0.927	0.269	0.894	0.461	0.79	0.857	0.847	0.9
PheCAP (80)	0.876	NA	NA	0.957	0.938	0.643	0.872	0.921	0.878	0.929	0.271	0.907	0.465	0.796	0.879	0.853	0.906
Xpress	0.902	0.621	0.946	0.966	0.976	0.548	0.851	0.914	0.864	0.845	0.609	0.898	0.484	0.547	0.659	0.882	0.904
DAPS (20)	0.683	0.539	0.725	0.558	0.606	0.418	0.633	0.824	0.793	0.651	0.289	0.473	0.392	0.587	0.43	0.478	0.518
DAPS (40)	0.698	NA	0.731	0.587	0.606	0.389	0.62	0.854	0.794	0.661	0.299	0.5	0.386	0.586	0.417	0.502	0.519
DAPS (60)	0.701	NA	0.723	0.604	0.609	0.395	0.627	0.867	0.792	0.661	0.304	0.484	0.375	0.574	0.456	0.521	0.52
DAPS (80)	0.708	NA	NA	0.587	0.598	0.395	0.623	0.86	0.793	0.68	0.308	0.49	0.365	0.595	0.451	0.535	0.52
NN (20)	0.778	0.879	0.942	0.924	0.848	0.666	0.853	0.914	0.844	0.763	0.443	0.79	0.59	0.696	0.607	0.816	0.786
NN (40)	0.804	NA	0.965	0.929	0.858	0.715	0.857	0.931	0.854	0.83	0.518	0.849	0.741	0.779	0.699	0.844	0.823
NN (60)	0.838	NA	0.979	0.93	0.91	0.779	0.86	0.961	0.868	0.905	0.555	0.892	0.869	0.845	0.779	0.86	0.872
NN (80)	0.888	NA	NA	0.954	0.932	0.856	0.882	0.968	0.886	0.961	0.586	0.913	0.952	0.911	0.848	0.877	0.918

Table 2. PPVs of disease status prediction - WSS-DL (n = 20, 40,60,80), PheCAP, WSS-DL (n = 20, 40,60,80), silver standard labels, raw ICD-9 codes, DAPS (n = 20, 40,60,80), and NN (n = 20, 40,60,80)

Method	Asthma	PARDS	BrCa	CAD	CD	COPD	Depression	Epilepsy	HTN	MS	pGout	RA	SCZ	Stroke	T1DM	T2DM	UC
WSS-DL (20)	0.843	0.917	0.892	0.971	0.845	0.905	0.851	0.93	0.79	0.924	0.873	0.925	0.967	0.907	0.957	0.897	0.934
WSS-DL (40)	0.886	NA	0.894	0.968	0.84	0.887	0.862	0.929	0.818	0.907	0.894	0.939	0.965	0.889	0.968	0.899	0.937
WSS-DL (60)	0.889	NA	0.923	0.979	0.853	0.904	0.883	0.942	0.792	0.924	0.901	0.946	0.973	0.906	0.973	0.904	0.95
WSS-DL (80)	0.904	NA	NA	0.98	0.878	0.911	0.881	0.956	0.793	0.937	0.905	0.954	0.982	0.931	0.984	0.911	0.962
ICD	0.766	0.949	0.91	0.981	0.886	0.878	0.815	0.915	0.696	0.903	0.858	0.903	0.975	0.904	0.935	0.893	0.922
Silver std.	0.911	0.971	0.862	0.863	0.879	0.854	0.861	0.851	0.65	0.936	0.927	0.915	0.982	0.904	0.956	0.888	0.937
PheCAP (20)	0.846	0.762	0.926	0.935	0.855	0.87	0.807	0.914	0.717	0.881	0.895	0.895	0.971	0.843	0.974	0.893	0.904
PheCAP (40)	0.9	NA	0.945	0.944	0.865	0.897	0.813	0.923	0.753	0.896	0.862	0.893	0.972	0.879	0.976	0.904	0.916
PheCAP (60)	0.907	NA	0.944	0.958	0.865	0.915	0.805	0.928	0.711	0.893	0.856	0.897	0.97	0.873	0.979	0.918	0.921
PheCAP (80)	0.909	NA	NA	0.966	0.871	0.91	0.798	0.929	0.758	0.888	0.858	0.905	0.973	0.873	0.981	0.931	0.923
Xpress	0.903	0.893	0.925	0.969	0.874	0.954	0.891	0.925	0.857	0.864	0.884	0.915	0.968	0.862	0.949	0.934	0.944
DAPS (20)	0.82	0.859	0.852	0.856	0.747	0.82	0.775	0.845	0.795	0.861	0.862	0.789	0.944	0.799	0.927	0.842	0.745
DAPS (40)	0.819	NA	0.804	0.861	0.734	0.801	0.778	0.837	0.734	0.881	0.866	0.771	0.946	0.788	0.934	0.845	0.749
DAPS (60)	0.841	NA	0.884	0.857	0.72	0.792	0.754	0.833	0.765	0.867	0.872	0.754	0.953	0.814	0.934	0.841	0.717
DAPS (80)	0.817	NA	NA	0.869	0.687	0.792	0.774	0.831	0.798	0.877	0.871	0.748	0.955	0.81	0.934	0.844	0.693
NN (20)	0.868	0.906	0.887	0.944	0.801	0.883	0.863	0.882	0.79	0.831	0.891	0.892	0.946	0.84	0.956	0.876	0.862
NN (40)	0.91	NA	0.904	0.949	0.838	0.889	0.868	0.907	0.824	0.888	0.894	0.903	0.955	0.852	0.967	0.886	0.893
NN (60)	0.905	NA	0.929	0.958	0.853	0.904	0.884	0.925	0.786	0.903	0.904	0.911	0.966	0.88	0.973	0.893	0.92
NN (80)	0.904	NA	NA	0.962	0.878	0.91	0.88	0.954	0.785	0.942	0.905	0.932	0.98	0.898	0.982	0.902	0.939

Table 3. NPVs of disease status prediction - WSS-DL (n = 20, 40,60,80), PheCAP, WSS-DL (n = 20, 40,60,80), silver standard labels, raw ICD-9 codes, DAPS (n = 20, 40,60,80), and NN (n = 20, 40,60,80)

Method	Asthma	PARDS	BrCa	CAD	CD	COPD	Depression	Epilepsy	HTN	MS	pGout	RA	SCZ	Stroke	T1DM	T2DM	UC
WSS-DL (20)	0.878	0.842	0.916	0.963	0.863	0.701	0.874	0.939	0.927	0.906	0.557	0.889	0.596	0.823	0.724	0.826	0.917
WSS-DL (40)	0.875	NA	0.939	0.96	0.866	0.73	0.872	0.94	0.911	0.909	0.564	0.894	0.73	0.836	0.757	0.835	0.92
WSS-DL (60)	0.884	NA	0.966	0.97	0.893	0.782	0.883	0.956	0.915	0.931	0.595	0.906	0.822	0.863	0.813	0.849	0.933
WSS-DL (80)	0.907	NA	NA	0.974	0.918	0.823	0.889	0.965	0.918	0.955	0.615	0.923	0.88	0.899	0.873	0.862	0.947
ICD	0.849	0.695	0.947	0.946	0.905	0.722	0.848	0.922	0.898	0.919	0.604	0.847	0.556	0.786	0.625	0.817	0.908
Silver std.	0.926	0.866	0.952	0.788	0.926	0.695	0.882	0.898	0.914	0.889	0.528	0.858	0.844	0.723	0.573	0.805	0.9
PheCAP (20)	0.86	0.597	0.894	0.882	0.891	0.657	0.799	0.907	0.905	0.889	0.379	0.822	0.542	0.776	0.798	0.816	0.878
PheCAP (40)	0.889	NA	0.905	0.908	0.91	0.705	0.826	0.922	0.909	0.907	0.383	0.844	0.555	0.797	0.85	0.836	0.896
PheCAP (60)	0.906	NA	0.92	0.939	0.916	0.718	0.837	0.928	0.911	0.911	0.389	0.857	0.561	0.798	0.869	0.855	0.906
PheCAP (80)	0.908	NA	NA	0.952	0.919	0.725	0.842	0.933	0.914	0.911	0.392	0.871	0.562	0.803	0.884	0.87	0.91
Xpress	0.917	0.711	0.957	0.96	0.935	0.69	0.883	0.93	0.917	0.86	0.566	0.878	0.575	0.674	0.659	0.888	0.92
DAPS (20)	0.783	0.634	0.826	0.663	0.745	0.533	0.745	0.853	0.881	0.764	0.406	0.591	0.44	0.671	0.43	0.611	0.667
DAPS (40)	0.793	NA	0.827	0.675	0.746	0.526	0.742	0.862	0.881	0.774	0.413	0.597	0.449	0.669	0.435	0.627	0.667
DAPS (60)	0.798	NA	0.826	0.679	0.745	0.527	0.741	0.865	0.88	0.778	0.418	0.59	0.451	0.669	0.445	0.636	0.666
DAPS (80)	0.798	NA	NA	0.676	0.741	0.522	0.74	0.866	0.881	0.785	0.416	0.587	0.448	0.673	0.444	0.644	0.666
NN (20)	0.839	0.846	0.924	0.918	0.841	0.708	0.869	0.911	0.898	0.807	0.521	0.802	0.574	0.733	0.652	0.804	0.815
NN (40)	0.864	NA	0.954	0.924	0.866	0.735	0.872	0.929	0.905	0.861	0.566	0.84	0.688	0.78	0.739	0.826	0.851
NN (60)	0.88	NA	0.971	0.933	0.893	0.786	0.879	0.951	0.906	0.907	0.6	0.867	0.809	0.828	0.802	0.839	0.89
NN (80)	0.902	NA	NA	0.948	0.918	0.831	0.89	0.966	0.913	0.953	0.617	0.898	0.9	0.875	0.87	0.855	0.924

Table 4. F-scores of disease status prediction - WSS-DL (n = 20, 40,60,80), PheCAP, WSS-DL (n = 20, 40,60,80), silver standard labels, raw ICD-9 codes, DAPS (n = 20, 40,60,80), and NN (n = 20, 40,60,80)