Appendix A. Tables

Table S1 Statistics of Corpus BiolarkGSC+ and COPD-HPO

Within each sample	Metric	BiolarkGSC+	COPD-HPO
Sentences	Min	1	1
	Max	23	30
	Average	7.2	5.3
	Standard deviation	3.7	3.6
Tokens	Min	19	1
	Max	359	687
	Average	148.9	122.7
	Standard deviation	70.9	77.8
Labels	Min	1	0
	Max	54	16
	Average	12.2	2.3
	Standard deviation	8.8	2.0

Table S2 Parameter settings of the re-ranking model, including the pre-trained language model BlueBERT and the other components

BlueBERT-related parameters		Other parameters	
Hidden size	768	Epochs	50
Number of hidden layers	12	Early stop delta	0.001
Number of attention heads	12	Early stop patience	10
Dropout probability	0.1	Learning rate	0.0002
Max length of text	512	Batch size	20
Vocabulary size	30,522		
Hidden activation	GELU		
Pooler	Output of ‘[CLS]’		

Table S3 Performance comparison on BiolarkGSC+ and COPD-HPO testing sets

	BiolarkGSC+			COPD-HPO		
Method/Metric	Precision	Recall	F1-score	Precision	Recall	F1-score
OBO Anotator [9]	0.780	0.428	0.553	0.212	0.154	0.178
NCBO [10]	0.553	0.396	0.462	0.502	0.647	0.603
MonarchInitiative [16]	0.536	**0.464**	0.498	0.472	0.642	0.589
Doc2hpo-Ensemble [15]	0.504	0.306	0.381	0.663	0.567	0.589
MetaMap [12]	0.531	0.440	0.481	0.478	**0.731**	0.646
Clinphen [11]	0.491	0.247	0.328	0.248	0.173	0.187
NeuralCR [14]	0.602	0.347	0.441	0.486	0.594	0.563
TrackHealth	0.592	0.313	0.409	0.587	0.486	0.532
PhenoTagger [17]	0.623	0.460	0.529	0.562	0.712	0.628
MMRerank	0.798	0.404	0.537	**0.710**	0.647	0.677
MNIRerank	**0.834**	0.444	**0.579**	0.667	0.594	0.629
PTRerank	0.716	0.432	0.539	0.706	0.703	**0.704**

Table S4 Performance comparison on BiolarkGSC+ and COPD-HPO without post-processing

	BiolarkGSC+			COPD-HPO		
Method/Metric	P	R	F1	P	R	F1
OBO Anotator	0.697	0.393	0.503	0.224	0.171	0.194
NCBO	0.549	0.474	0.509	0.493	0.650	0.560
MonarchInitiative	0.547	0.542	0.545	0.447	0.638	0.526
Doc2hpo-Ensemble	0.666	0.424	0.518	0.639	0.557	0.595
MetaMap	0.592	0.539	0.564	0.461	**0.707**	0.558
Clinphen	0.495	0.286	0.363	0.236	0.171	0.198
NeuralCR	0.651	0.440	0.525	0.470	0.586	0.522
TrackHealth	0.661	0.409	0.505	0.584	0.485	0.530
PhenoTagger	0.694	**0.734**	0.714	0.554	0.638	0.593
MMRerank	0.844	0.445	0.583	0.672	0.693	**0.682**
MNIRerank	**0.852**	0.515	0.642	0.667	0.594	0.629
PTRerank	0.808	0.673	**0.734**	**0.682**	0.616	0.647

Table S5 Performance comparison on a subset aligned with our previous study

	BiolarkGSC+ Subset

Method/Metric	P	R	F1
OBO Anotator	0.809	0.565	0.665
NCBO	0.468	0.485	0.476
MonarchInitiative	0.757	0.605	0.673
Doc2hpo-Ensemble	0.768	0.618	0.685
MetaMap	0.516	0.568	0.541
Clinphen	0.597	0.416	0.490
NeuralCR	0.741	0.602	0.664
TrackHealth	0.753	0.583	0.658
PhenoTagger	0.774	**0.740**	0.757
MMRerank	0.648	0.525	0.580
MNIRerank	0.785	0.557	0.652
PTRerank	**0.813**	0.710	**0.758**

Table S6 Calibrated precision on BiolarkGSC+ and COPD-HPO

	BiolarkGSC+			COPD-HPO		
Method/Metric	Precision	Error Rate	Calibration	Precision	Error Rate	Calibration
OBO Anotator	0.810	8.84%	0.824	0.318	4.21%	0.338
NCBO	0.777	5.97%	0.785	0.756	5.36%	0.764
MonarchInitiative	0.751	6.03%	0.763	0.741	4.97%	0.750
Doc2hpo-Ensemble	0.754	7.32%	0.767	0.779	6.68%	0.749
MetaMap	0.707	6.12%	0.723	0.640	7.28%	0.666
Clinphen	0.590	7.95%	0.603	0.377	6.52%	0.394
NeuralCR	0.736	8.37%	0.758	0.543	8.37%	0.580
TrackHealth	0.757	7.12%	0.772	0.719	7.12%	0.734
PhenoTagger	0.720	8.94%	0.783	0.623	6.45%	0.583
MMRerank	0.754	5.74%	0.811	0.822	5.74%	0.830
MNIRerank	0.789	5.89%	0.845	0.802	5.89%	0.811
PTRerank	**0.843**	5.56%	**0.849**	**0.836**	6.02%	**0.854**