SUPPLEMENTARY METHODS

Subject Selection

All patients seen at Boston Children's Hospital are eligible for enrollment in the PrecisionLink Biobank. Subjects were enrolled in both inpatient and ambulatory settings, including the preoperative clinic, emergency department, and intensive care units. Subjects were enrolled beginning in December 2015. For enrolled subjects, EHR data available in the local Informatics for Integrating Biology and the Bedside (i2b2) database,[1, 2] a research "sidecar" to the EHR, become part of the Biobank phenotyping data. All available historical data for enrolled subjects are included, and data are refreshed monthly.

MAP Data Preprocessing

International Classification of Diseases (ICD), 9th or 10th revision codes from the i2b2-based PrecisionLink data mart were grouped into "phecodes"[3-5] (available at https://phewascatalog.org/phecodes) to represent the primary diagnostic code(s) for each phenotype (Table 1 in the main text). Phecodes are hierarchical, and we include all codes that are more specific than those listed in Table 1. Each ICD code was only counted once per patient per day or once per patient per encounter for inpatient hospitalizations.

Creation of Custom Dictionaries

The custom dictionary for the clinical text features of MAP was created by matching the concept unique identifiers (CUIs) listed in the Unified Medical Language System (UMLS)[6] for the ICD codes matching the phecode, the CUIs generated from the ICD code strings in UMLS, and the CUIs generated from the phenotype string in the PheWAS catalog. No human input was used to curate the custom dictionaries. An illustrative example is given in Supplementary Figure 1. The primary and secondary CUIs listed in Table 1 are those CUIs for each phecode with the highest and second-highest frequency in the Biobank cohort.

MAP Evaluation

Physician-reviewed labels for 91 subjects with pulmonary hypertension (including both primary and secondary pulmonary vascular disease) were available from a prior study.[7] We randomly selected 20 subjects for each of the other nine phenotypes from among the filter-positive Biobank subjects for each phenotype. Physicians blinded to the MAP-predicted phenotype reviewed all available medical records for each subject and determined whether the subject had the phenotype of interest. These gold-standard labels were used to evaluate the classification performance of MAP and ICD codes or CUIs alone. When evaluating ICD codes or CUIs alone, in order to mimic the method used with MAP to determine thresholds for binary yes/no classification for each phenotype, we classified subjects as phenotype positive if their count of relevant codes was higher than a threshold chosen for each phenotype such that the percentage of phenotype positive subjects was closest to the phenotype prevalence estimated from the gold-standard labels.

SUPPLEMENTARY REFERENCES

1. Gainer VS, Cagan A, Castro VM, et al. The Biobank Portal for Partners Personalized Medicine: A Query Tool for Working with Consented Biobank Samples, Genotypes, and Phenotypes Using i2b2. *J Pers Med* 2016;6(1):e11.
2. Klann JG, Abend A, Raghavan VA, et al. Data interchange using i2b2. *J. Am. Med. Inform. Assoc.* 2016;23:909-15.
3. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 2010;26(9):1205-10.
4. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 2013;31(12):1102-10.
5. Wu P, Gifford A, Meng X, et al. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform* 2019;7(4):e14325.
6. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32(Database issue):D267-70.
7. Geva A, Gronsbell JL, Cai T, et al. A computable phenotype improves cohort ascertainment in a pediatric pulmonary hypertension registry. *J Pediatr* 2017;188:224-31.e5.

Supplementary Table 1. Complete list of concept unique identifiers (CUIs) included in the custom dictionary for each phenotype.

Phecode	Primary CUI	Secondary CUI	Remaining CUIs[a]
555.1	C0010346	C0156147	C0156146, C0267383, C0678202
555.2	C0009324	C2937222	C0267390, C0267392, C0348737, C0267388, C0267389, C0375359, C0375360
555.	C0010346	C0009324	C0678202, C0156146, C0156147, C0267383, C0267390, C0267392, C0348737, C0267388, C0267389, C2937222, C0375359, C0375360
495.	C0004096	C0038218	C0155877, C0155878, C0155880, C0155881, C1260416, C0694548, C0155886, C0155883, C0375333, C0375334, C1176341, C0155879, C1176339, C0155882, C1176340, C0015263, C1176342
250.1	C0011854	C0375114	C0375116, C0375123, C0375125, C0375127, C0375129, C0375146, C0375148, C0375150, C0375152, C0375118, C0375120, C0375131, C0375133, C0375135, C0375136, C0375138, C0375140, C0375142, C0375144
345.1	C0014544	C0037769	C0017332, C0154707, C1112693, C0270823, C0311335, C0154715, C0154716, C0085543, C0154717, C0154718, C1719409, C0154719, C0154720, C0154722, C0311334, C0154709, C0154710, C1719405, C1306246, C0154713, C1719407, C0154714, C0154712, C0553587, C0234974, C0014547
345.3	C0009951	C0009952	C0751057, C2921125, C0490011
345.	C0014544	C0009951	C1719410, C0154721, C0017332, C0154707, C1112693, C0270823, C0311335, C0037769, C0154715, C0154716, C0085543, C0154717, C0154718, C1719409, C0154719, C0154720, C0154722, C0311334, C0154709, C0154710, C1719405, C1306246, C0154713, C1719407, C0154714, C0154712, C0009952, C0751057, C2921125, C0490011, C0553587, C0234974, C0014547
714.2	C0553662	C0157917	C0409667, C0837691, C0157916, C0157918
425.	C0878544	C0007194	C0553980, C1959600, C0014117, C0340419, C0348615, C0340422, C0155699, C0036529, C0007192, C1739395
428.	C0018802	C0018801	C0155582, C0023212, C1135191, C2215291, C1135194, C2215175, C2882273, C2882274, C2882275, C2882276, C1135196, C2215111, C2074673, C2215174, C0810005
415.2	C0152171	C0238074	C0152102, C0856722, C0155673
747.1	C0041207	C0018818	C0158611, C0009995, C0869419, C0158623, C0149530, C0024649, C0477999, C0018798, C0152419, C0158629, C0158606, C0040761, C0013069, C0344616, C0158608, C0039685, C0152424, C2939192, C0014116, C0031192, C0029608, C0152238, C0158609, C0158610, C0158621, C0265830, C0242855, C0162164, C0477996, C0158616, C0013481, C0152417, C0158617, C0158618, C0158619, C0152101, C0034084, C0345010, C0013274, C0003492, C0478000, C0302467, C0009681, C3161124, C0241790, C3161125, C0158632, C0036400, C0158634, C0029520

[a] Unified Medical Language System Release 2012AA

Supplementary Table 2. Registry cohorts excluded from comparison to chart review. Registries with fewer than 200 subjects, those that combined multiple phenotypes (i.e., were not disease- or condition-specific), and those for phenotypes for which a specific phecode does not exist were excluded.

Registry	**Exclusion Reason**
Epilepsy	N = 89
Opsoclonus-myoclonus syndrome	N = 107
Congenital heart disease, cardiomyopathy, and heart failure	Includes multiple phenotypes
Pulmonary hypertension	N = 66
Cerebrospinal fluid	Includes multiple phenotypes
Elevated LDL cholesterol	No specific phecode
Cardiovascular and Critical Care	Includes multiple phenotypes
Immunological studies	Includes multiple phenotypes
Ehlers Danlos syndrome	N = 18
Disorders of Sex Development	Includes multiple phenotypes
Neonatal intensive care unit	Includes multiple phenotypes
Pulmonary	Includes multiple phenotypes
Hearing loss	N = 51
Bronchiectasis	N = 17
Interstitial lung disease	N = 8
Early onset childhood obesity	N = 35

Supplementary Table 3. Demographics of 14,303 subjects enrolled in the PrecisionLink Biobank and 36,800 subjects enrolled in the Partners Biobank.

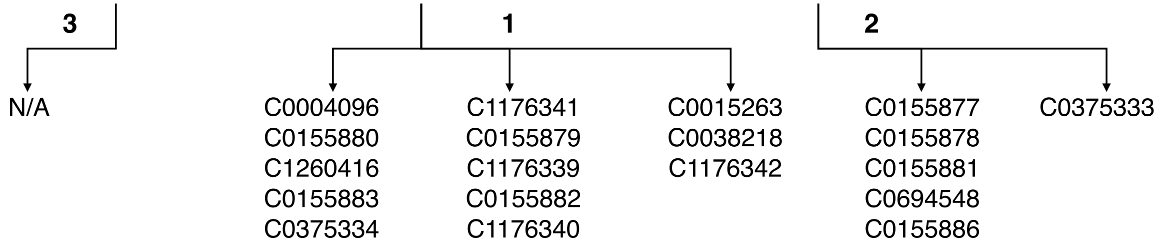
Variable	Frequency (percent) PrecisionLink	Frequency (percent) Partners Biobank
Sex		
Female	7283 (51)	21,413 (58)
Male	7019 (49)	15,387 (42)
Other/Unknown	1 (0)	0 (0)
Age (years)		
< 5	2220 (16)	10 (0)
5 – < 10	2568 (18)	61 (0)
10 – < 20	5702 (40)	140 (0)
20 – < 30	2386 (17)	2856 (8)
30 – < 40	628 (4)	3798 (10)
40 – < 50	336 (2)	4680 (13)
50 – < 60	230 (2)	7348 (20)
≥ 60	233 (2)	17907 (49)
Race		
White	9369 (66)	30854 (84)
Black or African American	826 (6)	2090 (6)
Other/Unknown	4108 (29)	3856 (10)
Ethnicity		
Hispanic or Latino	578 (5)	1601 (4)
Other/Unknown	13725 (95)	35199 (96)

Supplementary Table 4. Comparison of the most frequent diagnoses between Partners and PrecisionLink Biobanks.

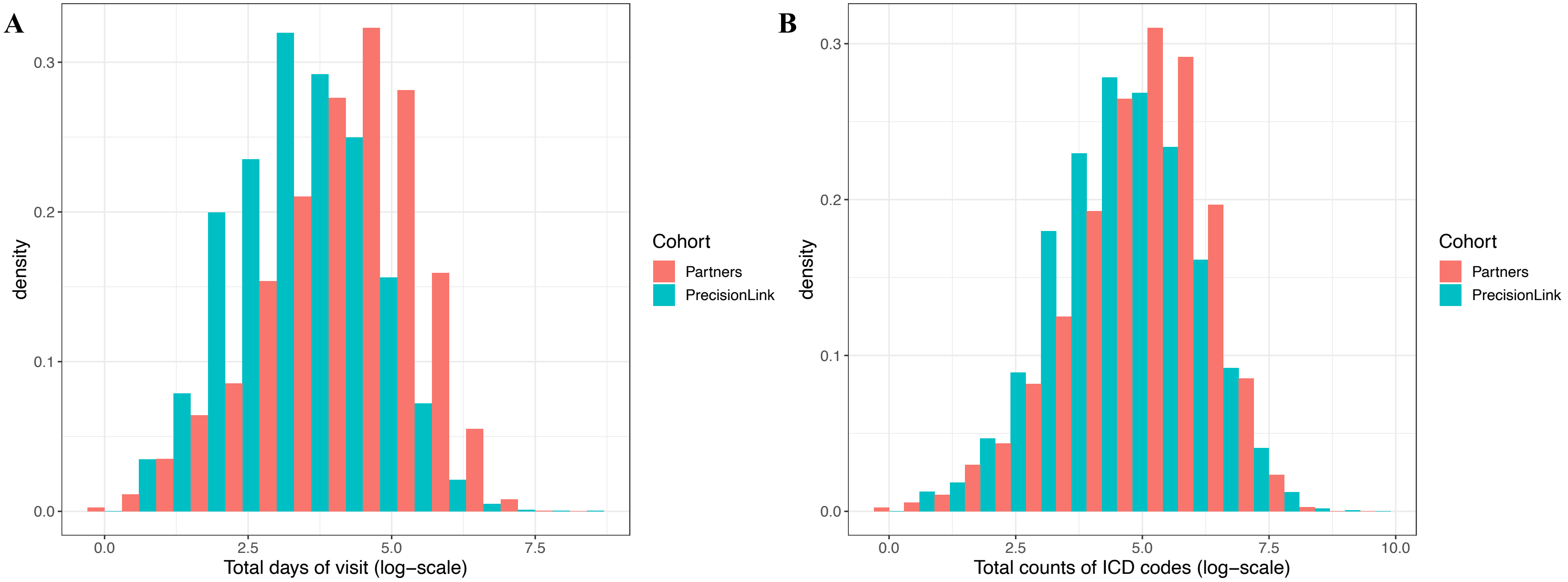
	PrecisionLink			**Partners**	
Phecode	**Description**	**Prevalence (%)**	**Phecode**	**Description**	**Prevalence (%)**
1010.	Other tests	70	1010.	Other tests	55
747.1	Cardiac congenital anomalies	30	745.	Pain in joint	52
785.	Abdominal pain	29	401.1	Essential hypertension	51
395.3	Nonrheumatic tricuspid valve disorders	29	272.1	Hyperlipidemia	49
563.	Constipation	28	773.	Pain in limb	41
1002.	Symptoms concerning nutrition, metabolism, and development	28	785.	Abdominal pain	40
426.7	Abnormal electrocardiogram	24	760.	Back pain	37
530.1	Esophagitis, GERD and related diseases	24	512.8	Cough	35
427.5	Arrhythmia (cardiac) NOS	23	740.9	Osteoarthrosis NOS	33
512.8	Cough	21	530.1	Esophagitis, GERD and related diseases	32

Supplementary Figure 1. Schematic of creation of the custom dictionary of concept unique identifiers (CUIs) for the asthma phenotype. International Classification of Diseases (ICD), 9th revision (ICD-9) codes matching the phenotype of interest are mapped to CUIs in the Unified Medical Language System (UMLS) (1). The string for the ICD-9 code description is also matched to the UMLS, and any additional CUIs are added to the dictionary (2). Finally, the phenotype string matching the phecode is matched to the UMLS and any additional CUIs are added to the dictionary (3). (In this example, all matching CUIs had already been added from steps 1 and 2.)

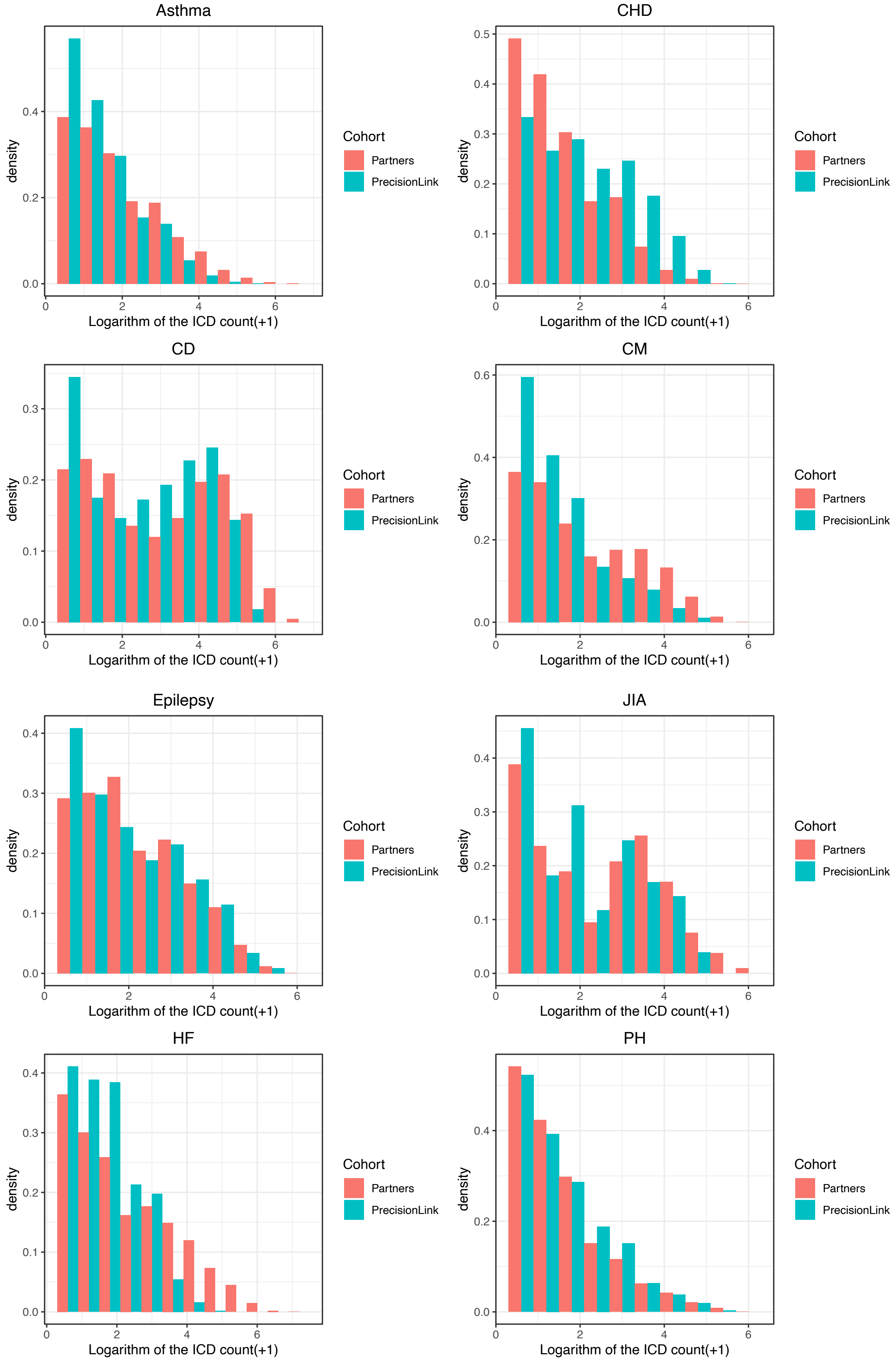
Phecode	Phenotype	ICD-9	ICD9 String
495.	Asthma	493	Asthma
495.	Asthma	493	Extrinsic asthma
495.	Asthma	493	Extrinsic asthma without mention of status asthmaticus
495.	Asthma	493.1	Intrinsic asthma
495.	Asthma	493.1	Intrinsic asthma without mention of status asthmaticus
495.	Asthma	493.8	Other forms of asthma
495.	Asthma	493.82	Cough variant asthma
495.	Asthma	493.9	Asthma, unspecified
495.	Asthma	493.9	Asthma, unspecified type, without mention of status asthmaticus
495.1	Chronic obstructive asthma	493.2	Chronic obstructive asthma
495.1	Chronic obstructive asthma	493.2	Chronic obstructive asthma, without mention of status asthmaticus
495.11	Chronic obstructive asthma with exacerbation	493.21	Chronic obstructive asthma, with status asthmaticus
495.11	Chronic obstructive asthma with exacerbation	493.22	Chronic obstructive asthma, with acute exacerbation
495.2	Asthma with exacerbation	493.01	Extrinsic asthma with status asthmaticus
495.2	Asthma with exacerbation	493.02	Extrinsic asthma with acute exacerbation
495.2	Asthma with exacerbation	493.11	Intrinsic asthma with status asthmaticus
495.2	Asthma with exacerbation	493.12	Intrinsic asthma with acute exacerbation
495.2	Asthma with exacerbation	493.81	Exercise induced bronchospasm
495.2	Asthma with exacerbation	493.91	Asthma, unspecified type, with status asthmaticus
495.2	Asthma with exacerbation	493.92	Asthma, unspecified type, with acute exacerbation



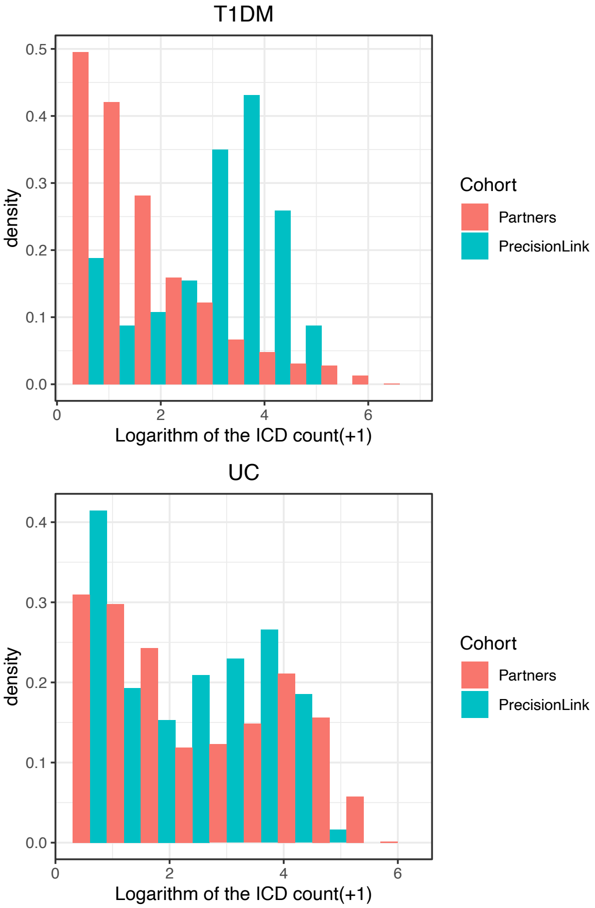
Supplementary Figure 2. Distribution of the per-patient length of stay (A) and counts of International Classification of Diseases (ICD) codes (B) for subjects enrolled in the PrecisionLink and Partners Biobanks.



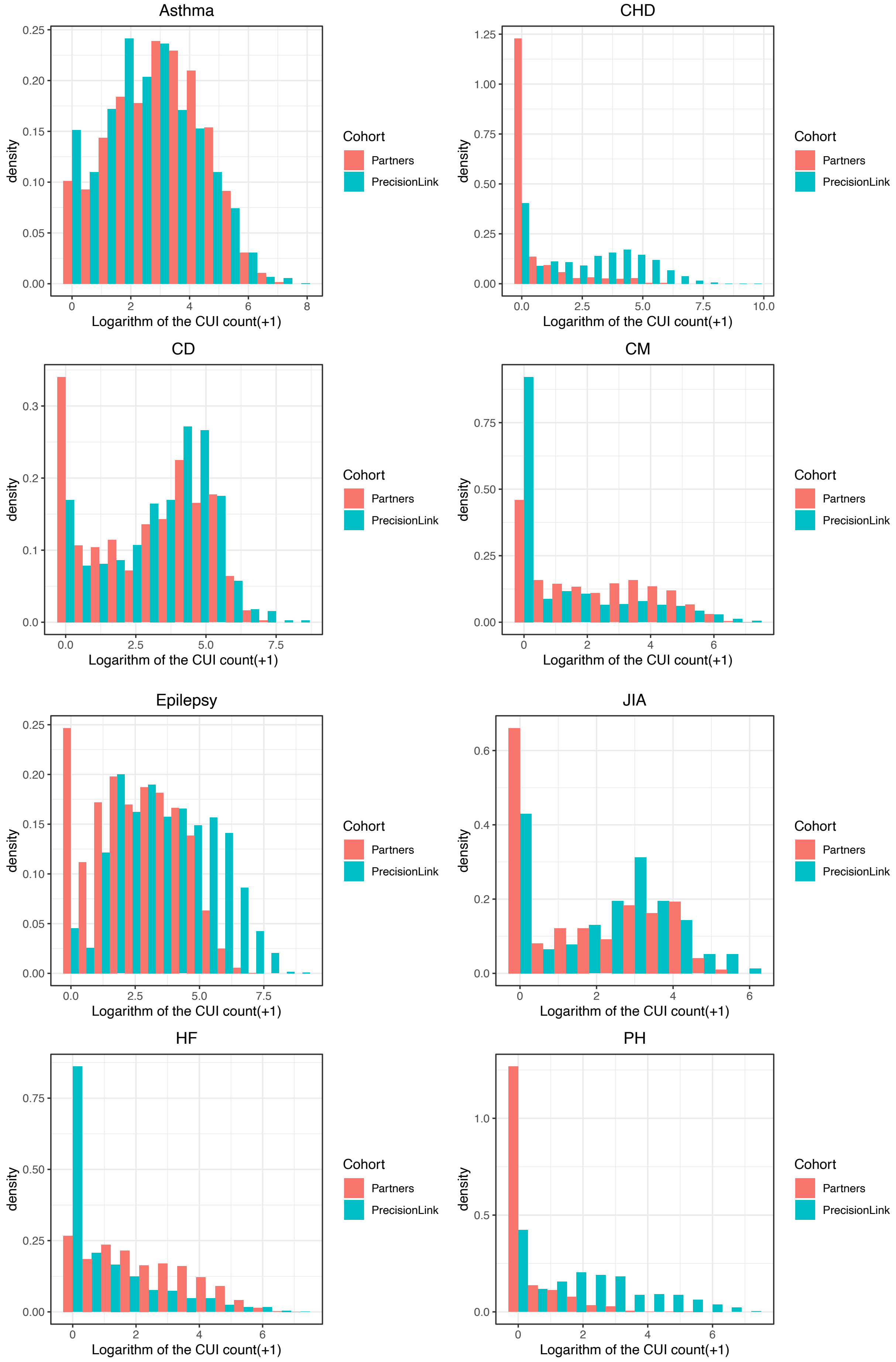
Supplementary Figure 3. Comparison of phenotype-specific counts of relevant International Classification of Diseases (ICD) codes between filter-positive subjects (ICD count ≥ 1) in the PrecisionLink and Partners Biobanks.



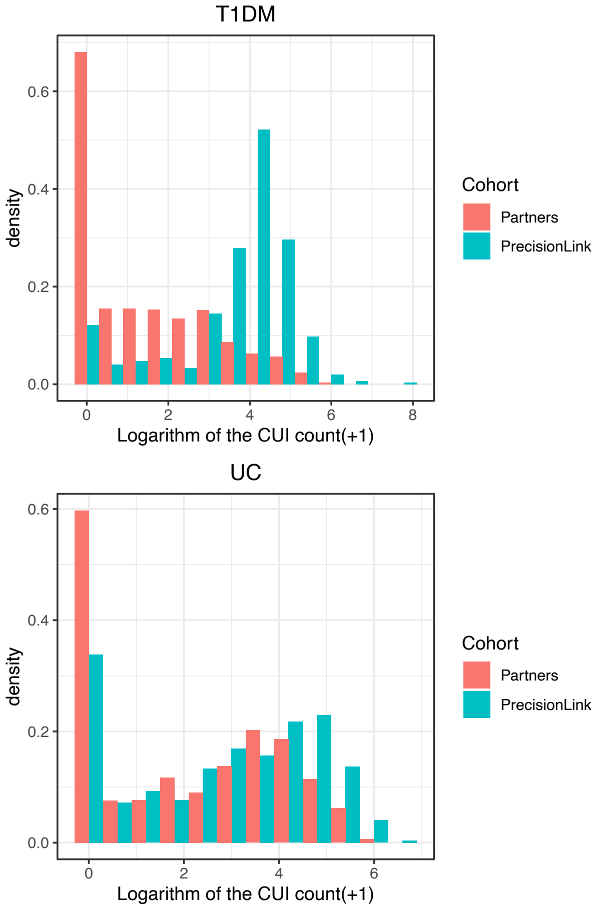
Supplementary Figure 3 (cont.). Comparison of phenotype-specific counts of relevant International Classification of Diseases (ICD) codes between filter-positive subjects (ICD count ≥ 1) in the PrecisionLink and Partners Biobanks.



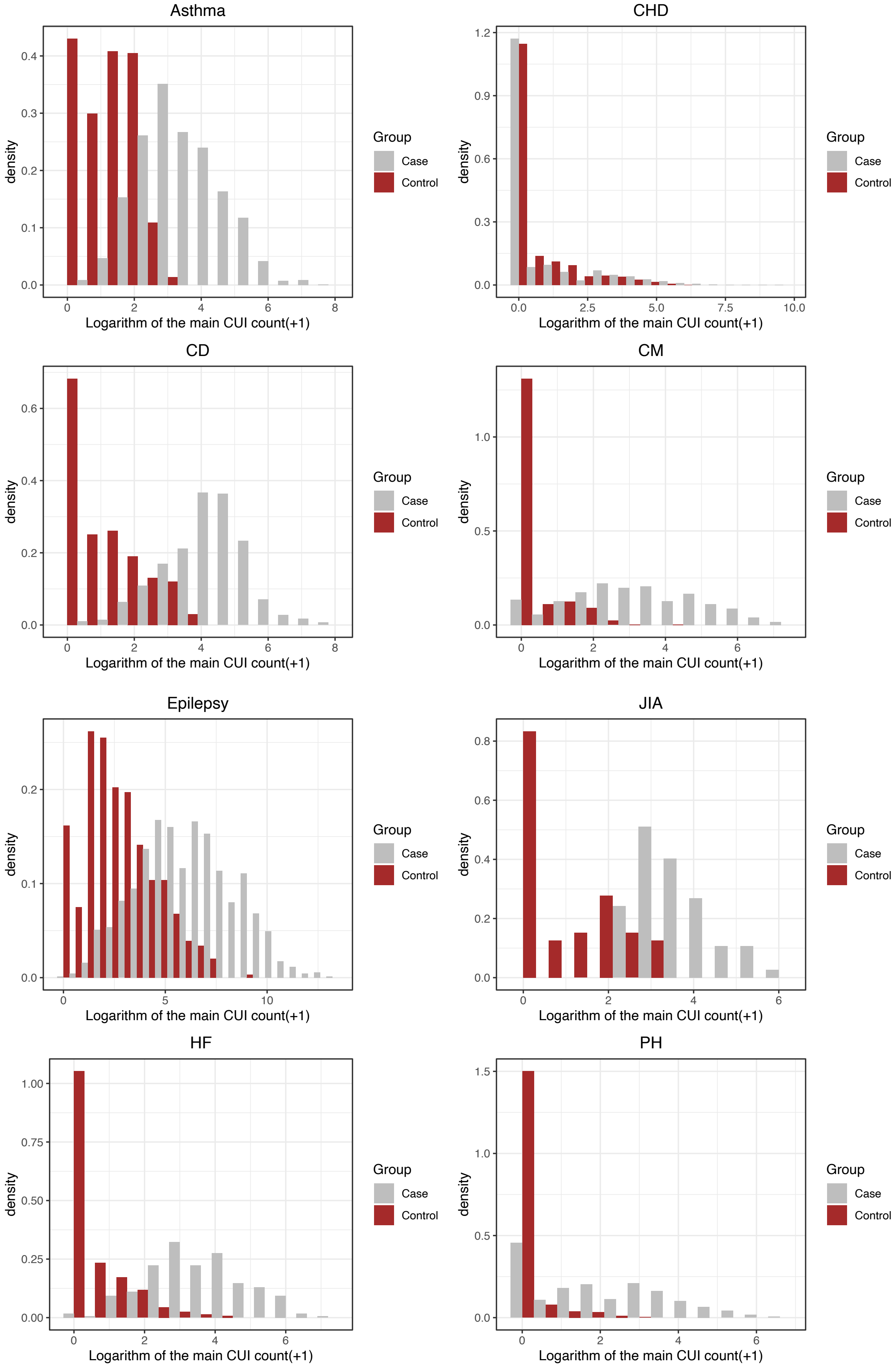
Supplementary Figure 4. Comparison of phenotype-specific counts of relevant concept unique identifier (CUI) codes between subjects in the PrecisionLink and Partners Biobanks.



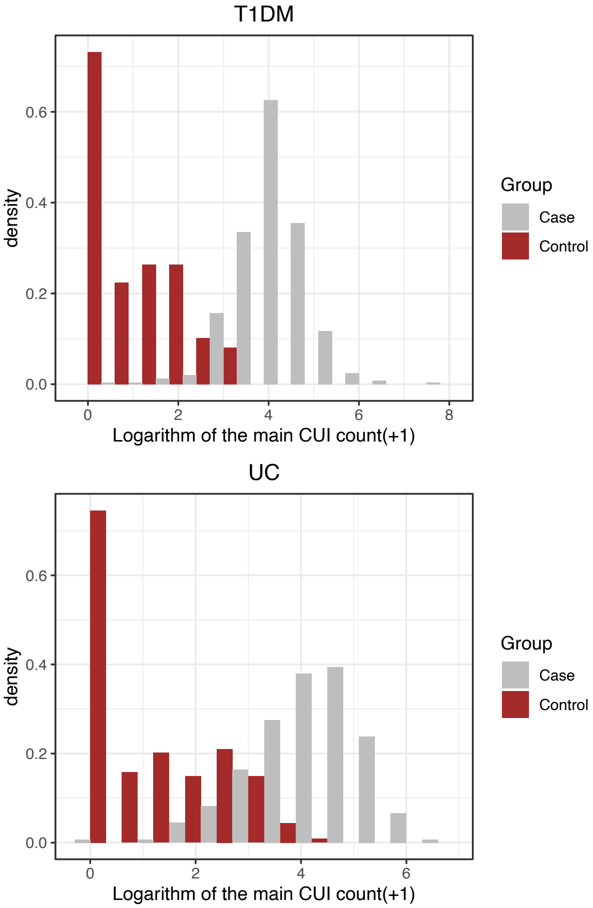
Supplementary Figure 4 (cont.). Comparison of phenotype-specific counts of relevant concept unique identifier (CUI) codes between subjects in the PrecisionLink and Partners Biobanks.



Supplementary Figure 5. Comparison of counts of relevant concept unique identifier (CUI) codes between MAP-predicted positive (cases) and negative (control) subjects for each phenotype.



Supplementary Figure 5 (cont.). Comparison of counts of relevant concept unique identifier (CUI) codes between MAP-predicted positive (cases) and negative (control) subjects for each phenotype.


T1DM
density
Logarithm of the main CUI count(+1)
Group
Case
Control
UC
density
Logarithm of the main CUI count(+1)
Group
Case
Control