Supplemental Figures

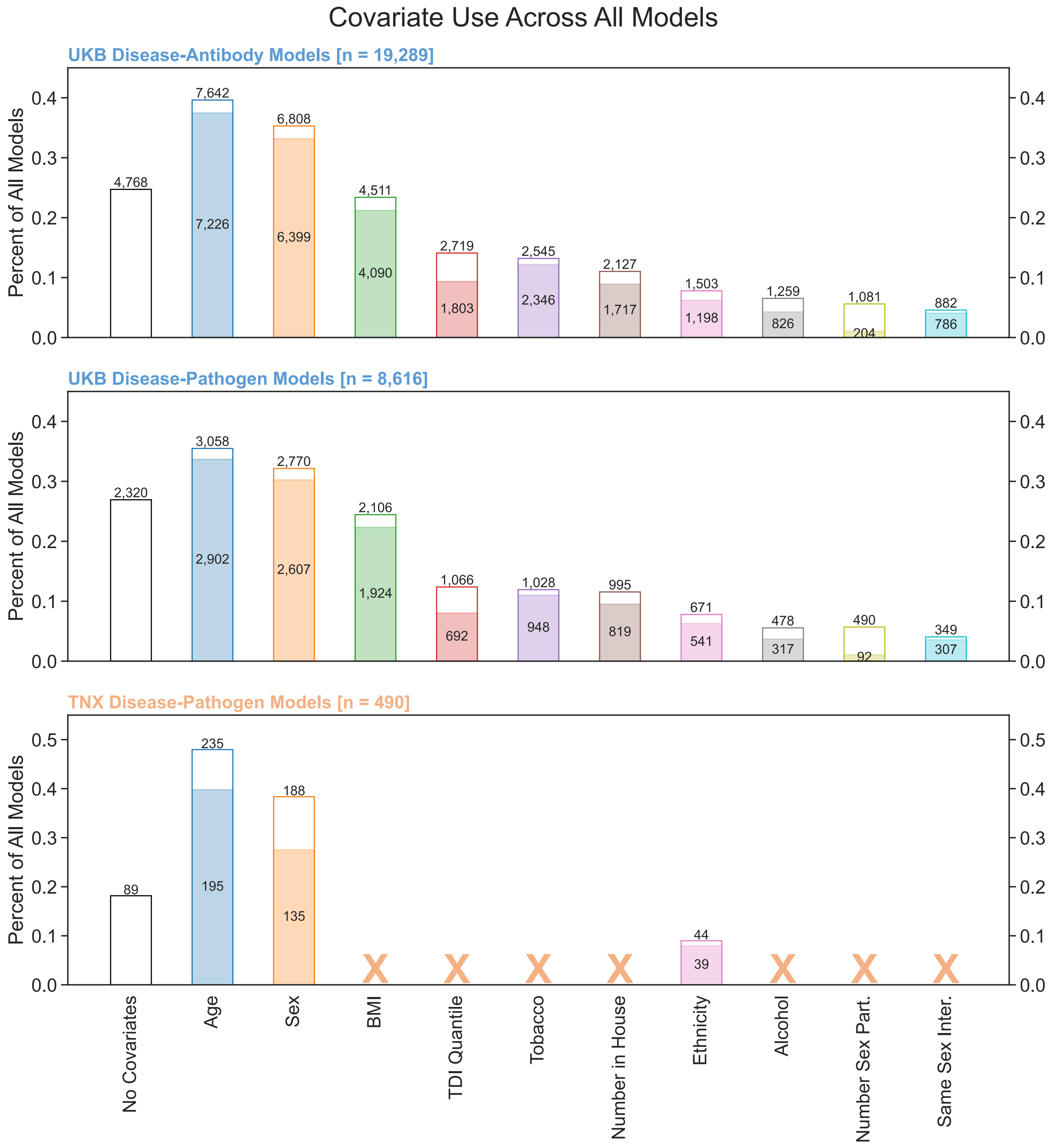


Figure S1 | Covariate use across models.

Bar charts summarizing covariate usage in antibody-disease and pathogen-disease tests across the discovery and replication cohorts. The height of each bar indicates the percentage of all models that were adjusted for the corresponding sociodemographic or health-related covariate. To be included in an antibody-disease model for adjustment, a covariate had to be significantly associated (unadjusted $p < 0.05$) in univariate tests with both the disease status and the titer level separately (Table S2). The filled portion of each bar indicates the percent of the final multivariate logistic regression models in which the covariate remained significant (unadjusted $p < 0.05$). The covariates that were unavailable in the TriNetX (TNX) cohort are marked by an orange “X”. Abbreviations: Part: Partners; Inter: Intercourse; TDI: Townsend deprivation index.

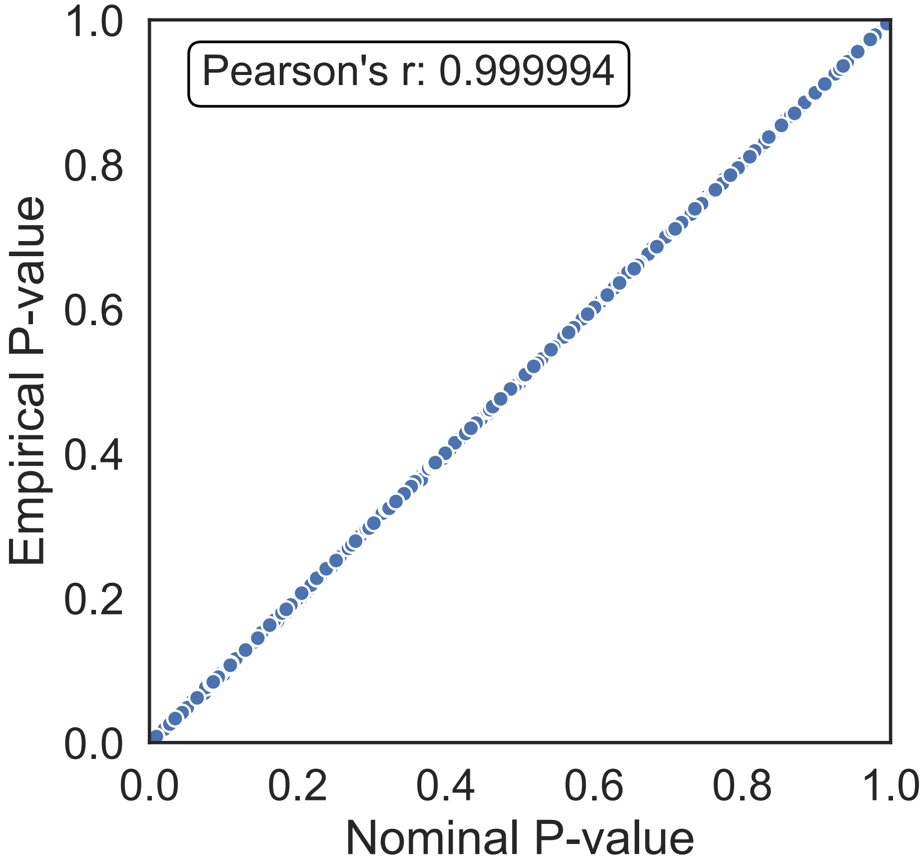


Figure S2 | Nominal versus empirical p-values across all UK Biobank antibody-disease models.

Scatter plot comparing the nominal and empirical p-values for all discovery cohort antibody-disease models. To calculate empirical p-values, 10,000 permutations of each antibody-disease model were performed. All permutations for a particular disease were combined into a per-disease null distribution, yielding 450,000 permutation results per disease (see Methods). The nominal p-value for a specific antibody-disease model was compared to the disease-specific null distribution to calculate the empirical p-value.

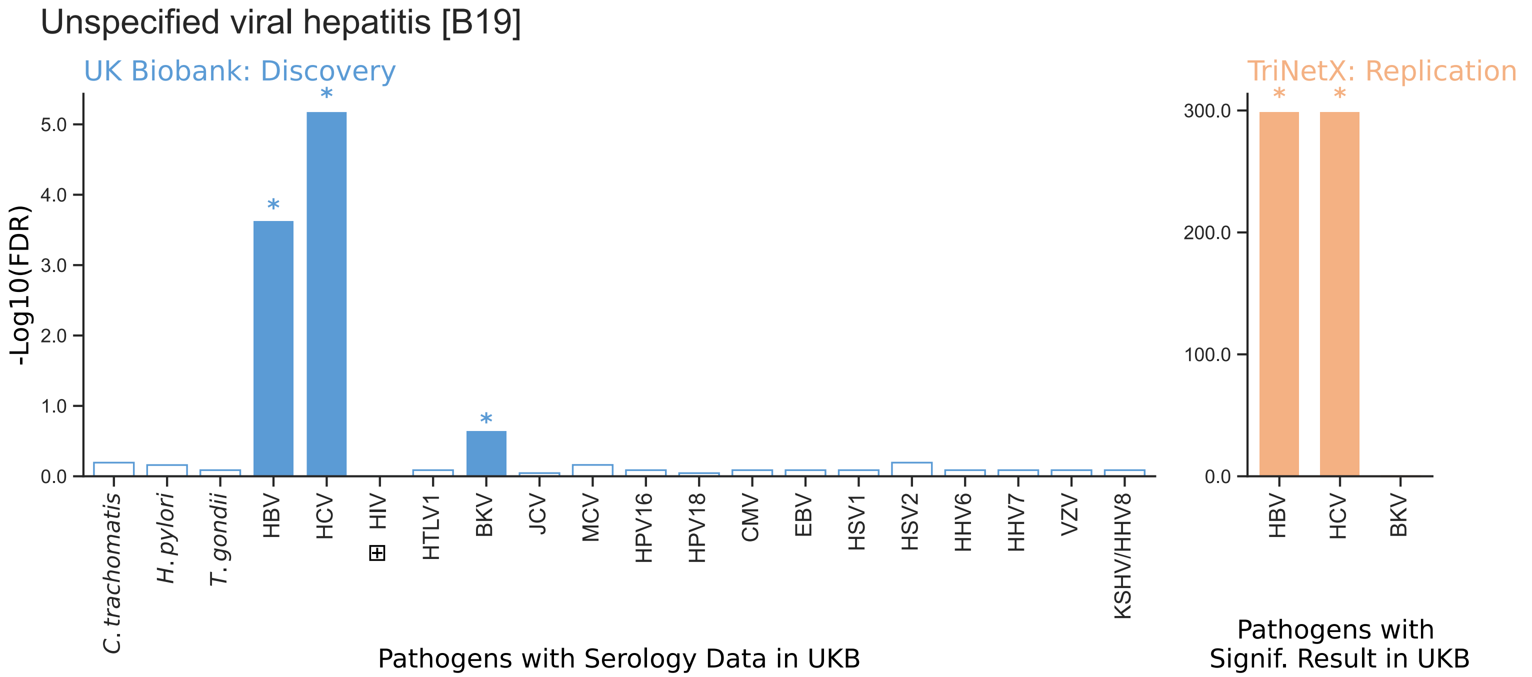


Figure S3 | Illustration of the two-step discovery-replication process.

Bar charts depicting our methods' results for a control disease. The left plot (blue) displays the significance of association (-Log10 transformed per-disease Benjamini-Hochberg false discovery rate (FDR)) for each pathogen with the disease of interest, here the control disease, “unspecified viral hepatitis”. Significant associations are depicted as filled bars with colored asterisks above. Hepatitis B (HBV), hepatitis C (HCV), and BK Virus (BKV) are all significantly associated with “unspecified viral hepatitis” in the UK Biobank (UKB) cohort. The right plot (orange) shows the results of testing only the significant UKB results in the replication cohort, TriNetX. Only HBV and HCV remain significant at the more stringent replication threshold of per-disease FDR < 0.01, leaving the pairs HBV-unspecified viral hepatitis and HCV-unspecified viral hepatitis the only replicated results.

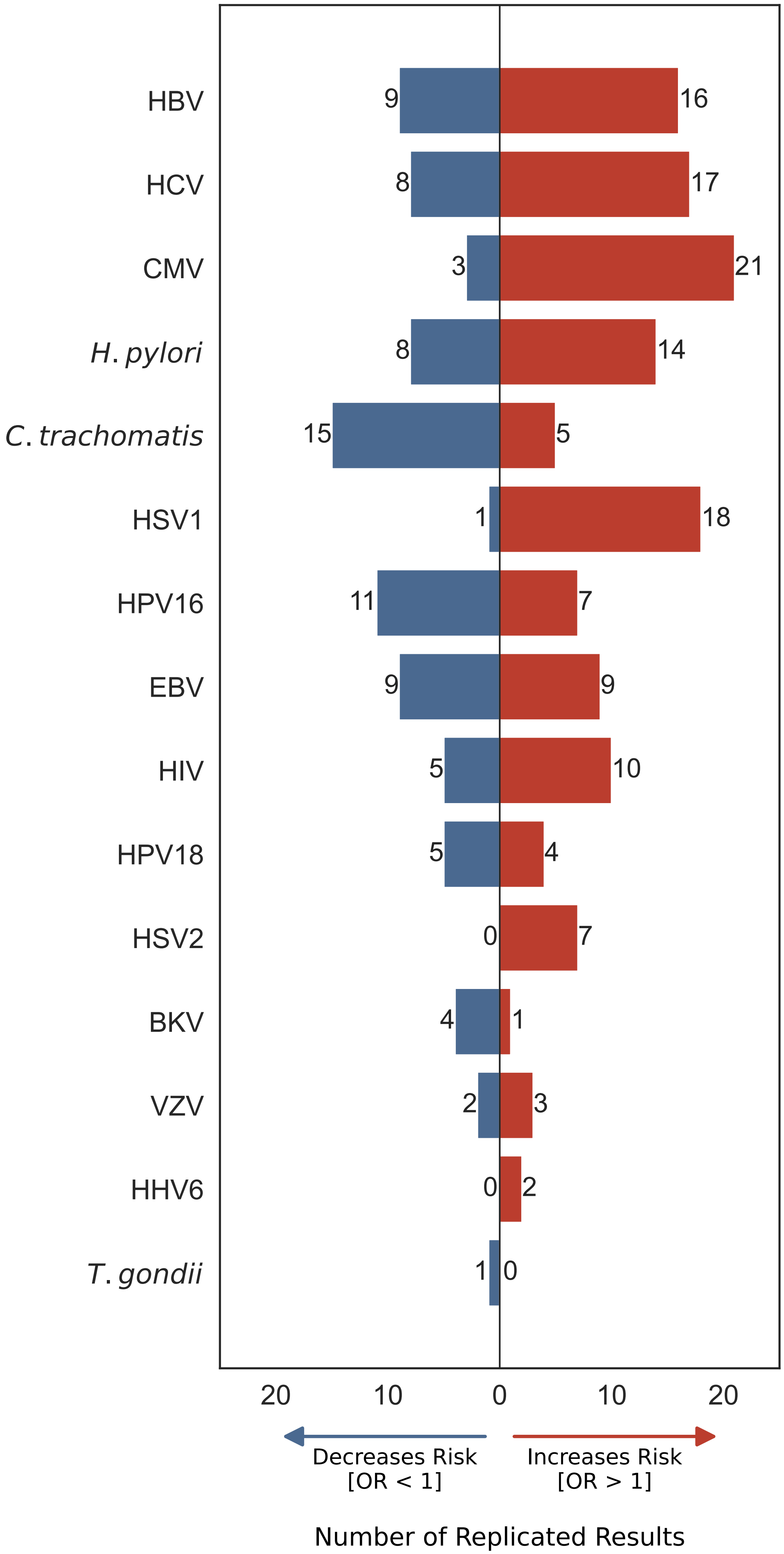
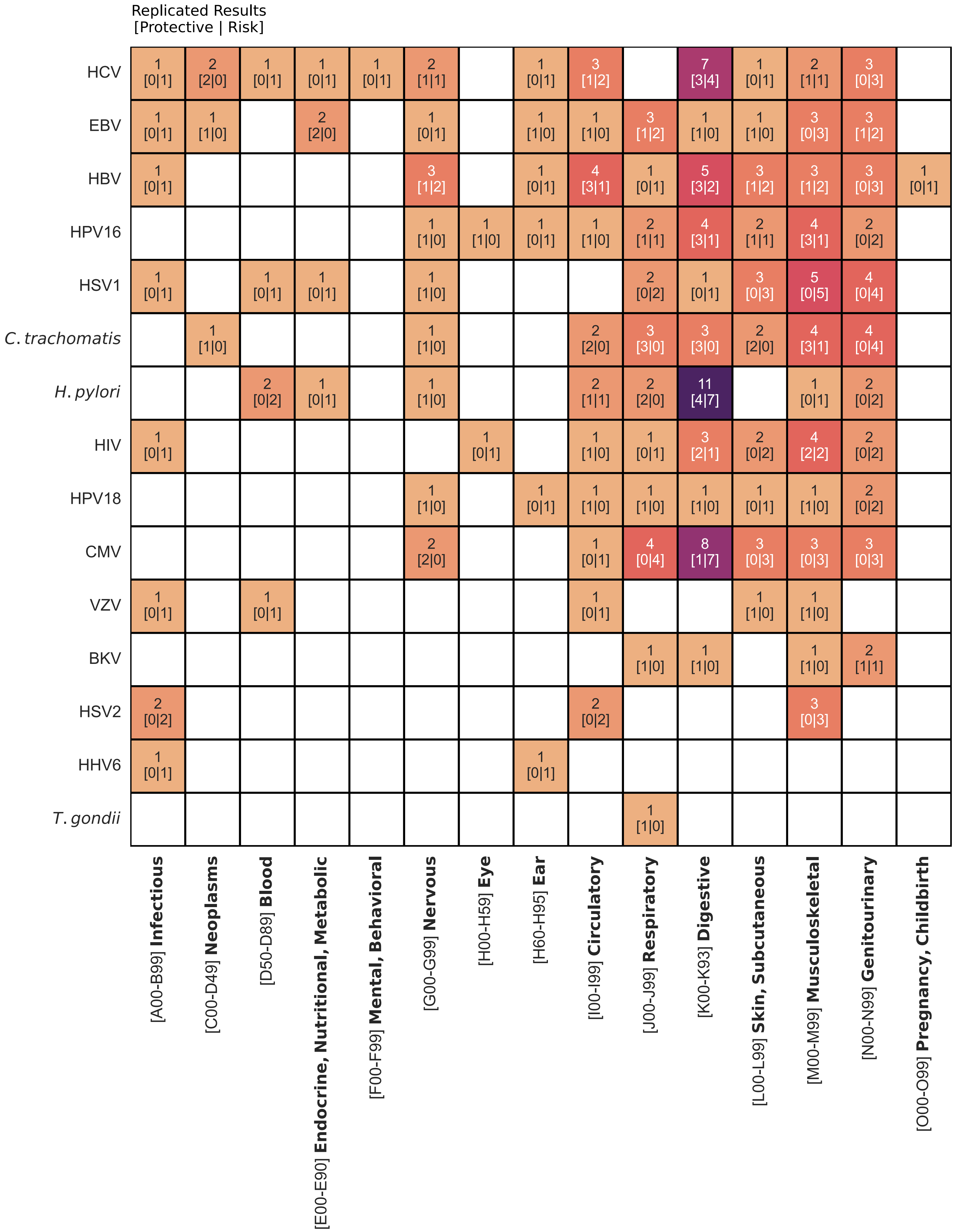

HBV
9
16
HCV
8
17
CMV
3
21
H. pylori
8
14
C. trachomatis
15
5
HSV1
1
18
HPV16
11
7
EBV
9
9
HIV
5
10
HPV18
5
4
HSV2
0
7
BKV
4
1
VZV
2
3
HHV6
0
2
T. gondii
1
0
20
10
0
10
20
Decreases Risk
[OR < 1]
Increases Risk
[OR > 1]
Number of Replicated Results

Figure S4 | Replicated results split by pathogen and effect direction.

A split bar chart showing the number of replicated “risk” and “protective” associations for each pathogen. The total number of replicated results for each pathogen is split by effect direction. Those with a UK Biobank (UKB) and TriNetX (TNX) odds ratio of less than one are represented by the blue bars to the left of the center line. While those with a UKB and TNX odds ratio greater than one are represented by the red bars to the right of the center line.



Replicated Results
[Protective | Risk]

	[A00-B99] **Infectious**	[C00-D49] **Neoplasms**	[D50-D89] **Blood**	[E00-E90] **Endocrine, Nutritional, Metabolic**	[F00-F99] **Mental, Behavioral**	[G00-G99] **Nervous**	[H00-H59] **Eye**	[H60-H95] **Ear**	[I00-I99] **Circulatory**	[J00-J99] **Respiratory**	[K00-K93] **Digestive**	[L00-L99] **Skin, Subcutaneous**	[M00-M99] **Musculoskeletal**	[N00-N99] **Genitourinary**	[O00-O99] **Pregnancy, Childbirth**
HCV	1 [0|1]	2 [2|0]	1 [0|1]	1 [0|1]	1 [0|1]	2 [1|1]		1 [0|1]	3 [1|2]		7 [3|4]	1 [0|1]	2 [1|1]	3 [0|3]	
EBV	1 [0|1]	1 [1|0]		2 [2|0]		1 [0|1]		1 [1|0]	1 [1|0]	3 [1|2]	1 [1|0]	1 [1|0]	3 [0|3]	3 [1|2]	
HBV	1 [0|1]					3 [1|2]		1 [0|1]	4 [3|1]	1 [0|1]	5 [3|2]	3 [1|2]	3 [1|2]	3 [0|3]	1 [0|1]
HPV16						1 [1|0]	1 [1|0]	1 [0|1]	1 [1|0]	2 [1|1]	4 [3|1]	2 [1|1]	4 [3|1]	2 [0|2]	
HSV1	1 [0|1]		1 [0|1]	1 [0|1]		1 [1|0]				2 [0|2]	1 [0|1]	3 [0|3]	5 [0|5]	4 [0|4]	
C. trachomatis		1 [1|0]				1 [1|0]			2 [2|0]	3 [3|0]	3 [3|0]	2 [2|0]	4 [3|1]	4 [0|4]	
H. pylori			2 [0|2]	1 [0|1]		1 [1|0]			2 [1|1]	2 [2|0]	11 [4|7]		1 [0|1]	2 [0|2]	
HIV	1 [0|1]						1 [0|1]		1 [1|0]	1 [0|1]	3 [2|1]	2 [0|2]	4 [2|2]	2 [0|2]	
HPV18						1 [1|0]		1 [0|1]	1 [1|0]	1 [1|0]	1 [1|0]	1 [0|1]	1 [1|0]	2 [0|2]	
CMV						2 [2|0]			1 [0|1]	4 [0|4]	8 [1|7]	3 [0|3]	3 [0|3]	3 [0|3]	
VZV	1 [0|1]		1 [0|1]						1 [0|1]			1 [1|0]	1 [1|0]		
BKV										1 [1|0]	1 [1|0]		1 [1|0]	2 [1|1]	
HSV2	2 [0|2]								2 [0|2]				3 [0|3]		
HHV6	1 [0|1]							1 [0|1]							
T. gondii										1 [1|0]					

Figure S5 | Heatmap of replicated results at the ICD10 block level.

A heatmap showing the total number of replicated results for each pathogen across all diseases in an International Classification of Diseases 10th revision (ICD10) block, which generally includes only diseases of a particular body system. All cells are annotated with the total number of replicated results above, in square brackets, the number of these replicated results with odds ratios less than one, followed by those with odds ratios greater than one. White squares with no annotation indicate that no replicated associations were found for that pathogen and any tested diseases in that ICD10 block.

Supplemental Tables

			UK Biobank	TriNetX
Sex				
	Female		5,276	7,410,388
	Male		4,153	3,807,258
	Other		0	5,827
	Total		9,429	11,223,473
Age				
Resolution:	Mean		67.5	46.3
UKB: Month	SD		8.1	17.3
TNX: Year	Range		[43.8 - 82.6]	[0 - 91]
Race/Ethnicity				
	White		8,883	4,654,725
	Asian		231	258,886
	Black		140	1,664,492
	Hispanic		-	860,830
	American Indian or Alaska Native		-	36,516
	Native Hawaiian or Other Pacific Islander		-	12,767
	Other		175	3,735,257
	Total		9,429	11,223,473
BMI				
	Mean		27.3	-
	SD		4.8	-
	Range		[16.2 - 61.3]	-
n Imputed			32	-
Townsend Deprivation Index				
	Lowest 1/6		1,572	-
	Middle 2/3		6,285	-
	Highest 1/6		1,572	-
	Total		9,429	0
n Imputed			8	-
Number in House				
	Alone		1,712	-
		2	4,331	-
		3	1,524	-
		4	1,282	-

	>= 5	514	-
	Unknown	66	-
	Total	9,429	0

Tobacco Use			
	Never	5,210	-
	Previous	3,218	-
	Current	949	-
	Unknown	52	-
	Total	9,429	0

Alcohol Use			
	Never	417	-
	Previous	332	-
	Current	8,661	-
	Unknown	19	-
	Total	9,429	

Number of Sex Partners			
	0 partners	88	-
	1 partner	2,265	-
	2-3	1,838	-
	4-5	1,255	-
	> 5	2,343	-
	Unknown	1,640	-
	Total	9,429	0

Same-sex Intercourse			
	No	8,204	-
	Yes	283	-
	Unknown	942	-
	Total	9,429	0

Table S1 | Summary statistics for additional covariates across both cohorts.

This table contains the summary statistics for the ten health-related and sociodemographic variables considered as possible confounders. The mean, standard deviation, and range are presented for continuous variables. For categorical variables, counts for each group are included. Note that the resolution for age differed between cohorts, with the UK Biobank having a resolution of one month, while TriNetX only provided the year of birth.

Covariate	Ab Titer Assoc Test	Disease Status Assoc Test
Sex	t-test	Chi-squared
Age	linear regression	t-test
BMI	linear regression	t-test
Race	ANOVA	Chi-squared
Townsend Deprivation Index	ANOVA	Chi-squared
Number in House	ANOVA	Chi-squared
Tobacco Use	ANOVA	Chi-squared
Alcohol Use	ANOVA	Chi-squared
Number of Sex Partners	ANOVA	Chi-squared
Same-sex Intercourse	ANOVA	Chi-squared

Table S2 | Univariate tests used to determine confounding.

This table shows the type of univariate test performed between each covariate and antibody titer and each covariate and disease status to determine if the covariate was confounding. For a covariate to be considered a confounder, it had to be significantly associated (unadjusted $p < 0.05$) separately with both disease status and antibody titer.