Supplementary Data Legends

All Supplementary Data legends listed here can also be found in the README sheet of each Supplementary Data XLSX file.

Supplementary Data 1:

Summary statistics for additional covariates across both cohorts.

This document contains the summary statistics for each of the ten additional health-related and sociodemographic variables that were considered as possible confounders. For continuous variables, the mean, standard deviation, and range are presented. 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 a resolution of one year.

Supplementary Data 2:

Full pathogen-Phecode results across both the UK Biobank and TriNetX.

This document contains the results from the Phecode analysis, where Phecodes replaced International Classification of Diseases 10th revision (ICD10) codes as the endpoint. The analysis was otherwise identical to our ICD10-based analysis, with association test results for all 18,934 pathogen-Phecode pair tests.

Supplementary Data 3:

Top 175 log product frequency ranked results that were manually reviewed to assemble "Tier 2" positive controls.

This document contains the top 175 most co-cited pathogen-disease pairs as determined by a negated form of the log product frequency (LPF), where the closer the LPF value is to zero, the more commonly the disease and pathogen are found to be co-cited rather than cited individually. After manual review, we limited this list to just pairs with a 'High' level of evidence, leaving us with 83 "Tier 2" positive controls. The column "Evidence" contains the evidence level classification, "Mechanism" indicates whether there is a known mechanism by which the pathogen is involved with the disease, "Notes" contains notes taken by M.L. while reviewing, and the "Reference" column includes information about which references were used in the filling of the previously mentioned columns.

Supplementary Data 4:

Automated PubMed querying and log product frequency calculation results.

This document contains several sheets, including the calculation of the negated form of the log product frequency (LPF) values for all (426 disease x 20 pathogens) 8,520 disease-pathogen pairs (LPF_data sheet). Because the LPF equation (**Figure S1b**) contains a log10 transform, the LPF value equals positive infinity for any disease-pathogen pairs with zero co-citations. All three sheets contain the PubMed queries run ("pair_query" on "LPF_data" sheet, "dis_query" on "disease_info" sheet, and "path_query" on "path_info" sheet) to get the PubMed IDs (PMIDs) containing either the co-citations or individual citations ("pair_PMIDs" column for co-citations and "dis_PMIDs" or "path_PMIDs" columns for citations for each disease or pathogen by itself), and the total count of PMIDs found for a co-citation or individual citations ("pair_count" column for number of co-citations, "dis_count" and "path_count" for number of citations for each disease or pathogen individually).

Supplementary Data 5:

NCBI GEO RNA-seq datasets used for VIRTUS analyses.

This document contains information on the NCBI GEO data sets used for our VIRTUS analyses. The first sheet, "SLE", includes information on the six systemic lupus erythematosus (SLE) case/control RNA-seq data sets performed in blood and B cell subsets used for the Epstein-Barr virus (EBV)-SLE VIRTUS analysis. The second sheet, "UC", shows the same information for the seven ulcerative colitis (UC) case/control RNA-seq data sets examined in the cytomegalovirus (CMV)-UC VIRTUS analysis.

Supplementary Data 6:

GWAS risk loci and RNA-seq data sets used for RELI analyses.

This document contains both the GWAS risk loci downloaded from The GWAS Catalog ("GWAS loci" sheet) and the RNA-seq data sets ("Gene Sets" sheet) that were used as inputs for our RELI analyses. Note that GSE99454, a CMV-infection experiment, used two different cell types (lung fibroblast and retinal pigment epithelium), which were analyzed separately. Also, GWAS loci with "-null" in the name indicate the disease-matched (number of SNPs) list of random independent SNPs sampled from snp151Common. The "UC-exclusive" and "CD-exclusive" variant lists represent the lists of variants where any overlapping GWAS loci between ulcerative colitis (UC) and Crohn's disease (CD) have been removed from each respective list.

Supplementary Data 7:

Overview of diseases outside of ICD10 chapter 1 with an infectious component.

All diseases were tested at the 3-character International Classification of Diseases 10th revision (ICD10) code level. This document contains a review of whether those diseases outside of ICD10 chapter I (“infectious diseases”) are either infectious at the 3-character level, at a sub-code level, or not infectious at all. For example, J11, “influenza, virus not identified” is infectious at the 3-character level, whereas N39, “other disorders of urinary system” is not infectious at the 3-character level but does include the sub-code N39.0, “Urinary tract infection, site not specified”, which is infectious, as well as several other non-infectious sub-codes. We classified each as falling into one of the following groups: "No known infectious agent" (NKIA), where the diagnosis has no infectious component, "Single Known Infectious Agent" (SKIA), which includes codes like J11, where the diagnosis "influenza" is caused by the influenza virus, and finally "Multiple/suspected/non-specific infectious agents" (MSNSIA), for which N39.0 would be an example as numerous infectious agents can cause urinary tract infections, even though N39 at the 3-character level would still be considered NKIA.

Legend	
SKIA	Single Known Infectious Agent
MSNSIA	Multiple/suspected/non-specific infectious agents
NKIA	No Known infectious agent

Final 3-Char Counts	
SKIA	1
MSNSIA	26
NKIA	399

Supplementary Data 8:

Full pathogen-disease (ICD10) results across both the UK Biobank and TriNetX.

This document contains the results from our International Classification of Diseases 10th revision (ICD10) analysis with association test results for all 8,616 pathogen-disease pair tests.

Supplementary Data 9:

Results of VIRTUS analysis of SLE and UC case/control cohorts.

This document contains the VIRTUS analysis results for systemic lupus erythematosus (SLE) and ulcerative colitis (UC). The normalized hit rate (NHR) normalizes the pathogen read counts by the number of human reads in the sample, allowing for the comparison of multiple samples in the same study and samples across studies. A cohort of 378 SLE cases and 74 controls was assembled from 6 separate publicly available

SLE case/control blood and B cell datasets. We compared the VIRTUS output NHR values between disease cases and controls using a Mann-Whitney test. The same was done for UC, where across a total of 7 case/control studies of intestinal biopsies, a cohort of 669 UC cases and 59 controls was built and run through VIRTUS and the Mann-Whitney test.

Supplementary Data 10:

RELI Results for EBV-SLE and CMV-UC

This document contains all RELI results using differentially expressed genes (DEGs) upon infection with either EBV or CMV and a random sample, containing the same number of genes as in the respective DEGs set, of genes whose expression remained unaltered upon infection for all datasets tested (unchanged). For a RELI result to be significant the “Corrected_P_val” must be less than 0.01 and the “Enrichment” must be greater than 2. EBV-infection-altered gene sets were tested against SLE GWAS loci from the GWAS catalog (“EBV_SLE” sheet) as well as a set of null loci made by sampling the same number of SNPs as seen in the SLE GWAS loci randomly from snp151Common (“EBV_SLE_null” sheet). In the same manner, CMV-infection altered gene sets and the respective unchanged gene sets were run against ulcerative colitis (UC) GWAS loci, Crohn’s disease (CD) GWAS loci, and Inflammatory bowel disease (IBD) GWAS loci, all with a corresponding analysis of disease-specific GWAS risk loci number matching null SNPs.