

25 **ABSTRACT**

26 Electronic health records (EHRs) contain rich temporal data about infectious diseases, but an
27 optimal approach to identify infections remains undefined. Using the *All of Us* Research
28 Program, we developed computable phenotypes for respiratory viruses by integrating billing
29 codes, prescriptions, and laboratory results within 90-day episodes. Phenotypes computed from
30 265,222 participants yielded cohorts ranging from 238 (adenovirus) to 28,729 (SARS-CoV-2)
31 cases. Virus-specific billing codes showed varied sensitivity (8-67%) and high positive predictive
32 value (90-97%), except for influenza virus and SARS-CoV-2 where lower PPV (69-70%)
33 improved with increasing billing codes. Identified infections exhibited expected seasonal
34 patterns and virus proportions when compared with CDC data. This integrated approach
35 identified episodic disease more effectively than individual components alone and demonstrated
36 utility in identifying severe infections. The method enables large-scale studies of host genetics,
37 health disparities, and clinical outcomes across episodic diseases.

38 INTRODUCTION

39 Respiratory infections are among the most common human diseases. Severity is influenced by
40 demographics, social determinants of health, comorbidities, immunosuppression, lifestyle,
41 exposures, and genetic factors.¹ Influenza virus, respiratory syncytial virus (RSV), and SARS-
42 CoV-2 are well-studied causes of lower respiratory tract infections, but over 25 known viruses
43 can cause such disease.² The advent of multiplex testing has revealed that rhinovirus (RV),
44 human metapneumovirus (hMPV), parainfluenza viruses (PIV), and common human
45 coronaviruses (hCoVs) can cause similar symptoms, detection rates, morbidity, mortality, and
46 healthcare costs in hospitalized patients.³⁻⁵

47
48 Risk factors for severe infection by influenza virus, RSV, and SARS-CoV-2 are well
49 characterized. These include age extremes, immunosuppression, male sex, smoking, and
50 comorbidities such as obesity, chronic lung disease, hypertension, diabetes, and heart
51 failure.^{1,6-8} Similarly, population-level genetic risk factors have been identified through genome-
52 wide studies for SARS-CoV-2, avian influenza virus (H7N9), pandemic influenza virus
53 (H1N1_{pdm09}), and influenza virus by survey report.⁹⁻¹⁴ While underrecognized viruses share
54 similar risk factors like advanced age and immunocompromised status, they remain
55 understudied.^{3,4,15,16}

56
57 Population-level studies of respiratory infections typically rely on administrative claims data,
58 laboratory surveillance data, or curated clinical cohorts.^{15,17} While laboratory results and some
59 pathogen-specific billing codes are highly specific, they have poor sensitivity due to infrequent
60 testing, and adding non-specific International Classification of Diseases (ICD) codes only
61 modestly improves sensitivity.^{18,19} In contrast, electronic health record (EHR)-based
62 phenotyping algorithms can integrate multiple data types to reliably identify disease cohorts for
63 observational studies. Although integrating billing codes, clinical notes, and medications in

64 disease phenotyping using EHRs improves performance, this approach has only rarely been
65 used for respiratory viruses.^{15,20–23}

66
67 This study aimed to create and assess a computable phenotype for identifying viral respiratory
68 infections in EHR data using the National Institutes of Health’s *All of Us* Research Program (*All*
69 *of Us*). We evaluated ICD billing codes, medications, and laboratory results to identify infections,
70 and assessed phenotype performance through specificity, positive predictive value (PPV) and
71 sensitivity calculations compared to gold-standard laboratory testing; and compared the
72 frequency and distribution of laboratory results against CDC surveillance data.

73

74 **MATERIALS AND METHODS**

75 **Data Acquisition**

76 We analyzed data from the *All of Us* Research Program, which digitally enrolls participants aged
77 18 years and older across the United States.²⁴ *All of Us* is a large, diverse national cohort where
78 participants contribute survey data, standardized physical measurements, biospecimens, and
79 EHR data including billing codes, prescriptions, and laboratory results [23]. Participants provide
80 consent to share health information, which includes physical measurements, surveys, genomic
81 data, and EHRs. The informed consent and enrollment process has been described, and
82 specific Institutional Review Board approval is not required for Controlled Tier use of de-
83 identified data, deemed nonhuman subjects research by the *All of Us* Institutional Review
84 Board.^{24,25} The program prioritizes recruitment of populations historically underrepresented in
85 biomedical research.²⁴ This analysis used Controlled Tier data (C2022Q4R13) from the *All of Us*
86 Researcher Workbench and was restricted to the 265,222 participants who had ICD codes,
87 medications, or laboratory results in their EHR data between 1/1/1981 and 7/1/2022. Data
88 linkage, follow-up completeness, quality assessment, privacy, and community engagement are

89 described in the *All of Us* Protocol.²⁶ This study meets all five of the CODE-EHR minimum
90 framework standards for the use of structured health care data in clinical research, with one out
91 of five standards meeting preferred criteria.²⁷ Participants' demographic data were derived from
92 the *All of Us* Researcher Workspace's "person" table and its "The Basics" survey.

93

94 **Phenotype Development**

95 We developed computable phenotypes for eight respiratory viruses: rhinovirus (RV); human
96 metapneumovirus (hMPV); respiratory syncytial virus (RSV); adenovirus (ADV); SARS-CoV-2,
97 parainfluenza (PIV); common human coronavirus (hCoV); and influenza virus. Patient
98 encounters were identified in the EHR if they had at least one of the following: a virus-specific
99 billing code (ICD-9-CM or ICD-10-CM), an antiviral indicated for the target pathogen, or a
100 positive laboratory test. We identified virus-specific billing codes and Logical Observation
101 Identifiers Names and Codes (LOINC) laboratory results by searching for the virus name and
102 related terms (*e.g.*, "adenovirus" and "adenoviral pneumonia"). We excluded codes and results
103 for zoonotic infections, vaccine-related events, and ICD codes for explicitly non-respiratory
104 conditions (*e.g.*, "ovine adenovirus", "enteritis due to adenovirus"). We also excluded codes and
105 results for pathogens sharing components of the virus name (*e.g.*, "Haemophilus influenzae").
106 Laboratory data included nucleic acid amplification, antigen, and culture results. For influenza
107 virus and SARS-CoV-2, we included antiviral medications (*i.e.*, oseltamivir, zanamivir, and
108 baloxavir for influenza virus; remdesivir, molnupiravir, and nirmatrelvir/ritonavir for SARS-CoV-2)
109 given for more than one day. For oseltamivir and zanamivir, prescriptions were removed if the
110 duration of treatment was greater than 6 days to exclude prophylaxis. Antivirals for other
111 respiratory viruses were not included due to poor specificity or their reserved use for severe or
112 immunocompromised cases (Figure 1A).

113

114 We computed phenotypes for each infection episode by grouping related clinical events (Figure
115 1A). An episode began with the first occurrence (t_0) of any virus-specific component: a virus-
116 specific ICD code, positive laboratory result, or qualifying antiviral prescription. All subsequent
117 components within 90 days of t_0 were considered part of the same episode, while events
118 beyond 90 days initiated new episodes.¹⁵ To capture false negatives, we included negative or
119 indeterminate laboratory results from t_0 -5 days through the episode's end. The per-episode
120 occurrence of constituent components (ICD codes, positive laboratory results, and medications)
121 were tallied (Figure 1B). Counts for all virus-specific ICD codes, laboratory results, and
122 medications used for phenotyping are provided (Tables S1-3). We de-duplicated row-level
123 entries; categorized visit types into major categories (*i.e.*, intensive care unit, inpatient,
124 emergency room, urgent care, post-acute care, outpatient, and unknown); and reclassified
125 laboratory results as positive, negative, or indeterminate (Table S4).

126

127 **Phenotype Sensitivity Analyses**

128 *Positive Predictive Value, Specificity, and Sensitivity Calculations*

129 Using non-antigen test results as reference standard, we calculated the performance (PPV,
130 sensitivity, and specificity) based on increasing counts of virus-specific ICD codes within each
131 episode (*e.g.*, multiple occurrences of J10.1, or combinations like J10.1 + J11 + 487). For a
132 given threshold N ($n = 0$, $n \geq 1$, $n \geq 2$, $n \geq 3$, $n \geq 4$), we defined true positives as episodes with N
133 virus-specific ICD codes and a positive test, false positives as episodes with N ICD codes and
134 only negative tests, and false negatives as episodes with fewer than N ICD codes and a positive
135 test.

136

137 For influenza virus and SARS-CoV-2, we additionally tested the performance of incorporating
138 antiviral prescriptions. We first assessed specificity, sensitivity, and PPV in cases requiring both

139 the specified ICD count and a prescription to be considered positive, or fewer than N ICD codes
140 and no prescription to be considered negative. Then, we evaluated full performance metrics
141 (sensitivity, specificity, PPV, negative predictive value (NPV), and phi coefficient) across
142 different combinations of ICD code thresholds and medication criteria for these two viruses.

143

144 *Temporal and Geographic Analysis of Phenotype-Positive Episodes*

145 For any episode that met positivity criteria, we analyzed temporal patterns by calculating three-
146 week moving averages of episode and constituent component counts from July 2017 (MMWR
147 week 201726) through June 2022 (MMWR week 202225). Because hCoV PPV was lower than
148 expected for non-influenza, non-SARS-CoV-2 viruses, and as hCoV ICD counts were disrupted
149 during the COVID-19 pandemic, hCoV ICD codes after February 1, 2020 were excluded from
150 the analysis (Figure S3). We assessed the geographic distribution of episode rates by three-
151 digit ZIP code prefixes (zip3).

152

153 *Level of Care Sensitivity Analysis*

154 Because of differences in testing and care by facility type and acuity, for each episode we
155 identified the highest acuity encounter (from lowest to highest: outpatient, post-acute care,
156 urgent care, emergency department, or inpatient) within a window spanning t_0-7 days through
157 t_0+14 days. We chose this window after analyzing the distributions of visit timing for phenotype-
158 related visits (those associated with virus-specific ICD codes, antivirals, or laboratory results)
159 and all visits (Figure S4). Use of all visits within the selected window, rather than only
160 phenotype-related visits, reduced the overall percentage of missing encounter data from 25.9%
161 to 15.4%; across viruses and ICD counts, median missing encounter data decreased from
162 17.6% (IQR 7.6-28.3%) to 11.3% (IQR 2.2-17.8%) (Table S5). Rarely (0.09%-0.43%),
163 phenotype-related maximum encounter acuity exceeded encounters within this time window
164 (e.g., hospitalization occurring >7 days prior to initial viral diagnosis, Figure S5). For PIV, hMPV,

165 and RSV, manual review of these visits revealed that the associated encounter start date
166 preceded the window period by a few days, and the phenotype component level of care was
167 retained. We analyzed level-of-care patterns across viruses, stratifying by ICD codes and test
168 positivity.

169

170 **Comparison with National Surveillance Data**

171 To understand the representativeness of respiratory illness data in *All of Us*, we compared the
172 seasonal percent positivity and test volume of *All of Us* EHR laboratory results to data from
173 three CDC/WHO sources: National Respiratory and Enteric Virus Surveillance System
174 (NREVSS), COVID-19 Data Tracker, and the Global Influenza Surveillance and Response
175 System (GISRS).

176

177 First, we assessed geographic coverage by comparing *All of Us* participant locations and testing
178 rates to deduplicated NREVSS clinical laboratory results from 2017-2021.²⁸ *All of Us* EHR
179 participant counts were visualized by aggregating data across three-digit ZIP code prefixes
180 (zip3). Zip3 information was missing for 2/265,222 (0.00%). *All of Us* EHR participants (per
181 1,000 zip3 2020 Census population) and *All of Us* participants tested (per 1,000 *All of Us*
182 participants with EHR data) were similarly aggregated by zip3 code. Zip3 regions with five or
183 fewer *All of Us* participants were removed and classified as “No Data.” We obtained zip3
184 boundaries from US Census Bureau zip code tabulation areas, 2017 cartographic state
185 boundaries from the Vega us-10m.json dataset, and 2020 zip code tabulation area populations
186 from the US Census Bureau.

187

188 Next, we assessed virus distribution by comparing proportions detected ($\% = N \text{ type} / N \text{ total}$
189 with known type) in *All of Us* to surveillance data from NREVSS and the GISRS.^{29–31} These

190 comparisons were limited to hCoV, influenza virus, and PIV, as these were the only viruses for
191 which syndromic multiplex panels routinely report type-specific results.

192
193 To evaluate temporal patterns, we compared weekly test positivity data between *All of Us* and
194 CDC surveillance from the first week of July, 2016 to the last week of June, 2022. For each
195 virus, we calculated the percentage of positive tests for each *MMWR* reporting week. We plotted
196 three-week moving averages for both percent positivity and total tests performed for *All of Us*
197 and CDC data. We obtained CDC comparison data from NREVSS for non-SARS-CoV-2
198 viruses, additional influenza virus data from FluView, and SARS-CoV-2 data from the COVID
199 Data Tracker.^{28,32,33}

200

201 **RESULTS**

202 **Cohort Characteristics**

203 Among 265,222 *All of Us* participants with ICD codes, medication entries, or laboratory results
204 recorded between 1/1/1981 and 7/1/2022, we identified respiratory virus episodes that varied
205 substantially in duration and composition (Figure 1B). SARS-CoV-2 (n [distinct
206 episodes]=28,729) and influenza virus (n=19,784) were the largest cohorts, followed by RV
207 (n=1,620), hCoV (n=1,437), and RSV (n=1,161) and the smallest cohorts, hMPV (n=486), PIV
208 (n=400), and ADV (n=238).

209

210 Across all cohorts, participants were predominantly female (61-68%) with median ages mostly
211 between 50 and 58 (Table S6). Participants who self-reported as White were the plurality for
212 every virus (32.9-60.1%), compared to participants self-reporting as Black (16.5-28.5%) or
213 Hispanic/Latino (17-32.1%). All other options (Asian, multiple selected, Middle Eastern or North
214 African, and Native Hawaiian or Other Pacific Islander) were rare (0-2.2%). SARS-CoV-2 and
215 influenza virus participant demographics most closely mirrored the overall *All of Us* cohort with

216 ICD, laboratory, or medication data. These two groups more frequently self-reported as White
217 (50.4-60.1%) and were more frequently employed with higher reported income, education, and
218 employer-provided insurance. Demographic data were only notably missing for insurance type
219 (46,487/265,222=17.5% for all participants with EHR data). Test count per person was higher
220 among infected cohorts than tested cohorts for each virus.

221
222 Episodes commonly consisted of either laboratory results alone (predominant for RV [74.7%],
223 PIV [65.0%], SARS-CoV-2 [31.3%], hMPV [32.9%]) or single ICD codes (predominant for ADV
224 [45.8%], hCoV [43.1%], RSV [35.4%], influenza [34.8%]) (Figure 1B). Antiviral use varied
225 markedly: SARS-CoV-2 episodes rarely included antiviral prescriptions (4.57%), while
226 medication-only episodes were frequently observed for influenza virus (22.9%), even after
227 excluding prophylactic prescriptions.

228

229 **Phenotype Performance for Detecting True Positives**

230 To understand the diagnostic performance of varying ICD codes for an episode, we calculated
231 sensitivity, specificity, and PPV for each virus across N ICD codes per episode using non-
232 antigen test results as a reference standard (Figure 2). The sensitivity of using one or more ICD
233 codes varied between viruses and decreased as the minimum N ICD codes increased.

234 Diagnoses using one or more ICD codes for influenza virus had the highest sensitivity (66.8%),
235 compared to moderate sensitivity for RSV (55.2%), SARS-CoV-2 (44.8%), ADV (42.4%), hMPV
236 (40.2%), and hCoV (33.4%), and minimal sensitivity for RV (9.2%) and PIV (8.3%).

237

238 Specificity and PPV demonstrated similar patterns, with exaggerated variation in PPV initially
239 demonstrating three groupings. First, for influenza virus and SARS-CoV-2, the PPV for one or
240 more ICD codes was lower (69.7% and 68.8%, respectively), but increased as minimum N ICD
241 count increased (78.1% and 76.7% for at least 2 ICD codes, respectively) (Figure 2). Second,

242 for non-influenza, non-SARS-CoV-2 viruses, the PPV was high and did not change substantially
243 as ICD count increased (89.7-97.3%). Third, hCoV initially demonstrated a high PPV (79.5%)
244 that decreased as ICD count increased (71.8% for at least 2 ICD codes) (Figure S3A).

245
246 During the COVID-19 pandemic, hCoV ICD code counts spiked above historical maxima despite
247 an absence of positive tests (Figure S3B). After removing hCoV ICD codes after February 1,
248 2020, the PPV trend for common hCoV became similar to other non-influenza, non-SARS-CoV-
249 2 viruses (Figure 2, Figure S3A).

250
251 Adding medication use to the phenotype had varying effects on performance. As with the
252 medication-exclusive phenotypes, specificity and PPV increased with each additional ICD code
253 for the medication-inclusive influenza and SARS-CoV-2 cohorts. While only a small proportion
254 (1,345/28,741=4.67%) of SARS-CoV-2 episodes included a prescription for remdesivir,
255 molnupiravir or nirmatrelvir, the addition of medication to the phenotype did increase PPV for
256 this subset of 1,345 participants (Figure 2). For influenza virus, medication use alone was poorly
257 predictive (PPV=46.8%), but combining medications with 1 ICD code improved PPV compared
258 to 1 or more codes alone (87.1% vs. 69.7%, respectively) (Figure 2).

259
260 Varying ICD code thresholds and antiviral requirements in the entire influenza and SARS-CoV-2
261 cohorts demonstrated an expected trade-off in performance (Table 1). For both viruses, at least
262 one ICD code or medication was the most sensitive phenotype (76.0% influenza virus and
263 45.1% SARS-CoV-2), but this caused the highest number of false positives and the lowest
264 PPVs (65.8% and 68.8%, respectively). For influenza virus, by requiring at least two ICD codes
265 or a medication accompanied by an ICD code, the lower sensitivity (47.7%) was accompanied
266 by a marked reduction in false positives (778 to 238) and increase in PPV (65.8% to 79.8%).

267 Similar trends were observed for SARS-CoV-2, and despite trade-offs, the phi coefficient was
 268 highest for the broadest phenotypes.
 269

Table 1 Phenotype Performance

	Phenotype	TP	Counts (N)			Phenotype Performance				
			FP	FN	TN	Sen.	Spec.	PPV	NPV	ϕ
Flu	1+ ICDs or medication	1496	778	472	32018	0.760	0.976	0.658	0.985	0.688
	1+ ICDs	1315	572	653	32224	0.668	0.983	0.697	0.980	0.664
	2+ ICDs or medication	1120	444	848	32352	0.569	0.986	0.716	0.974	0.619
	2+ ICDs or (1 ICD + medication)	939	238	1029	32558	0.477	0.993	0.798	0.969	0.600
	3+ ICDs or medication	941	339	1027	32457	0.478	0.990	0.735	0.969	0.574
	3+ ICDs or (1-2 ICDs + medication)	760	133	1208	32663	0.386	0.996	0.851	0.964	0.558
	2+ ICDs	691	194	1277	32602	0.351	0.994	0.781	0.962	0.506
	3+ ICDs	360	68	1608	32728	0.183	0.998	0.841	0.953	0.379
	1+ ICDs or medication	7298	3302	8894	177219	0.451	0.982	0.688	0.952	0.526
	1+ ICDs	7258	3298	8934	177223	0.448	0.982	0.688	0.952	0.524
COVID	2+ ICDs or medication	4483	1324	11709	179197	0.277	0.993	0.772	0.939	0.438
	2+ ICDs or (1 ICD + medication)	4443	1320	11749	179201	0.274	0.993	0.771	0.938	0.435
	3+ ICDs or medication	2598	546	13594	179975	0.160	0.997	0.826	0.930	0.345
	3+ ICDs or (1-2 ICDs + medication)	2558	542	13634	179979	0.158	0.997	0.825	0.930	0.342
	2+ ICDs	4309	1310	11883	179211	0.266	0.993	0.767	0.938	0.427
	3+ ICDs	2260	523	13932	179998	0.140	0.997	0.812	0.928	0.318

TP: true positive; FP: false positive; FN: false negative; TN: true negative; Sen: sensitivity; Spec: specificity; PPV: positive predictive value; NPV: negative predictive value; ϕ : phi coefficient (mean square contingency coefficient).

270
 271 We identified infection episodes nationwide, with SARS-CoV-2 and influenza virus
 272 demonstrating the broadest US coverage. While episodes generally matched the *All of Us* EHR
 273 subgroup distribution (Figure S1C), incorporating ICD codes and medications exhibited higher
 274 infection rates in the Southeast and Texas despite lower testing coverage in these regions
 275 (Figure 3). Temporally, episodes composed of 1-3 ICD codes showed seasonality patterns
 276 consistent with test-positive episodes for all frequently detected viruses (Figure 4). During the
 277 early COVID-19 pandemic (winter 2020 to spring 2021), only SARS-CoV-2 and RV were
 278 consistently identified.

279

280 **Patterns in Phenotype Composition by Level of Care**

281 Encounter level of care varied by virus and episode composition. For RV, hMPV, PIV, hCoV,
282 and SARS-CoV-2, episodes defined by at least one test without ICD codes were the most
283 frequent, while for RSV, ADV, and influenza virus, ICD-only episodes predominated (Figure 5).
284 Influenza virus episodes with antiviral prescriptions showed a similar distribution of visit types
285 compared to those without, while SARS-CoV-2 episodes rarely included prescriptions during our
286 study period.

287

288 By percentage, influenza virus and SARS-CoV-2 episodes showed a mix of outpatient, ER, and
289 inpatient encounters, while other viruses demonstrated higher rates of ER visits and
290 hospitalization. Episodes containing a positive test consistently showed higher rates of
291 hospitalization compared to test-negative episodes. Similarly, hospitalization rates increased as
292 the number of ICD codes within an episode increased. The cohorts included very few post-acute
293 care encounters and almost no urgent care encounters.

294

295 **Laboratory Result Comparison**

296 Using national epidemiological data from NREVSS, COVID Data Tracker, and GISRS, we
297 compared *All of Us* laboratory results by geographic coverage, virus type proportion, and
298 temporal trends.

299

300 We found broad US coverage of *All of Us* participants with relevant EHR data (265,222
301 participants), with enriched sampling near population centers in the Northeast megalopolis;
302 Western Pennsylvania; Great Lakes Region; Southeast; Arizona; California; and the
303 metropolitan areas of Austin/Dallas, Kansas City, Denver, and Seattle (Figure S1A, Table S7).
304 Only 3.7% of zip3 codes had no *All of Us* participants with ICD, laboratory, or medication data.

305

306 Testing patterns in the *All of Us* data overlapped with CDC clinical laboratories reporting to
307 NREVSS (Figure S1B) and mirrored participant distribution (Figure S1A) with a notable
308 decrease in testing for all respiratory viruses in the Southeast relative to participant density
309 (Figure S1C). Testing frequency varied substantially by virus; participants were more frequently
310 tested for influenza virus and SARS-CoV-2 compared to all other viruses.

311

312 Virus type distributions in *All of Us* were similar to national surveillance data from NREVSS and
313 GISRS.^{29–31} For PIV (2011-2019), HPIV-3 was most commonly detected and all other types
314 were less frequent (Figure S2A). For hCoV (2014-2021), OC43 was most common and 229E
315 was least common, while the order of NL63 and HKU1 differed (Figure S2B). Influenza virus
316 type proportions (2010-2020) were nearly identical, with influenza virus A more common than
317 influenza virus B (Figure S2C). Cross-dataset influenza subtype comparisons were not
318 available, but in *All of Us*, H3N2 and H1N1 pdm09 were markedly more common than H1N1
319 and H5N1, as expected.

320

321 Test positivity patterns from 2017 to 2022 matched CDC rates for most viruses (mean absolute
322 error 5.89 percent positive tests per week for RV and 1.18-2.82 for all other viruses) (Figure 6).
323 SARS-CoV-2, influenza virus, and RV showed the highest percent positivity, and most viruses
324 showed expected seasonal patterns: PIV and RV exhibited two seasonal peaks per year
325 (spring-dominant for PIV, fall-dominant for RV), while RSV, influenza, and hMPV demonstrated
326 single overlapping winter peaks. SARS-CoV-2 positivity matched expected variant waves (*e.g.*,
327 Alpha, Delta, and Omicron BA.1). Notable differences in the *All of Us* data include more
328 variability in RSV tests and positivity, undercounted positivity by ~10% during peak respiratory
329 season for influenza and RSV, and less ADV positivity, relative to CDC data.

330

331 **DISCUSSION**

332 This study demonstrates that combining virus-specific EHR data elements reliably identifies
333 respiratory viral infections in large biobank datasets. Using temporal, geographic, and sensitivity
334 analyses, we described important performance insights into respiratory virus phenotyping. All
335 phenotypes exhibited high specificity, though trade-offs exist between accuracy and sensitivity
336 for influenza virus and SARS-CoV-2.

337
338 We identified epidemiological patterns and cohort sizes that matched national surveillance data.
339 The combination of laboratory results, ICD codes, and medications increased case detection
340 beyond what any component alone could identify, which was valuable where laboratory testing
341 was less frequent (e.g., the Southeast). The cohorts varied in size: approximately 20,000-
342 30,000 episodes for influenza virus and SARS-CoV-2 compared to 200-1,600 for other viruses,
343 likely reflecting increased clinical suspicion and testing for common pathogens rather than true
344 differences in disease burden.¹⁵ These cohorts could be used to investigate host genetic
345 factors, health disparities, geographic and environmental risk factors, and clinical outcomes
346 across respiratory viral infections. The longitudinal available also supports study of lifestyle
347 factors through wearables and patient-reported outcomes from surveys.

348
349 Phenotype performance varied substantially between viruses, across episode composition, and
350 over time. Virus-specific ICD code sensitivity varied widely: higher for influenza (66.8% vs. 38-
351 95% in published studies) and RSV (55.2% vs. 24%) and moderate for SARS-CoV-2, ADV,
352 hMPV, and hCoV (33-44%), but very low for PIV (8.3% vs. 14% and RV (9.2% vs. 0%).^{18,19,34-36}
353 While non-influenza, non-SARS-CoV-2 virus-specific ICD codes showed high PPV regardless of
354 count (89.7-97.3%), single ICD codes for SARS-CoV-2 and influenza were less predictive (68.8-
355 69.7%).^{18,19,35,36} For medication-inclusive cases, ICD codes had a higher PPV for influenza virus
356 and SARS-CoV-2 episodes, and for SARS-CoV-2, antivirals alone were highly predictive of test

357 results, although receipt of remdesivir, molnupiravir or nirmatrelvir was rare (4.57% of all
358 episodes) during the study period.

359
360 The COVID-19 pandemic disrupted the seasonal transmission of most respiratory viruses,
361 granting the opportunity to assess ICD code performance in unexpected settings and
362 demonstrating the importance of evaluating code performance over time for seasonally variable
363 diseases. hCoV ICD codes were inappropriately used to identify concern for COVID-19
364 infection, with diminished but persistent effects throughout the study period. Apart from
365 rhinovirus, our phenotypes only rarely identified false positive episodes during the COVID-19
366 pandemic, mostly attributable to influenza virus episodes composed of a single ICD code. We
367 suspect that these episodes reflect clinical concern for infection rather than true infection, and
368 indeed, adjusting the influenza phenotype from any ICD code or medication to 2+ ICD codes or
369 medications accompanied by an ICD code markedly reduced false positives and increased
370 PPV. In this work, we suggest choosing phenotype characteristics that match the research
371 question and desire to maximize sensitivity vs. PPV.

372
373 Level-of-care analyses revealed patterns suggesting systematic detection biases toward higher
374 acuity settings. Our phenotype identified a high frequency of infections at emergency and
375 inpatient visits for RV, hMPV, RSV, ADV, PIV, and hCoV, compared to more outpatient visit
376 types for SARS-CoV-2 and influenza virus. These findings differ from established hospitalization
377 rates: CDC estimates suggest that only 1-2% of medically-attended influenza cases require
378 hospitalization, while COVID-19 hospitalization rates ranged between 2.1% and 68% over this
379 study period, with temporal trends showing a decrease from ~50% in the early pandemic to 20%
380 by July, 2022.³⁷⁻⁴¹ Moreover, prospective studies in adults have shown that other respiratory
381 viruses show either similar or lower hospitalization rates compared to influenza - the opposite of
382 our results.^{4,16} This discordance suggests that the *All of Us* computable phenotype oversamples

383 high levels of care, particularly for non-influenza, non-SARS-CoV-2 viruses, likely due at least
384 three factors: the lack of cost-effective outpatient assays, an absence of specific therapeutic
385 interventions that would justify multiplex testing costs in lower-acuity care settings, and the utility
386 of identifying an etiology in the inpatient setting, where cessation of antibiotics or discharge are
387 considerations.

388
389 Several limitations affect the interpretation and generalizability of these findings. The low
390 sensitivity for all viruses indicates that this method cannot be used to study disease prevalence,
391 as nonspecific syndromic coding likely predominates for upper respiratory illnesses.

392 Additionally, the requirement for data conversion to a common data model before release in
393 curated data repositories means this method cannot support real-time surveillance. In addition,
394 while *All of Us* provides broad national coverage, it highly sampled some states (AZ, MA, WI,
395 AL, PA, IL, MI, NY, MS, CA) and left approximately half the US population sampled below rates
396 of 1 in 10,000. The program's intentional oversampling of populations historically
397 underrepresented in biomedical research, while invaluable to health equity research, also
398 results in demographics that differ from the overall US population. Other known impediments to
399 generalizability compared to the US population include decreased representation of blind and
400 deaf participants; difficulty in linking EHRs from a considerable portion of *All of Us* participants;
401 and the decreased representation of persons of Asian, Middle Eastern or North African, and
402 Native Hawaiian or Other Pacific Islander heritage in this cohort.⁴²

403
404 Finally, this work faces challenges common to EHR-based research. These include labeling
405 bias, implicit clinician biases which could be influenced by demographics, and informed
406 presence bias where EHR inclusion typically reflects illness rather than routine care.^{43,44} The
407 minimal representation of urgent care and post-acute care in our data further underscore the

408 disconnected nature of healthcare in the US, and identifies a likely gap in infection detection in
409 this cohort.

410
411 Despite these limitations, our computable phenotypes reliably detected geographic and
412 temporal patterns of infection matching national surveillance, although severe infections are
413 oversampled and many mild infections are likely missed. This work supports a need for
414 expanded surveillance of non-influenza/non-SARS-CoV-2 pathogens in routine medical care.⁴⁵
415 Using a cohort that is expected to continually grow, our results enable future studies of genetic
416 susceptibility and clinical outcomes research across both well-studied and understudied
417 respiratory viruses. This work serves as a foundation for the creation and validation of other
418 computable phenotypes for episodic infectious diseases using EHR-based methods.

419

420 **DATA AND CODE AVAILABILITY**

421 The Community Workspace “Respiratory Virus Computable Phenotype” is available for all
422 approved *All of Us* users and includes all code and data used in this work

423 ([https://workbench.researchallofus.org/workspaces/aou-rw-](https://workbench.researchallofus.org/workspaces/aou-rw-ae307fda/respiratoryviralinfectionsinallofus/analysis)
424 [ae307fda/respiratoryviralinfectionsinallofus/analysis](https://workbench.researchallofus.org/workspaces/aou-rw-ae307fda/respiratoryviralinfectionsinallofus/analysis)).

425

426 **ACKNOWLEDGEMENTS**

427 S Olsen and A Winn (CDC) kindly provided percent positivity data and locations of CDC clinical
428 labs reporting to NREVSS. This work received an exception to the Data and Statistics
429 Dissemination Policy from the *All of Us* Resource Access Board for figures 4 and 5. The primary
430 author acknowledges the use of Claude (Anthropic) in the generation and revision of code to
431 produce this computable phenotype, and in the revisions of this manuscript.

432

433 The contents of this publication are the sole responsibility of the authors. The content of this
434 publication does not necessarily reflect the views, opinions, or policies of the NIH, the
435 Uniformed Services University of the Health Sciences, the US Department of Health and Human
436 Services, the US Department of Defense, or the US Government, nor does mention of trade
437 names, commercial products, or organizations imply endorsement by the US government. This
438 work was prepared by a military or civilian employee of the US Government as part of the
439 individual's official duties. Therefore, it is in the public domain and does not possess copyright
440 protection. Public domain information may be freely distributed and copied; however, as a
441 courtesy, it is requested that the authors be given an appropriate acknowledgement.

442
443 This work was supported by the National Human Genome Research Program Intramural
444 Research Program, grant numbers: ZIA HG200417, ZIC HG200420; and the Division of
445 Intramural Research of the National Institute of Allergy and Infectious Diseases. The All of Us
446 Research Program is supported by the National Institutes of Health, Office of the Director:
447 Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2
448 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2
449 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN
450 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121;
451 The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24
452 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and
453 Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2
454 OD025276. Funders played no role in study design, data collection, analysis and interpretation
455 of data, or the writing of this manuscript.

456

457 **AUTHOR CONTRIBUTIONS**

458 BJW and JCD conceived of the study design. BJW constructed the phenotype, conducted all
459 analyses, and drafted the manuscript. TCT and HM contributed foundational data analysis
460 support. FABC and EER offered appraisal of data analysis and early revision of the manuscript.
461 All authors contributed to the final revision of the manuscript.

462

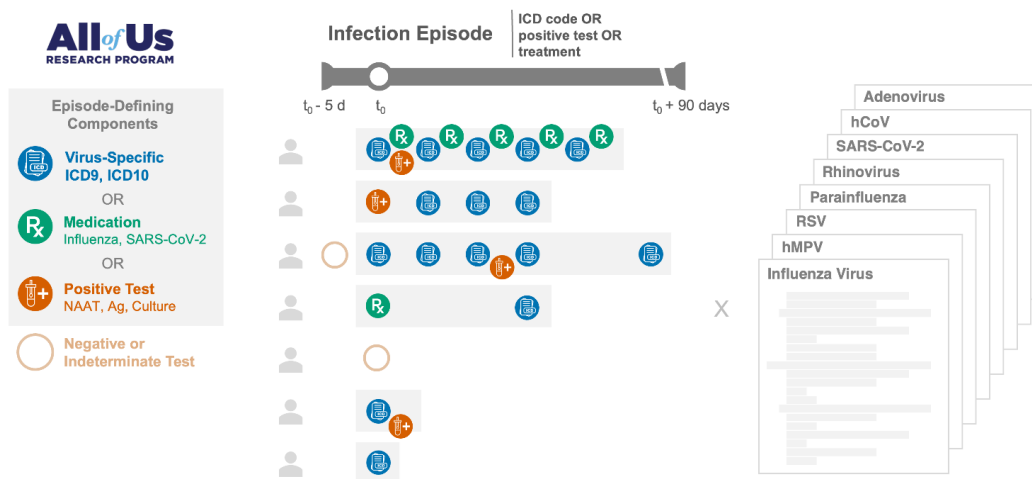
463 **COMPETING INTERESTS**

464 All authors declare no financial or non-financial competing interests.

465

466 **FIGURES AND FIGURE LEGENDS**

a Phenotype Definition



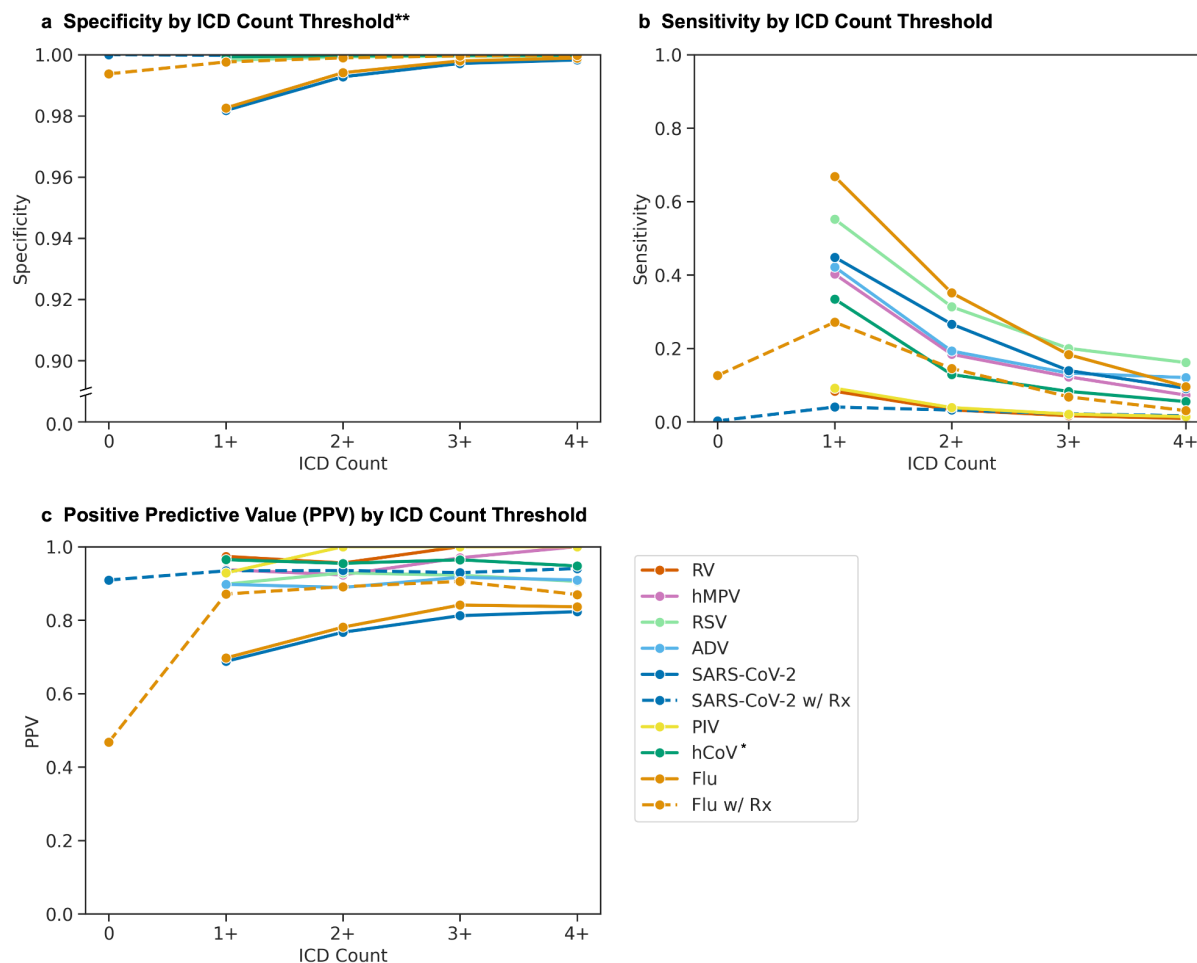
b Episode Proportion by Component



467

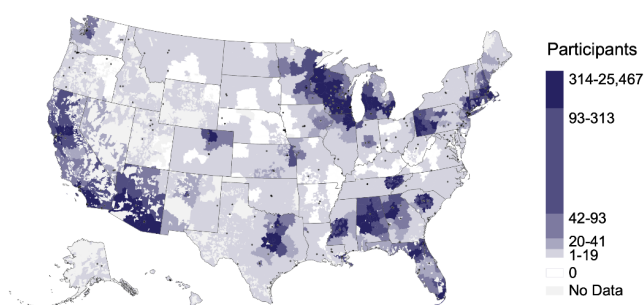
468 **Figure 1.** Computational phenotype for respiratory virus episodes using electronic health
 469 records (EHRs) with respiratory episode composition. A: Episodes were defined by (1) virus-
 470 specific ICD-9-CM or ICD-10-CM codes, (2) antiviral medications (for influenza and SARS-CoV-

471 2), and/or (3) positive laboratory results including nucleic acid amplification tests (NAAT),
472 antigen tests, or cultures. The first qualifying event is designated as time zero (t_0), and all
473 related, subsequent events within 90 days were grouped into the same episode. Negative or
474 indeterminate tests were also included, with a five-day lookback window to incorporate false
475 negative results. Phenotypes were computed for influenza, human metapneumovirus (hMPV),
476 respiratory syncytial virus (RSV), parainfluenza, rhinovirus (RV), SARS-CoV-2, common human
477 coronavirus (hCoV), and adenovirus (ADV). Heatmap (B) showing the breakdown of episode
478 types (columns) for each virus (row). Colors indicate the percentage of counts for each virus
479 (e.g., 74.9% of RV episodes (1,210/1,620) contained only positive laboratory results.
480 Percentage and corresponding color for counts below 20 (with a second count to prevent back-
481 calculation for hMPV), were censored per the *All of Us* participant privacy policy. N/A : not
482 applicable.
483

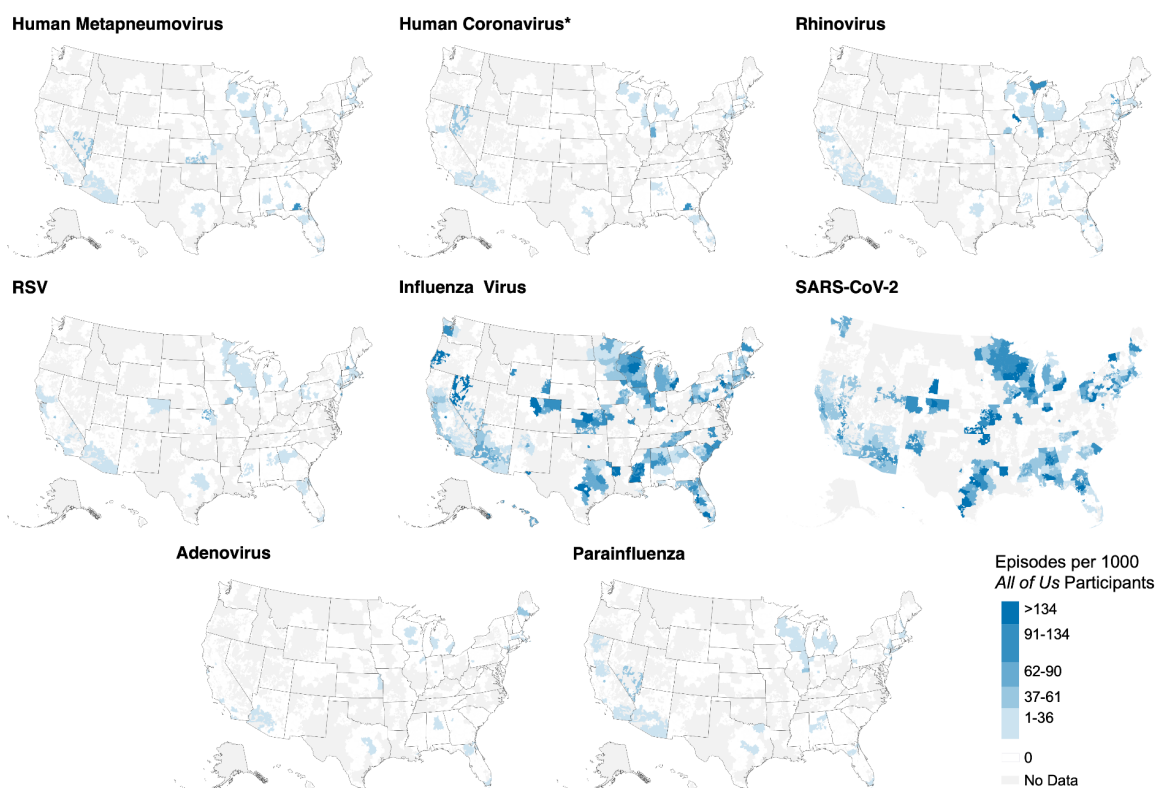


484
 485 **Figure 2.** Phenotype performance across different episode definitions. Specificity (A), sensitivity
 486 (B), and PPV (C) were calculated using non-antigen laboratory results as the reference
 487 standard. For each virus, colored lines show performance across episodes containing
 488 increasing numbers of ICD codes. For influenza virus (dashed orange line) and SARS-CoV-2
 489 (dashed dark blue line), additional lines show performance when episodes were restricted to
 490 episodes containing both ICD codes and antiviral prescriptions (Rx). *Human coronavirus
 491 (hCoV) episodes excluded ICD codes after February 1, 2020, due to loss of specificity during
 492 the COVID-19 pandemic (Figure S3A). **Y-axis for specificity (A) is broken to depict differences
 493 near 1.0. PIV: parainfluenza; hMPV: human metapneumovirus; RV: rhinovirus; ADV:
 494 adenovirus; RSV: respiratory syncytial virus; hCoV: human coronavirus; Flu: influenza virus.

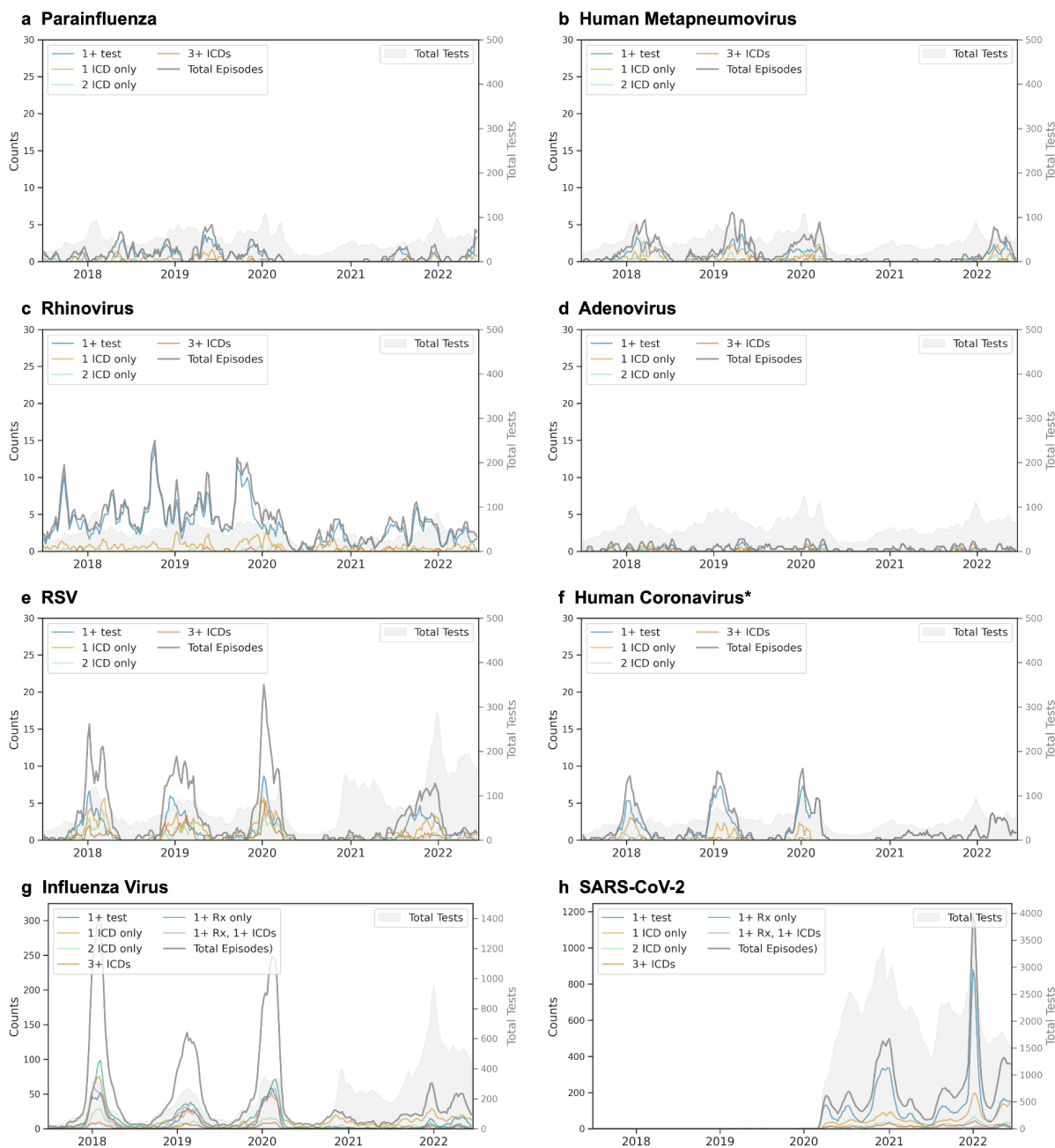
a All of Us Participants with Diagnosis, Drug, or Lab Result



b All of Us Respiratory Infection Episodes (per 1000 Participants)



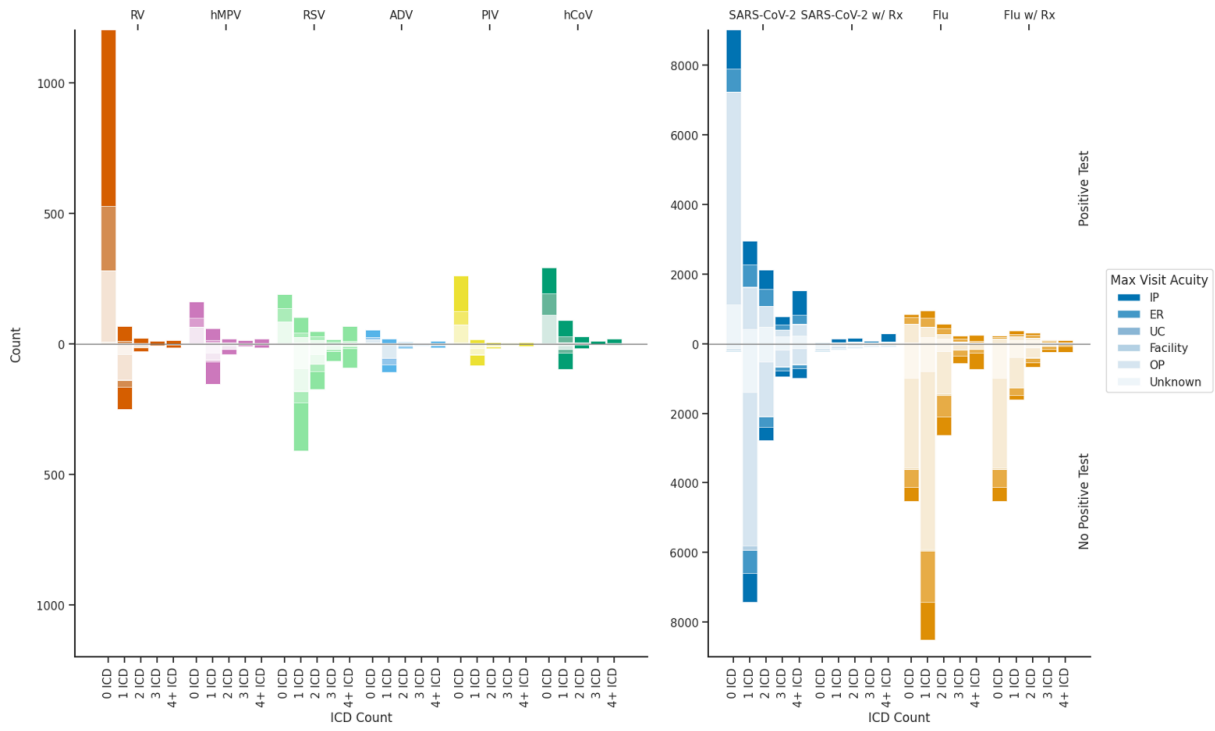
495
496 **Figure 3.** Geographic distribution of respiratory virus episodes. Heat maps show episode rates
497 per 1,000 *All of Us* participants with EHR data by three-digit zip code prefix. Colors represent
498 quintiles defined by SARS-CoV-2 rates, the largest cohort. Regions with five or fewer *All of Us*
499 participants are marked as "No Data" in (B). *Human coronavirus episodes were filtered as
500 described in methods. RSV: respiratory syncytial virus.
501



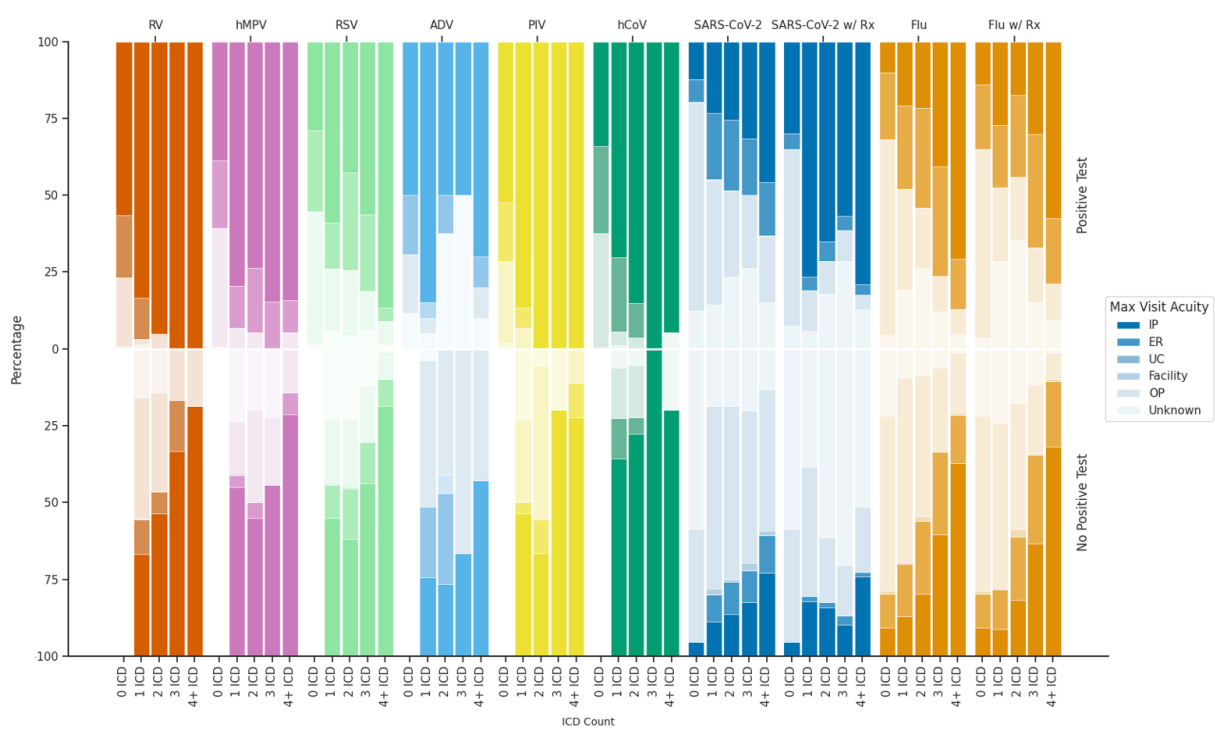
502
 503 **Figure 4.** Temporal patterns of respiratory virus episodes by composition. Three-week moving
 504 averages shown for (A) parainfluenza, (B) human metapneumovirus, (C) rhinovirus, (D)
 505 adenovirus, (E) respiratory syncytial virus (RSV), (F) human coronavirus, (G) influenza virus,
 506 and (H) SARS-CoV-2. For each virus, lines show total episodes (gray) and episodes by
 507 composition: one or more positive tests (blue), single ICD code (light orange), two ICD codes

508 (light green), and three or more ICD codes (red-orange). For influenza, and SARS-CoV-2,
509 additional lines show episodes with antiviral prescriptions alone (dark green) or with concurrent
510 ICD codes (pink). Gray shading indicates testing volume. Left y-axis corresponds to episode
511 counts; right y-axis shows total tests performed. *Episodes for hCoV are shown after applying
512 temporal filtering (unfiltered plot compared in [Figure S3](#)).
513

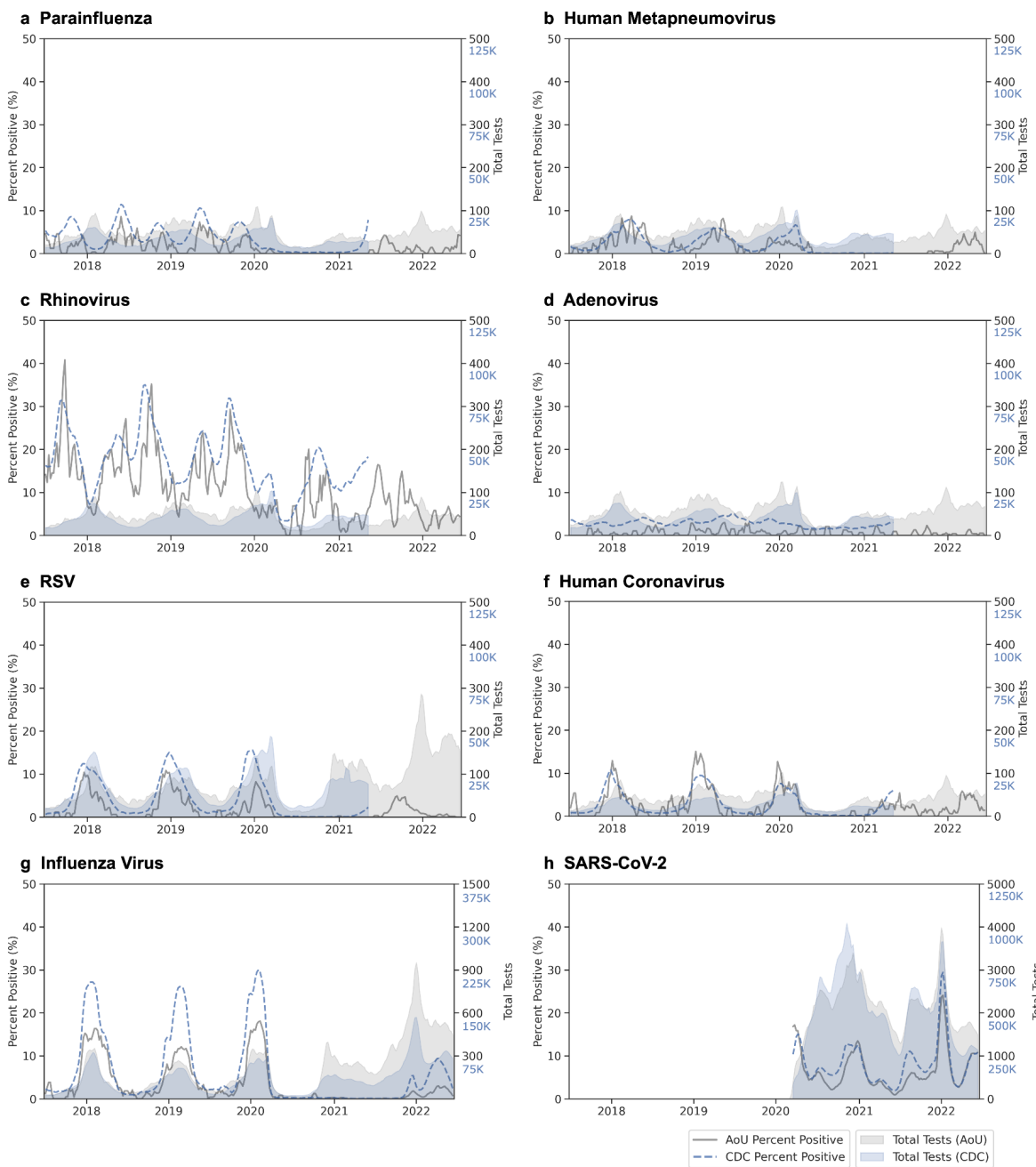
a Visit Type Count by Virus and ICD Code Count



b Visit Type Percentage by Virus and ICD Code Count



515 **Figure 5.** Level of care patterns by virus and episode composition. For each virus, (A) counts
516 and (B) percentages of maximum level of care recorded between seven days before and 14
517 days after episode start. Results are stratified by episode characteristics: episodes with positive
518 tests (upper panels) versus those without (lower panels), and by number of ICD codes
519 (columns). Color intensity corresponds to level of care: inpatient (IP, darkest), emergency room
520 (ER), urgent care (UC), outpatient (OP), and unknown (lightest). For influenza virus (Flu) and
521 SARS-CoV-2, additional columns show episodes containing antiviral prescriptions. PIV:
522 parainfluenza; hMPV: human metapneumovirus; RV: rhinovirus; ADV: adenovirus; RSV:
523 respiratory syncytial virus; hCoV: human coronavirus; Rx: antiviral prescription.
524



525
 526 **Figure 6.** Temporal validation of test positivity using CDC surveillance data. Three-week moving
 527 average of test positivity and test volume comparing *All of Us* (gray) to CDC surveillance data
 528 (blue) for eight respiratory viruses: (A) parainfluenza, (B) human metapneumovirus, (C)
 529 rhinovirus, (D) adenovirus, (E) respiratory syncytial virus (RSV), (F) common human
 530 coronavirus (hCoV), (G) influenza virus, and (H) SARS-CoV-2. In each case, percent positivity

531 (lines) corresponds to the left y-axis and total tests performed (shaded areas) corresponds to
532 the right y-axis. CDC data were obtained from NREVSS for panels A-F (ending 2021), FluView
533 for influenza virus, and COVID Data Tracker for SARS-CoV-2. AoU: *All of Us*.

534

535 REFERENCES

536 1. Clementi, N. et al. Viral Respiratory Pathogens and Lung Injury. *Clin. Microbiol. Rev.*
537 222, 563–569 (2019).

538 3. Boon, H., Meinders, A., Hannen, E. J. van, Tersmette, M. & Schaftenaar, E.
539 Comparative analysis of mortality in patients admitted with an infection with influenza
540 A/B virus, respiratory syncytial virus, rhinovirus, metapneumovirus or SARS-CoV-2.
541 *Influ. Other Respir. Viruses* 18, e13237 (2024).

542 4. Lee, N. et al. Burden of noninfluenza respiratory viral infections in adults admitted to
543 hospital: analysis of a multiyear Canadian surveillance cohort from 2 centres. *CMAJ*
544 193, E439–E446 (2021).

545 5. Korsten, K. et al. Burden of respiratory syncytial virus infection in community-dwelling
546 older adults in Europe (RESCEU): an international prospective cohort study. *Eur.*
547 *Respir. J.* 57, 2002688 (2021).

548 6. DeMartino, J. K. et al. Respiratory Syncytial Virus–Related Complications and
549 Healthcare Costs Among a Medicare-Insured Population in the United States. *Open*
550 *Forum Infect. Dis.* 10, ofad203 (2023).

551 7. Chenchula, S. et al. Global prevalence and effect of comorbidities and smoking
552 status on severity and mortality of COVID-19 in association with age and gender: a
553 systematic review, meta-analysis and meta-regression. *Sci. Rep.* 13, 6415 (2023).

- 554 8. Andrew, M. K. et al. Age Differences in Comorbidities, Presenting Symptoms, and
555 Outcomes of Influenza Illness Requiring Hospitalization: A Worldwide Perspective From
556 the Global Influenza Hospital Surveillance Network. *Open Forum Infect. Dis.* 10,
557 ofad244 (2023).
- 558 9. Lee, N. et al. IFITM3, TLR3, and CD55 Gene SNPs and Cumulative Genetic Risks for
559 Severe Outcomes in Chinese Patients With H7N9/H1N1pdm09 Influenza. *J. Infect. Dis.*
560 216, 97–104 (2017).
- 561 10. López-Rodríguez, M. et al. IFITM3 and severe influenza virus infection. No evidence
562 of genetic association. *Eur. J. Clin. Microbiol. Infect. Dis.* 209, 1028–1031 (2014).
- 563 12. Yang, X. et al. Interferon-Inducible Transmembrane Protein 3 Genetic Variant
564 rs12252 and Influenza Susceptibility and Severity: A Meta-Analysis. *PLoS ONE* 10,
565 e0124985 (2015).
- 566 13. Initiative, C.-19 H. G. et al. A first update on mapping the human genetic
567 architecture of COVID-19. *Nature* 608, E1–E10 (2022).
- 568 14. Kosmicki, J. A. et al. Genetic risk factors for COVID-19 and influenza are largely
569 distinct. *Nat. Genet.* 1–5 (2024) doi:10.1038/s41588-024-01844-1.
- 570 15. Hedberg, P. et al. Clinical phenotypes and outcomes of SARS-CoV-2, influenza,
571 RSV and seven other respiratory viruses: a retrospective study using complete hospital
572 data. *Thorax* . 206, 56–62 (2012).
- 573 17. Buda, S., Tolksdorf, K., Schuler, E., Kuhlen, R. & Haas, W. Establishing an ICD-10
574 code based SARI-surveillance in Germany – description of the system and first results
575 from five recent influenza seasons. *BMC Public Heal.* 17, 612 (2017).
- 576 18. Cai, W. et al. Evaluation of using ICD-10 code data for respiratory syncytial virus

- 577 surveillance. *Influ. Other Respir. Viruses* 14, 630–637 (2020).
- 578 19. Hamilton, M. A. et al. Validating International Classification of Disease 10th Revision
579 algorithms for identifying influenza and respiratory syncytial virus hospitalizations. *PLoS*
580 *ONE* 16, e0244746 (2021).
- 581 20. Wei, W.-Q. et al. Combining billing codes, clinical notes, and medications from
582 electronic health records provides superior phenotyping performance. *J. Am. Med.*
583 *Inform. Assoc.* 23, e20–e27 (2016).
- 584 21. Fathima, S., Simmonds, K., Invik, J., Scott, A. N. & Drews, S. Use of laboratory and
585 administrative data to understand the potential impact of human parainfluenza virus 4
586 on cases of bronchiolitis, croup, and pneumonia in Alberta, Canada. *BMC Infect. Dis.*
587 16, 402 (2016).
- 588 22. Cocoros, N. M. et al. Early Release - Electronic Health Record–Based Algorithm for
589 Monitoring Respiratory Virus–Like Illness - Volume 30, Number 6—June 2024 -
590 Emerging Infectious Diseases journal - *CDC. Emerg. Infect. Dis.* 30, 1096–1103 (2024).
- 591 23. Khera, R. et al. A multicenter evaluation of computable phenotyping approaches for
592 SARS-CoV-2 infection and COVID-19 hospitalizations. *npj Digit. Med.* 5, 27 (2022).
- 593 24. Investigators, A. of U. R. P. et al. The “All of Us” Research Program. *N. Engl. J.*
594 *Med.* 381, 668–676 (2019).
- 595 25. Office, A. of U. I. R. B. O. Not Human Subjects Research Determination: Research
596 Hub: Controlled Tier Data Access Process for the All of Us Research Program. (2021).
- 597 26. Investigators, A. of U. R. P. All of Us Research Program Protocol. Preprint at
598 https://allofus.nih.gov/sites/default/files/AOU_Core_Protocol_Redacted_Dec_2021.pdf
599 (2021).

- 600 27. Kotecha, D. et al. CODE-EHR best practice framework for the use of structured
601 electronic healthcare records in clinical research. *BMJ* 378, e069048 (2022).
- 602 28. Olsen, S. J. et al. Changes in Influenza and Other Respiratory Virus Activity During
603 the COVID-19 Pandemic — United States, 2020–2021. *Morb. Mortal. Wkly. Rep.* 70,
604 1013–1019 (2021).
- 605 29. DeGroot, N. P. et al. Human parainfluenza virus circulation, United States, 2011–
606 2019. *J. Clin. Virol.* 124, 104261 (2020).
- 607 30. Shah, M. M. et al. Seasonality of Common Human Coronaviruses, United States,
608 2014–2021 - Volume 28, Number 10—October 2022 - Emerging Infectious Diseases
609 journal - *CDC. Emerg. Infect. Dis.* 28, 1970–1976 (2022).
- 610 31. Malosh, R. E., McGovern, I. & Monto, A. S. Influenza During the 2010–2020 Decade
611 in the United States: Seasonal Outbreaks and Vaccine Interventions. *Clin. Infect. Dis.*
612 76, 540–549 (2022).
- 613 32. Prevention, C. for D. C. and. Influenza (flu) weekly U.S. influenza surveillance report
614 (FluView). <https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html>.
- 615 33. Prevention, C. for D. C. and. COVID Data Tracker. [https://covid.cdc.gov/covid-data-](https://covid.cdc.gov/covid-data-tracker/)
616 [tracker/](https://covid.cdc.gov/covid-data-tracker/).
- 617 34. Higgins, T. L. et al. Assessment of the Accuracy of Using ICD-9 Diagnosis Codes to
618 Identify Pneumonia Etiology in Patients Hospitalized With Pneumonia. *JAMA Netw.*
619 *Open* 3, e207750 (2020).
- 620 35. Sivakumaran, S., Alsallakh, M. A., Lyons, R. A., Quint, J. K. & Davies, G. A.
621 Estimating the contribution of respiratory pathogens to acute exacerbations of COPD
622 using routine data. *J. Infect.* 86, 233–238 (2023).

- 623 36. Alchikh, M. et al. Are we missing respiratory viral infections in infants and children?
624 Comparison of a hospital-based quality management system with standard of care. *Clin.*
625 *Microbiol. Infect.* 25, 380.e9-380.e16 (2019).
- 626 37. Services, U. D. of H. and H. & CDC. CDC. Influenza (flu): Flu Burden.
627 <https://www.cdc.gov/flu-burden/php/about/index.html> (2024).
- 628 38. Near, A. M., Tse, J., Young-Xu, Y., Hong, D. K. & Reyes, C. M. Burden of influenza
629 hospitalization among high-risk groups in the United States. *BMC Heal. Serv. Res.* 22,
630 1209 (2022).
- 631 39. Nikolla, D. A. et al. Defining Incidental Versus Non-incident COVID-19
632 Hospitalizations. *Cureus* 16, e56546 (2024).
- 633 40. Vu, C. et al. A More Accurate Measurement of the Burden of Coronavirus Disease
634 2019 Hospitalizations. *Open Forum Infect. Dis.* 9, ofac332 (2022).
- 635 41. Menachemi, N., Dixon, B. E., Wools-Kaloustian, K. K., Yiannoutsos, C. T. &
636 Halverson, P. K. How Many SARS-CoV-2–Infected People Require Hospitalization?
637 Using Random Sample Testing to Better Inform Preparedness Efforts. *J. Public Heal.*
638 *Manag. Pr.* 27, 246–250 (2021).
- 639 42. V, C. L., Huebner, J., Hripcsak, G. & Sabatello, M. Underrepresentation of blind and
640 deaf participants in the All of Us Research Program. *Nat. Med.* 29, 2742–2747 (2023).
- 641 43. Goldstein, B. A., Bhavsar, N. A., Phelan, M. & Pencina, M. J. Controlling for
642 Informed Presence Bias Due to the Number of Health Encounters in an Electronic
643 Health Record. *Am. J. Epidemiology* 184, 847–855 (2016).
- 644 44. Perets, O. et al. Inherent Bias in Electronic Health Records: A Scoping Review of
645 Sources of Bias. *medRxiv* 2024.04.09.24305594 (2024)

646 doi:10.1101/2024.04.09.24305594.

647 45. Tang, J. W. et al. Global epidemiology of non-influenza RNA respiratory viruses:
648 data gaps and a growing need for surveillance. *Lancet Infect. Dis.* 17, e320–e326
649 (2017).