

1 **Title:** A Caveat to Using Wearable Sensor Data for COVID-19 Detection: The Role of
2 Behavioral Change after Receipt of Test Results

3
4 **Authors:** Jennifer L. Cleary,^{1,2,5} Yu Fang^{1,5}, Srijan Sen^{1,3}, Zhenke Wu*⁴

5
6 **Affiliations**

7 ¹ Michigan Neuroscience Institute, University of Michigan, Ann Arbor, MI;

8 ² Department of Psychology, University of Michigan, Ann Arbor, MI;

9 ³ Department of Psychiatry, University of Michigan Medical School;

10 ⁴ Department of Biostatistics, University of Michigan, Ann Arbor, MI;

11 ⁵ These authors contributed equally: Jennifer L. Cleary, Yu Fang.

12
13 **Corresponding author**

14 Zhenke Wu, PhD. Department of Biostatistics, University of Michigan. 1415 Washington
15 Heights, Ann Arbor, MI, 48109-2029, USA; telephone (734) 764-7067; e-mail:

16 zhenkewu@umich.edu

17
18 Word, Table, and Figure Count

19 Abstract, 147 words; main, 1489 words; 2 figures; 2 extended data figures.

20

21

22

23

24 **Abstract:**

25 Recent studies indicate that wearable sensors have the potential to capture subtle
26 within-person changes that signal SARS-CoV-2 infection. However, it remains unclear
27 the extent to which observed discriminative performance is attributable to behavioral
28 change after receiving test results. We conducted a retrospective study in a sample of
29 medical interns who received COVID-19 test results from March to December 2020. Our
30 data confirmed that sensor data were able to differentiate between symptomatic
31 COVID-19 positive and negative individuals with good accuracy (area under the curve
32 (AUC) = 0.75). However, removing post-result data substantially reduced discriminative
33 capacity (0.75 to 0.63; $\Delta = -0.12$, $p = 0.013$). Removing data in the symptomatic period
34 prior to receipt of test results did not produce similar reductions in discriminative
35 capacity. These findings suggest a meaningful proportion of the discriminative capacity
36 of wearable sensor data for SARS-CoV-2 infection may be due to behavior change after
37 receiving test results.

38

39

40

41

42

43

44

45

46

47 **Main**

48 Recent studies ¹⁻⁴ suggest enormous public health potential of wearable sensors in
49 capturing subtle within-person changes that indicate an infection, such as by SARS-
50 CoV-2. Detection of infection via wearable data provides a potentially effective, scalable
51 method of infection surveillance, through passive, non-invasive methods ⁵. However,
52 the assessments of wearable sensors for SARS-CoV-2 infection to date conflate two
53 distinct streams of information - direct physiologic effects of infection and behavioral
54 changes secondary to learning confirmation of infection through receipt of test results.
55 Understanding the relative importance of these two streams of information in infection
56 detection is critical to determining if infection surveillance may be possible through
57 wearable technology. This paper seeks to further this understanding by leveraging a
58 unique data set with individual-level dates of receiving test results that are linked to
59 wearable data collected from a cohort of symptomatic COVID-19 positive or negative
60 medical interns.

61
62 The Intern Health Study is a prospective cohort study that assesses mental health
63 during the first year of residency training ^{6,7}. Individuals starting residency in the 2019
64 and 2020 cohorts were invited to take part. Participating interns received a Fitbit Inspire
65 HR or Charge 3 device (or \$50 if they already have a Fitbit, Fitbit Inc., San Francisco,
66 CA; or an Apple Watch, Apple Inc., Cupertino, CA) and \$60 in compensation. The
67 institutional review board at the University of Michigan approved the study. From April to
68 December 2020, participants were sent multiple surveys that assessed whether they 1)
69 exhibited any symptoms consistent with COVID-19 (e.g. fever, cough, shortness of

70 breath, headache); (2) were tested for SARS-CoV-2 infection; (3) tested positive. Daily
71 sleep duration, physical activity, and resting heart rate (RHR) were measured through
72 Fitbit or Apple Watch throughout the first internship year. We focused on interns
73 because this is a population that is likely to receive tests, receive test results quickly,
74 and are more adherent to quarantine measures.

75
76 A total of 3,532 subjects participated in the 2019 and 2020 cohorts of Intern Health
77 Study. Among them, 506 subjects experienced COVID-19-like symptoms between
78 March 15 and December 2020 and of these, 379 reported being tested for SARS-CoV-
79 2. There were 94 individuals who tested positive (“cases”) and 285 individuals who
80 tested negative (“controls”). We included in the analysis 22 cases and 83 controls who
81 had step, sleep, and RHR data available for more than 50% of the days during baseline
82 (21 to 7 days prior to symptom onset) and test (0-7 days after symptom onset) periods,
83 respectively (Extended Data Figure 1). Participants were on average 28.5 +/- 2.81 years
84 of age, and 50.5% (n = 53) of the sample were female.

85
86 Our results are consistent with those reported by Quer et al. (2020)², validating the
87 value of passive wearable sensor data in differentiating symptomatic COVID-19 positive
88 from negative individuals. In particular, we followed Quer et al. (2020)² and used
89 externally-constructed metrics that operationalize within-person changes in RHR, sleep,
90 steps and all three combined. The metrics effectively contrast an individual’s wearable
91 sensor data from the test period with those from the baseline period, which are then
92 used to discriminate cases and controls. Using all the data in baseline and test periods

93 (Figure 1a), we observed that metrics of within-individual change discriminated cases
94 from controls except for RHR (Figure 2, a-d). Sleep minutes increased more among
95 cases than controls after symptom onset (mean change: 47.9 in cases, 16.6 in controls
96 $p=0.044$; area under the curve, AUC, based on SLEEPmetric = 0.66, 95% confidence
97 interval, CI = 0.51-0.80). Cases reduced physical activity more than controls after
98 symptom onset (mean change: -3,703 in cases, -1,038 in controls, $p=0.002$; AUC based
99 on STEPmetric = 0.75, 95% CI = 0.63-0.87). Mean change in RHR is higher in the
100 cases (1.3 in cases, 0.4 in controls, $p=0.18$) with the lowest discriminative ability based
101 on RHRmetric (AUC = 0.63, 95% CI = 0.48-0.79). The combined metric based on all
102 wearable sensor data results in an AUC of 0.75 (95% CI = 0.62-0.89).

103
104 To test whether the realized AUCs were mainly driven by the subset of data after receipt
105 of test results, we conducted an analysis that removed data points on and after the
106 result delivery date (Figure 1b). Compared with the previous analysis, we observed
107 worse discriminative ability (Figure 2e-2h) by SLEEPmetric (AUC = 0.60, 95% CI =
108 0.42-0.76), STEPmetric (AUC = 0.63, 95% CI = 0.49-0.78), and combined sensor
109 metrics (AUC = 0.68, 95% CI = 0.50-0.82), but similar performance in RHR (AUC =
110 0.66, 95% CI = 0.51-0.86). The AUC based on STEPmetric experienced the largest
111 decrease ($\Delta = -0.12$).

112
113 To assess whether the observed decrease in AUC is consistent with random data
114 removal or systematic information loss, we further conducted one-sided conditional
115 permutation tests for each metric (see Methods). In particular, the test assesses the null

116 that, compared to random data removal, no additional decrease in AUC is caused by
117 systematically removing data after receipt of test results. For the STEPmetric, the
118 observed decrease in AUC (step 2, Methods) stands in the left tail of the reference
119 distribution of change in AUC (step 3, Methods; observed change in AUC: -0.12,
120 $p=0.013$; Extended Data Figure 2c), indicating the observed decrease in AUC by
121 removing post-result data is unlikely a chance event from data reduction and hence the
122 importance of post-result data. Although cases and controls reduced average daily step
123 counts after they became symptomatic, the reduction was significantly more among the
124 cases after receipt of test results (mean change: cases -4,012, controls -1,016;
125 $p=0.001$) and more so than during the symptomatic period before receipt of test results
126 (mean change: cases -2,894, controls -1,083, $p=0.03$). For RHR and sleep metrics, we
127 did not observe a statistically significant decrease in the AUC after removing the post-
128 result data.

129
130 Finally, when only using the post-result-data in the test period (Figure 1c), the AUCs is
131 comparable to the all-data AUC for all metrics (Figure 2i-2l, RHRmetric: 0.62 vs 0.63;
132 SLEEPmetric: 0.63 vs 0.66; STEPmetric: 0.75 vs 0.75; all-sensor: 0.72 vs 0.75),
133 indicating no substantial loss of discriminative accuracy is incurred by only using post-
134 result data when defining the metrics. We performed conditional permutation tests as
135 above, but with the number of random days removed being the number of days prior to
136 receipt of test results. No statistically significant decrease in AUCs was observed for
137 any of the metrics (Extended Data Figure 2e-2h).

138

139 Our analysis reveals the discriminative accuracy of wearable data in COVID-19
140 detection can be explained by behavior changes after receiving test results, more so
141 driven by subjects' within-person change in physical activity, less so by sleep or RHR. In
142 particular, when removing data on and after receipt of test results, the AUC based on
143 STEPmetric drops significantly from the all-data AUC. A small though non-statistically
144 significant drop was observed for SLEEPmetric. No decrease was observed for
145 physiology-based RHRmetric. This pattern is consistent with behavior change after
146 receiving COVID-19 test results. Compared to symptomatic individuals who tested
147 negative, symptomatic individuals who received a positive COVID-19 test may initiate
148 stricter quarantine measures thus reducing activity and aim to get more sleep. It
149 appears that in the short term sleep is more resistant to change than physical activity
150 during the test period, likely strongly regulated by circadian rhythms ⁸.

151
152 This study has some limitations. First, our sample is a small subset of symptomatic
153 subjects from a sizable cohort. In future studies, it is critical to aggregate data from
154 multiple studies to further validate and study the variation in AUC with factors that may
155 impact the propensity of behavioral change. Second, the cohort is likely not
156 representative of the entire spectrum of population that may have access to both
157 wearables and tests. However, the unique cohort of medical interns who are likely more
158 adherent to quarantine measures strengthened the specific investigation addressed
159 here. It is of interest to investigate the same question in a broader population. Third, the
160 SARS-CoV-2 tests are not perfectly sensitive or specific. Knowledge about these test-
161 related parameters will likely further improve AUC estimates. Fourth, recall of symptom

162 onset date and test date might not be entirely accurate, but this population of medical
163 interns is particularly primed to remember the dates due to workplace enforcements of
164 symptom screening, testing, and compulsory quarantines.

165
166 In a future pandemic, passively-collected wearable data linked with test results may
167 reveal distinct patterns of behavioral change across subpopulations. For example, lack
168 of appropriate behavioral changes upon receiving test results may hurt discriminative
169 accuracy based on wearable sensor data. Variation in the AUC of the step metric by
170 age group may indicate differential levels of within-person change in activity. Groups
171 with higher step-based AUC may have effectively quarantined after receiving their test
172 results; while groups with lower step-based AUC may indicate either delay in their
173 receiving the test results or difficulty and infeasibility in reducing physical activity.
174 Subpopulations with lower observed AUCs may benefit from more targeted public
175 health policy innovations that may promote behavioral change, such as self-quarantine
176 measures.

177

178 **Acknowledgements**

179 This study was supported by grants from the National Institute of Mental Health (R01
180 MH101459) to S.S. and Z.W. and an investigator grant from Precision Health Initiative
181 at University of Michigan, Ann Arbor to Z.W. and S.S.. J.C. was supported by
182 T32HD007109. We thank the interns and residency programs who took part in this
183 study.

184

185 **Author Contributions**

186 J.C., Y.F., S.S., Z.W. made substantial contributions to the study conception and
187 design. S.S. made substantial contributions to the acquisition of data. J.C., Y.F., Z.W.
188 conducted statistical analysis. J.C., Y.F., S.S., Z.W. made substantial contributions to
189 the interpretation of data. J.C., Y.F., Z.W. drafted the first version of the manuscript.
190 J.C., Y.F., S.S., Z.W. contributed to critical revisions and approved the final version of
191 the manuscript. J.C., Y.F., S.S., Z.W. take responsibility for the integrity of the work.

192

193 **Competing Interests Statement**

194 The authors have no competing interests to disclose.

195

196 **Data and code availability**

197 Deidentified data supporting the results and figures in this manuscript are available
198 upon reasonable request and completion of a data agreement with the Intern Health
199 Study team. Code for data preprocessing and statistical analysis is available upon
200 request.

201

202 **References**

- 203 1. Natarajan, A., Su, H.-W. & Heneghan, C. Assessment of physiological signs
204 associated with COVID-19 measured using wearable devices. *npj Digital Medicine*
205 **3**, 1–8 (2020).
- 206 2. Quer, G. *et al.* Wearable sensor data and self-reported symptoms for COVID-19

- 207 detection. *Nat. Med.* **27**, 73–77 (2020).
- 208 3. Ceren Ates, H., Yetisen, A. K., Güder, F. & Dincer, C. Wearable devices for the
209 detection of COVID-19. *Nature Electronics* **4**, 13–14 (2021).
- 210 4. Mishra, T. *et al.* Pre-symptomatic detection of COVID-19 from smartwatch data.
211 *Nat Biomed Eng* **4**, 1208–1220 (2020).
- 212 5. Radin, J. M., Wineinger, N. E., Topol, E. J. & Steinhubl, S. R. Harnessing wearable
213 device data to improve state-level real-time surveillance of influenza-like illness in
214 the USA: a population-based study. *Lancet Digit Health* **2**, e85–e93 (2020).
- 215 6. Sen, S. *et al.* A prospective cohort study investigating factors associated with
216 depression during medical internship. *Arch. Gen. Psychiatry* **67**, 557–565 (2010).
- 217 7. NeCamp, T. *et al.* Assessing Real-Time Moderation for Developing Adaptive Mobile
218 Health Interventions for Medical Interns: Micro-Randomized Trial. *J. Med. Internet*
219 *Res.* **22**, e15033 (2020).
- 220 8. Fang, Y., Forger, D. B., Frank, E., Sen, S. & Goldstein, C. Day-to-day variability in
221 sleep parameters and depression risk: a prospective cohort study of training
222 physicians. *npj Digital Medicine* **4**, 1–9 (2021).
- 223 9. The R Project for Statistical Computing. <https://www.R-project.org/>.

224

225

226

227

228

229

230 **Methods**

231

232 *Metrics Definition.* Participants were drawn from the 2019 and 2020 cohorts of the Intern
233 Health Study. Study recruitment and procedures are detailed elsewhere.⁸ Briefly,
234 incoming first-year medical residents were surveyed throughout the pandemic from April
235 to December 2020 and asked to report whether and when they experienced any
236 potential COVID-19 symptoms, were tested, and their test results. The sample for this
237 analysis included individuals who reported symptoms and a COVID-19 test, as well as
238 at least 50% of the wearable data (collected through Fitbit or Apple Watch) during both
239 baseline (21 to 7 days prior to symptom onset) and test (0 to 7 days after symptom
240 onset) periods.

241

242 Following Quer et al. (2020)², we calculated metrics for sleep, activity, and resting heart
243 rate (RHR), as well as an overall wearable sensor metric for each participant:

244

245 **RHRmetric** = $\max(\text{dailyRHR}[\text{test}]) - \text{median}(\text{dailyRHR}[\text{baseline}]) / \text{IQR}$

246 **SLEEPmetric** = $\text{mean}(\text{dailySLEEP}[\text{test}]) - \text{median}(\text{dailySLEEP}[\text{baseline}]) / \text{IQR}$

247 **STEPmetric** = $\text{mean}(\text{dailySTEP}[\text{test}]) - \text{median}(\text{dailySTEP}[\text{baseline}]) / \text{IQR}$

248 **SENSORmetric** = $\text{RHRmetric} / 10 + \text{SLEEPmetric} - \text{STEPmetric}$

249 *Discriminative Accuracy.* We calculated ROC curves, AUC, sensitivity (SE), specificity
250 (SP) for each metric to compare the intra-individual change in each metric with

251 symptom onset between COVID-19 positive and COVID-19 negative individuals. To
252 assess which part of the test period data is mainly responsible for the realized AUC, we
253 calculated these parameters in three data schemes: Scheme I - using all the data in
254 baseline and test periods; Scheme II - removing data on and after receipt of test results
255 in test periods; Scheme III - removing data before receipt of test results in test periods.

256

257 *Conditional Permutation Tests.* In order to test the statistical significance of the
258 observed AUC decrease in Scheme II and III, we designed the one-sided conditional
259 permutation tests in a way that breaks the link between the indices of days removed
260 during the test period and the dates of receiving the test results hence creating a null
261 distribution that is adequate for assessing the statistical significance of the observed
262 change in AUC. In particular, for each metric (RHR, sleep, activity, sensor) we perform
263 the following steps:

264

265 Step 1. Calculate AUC based on all the baseline and test data;

266 Step 2. Remove part of the test data (on/after receiving the test results as in
267 Figure 1b; OR before receiving the test results as in Figure 1c), and calculate a
268 single AUC and the change from the AUC in Step 1;

269 Step 3. Create $B=1000$ data sets, each by randomly removing the same amount
270 of data for each person as in Step 2; based on each of B random reduced data
271 sets, calculate an AUC and the difference from the AUC in Step 1, resulting in
272 $B=1000$ values of change in AUC;

273 Step 4. Compare the change of AUC in Step 2 against the null distribution of the
274 change of AUCs in Step 3; Calculate the p-value by the observed fraction among
275 the 1000 randomly reduced data sets that have AUC change less than or equal
276 to the observed change in Step 2.

277

278 All analyses were conducted using R 4.0.2⁹.

279

280

281

282

283

284

285

286

287

288

289

290

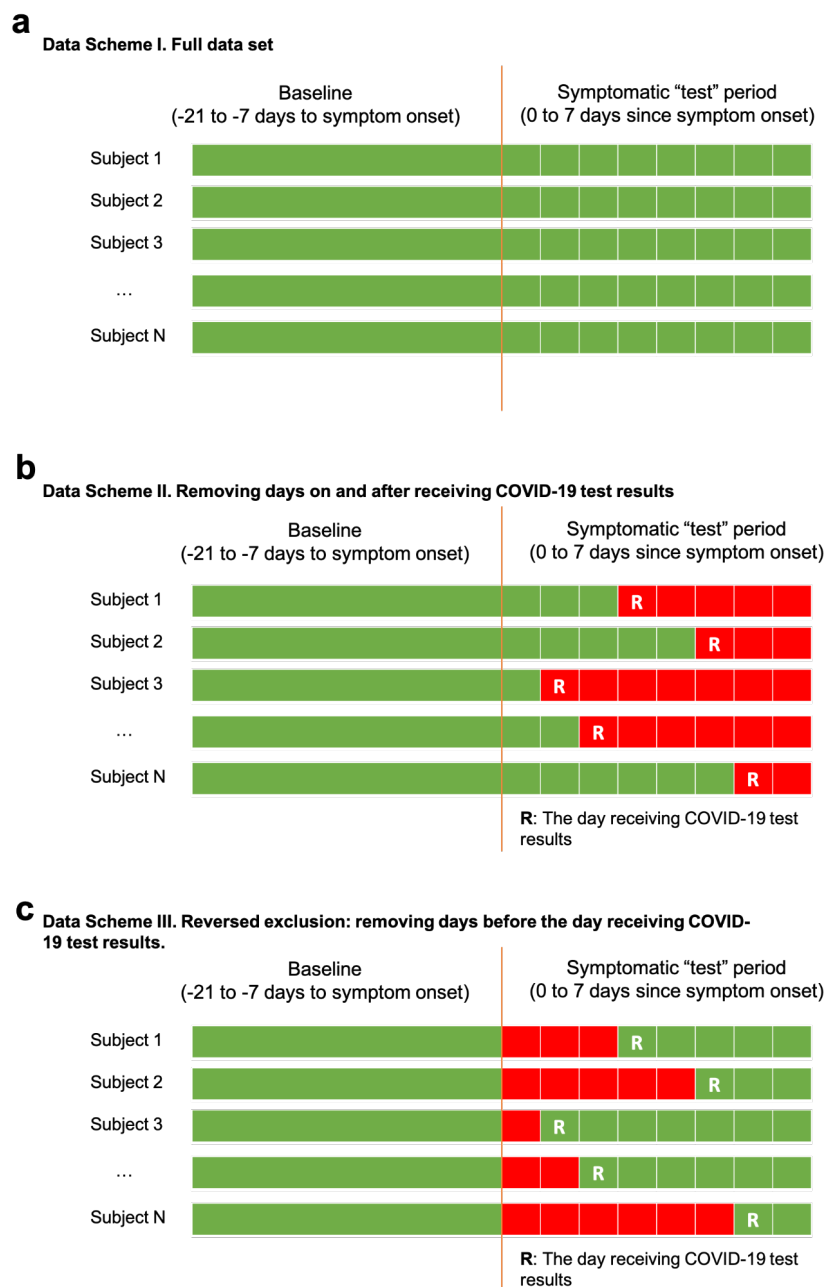
291

292

293

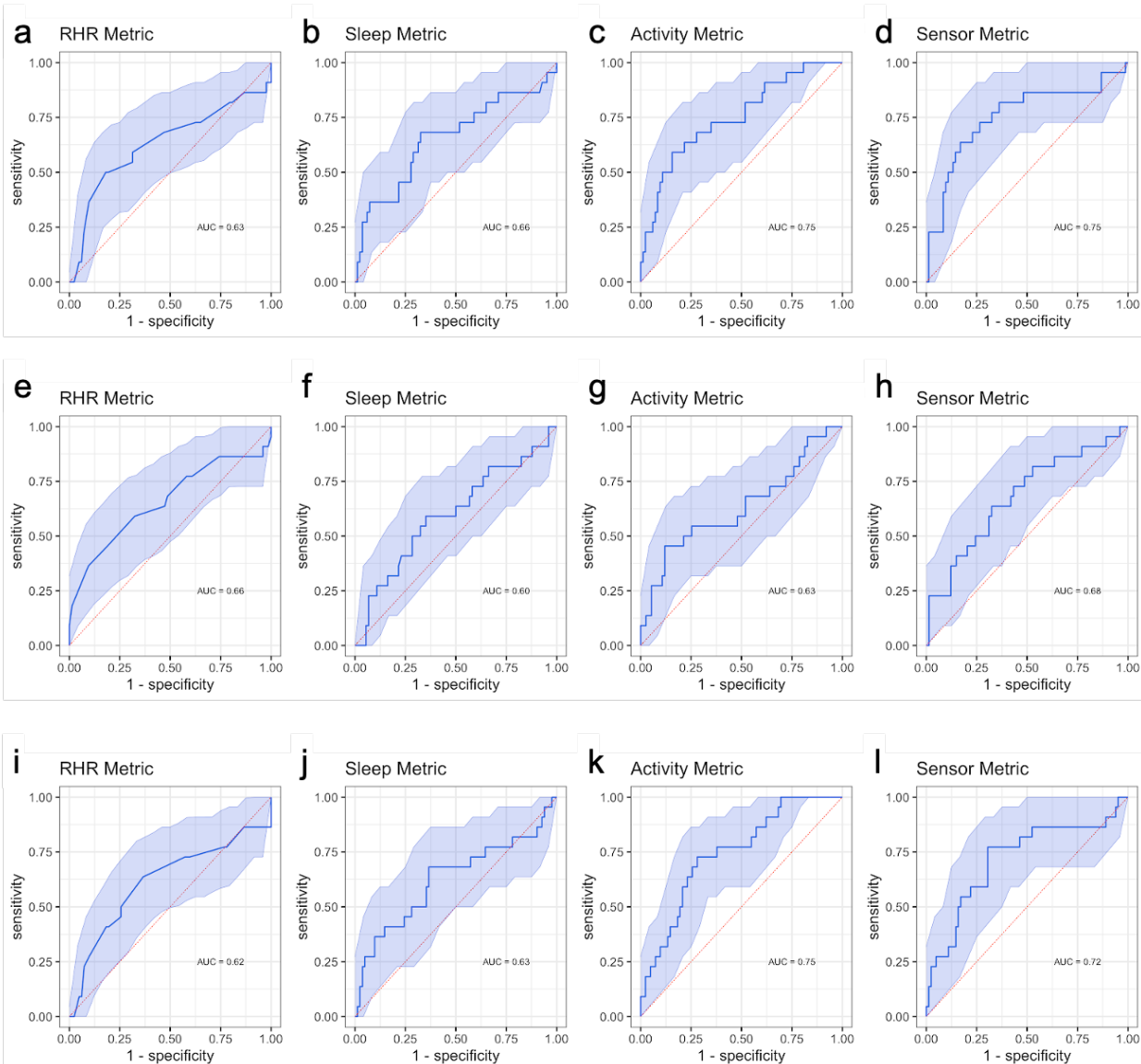
294

295



296

297 **Figure 1. Data Schemes. a-c**, green: included data; red: excluded data; R: the day
 298 receiving test results: (a) include all data; (b) exclude data on and after the day
 299 receiving test results; (c) exclude data before the day receiving test results since
 300 symptom onset. Ninety-two subjects (87.6%) received their results within the
 301 symptomatic period (0 to 7 days after symptom onset).

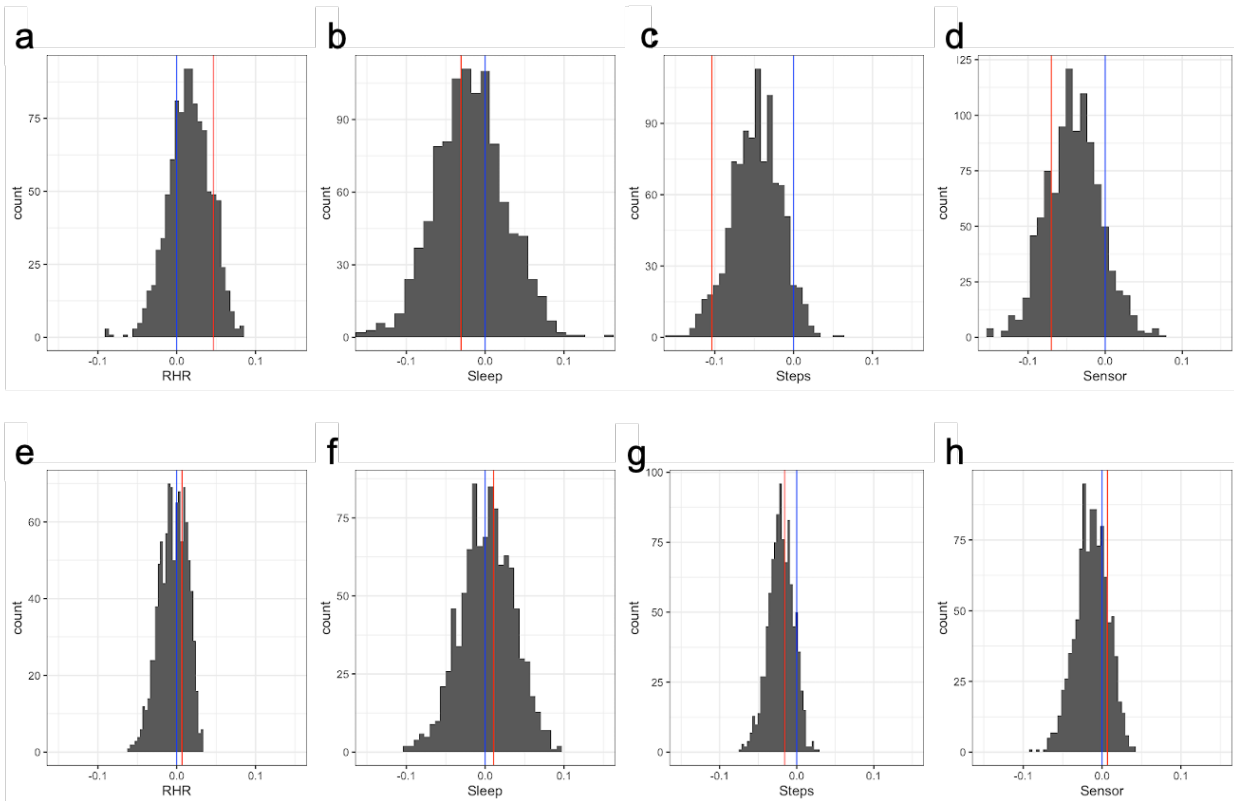


302
303

304 **Figure 2. AUCs based on RHR, sleep, activity and all-sensor metric derived from**
305 **wearable sensors to differentiate symptomatic subjects who were tested positive**
306 **and negative corresponding to data schemes I-III. (a-d): Scheme I - all data; (e-h)**
307 **Scheme II - remove data on and after knowing the test results; (i-l): Scheme III -**
308 **remove data since symptom onset and before the test results. For each data scheme in**
309 **the row, the four panels are for RHR, sleep, activity and sensor metrics, respectively.**

	Including days after the test result (Scheme I)			Excluding days after the test result (Scheme II)			Excluding days before the test result (Scheme III)		
	Test positive	Test negative	P value	Test positive	Test negative	P value	Test positive	Test negative	P value
Demographics									
Number of participants, N	22	83	-	22	75	-	22	82	-
Age, Mean (SD)	28.1 (3.2)	28.7 (2.7)	0.49	28.1 (3.2)	28.7 (2.7)	0.46	28.1 (3.2)	28.6 (2.7)	0.54
Female, N (%)	7 (31.8%)	46 (55.4%)	0.06	7 (31.8%)	41 (54.7%)	0.09	7 (31.8%)	45 (54.9%)	
Fitbit users, N (%)	22 (100%)	82 (98.8%)	-	22 (100%)	74 (98.7%)	-	22 (100%)	81 (98.8%)	-
AppleWatch users, N (%)	0 (0%)	1 (1.2%)	-	0 (0%)	1 (1.3%)	-	0 (0%)	1 (1.2%)	-
Available days during baseline (IQR)									
RHR	13.4 (11.5-15)	13.7(13-15)	0.55	13.4 (11.5-15)	13.6 (13-15)	0.66	13.4 (11.5-15)	13.7 (13-15)	0.54
Sleep	14.3 (14-15)	13.9(13.5-15)	0.3	14.3 (14-15)	13.9 (13-15)	0.24	14.3 (14-15)	13.9 (13-15)	0.3
Activity	14.9 (15-15)	14.7(15-15)	0.16	14.9 (15-15)	14.7 (15-15)	0.14	14.9 (15-15)	14.7 (15-15)	0.16
Available days during symptomatic period (IQR)									
RHR	7.1 (7.0-8)	7.3 (7-8)	0.66	3.8 (2-6.75)	3.0 (1-4)	0.21	5.1 (3.25-7)	5.2 (4-7)	0.9
Sleep	7.6 (7.3-8)	7.4 (7-8)	0.12	3.9 (2-6.75)	3.0 (1-4)	0.15	5.5 (4-7)	5.3 (4-7)	0.57
Activity	7.9 (8-8)	7.9 (8-8)	0.89	4.0 (2-6.75)	3.1 (1-4)	0.21	5.7 (4-7)	5.7 (4-7)	0.93
Baseline mean (SD)									
RHR (bpm)	60.2 (5.6)	65.3 (7.4)	<0.001	60.2 (5.6)	65.6 (7.4)	<0.001	60.2 (5.6)	65.3 (7.4)	<0.001
Sleep (min)	412 (38)	411 (40)	0.94	412 (38)	411 (41)	0.91	412 (38)	412 (40)	1
Activity (steps)	8650 (2808)	8382 (2366)	0.68	8650 (2808)	8378 (2305)	0.68	8650 (2808)	8400 (2375)	0.7
Symptomatic period mean (SD)									
RHR (bpm)	61.5 (5.6)	65.7 (7.7)	0.007	62.1 (5.5)	66.2 (7.5)	0.007	61.3 (5.7)	65.6 (7.8)	0.005
Sleep (min)	460 (67)	428 (50)	0.046	432 (83)	411 (85)	0.32	469 (75)	433 (57)	0.049
Activity (steps)	4948 (2223)	7344 (2710)	<0.001	5756 (2411)	7295 (3164)	0.02	4638 (2494)	7384 (2902)	<0.001
Mean change (SD)									
RHR (bpm)	1.3 (3.1)	0.4 (2.3)	0.18	1.9 (3.2)	0.5 (2.5)	0.07	1.1 (3.4)	0.3 (2.4)	0.3
Sleep (min)	47.9 (64.8)	16.6 (48.2)	0.044	20 (78)	0.5 (91)	0.33	56.9 (74)	21.4 (49.2)	0.043
Activity (steps)	-3703 (3422)	-1038 (2333)	0.002	-2894 (3327)	-1083 (3140)	0.03	-4012 (3632)	-1016 (2599)	0.001
Metric value (SD)									
RHR	0.36 (0.35)	0.26 (0.27)	0.2	0.26 (0.39)	0.14 (0.25)	0.18	0.33 (0.37)	0.28 (0.27)	0.24
Sleep	0.45 (0.60)	0.19 (0.41)	0.06	0.44 (0.95)	0.24 (0.53)	0.36	0.40 (0.62)	0.17 (0.55)	0.12
Activity	-0.75 (0.69)	-0.18 (0.48)	0.001	-0.63 (0.79)	-0.17 (0.62)	0.02	-0.72 (0.68)	-0.18 (0.51)	0.002
Sensor Metric	1.24 (1.1)	0.40 (0.72)	0.002	1.11 (1.43)	0.42 (0.88)	0.04	1.16 (1.03)	0.37 (0.85)	0.003

Extended Data Figure 1. Summary of key characteristics, metrics and COVID-19 test results.



Extended Data Figure 2. One-sided conditional permutation test for assessing the null that, compared to random data reduction, no additional change in AUC caused by removing data: (a-d) on or after receipt of test result and (e-h) in the symptomatic period and prior to receipt of test results. The random data removal and AUC calculation are done for RHR, sleep, step and the all-sensor data, respectively (shown in four panels in each row). In each panel, the red line indicates the observed change of AUC; the blue line is at zero, indicating no change. The reference distributions are not centered at zero despite data removal being random; because on average there are more days after the receipt of test results than before, random data removal may still impact AUCs. For each metric, if a red line is at the left tail of the histogram, we conclude a statistically significant additional decrease in AUC.