

1 **An infection prediction model developed from inpatient data can predict out-of-hospital**
2 **COVID-19 infections from wearable data when controlled for dataset shift**

3 Ting Feng^{1,*}, Sara Mariani¹, Bryan Conroy¹, Robert Damiano¹, Ikaro Silva¹, Dennis

4 Swearingen^{2,3}, Daniel C. McFarlane

5 ¹ Philips North America, Cambridge MA, USA

6 ²Department of Medical Informatics, Banner health, Phoenix AZ, USA

7 ³Department of Biomedical Informatics, University of Arizona College of Medicine, Phoenix

8 AZ, USA

9 * Corresponding author

10 Corresponding author email: ting.feng@philips.com

11 **ABSTRACT**

12 The COVID-19 pandemic highlighted the importance of early detection of illness and the need
13 for health monitoring solutions outside of the hospital setting. We have previously demonstrated
14 a real-time system to identify COVID-19 infection before diagnostic testing ¹, that was powered
15 by commercial-off-the-shelf wearables and machine learning models trained with wearable
16 physiological data from COVID-19 cases outside of hospitals. However, these types of solutions
17 were not readily available at the onset nor during the early outbreak of a new infectious disease
18 when preventing infection transmission was critical, due to a lack of pathogen-specific illness
19 data to train the machine learning models. This study investigated whether a pretrained clinical
20 decision support algorithm for predicting hospital-acquired infection (predating COVID-19)
21 could be readily adapted to detect early signs of COVID-19 infection from wearable
22 physiological signals collected in an unconstrained out-of-hospital setting. A baseline
23 comparison where the pretrained model was applied directly to the wearable physiological data
24 resulted a performance of AUROC = 0.52 in predicting COVID-19 infection. After controlling
25 for contextual effects and applying an unsupervised dataset shift transformation derived from a
26 small set of wearable data from healthy individuals, we found that the model performance
27 improved, achieving an AUROC of 0.74, and it detected COVID-19 infection on average 2 days
28 prior to diagnostic testing. Our results suggest that it is possible to deploy a wearable
29 physiological monitoring system with an infection prediction model pretrained from inpatient
30 data, to readily detect out-of-hospital illness at the emergence of a new infectious disease
31 outbreak.

32 **KEYWORDS:** Infection Prediction, Wearable Physiological Monitoring, Clinical Decision
33 Support (CDS), Dataset Shift, Machine Learning, COVID-19 infection, Infectious Disease,
34 Public Health

35 INTRODUCTION

36 The COVID-19 pandemic highlighted the importance of early disease detection and isolation in
37 order to prevent the spread of infection ²⁻⁴. It is desirable, therefore, to have an effective system
38 to continuously monitor an individual's health state. Health monitoring systems consisting of
39 wearable devices and artificial intelligence (AI) tools are portable, minimally invasive, and were
40 shown to be able to detect COVID-19 infections ^{1,5-13}. For example, we developed a real-time
41 infection prediction system using commercial-off-the-shelf (COTS) wearable devices and AI,
42 which was capable of identifying COVID-19 infection on average 2.3 days before diagnostic
43 testing with an Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.82 ¹.
44 Two other studies reported comparable performance from wearable physiological monitoring
45 with AUROC=0.80 ⁵ and AUROC=0.77 ¹⁴, respectively.

46 These health monitoring systems are typically powered by machine learning (ML) models ^{1,5-7,14}
47 or statistical methods ¹⁰ that are sensitive to physiological changes caused by COVID-19
48 infection. The models gain intelligence through supervised learning on physiological data
49 collected from the target populations of COVID-19 infection cases. However, training these
50 models require data from a significant number of COVID-19 positive cases, which is challenging
51 because infection data collection is time consuming and costly. Additional challenges of data
52 collection include user compliance, physiologic context effects (such as traveling, intense
53 exercises, etc.), and uncertainties in the timing of infection onset. These models cannot therefore
54 be easily developed when they are most needed, such as at the beginning of novel infection
55 outbreaks like the COVID-19 pandemic.

56 To this end, we propose that clinical decision support algorithms developed from data collected
57 in hospitals can be utilized to significantly accelerate or provide a minimum viable starting point
58 for wearable systems to monitor for infections in unconstrained, real-world environments. We
59 previously developed a machine learning model that can identify hospital-acquired infection
60 (HAI) patients up to 48 hours before clinical suspicion of infection. The model used
61 physiological measurements from hospital grade devices and demographic information collected
62 in the hospital ¹⁵. Here, we hypothesized that such infection prediction algorithms trained from
63 hospital dataset (referred to in this article as the “hospital model”) can predict COVID-19
64 infection from the same set of physiological measurements collected through wearables outside
65 of hospitals, provided that dataset shift ¹⁶ – the changes in the joint distribution of the
66 physiological features and the infectious disease labels between hospital and wearable datasets –
67 are properly addressed. More specifically, if we define our physiological input features as X and
68 our infectious disease labels as Y , we can typically have three types of dataset shifts:

- 69 1. Covariate shift: $P(X)$ changes but $P(Y|X)$ and $P(Y)$ remain the same.
- 70 2. Label shift: $P(Y)$ changes but $P(Y|X)$ and $P(X)$ remain the same.
- 71 3. Concept drift: $P(Y|X)$ changes but $P(X)$ and $P(Y)$ remain the same.

72 where $P(X)$, $P(Y)$ and $P(Y|X)$ are the probability distribution of X , probability distribution of Y ,
73 and the conditional probability distribution of Y given X , respectively.

74 In this study, we first performed retrospective analyses to identify sources of dataset shift
75 between hospital dataset and wearable dataset, and then described two correction techniques –
76 removing contextual confounders and applying a monotonic feature transformation – to reduce
77 the differences in data distribution between the two datasets. We found that our infection

78 prediction model trained from the hospital dataset performed best after applying both correction
79 techniques, with an AUROC of 0.74, and detection of COVID-19 infection on average 2 days
80 prior to diagnostic testing. Only a small sample of wearable data from healthy subjects (2 weeks
81 of data from 25 healthy subjects) was required for the feature transformation. Our results suggest
82 that a minimum viable wearable physiological monitoring system that detects early signs of
83 COVID-19 infection can be developed and deployed without the need for data from COVID-19
84 cases.

85 **METHODS**

86 *Description of datasets*

87 The two hospital datasets - MIMIC-III (Medical Information Mart for Intensive Care III) ¹⁷ and
88 Banner Health data - used to train the infection prediction model were described previously ¹⁵.
89 The two datasets were combined in this study to create a single hospital dataset to train the
90 infection prediction model. Both MIMIC-III and Banner Health data comprise de-identified
91 health-related data from patients during their hospital stay. The MIMIC-III data we used was
92 from patients who stayed in critical care units of the Beth Israel Deaconess Medical Center
93 (Boston, MA) between 2001 and 2012. Each patient encounter included in this study was from
94 the MIMIC-III Waveform Database Matched Subset ¹⁸. The Banner Health data was from
95 patients who stayed in critical care units or low-acuity settings such as general wards in Banner
96 Health hospitals (Phoenix, AZ). The patient cohort included in this study was collected between
97 2016 and 2017, where waveform records were available for a subset of the patient encounters.

98 The wearable dataset used to test hospital model's ability to detect COVID-19 infection was
99 collected in the framework of a study described previously ¹ and an extension of the study which
100 focused on algorithm improvement and augmentation. This dataset comprises de-identified
101 COTS wearable physiological and activity data from Garmin watch and Oura ring devices,
102 collected from active military personnel recruited from multiple US Department of Defense
103 (DoD) sites between June 2020 and May 2022. This dataset also included symptoms and
104 diagnostic tests information from self-reported daily survey questionnaires.

105 ***Ethical approval***

106 The MIMIC-III project was approved by the Institutional Review Boards of Beth Israel
107 Deaconess Medical Center and the Massachusetts Institute of Technology (Cambridge, MA).
108 The use of Banner Health data was a part of a retrospective deterioration detection study
109 approved by the Institutional Review Board of Banner Health and by the Philips Internal
110 Committee for Biomedical Experiments. For both hospital datasets, requirements for individual
111 subject consent were waived because the project did not impact clinical care, was no greater than
112 minimal risk, and all protected health information was removed from the limited dataset used in
113 this study.

114 The collection and use of the wearable dataset was approved by the Institutional Review Boards
115 of the US Department of Defense. Informed consent was obtained from all participants.

116 ***Cohort selection***

117 Patient encounters used in this study to train the hospital-acquired infection prediction model
118 were selected using the same methodology described previously ¹⁵, namely a set of MIMIC-III
119 and Banner Health patient encounters that had high-sampling frequency waveform recordings
120 around the time of clinical suspicion of hospital-acquired infection. These data were acquired
121 prior to the COVID-19 outbreak therefore did not include instances of COVID-19 infections. We
122 focused on patient encounters with waveform recordings, because we wanted to match the
123 temporal resolution of the vital sign measurements from which the hospital model was trained,
124 with the temporal resolution of the vital sign measurements in the wearable dataset to which the
125 hospital model would be applied. The infection patients, as described previously, were those who
126 had confirmed infection diagnoses and whose timing of clinical suspicion of infection could be
127 localized by a microbiology culture test order. Note that we used as our reference the time when
128 the microbiology culture test was ordered, not the time when the test result was returned. These
129 infection patients were further screened into a hospital-acquired infection cohort if the earliest
130 timing of the microbiology culture test order occurred at least 48 hours after hospital admission.

131 Subjects used to validate the performance of the hospital model in predicting COVID-19
132 infection were extracted from the wearable dataset, as described previously ¹. Specifically,
133 COVID-19 positive subjects were those who reported positive test results and symptoms, and
134 COVID-19 negative subjects were those who reported at least 1 symptom-free negative test
135 result, but no positive results. Condition for inclusion in both classes was the presence of data
136 from a Garmin watch and an Oura ring simultaneously, and that at least 10 nights of physiologic
137 data were collected during sleep within the 21-day period prior to their COVID-19 test (subjects
138 were excluded post-hoc if they did not meet these criteria).

139 *Feature extraction*

140 The trained hospital model in this study used features derived from a subset of the demographics
141 and vital sign measurements described previously¹⁵. We chose this subset because the same set
142 of demographics and vital sign measurements were available and reliable in the wearable dataset.
143 Specifically, the feature vector for training was composed of demographics (age, sex) and four
144 statistic features - average, minimum, maximum and the standard deviation – of core body
145 temperature, respiratory rate, heart rate, and RMSSD (Root Mean Square of Successive
146 Differences between normal heartbeats – a standard measure of heart rate variability), collected
147 in a 24-hour observation window prior to the observation time of 1-hour before clinical suspicion
148 of infection. This resulted in a total set of 18 features in the feature vector. We required the
149 feature vector to contain no missing values, and thus excluded patient encounters that had one or
150 more types of vital sign measurements missing in the observation window. The majority of the
151 vital sign measurements, except for temperature which was sporadically measured at the bedside,
152 were derived from high temporal resolution waveforms and matched to the temporal resolution
153 of the corresponding measurements provided by a Garmin watch and an Oura ring. In particular,
154 heart rate and RMSSD were calculated after extracting inter-beat interval from
155 photoplethysmography (PPG), and respiratory rate was derived from impedance-based
156 measurements.

157 The same set of demographics and vital sign features were extracted from the wearable dataset.
158 The Oura rings provided skin temperature and RMSSD measurements. Respiratory rate was
159 measured from the Garmin watches. Concurrent heart rate measurements from the Garmin watch

160 and Oura ring were combined before feature extraction. Plausibility filters were applied so that
161 unrealistic values outside of a very broad physiological range were discarded¹. For each subject,
162 we extracted statistic features in 24-hour intervals within a 14-day window prior to their COVID-
163 19 test (hence 14 observation times). Statistic features were derived from measurements
164 collected in a 24-hour observation window prior to the observation time, similar to those used to
165 train the hospital model. We extracted more than one day of features because we wanted to
166 assess how early our model could detect COVID-19 infection prior to diagnostic testing. To
167 examine the impact of daytime activity and other contextual factors on physiology, we extracted
168 two sets of features: the first set used all vital sign measurements collected in the 24-hour
169 observation window (“daily features”), and the second set used vital sign measurements
170 collected during sleep in the 24-hour observation window (“sleep-only features”). Hypnogram
171 information from the wearable devices were used to identify the sleep segments where the sleep-
172 only features were extracted.

173 ***Hospital model training***

174 The model for hospital-acquired infection prediction was trained using the same methodology
175 described previously¹⁵. Specifically, we used the XGBoost algorithm¹⁹ to train and test the
176 hospital model with 5-fold cross-validation. Hyperparameters were optimized using grid search.
177 The set of hyperparameter that yielded the best model performance averaged from the 5
178 validation folds were used to train the final model for assessing its performance in the wearable
179 dataset.

180 ***Testing the hospital model in the wearable dataset***

181 To assess the performance of the hospital model in predicting COVID-19 infection, we defined a
182 true positive as being a positive model prediction within the 14-day period prior to a positive
183 COVID-19 test for the positive class, and a true negative as being a negative model prediction
184 within the 14-day period prior to a negative COVID-19 test for the negative class. Because
185 infection risk scores from the model were calculated in 24-hour intervals within a 14-day period,
186 a positive model prediction was defined as one with at least one prediction within the 14-day
187 period above the defined risk threshold, and a negative model prediction was defined as one with
188 all predictions within the 14-day period below the defined risk threshold. In other words, we
189 computed the hospital model outputs - which were probabilistic scores that estimated the
190 likelihood of a given subject being infected – from the demographics and vital sign features for
191 each day (or each sleep segment) and took the maximum score during the 14-day window for
192 each subject. We then compared the maximum scores between COVID-19 positive and negative
193 subjects and reported the model performance using the following metrics:

- 194 • Area under the Receiver Operating Characteristic curve (AUROC),
- 195 • Average Precision (AP),
- 196 • True Negative Rate (Specificity),
- 197 • True Positive Rate (Sensitivity, or Recall), including:
 - 198 ○ Sensitivity(Break-Even): Sensitivity at the break-even point, where Sensitivity
199 and Precision are equal,
 - 200 ○ Sensitivity(80%): Sensitivity when Specificity=0.8,
 - 201 ○ Sensitivity(90%): Sensitivity when Specificity=0.9.

202 The significance of an AUROC value was assessed by performing a permutation test. The class
203 labels were randomly permuted 1000 times to estimate the empirical distribution of a “random”
204 AUROC. The observed AUROC value was then compared with this bootstrapped empirical
205 distribution to calculate the p-value.

206 To estimate the overall lead time of positive classification, we identified the days (interpolated)
207 in which the hospital model prediction exceeded a predefined threshold of sensitivity = 0.6
208 within the 14-day window prior to COVID-19 testing. The threshold was suggested by the study
209 principal investigators in the US DoD sites. The lead time was then defined as the average across
210 these positive days for each user and then aggregated across the cohort for the final mean
211 estimate of the lead time for COVID-19 classification (False Negatives have lead time of 0
212 days). We also overlaid risk scores with time to have a visual representation of risk score
213 elevation during the infection period.

214 To reduce the impact of dataset shift on model performance, we performed a monotonic feature
215 transformation by first calculating percentile values of each feature in the hospital dataset and the
216 wearable dataset respectively, and then replacing wearable feature values with the hospital
217 feature values that shared the same percentile. The percentile values of a given feature were
218 calculated in each dataset using all samples without distinguishing between positive and negative
219 class labels. This way, we calibrated features from wearables to match the distribution in the
220 hospital dataset without knowledge of the class labels. We then validated the performance of the
221 hospital model on the calibrated wearable features and compared it with model performance on
222 wearable features before feature transformation. To understand the data requirements for feature

223 transformation, we performed additional benchmarking experiments with restrictions on the type
224 and size of wearable data used for feature transformation, including: using wearable data
225 acquired when the subjects were not under impact of COVID-19 infection; using wearable data
226 from subjects that were not used to test the model performance; using the most recent days of
227 wearable data prior to diagnostic testing; and using wearable data from randomly down-sampled
228 cohorts or subject days (without replacement, 10 iterations).

229 **RESULTS**

230 *Cohorts and Features for Training and Testing*

231 The cohort selection criteria for training the hospital model resulted in a total dataset size of
232 9,517 patient encounters with waveform recordings around the time of clinical suspicion of
233 hospital-acquired infections (not including COVID-19). Of these patient encounters, 3,951
234 (3,665 controls and 286 HAIs; 51% Banner Health and 49% MIMIC-III) had overlapping PPG
235 waveforms and impedance-based measurements with good data quality, and therefore had the
236 full set of 18 demographics and vital sign features (see METHODS) available at 1-hour before
237 clinical suspicion of infection. These 3,951 patient encounters were used to train the hospital
238 model of hospital-acquired infection prediction.

239 The cohort selection criteria for testing the trained hospital model resulted in 301 COVID-19
240 positive subjects and 2,111 COVID-19 negative subjects from the wearable dataset. Within the
241 14-day windows prior to COVID-19 tests from these subjects, a total of 33,164 subject days and
242 31,269 subject sleep segments had vital sign measurements that passed our plausibility filter.
243 From these subject days, we extracted the feature vectors comprising the same set of 18 features

244 that was used to train the hospital model, using either all available vital sign measurements or
245 those measured during sleep (“daily features” and “sleep-only features”, see METHODS), to
246 quantify the performance of the trained hospital model in predicting COVID-19 infections.

247 *Differences between training and testing datasets*

248 The joint distribution of inputs and outputs of the infection prediction model differed between
249 the training scenario in the hospital dataset and the testing scenario in the wearable dataset – a
250 problem known as “dataset shift”¹⁶. Here we describe five sources of dataset shift in our study.

251 First, the demographics of the training and testing cohorts were different. The patients from the
252 hospital dataset were older than the subjects from the wearable dataset (Figure 1A), and the
253 wearable dataset had an imbalanced sex ratio than the hospital dataset (Figure 1B, 20% female in
254 the wearable dataset versus 47% female in the hospital dataset). Both age and sex may result in
255 differences in physiology^{20–30}.

256 Second, the health states of the training and testing cohorts were different. Patients in the
257 hospital dataset are those who developed hospital-acquired infections during their stays in
258 general wards or in some cases intensive care units, and are likely older adults with
259 comorbidities and under medical treatments, therefore the physiological measurements in the
260 hospital dataset were more likely to be abnormal and unstable compared to the physiological
261 measurements in the wearable dataset where healthy young military personnel performing their
262 daily duty were monitored. We found that patients in the hospital dataset had higher heart rate
263 and higher respiratory rate than the subjects in the wearable dataset (see the Average and

264 Maximum statistic feature in Supplementary Table 1), which were consistent with an overall
265 declined health state³⁰⁻³³. The hospital patients also had larger variations in heart rate and
266 respiratory rate than the subjects in the wearable dataset (see the Standard Deviation statistic
267 feature in Supplementary Table 1).

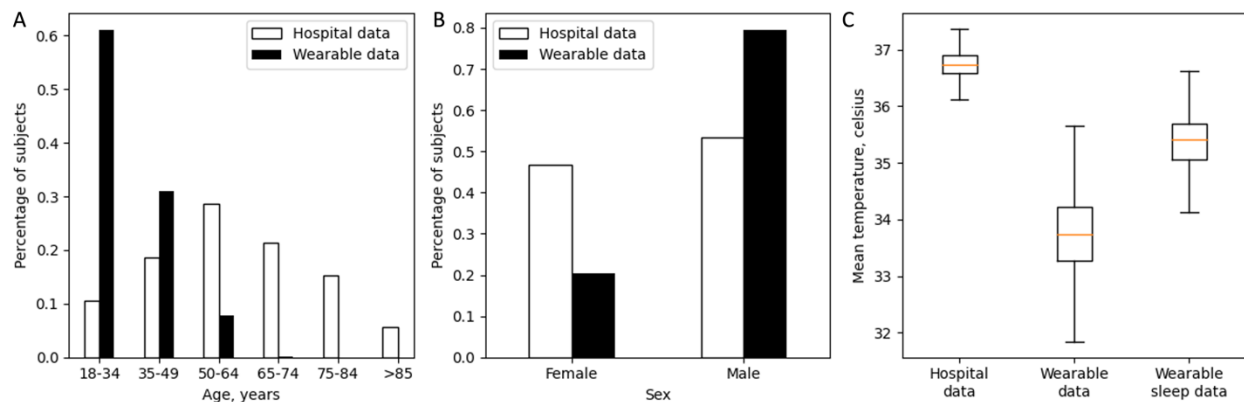
268 Third, the data sources where the physiological features were extracted from were different
269 between the hospital dataset and the wearable dataset. Temperature features were extracted from
270 core body temperatures in the hospital dataset, whereas in the wearable dataset skin temperatures
271 measured at the fingers were used. We found that skin temperature had lower values and larger
272 variance compared with core body temperature (Figure 1C, Supplementary Table 1), which was
273 consistent with the literatures³⁴⁻³⁷.

274 Fourth, the processing methods to extract physiological signals were different between the two
275 datasets. Heart rate variability measurement RMSSD were computed based on pulse estimates of
276 heart beats. However, the signal processing algorithms that Oura ring used could be different
277 from ours in detecting the fiducial points on the pulse waveforms, and in the validation of the
278 resulted inter-beat intervals. We suspected that differences in the signal processing algorithms to
279 obtain RMSSD also contributed to the distribution differences in the RMSSD features between
280 the hospital dataset and the wearable dataset (Supplementary Table 1), in addition to the
281 demographics and health state differences mentioned above.

282 Finally, wearable physiological data is acquired in an unconstrained, real-world environment,
283 which is influenced by everyday activities and other contextual factors. In contrast, hospital

284 physiological data is typically acquired when the patient is sedentary. Daytime activity such as
285 physical exercise increases heart rate and respiratory rate^{30,31}, which is a confounding factor to
286 infection prediction because infections cause similar changes in vital signs^{32,38}. Skin temperature
287 also changes dynamically upon physical exercise, and the directionality of change depends on
288 the intensity level of the exercise and whether the skin temperature is measured over active or
289 non-active muscles³⁹. When limiting feature extraction to wearable physiology data acquired
290 during sleep, we found that sleep-only features have different data distributions compared to the
291 daily features (Supplementary Table 1). For example, the data distribution of the mean
292 temperature feature was shifted towards higher values when restricted to measurements during
293 sleep (Figure 1C).

294 We included a full comparison of feature values in Supplementary Table 1.



295 *Figure 1: Comparison between hospital dataset and wearable dataset. (A) Age distribution of*
296 *hospital dataset (white) and wearable dataset (black). (B) Sex distribution of hospital dataset*
297 *(white) and wearable dataset (black). (C) Boxplot of mean temperature feature value from*
298 *hospital dataset (left), wearable dataset (middle), and wearable dataset during sleep (right).*
299

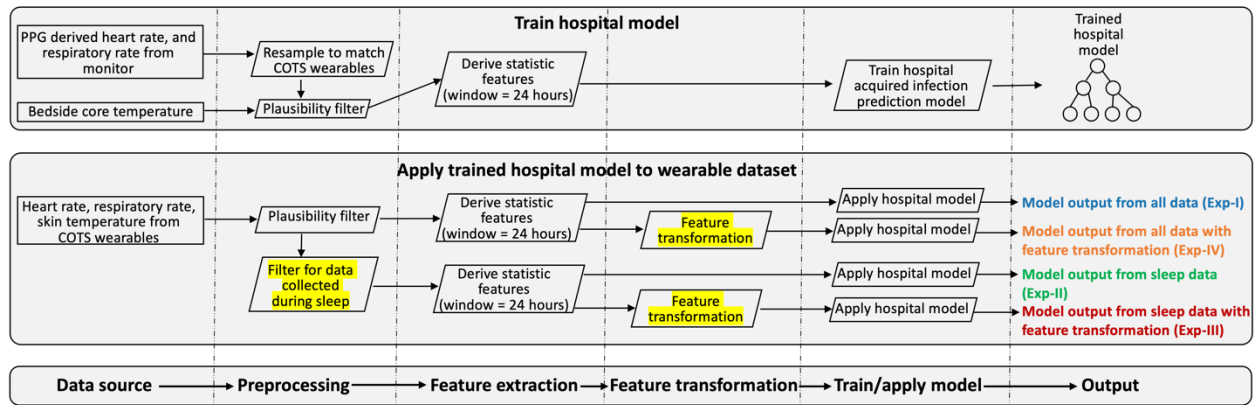
300 **Experiment design**

301 We explored two approaches to correct for differences in data distributions between hospital and
302 wearable datasets. First, we limited feature extraction to wearable physiological data from

303 wearable sensors acquired when the subject was sleeping. This approach directly mitigated
304 dataset shift by removing contextual confounders of daytime activities. Second, we explored a
305 monotonic feature transformation method to convert the data distribution of physiological
306 features in the wearable dataset to match the data distribution in the hospital dataset. This
307 approach addressed covariate shift – one of the three types of dataset shift (see
308 INTRODUCTION) - due to differences in demographics and health state between hospitalized
309 patients and subjects in the wearable dataset, as well as differences in physiological
310 measurements between COTS wearables and hospital grade devices. We compared model
311 performances with or without using such correction techniques (Experiments I, II, III, IV in
312 Figure 2), and in addition benchmarked data requirements (Experiments V, VI, VII):

- 313 • Experiment I: a baseline comparison where the trained hospital model was
314 directly applied to the daily features from the wearable dataset. Physiological
315 measurements during both awake and sleep were used to extract the daily
316 features.
- 317 • Experiment II: the trained hospital model was tested on sleep-only features from
318 the wearable dataset. Sleep-only features were extracted from the same window
319 and time interval as the daily features but only using measurements during sleep
320 segments.
- 321 • Experiment III: the trained hospital model was tested on sleep-only features after
322 the sleep-only features were transformed to match the distribution of the hospital
323 dataset.
- 324 • Experiment IV: the trained hospital model was tested on daily features after the
325 daily features were transformed to match the distribution of the hospital dataset.

- 326
- Experiments V, VI, VII: benchmarking the amount and type of wearable data
- 327 needed for the monotonic feature transformation.



328

329 *Figure 2: Schematic view of pipelines for training the hospital model (top box) and for testing*
330 *the trained model in the wearable dataset (middle box). Similar steps of the two pipelines are*
331 *aligned (bottom box). The trained hospital model was applied to the wearable dataset with or*
332 *without the two dataset shift corrections (highlighted, middle box), which resulted in four*
333 *experiments (Exp-I, II, III, IV in middle box) to compare model performance.*

334 **Baseline comparison (Experiment I)**

335 We directly applied the hospital model trained for hospital-acquired infection prediction to the
336 wearable daily features and quantified its performance in predicting COVID-19 infections. We
337 hypothesized that the hospital model would not generalize well in predicting COVID-19
338 infections, due to the differences between hospital and wearable physiological feature spaces.
339 We found that the hospital model performed at Area under ROC Curve (AUROC) = 0.527,
340 Average Precision (AP) = 0.132, Sensitivity = 0.163 and Specificity = 0.866 at break-even point,
341 Sensitivity = 0.193 and 0.113 respectively when Specificity was at 0.8 and 0.9. This performance
342 was at chance level ($p=0.07$ for AUROC), suggesting that the hospital model failed to generalize
343 when directly applied to wearable dataset.

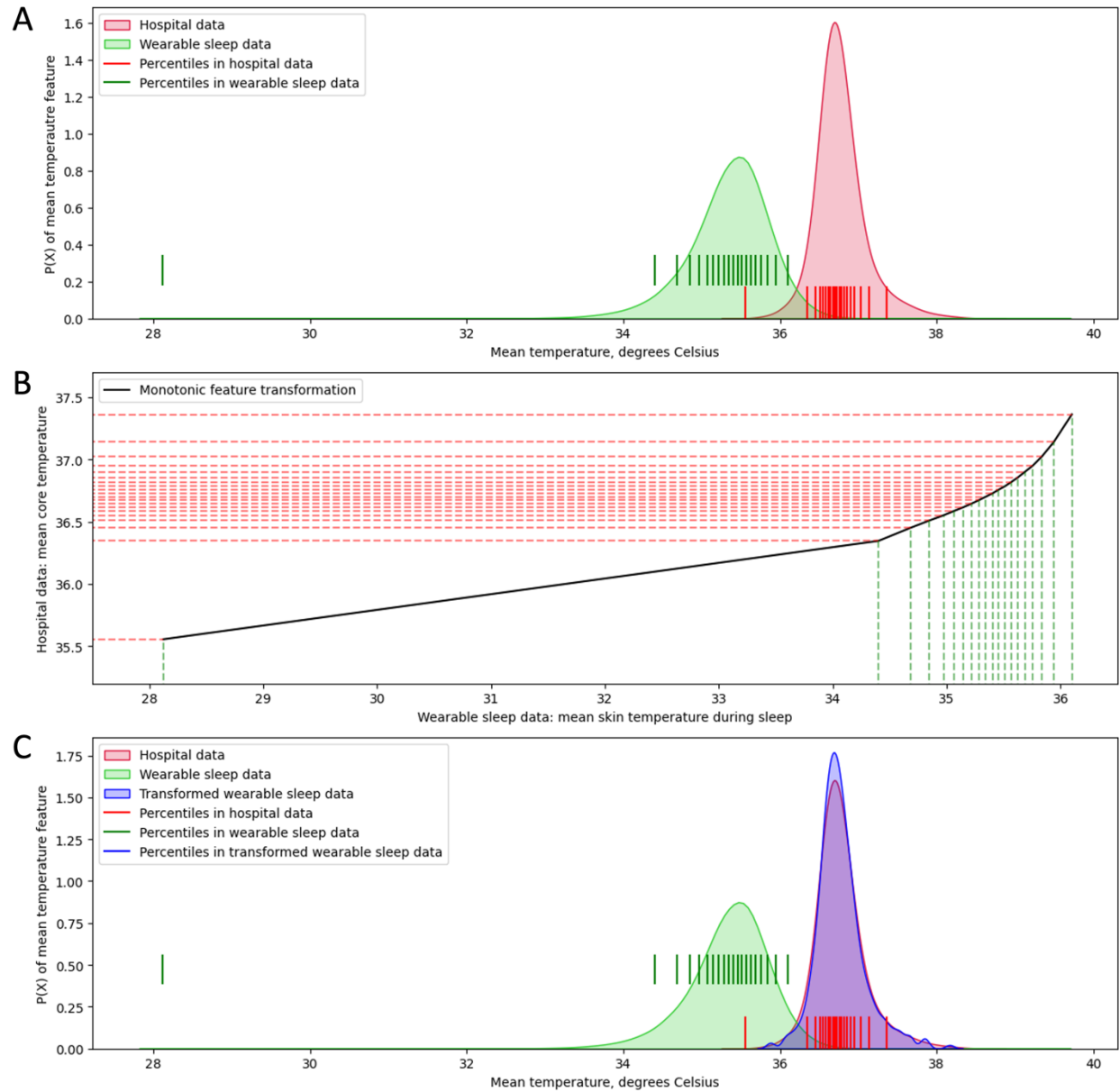
344 **Removing contextual confounders (Experiment II)**

345 When controlling for contextual factors like daytime activity, we found that the hospital model
346 using the sleep-only features performed at AUROC = 0.644 ($p < 0.001$), AP = 0.260, Sensitivity =
347 0.279 and Specificity = 0.897 at break-even point, Sensitivity = 0.402 and 0.269 respectively
348 when Specificity was at 0.8 and 0.9. Thus, using sleep-only features resulted a 22% boosting of
349 performance in terms of AUROC, suggesting the importance of controlling for contextual
350 confounders when extracting the likelihood of infection from wearables physiological data.

351 *Applying feature transformation after removing contextual confounders (Experiment III)*

352 We hypothesized that a monotonic feature transformation procedure which transforms the
353 wearable feature values to match the distribution in hospital dataset (see METHODS) could
354 improve performance of the hospital model. Using mean temperature feature as an example
355 (Figure 3), the feature transformation procedure based on matching feature values that share the
356 same 0-100 percentile value in their corresponding datasets resulted in an almost identical data
357 distribution of the mean temperature feature between the two datasets, despite large
358 discrepancies in the data distributions before transformation. Hence, we performed the same
359 feature transformation procedure independently on each feature, and evaluated the performance
360 of hospital model on the wearable dataset after all the features were transformed. We found that
361 the hospital model performed at AUROC = 0.740 ($p < 0.001$; Figure 4A, red), AP = 0.330 (Figure
362 4B, red), Sensitivity = 0.379 and Specificity = 0.910 at break-even point, Sensitivity = 0.588 and
363 0.409 respectively when Specificity was at 0.8 and 0.9, using transformed wearable sleep-only
364 features. Applying feature transformation on the sleep-only features resulted an additional 15%
365 boosting of performance in terms of AUROC (0.740 versus 0.643, red versus green in Figure
366 4A).

It is made available under a [CC-BY 4.0 International license](https://creativecommons.org/licenses/by/4.0/).



367

368 *Figure 3: Monotonic feature transformation of mean temperature feature. Red, hospital dataset;*
369 *green, wearable dataset (sleep-only features); blue, transformed wearable dataset (sleep-only*
370 *features). (A) Data distribution of mean temperature feature: red and green shaded areas*
371 *describe data distribution from hospital and wearable sleep data respectively. Vertical lines*
372 *mark the 0-100 percentile values in 5% intervals on the x-axis corresponding to each dataset.*
373 *(B) Monotonic feature transformation curve (black) where feature values with the same*
374 *percentile value are mapped between two datasets. Dashed lines mark the 0-100 percentile*
375 *values in 5% intervals on the x-axis for wearable sleep data (green) and on the y-axis for*
376 *hospital data (red). (C) Data distribution of mean temperature feature: red, green and blue*
377 *shaded areas describe data distribution from hospital dataset, wearable sleep dataset and*
378 *transformed wearable sleep dataset respectively. Vertical lines mark the 0-100 percentile values*
379 *in 5% intervals on the x-axis corresponding to each dataset; blue vertical lines are overlapped*
380 *with red vertical lines.*

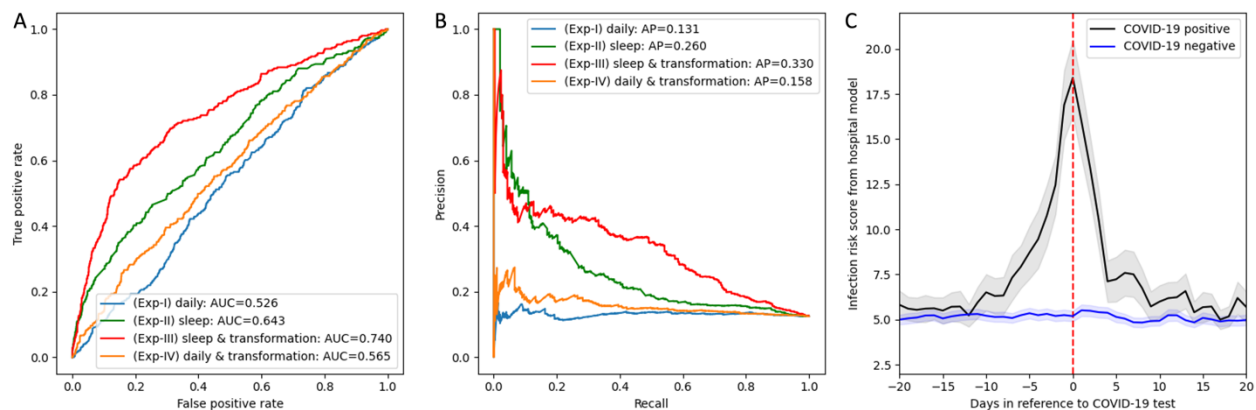
381 ***Applying feature transformation without removing contextual confounders (Experiment IV)***

382 We further investigated whether the same feature transformation procedure could improve the
383 performance of the hospital model on wearable features without removing the contextual
384 confounder of awake versus sleep. Similarly, we calculated percentile values of daily wearable
385 features derived from awake and sleep data combined, replaced the feature value with the
386 corresponding value from the hospital dataset, and evaluated the performance of the hospital
387 model on the transformed features. The model had an AUROC = 0.566 ($p < 0.001$; Figure 4A,
388 orange), AP = 0.158 (Figure 4B, orange), Sensitivity = 0.256 and Specificity = 0.844 at break-
389 even point, Sensitivity = 0.296 and 0.146 respectively when Specificity was at 0.8 and 0.9, when
390 applied to the transformed wearable features without using sleep data exclusively. The model
391 performance was slightly better than before feature transformation (AUROC: 0.565 versus 0.526,
392 orange versus blue in Figure 4A), but the improvement was not as substantial as when applying
393 feature transformation to the sleep-only features (AUROC: 0.740 versus 0.643, red versus green
394 in Figure 4A). These results suggested that both controlling for contextual cofounders and
395 applying feature transformation to address dataset shift were important to enable good model
396 performance.

397 ***Comparison with previous work***

398 We have shown that the hospital model trained for hospital-acquired infection prediction
399 performed the best in detecting early signs of COVID-19 infection on wearable dataset when
400 feature transformations were performed and when only sleep data were considered (AUROC =
401 0.740, Experiment III). Although this performance is viable for a system, it was lower than our

402 previously reported solution using a model trained directly on wearable dataset with COVID-19
403 labels (AUROC = 0.82)¹. This was expected because the hospital model was designed to be an
404 economical minimal viable solution that uses no COVID-19 labels for training, and thus not
405 capable of controlling for concept drifts and/or label shifts. When overlaying risk scores with
406 time from the Experiment III hospital model, on average subjects with positive COVID-19 test
407 results showed risk score elevations around COVID-19 test time (Figure 4C, black), whereas
408 subjects with negative COVID-19 test maintained their baseline risk scores (Figure 4C, blue).
409 Based on a cut-off risk threshold of 15 (yielding 60% sensitivity and 78% specificity), we
410 identified the days in which the model output exceeded the defined threshold within the 14-day
411 window prior to COVID-19 testing to estimate the lead time of positive classification (see
412 METHODS). We found that the Experiment III hospital model successfully predicted COVID-
413 19 infection, on average, 2.2 days prior to testing. This lead time was slightly lower but
414 comparable to our previously reported wearable solution of 2.3 days prior to testing¹.



415

416 *Figure 4: Hospital model performance and risk scores in detecting COVID-19 infection from*
417 *wearable dataset. (A) Receiver Operating Characteristic (ROC) curves. Experiment I (blue):*
418 *hospital model directly applied to wearable daily features. Experiment II (green): hospital model*
419 *applied to wearable sleep-only features. Experiment III (red): hospital model applied to*
420 *wearable sleep-only features after feature transformation. Experiment IV (orange): hospital*
421 *model applied to wearable daily features after feature transformation, without using sleep data*
422 *exclusively. Area under the ROC curve (AUROC) for each experiment is included in the figure*
423 *legend. (B) Precision-recall curve. Colors are the same as described in subplot A. Average*

424 *Precision (AP) score for each experiment is included in the figure legend. (C) Mean infection*
425 *risk score based on the output of the best generalized hospital model (Experiment III: sleep-only*
426 *features + feature transformation) in 301 COVID-19 positive subjects (black) and 2,111*
427 *COVID-19 negative subjects (blue) as a function of number of days relative to the COVID-19*
428 *test time (red). Grey and light-blue shaded area depicts 95% confidence interval.*

429 ***Data requirements for feature transformation (Experiments V, VI, VII)***

430 Given that our best generalized hospital model (Experiment III) performed reasonable but
431 inferior to our previous wearable model ¹, it is most sensible to use a hospital model for
432 predicting COVID-19 in the absence of the wearable model, e.g. at the onset and during the early
433 stage of the outbreak when data from COVID-19 positive cases were limited or unavailable to
434 train a wearable model. Therefore, we investigated the data requirements of the generalized
435 hospital model in Experiment III, in particular, the type and amount of wearable sleep data
436 needed for the feature transformation. A favorable solution should require minimal COVID-19
437 positive instances. We performed three sets of additional experiments.

438 First, we asked whether illness data of COVID-19 were required for feature transformation
439 (Experiment V). Interestingly, we found that baseline healthy data was sufficient because 1)
440 using wearable sleep data from subjects that only reported negative test results for the feature
441 transformation resulted in similar AUROC of 0.741 (Experiment V-a, Supplementary Table 2),
442 and 2) using wearable sleep data 4 weeks to 2 weeks before COVID-19 test – a time range when
443 subjects were not infected - achieved similar results (AUROC = 0.741; Experiment V-b,
444 Supplementary Table 2).

445 Second, we asked whether wearable sleep data for feature transformation needed to be from the
446 same subjects (Experiment VI, Supplementary Table 2). We randomly split the subjects in the

447 wearable dataset into 5 folds, and for each subject we used subjects from the other four folds to
448 transform the features of the given subject. The model performed with an AUROC of 0.741,
449 suggesting that the wearable data used for feature transformation do not need to come from the
450 same subjects used to test the model.

451 Third, we examined the minimum sleep data needed for feature transformation by benchmarking
452 model performance against using sleep data from reduced number of days or from reduced
453 number of subjects (Experiment VII). We gradually decreased the number of days from the 14
454 days prior to COVID-19 test where the wearable data were used for feature transformation
455 (Experiment VII-a, Supplementary Table 2). We found that the model performed at AUROC of
456 0.74 when more than 2 days immediately preceding the COVID-19 test were used for feature
457 transformation, and the model performed at AUROC = 0.73 when using data from the day before
458 or two days before COVID-19 test for feature transformation. We also benchmarked against data
459 from randomly selected days within the 14-day window prior to the COVID-19 test for feature
460 transformation and found that the model performed at AUROC of 0.74 for all experiments -
461 randomly selecting number of N days where N ranges from 1 to 13 days (Experiment VII-b,
462 Supplementary Table 2). Further, we pooled all subject days and used random down-samples for
463 feature transformation (Experiment VII-c, Supplementary Table 2). We found that the model
464 performed at AUROC of 0.74 for all experiments of reduced subject days (number of reduced
465 subject days: 25,000, or 20,000, or 15,000, or 10,000, or 7,500, or 5,000, or 3,000, or 1,000, or
466 500, or 300), even when only 300 subject days were used. Regarding the number of subjects
467 needed, we used data from randomly down-sampled subjects for feature transformation and
468 found that the model performed at AUROC of 0.74 for all experiments of reduced number of

469 subjects (number of reduced subjects: 2,000, or 1,500, or 1,000, or 500, or 250, or 100, or 50, or
470 25), even when the number of subjects was reduced to 25 (Experiment VII-d, Supplementary
471 Table 2).

472 Summarizing all the experiments (Supplementary Table 2), we concluded that healthy wearable
473 data from 25 subjects collected in a period of 14 days for feature transformation would be
474 sufficient to ensure the same model performance of AUROC = 0.74.

475 **DISCUSSIONS**

476 This study demonstrated the feasibility of applying a machine learning model trained on hospital
477 data to detect early signs of COVID-19 infection in physiological data from COTS wearables
478 outside of hospitals. Our hospital model was trained from hospitalized patients and vital signs
479 collected from hospital grade devices to test against a set of common hospital-acquired infections
480 (prior to the COVID-19 outbreak), therefore had no prior knowledge of COVID-19 infections
481 and no exposure to physiological data collected through COTS wearables. Nevertheless, after
482 controlling for dataset shift, the hospital model performed at AUROC = 0.74 in alerting COVID-
483 19 infection before diagnostic testing from wearable physiology monitoring in military personnel
484 under unrestrained use. This performance was lower than our previously reported solution using
485 a model trained directly on wearable dataset with COVID-19 labels ¹, but is nevertheless viable
486 for a system, and can detect COVID-19 infection 2 days before diagnostic testing, with no need
487 of model retraining. Importantly, our approaches in addressing dataset shift did not require any
488 labeled data of COVID-19 cases; rather, a small dataset from healthy subjects – e.g. 2 weeks of
489 wearable data from 25 subjects – was sufficient to generalize the hospital model to predict

490 COVID-19 infection from wearable data with an AUROC of 0.74. Therefore, our efficient
491 solution of generalizing the hospital model of infection prediction to wearable physiological
492 monitoring would be most economical and useful at early onset of outbreaks of novel infections
493 when data from positive cases are limited or unavailable to train an pathogen-specific model –
494 such as our previously reported COVID-19 wearable model ¹. Because a small amount of healthy
495 baseline data is feasible to collect prior to any infection outbreak, the transformation function to
496 calibrate the feature values can be derived to enable rapid deployment of a pre-trained model.
497 We anticipate such a solution could create a big impact in infectious disease control, as
498 transmission prevention at the onset and during the early outbreak of an infectious disease is
499 critical.

500 The two enablers of our solution of generalized hospital model were 1) the isolation of
501 contextual confounders, focusing on sleep-only wearable data, and 2) feature transformations
502 that calibrated the wearable feature values to match the distribution of the hospital model training
503 data and that do not rely on positive labels. Both reduced the differences in the joint distribution
504 of the physiological features X and the infectious disease labels Y between the hospital dataset
505 and the wearable dataset, therefore mitigating dataset shift. The model performed at chance level
506 without these two corrections and performed at AUROC of 0.74 when and only when both
507 corrections were used. This is likely because the two methods controlled for different aspects of
508 dataset shift. Feature transformation is a correction technique for covariate shift (see
509 INTRODUCTION) because it modifies the probability distribution of the physiological features
510 $P(X)$. Removing contextual confounder of daytime activities, on the other hand, controls for
511 both covariate shift $P(X)$ and to some extent concept drift $P(Y|X)$. For example, increases in

512 heart rate in hospitalized patients are associated with increased risk of infection ³⁸; in contrast,
513 increases of heart rate in the subjects from the wearable monitored cohort could be normal
514 physiology change, e.g. if the subjects are exercising ³⁰. Therefore, daytime activities such as
515 physical exercises affect the wearable physiology data in such a way that increases the likelihood
516 that they will be misclassified by the hospital model as infection cases. Hence it is beneficial to
517 use sleep-only features in our study, and that it is not sufficient to perform feature
518 transformations on the daily features without isolating sleep periods.

519 In our study we used hypnogram information from wearables to identify measurements during
520 sleep to compute sleep-only features so that both $P(X)$ and $P(Y|X)$ were more similar to the
521 hospital dataset, where the physiological measurements were acquired when patients were
522 sedentary. We could also apply the hospital model to the wearable dataset in other similar
523 scenarios such as during wakefulness, but limited to resting/sedentary states. It is possible that
524 there are other contextual confounders that we could identify and isolate from the wearable
525 dataset to further improve the model performance of the generalized hospital model. Identifying
526 contextual factors does not require any explicit knowledge of data distributions of the training
527 nor testing datasets but relies on domain knowledge of the model training and application
528 scenarios. Removing contextual factors, however, relies on the availability of data elements that
529 can be used to isolate the contextual factors.

530 It is challenging to address all aspects of dataset shift. In particular, label shift and concept drift
531 would require labels to be properly addressed. Previous work that corrected dataset shift using
532 unlabeled data typically addressed covariate shift, and involved re-training using resampling

533 weights that were either estimated from the biasing densities ⁴⁰⁻⁴² or inferred by comparing
534 nonparametric distributions between training and testing samples ⁴³. In contrast, the monotonic
535 feature transformation technique described in our study requires no labels, no re-training, and is a
536 straightforward mathematical operation that preserves the rank order of data but modifies the
537 shape of the distribution. By doing so, we are minimizing the dataset differences in physiological
538 signals caused by the differences in individual and group baselines, and by the differences in
539 measurement devices, yet preserving the relative rank of infection risk among individuals. Our
540 hospital model was based on ensembles of decision tree which makes aggregated decisions from
541 individual features on each tree split. This makes it possible for us to manipulate the distribution
542 of each feature independently without altering the overall decision from the tree ensembles based
543 on the feature ranks (e.g. $P(Y|X)$ is unchanged for monotonic transformations of X , where X is
544 the physiological features and Y is the infection labels). Algorithms based on decision trees are
545 particularly suitable for disease modeling, as typically lower and/or higher clinical measurements
546 are associated with declined health. In other words, infection risk as a function of clinical
547 measurements resembles a U-shape curve or a monotonic function. This is the reason why
548 preserving the rank of feature values worked in our solution as it preserved the rank of infection
549 risk, e.g. both a high rank of skin temperature and a high rank of core temperature are associated
550 with high infection risk, therefore the conditional probability of COVID-19 infection risk given
551 skin temperature $P(Y_{\text{covid}}|X_{\text{skin}})$ can be monotonically mapped to the conditional probability
552 of infection risk given core temperature $P(Y_{\text{infection}}|X_{\text{core}})$.

553 Our feature transformation technique requires no labels (therefore is “unsupervised”), no re-
554 training, and is computationally inexpensive and interpretable, compared with previous work that

555 corrected dataset shift^{40,41,41-43}. It is device-agnostic by nature, and we demonstrated its
556 effectiveness in addressing the dataset shift due to differences in measurement devices, e.g. the
557 skin temperature feature from Oura ring was transformed to have almost identical distribution as
558 the core temperature feature from hospital grade device (Figure 3). Removing context
559 confounders have its challenges in first identifying the relevant context and then finding data
560 elements that can be used to isolate the context, but theoretically has the potential to make our
561 solution context-agnostic. Our generalized hospital model of infection prediction performed well
562 in detecting COVID-19, despite pathogen differences in COVID-19 infection and the set of
563 hospital-acquired infections used to train the hospital model. Therefore, we believe the
564 generalized hospital model can be easily adapted to deploy in other scenarios of infection
565 prediction, and it is not restricted to a specific set of wearable devices, a specific population, or a
566 specific context. For example, the hospital model of infection prediction may be used to track the
567 health state of healthcare professionals during flu season with a different set of wearables, given
568 that similar types of vital sign signals are collected, and appropriate dataset shift transformations
569 are applied.

570 **CONCLUSTIONS**

571 We found that an infection prediction model developed for hospitalized patients can detect early
572 signs of COVID-19 infection from wearable physiological monitoring (AUROC=0.74), on
573 average 2 days earlier than diagnostic testing, provided that a small sample (e.g. 25 subjects in a
574 period of 14 days) of wearable data from healthy subjects is available to address the dataset shift
575 between hospital dataset and wearable dataset, and that sleep markers can be extracted to control
576 for contextual effects in wearable dataset. Our approaches to transform features between datasets

577 and isolate contextual confounders can enable rapid deployment of a pre-trained infection
578 prediction model at the onset of novel infection outbreaks.

579 **ACKNOWLEDGEMENTS**

580 This study is sponsored by the US Department of Defense (DoD), Defense Threat Reduction
581 Agency (DTRA) under contracts: W15QKN-18-9-1002 (CB10560), HDTRA1-20-C-0041,
582 HDTRA121C0006. The views, opinions and/or findings expressed are those of the authors and
583 should not be interpreted as representing the official views or policies of the Department of
584 Defense or the US Government. We appreciate the vision, leadership, and sponsorship from the
585 US Department of Defense and the US Government: Edward Argenta, Christopher Kiley and
586 Katherine Delaveris. We recognize our former Philips North America colleague Saeed
587 Babaeizadeh for PPG signal processing.

588 **LIST OF ABBREVIATIONS**

589 AI: Artificial Intelligence
590 ML: Machine Learning
591 CDS: Clinical Decision Support
592 Spec: Specificity
593 Sens: Sensitivity
594 AUROC: Area under the Receiver Operating Characteristic Curve
595 AP: Average Precision
596 COTS wearables: Commercial-off-the-shelf wearables
597 HAI: Hospital-acquired Infection

598 **DECLARATIONS**

599 *Ethics approval and consent to participate:* The MIMIC-III project was approved by the
600 Institutional Review Boards of Beth Israel Deaconess Medical Center and the Massachusetts
601 Institute of Technology. Banner Health data use was a part of a retrospective deterioration
602 detection study approved by the Institutional Review Board of Banner Health and by the Philips
603 Internal Committee for Biomedical Experiments. For both hospital datasets, requirement for
604 individual patient consent was waived because the project did not impact clinical care, was no
605 greater than minimal risk, and all protected health information was removed from the limited
606 dataset used in this study.

607 The collection and use of the wearable dataset was approved by the Institutional Review Boards
608 of the US Department of Defense. Informed consent was obtained from all participants.

609 *Availability of data and materials:* MIMIC-III dataset is available in PhysioNet repository,
610 <https://mimic.physionet.org/>. The Banner Health dataset is a proprietary dataset that is not
611 publicly shareable. The wearable dataset is from US military personnel and is not publicly
612 shareable.

613 **Conflicts of Interest Statement:** Authors TF, SM, BC, RD and IS are employees of Philips
614 North America. Author DM was employee of Philips North America. Author DS is employee of
615 Banner Health. All authors declare no other competing interests.

616 **Funding Statement:** This study is sponsored by the US Department of Defense (DoD), Defense
617 Threat Reduction Agency (DTRA) under contracts: W15QKN-18-9-1002 (CB10560),
618 HDTRA1-20-C-0041, HDTRA121C0006. The funding body did not play a role in the study
619 design, collection, analysis, interpretation of data, the writing of this article or the decision to
620 submit it for publication. The views, opinions and/or findings expressed are those of the authors
621 and should not be interpreted as representing the official views or policies of the Department of
622 Defense or the US Government.

623 **Authors' contributions:** TF, DM and BC participated in the conception of the study. TF
624 analyzed the data, trained, validated the models, and wrote the first draft. SM extracted
625 waveform numeric, processed PPG waveforms and extracted heart rate variability features. RD
626 extracted labels from the wearable dataset. BC and IS set up the ETL pipeline for wearable data
627 processing. DS provided clinical consultation and reviewed the manuscript. All authors
628 participated in interpreting the results, writing, and revising the manuscript. All authors have
629 read and approved the manuscript.

630 REFERENCES

- 631 1. Conroy, B. *et al.* Real-time infection prediction with wearable physiological monitoring and
632 AI to aid military workforce readiness during COVID-19. *Sci. Rep.* **12**, 3797 (2022).
- 633 2. Pascarella, G. *et al.* COVID-19 diagnosis and management: a comprehensive review. *J.*
634 *Intern. Med.* **288**, 192–206 (2020).
- 635 3. Zhai, P. *et al.* The epidemiology, diagnosis and treatment of COVID-19. *Int. J. Antimicrob.*
636 *Agents* **55**, 105955 (2020).

- 637 4. Jin, Y. *et al.* Virology, epidemiology, pathogenesis, and control of COVID-19. *Viruses* **12**,
638 372 (2020).
- 639 5. Quer, G. *et al.* Wearable sensor data and self-reported symptoms for COVID-19 detection.
640 *Nat. Med.* **27**, 73–77 (2021).
- 641 6. Mishra, T. *et al.* Pre-symptomatic detection of COVID-19 from smartwatch data. *Nat.*
642 *Biomed. Eng.* **4**, 1208–1220 (2020).
- 643 7. Miller, D. J. *et al.* Analyzing changes in respiratory rate to predict the risk of COVID-19
644 infection. *PloS One* **15**, e0243693 (2020).
- 645 8. Hasty, F. *et al.* Heart rate variability as a possible predictive marker for acute inflammatory
646 response in COVID-19 patients. *Mil. Med.* **186**, e34–e38 (2021).
- 647 9. Zhu, G. *et al.* Learning from Large-Scale Wearable Device Data for Predicting the Epidemic
648 Trend of COVID-19. *Discrete Dyn. Nat. Soc.* **2020**, 1–8 (2020).
- 649 10. Hirten, R. P. *et al.* Use of Physiological Data From a Wearable Device to Identify SARS-
650 CoV-2 Infection and Symptoms and Predict COVID-19 Diagnosis: Observational Study. *J.*
651 *Med. Internet Res.* **23**, e26107 (2021).
- 652 11. Radin, J. M., Quer, G., Jalili, M., Hamideh, D. & Steinhubl, S. R. The hopes and hazards of
653 using personal health technologies in the diagnosis and prognosis of infections. *Lancet Digit.*
654 *Health* **3**, e455–e461 (2021).
- 655 12. Mitratza, M. *et al.* The performance of wearable sensors in the detection of SARS-CoV-2
656 infection: a systematic review. *Lancet Digit. Health* **4**, e370–e383 (2022).
- 657 13. Yang, D.-M. *et al.* Smart healthcare: A prospective future medical approach for COVID-19.
658 *J. Chin. Med. Assoc.* **86**, 138 (2023).

- 659 14. Natarajan, A., Su, H.-W. & Heneghan, C. Assessment of physiological signs associated with
660 COVID-19 measured using wearable devices. *NPJ Digit. Med.* **3**, 156 (2020).
- 661 15. Feng, T. *et al.* Machine learning-based clinical decision support for infection risk prediction.
662 *Front. Med.* **10**, (2023).
- 663 16. Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A. & Lawrence, N. D. *Dataset Shift in*
664 *Machine Learning*. (MIT Press, 2022).
- 665 17. Johnson, A. E. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9
666 (2016).
- 667 18. Moody, B., Moody, M., Villarroel, M., Clifford D., G. & Silva, I. MIMIC-III Waveform
668 Database Matched Subset. (2020).
- 669 19. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the*
670 *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*
671 785–794 (ACM, San Francisco California USA, 2016). doi:10.1145/2939672.2939785.
- 672 20. Geovanini, G. R. *et al.* Age and sex differences in heart rate variability and vagal specific
673 patterns–Baependi heart study. *Glob. Heart* **15**, (2020).
- 674 21. Almeida-Santos, M. A. *et al.* Aging, heart rate variability and patterns of autonomic
675 regulation of the heart. *Arch. Gerontol. Geriatr.* **63**, 1–8 (2016).
- 676 22. Reardon, M. & Malik, M. Changes in Heart Rate Variability with Age. *Pacing Clin.*
677 *Electrophysiol.* **19**, 1863–1866 (1996).
- 678 23. Bonnemeier, H. *et al.* Circadian Profile of Cardiac Autonomic Nervous Modulation in
679 Healthy Subjects: Differing Effects of Aging and Gender on Heart Rate Variability. *J.*
680 *Cardiovasc. Electrophysiol.* **14**, 791–799 (2003).

- 681 24. Stein, P. K., Kleiger, R. E. & Rottman, J. N. Differing effects of age on heart rate variability
682 in men and women. *Am. J. Cardiol.* **80**, 302–305 (1997).
- 683 25. Abhishekh, H. A. *et al.* Influence of age and gender on autonomic regulation of heart. *J. Clin.*
684 *Monit. Comput.* **27**, 259–264 (2013).
- 685 26. Peng, C.-K. *et al.* Quantifying Fractal Dynamics of Human Respiration: Age and Gender
686 Effects. *Ann. Biomed. Eng.* **30**, 683–692 (2002).
- 687 27. Migliaro, E. R. *et al.* Relative influence of age, resting heart rate and sedentary life style in
688 short-term analysis of heart rate variability. *Braz. J. Med. Biol. Res.* **34**, 493–500 (2001).
- 689 28. Ogliari, G. *et al.* Resting heart rate, heart rate variability and functional decline in old age.
690 *Cmaj* **187**, E442–E449 (2015).
- 691 29. Umetani, K., Singer, D. H., McCraty, R. & Atkinson, M. Twenty-Four Hour Time Domain
692 Heart Rate Variability and Heart Rate: Relations to Age and Gender Over Nine Decades. *J.*
693 *Am. Coll. Cardiol.* **31**, 593–601 (1998).
- 694 30. Altini, M. & Plews, D. What is behind changes in resting heart rate and heart rate variability?
695 A large-scale analysis of longitudinal measurements acquired in free-living. *Sensors* **21**, 7932
696 (2021).
- 697 31. Nicolò, A., Massaroni, C., Schena, E. & Sacchetti, M. The importance of respiratory rate
698 monitoring: From healthcare to sport and exercise. *Sensors* **20**, 6396 (2020).
- 699 32. Loughlin, P. C., Sebat, F. & Kellett, J. G. Respiratory rate: The forgotten vital sign—Make it
700 count! *Jt. Comm. J. Qual. Patient Saf.* **44**, 494–499 (2018).
- 701 33. Hill, B. & Annesley, S. H. Monitoring respiratory rate in adults. *Br. J. Nurs.* **29**, 12–16
702 (2020).

- 703 34. Gradisar, M. & Lack, L. Relationships between the Circadian Rhythms of Finger
704 Temperature, Core Temperature, Sleep Latency, and Subjective Sleepiness. *J. Biol. Rhythms*
705 **19**, 157–163 (2004).
- 706 35. Henane, R., Buguet, A., Roussel, B. & Bittel, J. Variations in evaporation and body
707 temperatures during sleep in man. *J. Appl. Physiol.* **42**, 50–55 (1977).
- 708 36. Hasselberg, M. J., McMahon, J. & Parker, K. The validity, reliability, and utility of the
709 iButton® for measurement of body temperature circadian rhythms in sleep/wake research.
710 *Sleep Med.* **14**, 5–11 (2013).
- 711 37. Caroline Kryder. How accurate is Oura’s temperature data?
712 <https://ouraring.com/blog/temperature-validated-accurate/> (2020).
- 713 38. Radin, J. M. *et al.* Assessment of Prolonged Physiological and Behavioral Changes
714 Associated With COVID-19 Infection. *JAMA Netw. Open* **4**, e2115959 (2021).
- 715 39. Neves, E. B. *et al.* Different responses of the skin temperature to physical exercise:
716 Systematic review. in *2015 37th Annual International Conference of the IEEE Engineering*
717 *in Medicine and Biology Society (EMBC)* 1307–1310 (IEEE, 2015).
- 718 40. Zadrozny, B. Learning and evaluating classifiers under sample selection bias. in *Twenty-first*
719 *international conference on Machine learning - ICML '04* 114 (ACM Press, Banff, Alberta,
720 Canada, 2004). doi:10.1145/1015330.1015425.
- 721 41. Dudík, M., Phillips, S. & Schapire, R. E. Correcting sample selection bias in maximum
722 entropy density estimation. *Adv. Neural Inf. Process. Syst.* **18**, (2005).
- 723 42. Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-
724 likelihood function. *J. Stat. Plan. Inference* **90**, 227–244 (2000).

- 725 43. Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B. & Smola, A. Correcting sample
726 selection bias by unlabeled data. *Adv. Neural Inf. Process. Syst.* **19**, (2006).

727 *Supplementary Table 1: Mean and standard deviation (std) of the feature values by dataset.*
 728 *Hospital data – from 9,517 hospitalized patients. Wearable data – from 33,164 subject days.*
 729 *Wearable sleep data – from 31,269 subject sleep segments.*

Physiological signal	Feature name	Hospital data (mean±std)	Wearable data (mean±std)	Wearable sleep data (mean±std)
Heart Rate, Beats per Minute	Mean(Heart Rate)	81.40±15.58	71.57±9.80	59.03±8.54
	Std(Heart Rate)	15.00±8.75	13.54±5.20	4.58±1.89
	Max(Heart Rate)	151.9±34.52	125.4±28.23	78.05±12.62
	Min(Heart Rate)	49.50±16.44	49.92±7.29	49.70±7.31
Respiratory Rate, Breaths per Minute	Mean(Respiratory Rate)	17.89±3.49	13.90±0.89	14.51±1.58
	Std(Respiratory Rate)	3.07±0.99	1.67±0.55	1.63±0.63
	Max(Respiratory Rate)	30.24±6.63	19.91±2.64	20.01±2.89
	Min(Respiratory Rate)	9.36±3.02	9.83±1.14	10.16±1.59
Temperature, Celsius	Mean(Temperature)	36.76±0.32	33.72±0.90	35.34±0.55
	Std(Temperature)	0.30±0.16	2.04±0.47	0.72±0.32
	Min(Temperature)	36.30±0.35	28.16±0.56	32.07±1.94
	Max(Temperature)	37.26±0.54	37.07±1.45	36.40±0.45
Root Mean Square Successive	Mean(RMSSD)	0.118±0.10	0.060±0.035	0.060±0.034
	Std(RMSSD)	0.040±0.036	0.016±0.009	0.017±0.009
	Max(RMSSD)	0.212±0.133	0.105±0.050	0.111±0.051

Difference, Milliseconds	Min(RMSSD)	0.061±0.073	0.028±0.020	0.025±0.018
-----------------------------	------------	-------------	-------------	-------------

730 *Supplementary Table 2: model performance. Six performance metrics were calculated: AUC*
 731 *(Area under ROC Curve), AP (Average Precision), Sens.@Break-even (Sensitivity at Precision-*
 732 *Recall break-even point), Spec.@Break-even, (Specificity at Precision-Recall break-even point),*
 733 *Sens.@Spec.=0.8 (Sensitivity when Specificity is 0.8), Sens@Spec.=0.9 (Sensitivity when*
 734 *Specificity is 0.9). Experiment I, hospital model directly applied to wearable daily features.*
 735 *Experiment IV, hospital model applied to wearable daily feature after feature transformation.*
 736 *Experiment III, hospital model applied to wearable sleep-only feature. Experiment IV, hospital*
 737 *model applied to wearable sleep-only feature after feature transformation. Experiment V-a,*
 738 *hospital model applied to wearable sleep-only feature with feature transformation using data*
 739 *from only subjects who reported negative test. Experiment V-b, hospital model applied to*
 740 *wearable sleep-only feature with feature transformation using data collected 28 days before to*
 741 *14 days before COVID-19 test. Experiment VI, hospital model applied to wearable sleep-only*
 742 *feature with cross-validated feature transformation. Experiment VII-a, hospital model applied to*
 743 *wearable sleep-only feature with feature transformation using data from the recent n days prior*
 744 *to testing (n ranges from 1 to 13). Experiment VII-b, hospital model applied to wearable sleep-*
 745 *only feature with feature transformation using data from randomly selected number of n days*
 746 *within 14 days prior to testing (n ranges from 1 to 13), mean(std) from 10 iterations is shown.*
 747 *Experiment VII-d, hospital model applied to wearable sleep-only feature with feature*
 748 *transformation using data from randomly selected number of n subject days*
 749 *(n=[25000,20000,15000,10000,7500,5000,3000,1000,500,300]), mean(std) from 10 iterations is*
 750 *shown. Experiment VII-d, hospital model applied to wearable sleep-only feature with feature*
 751 *transformation using data from randomly selected number of n subjects*
 752 *(n=[2000,1500,1000,500,250,100,50,25]), mean(std) from 10 iterations is shown.*

Exp.	Features	Wearable data used for feature transformation	AUC	AP	Sens. @ Break- even	Spec. @ Break- even	Sens. @ Spec. = 0.8	Sens. @ Spec. = 0.9
I	Daily features	None	0.527	0.132	0.163	0.866	0.193	0.113
IV	derived from both awake and sleep data	[-14,0] day	0.566	0.158	0.256	0.844	0.296	0.146

II	Sleep-only	None	0.644	0.260	0.279	0.897	0.402	0.269
III	features	[-14,0] day	0.740	0.330	0.379	0.910	0.588	0.409
V-a	Sleep-only	[-14,0] day from COVID- 19 negatives	0.741	0.313	0.369	0.910	0.578	0.382
V-b	features	[-28,-14] day	0.741	0.316	0.375	0.910	0.591	0.389
VI	Sleep-only features	Cross-validated [-14,0] day	0.741	0.330	0.385	0.911	0.585	0.415
VII-a	Sleep-only features	[-13,0] day	0.741	0.332	0.399	0.907	0.588	0.419
		[-12,0] day	0.742	0.333	0.399	0.908	0.585	0.422
		[-11,0] day	0.741	0.336	0.392	0.910	0.575	0.415
		[-10,0] day	0.741	0.338	0.395	0.911	0.575	0.422
		[-9,0] day	0.741	0.338	0.392	0.913	0.578	0.429
		[-8,0] day	0.739	0.342	0.425	0.902	0.568	0.429
		[-7,0] day	0.739	0.338	0.425	0.903	0.561	0.432
		[-6,0] day	0.737	0.337	0.399	0.911	0.555	0.429
		[-5,0] day	0.739	0.341	0.429	0.906	0.555	0.435
		[-4,0] day	0.740	0.341	0.425	0.907	0.555	0.435
		[-3,0] day	0.738	0.340	0.419	0.912	0.568	0.435
		[-2,0] day	0.734	0.334	0.395	0.913	0.558	0.422
[-1,0] day	0.733	0.341	0.389	0.910	0.565	0.412		
VII-b	Sleep-only features	Random 1 day	0.738 (0.002)	0.325 (0.012)	0.391 (0.022)	0.909 (0.005)	0.579 (0.011)	0.407 (0.013)

		Random 2 day	0.740 (0.002)	0.327 (0.009)	0.406 (0.035)	0.906 (0.008)	0.579 (0.01)	0.411 (0.015)
		Random 3 day	0.740 (0.001)	0.329 (0.01)	0.388 (0.019)	0.911 (0.002)	0.578 (0.015)	0.410 (0.019)
		Random 4 day	0.739 (0.002)	0.327 (0.009)	0.383 (0.011)	0.911 (0.001)	0.580 (0.01)	0.410 (0.014)
		Random 5 day	0.740 (0.002)	0.324 (0.008)	0.380 (0.006)	0.911 (0.001)	0.581 (0.01)	0.409 (0.009)
		Random 6 day	0.739 (0.001)	0.329 (0.009)	0.393 (0.026)	0.909 (0.004)	0.577 (0.011)	0.415 (0.017)
		Random 7 day	0.741 (0.002)	0.334 (0.004)	0.398 (0.029)	0.908 (0.007)	0.578 (0.01)	0.420 (0.01)
		Random 8 day	0.740 (0.002)	0.330 (0.009)	0.388 (0.011)	0.910 (0.002)	0.58 (0.009)	0.417 (0.011)
		Random 9 day	0.741 (0.001)	0.336 (0.005)	0.408 (0.029)	0.906 (0.007)	0.577 (0.009)	0.422 (0.01)
		Random 10 day	0.740 (0.001)	0.330 (0.008)	0.389 (0.022)	0.909 (0.005)	0.582 (0.007)	0.412 (0.01)
		Random 11 day	0.741 (0.001)	0.332 (0.003)	0.387 (0.006)	0.910 (0.001)	0.583 (0.006)	0.416 (0.005)
		Random 12 day	0.740 (0.001)	0.328 (0.006)	0.393 (0.027)	0.908 (0.006)	0.582 (0.005)	0.412 (0.007)

		Random 13 day	0.740 (0.001)	0.331 (0.006)	0.396 (0.032)	0.908 (0.008)	0.583 (0.008)	0.415 (0.011)
VII-c	Sleep-only features	Random 25,000 subject days	0.742 (0.001)	0.331 (0.003)	0.388 (0.013)	0.910 (0.002)	0.585 (0.005)	0.411 (0.01)
		Random 20,000 subject days	0.741 (0.001)	0.333 (0.001)	0.392 (0.008)	0.909 (0.003)	0.583 (0.005)	0.414 (0.006)
		Random 15,000 subject days	0.741 (0.002)	0.331 (0.004)	0.395 (0.022)	0.907 (0.007)	0.582 (0.004)	0.413 (0.007)
		Random 10,000 subject days	0.741 (0.001)	0.331 (0.003)	0.384 (0.009)	0.911 (0.003)	0.582 (0.006)	0.411 (0.007)
		Random 7,500 subject days	0.742 (0.001)	0.330 (0.004)	0.384 (0.01)	0.910 (0.002)	0.584 (0.005)	0.409 (0.011)
		Random 5,000 subject days	0.742 (0.002)	0.331 (0.005)	0.388 (0.01)	0.910 (0.002)	0.586 (0.01)	0.412 (0.008)
		Random 3,000 subject days	0.741 (0.001)	0.329 (0.006)	0.394 (0.027)	0.909 (0.006)	0.586 (0.008)	0.415 (0.01)
		Random 1,000 subject days	0.741 (0.003)	0.325 (0.008)	0.384 (0.018)	0.910 (0.003)	0.588 (0.008)	0.407 (0.015)
		Random 500 subject days	0.740 (0.003)	0.332 (0.011)	0.394 (0.031)	0.908 (0.008)	0.585 (0.012)	0.407 (0.013)
		Random 300 subject days	0.739 (0.003)	0.324 (0.009)	0.396 (0.029)	0.907 (0.007)	0.584 (0.008)	0.407 (0.018)

VII-d	Sleep-only features	Random 2000 subjects	0.742 (0.001)	0.332 (0.003)	0.386 (0.008)	0.911 (0.002)	0.586 (0.007)	0.416 (0.011)
		Random 1500 subjects	0.743 (0.002)	0.331 (0.004)	0.391 (0.012)	0.910 (0.002)	0.584 (0.007)	0.417 (0.008)
		Random 1000 subjects	0.740 (0.002)	0.329 (0.004)	0.386 (0.01)	0.910 (0.003)	0.581 (0.01)	0.416 (0.012)
		Random 500 subjects	0.741 (0.002)	0.328 (0.007)	0.389 (0.017)	0.909 (0.005)	0.585 (0.011)	0.407 (0.016)
		Random 250 subjects	0.739 (0.003)	0.327 (0.011)	0.397 (0.027)	0.908 (0.005)	0.586 (0.007)	0.412 (0.016)
		Random 100 subjects	0.738 (0.006)	0.328 (0.015)	0.386 (0.013)	0.910 (0.004)	0.570 (0.014)	0.41 (0.023)
		Random 50 subjects	0.738 (0.004)	0.327 (0.019)	0.386 (0.018)	0.910 (0.004)	0.584 (0.015)	0.407 (0.02)
		Random 25 subjects	0.739 (0.006)	0.325 (0.014)	0.390 (0.021)	0.901 (0.004)	0.591 (0.018)	0.412 (0.016)