Exploring the impact of pulse oximeter selection within the COVID-19 home-use pulse oximetry pathways

David Stell,1 Jonathan James Noble,2 Rebecca Hazell Kay,1,3 Man Ting Kwong,1 Michael John Russell Jeffries2,3,4 Liam Johnston,3,4 Guy Glover,5 Emmanuel Akinluyi1,3

ABSTRACT
Background During the COVID-19 pandemic, portable pulse oximeters were issued to some patients to permit home monitoring and alleviate pressure on inpatient wards. Concerns were raised about the accuracy of these devices in some patient groups. This study was conducted in response to these concerns.

Objectives To evaluate the performance characteristics of five portable pulse oximeters and their suitability for deployment on home-use pulse oximetry pathways created during the COVID-19 pandemic. This study considered the effects of different device models and patient characteristics on pulse oximeter accuracy, false negative and false positive rate.

Methods A total of 915 oxygen saturation (\text{SpO}_2) measurements, paired with measurements from a hospital-standard pulse oximeter, were taken from 50 patients recruited from respiratory wards and the intensive care unit at an acute hospital in London. The effects of device model and several patient characteristics on bias, false negative and false positive likelihood were examined using multiple regression analyses.

Results and conclusions All five portable pulse oximeters appeared to outperform the standard to which they were manufactured. Device model, patient \text{SpO}_2 and patient skin colour were significant predictors of measurement bias, false positive and false negative rate, with some variation between models. The false positive and false negative rates were 11.2% and 24.5%, respectively, with substantial variation between models.

INTRODUCTION
The COVID-19 pandemic created unprecedented demand for bed space within National Health Service (NHS) hospitals. Home-use pulse oximeters were introduced to allow home monitoring of some patients who would otherwise occupy a hospital bed. This was recommended by WHO guidance.1 There is evidence of uptake of this recommendation in various health systems.2 3 Use of similar pathways is likely to increase,4 including for non-COVID-19 patients.5

There are two virtual wards created by the NHS in England to manage the surge of COVID-19 patients: ‘COVID-19 Virtual Ward-s’ and ‘COVID-19 Oximetry @Home’. Both virtual wards were supported by portable pulse oximeters (as opposed to hospital grade pulse oximeters) due to their wider availability. For the rest of the article portable pulse oximeters will be referred to as pulse oximeters for simplicity.

COVID-19 Virtual Wards were operated by secondary care providers, and it contains an ascending and a descending pathways. The descending pathway included patients in the recovery phase, deemed fit for transfer to a virtual ward for home-monitoring. The
ascending pathway included patients with a COVID-19 diagnosis deemed appropriate for home monitoring. A pulse oximeter was issued to patients on both pathways. Patients were asked to record their pulse oximetry derived oxygen saturation ($s_{O_2}$) three times daily, and they were proactively contacted by phone daily, and asked to contact the hospital if their $s_{O_2}$ fell below 92% or if other symptoms worsened. Asymptomatic patients were considered for discharge by fourteen days. The other virtual ward, COVID-19 Oximetry @Home, was led by primary care providers, where pulse oximeters were issued to COVID-19 positive patients. Patients were supported to self-escalate if their $s_{O_2}$ fell to 92% or below, and to call a low-acuity hotline or their general practitioner if it fell to 93% or 94% or if other symptoms worsened.

Various media articles suggested that the public purchase pulse oximeters and monitor themselves, hence forming a self-referral pathway. The recommended decision boundaries were the same as for the two virtual wards.

**Pulse oximeter performance**

The ISO 80601-2-61:2019 standard gives performance requirements for pulse oximeters. Pulse oximeter manufacturers are not mandated to comply with this standard, but compliance is generally expected for NHS devices. All of the virtual ward devices claimed compliance with this standard.

The standard requires that manufacturers perform a study to quantify their pulse oximeter’s accuracy. This can be a desaturation study on healthy volunteers, using either a CO-oximeter or another pulse oximeter as reference. Patient studies are also permitted provided that a CO-oximeter is used as reference. Root-mean-square error may not exceed 4% for $s_{O_2}$ in the range 70%-100%. The standard recommends that study subjects ‘should vary in their physical characteristics to the greatest extent possible’ to permit broad application to different patient groups. It does not prescribe sample size or analysis methodology. Publication and peer-review of these studies is not required.

Assurance of the accuracy of the pulse oximeters deployed during the pandemic was sought, since they had little history of NHS use and the results of previous investigations of pulse oximeters performance have been variable.

A desaturation study on six non-US Food and Drug Administration (FDA)-approved low-cost pulse oximeters found that 4/6 did not meet ISO 80601-2-61 requirements. The study included two device models that are earlier versions of devices used for the virtual wards: the Contec CMS50DL (which did meet the requirements) and the Beijing Choice MD300C23 (which did not). Two further clinical studies found that several portable pulse oximeters (including the Contec CMS50D, CMS50DL and ChoiceMMed MD300C52) performed acceptably in patients with relatively high $s_{O_2}$, but offered no definitive comment for patients with lower $s_{O_2}$.

Other studies which have considered the accuracy of portable pulse oximeters similar to those investigated for this study, have found that their measurements agree closely with gold standard measurements from CO-oximeters. These include studies which consider these devices’ use at high altitude (a common use case for these devices, due to their easy portability).

Concerns had also been raised about pulse oximeter performance for patients with darker skin. A California-based research group has undertaken desaturation studies to evaluate the relationship between pulse oximeter performance and skin tone, finding that skin tone, gender and $s_{O_2}$ range were consistent predictors of bias for various pulse oximeters. They found that $s_{O_2}$ was overestimated in dark-skinned, hypoxic individuals ($s_{O_2} <80\%$), and concluded that this effect could be clinically significant. Similar findings were presented in a letter to the *New England Journal of Medicine* and an early study of ear pulse oximetry also found that performance was poorer for darker skinned patients. Other studies have found no relationship between pulse oximeter performance and skin colour, including one study on the Contec CMS50D.

Factors other than skin tone may affect pulse oximeter performance. These devices rely on arterial flow pulsatility to measure $s_{O_2}$, so conditions like peripheral arterial disease (PAD), which reduce pulsatility, may affect function. The presence of carboxyhaemoglobin in the blood (for instance, due to smoking) also causes pulse oximeters to overestimate $s_{O_2}$.

While questions remain about the performance characteristics of home-use pulse oximeters, there is evidence that they reduce COVID-19 mortality in some populations. A South African retrospective cohort study of high risk COVID-19 positive individuals found that mortality was 48% lower in individuals issued with a pulse oximeter than in those without one.

Previous studies have not investigated the performance of pulse oximeters used within the context of these virtual wards, considering the 92% decision threshold. For this study ‘false positive’ was defined as a $s_{O_2}$ reading less than 92% for a patient whose ‘true’ $s_{O_2}$ is greater than or equal to 92%. False positive rate (FPR) was the proportion of negative results recorded as positive by the pulse oximeter. A ‘false negative’ was defined as a $s_{O_2}$ reading greater than or equal to 92% for a patient whose true $s_{O_2}$ is less than 92%. False negative rate (FNR) was defined as the proportion positive results recorded as negative by the pulse oximeter. The objective of this work is to estimate the impact of performance characteristics of these devices on the home-use pathways in terms of FPR and FNR. This work also investigated the effects of factors such as skin tone, smoking status and PAD status on FPR and FNR.
METHODS

The accuracies of the five pulse oximeters were evaluated by comparison with a hospital-use pulse oximeter, an M191BL digital probe (Philips, Eindhoven, the Netherlands) connected to an IntelliVue X3 Patient Monitor (Philips, Eindhoven, Netherlands). The test pulse oximeters were: (1) Oxywatch MD300C19 (ChoiceMMed, Beijing, China), (2) Oxywatch MD300C29 (ChoiceMMed, Beijing, China), (3) PC-60F (Creative Medical, Shenzhen, China), (4) Contec CMS50D (Contec Medical Systems, Hebei, China) and (5) AM801 (Shenzhen Med-link Electronics Tech, Shenzhen, China).

Use and maintenance of the M191BL digital probe and IntelliVue X3 Patient Monitor was in accordance with the manufacturer’s instructions and standard hospital practice. Measurements made using this device therefore represent ‘standard’ measurements that would be performed in a hospital setting in the absence of any home-oximetry pathway.

This evaluation was authorised by the Guy’s and St Thomas’ NHS Foundation Trust Quality Improvement and Patient Safety team. Informed consent was taken from all conscious participants, and the responsible nursing team advised on the inclusion of sedated patients. Participants’ clinical care was unaffected by inclusion in the study.

Patient and public involvement

Patients and members of the public were involved in the determination of this study’s value, and in the design of the information sheets issues to participants.

Data collection

Patients with \( s_{O_2} < 85\% \) (according to inpatient monitoring equipment) were excluded, together with unstable patients likely to experience acute changes to \( s_{O_2} \) (eg, due to respiratory support or positional changes). Each participant underwent up to three sets of measurements, at least 90 min apart.

Patients, or their nursing teams, who met the inclusion criteria and were inpatients on the intensive care units, recovery and respiratory wards, were approached for inclusion in the study. Fifty patients were recruited in total.

Measurements were collected in pairs from a test device and reference device, placed simultaneously on the ring and index finger. The measurement was taken 30s after placement of the pulse oximeter, in-line with the greatest stabilisation period recommended by any of the device manufacturers. This was repeated three times for each test device. Test order and finger selection were randomly determined for each session.

Skin pigmentation was quantified using the Fitzpatrick Skin Pigmentation (FSP) scale and recorded alongside \( s_{O_2} \) measurements. Patients’ ages, genders, PAD and smoking statuses were extracted from their medical notes.

Data analysis

The effects of device model, gender, age, smoking status, PAD status, FSP score and \( s_{O_2} \) (as recorded using the reference device) on bias, and on the likelihoods of a false positive or false negative result were evaluated using multiple effects analyses.

The effect on bias was modelled using linear multiple regression analysis, with significance calculated using t-tests. The effects on false positive and false negative likelihoods were modelled using binary logistic regression, with significance calculated using the Wald test. Results were considered significant where \( p<0.05 \).

In all cases, an initial analysis was conducted on the data set as a whole with subsequent analyses performed for each test device model individually. All analyses were performed using the IBM SPSS software, release 27.0.1.0.

RESULTS

Subject characteristics are shown in table 1.

Multiple regression analyses

Aggregated data set

Table 2 shows the results of the multiple effects analyses with the aggregated data from all five test pulse oximeters. This allows investigation of the differences between the reference device and the general performance of test pulse oximeters.

The relationships between bias and its significant predictors are shown in figure 1.

Bias is plotted against device model (top left), FSP (top right), subject \( s_{O_2} \) as measured by the reference device...

On average, measurements from all five test pulse oximeters were lower than those by the reference device. The mean biases were −1.1% (AM801), −2.1% (CMS50D), −0.6% (MD300C29), −1.3% (MD300C29) and −0.1% (PC-60F). The associated SD were 2.2%, 3.3%, 2.8%, 3.7% and 2.6%, respectively. These SD overestimate devices’ RMS error, as they also include reference device error (quoted by the manufacturer as 2.5% RMS).

Subject $s_O^2$ (as measured using the reference device) was a significant predictor of bias, false negative likelihood and false positive likelihood. Bias became more

### Table 2  Significance and effect size of factors on bias, false positive (FP) likelihood and false negative (FN) likelihood

<table>
<thead>
<tr>
<th>Factors</th>
<th>Bias P value</th>
<th>Coefficient (95% CI)</th>
<th>FP likelihood P value</th>
<th>OR (95% CI)</th>
<th>FN likelihood P value</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>0.735</td>
<td>−0.076 (−0.514 to 0.363)</td>
<td>0.579</td>
<td>0.863 (0.512 to 1.453)</td>
<td>0.938</td>
<td>1.055 (0.272 to 4.098)</td>
</tr>
<tr>
<td>Age</td>
<td>0.017</td>
<td>−0.016 (−0.029 to −0.003)</td>
<td>0.261</td>
<td>1.010 (0.992 to 1.029)</td>
<td>0.062</td>
<td>1.038 (0.998 to 1.080)</td>
</tr>
<tr>
<td>Smoking status</td>
<td>0.089</td>
<td>0.528 (−0.081 to 1.136)</td>
<td>0.933</td>
<td>1.032 (0.496 to 2.146)</td>
<td>0.007</td>
<td>0.121 (0.026 to 0.560)</td>
</tr>
<tr>
<td>PAD status</td>
<td>0.672</td>
<td>−0.121 (−0.680 to 0.439)</td>
<td>0.736</td>
<td>1.115 (0.593 to 2.098)</td>
<td>0.080</td>
<td>7.872 (0.779 to 79.58)</td>
</tr>
<tr>
<td>FSP</td>
<td>&lt;0.001</td>
<td>−0.455 (−0.639 to −0.279)</td>
<td>0.071</td>
<td>1.263 (0.980 to 1.628)</td>
<td>0.006</td>
<td>0.344 (0.161 to 0.735)</td>
</tr>
<tr>
<td>Device model</td>
<td>&lt;0.001</td>
<td>0.309 (0.167 to 0.451)</td>
<td>0.013</td>
<td>0.801 (0.672 to 0.953)</td>
<td>0.014</td>
<td>1.511 (1.087 to 2.100)</td>
</tr>
<tr>
<td>Ref. $s_O^2$</td>
<td>&lt;0.001</td>
<td>−0.155 (−0.209 to −0.100)</td>
<td>&lt;0.001</td>
<td>0.694 (0.613 to 0.785)</td>
<td>&lt;0.001</td>
<td>2.415 (1.604 to 3.636)</td>
</tr>
</tbody>
</table>

Bias results are from multiple regression analysis. FP and FN likelihood results are from binary logistic regression analysis. Significant p values are shown in bold. 95% CIs are shown in parentheses after coefficient and OR values.

FSP, Fitzpatrick Skin Pigmentation; PAD, peripheral arterial disease; $s_O^2$, oxygen saturation.

Figure 1  Plots showing the relationship between measurement bias (difference between test oximeter and reference oximeter measurements) vs (1) test pulse oximeter models (top left), (2) Fitzpatrick skin pigmentation scale (top right), (3) subject $s_O^2$, as measured by the reference oximeter (bottom left), and (4) subject age (bottom right). Zero bias is shown for each plot with a red dashed line. Error bars show the bias SD associated with each data group. $s_O^2$, oxygen saturation.
negative with increasing $s_pO_2$. The directions of the effects on false positive and FNR indicate that false results are more likely for patients whose $s_pO_2$ is close to the threshold value of 92%, as would be expected. Device model was also a significant predictor of bias, false negative likelihood and false positive likelihood.

Skin tone was a significant predictor of bias and false negative likelihood. Bias was more negative for subjects with darker skin. False negative likelihood was lower for subjects with darker skin, this is consistent with a stronger tendency for test pulse oximeters to under-read for these subjects.

A significant effect existed between smoking status and FNR, with current smokers less likely to receive a false negative result than non-smokers.

The only other significant effect was between subject age and bias. The test pulse oximeters’ tendency to under-read was greater for older subjects. However, the effect size was too small to be clinically significant.

### Results for individual test pulse oximeters

Factors (such as age, gender, FSP score smoking status, PAD and reference $s_pO_2$ range) that may affect measurement bias, FPR and FNR were considered, and Table 3 summarises the statistically significant results from the multiple effects analyses for individual test pulse oximeters. A summary of all the results (including non-significant results) is given in the online supplemental material.

Subject $s_pO_2$ (as measured using the reference device) had a significant effect on bias for all PPOs except the CMS50D and MD300C29. The effect direction was the same in all cases, with bias becoming more negative for greater $s_pO_2$. The effect of $s_pO_2$ on false positive likelihood was also fairly consistent, being significant for all devices other than the MD300C19 and MD300C29. The direction of this effect was also the same in all cases, indicating that false positive likelihood increases for subjects whose $s_pO_2$ is close to the 92% threshold value. The equivalent effect on false negative likelihood (with likelihood increasing for $s_pO_2$ close to the threshold) was significant for the AM801 only.

Skin tone had a significant effect on bias for the MD300C19 and MD300C29, with bias becoming more negative for subjects with darker skin. PAD status had a significant effect on bias for the MD300C29 and PC-60F. The effects were in opposite directions, with the MD300C29 tending to underestimate saturation level compared with the reference device for subjects with a PAD diagnosis, and the PC-60F tending to overestimate saturation level compared with the reference device.

The only other significant effect was between smoking status and bias for the MD300C19. This PPO was more likely to over-read in current smokers.

### Observed error rates

The observed FNRs and FPRs in the study population are shown in Table 4. Any attempt to generalise the results shown in Table 4 should account for variation in population characteristics, particularly $s_pO_2$ distribution. The $s_pO_2$ distribution of the study population is therefore presented in online supplemental figure 2.

False negative and false positive likelihood vary depending on subject $s_pO_2$, being greatest for subjects with $s_pO_2$ close to the threshold value of 92%. The PPOs’ tendency to under-read might be expected to produce a greater FNR than FPR, however the opposite was true in this case. This was due to the high proportion of positive cases with $s_pO_2$ close to the threshold value (46.5% within
2% of this value), so small biases were more likely to cause false negative results. The proportion of negative cases close to the threshold was smaller (20.1% within 2% of this value) so larger biases were generally required to produce false positive results.

The results indicate that, for the study population, both the FPR and FNR were appreciable, with approximately one in four positive cases incorrectly identified as negative, and one in nine negative cases incorrectly identified as positive by the portable pulse oximeters.

**DISCUSSION**

**Level of agreement**

All five portable pulse oximeters tended to under-read $s_{O_2}$ relative to the reference device. The bias SD was between 2% and 4% in all cases. This would be expected to include the error associated with the portable pulse oximeters, as well as that associated with the reference device, suggesting that all devices meet the 4% RMS error requirement in the ISO 80601-2-61 standard. However, this error would be enough to affect clinical decisions, particularly for patients with $s_{O_2}$ close to a decision threshold (as a small bias may be sufficient to push these patients’ measurements across the threshold).

**Pulse oximeter performance**

This study evaluated the individual performance of five portable pulse oximeters, and also their aggregated performance, to identify performance attributes they shared in common. More significant effects were identified for the aggregated data set than for the individual devices. This is likely partly due to the smaller sample size, and consequently lower analytical power, for the individual models. Power was particularly limited for the analyses of false negative likelihood, due to the small individual models. Power was particularly limited for the size, and consequently lower analytical power, for the devices. This is likely partly due to the smaller sample shared in common. More significant effects were identified for the aggregated data set than for the individual devices. These devices appear to under-read $s_{O_2}$ more for skinned subjects whose $s_{O_2}$ is very low.

The aggregate analysis showed that pulse oximeter error (whether expressed in terms of bias, false positive likelihood or false negative likelihood) varied with patient characteristics and, independently, with device model. A personalised approach to $s_{O_2}$-based decision-making could account for both effects by modelling these dependencies.

**Dependency on $s_{O_2}$**

The variation in pulse oximeter performance with $s_{O_2}$ (as measured by CO-oximetry) is well documented, and is permitted by the ISO 80601-2-61 standard. Pulse oximeters are known to perform more poorly at lower $s_{O_2}$. This study suggests that a separate effect exists between the performance of the portable pulse oximeters, and the ‘hospital standard’ device used as reference. The direction of this dependency was the same for all five pulse oximeters, with all five under-reading by a greater amount at higher $s_{O_2}$. However, this effect only reached statistical significance for three of the devices, when considered individually.

Within the home-use oximetry pathways, the effect of this dependency is likely to be modest. Irrespective of this effect, the patients most likely to be wrongly classified remain those whose $s_{O_2}$ is closest to the decision threshold. The relationship between bias and $s_{O_2}$ is not strong enough to much increase the likelihood of false positive results among patients with high $s_{O_2}$ (and consequently higher bias).

**Dependency on skin colour**

Previous studies have reported that pulse oximeters tend to over-read $s_{O_2}$ relative to CO-oximeters, for dark-skinned subjects whose $s_{O_2}$ is very low. This study did not include subjects with $s_{O_2}$ low enough to be influenced by this effect. However, the data do suggest that a separate effect exists for the portable pulse oximeters. These devices appear to under-read $s_{O_2}$ more for subjects with darker skin than those with lighter skin. This effect would be expected to increase the likelihood of false positive results, and reduce the risk of false negative results in this patient group. There was evidence of this effect in the data, although it did not achieve significance in all cases.

In the context of the home oximetry pathways, this effect could mean that failure to appropriately escalate is
less likely for darker-skinned patients, and that inappropriate escalation is more likely.

Other dependencies
Other dependencies were found to exist between age and bias, between smoking status and FNR, between PAD status and bias for specific pulse oximeter models, and between smoking status and bias for one model. These effects were not consistent for different devices or between the aggregate data set and the device specific data sets. It is difficult to draw firm conclusions from these results, however, they would merit further investigation if widespread deployment of the affected devices were planned.

Observed and forecasted error rates
The observed FNR and FPR for all five PPOs are given in table 4. These parameters are strongly dependent on population characteristics, particularly $\text{SpO}_2$ distribution. Caution should be exercised before generalising these results to other populations. This should be informed by a comparison of respective population characteristics (see table 1 and the online supplemental material). These results provide a first estimate of the likely practical implications of deployment of these devices.

FNR implications
Within a home-oximetry pathway, a high FNR may result in failure to detect clinical deteriorations, representing a patient safety risk. This study found the FNR to be appreciable in the study population (table 4). The associated risks may be reduced by emphasising the importance of other information available to the responsible clinical team, including regular communications and the symptom diary—in line with elaborated advice.2

FPR implications
The home oximetry pathways were intended to reduce hospital bed occupancy by allowing some COVID-19 patients, who would otherwise occupy a hospital bed, to instead be monitored at home. Use of pulse oximeters with high FPR would tend to increase (re)admission and reduce this benefit. The results of this study suggest that this effect would be modest in size for these pulse oximeters.

Limitations
Choice of reference measurement device
Due to practical constraints on the protocol, a widely used hospital pulse oximeter was used as the reference device in this study, rather than a CO-oximeter. This study, therefore, gives an indication of how portable pulse oximeters compare to hospital-based oximetry. It is not possible to definitively determine the devices’ accuracies from these results.

$s\text{O}_2$ range
The relevant FDA and ISO standards require that pulse oximeters are validated across an $\text{SpO}_2$ range of 70%–100%. This study only included patients with $\text{SpO}_2 \geq 85\%$. It is therefore not possible to comment on these devices’ performance for $\text{SpO}_2 < 85\%$. In practice, errors large enough for a patient with $\text{SpO}_2 < 85\%$ to receive a measurement $\geq 92\%$ are unlikely; this limitation is therefore unlikely to affect interpretation with respect to the home oximetry pathways.

Practical challenges
This study considered a small subset of the home-use oximeters available on the market. A comprehensive analysis of all available models would be logistically challenging and might, for example, require aggregation of data from multiple centres over longer time periods.

FURTHER WORK
Further research, with a larger sample size and use of CO-oximeter measurements as reference, is required to precisely determine these devices’ bias. A simple approach to accounting for this bias would be to apply the offset back to observed data, or to modify the threshold value. More sophisticated techniques, which account for the test’s receiver-operator characteristics, are also available.26 27 Another alternative would be to use a longitudinal deterioration model as described by Prower et al.28

CONCLUSIONS AND FINAL REMARKS
All the portable pulse oximeters investigated outperformed the standard to which they were manufactured. When used within a home-use pathway, pulse oximetry measurements must be interpreted in isolation, without the benefit of the other information available to a clinician reviewing a patient who is physically present (such as observable breathlessness and work of breathing). This creates an increased risk of error. Notwithstanding methodological limitations discussed, the variation between the pulse oximeters upholds this consideration.

This indicates the use of trend analysis for more robust indications that might mitigate inter-oximeter variation. The practice of observing baselines and trends goes beyond the initial SOP documents but is clear in elaborated guidance.2 Other risk mitigations include use of conservative escalation thresholds, and inclusion of other clinical information from patients’ symptom diaries and from telephone contact with clinicians. Advanced approaches might include use of a-priori data and analyses to tailor these approaches for individual patients and/or devices.

There was also evidence of performance being dependent on subject skin tone. This would tend to increase the likelihood of inappropriate escalations of patients with darker skin, and reduce the risk of these patients not being escalated when appropriate.
References


