Study Title: Performance evaluation study on AI aided interpretation of results from Innova COVID-19 Antigen Lateral Flow Test for asymptomatic users

Internal Reference No: PER-001 - AI Performance Evaluation Study Protocol

Ethics Ref: N/A

Version No: v1.0

Chief Investigator:	**Andrew Beggs**
Sponsor:	**DHSC**
Signatures:	**Tom Fowler**
Date:	**04 March 2021**

Confidentiality Statement

This document contains confidential information that must not be disclosed to anyone other than the Sponsor, the Investigator Team, host organisation, and members of the Research Ethics Committee and Regulatory Authorities unless authorised to do so.

PROTOCOL AGREEMENT

I confirm that I have read and understand the protocol specified below. In my formal capacity as Chief Investigator I agree to conduct this study in accordance with the requirements of this protocol, and my duties include ensuring the rights, safety, privacy and well-being of the study subjects enrolled under my supervision and providing DHSC with complete and timely information, as outlined in the protocol. It is understood that all information pertaining to the study will be held strictly confidential and that this confidentiality requirement applies to all study staff at this site. Furthermore, on behalf of the study staff and myself, I agree to maintain the procedures required to carry out the study in accordance with accepted GCP principles and to abide by the terms of this protocol.

Chief Investigator:

Signature: 

Date: 04 March 2021

Name: Andrew Beggs

AMENDMENT HISTORY

Amendment No.	Protocol Version No.	Date issued	Author(s) of changes	Details of Changes made
1	1.0	04 Mar 21	Nataliya Kuras	First issue

Table of Contents

1 ABBREVIATIONS

Abbreviation	Description
Ag-RDTs	Antigen Rapid Diagnostic Tests
AE	Adverse event
ADE	Adverse Device Effect
API	Application Programming Interface
AI	Artificial Intelligence
ATS	Asymptomatic Testing Site
CFR	Case Report Form
DHSC	Department of Health and Social Care
EC	Ethics Committee (see REC)
GCP	Good Clinical Practice
GDPR	General Data Protection Regulation
LFD	Lateral Flow Device
MHRA	Medicines and Healthcare products Regulatory Agency
NHSD	National Health Service Digital
RDT	Rapid Diagnostic Test
REC	Research Ethics Committee
SAE	Serious Adverse Event
SADE	Serious Adverse Device Effect
UADE	Unanticipated Adverse Device Effect
UON	Unique Organisation Number

2 BACKGROUND AND RATIONALE

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a betacoronavirus responsible for coronavirus disease-19 (COVID-19)[1]. Assessment of the extent of infection has largely been based on real-time polymerase chain reaction (RT-PCR) that identifies the virus in those with an active or recent infection. However, in the UK, antigen testing is predominantly performed in those presenting with symptoms (which may represent a small fraction of those infected with SARSCoV-2) and it cannot detect past infection. Moreover, it has been reported that early estimates of 80% of the infections being asymptomatic were overestimated, and such figures were subsequently revised to 17-20%[2].

Historically, a test has always been considered a support for clinical diagnosis, not a substitute, and to this extent the timing of the infection and the viral load at time of test are variables that should be considered in the overall assessment. Recent evidence demonstrated that 60% of infected individuals without symptoms were undetected, of which more than 1/3 had a high enough viral load to represent a high risk of infection to others[3].

LFDs are relatively easy-to-use rapid test that can be performed at home settings by lay user or near the point of care, without the need for laboratory infrastructure or expensive equipment. There are two types of SARS-CoV-2 LFDs: SARS-CoV-2 virus antigen(s) tests, and antibody tests that detect one or more types of antibodies produced by the host immune response against the virus.

Preliminary results conducted by a Joint PHE Porton Down and the University of Oxford[4] study showed that performance characteristics of such LFD appear to be good with a low failure rate with a demonstrated specificity of 99.6%. A joint University of Liverpool / DHSC pilot demonstrated that the accuracy of self-read LFDs was lower than those read by trained operatives.

Most SARS-CoV-2 antigen LFDs require nasal and/or oral fluids samples. A key step of the entire testing strategy that this study is intending to address and research is the mechanism of reporting the LFD results. This is typically read by the individual performing the test and in this study scenario, such individual would need to input the result of the test following the LFD manufacturer's test interpretation instructions, manually into the process flow (by answering a question in the digital test reporting service).

DHSC and NHSD intend to enhance the existing Self-Test LFD service, (using an Innova device – purchased by DHSC and MHRA derogated for lay person use (MHRA ref. number: DEU/012/2020/003) and accessed from the gov.uk website)) with additional functionality. In order to render the interpretation of the test less subjective and less prone to user error and to decrease the risk of falsification of reported results, it is proposed to use an AI based technology that will request the user take a picture of the test result. The above process will allow the user to submit their LFT result via the camera on a mobile device and for the result interpretation to be made digitally by AI reader technology and for the result to be returned to the user.

This Performance Evaluation Study is aiming to collect data to verify the performance claims under the anticipated conditions of use.

The whole Performance Evaluation Study will be split into two parts – Sub-studies, which will be run in parallel.

Sub-study 1 (source: ATS photos) – Image collection which were captured in Asymptomatic Test Sites (ATSs). LFT will be performed as per manufacturer's IFU. Then the fully trained operative will read the test results, take a photo of completed test and enter/upload results on NHSD web service as per written instruction. The picture of completed test will be taken at the point at which the result is interpreted/decided.

The main objective of this Sub-study is to test and validate that AI algorithm can meet the primary objective analytical performance measures detailed in Section 3.

Sub-study 2 (source: photos from NHS and ASC users) - Antigen Lateral Flow Device (LFD) test subjects who have identified themselves as working within NHS primary care (GP's, NHS dentistry, community pharmacists and NHS optometrists etc), or are linked to a subset of adult social care homes (staff, visitors and visiting professionals) will be invited to take images on their mobile devices (such as mobile phone or tablet) and upload these images using the NHSD web service. Study participants will be self-tested using an Innova LFT device (on naseal swab samples) – purchased by DHSC and MHRA derogated for lay person use (MHRA ref. number: DEU/012/2020/003) and accessed from the gov.uk website)). Self-testing will be performed as per manufacturer's instructions for use (IFU). Non-Innova LFT devices will be removed from the data set analysed.

The picture of completed test will be taken at the point at which the result is interpreted/decided.

The main objectives of this Sub-study is to test and validate that the digital service and integrated AI algorithm which wll be used by subjects can meet the observational objectives measures detailed in Section 3.

This performance evaluation study will use AI technology (provided by a third party) to interpret the results of LFD tests from photos. Study participants in Sub-study 2 submit their photo when they register and report a test result. This performance evaluation study will test the hypothesis that a digital reader will:

- Facilitate and, in some cases, improve consistency and accuracy of result interpretation compared to a self-read test.

- potentially identify more positive results in asymptomatic subjects than a human read.

In addition to above, this performance evaluation study will improve our understanding of the use of the online “photograph taking process” in practice and to determine whether the product is easy to use and safe in the intended context of use.

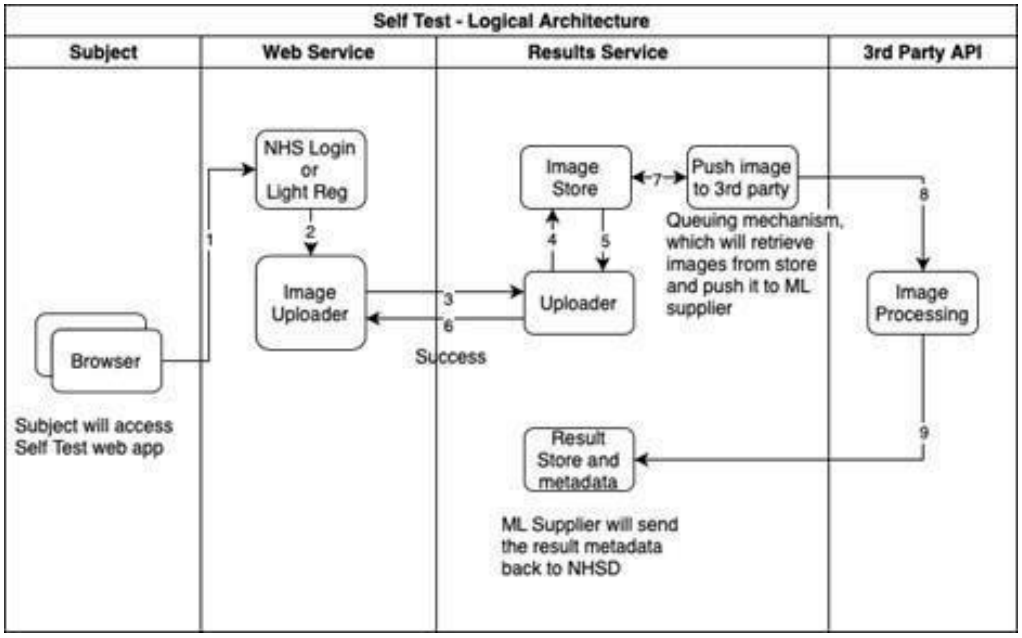


Figure 1: Self-Test – Logical Architecture

If the algorithm is found to be at least as good as the average user read, this evidence will support an application to MHRA to allow the use of the algorithm to help the user interpret tests for mass testing, using the same LFD self-test digital service.

3 STUDY OBJECTIVES

Note: Any monitoring of the efficacy of the physical Lateral Flow test device is out of scope of this study.

Objectives:	Outcome Measures:
Primary Objective:	
1) To evaluate the analytical performance of the AI (decision support software)	1) The outcome of the AI interpretation of uploaded images will be compared to the individually reported results. Where the test line is visible and interpretable, comparison between the AI-enabled read and human read of a lateral flow test shows that the algorithm reliably matches correctly interpreted human self-reported results with > 95% sensitivity and specificity. Void returned results (no control line) and images which fail to meet all the quality criteria (defined below) will be excluded from this calculation. We will report this accuracy separately for positive and negative classes of tests. In those cases where there is a discrepancy between the human and algorithm reads, a further manual assessment will be conducted to determine 'ground truth' of the read (i.e. whether the test is positive, negative or void).

Observational Objectives:	
1**)** To assess the ability of the AI algorithm to perform its intended image recognition function in real world settings (image quality)	1) 95% of the images received from users are of "sufficiently quality", defined as follows: • Taken from directly above the device; • Without bright reflections or dark shadows falling over the test or control lines of the LFT; • Non-blurry (C and T can be distinguished clearly); • Taken right side up (i.e., with the barcode at the top and specimen collection at the bottom); • The test comprises at least 50% of the overall image, and the entirety of the test is present in the image; • The resolution of the image is at least 1920 pixels (on the vertical); • Oriented at less than a 30-degree angle to vertical. These quality determinants allow the image to be readable by the algorithm and by an independent evaluator if needed. We expect this to increase over time as users get more experienced with the process

2) To assess the level of image anomaly detection (i.e., test not present, picture of something that is NOT a LFD, picture of an LFD with counterfeit lines, etc.)	2) Of those images reported as "anomalies", an independent reader will assess the images and decide whether the image is a potentially fraudulent image (for example lines deleted and/or added with pen/marker/tippexed) or genuine user error mistake. The categories above will be calculated as a single % anomaly detection as a proportion of the overall number of anomalous images.
3 To assess the effectiveness and efficiency of the digital service in its intended use setting by the intended user population (reference to the SIU	3) At least 95% of effectiveness and efficiency will confirm the usability of the image-capture element of the digital service, regardless of the device used for image capture (i.e., phone, tablet)
4) To assess the performance of the algorithm in identifying void tests (i.e., tests performed incorrectly), where a control line is missing.	4) The % of tests reported as void by the algorithm will be compared to the % reported by individuals as void (study participants will be asked to submit photos of void tests. In addition, where there is a discrepancy (whereby the user's response differs from an algorithm's read, these will be referred to an independent off-line review of the image. The % of those images correctly identified as void will be reported.

5) To assess the implication of the types of devices used (Tablets, PCs, phones of various models) on the image quality and therefore implication for algorithms	5) A breakdown of the device types, with % of images which meet the quality requirements defined above, particularly linked to image quality (resolution, pixels, lighting/flash)

4 STUDY DESIGN AND PROCEDURE

This Performance Evaluation Study is aiming to collect data to verify the performance claims under the anticipated conditions of use.

The whole Performance Evaluation Study will be split into two parts – Sub-studies, which will be run in parallel.

Sub-study 1 (source: ATS photos) – Images captured in Asymptomatic Test Sites (ATSs). LFT will be performed as per manufacturer's IFU. Then the fully trained operative will read the test results, take a photo of completed test and enter/upload results using a NHSD web service as per written instructions. The picture of completed test will be taken at the point at which the result is interpreted/decided.

The images will be analysed by the AI algorithm which will return an interpreted result. This result will be stored in the NHSD image store.

Sub-study 2 (source: photos from NHS and ASC users) - Antigen Lateral Flow Device (LFD) test subjects who have identified themselves as working within NHS primary care (GP's, NHS dentistry, community pharmacists and NHS optometrists etc), or are linked to a subset of adult social care homes (staff, visitors and visiting professionals) will be invited to take images on their mobile devices (such as mobile phone or tablet) and upload these images using the NHSD web service. Study participants will be self-tested using an Innova LFT device (on naseal swab samples) – purchased by DHSC and MHRA derogated for lay person use (MHRA ref. number: DEU/012/2020/003) and accessed from the gov.uk website)). Self-testing will be performed as per manufacturer's instructions for use (IFU). The picture of completed test will be taken at the point at which the result is interpreted/decided.

The webservice for Sub-study 2 is written in React.j and is capable of processing requests from various types of mobile devices. The webservice will allow the user to use their device's camera to take a photograph of their LFD and submit this to NHSD. Upon receipt of the image, it is written to an image store prior to being processed and analysed by the vendors' AI component.

The AI's interpretation is then passed to a results database where is it logged alongside the test subject's own asserted result.

When an void result is observed (no control line), the study participant will be instructed to take another test with a new test kit. Study subjects will be asked to take a picture of the void test before proceeding with a new test.

AI interpreted result will not be returned to the user during either sub-study, but instead analysed and kept as evidence to assess the accuracy of the algorithm reading against the reported outcome by the lay user and trained operator.

During this study we have an intention collecting data from four (A, B, C and D) different variations of Innova device (see Figure 2 below).

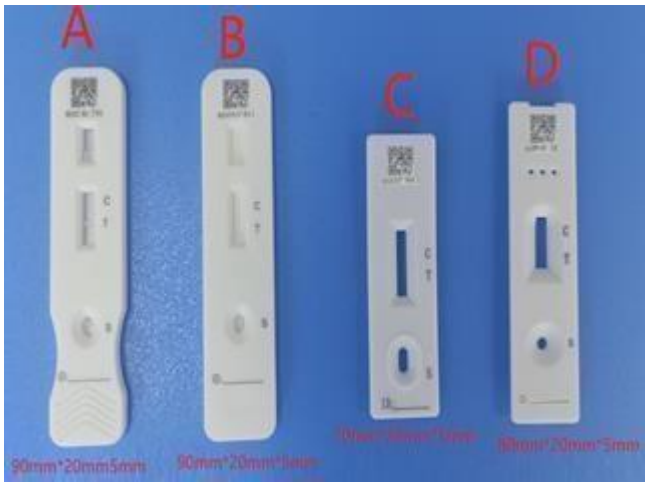


Figure 2: Four variations of Innova device

In addition to the manually read and AI generated results, a third-party independent inspection will be performed for a subset of images to create a 'most trusted' result against which the sensitivity and specificity of the AI read can be measured and to demonstrate the effectiveness of the AI algorithm, showing that it is 'at least as good as the average user read'. This will be performed for both sub-studies.

The samples to be reviewed by the independent inspectors will fall into two categories. Firstly, all samples where there is a discrepancy between the AI result and the test subjects reported result.

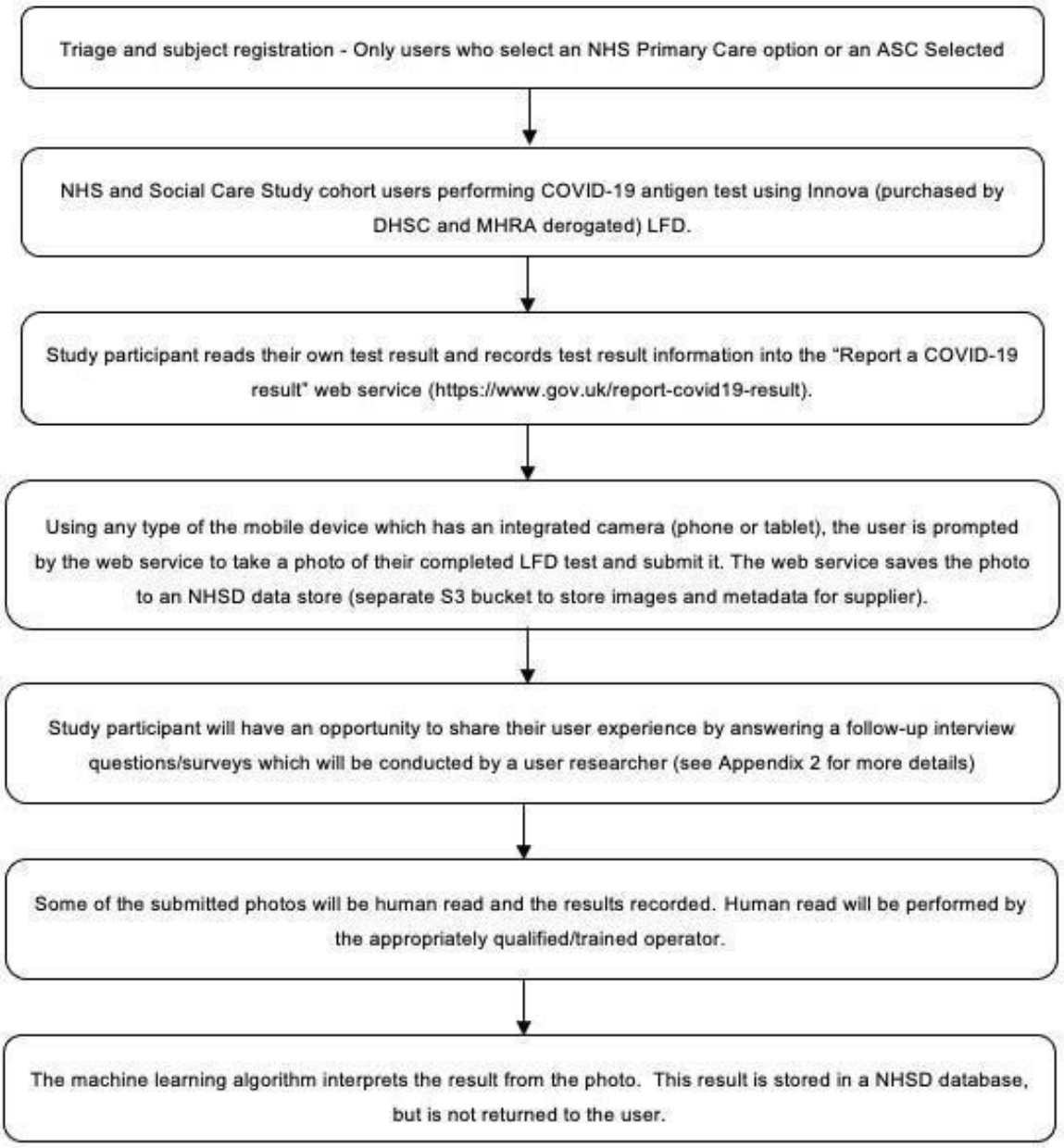
Secondly, a random sample of images where there is consensus between all the results (human and AI). This will be a selection of 10% of the consensus positives alongside at least 50 consensus negatives per week.

Assessment will take place via interpretation of the submitted image of the test subjects LFD. Each inspector will individually assess the image and record their interpretation of the result against the unique image identifier. The individual assessments will subsequently be compared to determine a consensus amongst the inspectors and a final 'third party review' data point will be recorded against the unique image identifier.

Note that this data point may not represent a unanimous view amongst the three inspectors, but will be a majority view at least. On a daily basis the results of each individual assessment and the 'third party review' data point will be provided to the AI vendor and to DHSC for review.

These individual inspectors should be experienced in making clinical assessments and understand the rigour required in order to achieve a reliable outcome. To perform the assessment, 3 individuals shall be trained by interpreting a practice suite of images, recording their verdicts and comparing these to determine a consensus view. Each individual will require an internet connected computer which is able to receive the images to be assessed from the AI vendor and return their results to the AI vendor and to DHSC.

Sub-study 2 procedure ilustrated in Flowchart 1 below.



Flowchart 1: Sub-study 2 process flow

5 STUDY PARTICIPANTS

The study will be undertaken within 2 settings:

- Sub-study 1: Asymptomatic Testing Sites (ATSs) and
- Sub-study 2: NHS primary care staff, and Adult Social Care staff, visiting professionals and visitors.

We expect widespread usage of the webservice during sub-study 2 by a broad range of, healthcare professionals within the NHS cohort and by an equally diverse group of care home staff and visitors from the general public.

User age ranges across all cohorts will likely be 18-65 years old (care home visitors may also be outside of this range), representing all gender identities and a wide cross-section of ethnicities.

Geographic spread will be largely dependent on the number of Asymptomatic Test Sites, NHS trusts and Care Homes that participate in the pilot but will represent all English regions.

Inclusion Criteria:

- Participant is willing to participate in the study and agrees with the privacy statement
- Aged 18 years or above
- Adolescents aged 12-17 (self-test and report with adult supervision)
- Without any common COVID-19 symptoms
- Able (in the Investigators opinion) and willing to comply with all study requirements
- Children under 12 (should be tested and reported by an adult)

Exclusion Criteria

The participant may not enter the study if ANY of the following apply:

- Participant does not agree with privacy statement.
- Has any common COVID-19 symptoms.
- Any other significant disease or disorder which, in the opinion of the Investigator, may either put the participants at risk because of participation in the study, or may influence the result of the study, or the participant's ability to participate in the study.

6 RECRUITMENT AND STUDY SITES

We will be using images captured at Asymptomatic Testing Sites by trained operatives as part of the ATS LFD current registration and reporting service. We will recruit NHS staff and Adult Social Care staff and visitors for the pilot cohort by filtering responses given during the current Self-report LFD web registration journey.
For NHS use case, we have the opportunity to filter based on NHS setting (eg General Practice, Dentistry, Acute (hospital) trust etc).

For Adult Social Care homes, we will filter on the Adult Social Care homes that have agreed to be part of the Sub-study 2. We will select such care homes that have a significant number of full-time care staff in order to provide the scale of staff and visitors required for the study.

7 INFORMED CONSENT

User consent for using the photo is covered in the privacy statement.
https://www.gov.uk/government/publications/coronavirus-covid-19-testingprivacyinformation/testingfor-coronavirus-privacy-information-quick-read--2

8 DEFINITION OF END OF STUDY

The end of trial is the point at which all the data has been entered and queries resolved. This is anticipated to be in March 2021.

9 WITHDRAWALS

Participants are free to withdraw their consent from this study at any time, for any reason. There is no obligation to give the reason for withdrawal. If the participant is withdrawn due to any adverse event (LFT device or study related), the investigator

will arrange for follow-up visits or telephone calls until the adverse event has resolved or stabilised.

10 RISK ANALYSIS

The anticipated risks associated with this study are given in the Risk Assessment Table (see Appendix 1).

The risk-to-benefit assessment is preferable to a ratio and is based on anticipated risks associated with this study and how harm is minimised through the careful study design and importance of the objectives.

This describes all perceived risks to which the participants will be exposed as a result of their participation in this study and how these risks will be minimised. The risks are considered to be low-medium and do not outweigh the benefit, to enable us to continue with the study.

A FMEA analysis based on the features and functionality available for the study is being prepared to assess the potential risks which can arise during the study and will be finalised according to risk management procedure QP08 RMF.

Separate FMEAs are available for camera interpretation of Lateral Flow Testing software and web-based tests results reporting process.

We are utilising existing use cases with established operating procedures which are clinically safe e.g., use of PPE and where the appropriate approvals for use of the LFDs already exist.

The study does not introduce any new additional processes with the exception of the use of a mobile device to capture the image.

We are not looking to introduce new LFD use cases to facilitate this trial.

11 SAFETY REPORTING

As identified within Department of Health and Social Care Quality Management System (QMS) the safety reporting process is controlled under the QP21 Medical Device Reportability Requirements and FSCA Procedure.

11.1 Definitions[5]

Adverse Event (AE):

Any untoward medical occurrence, inappropriate patient management decision, unintended disease or injury, or untoward clinical signs in subjects, users or other persons, with any connection to study related activities, whether or not related to the IVD medical device under investigation.

Adverse Device Effect (ADE):

Adverse event related to the use of an IVD medical device under investigation.

Serious Adverse Event (SAE):

Adverse Event that led to any of the following

a) death,
b) serious deterioration in the health of the subject, users, or other persons as defined by one or more of the following:

    1. a life-threatening illness or injury, or
    2. a permanent impairment of a body structure or a body function including chronic diseases, or
    3. in-patient or prolonged hospitalization, or
    4. medical or surgical intervention to prevent life-threatening illness or injury, or permanent impairment to a body structure or a body function,

5. foetal distress, foetal death, a congenital abnormality, or birth defect including physical or mental impairment

Serious Adverse Device Effects (SADE):

Adverse device effect that has resulted in any of the consequences characteristic of a serious adverse event.

Unanticipated Serious Adverse Device Effect (USADE):

Serious adverse device effect which by its nature, incidence, severity or outcome has not been identified in the current version of the risk analysis report.

Device deficiency:

Inadequacy of a medical device with respect to its identity, quality, durability, reliability, usability, safety or performance, such as malfunction, misuse or use error and inadequate labelling.

Malfunction:

Failure of an IVD medical device under investigation to perform in accordance with its intended use.

11.2 Reporting of AE

All study personnel will be aware of the requirements for reporting adverse events and will be responsible for informing Study Chief Investigator and DHSC's Regulatory Affairs Department if they become aware of a suspected event.

Study Chief Investigator and DHSC's Regulatory Affairs Department will review each incident to check if it meets reportability requirements and, if so, will identify the

authorities to whom the incident needs to be reported and will perform this report in conformity with the process and timescales required by the regulatory authorities, as follows:

- a SAE which indicates an imminent risk of death, serious injury, or serious illness and that requires prompt remedial action for other patients/subjects, users or other persons or a new finding to it - immediately, but not later than 2 calendar days after awareness by sponsor of a new reportable event or of new information in relation with an already reported event.
- any other reportable events or a new finding/update to it: immediately, but not later than 7 calendar days following the date of awareness by the sponsor of the new reportable event or of new information in relation with an already reported event.

If the incident is deemed to be a reportable adverse event (AE) the DHSC Regulatory Affairs Department will initiate an AE reporting process as per QP21.

All AE's occurring during the study observed by the investigator or reported by the participant, whether or not attributed to the device under investigation will be recorded in the AE Form (QP20-F01).

The relationship of AEs to the device will be assessed by a medically qualified investigator or the sponsor/manufacturer and will be followed up until resolution or the event is considered stable.

All ADE that result in a participant's withdrawal from the study or are present at the end of the study, should be followed up until a satisfactory resolution occurs.

11.3 Reporting Procedures for All SAEs/ SADEs/ UADEs

For Non-CE marked device: All SAE/SADE/UADEs need to be reported to the sponsor/legal representative and manufacture **immediately**; regardless of relationship to the device.

For studies of CE marked devices: All incidents need to be reported to the sponsor/legal representative and manufacture **within one working day** of the investigator team becoming aware of them.

Reports of related and unexpected incidents should be submitted to competent authority and ethics committee within reporting timelines summarised below of the Investigator becoming aware of the event, using the applicable report method:

- Serious Public Health Threat – Immediately and not later than 2 calendar days
- Death or Unanticipated serious deterioration in state of health – Immediately after DHSC established a link between the device and the event and not later than 10 elapsed calendar days following awareness of the event.
- Others - Immediately after DHSC established a link between the device and the event and not later than 30 elapsed calendar days following awareness of the event.

12 STATISTICS[9,10,11]

The plan for the statistical analysis of the study is outlined below. There is not a separate Statistical Analysis Plan document in use for the study.

12.1 Sub-study 1: External validation of algorithms

12.1.1 Calibration:

The model was calibrated using samples with varying viral loads and therefore mostly positive. This approach is reflected in the proposal for retraining the algorithm to take into account new testing devices. By construction, the proportion of positive

"individuals" used to train the model exceeds, in proportion, what we expect to see in a real setting. Therefore, underprediction is unlikely to occur. The model was then used in a real-world setting using the Liverpool pilot data. The positivity rate at the time was low and therefore few positives were observed via LFT and PCR, but given the circumstances, the performance achieved was adequate. The specificity is very likely to be accurate while the sensitivity is likely to be inflated due to the small sample number of positives.

12.1.2 Level of statistical significance:

The aim is 95% sensitivity and 95% specificity within 1% margin. If we see indication of convergence of either of these two values to a value lower than 95%, then we might need to reassess the model calibration. This normally only involves refitting using more data (which could be the pilot data) without changing the structure of the model. We then need to sample more for external validation.

12.1.3 External validation and initial sample size:

Since training and calibration of the algorithm have already taken place, we only need to focus on evaluation and external validation at this point. We will treat specificity and sensitivity separately to ensure accuracy is dealt with accordingly. The focus will be primarily on specificity to ensure an adequate number of positives are observed in the validation process. Ideally a minimum of 100 positives should be observed. However, the expected performance of 95% sensitivity and 95% specificity might be achieved earlier. It is worth noting that a minimum of 100 negatives is also necessary but this will almost certainly be achieved in the first ATS batch.

- It is almost certain that we will have a magnitude larger in terms of negative images due to low prevalence, but we would be looking for a minimum of 2500 negative samples to demonstrate specificity, although this is likely to be larger due to the minimum requirement of positive samples above

- We will review and calculate sensitivity / specificity for each image batch, reestimating the values and our projection as sample sizes increase. If we need to use more than one batch of ATS images, this will allow us to rebuild the ROC curves at each stage to determine our ability to meet the benchmarks set by our primary objectives.
- Any contested results will be reviewed by a third party.

12.1.4 Adaptive sampling

Based on the Liverpool data study, it is clear that the algorithm is adequately sensitive and now we need to primarily focus on obtaining an adequate sample of positives and therefore assess its specificity. Instead of using the more traditional and passive approach for specifying a minimum sample size, we will use an adaptive approach to re-evaluate the necessary sample size as data comes through for external validation.

Since prevalence of Covid-19 is currently low (March 2021), if we were to use a classic sample calculation approach, we would likely need around 10-20k samples before we can observe a reasonable number of positives in order to produce a good estimate of the algorithm's sensitivity. As an alternative, we propose the use of sample images from Asymptomatic Testing Sites (ATS) for external validation in batches. While the ATS staff are trained operators and have consistently captured images of test devices, they are not instructed to follow the same set of instructions for capture so some images might not be suitable for processing. It is likely that we will observe at least 100 positives within one batch (around 40k images) and complete the external validation exercise, however some images might not be necessarily fit for purpose (e.g., blurry images, out of frame, rotated) and might require further sampling.

We will review the external validation approach after each batch is processed to determine whether further sampling is needed. Unless anomalies are detected during the study, we will adopt a simple sequential stopping design to ensure bias due to design choices is reduced. We do not plan to adopt an allocation design but acknowledge that the target population of this pilot study carries its own bias. The principles adopted in this study should then be transferred and reviewed when the

model we are assessing goes live. Sampling will continue until sensitivity and specificity estimates converge to or above the target of 95%. Below we also discuss other potential criteria that can lead to a stop or review of the process.

The initial focus will be on estimating the Receiver Operating Characteristic curve (ROC curve) and the area under the ROC (AUCROC). Since specificity is likely to be high, we will need to aim for a large AUCROC to ensure minimum sensitivity of 95%. We will use bootstrap to assess the variation of the AUCROC and derive confidence intervals for the measure but also for specificity and sensitivity. Once the confidence intervals for these measures are small enough or show signs of convergence, we will reassess the study. If the minimum sensitivity and specificity is achieved, a proposal to go live should be made. If, for large sample sizes (50 positives and 100 negatives as a minimum), sensitivity or specificity start to converge to a lower value than expected, model calibration and validation data quality should be reviewed.

12.1.5 Other monitoring tools:

a) Brier's score: we will observe and note the Brier score for this model. If we note substantial deviation from what is considered appropriate (0 to 0.25), we will reassess the model for overfitting and try to identify other deviations that could have led to issues with discrimination. This tool will be used for monitoring purposes only.
b) Cross entropy loss: We will use the log loss approach to monitor the discrepancy between label and predicted probability. We don't expect the model to be perfect (log-loss of 0), but we will use this approach to identify when and if the model is possibly confidently wrong.

12.2 Sub-study 2: Human Usability Study

Sub-study 1 will be used to demonstrate the primary objective in this study and validate the algorithm used for classification. In Sub-study 2, the focus turns to user experience, adoption and adequacy of the service provided.

We will monitor the web service metrics and measure effectiveness and efficiency primarily. We will also use a survey and interviews to assess user satisfaction and identify areas for service improvement.

We will also continue to monitor the algorithm accuracy, investigate contested cases, and monitor the reported values used for classification to ensure changes in performance after the service goes live are identified early. We expect to see a small sample of positives (at least 50 if possible) and negatives (at least 100) to be inspected weekly as well as the contested cases. If a systemic bias is identified, the protocol for retraining should be followed to ensure the algorithm can adequately account for new cases (e.g., new test devices, unforeseen user errors).

Effectiveness: we will observe completion rate and number of errors to assess effectiveness. Completion rate constitutes successful submission of an image by the user and receipt of test interpretation by the algorithm. We will record and report failed attempts at different points of the user journey and identify potential areas for review. We expect completion rate to be at least 95% and to increase with time as users adapt to the service and reviews take place to address any systemic issues.

Efficiency: We will focus on time to completion to measure efficiency and reflect on user feedback (survey and interviews) to assess whether the total journey time is appropriate for the service provided. We will investigate outliers to determine the main barriers affecting completion times.

Satisfaction: We will deploy a survey to capture the overall user experience and followup with interviews to further explore barriers and facilitators in the current service. We will observe the response rate and note the sentiment and level of satisfaction captured but will not use these as a direct measure of satisfaction since we will not be able to adequately control for self-selection and non-response biases. We will, however, use the survey data to potentially identify cohorts and investigate whether they need to be targeted for further feedback to ensure they are being adequately catered for.

12.3 Analysis Population

For Sub-study 1, the target population in the study is representative of the general population and they are sufficient to show accuracy and some level of robustness.

For Sub-study 2, the initial target population will be, in some cases, more skilled than the general population but this bias can be partially corrected by capturing basic demographic information. The regular re-evaluation and monitoring after it goes live is essential to review performance.

12.4 Procedure for Accounting for Missing, Unused, and Spurious Data

Sub-study 1: we will note and report the number of images that don't meet minimum quality requirements and the reason for their removal. Since the ATS staff are not using the same portal as the final user, image quality is likely to be lower than expected. We don't expect any missing data as an output from the algorithm and we will be able to complete the external validation process once enough suitable images are observed.

Sub-study 2: Survey – we will first investigate whether missingness or non-response appears to be random based by first analysing demographic and cohort-based questions. If deemed to be random, we will use multiple imputation where possible. If survey response does not seem to be missing at random and high missingness is present, we will report at first by cohort and use multiple imputation where possible. If possible, we will then identify cohorts for interview before proceeding with imputation or deletion when reporting for the full sample.

We are likely to observe incorrect or spurious values in free text questions. Age and other numerical answers will be imputed by median. We might also observe nonresponse for some demographic-based questions such as gender and ethnicity which we will then impute to "Prefer not to say" or similar. It is worth noting that the survey is being used as a sentiment and user satisfaction tool rather than a tool for estimating specific characteristics in a population so incorporating other tools such as

interviews as part of the analysis process is adequate. Deletion will only be used when imputation proves to not be possible (e.g., most answers missing).

Service-related failures: We don't expect missing data at this stage. There are instead a number of expected errors to be observed and monitored as part of the user journey. The service around the algorithm may return: error 400 - likely to be caused by image corruption and invalid requests, or error 500 – likely to be linked to write issues to database or image store. See Table 1 below. These cases cannot be used for the ongoing assessment of the algorithm's performance per se but need to be monitored as part of service usability. We will also need to assess how different errors are presented to the user and whether they have any impact on their journey.

Status Code	Error Message	Cause
400	Request type is not Json	Invalid request header
400	Request body is empty	
400	Request body does not contain image_id	No image_id key present in json
400	Request body does not contain image_data	No image_data key present in json
400	Image_id is too short	Image_id should be at least 5 characters
400	Image_data is empty	Image_data is empty string
400	Failed to decode image	Cannot parse the image data into matrices
500	Prediction failed	Could not get results from model 1
500	Failed to store image	Error with writing image to blob store
500	Failed to store image	Error with writing to database

Table 1: Service-related failures coding

13 PROCEDURES FOR REPORTING ANY DEVIATION(S) FROM THE ORIGINAL STUDY PROTOCOL

Any changes, divergence or departure from this study protocol shall be immediately reported, in order for appropriate corrective and preventative actions to be taken and/or to ensure that these deviations are included and considered when the pilot study report is produced, as they may have an impact on the analysis of the data.

It is important to inform the NHSD team:

- Jack Dix jack.dix1@nhs.net
- Robert Banathy robert.banathy1@nhs.net

of deviations at the time, they are identified.

Protocol deviations may be:

- Reported directly by the Investigator or member of the study research team - Result from whistle-blowing by another source or indirectly via the NHSD.

14 DATA MANAGEMENT

The plan for the data management of the study is outlined below. There is not a separate Data Management document in use for the study.

Personal data recorded on all documents will be regarded as strictly confidential and will be handled and stored in accordance with the General Data Protection Regulation (GDPR) 20188 and Data Protection Act 20187.

The images of the test results will be sent to NHSD by using the LFD self-report service, or captured by the ATS image capture service.

The “Image ID” will be randomly generated by the Kainos API when the image is passed to the S3 bucket (add brief diagram).

The “Subject ID” is created at the end of the self-report journey (on submission of results) when an entry is written to the “Subject Data Table”.

Data will be stored by the NHSD in a de-identified form (such that no participant can be identified from their data or test result).

During and at the end of the study the de-identified data will be passed to AI vendor for further analysis and as evidence of raw data for regulatory purposes.

15 DIRECT ACCESS TO SOURCE DATA/DOCUMENTS

Direct access will be granted to authorised representatives from the sponsor, host institution and the regulatory authorities to permit trial-related monitoring, audits and inspections.

16 ETHICAL AND REGULATORY CONSIDERATIONS

The study will comply with the General Data Protection Regulation (GDPR)[8] and Data Protection Act 2018[7], which require data to be de-identified as soon as it is practical to do so. The processing of the personal data of participants will be minimised by making use of a unique participant study number only on all study documents and any electronic databases. Data and all appropriate documentation will be stored for 15 years after the completion of the study.

If required, this study protocol will be submitted for consideration, comment, guidance and approval to the concerned research ethics committee before the study begins. This committee must be transparent in its functioning, must be independent of the researcher, the sponsor and any other undue influence and will be duly qualified.

The committee must have the right to monitor ongoing studies. Where applicable, the researcher will provide monitoring information to the committee, especially information about any serious adverse events. No amendment to the protocol may be made

without consideration and approval by the committee. After the end of the study, the researchers will submit a final report to the committee containing a summary of the study's findings and conclusions.

NHSD is committed to ensuring that its research involving human participants is conducted in a way that respects the dignity, rights, and welfare of participants, and minimises risk to participants and researchers. The Investigator will ensure that this study is conducted in accordance with the principles of the Declaration of Helsinki[6].

17 DISSEMINATION OF RESULTS

Results of this trial will be owned by the sponsoring organisation and submitted to MHRA for the purpose of applying for derogation (to ensure continuation of service provision to NHSD).

18 REFERENCES

1. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet. 2020;395(10223):507-513. doi:10.1016/S0140-6736(20)302117
2. BMJ 2020; 371 doi: https://doi.org/10.1136/bmj.m4851
3. https://blogs.bmj.com/bmj/2021/01/12/covid-19-governmentmusturgentlyrethink-lateral-flow-test-roll-out/
4. https://www.ox.ac.uk/sites/files/oxford/media_wysiwyg/UK%20evaluation_PHE%20Porton%20Down%20%20University%20of%20Oxford_final.pdf
5. BS ISO 20916:2019 In vitro diagnostic medical devices – Clinical performance studies using specimens from human subjects – Good study practice
6. Declaration of Helsinki - https://www.wma.net/policies-post/wmadeclarationofhelsinki-ethical-principles-for-medical-research-involving-humansubjects/
7. Data Protection Act 2018 - https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

8. Guide to the General Data Protection Regulation GDPR - https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/711097/guide-to-the-general-dataprotectionregulationgdpr-1-0.pdf
9. Assel, Melissa, Daniel D. Sjoberg, and Andrew J. Vickers. "The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models." Diagnostic and prognostic research 1.1 (2017): 1-7. https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-017-0020-3
10. Hickey, Graeme L., and Eugene H. Blackstone. "External model validation of binary clinical risk prediction models in cardiovascular and thoracic surgery." (2016): 351-355. Ehttps://www.sciencedirect.com/science/article/pii/S0022522316300757
11. Steyerberg, Ewout W. Clinical prediction models. Cham: Springer International Publishing, 2019.

APPENDIX 1: STUDY RISK ASSESSMENT

Item	Process/Function	Potential failure mode	Effect	Severity	Potential cause/failure mechanism	Probability of Occurrence	Current controls	Risk Matrix	Risk Control Action(s) (Mitigations)	Responsibility	Date required	Severity revised	Probability of Occurrence revised	Risk Matrix revised
FS-001	Operational	Care homes and suitability to support pilot.	Non-compliance to pilot or insufficient numbers.	Minor	Care homes are already overburdened with regular testing, bulk registration, and self-testing.	Remote	Agree pilot participation with a select number of care homes. Minimise length of pilot/data capture requirements. Ensure operations are minimally impacted.	Low	N/A	N/A	N/A	N/A	N/A	N/A
FS-002	Operational	Submitted result is negative, AI and independent adjudicator reads the result as positive but does not communicate this back due to study design.	Public health risk as NHS Staff continue to work around vulnerable patients or ASC visitor enters care home without knowing they are positive.	Minor	Current study design when AI interpretation happens later on the day or next day	Occasional	This study is only to ascertain accuracy of AI reading. The data is not going to be used to make a decision on positive or not that is going to be fed back.	Medium	Obtain Clinical Assurance Approval for the study	Nishali Patel	05-Mar-21	N/A	N/A	N/A
FS-003	Study specific / data collection	No established time for taking picture of test – when is it clinically appropriate? Do we accept pictures being uploaded 12 hours later?	Could this conflate data being collected to detect viral loads depending on faintness of line? How does the time after which the picture is taken impact faintness of line/deterioration of result?	Minor	Gaps in study design, webpage development and training	Probable	Photo needs to be taken at the point at which the result is interpreted/decided. This requirement is captured in the study protocol.	Medium	Current control measures is sufficient	N/A	N/A	N/A	N/A	N/A
FS-004	Study specific / data collection	Establishing an acceptable percentage to measure sensitivity and specificity –(primary outcome)	Implications on supporting evidence to show efficacy of AI tool	Serious	Lack of clinical and research expertise during protocol development	Remote	Clinical and research expertise has set a clinically acceptable benchmark for primary outcome measurement.	Medium	Current control measures is sufficient	N/A	N/A	N/A	N/A	N/A
FS-005	Study specific / data collection	Method of measuring AI effectiveness is unclear in protocol – what is the ultimate determinant of the 'correct' result interpretation?	Affects robustness of data. Study aims to understand if AI is better at results interpretation vs humans – how is it doing this?	Minor	Gaps in study design, data analysis process	Remote	We will use the third party independent readers to assess any cases where there is a discrepancy between the human and machine read. They will also be supplied with random ~10% of images where there is agreement.	Low	N/A	N/A	N/A	N/A	N/A	N/A
FS-006	Study specific / data collection	What data has been used to train the AI? Is this sufficient and of high enough quality to train AI to be effective? E.g. image clarity.	Affects robustness of study data.	Minor	Gaps in design and development process	Remote	User Requirements Specification and Software Requirements Specifications will be created and reviewed and approved by NHSD and vendor. Verification and Validation plan and protocols will be created and reviewed and approved by NHSD and vendor. Before the study commenced analytical (bench testing) report will be reviewed and approved by NHSD.	Low	N/A	N/A	N/A	N/A	N/A	N/A
FS-007	Digital	Inappropriate users land on pilot journey	Users may see this as a barrier for self-reporting and not proceed.	Negligible	Data collection is not for the intended use cases and outside the remit of study pilot inclusion criteria.	Remote	Mitigate inappropriate users by design. Only LFD test subjects who have identified themselves as working with NHS primary care or are linked to a subset of adult social care homes will be invited to take images for the study.	Low	N/A	N/A	N/A	N/A	N/A	N/A
FS-008	Digital	No consent captured or privacy information updated for study participation	IG risks and study protocol breach	Serious	Gaps in study design	Remote	Ensure DPIA complete before pilot and IG input into digital design and privacy/data handling information	Medium	Current control measures is sufficient	N/A	N/A	N/A	N/A	N/A
FS-009	Digital	Photo capture functionality affecting self-reporting portal	Risk of self-report tool being affected	Negligible	Delays or downtime due to insertion of photo capture functionality	Occasional	Validation testing will be done before go-live and it will be monitoring in live with contingency plan to toggle functionality on/off if required	Medium	Current control measures is sufficient	N/A	N/A	N/A	N/A	N/A
FS-010	Digital	Image storage	No defined time period or location for image storage for pilot data capture.	Minor	Gaps in study design	Occasional		Medium	Needs a defined time period for image storage and documented somewhere secure to support pilot data.	Robert Banathy	05-Mar-21	Minor		
FS-011	Study specific / data collection	Study does not reflect 'end state' use case	Study cannot draw data on how AI interpretated results will be fed back to user – i.e. is this before user submits their interpretation? Is it after they submit their interpretation? How does this affect usability/uptake and behaviours.	Serious	Not the final version of the product will be used during the study	Frequent	User research will be conducted as part of the study to understand user behaviour	High	Additional Post Market Surveillance human usability stusy will be conducted	Rachel Abbot	05-May-21	Serious		
FS-012	Operational	Study doesn't capture time taken for AI to read result	This has operational and user impact in real use cases – how long does it take for a result to be communicated back to the user? Is it instantaneous? Can this be captured in the study?	Negligible	Gaps in design and development process	Remote	User Requirements Specification and Software Requirements Specifications will be created and reviewed and approved by NHSD and vendor. Verification and Validation plan and protocols will be created and reviewed and approved by NHSD and vendor. Before the study commenced analytical (bench testing) report will be reviewed and approved by NHSD.	Low	N/A	N/A	N/A	N/A	N/A	N/A
FS-013	Operational	AI is only trained to support use of Innova LFD tests. No way of distinguishing different kits	Newer LFD kits are coming out and need pictures to support AI training. Need to be able to distinguish different kit brands (using barcode) to apply appropriate AI reading rules.	Serious	Gaps in design and development process	Probable	The system will not allow a kit to be used in Self-report more than once.	High	Additional control needed to stop people submitting a picture of a kit that is not the one whose barcode they have entered	Mark Branigan	05-May-21	Serious		
FS-014	Operational	Not sufficient storage period of pictures collected during the study	Storing pictures is helpful for: incident/conflict management, continued provision of images for ongoing AI training, audit. How long and where will these be stored. Is there a risk to be considered around storage space?	Minor	Gaps in QMS related to data storage requirements	Remote	It's a regulatory requirement to store data for 10 years. Study protocol specifies 10 years of data storage requirements.	Low	N/A	N/A	N/A	N/A	N/A	N/A
FS-015	Operational	No defined "post-go live" evaluation plan	How will AI reader in live settings be monitored/measured for effectiveness? What are the post-go live monitoring requirements for MHRA and NHS T&T?	Serious	Gap in project plan	Remote	"post-go live" evaluation process documented in PMS plan	Medium	Current control measures is sufficient	N/A	N/A	N/A	N/A	N/A
FS-016	Operational	Pilot users are handling LFDs and mobile phones, and then care home manager handles the same phone	Potential cross contamination	Minor	Gaps in study design, training	Remote	Care home manager will be asked to use appropriate PPE when handling study participant's LFD and whoever takes the photo will be using their own device	Low	N/A	N/A	N/A	N/A	N/A	N/A
FS-017	Operational	Use of non-Inova device during the study	Validity of the study data can be jeopardised	Minor	Gaps in study design	Remote	Innova device barcode validation will be done and non Innova LFT devices will be removed from the data set analysed.	Low	N/A	N/A	N/A	N/A	N/A	N/A
FS-018	Operational	Failure to follow the instructions for test procedure and interpretation of test results	This may adversely affect test performance and/or produce invalid results.	Minor	IFU have not been validated for lay person use	Remote	The likelihood if this happening is reduced by initial observed performance of those staff who require it, ongoing support as required, and access to an instruction booklet and video.	Medium	Current control measures is sufficient	N/A	N/A	N/A	N/A	N/A
FS-019	Operational	Images containing user's personal information transferred to vendors	Violation of GDPR	Serious	Gaps in study design	Remote	DPIA undertaken and DSA in place.	Medium	Current control measures is sufficient	N/A	N/A	N/A	N/A	N/A

Refer to FMEA-001 AI Performance Evaluation Study FMEA v1.0 for additional information

APPENDIX 2: HUMAN FACTORS AND USABILITY STUDY DESIGN PROTOCOL

Background and Rationale

This study concerns the use of an AI reader function embedded within a digital reporting service for interpreting Covid-19 lateral flow tests at point of use.

The end assumptions and hypotheses for the final product are:

Better user experience - The user will not have to interpret the result and enter their result, so using the AI reader will reduce anxiety for the user to correctly interpret the result.

Greater accuracy - The AI interpreted result will be more accurate than the selfreported result as it will reduce human error and fraud. This will help to identify more asymptomatic test subjects and reduce the transmission rate of the virus.

This pilot will improve our understanding of the use of the online "photograph taking process" in practice and to determine whether the product is easy to use and safe in the intended context of use.

Aims

- To understand who the users, their experience, the tasks they have to perform and the contexts within which they work
- To assess certain human factors (lay and professional user) associated with the use of the decision support software in order to understand the usability and accessibility of the service
- Assess completion of tasks
- Gather subjective data on safety and ease of use
- Identify any use errors
- Evaluate the user experience and user expectations

Tasks

1. Take a photograph of the lateral flow test strip

Subtasks of task 1.0
1.1a User opts out of providing a photograph

1.1b User opts in to provide a photograph image of the test strip
1.2 User reads guidance on how to correctly take a photograph image of the test strip
1.3 User takes the photograph of the test strip
1.4 User checks the photograph is acceptable quality
1.5a User submits the photograph (1st attempt)
1.5b User submits the photograph (subsequent attempts)
1.6 User checks all answers and reports result

Participants

Participants for the trial will be recruited from two use case groups:

- Adult Social Care
- Primary Care (Community pharmacy, NHS Optometry, NHS Dental and General Practice)

Methodology

A survey will be distributed via comms teams to the use case groups to be completed during the trial period.

This survey will collect data on the participant's themselves and their experience of using the product during the trial, as well as operate as a recruitment tool for qualitative interviews.

Interviews to be conducted with 15 participants from each use case group (Adult Social Care and Primary Care) who consent to be contacted for a follow-up interview. The interview will cover their survey responses in more depth, and to attempt to establish the root cause of any user errors or difficulties encountered.

Adobe analytics will also be used to identify user interactions across the digital journey, including any drop off points.

Survey Questions

User Profile

- Sector of work
- Job Title
- Context of Use
- Gender
- Age
- Ethnicity
- Internet Use
- Accessibility software used
- Disabilities or long-term health conditions

Usability Questions

- How many Covid-19 Lateral Flow tests have you taken so far during the trial? *(If you are a nominated person responsible for reporting all staff lateral flow test results, please tell us how many test results you have reported in total)*
- Of these, how many times did you take and upload a photo of the lateral flow test strip?
- If you did not take and upload a photo every time, please tell us why
- Did you ever have to try more than once to take and upload a photo?
- How many times did you have to try more than once?
- If you had to try more than once, at any time, please tell us more about what happened
- How easy did you find it to take and upload a photo of the lateral flow test strip?

Interview Script

Welcome and Intro

- Purpose
- Consent and confidentiality
- Recording
- Observers

Recap

Refer back to their survey answers for prompts:

- Can you tell me a little bit about how you found taking and uploading a photo of your test?
- What device did you use?
- Where were you when you were taking the test/taking the photo?
- How did you approach it?
- Did you experience any difficulties doing this?
- Did you need to get any help or assistance from anyone else?

Improvements

- What benefits (if any) do you see in taking a photo of the lateral flow strip and uploading it?
- If you were to do this process again in the future, would you take a photo of the test strip on those occasions?
- If not, why not?
- How might we improve the process of taking and uploading a photo of the lateral flow test device?

Usability Metrics

The ISO 9241-11 standard defines usability as “the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use”.

The ISO/IEC 9126-4 Metrics recommends that usability metrics should include:

- Effectiveness: The accuracy and completeness with which users achieve specified goals
- Efficiency: The resources expended in relation to the accuracy and completeness with which users achieve goals.
- Satisfaction: The comfort and acceptability of use.

Usability Metrics for Effectiveness

a) Completion Rate

Effectiveness can be calculated by measuring the completion rate – uploading image of the completed LFD test. Referred to as the fundamental usability metric, the completion rate is calculated by assigning a binary value of '1' if the test participant manages to complete a task and '0' if he/she does not.

Effectiveness can thus be represented as a percentage by using this simple equation:

$$Effectiveness = \frac{\text{Number of tasks completed successfully}}{\text{Total number of tasks undertaken}} \times 100\%$$

Our aim for a completion rate of at least 95%

b) Number of Errors

This measurement involves counting the number of errors the study participant makes when attempting to take and upload the image.

Usability Metrics for Efficiency

Efficiency is measured in terms of submitted image quality. The Efficiency rate is calculated by assigning a binary value of '1' if the uploaded image was successfully interpreted by AI, and value of '0' will be assigned if AI failed to interprets the image.

Efficiency can be represented as a percentage by using this simple equation:

$$Efficiency = \frac{\text{Number of images successfully analysed by AI}}{\text{Total Number of images submitted for AI interpretation}} \times 100\%$$

Our aim for an efficiency rate of at least 95%

Usability Metrics for Satisfaction

a) Task Level Satisfaction

After users attempt a task (irrespective of whether they manage to achieve its goal or not), they will be given a questionnaire so as to measure how difficult that task was.

Users will be asked to use the Single Ease Questions score for response:

Extremely easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Extremely difficult
☐	☐ ☐		☐ ☐	

b) Test Level satisfaction

Test Level Satisfaction is measured by giving a formalized questionnaire to each test participant at the end of the test session. This serves to measure their impression of the overall ease of use of the system being tested.