Facial Recognition Technology in Law Enforcement Equitability Study

Test Strategy

Tony Mansfield
National Physical Laboratory

Draft 3.1
26/04/2022

Contents

1 Background

1.1 MPS and SWP deploy Facial Recognition Technology (FRT) to assist officers in identifying persons of interest. There are three types of Operational Use Case for Facial Recognition Technology:

a) **Live Facial Recognition (LFR)** compares a live camera feed of faces against a predetermined Watchlist to find a possible match that generates an Alert.

b) **Retrospective Facial Recognition (RFR)** is a post-event use of facial recognition technology, which compares still images of faces of unknown subjects against a reference image database in order to identify them.

c) **Operator Initiated Facial Recognition (OIFR)** is a near-real-time use of facial recognition technology, where an officer takes photograph of a subject via a mobile device and submits it for immediate search against a reference image database.

1.2 An accurate Live Facial Recognition system requires that:

a) whenever an individual on the Watchlist passes the system the LFR system should generate an Alert, and

b) whenever an individual who is not on the Watchlist passes the system, the LFR system should not generate an Alert and should automatically delete all biometric data relating to that individual.

1.3 An accurate RFR / OIFR system requires that, if the person in the still image or being photographed also has a face image in the reference dataset, this mated image and linked identifiers should be among the list of candidate images returned, preferably at the front of the ranked list of candidates.

1.4 In addition to having good accuracy, it is also important that the FR systems perform well in terms of minimising algorithmic and system biases, avoiding any performance differentials between different demographic groups that would disadvantage one demographic group in comparison to another.

1.5 The requirement of understanding accuracy and whether this varies by demographic is critical to the successful use of FRT by law enforcement. It allows decisions to be made as to where and how the technology is to be used and to consider its likely effectiveness. Ultimately, this enables law enforcement to prevent and detect crime more effectively and to minimise unnecessary data processing and collateral intrusion wherever possible.

1.6 The MPS and SWP have undertaken steps to understand their algorithms in line with their Public Sector Equality Duties (PSED) under the Equality Act 2010. This includes having regard to the FRT evaluation programme undertaken by the National Institute of Science and Technology (NIST) as well as their own trials of FRT to date:

a) NIST runs an ongoing evaluation programme FRVT (Face Recognition Vendor Test) assessing the accuracy of facial recognition algorithms from many suppliers. The MPS and SWP Facial Recognition Systems use algorithms supplied by NEC, one of the top performing vendors in FRVT, and NIST's ongoing evaluations have enabled MPS and SWP to monitor the improvements in baseline performance of the algorithms over time.

b) The MPS have reviewed the effectiveness of the M-20 NEC algorithm (the immediate predecessor to the M-30 algorithm) from a demographic perspective.

1.7 This diligence allows the MPS and SWP to take assurance in the performance of their algorithms in line with the 'PSED' responsibilities. Nevertheless, there is an ongoing duty to

keep taking reasonable steps to better understand algorithm performance, particularly in the operational context, in order to address requirements to demonstrate fair and equitable use of the technology. MPS documentation recognises the need to undertake further operationally relevant testing to assure and augment the available FRVT results. Such diligence can only take place in the operational environment and is needed now because:

a) The FRVT test datasets are not sufficiently representative of the face image data of the MPS and SWP operational use cases. In particular, face recognition from video sequences – the LFR Operational Use Case – is out of scope of the recent FRVT evaluations.
b) Though based on the same underlying facial recognition technology, the NEC algorithms submitted to NIST are not identical to the versions used by MPS and SWP.
c) The NIST Face In Video Evaluation (FIVE) published in 2017 reports on algorithms submitted for testing in 2015, and may therefore be regarded as somewhat historic.
d) The NIST FIVE report does, however, acknowledge that the challenge of finding and segmenting faces from the video prior to face comparison adds further causes for failed recognition.

1.8 As yet there are no suitable datasets representative of the policing use cases that could be used for bench testing which would enable the objectives without the need to capture data from the operational environment. This means that data capture and processing for the purposes of testing in parallel to an operational deployment is unavoidable and necessary to fulfil PSED responsibilities towards understanding algorithmic performance

1.9 The evaluation described in this Test Strategy document will bridge the gap between NIST FRVT results and MPS/SWP diligence and testing to date and the ongoing MPS/SWP Operational Use Cases of Facial Recognition Technology. Moreover, the evaluation will also produce a test corpus that can be used to test other Policing face recognition algorithms without the need to run further data collection from volunteers and the general public, minimising the data collected for further testing purposes.

2 Test objectives

Table 1 – The objectives of the evaluation

Objective	Why objective is necessary
A FR algorithms currently used/intended for use in policing	
To evaluate the performance of facial recognition technologies in an operational setting in terms of (i) accuracy and (ii) equitability (bias) related to subject demographics	Section 1 of this report explains why it is critical to understand accuracy and equitability. The evaluation results will enable law enforcement to: • based on understanding how the algorithm performs decide with greater assurance whether and how best to configure FR technology for effective deployment on operational use cases; • ensure unnecessary data processing and minimizing collateral intrusion including through use of other technologies where less likely to be effective; • continue to discharge of PSED obligations by continuing to take all reasonable steps to understand the algorithms.
For each Operational Use Case:	
(a) What is the accuracy of the facial recognition algorithms? • LFR accuracy: True Recognition Rate and False Alert Rate as a function of the alert threshold • RFR/OIFR accuracy: True Recognition Rate as a function of the number of top matches returned	Assures policing on (i) the accuracy of the FRT algorithms, and (ii) provides information on selection of threshold and other parameters to tune performance to the operational requirement.
(b) What is the variation in accuracy between the demographic groups?	Identifies the extent of "demographic bias"
(c) Are the variations in accuracy large enough to be "statistically significant"	
(d) Are demographic performance variations similar over the 3 Operational Use Cases?	Alerts policing to potential differences between LFR/OIFR/RFR which impact on operational deployment of use cases.
(e) Are variations in accuracy affected by environmental factors (e.g., weather, illumination level, crowd density).	Provides information of factors that could increase/reduce accuracy and bias
(f) How is the variation in accuracy affected by system factors (e.g., algorithmic thresholds, composition of watchlist/reference database)?	Informs on effects of choice of threshold / watchlist composition on accuracy and bias.
B Building capability to evaluate future Policing FR algorithms using representative data	
To collect a ground-truth dataset the UK Law Enforcement Community can use for future testing of other FR algorithms.	Reduces/avoids the need to collect further personal data to run similar evaluations going forward.

Table 2 - Key points and terminology relating to this evaluation

Key Point /	Terminology	Narrative
Algorithms to be tested	NEC NeoFace Watch NEC NeoFace Reveal	Algorithms are those used by MPS and SWP in their Operational Use Cases
Operational Use Cases	LFR Live Facial Recognition RFR Retrospective Facial Recognition OIFR Officer Initiated Facial Recognition	See MPS Facial Recognition Terminology Overview[1]
LFR Measures of Accuracy	TRR True Recognition Rate FAR False Alert Rate FTAR Failure-To-Acquire Rate (measures cases where the LFR system fails to acquire for biometric comparison the face image of a person in its zone of recognition)	The Biometric test standards use alternative terms: TRR = TPIR True Positive Identification Rate FAR = FPIR False Positive Identification TRR and FAR are dependent on a configurable alert threshold and should be measured and reported over a range of plausible thresholds.
RFR / OIFR Measures of Accuracy	TRR True Recognition Rate	TRR is dependent on the number of top matches returned. (Up to 200 in case of RFR) and should be reported over the range of values from 1 to 200.
Equitability of performance / Demographic factors:	Ethnicity Gender Age	Equitability of performance considers the extent and significance of differences in accuracy with respect to defined demographic characteristics (and intersections)
Data subjects of the evaluation	**Cohort Subjects** – recruited to provide corpus of face images and video to be recognised **Crowd Subjects** – member of the public passing through the zone of recognition of the LFR system **Filler Subjects** – data subjects of the Filler Dataset drawn from MPS holdings of custody images, and used as a reference dataset in the evaluation-	

[1] https://www.met.police.uk/SysSiteAssets/media/downloads/force-content/met/advice/lfr/other-lfr-documents/terminology-overview.pdf

Key Point /	Terminology	Narrative
Data controller/ Data processor	Operational LFR deployments processing data for a research purpose pursuant to this trial plan. **Data controller** MPS (London deployments) SWP (Cardiff deployments) **Data processor**: NPL (for collection of facial images of cohort subjects) Post deployment evaluation and curation of dataset for future Policing evaluations **Data controller**: MPS **Data processor**: NPL	• The MPS and SWP remain data controllers for their respective operational deployments (in line with their respective LFR policy documents). • SWP will controller-to-controller transfer to the MPS, for the purposes of testing only, LFR footage of from their deployment(s). • The data processing arrangements will be further outlined in the respective MPS and SWP data protection documentation

3 Summary of evaluation methodology

3.1 The evaluation will be conducted in accordance with international standards for testing and reporting the performance of biometric recognition systems: ISO/IEC 19795-1 [2] and ISO/IEC 19795-2 [3]. In the terminology of the standards for biometric performance testing and reporting, the performance will be evaluated as a "technology evaluation".

3.2 The evaluation will draw on a corpus of face images, and video comprising:

a) Face Images and metadata of Cohort Subjects recruited for the evaluation
b) LFR video recorded at operational deployments featuring both Cohort subjects and Crowd subjects), and
c) Face images and metadata of a Filler Reference Dataset drawn from MPS holdings.

3.3 Given the importance of testing in a realistic operational environment, the images/video used, and the quality and nature of the media used must closely emulate those that would be used by policing for their facial recognition use cases. This is achieved through recording LFR video at an operational deployment and using MPS custody images as Filler data for the reference dataset.

3.4 The stages of the evaluation outlined in Table 3 below.

[2] ISO/IEC 19795-1 Biometric testing and reporting – Part 1: Principles and framework
[3] ISO/IEC 19795-2 Biometric testing and reporting – Part 2: Testing methodologies for technology and scenario evaluation

Table 3 – Overview of evaluation plan

Evaluation Stage	Process	Detail
Stage 1.	**Collect test corpus of face images & video, and metadata**	
Cohort image data	**Cohort Reference Dataset** Reference face images representative of custody images (taken using MPS kit for custody images). (See 5.1.3)	Collected by NPL from Cohort subjects present at LFR deployments.
	Cohort Probe Dataset Probe face images representative of OIFR and RFR use cases	
	Cohort metadata See 7.1 for list of metadata	Collected by NPL at LFR deployments.
Filler reference data	**Filler Reference Dataset** This dataset, provided by MPS comprises MPS/SWP custody images together with ethnicity, gender and age metadata. Provided by MPS. (See 5.4)	Subset of MPS/SWP custody image dataset. These data subjects are referred to as "Filler subjects" in this plan.
LFR video data	**LFR videos** Probe video featuring both Cohort and Crowd subjects collected by MPS/SWP NeoFace Watch system at LFR deployments	Recorded by MPS/SWP at operational deployment
	LFR video metadata Metadata regarding Cohort appearance in the video, environmental factors and system factors. (See 7.4)	Collected by NPL at LFR deployments
Crowd metadata	The LFR videos will passively collect images of Crowd subjects. Metadata on the video footage of subjects forming part of the crowd will be collected on an aggregate basis, (and NOT per Crowd subject). The video footage will be sampled to survey the gender, age and ethnicity balance, and the total number of Crowd subjects appearing in the video. The aggregate numbers are necessary for analysis of the False Alert Rate.	Crowd metadata established at Stage 2c of the evaluation.
Stage 2	**Collation of test data** Organisation of the collected data to be ready for offline batch FRT processing	
	a. Transfer of data to NPL i) Filler Reference Dataset, from MPS ii) LFR video from MPS/SWP deployment iii) Cohort image data on camera devices and PC used at collection site b. Labelling of Cohort Data images for curated dataset. c. Establishing and tabulating metadata for curated dataset. d. Generation of lists of probes, references for offline running of LFR, RFR, OIFR	a.i) & a.ii) on encrypted hard drive, by trusted hand a.iii) by trusted hand b) &c) requires test staff inspection of images and footage and Cohort metadata. d) As per Section 8
Stage 3	**Offline running of LFR, RFR, OIFR Facial recognition algorithms**	

Evaluation Stage	Process	Detail
	The LFR, OIFR, RFR algorithms are run offline replicating operational deployment though without requiring operator involvement or adjudication of the process. The NeoFace Watch and NeoFace Reveal algorithms are run on MPS hardware at NPL. (See Section 8)	The separation of data capture from algorithmic processing provides the ability to re-run the video data to assess LFR performance at different parameter settings or with different watchlist composition
Stage 4	**Organisation of outputs of offline running for performance analysis** An intermediate stage necessary before detailed analysis	
	• Collation of outputs of offline running. • Resolution of anomalies arising in outputs of offline running.	Resolution of anomalies may require review of images. (See e.g., para 8.5)
Stage 5	**Performance analysis** Detailed analysis of the results of offline testing on the collected data	
	The outputs from offline running of the face recognition algorithms provide data (comparison scores, rank of mated reference, alert/non-alert outcome) for calculation of accuracy and analysis (i) accuracy measured over all demographics, (ii) a breakdown comparing accuracy for different demographic categories, and (iii) assessment of the statistical significance of any demographic variations in performance	
Stage 6	**Final report & Closure**	
	• Reporting of findings to MPS/SWP • Publishable version of final report • Transfer to MPS the collated dataset for future Policing testing of FRT • Deletion of any residual evaluation images/videos/subject data from NPL hardware	

4 Demographic categories

4.1 Ethnicity

4.1.1 Ethnicity will be classified in accordance with the MPA self-defined ethnicity codes[4]. For sourcing of Cohort subjects, for selection of face images for the filler reference dataset, and for analyses of performance the grouping of ethnicities will be:

- Asian or Asian British (A1 Indian, A2 Pakistani, A3 Bangladeshi, A9 Any other Asian background)

[4] http://policeauthority.org/metropolitan/publications/briefings/2007/0703/index.html

- Black or Black British (B1 Caribbean, B2 African, B9 Any other Black Background)
- White (W1 British, W2 Irish, W9 Any other White background)

4.1.2 These ethnic groups have been selected because (i) they are the largest in the national population and (ii) they are similarly the largest in the MPS custody records. The results of differential in demographic performance therefore have the greatest relevance to the MPS. Limiting the ethnicities considered to these three categories is therefore a proportionate, technically-considered and risk-based approach in line with the 'reasonable steps' requirement to understand the algorithms.

4.1.3 The groups identified roughly correspond the perceived ethnic categories IC4, IC3 and IC1 for cases where a self-declared ethnicity is not possible (e.g. categorising ethnicity of people in the crowd).

4.2 Gender

4.2.1 Gender categories for assessment are female and male

4.3 Age

4.3.1 Policing have the need to use FR to progress investigations and locate people across a wide spectrum of age ranges. This evaluation plan responds to the policing need to use FR across the age range and reflects a reasonable step to further test algorithm performance in an operational environment. The test plan will therefore assess the operational performance of algorithms and whether the performance varies by age. It will focus on age from 14 through 70+. Focus will be placed on younger ages. This reflects the age category of greatest policing need, and FR risk of disproportionality is most likely to be relevant. It is representative of the age profile of custody images.

4.3.2 The age profile of Filler Reference Dataset (to be provided from MPS/SWP), will be as per the random selection from the face images of the requested ethnicities/gender

4.3.3 For analysis of performance differences by age, subjects will be grouped by age, the boundaries being determined based on the age profile of the Filler reference dataset, which should be similar to the age profile of police custody images.

4.4 Other ethnicity and gender related demographic factors

4.4.1 Additional demographic factors recorded in data collection for potential analyses include:
(i) Skin tone (classified using the 6-point Fitzpatrick scale for human skin colour), and
(ii) Subject height.
These factors are correlated with ethnicity and with gender and need to be recorded to clarify possible causes for any variation in face detection performance.

4.5 Demographic categories outside the scope of this evaluation

4.5.1 In this evaluation we are not testing effects of disability, disfigurement, gender transition on facial recognition performance. Such individuals are not excluded from participating in the trials as Cohort subjects, though data regarding disability, or disfigurement is not collected (and gender classification is self-defined male or female).

4.5.2 Disabilities do not readily classify into a small number of similar categories or levels (unlike ethnicity, gender and age), and the statistical considerations to assess performance differentials would require a far larger cohort than can be addressed in this evaluation. Any evaluation that could viably be undertaken by law enforcement in a way that would generate representative and applicable results regarding the impact of disabilities on facial

recognition performance would be challenging in the volume of data processing and number of subjects that would need to be involved. When set against the typical law enforcement use case, the limited further assurance does not justify the need for such research as a priority.

5 Data subjects, images and video footage

5.1 Cohort: Size and composition

5.1.1 Cohort subjects will be drawn from two sources (i) an actor's agency and (ii) an under 18 cohort of volunteers from the Police Cadets.

5.1.2 400 Cohort Subjects will be needed to achieve the Trial Objectives. This can be broken down as in Table 4 below[5]. The rationale for total Cohort size and breakdown is provided in Section 6.

Table 4 – Composition of Cohort

Cohort Subjects – Extras Agency:			
Ages: Record ages, adjust specification for cohort recruitment as study progresses to maintain similar age profile for age and ethnicity categories, with regard that of filler reference dataset			
Self-declared ethnicity 16+1	A1+A2+A3+A9	B1+B2+B9	W1+W2+W9
Perceived ethnicity 6+1	*Similar to IC4*	*Similar to IC3*	*Similar to IC1*
Female	60	60	60
Male	60	60	60

Cohort Subjects – 13-17 year-old Police Cadets			
Ages 13-17			
Self-declared ethnicity 16+1	A1+A2+A3+A9	B1+B2+B9	W1+W2+W9
Perceived ethnicity 6+1	*Similar to IC4*	*Similar to IC3*	*Similar to IC1*
Female	20	20	20
Male	20	20	20

5.1.3 Probe and Reference facial images collected from the Cohort subjects for the purposes of the evaluation objectives will be representative of the MPS and SWP operational use cases. Reference frontal face images will be taken using MPS supplied image capture equipment. Probe images will replicate those typical used for RFR or OIFR purposes.

5.1.4 Facial images and video will be collected from Cohort subjects in two ways (i) photographing of a probe face images and (ii) video footage of the Cohort subject passing the zone of recognition of an LFR system.

5.2 Informing the Cohort

5.2.1 Agency Cohort subjects will be informed of the data to be collected and how this data will be used. These subjects will be paid, and as a condition of payment they will be given full details as their role and the data processed and invited to provide their consent to the activity and data processing. The data will be processed on the basis of strict necessity. The subjects will be informed that having agreed to participate in consideration for a payment,

[5] The numbers in Table 4 total 480, allowing some drop out/no shows.

their personal data will be processed for the purposes of testing. They will remain able to exercise their individual data rights in relation to their personal data and their right to do so will be flagged to them as part of the consent process, and a point of contact with MPS (as Data Controller) provided for this purpose.

5.2.2 Policy Cadets: Parents and cadets will be provided with age-appropriate information to allow parents and cadets to understand their role, how their data is used, and their rights. It will be an opportunity to gain a wider insight into how new technology cold help shape the policing role in the future when they may hope to become an officer. Whilst consent will be sought and this aids proportionality, data will be processed on the basis of strict necessity. Their involvement reflects the use cases of policing: a significant number of investigations where facial recognition would be used relate to those under 18. The recognition accuracy performance characteristics of this category can differ from the adult working-age population[6] and this testing is necessary to understand this in an operational realistic context. The output will allow Policing to know where Face Recognition may or may not provide an effective law enforcement tactic.

5.3 Protecting Cohort identity and enabling individual data rights to be exercised

5.3.1 The images and ground truth metadata regarding the demographics of the test subject will be pseudonymised via the use of test subject IDs. This segregation embeds a designed in privacy approach in terms of what NPL will access as the reconciliation process from test subject ID to Cohort subject will not be available to those undertaking facial recognition but will be held on a need-to-know basis via the MPS.

5.3.2 The MPS Data Office will hold the reconciliation table to allow test subject names and IDs to be reconciled for the purposes of exercising data rights and how to do this will be flagged to the Cohort.

5.4 Filler reference dataset: Need, size and composition

5.4.1 Further to the requirement that evaluation use Cohort images and video emulating that of operational deployments, it is also important in testing that the reference databases are representative of the scale and image types of operational deployments. It is relatively easy for a face recognition algorithm to match against the true mate when the reference dataset is small. With larger datasets, especially those with variation in image capture camera specification or capture processes, there are more possibilities for false matches and error rates increase. Moreover, a test which relies on a reference dataset somewhat smaller than that of the operational use case may miss effects that would be detectable at a representative scale. For these reasons, to address the evaluation objectives, it is necessary to supplement the reference dataset collected from a cohort with a filler dataset representative of the operational use cases.

5.4.2 The reference dataset used for evaluation of equitability should be balanced between demographics. If the reference dataset is imbalanced, with one demographic category being over-represented in comparison to others, then it is to be expected that false alerts are more likely to occur with the over-represented category, and this would hide any algorithmic biases in performance.

[6] See e.g. NIST FRVT reports

5.4.3 The filler reference dataset is to be drawn from MPS holdings of frontal custody images and so is truly representative of the operational use case. Our test design requires an evenly balanced demographic in the data set with equal numbers in each gender/ethnicity category. This necessarily requires the size of the filler set to approximately 180,000 based on the smallest of gender/ethnicity class in the MPS holdings.

5.4.4 From 2016 MPS introduced revised procedures for collection of custody images to improve to improve image quality. Automated facial recognition technology is generally more accurate with higher quality face images. Testing using only post 2016 custody reference images is needed to inform of on accuracy and equitability using current custody image practice. However, in operational scenarios for policing facial recognition both pre- and post-2016 custody images may be used. Testing to inform policing on how differences in image quality between the Pre-2016 and Post 2016 eras affect recognition accuracy and equitability requires additional filler dataset comprising pre-2016 custody images (that can be combined with a portion of the Post 2016 filler images to emulate the mix of pre- and post-2016 images of the operational use case).

5.4.5 Table 5 shows the requested composition for the Filler Reference Dataset.

Table 5 — Composition of Filler (Non-Mated) Reference Datasets

Filler (non-mated) reference dataset: Post-2016			
Random selection from the MPS custody images of Filler Subjects, with the ethnicity and gender metadata, to meet the criteria below. Post 2016 custody images (with exception of Female A1/A2/A3/A9 where pre-2016 images must be included to meet requirement)			
Self-declared ethnicity 16+1	A1+A2+A3+A9	B1+B2+B9	W1+W2+W9
Perceived ethnicity 6+1	*Similar to IC4*	*Similar to IC3*	*Similar to IC1*
Female	25,000 – 30,000	25,000 – 30,000	25,000 – 30,000
Male	25,000 – 30,000	25,000 – 30,000	25,000 – 30,000

Filler (non-mated) reference dataset: Pre 2016 & Post 2016				
Random selection from the Post-2016 Filler subset, and Random selection from Pre-2016 MPS custody image holdings with ethnicity, and gender and metadata, to meet the criteria below.				
Self-declared ethnicity 16+1		A1+A2+A3+A9	B1+B2+B9	W1+W2+W9
Perceived ethnicity 6+1		*Similar to IC4*	*Similar to IC3*	*Similar to IC1*
Female	Post-2016	11,000 – 13,000	11,000 – 13,000	11,000 – 13,000
	Pre-2016	14,000 – 17,000	14,000 – 17,000	14,000 – 17,000
Male	Post-2016	11,000 – 13,000	11,000 – 13,000	11,000 – 13,000
	Pre-2016	14,000 – 17,000	14,000 – 17,000	14,000 – 17,000

5.4.6 This data is requested as images with pseudonymised ID, gender, self-defined ethnicity code, and age at time image taken, and date of image capture (pre-2016 or post 2016) metadata.

5.4.7 Storage will be on encrypted hard drive and stored in approved secure cabinet when not in use.

5.5 Other subjects in crowd

5.5.1 There are two points to address in the context of members of the public passing through the zone of recognition of the LFR system: (i) numbers and (ii) realism. The number of Crowd

subjects is important as these subjects provide the majority of recognition transactions for calculation of the False Alert Rate (FAR). Crowd subjects are also necessary to truly replicate typical operational conditions for LFR.

5.5.2 Numbers: Based on MPS policing deployments to date, a typical operation deployment sees around 8000 crowd recognition transactions per day generated by those passing the LFR system. To achieve the evaluation objectives, between 40,000 and 60,000 non-mated recognition transactions are needed. At the lowest level, an expected 40 alerts are generated if the FPIR is 1 in 1000. This will enable evaluation objectives in relation to accuracy and demographic differential performance. To minimize the amount of Cohort and Crowd subject transaction data that needs to be collected, offline running of NeoFace Watch will enable alternative thresholds to be applied increasing FPIR and reducing FNIR, or vice versa. This will assist in revealing any demographic bias that may exist, but such biases may not be discernible at typical operating thresholds. This will also inform how police forces set their thresholds for future deployments.

5.5.3 Realism: To achieve the evaluation objectives, this plan outlines the importance of testing in operationally realistic circumstances. This necessitates the use of the passing public – the level of realism cannot be viably achieved with consenting volunteers alone. Accordingly, whilst the evaluation cannot (and for realism reasons would not) seek to control those passing the system, by monitoring flow count, and controlling when footage is gathered, it is possible to ensure the evaluation works within the tolerance outlined for non-mated recognition transactions.

6 Size and composition of cohort

6.1 Dataset size and statistical significance

6.1.1 The MPS and SWP have already reviewed algorithm performance tested by NIST[7] and recognized a high level of performance. This means that differences in face comparison accuracy and demographic performance can be expected to be small. Accordingly, this requires sufficient numbers of subject captures and FR decisions to enable statistically significant conclusions to be drawn.

6.2 How many subject captures & FR decisions are required to achieve the trial objectives?

6.2.1 To achieve the trial objectives a balanced cohort of at least 400 Cohort subjects will be required, with each subject undertaking ten LFR recognition transactions. (A total of 4000 recognition transactions). The need for 4000 Cohort Subject recognition transactions to determine accuracy and bias is made on the basis that the FNIR (i.e., 1–TRR) for LFR could be in the region of 10%. (The 10% FNIR figure is a reasonable basis on which to proceed given that the FNIR ranged between 11% and 46% in the 2016-19 trials with previous versions of NeoFace Watch, and algorithm performance (based on NIST testing) has improved since).

[7] E.g., NIST-IR 8271: Face Recognition Vendor Test (FRVT) Part 2 Identification. Draft supplement June 2021

6.2.2 To work through the point:

- Suppose the cohort comprises two demographic groups A and B of equal. Comparison of FNIR error rates for A and B would be based on the difference *D* in the observed error rates. The null hypotheses, that there is no difference in FNIR for the two demographics, will be rejected if the statistic $|D|$ exceeds the critical value *c*. The calculation of *c* depends on the significance level of the test, and the number of recognition transactions *n* per demographic group.
- Let $FNIR_A$ and $FNIR_B$ be the underlying FNIR for demographics A and B.
  If $FNIR_A = FNIR_B$ (the null hypothesis) then
  Probability[$|D| > c$] must be less than 0.05 (5% significance level).
  Furthermore, when $FNIR_A$ and $FNIR_B$ are not close in value (e.g., if the error rate for A is 1.5 times that for B) then the null hypothesis should be rejected.
  I.e., if $|FNIR_A - FNIR_B| > 4\%$ (*FNIR = 10%*, $FNIR_A = 12\%$, $FNIR_B = 8\%$), then
  Probability[$|D| > c$] must be greater than 0.8 (power of the test)
- The probabilities can be calculated using the normal approximation assuming transactions are independent and identically distributed. For these two inequalities to hold, a minimum number of transactions per group can be calculated.

6.2.3 In the above example case, the number of transactions per demographic group, n > 884. However, the actual number of transactions needed will be higher than 2n (2 x 884) as:

- Whilst the evaluation plan proposes a proportionate approach to testing, in some cases comparisons are made between more than two groups. For example, in relation to ethnicity the trial objectives necessitate cohort splits into three demographic groups.
- There are statistical dependencies between repeat transactions of the same test subject.
- FNIR error rates may be other than 10%.

6.2.4 As a result of the above analysis, 4000 transactions from 400 Cohort subjects achieves the required significance level to achieve the evaluation objective – a key point from a necessity perspective. The proposal is also proportionate, it ensures that excessive data processing is not taking place, whilst being consistent with the duty to take reasonable steps to understand algorithm performance. The proposal is also a practical one in terms of recruitment of a representative Cohort.

7 Subject, Image and LFR Video metadata

7.1 Cohort subject metadata

7.1.1 Metadata to be collected for each Cohort subject is detailed in Table 6 and Table 7 below. and Table 6 lists metadata that is necessary for the evaluation objectives, i.e. the metadata that provides ground truth as to matches between probe images and reference images, and ground truth regarding subject demographic. Table 7 lists metadata needed for DPA compliance. Some metadata is in both categories.

Table 6 – Cohort subject metadata: Necessary for the evaluation objectives

Metadata	How obtained	Notes
a) URN (unique reference number)	Assigned	Pseudonymous identifier of data subject.

b) Demographics: Ethnicity	Self-declared: A1, A2, A3, A9, B1, B2, B9, W1, W2, W9	Ground truth data on demographics
c) Demographics: Gender	Self-declared: F, M	
d) Demographic: Age	Self-declared: Age in years	
e) Skin tone	Self-declared: Fitzpatrick scale[8]	
f) Height	Self-declared: Height in cm or ft and ins - (measure if unknown)	
g) List of face images of subject	Recorded during data collation	
h) List of subject's LFR video appearances	Determined in data collation	These are the subject's recognition transactions

Table 7 – Cohort subject metadata: Necessary for DPA compliance

Metadata	Notes	
a) Subject Name		
b) Consent form	Location of signed consent form	
c) URN	Same as Table 6.a)	Links to imagery and metadata of cohort subject
d) List of face images of subject	Same as Table 6.g)	
e) List of subject's LFR video appearances	Same as Table 6.h)	
f) Supplying Agency	Necessary for commercial reasons and to provide a method of contact	

7.2 Filler subject metadata

7.2.1 Metadata to be collected for each Filler subject is detailed in Table 8 and Table 7 below. and Table 8 lists metadata that is necessary for the evaluation objectives, i.e. the metadata that provides ground truth as to matches between probe images and reference images, and ground truth regarding subject demographic. Table 9 lists metadata needed for DPA compliance. Some metadata is in both categories.

7.2.2 Note that Table 9 allows for a renaming of images and assigning different pseudonymous IDs should it be needed for consistent naming conventions in the curated test data set being returned to MPS. This metadata allows the data to be reconciled against the source MPS records.

Table 8 – Filler subject metadata: Necessary for the evaluation objectives

Metadata	How obtained	Notes
a) (Filler)URN	Assigned URN for filler subject shall be distinct from cohort subject URNs	Pseudonymous identifier of data subject.
b) Filename of custody image		

[8] https://en.wikipedia.org/wiki/Fitzpatrick_scale

c) Demographics: ethnicity	As per MPS source records	
d) Demographics: gender	As per MPS source records	
e) Demographics: age	As per MPS source records	
f) Pre 2016/Post 2016 image capture	As per MPS source records	From 2016 controls in place to achieve better quality custody images.

Table 9 – Filler subject metadata: Necessary for DPA compliance

Metadata	How obtained	Notes
a) MPS assigned URN b) MPS assigned filename of custody image	Assigned by MPS on when providing the filler dataset	MPS maintains link from this URN and filename to subject details in MPS custody image holdings.
c) (Filler)URN d) Filename of custody image	Same as Table 8.a) & b) The URN and custody image filename of filler subject in the curated corpus of test data assembled in the evaluation.	If there is a need to assign a new set of URNs / image filenames for consistency of labelling in the resulting curated corpus, this data provides the mapping against the original MPS data.

7.3 Crowd subject metadata

7.3.1 No individualised metadata is gathered for crowd subjects. Crowd image data will be collected passively during the operational deployments. Data subjects forming part of the crowd will NOT be separated into a specific dataset. LFR video footage will be sampled to survey the gender, age and ethnicity balance, and the total number of Crowd subjects appearing in the video. The aggregate numbers are necessary for analysis of the False Alert Rate, and variations by demographic category. See Table 10.

Table 10 – Aggregated crowd subject metadata: Necessary for evaluation objectives

Metadata	How obtained	Notes
Count of Crowd subject appearances in LFR video (recognition transactions)	Estimated though test staff review of samples of video footage, and counting people in Zone of Recognition, and perceived ethnicity, gender, and age. This will be undertaken on a survey basis, with no details being recorded against individual Crowd subjects.	The count of Crowd subject recognition transaction by demographic and in aggregate is necessary for calculation of the false alert rate. See 6.5.2 for the need to measure false alert rate over the large number of non-mated transactions.
Count of Crowd subjects by demographic Facemask: N/Y Gender: F/M Ethnicity: IC1, IC3, IC4, Other Approx Age: 0-10, 10-20, 20-40, 40+		

7.4 Video and face image metadata

7.4.1 Table 11 and Table 12 list the metadata associated with the LFR video, Cohort Subject facial images, and Filler facial images. Metadata is more complex in the case of LFR video, which contains facial images of many Cohort subjects, and unknown crowd subjects.

7.4.2 Most cameras will store EXIF metadata within the image file. Table 11 and Table 12 do not list this EXIF metadata. It is possible that some of this meta data is pertinent to FRT algorithm performance. The EXIF metadata will be reviewed, and a decision made on what, if any, of this metadata should be retained in the final collated image set that is returned to MPS for future testing purposes.

7.4.3 It is quite possible that the Filler Facial Images vary in terms of settings such as camera resolution, image compression. If these factors seem pertinent to variations in performance, consideration will be given to inclusion of such metadata for these images.

Table 11 – LFR video metadata: Necessary for evaluation objectives

Metadata	How obtained	Notes
Filename		
Date & start time	Time stamp on video	
Duration		
Camera details & settings	Record per camera at deployment	
Dimension of zone of recognition	Record per camera at deployment	
Periodic log of weather conditions	Record hourly or on significant change in conditions	
Periodic log of illumination level	Record hourly or on significant change in conditions	
Cohort recognition transactions	Logged as cohort subjects walk through zone of recognition (bar code reading app)	URN & time
Count of crowd recognition transactions & demographics	See Table 10	
How crowded is zone of recognition (Estimate of faces processed per minute)	NEC analysis tools	Aim is to know whether the crowd flow is comparatively busy or quiet. A precise count of the crowd may not be possible.

Table 12 –Cohort facial image metadata – Necessary for evaluation objectives

Metadata	How obtained	Notes
Filename		
URN of cohort subject	Logged at time photo is taken	
Date and time		
Camera device		
Image type	Image types include • Reference custody image • OIFR mobile device image • Selfie Image • Mobile phone camera • Ad-hoc digital camera photo • Surveillance photo	Image types are those typical for the MPS and SWP use cases. With cohort attending on two adjacent days, a second time separated image can be taken of most image types. This is necessarily a subset of the diverse set of possible image types for RFR.

	• Crop of low-resolution CCTV video	
Environmental factors	• Face mask / None • Indoor / Outdoor	

8 Offline running of facial recognition algorithms

8.1 This section specifies the sets of mated and not-mated recognition transactions to be conducted for offline processing by the NeoFace Watch and NeoFace Reveal algorithms.

8.2 Collected results (i.e., recognition results and comparison scores) can then be analysed per demographic group, per image type (for RFR), per environmental condition (for LFR), providing comparisons of recognition performance between demographic groups etc, statistical analysis of significance, and graphics representation of results.

8.3 The tests are designed to minimise processing to that necessary to accomplish the objectives of the evaluation. The offline batch-running of the LFR, RFR, OIFR replicates the live running of the facial recognition algorithms but eliminates (most) test staff involvement in examination and adjudication of face images and facial recognition results. 'Ground truth' regarding identities/demographics are, as far as is possible, established as part of the data collection processes.

8.4 Output from the offline OIFR/RFR processes does not include images. However, output of the offline LFR process shows, for each alert, the alerted face image of the video feed, and the Candidate image of the matched face. Prior to the analysis step, these images should be replaced. If the alerted face is of a Cohort subject, the face image should be replaced with the Cohort subject URN. If the alerted face is of a Crowd subject, the face image should be replaced with an anonymous ID based on perceived demographics (e.g., the alerted subject is labelled AF40 if they are perceived to be an Asian Female aged between 20 and 40). This preserves the ability to count the number of false alerts per demographic.

8.5 Note that in the case of LFR Non-mated recognition transactions (Table 13) there is a possibility that a person featuring on a LFR video-stream might also have an image in the Filler Reference Dataset. In such a case the test would inadvertently have included a mated recognition transaction. To avoid over-reporting of the false alert rate, when an alert in this test anomalous non-mated comparison scores (well beyond the normal range for "non-mated transactions" scores and well within the range of "mated transactions scores", a visual comparison of candidate and image causing the alert may help resolve whether the alert should be considered a true alert or false alert. Any such corrections shall be reported with the results (in accordance with the standard ISO/IEC 19795-1)

8.6 Offline running will take place on the hardware platform provided by MPS for this purpose, operating as a standalone computer unconnected to the internet, under the supervision of the NPL lead scientist for this project. This computer will be password protected and only project staff will have access to the computer, and the office will be locked when unoccupied.

8.7 Test transactions for LFR performance

8.7.1 Table 13, Table 14 and Table 15 detail the test transactions for determination of FAR, TRR, and FTAR performance of Live Facial Recognition.

Table 13 – LFR Non-mated recognition transactions for FAR

Algorithm	NeoFace Watch	Discussion
Watchlist	Filler non-mated reference dataset	Test counts False Alerts. With non-mated transactions no "true-alerts" should occur. See Note at 8.5.
Probe videos	Each LFR video (results from overlapping video streams to be combined before analysis) These contain recognition transactions for each person appearing on the video in the Zone of Recognition.	Combine alert list from overlapping video streams, to avoid double counting of alerts when a person is in the area of overlap.
Output	For each alert system provides • Watchlist Candidate ID • Comparison score • Time of alert in video stream • Alert face image (from video stream) • Candidate face image (from watchlist)	See Paragraph 8.4 above, regarding replacement of Alert Face Image / Candidate face image before passing results through to
Repeats	Repeat run over a range of alert thresholds	The alert threshold is configurable, and FAR should be calculated over a range of threshold values.
	Run with watchlist of post 2016 filler images Run with watchlist of pre- and post-2016 filler image	
Used to determine	FPIR(Watchlist size, Threshold)	

Table 14 – LFR mated recognition transactions for TRR

Algorithm	NeoFace Watch	
Watchlist	Cohort reference dataset PLUS Filler non-mated reference dataset	
Probe videos	Each LFR video (results from overlapping video streams to be combined before analysis)	
Output	For each alert system provides • Watchlist Candidate ID • Comparison score • Time of alert in video stream • Alert face image (from video stream) • Candidate face image (from watchlist)	Here we are only interested in whether the Cohort subjects are recognised at the selected alert threshold. Images in the output can be discarded.
Repeats	Repeat run over a range of alert thresholds	Same thresholds as Table 13
Used to determine	TPIR(Watchlist size, 1, Threshold)	

Table 15 – LFR mated recognition transactions for assessment of FTAR

Algorithm	NeoFace Watch	
Watchlist	Reference cohort images (set collected on day of enrolment) ONLY	
Probe videos	Each LFR video (results from overlapping video streams to be combined before analysis)	
Output	For each alert system provides • Watchlist Candidate ID • Comparison score • Time of alert in video stream • Alert face image (from video stream) • Candidate face image (from watchlist)	From the collected ground truth as to when cohort subjects are in the zone of recognition, we know when alerts for cohort subjects should occur. Running LFR at the lowest thresholds, a missing alert is most likely due to a failure to detect the subject's face. The candidate images, and alert face images can be removed from this output prior to analysis. These images provide no information on missed alerts!
Repeats	Run once at very low threshold (likely to match provided face is detected)	
Used to determine	FTAR Failure to acquire rate (face detection failures)	

8.7.2 For a fair comparison of LFR FAR between demographics, the non-mated comparison trials of Table 13 use a demographically balanced watchlist. In operational LFR the demographic composition of the watchlist varies according to the intelligence case for the deployment.

8.7.3 Repeating the offline LFR process of Table 13 using a subset of the Filler Reference Dataset, selected to match the demographic balance of the full set of MPS custody images will show the extent to which an imbalanced watchlist may alter the differences in FAR between demographics.

8.8 Test transactions for RFR performance

8.8.1 Table 16 and Table 17 detail the test transactions for determination of TRR, and Selectivity performance of Retrospective Facial Recognition

Table 16 – RFR mated recognition transactions for TRR

Algorithm	1) NeoFace Watch 2) NeoFace Reveal	
Reference database	Filler non-mated reference dataset PLUS Cohort reference dataset (from day 1)	
Probe image dataset	Cohort probe dataset (from day 2)	Split analysis by image type
Output	For each probe image: • Candidate image name and comparison score for top 200 matches	
Repeat	Repeat process, but use Day2 Reference custody images of cohort and Day 1 cohort images for probes	
Used to determine	TPIR(reference dataset size, rank, 0) for ranks 1 to 200	(i.e. determines the CMC cumulative match characteristic)

8.8.2 For systems such as RFR and OIFR that always return a number of top matching candidates, the biometric testing and reporting standard ISO/IEC 19795-1 recommends reporting of Selectivity, that measures the average number of candidates matched above comparison score threshold when the subject is NOT in the reference dataset.

Table 17 – RFR non-mated recognition transactions for Selectivity

Algorithm	NeoFace Watch NeoFace Reveal	
Reference database	Filler non-mated reference dataset	
Probe image dataset	Cohort face image taken on day 1 or 2	Split analysis by image type
Output	For each probe image: • Candidate image name and comparison score for top 200 matches	Over a range of comparison score thresholds, for each probe image, the number of candidates above threshold are counted, and averaged over all probes.

Used to determine	Selectivity SEL(reference dataset size, 200, threshold)	

8.9 Test transactions for OIFR performance

8.9.1 Table 18 details the test transactions for determining TRR performance for Operator Initiated Facial Recognition

Table 18 – OIFR mated recognition transactions for TRR

Algorithm	NeoFace Watch NeoFace Reveal	
Reference database	Filler Reference Dataset PLUS (Day1) Cohort Reference Dataset	
Probe image dataset	(Day2) OIFR Cohort face images	
Output	Probe image name Reference image name and comparison score for top 200 matches	
Repeat	Repeat process, but use Day2 Cohort Reference custody images And Cohort Day 1 OIFR for probes	
Used to determine	TPIR(N, 6, 0)	TRR with N = reference dataset size Results up to rank 6 returned No comparison score threshold

8.10 Effect of Demographic Imbalance in Reference Dataset

8.10.1 Project partner Ingenium has a paid license for use of the Morph Longitudinal Dataset (a large dataset of demographically labelled face images) for face recognition product evaluation. While this dataset is not fully representative of the MPS/SWP LFR, RFR and OIFR use cases, i.e. it cannot substitute for the use of Cohort Subjects, or Filler database, its use will add to the understanding of how an imbalance between subject demographics in a reference dataset affects bias in recognition accuracy for different demographic groups, and also addresses the question of whether evaluation findings based on the MPS/SWP images are replicated in when using datasets from other locations or use cases. Using this dataset as described below will add to MPS/SWP understanding of their algorithms

8.10.2 The experiment described here will compare RFR performance for a demographically balanced reference, with that when the demographics are imbalanced.

Table 19 – Test transactions for investigating demographic imbalance in reference dataset

Algorithm	NeoFace Watch NeoFace Reveal
Probe image dataset	Demographically balanced selection of probe data from Morph dataset (Gender: Male/Female & Ancestry: European/African)
Reference databases	**(a) Balanced demographic references:** Mated, demographically balanced reference data from Morph dataset (mated references for selected probe images) PLUS, demographically balanced non-mated reference dataset (from Morph dataset)
	(b) Imbalanced demographic references Mated, demographically balanced reference data from Morph dataset (mated references for selected probe images) PLUS, demographically imbalanced non-mated reference dataset from Morph dataset
Output	For each probe image name Reference image name and comparison score for top 200 matches (count number of scores above threshold)
Repeat?	Run for both Balanced and Imbalanced reference datasets
Used to determine	TPIR(reference dataset size, rank, 0) for ranks 1 to 200 Resulting CMCs (for male gender/ female gender/ European ancestry/ African ancestry) for the demographically balanced case can be compared with those of the imbalanced case. Does the imbalanced case show a great spread?

9 Performance analyses

9.1 Collation of offline results

9.1.1 Following offline running of the LFR, RFR, OIFR tests, there is a stage of collation of the results into a form suitable for performance analyses. This involves

- Addressing any anomalous results (that may require a test to be re-run)
- Removing any images from the results file. (LFR uses images in its output log, but these are no longer required for analyses)
- Linking the mated and non-mated recognition transactions to metadata of subjects, and video (LFR), image types (RFR).

9.1.2 At this point conceptually we have a list of transactions for a test giving the results and metadata of each transaction. E.g., for Live Facial Recognition TRR:

Date Time of Mated Transaction:	Video camera:	Cohort subject ID:	Ethnicity:	Gender:	Age:	Illumination level at Time of Transaction:	Threshold setting:	Alert (Y/N):	Comparison Score	…

From such a list, it is easy to select and count the number of mated transactions alerted for subjects meeting a given demographic to generate the accuracy measures for that demographic.

9.2 LFR Accuracy & Equitability

9.2.1 Accuracy metrics:

- FAR – False Alert Rate at threshold T
- TRR – True Recognition Rate at threshold T
- Measure over default and other operationally plausible thresholds T
- Graphically plot recognition error trade-off.

9.2.2 Accuracy measure for:

- Over all demographics
- By ethnicity class (A, B, W)
- By gender class (F, M)
- By age class (Determine class boundaries once age distribution of cohort and filler dataset known)
- By height (shortest 33 percentile, middle 33 percentile, tallest 33 percentile)
- By Fitzpatrick scale
- Where relevant, measure accuracy over intersections in demographic class,

9.2.3 Statistical assessment of significance of accuracy differences between classes

9.3 RFR/OIFR Accuracy & Equitability

9.3.1 Performance metrics:

- TRR – True recognition rate at rank R from R=1 to R=200 (R=6 for OIFR)
- Calculate TRR separately for different types of probe image
- Plot Cumulative Match Characteristic: TRR-at-Rank-R against R.

9.3.2 Accuracy measure for:

- Over all demographics
- By ethnicity class (A, B, W)
- By gender class (F, M)
- By age class (Determine class boundaries once age distribution of cohort and filler dataset known)
- Where relevant, measure accuracy over intersections in demographic class

9.3.3 Statistical assessment of significance of accuracy differences between classes

9.4 Effect of composition of Watchlist

9.4.1 Analyse accuracy and equitability results for the imbalanced watch list tests

- For LFR, test described at 8.7
- For RFR, test described at 8.8

9.4.2 Analyse accuracy and equitability differences between post-2016 custody images, and pre-2016 custody images.

10 Completion

10.1 The findings of the evaluation will be reported to MPS/SWP, along with a final report suitable for publication.

10.2 The project is expected to complete during the 3rd quarter of 2022 (An exact date is not yet known, due to dependencies e.g. on the running of deployments at which data collection occurs).

10.3 On completion of the evaluation, the curated datasets assembled in the evaluation will be delivered to MPS for future Policing use in testing of Face Recognition Technology, with documentation describing the dataset, its structure and use,

10.4 MPS hardware, encrypted hard drives used for the evaluation will be returned to MPS.

10.5 Any residual image/video data on NPL devices will undergo a secure wipe.