National Physical Laboratory

REPORT 4674

REPORT FOR HOME OFFICE AND OFFICE OF THE POLICING CHIEF SCIENTIFIC ADVISER

ACCURACY AND EQUITABILITY EVALUATION OF IDEMIA FACIAL RECOGNITION ALGORITHM FOR HOME OFFICE STRATEGIC FACIAL MATCHING

OCTOBER 2025

Accuracy and Equitability Evaluation of Idemia Facial Recognition Algorithm for Home Office Strategic Facial Matching

Data Science and AI Department

National Physical Laboratory
Hampton Road, Teddington, Middlesex, TW11 0LW

CONTENTS

1 INTRODUCTION

1.1 ABOUT THE EVALUATION

A key strategic priority for policing is the continued development, use and assessment of facial recognition (FR) technology. Critical to use of the technology is ensuring it is implemented in a responsible, transparent, and ethical way: doing so requires an understanding of the accuracy and demographic equitability of the technology.

The objective of the evaluation described in this report is to assess the performance of a specific FR algorithm in an operational setting in terms of accuracy and equitability related to subject demographics. The FR technology and version evaluated is the Idemia facial recognition algorithm which will be used by the Home Office Strategic Facial Matching service, for which the predominant use case is Retrospective Facial Recognition (RFR). The algorithm may also be used for Operator Initiated Facial Recognition (OIFR).

NPL was commissioned by the Home Office in collaboration with the Office of the Policing Chief Scientific Adviser. This assessment will add to Home Office's and Law Enforcement's understanding on how their FR systems perform and will provide information to help configure FR technology for effective and fair deployment on operational use cases.

The evaluation, conformant with the standards ISO/IEC 19795-1 [1] and ISO/IEC 19795-2 [2], was conducted as a 'technology evaluation' using the Metropolitan Police Service (MPS) 'Equitability Study Dataset' collected in 2022 for evaluation of the NEC NeoFace algorithm [3].

For the operational use cases, the evaluation:

- determines the recognition accuracy of the FR algorithm for images of the Equitability Study Dataset
- determines and assesses the statistical significance of variations in recognition accuracy between demographic groups
- assesses the equitability of the FR algorithm through the consideration of how variations in accuracy between demographic groups impact outcomes for data subjects involved in operational use.

This report sets out the findings of the evaluation, and is organised as follows:

- The remainder of this section summarises key findings of the evaluation.
- Section 2 provides background on the operational use cases, the demographics addressed in the evaluation, and the assessment of equitability.
- Section 3 outlines the data of the Equitability Study Dataset used in the evaluation.
- Section 4 provides details regarding the evaluation methodology.
- Section 5 reports the recognition accuracy for the evaluation data.
- Section 6 examines the variation in performance between demographic categories.
- Section 7 addresses equitability of the FR algorithm on the operational use case.
- Section 8 provides a glossary of the key terms and abbreviations used in this report.

1.2 KEY FINDINGS

Retrospective Facial Recognition (RFR) is a post-event use of FR technology which compares still images of faces of unknown subjects against a reference image database in order to identify them. Operator Initiated Facial Recognition (OIFR) is a near-real-time use of FR technology, where an officer takes a photograph of a subject via a mobile device and submits it for immediate search against a reference image database.

In both use cases the identification produces a candidate list of the reference images that best match the submitted image. The identification search generally has two configurable parameters: (i) the number of top matches to be returned, and (ii) a face-match threshold that specifies how similar the submitted image (referred to as a probe image) and reference image must be for the reference to be included in the candidate list.

An accurate FR system requires that, if the person in the submitted image also has a face image in the reference database, this mated image and associated identity should be among the candidates returned by the identification search, preferably first in the ranked list of candidates. Also, to avoid false matches, if the reference database does not hold a mated face image, then none of the returned candidates should have a similarity score exceeding the face-match threshold.

Recognition accuracy for RFR and OIFR is stated in terms of the True Positive Identification Rate (TPIR), which is measured over mated identification searches (where the person in the probe image has a reference image in the reference database) and the False Positive Identification Rate (FPIR) which is measured over non-mated identification searches (where the person in the probe image does not have a reference image in the reference database).

- The True Positive Identification Rate, TPIR(N, R, T), is the proportion of mated identification searches that include the mated reference among the candidates returned. TPIR depends on the number of candidates to be returned *(R*), the number of records in the reference image database (N), and the face-match threshold (T). If $T = 0$, no threshold is applied.
- The False Positive Identification Rate, FPIR(N, T), is the proportion of non-mated identification searches that return a candidate with a candidate score exceeding the face-match threshold (T).

To understand performance for non-mated identification searches (where the data subject is not in the reference database) it is necessary to consider FPIR together with TPIR at appropriate face-match thresholds.

1.2.1 Recognition accuracy

Recognition accuracy was measured over mated and non-mated identification searches using facial images from the Equitability Study Dataset. The same test data was used for the 2022 evaluation [3].

On this image set, with no thresholding applied, identification performance was perfect: of the 3513 mated identification searches the top match returned by the algorithm was the correct reference in every case.

- TPIR(180000, 1, 0) = 100.0 %

On the same image set, applying a face-match threshold in the range 3052 to 3981 (e.g., 3500):

(i) all mated identification searches returned the correct reference as the top match with candidate score exceeding the face-match threshold, and

(ii) no non-mated identification searches returned candidates with score above the face-match threshold.

- TPIR(180000, 1, 3500) = 100.0 %
- FPIR(180000, 3500) = 0.0 %

Further details are provided in Section 5.2.

1.2.2 Demographic variation in TPIR and FPIR

At face-match threshold values between 3052 and 3981, there is no demographic variation in performance. For all demographic groups of the Equitability Study Dataset, TPIR is 100 % and FPIR is 0 %.

Demographic variation in performance was analysed for some face-match thresholds outside this optimum range.

- At a face-match threshold 2010 (giving an overall FPIR of 5 %) the following observed demographic variations in FPIR are statistically significant at the 0.05 significance level:
  - the FPIR for subjects of Age 42-76 (2.8 %) was lower than that for subjects of Age 12-20 (4.6 %), Age 21-29 (8.2 %) and Age 30-41 (4.4 %).
- At a face-match threshold 7727 (giving an overall TPIR of 95 %) the following observed demographic variations in TPIR are statistically significant at the 0.05 significance level:
  - the TPIR for Asian subjects (97.4 %) is higher than that for White subjects (95.2 %) and Black subjects (92.5 %);
  - the TPIR for Male subjects (96.4 %) is higher than that for Female subjects (93.6 %);
  - the TPIR for Black Male subjects (95.3 %) is higher than that for Black Female subjects (90.2 %).

Further details are provided in Section 6.

1.2.3 Equitability

At face-match thresholds in the range 3052-3981, the recognition accuracy of the Idemia algorithm is perfect, there is no demographic variation, and the algorithm is considered equitable for both RFR and OIFR.

Further details are provided in Section 7.

2 BACKGROUND

2.1 RETROSPECTIVE FACIAL RECOGNITION

The predominant operational use case of the evaluation is Retrospective Facial Recognition (RFR). This is a post-event use of FR technology, where still images of faces of unknown subjects are compared against a reference image database in order to identify them. In operational use the candidate images returned by an identification search are adjudicated, typically by more than one officer.

The number of top matching candidates to be returned in an identification search will depend on the case under investigation. For the purposes of this evaluation, it is assumed that up to 200 top matches will be returned. Operationally, candidate lists may be somewhat shorter.

Probe images for RFR will vary in quality over operational use cases. The range of mated and non-mated comparison scores can change with image quality possibly requiring the face-match threshold to be tuned to the operational use case.

In assessing equitability of RFR, there are two classes of data subjects to consider: (i) the unknown subjects of the probe images submitted for identification search, and (ii) the individuals enrolled onto the reference database who may be added to a candidate list of an identification search.

The demographic composition of the reference database can affect the demographic variation in false positive identification rates. Also, the demographic profile of identification search probe images can affect the demographic balance of the reference database subjects being added to candidate lists.

2.2 OPERATOR INITIATED FACIAL RECOGNITION

Operator Initiated Facial Recognition (OIFR) is a near-real-time use of FR technology, where an officer takes a photograph of a subject via a mobile device and submits it for immediate search against a reference image database. For the purposes of the evaluation, it is assumed that up to six top matches will be returned. The officer adjudicates as to whether any of these is a correct identification.

In the case of OIFR, the officer-taken photographs will usually be of good image quality. If an image is of poor quality, the officer can request to take a further image.

A false positive recognition may impact the data subject photographed by OIFR, leading to further engagement between subject and officer.

The demographic composition of the reference database can affect demographic variation in false positive recognition rates.

2.3 DEMOGRAPHIC CATEGORIES

The demographic categories considered in the evaluation are those of the Equitability Study Dataset, and are limited to gender, age, and the ethnicities given below.

2.3.1 Ethnicity

Ethnicity is classified in accordance with the MPS self-defined ethnicity codes [4]. The data sets and analyses grouping of ethnicities are:

- Asian or Asian British (A1 Indian, A2 Pakistani, A3 Bangladeshi, A9 Any other Asian background);
- Black or Black British (B1 Caribbean, B2 African, B9 Any other Black Background);
- White (W1 British, W2 Irish, W9 Any other White background).

These ethnic groups were selected because they are: (i) the largest groups in the national population [5], and similarly (ii) the largest groups in the Policing arrest records [6]. Differences in performance of the FR algorithm for these demographics therefore have the greatest relevance to Policing.

2.3.2 Gender[1]

Self-defined gender categories for assessment are Female and Male.

2.3.3 Age

The data subjects in the Equitability Study Dataset have ages that range from 12 to 70+ years old. Age distribution of the data subjects is similar for both the Cohort and Filler portions of the Dataset, and approximately replicates the age distribution in MPS custody images.

2.4 ASSESSING EQUITABILITY

The evaluation is concerned with assessing FR accuracy, variations in performance for different demographics, and equitability between demographics in operational settings. Demographic variation in accuracy performance may be more easily observable at settings outside the normal operational parameters.

Equitability between demographics requires that, in the operational setting, the outcomes for the subjects (i.e., recognition rates and false alert rates) should be broadly equivalent for demographics considered. We cannot require exact equivalence as, even when there is no demographic variation in performance, due to the statistical nature of biometrics small deviations in observed performance can be expected. Thus, in assessing whether a system is equitable, criteria are needed for broad equivalence of performance figures.

In this evaluation we use the following criteria to determine demographic equitability.

The system is considered equitable if:

a) there is no variation in performance between demographic groups (e.g., when there are no false positives at the face-match threshold applied),
or

b) the variation in performance between demographic groups is not statistically significant,
or.

c) the variation in performance between demographics is inconsequential for the data subject in the operational setting (i.e., with operational thresholds, composition of reference database, etc.).

[1] **Gender:** classification as male, female, or another category based on social, cultural or behavioural factors. (Gender is generally determined through self-declaration or self-presentation and may change over time.)

Note that the evaluation is assessing equitability of the algorithmic outcomes only. Mitigation of bias by operator adjudication is beyond the scope of this evaluation.

2.5 STATISTICAL SIGNIFICANCE

Statistical significance quantifies whether the observed performance difference is likely due to chance, or due to some underlying factor of interest. For testing of statistical significance in the evaluation, we use the conventional significance level of 0.05 (5 %). The significance level relates to the probability of falsely rejecting the ‘null’ hypothesis of no underlying difference in performance rates.

Note that testing for performance variation over different demographic attributes involves multiple hypothesis tests. A multiplicity of tests each at 0.05 significance level can increase the probability of falsely rejecting the null hypothesis to a value higher than 0.05. Methods are available that address this issue requiring stricter significance thresholds for each individual test; however, use of these methods can increase the chance of falsely accepting the null hypothesis when there is, in fact, an underlying difference in performance rates. Such methods were not needed in this evaluation.

3 TEST CORPUS

3.1 DATA SUBJECTS

Cohort subjects in the Equitability Study Dataset were drawn from two sources: (i) acting/extras agencies, and (ii) a group of under 18-year-old volunteers from the Police Cadets.

These data subjects provided informed consent to the use of their images and metadata for further evaluation of FR for policing purposes. The MPS Data Office holds the reconciliation table to allow test subject names and identifiers to be reconciled for the purposes of exercising data rights.

To achieve the Evaluation Objectives, data was collected from 401 cohort subjects. A smaller cohort would reduce the power of the evaluation to reveal statistically significant variation in performance between demographics at anticipated accuracy levels. The composition of the cohort is shown in Table 1 below.

Table 1: Demographic composition of Cohort – 401 data subjects

		Female	Male
Self-defined ethnicity[2]	Asian (A1, A2, A3, A4, A9)	53	45
	Black (B1, B2, B9)	60	51
	White (W1, W2, W3, W9)	82	86
	Mixed and Other	8	16
Age	Age Range	12-76 years	
	Lower quartile	21 years	
	Median	30 years	
	Upper quartile	42 years	

3.2 COHORT SUBJECT FACE IMAGES

Face images in the Equitability Study Dataset can be partitioned by device and conditions for image capture.

Custody-style images were taken in accordance with the Police Standard for capture of Facial mugshots [7], i.e., (i) image of the full head with all hair, neck, shoulders, and ears; (ii) subject facing square to the camera, looking directly at camera; (iii) indoor photos with diffuse lighting providing uniform illumination across the face without hot spots or shadows, and with a plain flat background with 18 % shade of grey. Images were taken with a Canon EOS850D camera. In cases of non-conformance a further image was taken, and the non-conformant image relegated to a set of outtakes. The custody images were used for enrolling reference images of the cohort, and as probe images for non-mated identification searches.

OIFR-style facial images were taken on a Samsung XcoverPro smartphone. Images were taken both indoors and outdoors. In cases where the image is out-of-focus, there is motion blur, or the subject's eyes were closed, a further image would be taken and the original image relegated to a set of outtakes.

Adhoc digital camera photographs of subjects taken both indoors and outdoors using a Nikon D40 or Nikon 1J5 camera. Background and lighting conditions were uncontrolled, and images were taken without flash or supplementary lighting. The adhoc camera images are similar to the OIFR-style images though the depth of field is shallower.

[2] Self-defined ethnicity based on Office for National Statistics five high-level ethnic groups [5].

Selfie photos were taken by the cohort subject themselves using a Motorola G60S smartphone. Subjects were allowed to pose as they wished, a neutral expression was not required, and the subject did not need to pose 'square to the camera'. Selfies were taken both indoors and outdoors.

These four image types form the 'initial portion' of the images of Equitability Study Dataset. These images were used for the 2022 evaluation [3]. The combined Custody-style/ OIFR-style/Adhoc/Selfie image data from the Equitability Study Dataset form the probe data dataset previously used in the evaluation of RFR using the NEC NeoFace algorithm in 2022. The Cohort demographic for these image types is balanced. Photographic image quality and subject pose are generally good. Other than for the custody-style image, illumination, backgrounds and subject expression were uncontrolled, and the wearing of sunglasses and sunhats was permitted.

These images have been cropped and downsized to a head and shoulder image of 1000 x 1000 pixels, with an inter-eye distance (IED) of approximately 200 pixels. Table 2 gives image quality details for the different types of images.

Table 2: Face image quality of Cohort image types

	Custody-style	OIFR-style	Adhoc	Selfie
Number of images …	684	1135	1080	1015
… from number of subjects	401	401	401	400[3]
Inter-eye distance (pixels)	~200	~200	~200	~200
Frontal pose	Y	Y	Y	
Neutral expression	Y			
All parts of face visible, Entire face in image	Y	Y	Y	Y
Eyes open, eyes visible	Y	Y	Y	Y
No squinting	Y			
Controlled illumination, Even lighting on face	Y	N	N	N
Plain background	Y			
No sunglasses or thick rims No hats	Y			
Face in focus, no blur	Y	Y	Y	Y
Note: Y indicates the quality aspect was mandatory for photographing the image type. N indicates that the condition was not met for most images of the given type. Blanks indicate that the condition was not required and not met on some occasions.				

[3] Selfie image not collected for one subject.

3.3 FILLER IMAGES

To provide a large demographically balanced reference database, Cohort data was supplemented by approximately 180000 Filler images, provided from MPS holdings of custody image photos. The composition of the Filler set is shown in Table 3.

Table 3: Demographic composition of Filler Image Set, ~180000 facial images from ~116000 individuals

		Female	Male
Self-defined ethnicity[2]	Asian (A1, A2, A3, A4, A9)	~30000	~30000
	Black (B1, B2, B9)	~30000	~30000
	White (W1, W2, W3, W9)	~30000	~30000
Age at date image taken	Age range[4]	12-76 years	
	Lower quartile	22 years	
	Median	31 years	
	Upper quartile	40 years	

Each Filler image corresponds to a custody record, and the Filler dataset contains multiple images for some individuals who have multiple custody records. Guidance on reference database composition [8] suggests that, if multiple different images of a subject are available, consideration be given to including these to improve the likelihood of a match. Thus, the inclusion of multiple images for some individuals is not atypical of the operational use case.

[4] Age range of the Filler dataset taken as the 0.1^{th} – 99.9^{th} percentiles.

4 METHODOLOGY

Methodology follows the standards ISO/IEC 19795-1: 2021 [1] and ISO/IEC 19795-2: 2015 [2] on biometric performance testing and reporting. Testing was conducted as a technology evaluation using the Equitability Study Dataset collected for MPS in 2022 as the corpus of test data.

Reference and Probe image data
The reference image dataset consists of custody image photos from the Filler dataset with the addition of reference images for Cohort subjects (a Custody-style image of each Cohort subject taken without facemask or glasses). Probe image data comprised the remaining Cohort subject images. The images provide 3513 mated identification searches, and 3914 non-mated identification searches against a reference database of approximately 180000 images.

Processing of identification searches
For the purposes of the evaluation of the FR algorithm Idemia provided an executable script for encoding face images and conducting identification searches. This enabled batch processing of the enrolments, mated/non-mated identification searches for the evaluation.

Identical twins
In the Cohort there was a pair of identical twins: identical twins are known to be a challenging case for FR. In the evaluation neither twin was ever mistakenly recognised as the other. The non-mated comparison scores between the identical twins were (unsurprisingly) higher than that usual for non-mated comparison scores, and within the range typical for mated comparison scores.

The non-mated comparisons between identical twins are omitted from performance analysis of non-mated identification searches.

Failure-to-extract cases
In the analyses, failure-to-extract cases are treated as if the identification search had been completed with no candidate being returned. This approach removes the need to analyse demographic difference in failure-to-extract cases separately from demographic difference in TPIR and FPIR**.**

Size of reference image database
The number of filler images enrolled into the reference database is slightly below 180000 due to cases where the FR algorithm cannot extract facial details to create a matchable reference. The algorithm may record a failure-to-enrol or instead may create a non-matchable (possibly null) reference to the same effect. Both approaches are consistent with treatment of failure-to-extract cases in the evaluation, though result in slightly different number of reference images. Such differences in gallery size have negligible effect on measured performance. In this report we will write TPIR(180000, *R*, *T*) and FPIR(180000, *T*).

For this evaluation, combining filler images with those form Cohort custody-style provides 180183 references for mated identification searches, and 180182 for non-mated identification searches (which exclude the mated reference).

5 PERFORMANCE ANALYSES

5.1 MATED IDENTIFICATION SEARCHES WITHOUT THRESHOLDING

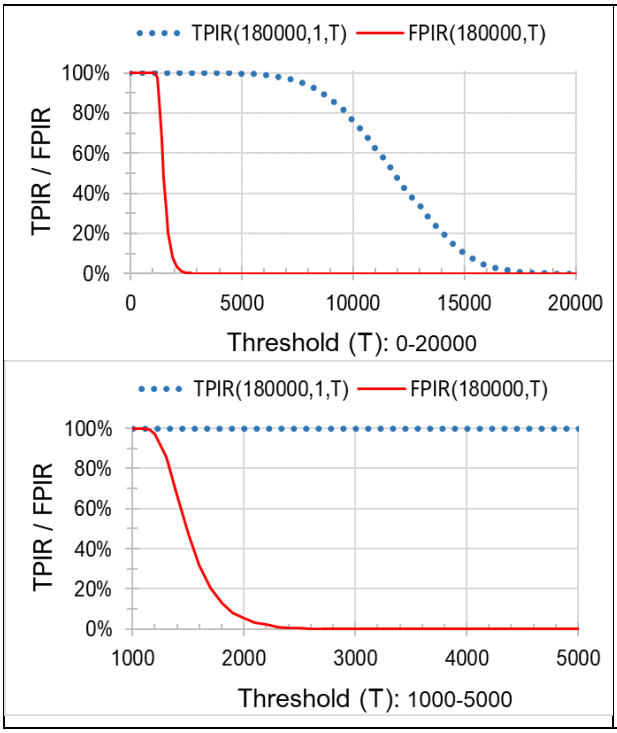
Combining the facial image types, all 3513 mated identification searches returned the correct reference at rank 1. It follows that for any demographic subset of the evaluation dataset, the recognition accuracy is identical.

- TPIR(180000, 1, 0) = 100.0 %

5.2 MATED AND NON-MATED IDENTIFICATION SEARCHES WITH FACE-MATCH THRESHOLD

The unthresholded True Positive Identification Rate TPIR(*N*, *R*, 0) informs on the recognition performance for mated identification searches. If an unmated identification search is performed, unless a face-match threshold is applied, any candidates returned would be false positives. To understand performance for non-mated identification searches it is necessary to consider FPIR together with TPIR at appropriate face-match thresholds.

In the evaluation, the mated comparison scores ranged between 3981 and 18922 while the rank 1 non-mated comparison score ranged from 1077 to 3051. These ranges do not overlap. Figure 1 shows how the TPIR (for mated identification searches) and FPIR (for non-mated identification searches) vary as a function of the face-match threshold.



Face-match threshold	TPIR	FPIR
1000	100.0 %	100.0 %
1860	100.0 %	10.0 %
2010	100.0 %	5.0 %
2212	100.0 %	2.0 %
2343	100.0 %	1.0 %
3052	100.0 %	0.0 %
3981	100.0 %	0.0 %
6040	99.0 %	0.0 %
6660	98.0 %	0.0 %
7727	95.0 %	0.0 %
8640	90.0 %	0.0 %

Figure 1: TPIR & FPIR by threshold

Using a face-match threshold from the gap between the ranges (e.g., 3500) gives optimal performance for the Equitability Study Dataset, all mated identification searches find the mated reference, and no non-mated identification search returns a false match.

- TPIR(180000, 1, 3500) = 100.0 %
- FPIR(180000, 3500) = 0.0 %

With perfect performance there is no variation in TPIR or FPIR across demographics.

6 VARIATION IN PERFORMANCE BETWEEN DEMOGRAPHICS

As noted in Section 5, selecting a face-match threshold between 3052 and 3981 achieves perfect performance (TPIR = 100.0 % and FPIR = 0.0 %) on the initial portion of the Equitability Study Dataset. At such setting there is no demographic variation in performance.

6.1 DEMOGRAPHIC VARIATION IN FPIR (thresholds below 3052)

At thresholds below 3052, there can be demographic variation in FPIR which is nonzero but there is no demographic variation in observed TPIR.

Table 4 shows the observed demographic variation in FPIR results over a range of face-match thresholds selected to achieve overall FPIR of 0 %, 1 %, 2 %, 5 % and 10 %. The extent of demographic variation is different at different face-match thresholds.

Table 4: FPIR by demographic and face-match threshold

Demographic	FPIR(180000, *T*)				
	T = 3052	*T* = 2343	*T* = 2212	*T* = 2010	*T* = 1860
All	0.0 %	1.0 %	2.0 %	5.0 %	10.0 %
Female	0.0 %	0.9 %	2.3 %	5.7 %	11.1 %
Male	0.0 %	1.0 %	1.7 %	4.2 %	8.8 %
Asian	0.0 %	1.1 %	2.4 %	6.3 %	11.8 %
Black	0.0 %	1.2 %	2.4 %	6.6 %	11.9 %
White	0.0 %	0.8 %	1.5 %	3.5 %	7.8 %
Asian/Female	0.0 %	0.4 %	2.1 %	5.9 %	11.1 %
Asian/Male	0.0 %	1.9 %	2.8 %	6.7 %	12.6 %
Black/Female	0.0 %	1.9 %	3.8 %	9.1 %	15.6 %
Black/Male	0.0 %	0.5 %	0.8 %	3.6 %	7.5 %
White/Female	0.0 %	0.5 %	1.1 %	3.2 %	7.9 %
White/Male	0.0 %	1.1 %	1.9 %	3.8 %	7.8 %
Age 12-20	0.0 %	0.8 %	2.2 %	4.6 %	12.0 %
Age 21-29	0.0 %	1.9 %	3.2 %	8.2 %	12.9 %
Age 30-41	0.0 %	0.7 %	1.8 %	4.4 %	9.5 %
Age 41-76	0.0 %	0.6 %	0.9 %	2.8 %	5.8 %

To determine the statistical significance of observed demographic variations, a *t*-test (Welch's unequal variance *t*-test [9]) was used when comparing FPIR of two demographic groups and analysis of variance (ANOVA) was used when comparing FPIR of three or more groups.

At face-match thresholds of 2343 (overall FPIR of 1 %), 2212 (overall FPIR of 2 %), and 2120 (overall FPIR of 3 %), the observed demographic variations in FPIR were not found significant at the 0.05 level.

Table 5: Statistical significance of demographic variation in FPIR at threshold 2010

Demographic		FPIR(180000, T = 2010)		Significance of variation	
		Mean for demographic	Standard error of mean	p-value	at significance level 0.05
All cohort subjects		5.0 %	0.6 %		
Gender	Female	5.7 %	0.9 %	p = 0.22	
	Male	4.2 %	0.8 %		
Ethnicity	Asian	6.3 %	1.3 %	p = 0.088	
	Black	6.6 %	1.4 %		
	White	3.5 %	0.9 %		
Ethnicity & Gender	Asian/Female	5.9 %	1.6 %	p = 0.068	
	Asian/Male	6.7 %	2.2 %		
	Black/Female	9.1 %	2.3 %		
	Black/Male	3.6 %	1.1 %		
	White/Female	3.2 %	1.2 %		
	White/Male	3.8 %	1.3 %		
Age	12-20 years	4.6 %	1.1 %	p = 0.017	statistically significant Ages ≥ 42 v Ages < 42 (p = 0.014)
	21-29 years	8.2 %	1.9 %		
	30-41 years	4.4 %	1.0 %		
	42-76 years	2.8 %	0.9 %		

Table 5 considers the demographic variation in FPIR at face-match threshold 2010 (overall FPIR of 5 %). At this threshold the following demographic variations in FPIR are statistically significant at the 0.05 level.

- The FPIR for subjects of Age 42-76 (2.8 %) is lower than that for the combined group of subjects of Age 12-20 (4.6 %), Age 21-29 (8.2 %), and Age 30-41 (4.4 %).

6.2 DEMOGRAPHIC VARIATION IN TPIR (thresholds above 3981)

At thresholds above 3981 there can be demographic variation in TPIR which will be below 100 %, but there is no demographic variation in observed FPIR.

Table 6 shows the observed demographic variation in TPIR results over a range of face-match thresholds selected to achieve an overall TPIR of 100 %, 99 %, 98 %, 95 % and 90 %. The extent of demographic variation is different at different face-match thresholds.

Table 6: TPIR by demographic and face-match threshold

Demographic	TPIR(180000, 1, *T*)				
	T = 3981	*T* = 6040	*T* = 6660	*T* = 7727	*T* = 8640
All	100.0 %	99.0 %	98.0 %	95.0 %	90.0 %
Female	100.0 %	98.9 %	97.8 %	93.6 %	87.5 %
Male	100.0 %	99.1 %	98.1 %	96.4 %	92.6 %
Asian	100.0 %	99.5 %	98.8 %	97.4 %	95.2 %
Black	100.0 %	98.7 %	97.7 %	92.5 %	85.1 %
White	100.0 %	98.8 %	97.5 %	95.2 %	89.7 %
Asian/Female	100.0 %	99.4 %	98.7 %	96.7 %	93.5 %
Asian/Male	100.0 %	99.5 %	98.9 %	98.3 %	97.2 %
Black/Female	100.0 %	97.9 %	96.8 %	90.2 %	82.0 %
Black/Male	100.0 %	99.7 %	98.9 %	95.3 %	88.8 %
White/Female	100.0 %	99.2 %	97.7 %	94.3 %	87.4 %
White/Male	100.0 %	98.5 %	97.3 %	96.0 %	92.0 %
Age 12-20	100.0 %	98.3 %	96.3 %	93.3 %	88.3 %
Age 21-29	100.0 %	98.4 %	97.5 %	94.0 %	88.8 %
Age 30-41	100.0 %	99.4 %	98.8 %	95.9 %	90.9 %
Age 41-76	100.0 %	99.8 %	99.1 %	96.6 %	91.7 %

To determine the statistical significance of observed demographic variations, a *t*-test (Welch's unequal variance *t*-test) was used when comparing TPIR of two demographic groups and analysis of variance (ANOVA) was used when comparing TPIR of three or more groups.

At face-match thresholds of 6040 (overall TPIR of 99 %) and 6660 (overall TPIR of 98 %) the following observed demographic variations in TPIR are statistically significant at the 0.05 level:

- The TPIR for subjects of age 41-76 is higher than that for subjects of ages 12-20, 21-29, and 30-41).

Table 7: Statistical significance of demographic variation in TPIR at threshold 7727

Demographic		TPIR(180000, 1 *T* =7727)		Significance of variation	
		Mean for demographic	Standard error of mean	*p*-value	at significance level 0.05
All cohort subjects		95.0 %	0.5 %		
Gender	Female	93.6 %	0.8 %	$p = 0.008$	statistically significant
	Male	96.4 %	0.7 %		
Ethnicity	Asian	97.4 %	0.7 %	$p = 0.004$	statistically significant Asian v [White ∪ Black] ($p = 0.001$)
	Black	92.5 %	1.3 %		
	White	95.2 %	0.8 %		
Ethnicity & Gender	Asian/Female	96.7 %	1.1 %	$p = 0.002$	
	Asian/Male	98.3 %	0.9 %		
	Black/Female	90.2 %	1.8 %		statistically significant Black/Female v Black/Male ($p = 0.04$)
	Black/Male	95.3 %	1.6 %		
	White/Female	94.3 %	1.2 %		
	White/Male	96.0 %	1.1 %		
Age	12-20 years	93.3 %	1.3 %	$p = 0.092$	
	21-29 years	94.0 %	1.2 %		
	30-41 years	95.9 %	0.9 %		
	42-76 years	96.6 %	0.8 %		

Table 7 considers the demographic variation in TPIR at face-match threshold 7727 (overall TPIR of 95 %). The following demographic variations in TPIR are statistically significant at the 0.05 significance level.

- The TPIR for Male subjects (96.4 %) is higher than that for Female subjects (93.6 %).
- The TPIR for Asian subjects (97.4 %) is higher than that for White subjects (95.2 %) and Black subjects (92.5 %).
- The TPIR for Black Male subjects (95.3 %) is higher than that for Black Female subjects (90.2 %).

6.3 OIFR-STYLE IMAGES: DEMOGRAPHIC VARIATION IN FPIR, TPIR

The observed FPIR and TPIR for OIFR-style probe images is very similar to that for probe images from the whole initial portion of the Equitability Study Dataset (which includes OIFR-style images).

For OIFR-style probe images, Table 8 shows the demographic variation in FPIR at threshold 2010 1 and Table 9 shows the demographic variation in TPIR at threshold 7727.

Table 8: OIFR-style images: Demographic variation in FPIR at threshold 2010.

Demographic		FPIR(180000, T = 2010)		Significance of variation	
		Mean for demographic	Standard error of mean	p-value	at significance level 0.05
All cohort subjects		3.9 %	0.7 %		
Gender	Female	4.4 %	1.1 %	p = 0.41	
	Male	3.3 %	0.9 %		
Ethnicity	Asian	4.8 %	1.5 %	p = 0.47	
	Black	5.1 %	1.6 %		
	White	3.1 %	1.0 %		
Ethnicity & Gender	Asian/Female	3.7 %	1.5 %	p = 0.39	
	Asian/Male	6.1 %	2.7 %		
	Black/Female	7.3 %	2.8 %		
	Black/Male	2.5 %	1.4 %		
	White/Female	3.3 %	1.5 %		
	White/Male	2.9 %	1.2 %		
Age	12-20 years	1.8 %	0.9 %	p = 0.009	statistically significant Ages 21-29 v Other Ages (p = 0.021)
	21-29 years	8.0 %	2.3 %		
	30-41 years	3.3 %	1.1 %		
	42-76 years	2.5 %	1.0 %		

Table 9: OIFR-style images: Demographic variation in TPIR at threshold 7727

Demographic		TPIR(180000, 1, T = 7727)		Significance of variation	
		Mean for demographic	Standard error of mean	p-value	at significance level 0.05
All cohort subjects		95.3 %	0.8 %		
Gender	Female	93.5 %	1.2 %	p = 0.018	statistically significant
	Male	97.1 %	0.9 %		
Ethnicity	Asian	97.9 %	0.9 %	p = 0.027	statistically significant Asian v [White ∪ Black] (p = 0.001)
	Black	92.9 %	2.0 %		
	White	95.5 %	1.0 %		
Ethnicity & Gender	Asian/Female	96.5 %	1.5 %	p = 0.015	
	Asian/Male	99.4 %	0.6 %		
	Black/Female	89.3 %	3.2 %		
	Black/Male	95.6 %	2.4 %		
	White/Female	94.1 %	1.7 %		
	White/Male	96.8 %	1.2 %		
Age	12-20 years	96.5 %	1.3 %	p = 0.046	statistically significant Age 42-76 v Ages < 42 (p = 0.02)
	21-29 years	91.8 %	2.1 %		
	30-41 years	95.2 %	1.5 %		
	42-76 years	97.5 %	0.8 %		

For OIFR-style images:

- At face-match threshold 2010 the following demographic variations in FPIR are statistically significant at the 0.05 level:
    - The FPIR for subjects of Age 21-29 (8.0 %) is higher than that for the combined group of subjects of Age 12-20 (1.8 %), Age 30-41 (3.3 %), and Age 42-76 (2.5 %).
- At face-match thresholds 2120, 2212 and 2343 the observed demographic variations in FPIR were not found significant at the 0.05 level.
- At face-match thresholds between 2751 and 3981, observed FPIR = 0 % and observed TPIR = 100 %.
- At face-match thresholds 6040 and 6660 the observed demographic variations in TPIR were not found significant at the 0.05 level.
- At face match threshold 7727 the following demographic variations in FPIR are statistically significant at the 0.05 level.
    - The TPIR for Male subjects (97.1 %) is higher than that for Female subjects (93.5 %).
    - The TPIR for Asian subjects (97.9 %) is higher than that for White subjects (95.5 %) and Black subjects (92.9 %).
    - The TPIR for subjects of Age 42-76 (97.5 %) is higher than for subjects of other Ages below 42.

These findings are consistent with the statistically significant variations on the whole initial portion of the Equitability Study Dataset.

7 EQUITABILITY

In assessing equitability of FR for RFR, or OIFR we consider the potential impact of a false positive, or a missed recognition, on data subjects of an identification search.

Generally, a missed recognition has little impact on the subject in the probe image (the subject is treated as if they do not have a face image in the reference dataset) and has no impact on data subjects of the reference database.

In RFR, a false positive affects the individuals enrolled on the reference database who are added to a candidate list of an identification search, for example, being identified as a potential suspect in a police investigation.

In OIFR, the subjects of the probe images submitted for identification search, may be misidentified due to the false positive, leading to further action by the officer.

The test results show that a face-match threshold (e.g., 3500) can be set that will achieve correct recognition (TPIR = 100.0 %, FPIR = 0.0 %) for all test cases regardless of demographic. In this case the facial recognition performance can be considered equitable under the evaluation criteria for both RFR and OIFR.

In the case of RFR, if a face-match threshold were not applied or was set at a level where FPIR > 0, demographic variation in FPIR could result in some demographics being more likely to be selected for the candidate list than others. The evaluation data allows analysis of this effect. For each probe image identification search, count the number of candidates of each demographic category. Then take a weighted average over all the probe images. The weighting is needed to correct any imbalance in the number of probe images of each demographic because returned candidates tend to belong to the same demographic category as the probe image.

Table 10 and Table 11 show the average demographic composition of the top scoring candidate of non-mated identification searches, (i) when the demographic profile of probe images is uniform, and (ii) when the demographic profile of probe images follows that of the 2001 UK Census[5]. For Table 10, the identification searches applied no face-match threshold, and for Table 11 a face-match threshold of 2010 was used. The tables show that there is indeed an imbalance in the candidate demographics, and moreover there is strong dependence on the demographic profile of the probes being searched.

Table 10: Ethnicity/gender balance of rank 1 candidates (non-mated identification searches without threshold)

Ethnicity/gender profile of probe image set			Proportion of returned candidates for each Ethnicity/Gender					
			Asian Female	Asian Male	Black Female	Black Male	White Female	White Male
Uniform	AF	17 %	16 %	19 %	19 %	16 %	18 %	12 %
	AM	17 %						
	BF	17 %						
	BM	17 %						
	WF	17 %						
	WM	17 %						
Per UK Census	AF	5 %	10 %	11 %	5 %	5 %	40 %	29 %
	AM	5 %						
	BF	2 %						
	BM	2 %						
	WF	43 %						
	WM	43 %						

[5] (CENSUS 2021 [5]): White 81.7 %, Asian, 9.3 %, Black 4.0 %

Table 11: Ethnicity/gender balance of rank 1 candidates (non-mated identification searches with face-match threshold 2010)

Ethnicity/gender profile of probe image set			Probability a candidate returned	Proportion of returned candidates for each Ethnicity/Gender					
				Asian Female	Asian Male	Black Female	Black Male	White Female	White Male
Uniform	AF	17 %	0.05	16 %	22 %	29 %	12 %	13 %	8 %
	AM	17 %							
	BF	17 %							
	BM	17 %							
	WF	17 %							
	WM	17 %							
Per UK Census	AF	5 %	0.034	9 %	10 %	5 %	3 %	46 %	27 %
	AM	5 %							
	BF	2 %							
	BM	2 %							
	WF	43 %							
	WM	43 %							

Note that applying any face-match threshold above 3052, FPIR = 0 % even when TPIR < 100 %, and no candidates are returned from a non-mated identification search, i.e., the algorithm would be equitable in terms of false matches.

8 TERMINOLOGY AND ABBREVIATIONS

Candidate: Image of a person from reference database returned as result of an identification search.

Cohort: Subjects recruited to provide a corpus of facial images and video for recognition in the evaluation.

Comparison score: Numerical value of the similarity between compared probe and reference facial images.

Equitable: Demographic equitability of an operational deployment requires that any differences in data subject outcomes arising from subject demographics are inconsequential.

Face-match threshold: A comparison score threshold above which the compared images will be considered to match.

FR: Facial recognition

FPIR: False Positive Identification Rate is the proportion of non-mated identification searches which return a (false positive) match against a candidate on the reference database.

$$\text{FPIR}(N, T) = \frac{\text{Number of non-mated identification searches returning a candidate with score above threshold}}{\text{Number of non-mated identification searches}}$$

where N represents the number of images in the reference database, and T the face-match threshold.

Filler dataset: Dataset drawn from MPS holdings of custody images and used to supplement Cohort images and metadata.

Identification search: Biometric comparison of a probe image against a reference database to return a candidate list of the best matching reference images.

Mated: A mated identification search is one in which the subject in probe image also has a reference image in the reference database. Similarly, a mated comparison score is produced from comparisons of two face images of the same individual.

MPS: Metropolitan Police Service

Non-mated: A non-mated identification search is one in which the subject in probe image does not have a reference image in the reference database. Similarly, a non-mated comparison score is produced from comparison of face images of different individuals.

OIFR: Operator Initiated Facial Recognition

Probe image: A facial image that is searched against the reference database.

Rank: The rank of a candidate facial image is its position in the set of candidates returned by an identification search listed in decreasing order of similarity to the probe image (i.e., the rank 1 candidate is the best matching candidate)**.**

Reference image: A facial image in the reference database.

RFR: Retrospective Facial Recognition

SEL: Selectivity is the average number of candidates returned in a non-mated identification search for which the candidate comparison score exceeds the face-match threshold.

$$\mathrm{SEL}(N, R, T) = \frac{\text{Total number of candidates returned with score above threshold } T \text{ in the set of non-mated identification searches}}{\text{Number of non-mated identification searches}}$$

where N represents the number of images in the reference database, and where only candidates up to rank R with candidate score greater than the face-match threshold T are returned

TPIR: True Positive Identification Rate is the proportion of mated identification searches that include the mated reference among the candidates returned.

$$\mathrm{TPIR}(N, R, T) = \frac{\text{Number of mated identification searches where the mated reference is among the candidates returned}}{\text{Number of mated identification searches}}$$

where N represents the number of images in the reference database, R the number of best matching candidates returned and T the face-match threshold (T=0 if no threshold is applied).

9 BIBLIOGRAPHY

[1] ISO/IEC 19795-1:2021 Biometric performance testing and reporting - Part 1: Principles and framework

[2] ISO/IEC 19795-2: 2007 Biometric performance testing and reporting - Part 2: Testing methodologies for technology and scenario evaluation

[3] Facial recognition technology in law enforcement – Equitability study, NPL Report MS43, 2023

[4] The code systems used within the Metropolitan Police Service (MPS) to formally record ethnicity, 2007 http://policeauthority.org/metropolitan/publications/briefings/2007/0703/index.html

[5] Office for National Statistics, 'Ethnic group, England and Wales: Census 2021', https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/bulletins/ethnicgroupenglandandwales/census2021

[6] Gov.uk, 'Arrest Data March 2020 to March 2022', https://www.ethnicity-facts-figures.service.gov.uk/crime-justice-and-the-law/policing/number-of-arrests/latest/downloads/arrests-data-2020-to-2022.csv

[7] NPIA, 'Police Standard for Still Digital Image Capture and Data Interchange of Facial/Mugshot and Scar, Mark & Tattoo Images', National Policing Improvement Agency, 2007

[8] ISO/IEC 30137-1:2024 Use of biometrics in video surveillance systems - Part 1: System design and specification

[9] Welch B.L, The significance of the difference between two means when the population variances are unequal. Biometrika, Vol. 9, No. 3/4 pp.350-362, 1938