National Physical Laboratory

REPORT 4153

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

FACIAL RECOGNITION TECHNOLOGIES: ACCURACY AND EQUITABILITY EVALUATION OF COGNITEC FACEVACS-DBSCAN ID V5.5

OCTOBER 2025

Facial Recognition Technologies: Accuracy and Equitability Evaluation of Cognitec FaceVACS-DBScan ID v5.5

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 objectives of the evaluation described in this report are 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 Cognitec FaceVACS-DBScan ID v5.5. This is the FR software currently used in the Police National Database (PND) for which the operational use case is Retrospective Facial Recognition (RFR).

NPL was commissioned by the Home Office in collaboration with the Office of the Policing Chief Scientific Adviser. This assessment will add to 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 case, 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 on 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 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. 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 RFR 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 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 nearly perfect (i.e. a TPIR of nearly 100 %). Of the 3513 mated identification searches the top match returned by the algorithm was the correct reference in 3510 cases. Of the three exceptions, two probes were recognised as the second-best match. The algorithm could not extract features needed for comparison (referred to as a failure-to-extract) from the remaining probe image.

- TPIR(180000, 1, 0) = 99.91 %
- TPIR(180000, 2, 0) = 99.97 %

At a face-match threshold of 0.9 there were no false positives, and TPIR exceeded 90 %.

- TPIR(180000, 1, 0.9) = 91.9 %
- FPIR(180000, 0.9) = 0.0 %

Further details are provided in Section 5.2.

1.2.2 Demographic variation in TPIR and FPIR

Demographic variation in performance was analysed for the test. Of the observed variations in TPIR and FPIR, the following variations were found to be statistically significant at the 0.05 significance level.

At a face-match threshold of 0.9:

- The TPIR for Asian subjects (98 %) is higher than that for White subjects (91 %) which in turn is higher than that for Black subjects (87 %).

At a face-match threshold of 0.8:

- The FPIR for Male subjects (1.5 %) is lower than that for Female subjects (3.8 %).
- The FPIR for White subjects (0.04 %) is lower than that for Asian subjects (4.0 %) and Black subjects (5.5 %).
- The FPIR for Black Male subjects (0.4 %) is lower than that for Black Female subjects (9.9 %).
- The FPIR for subjects of age 42-76 (0.1 %) is lower than that for subjects of age 12-20 (3.3 %), age 21-29 (5.2 %), and age 30-41 (2.2 %).

Further details are provided in Section 6.

1.2.3 Equitability

In Retrospective Facial Recognition, the demographic variations in performance do not impact on the unknown subject in the probe image as they are not directly affected by a false match. There is, however, a potential impact for data subjects whose face images are in the reference database, as some demographics may be more prone to inclusion in the candidate list of an identification search. The effect is larger when a threshold is applied to the candidate list.

Note that the demographics of subjects in the set of probe images for RFR also impacts the demographics of candidate lists. In Section 7 we examine the effect of the two sources of bias, using a hypothetical demographic profile of probe image subjects more representative of operational use cases than an even balance.

2 BACKGROUND

2.1 RETROSPECTIVE FACIAL RECOGNITION

The operational use case of the evaluation is 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. The candidate images returned by an identification search are adjudicated typically by more than one officer, e.g., the RFR officer and an investigating 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 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.

2.2.1 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.

2.2.2 Gender[1]

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

[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.)

2.2.3 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.

2.3 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.

The system is considered equitable if:

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

b) The system is considered equitable if the variation in performance between demographic groups is not statistically significant.

c) The system is considered equitable if 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.).

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.

2.4 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 representative of those used for Operator Initiated Facial Recognition 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 wearing of sunglasses, and sunhats was not prohibited.

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.				

3.3 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

[3] Selfie image not collected for one subject.

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 180,000 images.

Processing of identification searches
The FaceVACS-DBScan ID system provides for batch processing of identification searches without requiring operator involvement. This capability was used for all test identification searches.

The system also provides the capability to manually annotate eyes positions in any face image and then run or re-run the identification search. This capability is useful in operational deployments and was used for 28 probe images where the initial identification search outcomes seemed anomalous (e.g., failure-to-extract cases with an ‘eyes not found’ error, and cases where the presence of a small additional face in the background caused a false recognition, requiring checking for potential mislabelling of test data).

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 algorithms 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 179916 references for mated identification searches, and 179915 for non-mated identification searches (which exclude the mated reference).

5 PERFORMANCE ANALYSES

5.1 MATED IDENTIFICATION SEARCHES WITHOUT THRESHOLDING

The outcomes of the conducted mated identification searches are summarised in Table 4. Possible outcomes are:

Rank 1: the algorithm returns the correct mated reference as the best (rank 1) match.
Rank 2-200: the mated reference is found at rank between 2 and 200. (In the evaluation identification searches returned the top 200 candidates.)
No match: the mated reference was not returned in the top 200 matches.
Fail to extract: the algorithm could not process the probe image, e.g., failed to extract features for comparison.

Table 4: Recognition outcomes for mated identification searches

Face image type	Rank 1	Rank 2-200	No match	Fail to extract	TOTAL
Custody-style	283	0	0	0	283
OIFR-style	1133	2	0	0	1135
Adhoc	1080	0	0	0	1080
Selfie	1014	0	0	1	1015

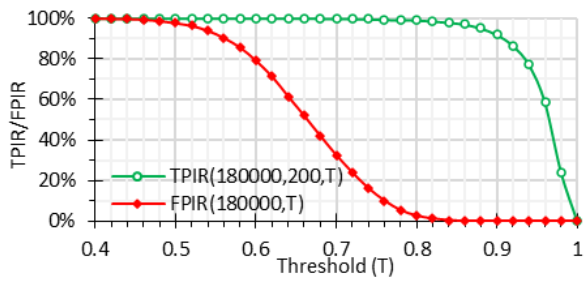
Combining the facial image types, 3510 mated identification searches returned the correct reference at rank 1. Of the remaining 3 cases: for one White Female probe image and for one Black Female probe image, the correct reference was returned at rank 2; and the algorithm failed to extract features for one Black Male probe image. It follows that for any demographic subset of the evaluation dataset, the observed recognition accuracy is almost identical.

- TPIR(180000, 1, 0) = 99.91 %
- TPIR(180000, 2, 0) = 99.97 %

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.

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[5]. The figure might be used to select a threshold to optimise the trade-off between a high TPIR and a low FPIR for face images of the type used in the evaluation. To achieve the lowest FPIR values it is necessary to sacrifice TPIR.



Face-match threshold	TPIR	FPIR
0.90	91.9 %	0.0 %
0.85	97.4 %	0.2 %
0.80	99.0 %	2.7 %
0.75	99.5 %	13.1 %
0.70	99.8 %	32.6 %

Figure 1: TPIR & FPIR by threshold

[5] Note that comparison scores are dependent on the algorithm, and face-match thresholds for differing algorithms are not directly comparable.

6 VARIATION IN PERFORMANCE BETWEEN DEMOGRAPHICS

Table 5 shows the observed demographic variation in TPIR and FPIR over a range of face-match thresholds from 0.7 to 0.9. The extent of variation is different at different face-match thresholds. To analyse the statistical significance of any demographic variation, thresholds should be chosen: (i) which are within the range of operational deployment, and (ii) where the variation is large enough to be operationally relevant.

For TPIR we choose threshold 0.9, a threshold at which there were no false positives, but some variation in TPIR. For FPIR we choose threshold 0.8, a threshold at which there is observable variation between the demographic categories. (At threshold 0.85 the observed demographic variation in FPIR is too small to be assessed as statistically significant.)

Table 5: TPIR and FPIR by threshold setting and demographic

	TPIR(180000, 2, *T*)				
Demographic	T = 0.9	T = 0.85	T = 0.8	T = 0.75	T = 0.7
All	91.9 %	97.4 %	99.0 %	99.5 %	99.8 %
Female	92.9 %	97.6 %	98.9 %	99.4 %	99.7 %
Male	90.8 %	97.3 %	99.0 %	99.6 %	99.9 %
Asian	97.9 %	99.2 %	99.4 %	100.0 %	100.0 %
Black	86.9 %	96.8 %	99.2 %	99.5 %	99.8 %
White	91.3 %	96.9 %	98.7 %	99.1 %	99.7 %
Asian/Female	96.9 %	98.8 %	99.2 %	100.0 %	100.0 %
Asian/Male	99.0 %	99.6 %	99.6 %	100.0 %	100.0 %
Black/Female	90.1 %	98.0 %	99.4 %	99.5 %	99.7 %
Black/Male	83.0 %	95.3 %	98.9 %	99.5 %	99.9 %
White/Female	92.1 %	96.8 %	98.6 %	98.9 %	99.5 %
White/Male	90.5 %	97.0 %	98.8 %	99.3 %	99.8 %
Age 12-20	90.1 %	95.6 %	97.9 %	98.9 %	99.5 %
Age 21-29	92.1 %	97.2 %	99.3 %	99.4 %	99.7 %
Age 30-41	93.9 %	99.1 %	99.4 %	99.8 %	100.0 %
Age 41-76	91.4 %	97.8 %	99.3 %	99.7 %	100.0 %

	FPIR(180000, *T*)				
Demographic	T = 0.9	T = 0.85	T = 0.8	T = 0.75	T = 0.7
All	0.0 %	0.2 %	2.7 %	13.1 %	32.6 %
Female	0.0 %	0.4 %	3.8 %	18.7 %	41.7 %
Male	0.0 %	0.1 %	1.5 %	7.3 %	23.2 %
Asian	0.0 %	0.2 %	4.0 %	18.6 %	43.7 %
Black	0.0 %	0.6 %	5.5 %	23.7 %	51.0 %
White	0.0 %	0.0 %	0.04%	2.4 %	11.6 %
Asian/Female	0.0 %	0.0 %	3.0 %	20.9 %	48.0 %
Asian/Male	0.0 %	0.4 %	5.3 %	15.8 %	38.6 %
Black/Female	0.0 %	1.2 %	9.9 %	38.5 %	71.6 %
Black/Male	0.0 %	0.0 %	0.4 %	6.2 %	26.8 %
White/Female	0.0 %	0.0 %	0.1 %	3.2 %	14.9 %
White/Male	0.0 %	0.0 %	0.0 %	1.6 %	8.4 %
Age 12-20	0.0 %	0.3 %	3.3 %	13.5 %	33.8 %
Age 21-29	0.0 %	0.7 %	5.2 %	20.0 %	44.5 %
Age 30-41	0.0 %	0.0 %	2.2 %	15.2 %	36.9 %
Age 41-76	0.0 %	0.0 %	0.1 %	3.9 %	15.7 %

Table 6: Statistical significance of demographic variation in TPIR

Demographic		TPIR(180000, 2, 0.9)		Significance of variation	
		Mean for demographic	Standard error of mean	*p*-value	at significance level 0.05
All cohort subjects		91.9 %	0.7 %		
Gender	Female Male	92.9 % 90.8 %	0.9 % 1.1 %	$p = 0.146$	
Ethnicity	Asian Black White	97.9 % 86.9 % 91.3 %	0.6 % 1.8 % 1.0 %	$p < 0.001$	statistically significant Asian v White ($p < 0.001$) White v Black ($p = 0.033$) Asian v Black ($p < 0.001$)
Ethnicity/ Gender	Asian/Female Asian/Male	96.9 % 99.0 %	1.0 % 0.5 %	$p = 0.062$	
	Black/Female Black/Male	90.1 % 83.0 %	2.1 % 2.9 %	$p = 0.051$	
	White/Female White/Male	92.1 % 90.5 %	1.3 % 1.5 %	$p = 0.421$	
Age	12-20 years 21-29 years 30-41 years 42-76 years	90.1 % 92.1 % 93.9 % 91.4 %	1.5 % 1.5 % 1.3 % 1.5 %	$p = 0.281$	

Table 6 shows the variation in TPIR between genders, ethnicities, and ages at a face-match threshold of 0.9.

To determine the statistical significance of observed demographic variations, a *t*-test (Welch's unequal variance *t*-test [9]) 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 a face-match threshold of 0.9, the following demographic variations in TPIR are statistically significant at the 0.05 significance level.

- The TPIR for Asian subjects (97.9 %) is higher than that for White subjects (91.3 %) which in turn is higher than that for Black subjects (86.9 %).

Table 7: Statistical significance of demographic variation in FPIR

Demographic		FPIR(180000, 0.8)		Significance of variation	
		Mean for demographic	Standard error of mean	*p*-value	at significance level 0.05
All cohort subjects		2.7 %	0.5 %		
Gender	Female Male	3.8 % 1.5 %	0.9 % 0.5 %	$p = 0.025$	statistically significant
Ethnicity	Asian Black White	4.0 % 5.5 % 0.04 %	1.1 % 1.5 % 0.04 %	$p < 0.001$	statistically significant White v [Asian ∪ Black] ($p < 0.001$)
Ethnicity & Gender	Asian/Female Asian/Male	3.0 % 5.3 %	0.9 % 2.2 %	$p = 0.339$	
	Black/Female Black/Male	9.9 % 0.4 %	2.6 % 0.3 %	$p < 0.001$	statistically significant
	White/Female White/Male	0.1 % 0.0 %	0.1 % 0.0 %	$p = 0.320$	
Age	12-20 years 21-29 years 30-41 years 42-76 years	3.3 % 5.2 % 2.2 % 0.1 %	1.2 % 1.5 % 0.6 % 0.1 %	$p = 0.005$	statistically significant Ages ≥ 42 v Ages < 42 ($p < 0.001$)

Table 7 shows the variation in FPIR between genders, ethnicities, and ages at a face-match threshold of 0.8.

To determine the statistical significance of observed demographic variations, a *t*-test (Welch's unequal variance *t*-test) 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 a face-match threshold of 0.8, the following demographic variations in FPIR are statistically significant at the 0.05 significance level.

- The FPIR for Male subjects (1.5 %) is lower than that for Female subjects (3.8 %).
- The FPIR for White subjects (0.04 %) is lower than that for Asian subjects (4.0 %) and Black subjects (5.5 %). (The difference in FPIR between Asian and Black subjects is not significant.)
- The FPIR for Black Male subjects (0.4 %) is lower than that for Black Female subjects (9.9 %).
- The FPIR for subjects of age 42-76 (0.1 %) is lower than that for subjects of age 12-20 (3.3 %), age 21-29 (5.2 %), and age 30-41 (2.2 %). (The differences in FPIR between subjects of age 12-20, of age 21-29, and of age 30-41 are not significant.)

7 EQUITABILITY

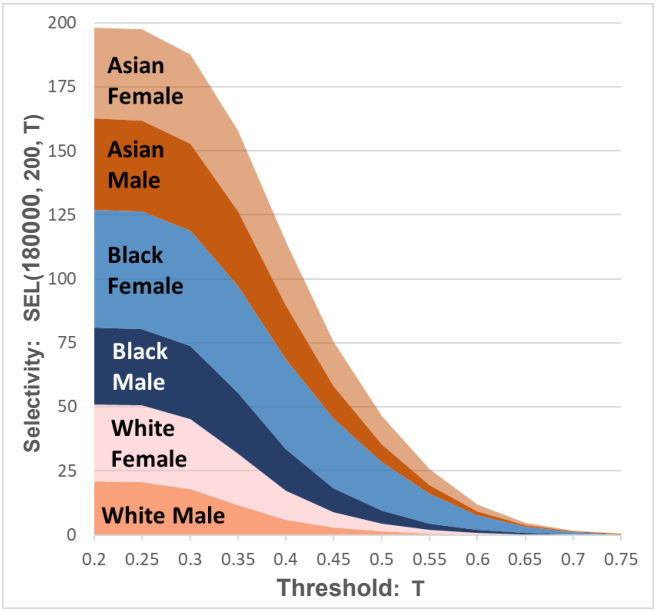
In the case of RFR, the data subjects that may be affected by algorithm performance are the individuals enrolled onto the reference database who may be added to a candidate list of an identification search as a result of a false positive. (A consequence, for example, could be being identified as a potential suspect in a police investigation.)

Given the demographic variation in performance shown in Section 6, are some demographics more likely to be selected for the candidate list than others?

The evaluation data allows analysis of this question. 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. Most, but not all candidates will belong to the same demographic category as the probe image. The weighting is needed to correct any imbalance in the number of probe images of each demographic.

Figure 2 plots, as a function of face-match threshold, the average number of candidates of each demographic returned in the non-mated identification searches. The figure does show an imbalance in the number of returned candidates of each demographic. For example, at threshold 0.45 the Black Female candidate layer is thicker than the White Male layer, i.e., there were more Black Female candidates than White Male candidates. This imbalance grows as the face-match threshold increases.

Figure 2: Selectivity plot – ethnicity/gender of candidates returned in non-mated identification searches



Note that though applying a threshold increases the relative imbalance between the demographics, it nevertheless reduces the frequency of a false positive identification for all demographics.

If the ethnicity/gender balance of the probe image set was not uniform, but instead followed the national demographics for England and Wales[6], or Policing arrest numbers for 2021/22[7], we can reweight probes accordingly, and the average demographic balance of the returned rank 1 candidates would change as shown in Table 8 (no threshold) and Table 9 (with face-match threshold 0.7).

The examples show that demographics of the probe image set can have a larger effect on equitability than algorithmic bias.

Table 8: Ethnicity/gender balance of rank 1 candidates (non-mated identification searches without threshold)

Probe image set: Ethnicity/gender profile	Asian Female	Asian Male	Black Female	Black Male	White Female	White Male
Uniform	17 %	19 %	20 %	15 %	17 %	13 %
Per UK Census	11 %	12 %	5 %	5 %	36 %	30 %
Per Policing arrest numbers	4 %	19%	4 %	10 %	16 %	47 %

Table 9: Ethnicity/gender balance of rank 1 candidates (non-mated identification searches with face-match threshold 0.7)

Probe image set: Ethnicity/gender profile	Probability of non-empty candidate list	Asian Female	Asian Male	Black Female	Black Male	White Female	White Male
Uniform	0.34	20 %	19 %	37 %	12 %	8 %	4 %
Per UK Census	0.15	17 %	15 %	15 %	5 %	33 %	16 %
Per Policing arrest numbers	0.13	4 %	27 %	9 %	16 %	14 %	29 %

Similarly, Table 10 shows the age balance of rank 1 non-mated candidates assuming uniform probe demographic and numbers in each age category (12-20, 21-39, 30-41, 42-76) of the evaluation.

Table 10: Age balance of rank 1 non-mated candidates

	Probability of a non-empty candidate list	Age 12-20	Age 21-29	Age 30-41	Age 42-76
Identification searches without threshold	1.0	26 %	29 %	24 %	21 %
Identification searches with threshold 0.7	0.3	33 %	35 %	23 %	10 %

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

[7] (Arrest data March 2020 to March 2022 [6]) Asian Female 3582, Asian Male 38384, Black Female 5371, Black Male 45536, White Female 70589, White Male 367019

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.

PND: Police National Database

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