Facial recognition technologies: Accuracy and Equitability Evaluation of Cognitec FaceVACS-DBScan ID v5.5

S.40(2) National Physical Laboratory

Draft 4

August 2024

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 successful 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 system in an operational setting in terms of accuracy and equitability related to subject demographics. This 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.

This work was commissioned by the Home Office in collaboration with the Office of the Policing Chief Scientific Adviser.

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). The operational use cases considered in the evaluation are Retrospective Facial Recognition and Operator Initiated Facial Recognition.

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 different types of image in 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:

- In the remainder of this section, key findings of the evaluation are summarised.
- 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 difference in performance between demographic categories.
- Section 7 addresses equitability of the FR algorithms on operational use cases.
- Section 8 provides further discussion on evaluation findings, and on operational implications and considerations.
- Section 9 provides a glossary of the key terms and abbreviations used in this report.

1.2 Key findings: Retrospective Facial Recognition

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. The identification produces a candidate list of the reference images that best match the submitted probe 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.

Recognition accuracy for RFR is measured in terms of:

- True Positive Identification Rate: TPIR(*N*, *R*, *T*), the proportion of 'mated' identification searches (i.e., where the person in the probe image has a reference image in the reference database) 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.
- False Positive Identification Rate: FPIR(*N*, *T*), the proportion of 'non-mated' identification searches (i.e., where the person in the probe image does not have a reference image in the reference database) 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 of Custody, OIFR, Adhoc camera and Selfie style images

The initial portion of the Equitability Study Dataset (the data used in the 2022 evaluation) comprises Custody style images, ad-hoc digital camera images, and OIFR style and Selfie images taken on mobile phones.

On this image set, with no thresholding applied, identification performance was nearly perfect. For all except three probe images, the top match returned by the algorithm was the correct reference. 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.

- $TPIR_{Custody + Oifr + Adhoc + Selfie}(180000, 1, 0) = 99.9\ \%$
- $TPIR_{Custody + Oifr + Adhoc + Selfie}(180000, 2, 0) = 99.97\ \%$

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

- $TPIR_{Custody + Oifr + Adhoc + Selfie}(180000, 1, 0.9) = 91.9\ \%$
  $FPIR_{Custody + Oifr + Adhoc + Selfie}(180000, 0.9) = 0\ \%$

Further details are provided in Section 5.2.1.

S.31(1) [redacted]

S.31(1)

[redacted]

1.2.5 RFR Demographic variation in TPIR and FPIR

Demographic variation in performance was analysed for the test results for the initial portion of the Equitability Study Dataset. Of the observed variations in TPIR and FPIR, the following variations were found to be statistically significant at the .05 significance level.

At a face-match threshold of 0.9

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

At a face-match threshold of 0.8

- The FPIR for Female subjects (3.8 %) is higher than that for Male subjects (1.5 %).

- 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 Females (9.9 %) is higher than that for Black Males (0.4 %).
- 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 %).

1.2.6 RFR Equitability

In RFR, 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, an impact for data subjects with face images in the reference database, causing some demographics to be more likely to be included in the candidate list of an identification search. The effect is larger when a threshold is applied to the candidate list.

The demographics of subjects in the set of probe images for RFR also impact the demographics of candidate lists. In Section 7.1 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.

1.3 Key findings: Operator Initiated Facial Recognition

Operator Initiated Facial Recognition (OIFR) is a near-real-time use of facial recognition technology. A photograph of a person's face is taken on an officer's mobile phone and is submitted for immediate search against a reference image database. The identification search produces a candidate list of reference images that best match the submitted probe image, and this candidate list is returned to the officer's phone to assist in identifying the subject.

As for RFR, recognition accuracy for OIFR is also measured in terms of TPIR and FPIR.

1.3.1 Recognition accuracy for OIFR

For OIFR the evaluation used only the initial portion of the Equitability Study Dataset (the dataset used in the 2022 evaluation). The other images are not representative of the types of images taken by the officers' OIFR mobile device.

The findings on recognition accuracy are the same as for RFR as described in Section 1.2.1. Identification performance was nearly perfect. All but three probe images were correctly recognised as the top match by the algorithm, and a further two probe images were recognised at rank 2. The algorithm could not extract features needed for comparison from the remaining probe image..

- $TPIR_{Custody + Oifr + Adhoc + Selfie}(180000, 1, 0) = 99.9\ \%$
- $TPIR_{Custody + Oifr + Adhoc + Selfie}(180000, 2, 0) = 99.97\ \%$

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

- $TPIR_{Custody + Oifr + Adhoc + Selfie}(180000, 1, 0.9) = 91.9\ \%$
  $FPIR_{Custody + Oifr + Adhoc + Selfie}(180000, 0.9) = 0\ \%$

There was no demographic variation in FPIR at threshold 0.9.

TPIR at threshold 0.9 varied between ethnicity groups (Asian: 98 %, White: 91 %, Black: 87 %) which is statistically significant at the .05 significance level.

1.3.2 OIFR Demographic variation in TPIR and FPIR.

The findings on demographic variation in TPIR and FPIR given in section 1.2.5 for RFR also apply to OIFR.

Assuming a threshold of 0.9, which gives good TPIR and FPIR performance, the relevant significant demographic variation is that the TPIR for Asian subjects (98 %) is higher than for White subjects (91 %) which in turn is higher than for Black subjects (87 %).

1.3.3 OIFR Equitability

If the OIFR system is operating without threshold, it would be considered equitable under our evaluation criteria (no statistically significant demographic variation in performance).

If the OIFR system is operating with threshold 0.9, we must consider the effect of the differences in TPIR between Asian, Black, and White subjects. From the subject's perspective the outcome is the same as if the subject did not have a record in the reference database, and in most cases is unlikely to be of consequence to the individual.

We should also note that the demographic composition of the reference database influences demographic variation in TPIR. In operational cases the reference database is unlikely to have the same numbers in each ethnicity/gender category. This situation is discussed further in Section 7.2.

2 Background

2.1 Operational use cases

2.1.1 Retrospective Facial Recognition

Retrospective Facial Recognition 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 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, with some images being of very low quality. The range of mated and non-mated comparison scores can change with image quality and selecting an appropriate face-match threshold can be difficult.

A false positive identification does not directly affect the unidentified subject in the submitted image. 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.

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

2.1.2 Operator Initiated Facial Recognition

Operator Initiated Facial Recognition 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 dataset can affect demographic variation in true positive and false positive recognition rates.

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 as:

- 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.

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 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:

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.

[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.)

2.4 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 .05 (5%). The significance level relates to the probability of falsely rejecting the 'null' hypothesis that there is 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 5% significance level can increase the probability of rejecting the null hypothesis to a value much higher than 5%. Methods are available that address this issue requiring stricter significance thresholds for each individual test; however, use of these methods increases the chance of failing to detect a demographic effect. Such methods were NOT applied 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 IDs 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 based on ONS 5+1 codes [5]	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]. Image of the full head with all hair, neck, shoulders, and ears. Subject facing square to the camera, looking directly at camera. 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 taken with Canon EOS850D camera. In cases of non-conformance a further image was taken, and the nonconformant image added to the set of outtakes. The custody images were used for enrolling reference images of the cohort, and as probe images for non-mated identification searches.

S.31(1) [redacted]

OIFR-style facial images taken on Samsung XcoverPro S.31(1) [redacted] Images 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.

S.31(1) [redacted]

Selfie photos were taken by the cohort subject themselves using a Motorola G60s mobile phone. Subjects were allowed to pose as they wished, a neutral expression was not required, and

subject did not need to pose 'square to the camera'. Selfies taken both indoors and outdoors.
S.31(1)

Ad-hoc digital camera photographs of subjects taken with a digital camera (Nikon D40, Nikon 1J5). Uncontrolled conditions and background, no flash or supplementary lighting, taken indoors and outdoors. Photos are similar to the OIFR images though the depth of field is shallower.
S.31(1)

S.31(1)

The first four image types, (i.e., Custody, Oifr, Adhoc and Selfie), form the 'initial portion' of the images of Equitability Study Dataset. These images were used for the 2022 evaluation [3].

S.31(1)

Table 2: Face image quality of Cohort image types

	Custody	OIFR	Adhoc	Selfie	S.31(1)
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, thick rims, hats,	Y				
Face in focus, no blur	Y	Y	Y	Y	
Number of images ...	684	1135	1080	1015	
... from number of subjects	401	401	401	400	
Under-represented demographic groups					

Note:
Y indicates the quality aspect was mandatory for photographing the image type.
N indicates that the condition was not met for all or almost all images of the given type.
Blanks indicate that the condition was not required, or 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 based on ONS 5+1 codes	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[2]	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.

[2] 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.

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 29 images of the OIFR/Adhoc/Selfie categories 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). The analysis of performance includes the updated results for these cases. (Manual annotation was also tested on some images from the lower quality categories, but far fewer cases were improvable, and the much greater number of cases needing manual processing was not practical 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 running of RFR/OIFR 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.

Reporting size of reference image database

In this report we will write TPIR(180000, *R*, *T*) and FPIR(180000, *T*), though the number of images in the reference dataset is just under 180000. The reference dataset comprises enrolments of 179515 Filler images and 401 Cohort custody-style images, giving 179916

faces for mated identification searches, and 179915 for non-mated identification searches (after removing 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 2-6: the mated reference is found at rank between 2 and 6 (OIFR returns up to 6 candidates).

Rank 7-200: the mated reference is found at rank between 7 and 200 (RFR returns up to 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 2-6	Rank 7-200	No match	Fail to extract	TOTAL
Custody	283	0	0	0	0	283
Oifr	1133	2	0	0	0	1135
Adhoc	1080	0	0	0	0	1080
Selfie	1014	0	0	0	1	1015
S.31(1)	[redacted]					

For the combined set of Custody, Oifr, Adhoc and Selfie type images, all but only three identification searches returned the correct reference rank 1; for one White Female probe image and 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 Custody, Oifr, Adhoc and Selfie type images, the recognition accuracy is almost identical.

- $TPIR_{Custody + Oifr + Adhoc + Selfie}(180000, 1, 0) = 99.9\ \%$
- $TPIR_{Custody + Oifr + Adhoc + Selfie}(180000, 2, 0) = 99.97\ \%$

S.31(1) [redacted]

FR Equitability Testing
FaceVACS-DBScan ID v5.5

S.31(1) [redacted]

S.31(1) [redacted]

5.2 Mated and non-mated identification searches with face-match threshold

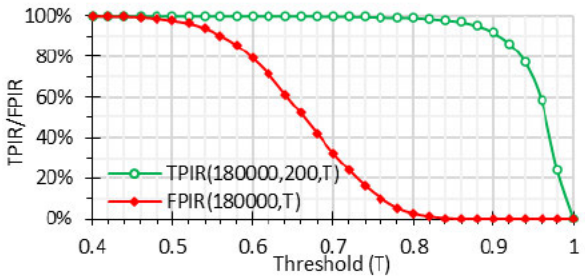
The unthresholded True Positive Identification Rate TPIR(N, R, 0) informs on the recognition performance for mated identification searches. If an unmated identification search is performed, unless there is a non-zero threshold for candidate comparison scores, all the candidates returned will be false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

5.2.1 Custody/Oifr/Adhoc/Selfie probe images

The combined Custody/Oifr/Adhoc/Selfie image data from the Equitability Study Dataset is the portion of the dataset previously used in evaluation of the NEC NeoFace algorithm in 2022 for RFR. Photographic image quality and subject pose are generally good.

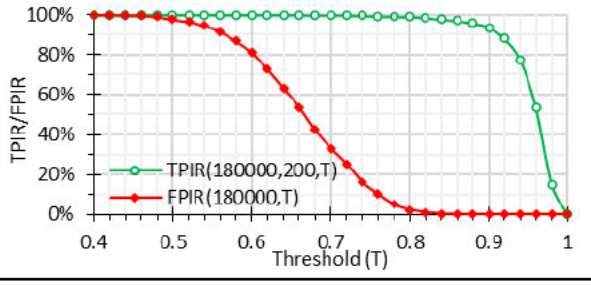
Figure 2 shows how the TPIR (for mated identification searches) and FPIR (for non-mated identification searches) varied as a function of the face-match threshold[3]. Figure 3 shows TPIR and FPIR for just the Oifr-type images, i.e., those face images taken on the mobile phones of the model used by South Wales Police for operational OIFR. The TPIR and FPIR values for these two cases are very close, and the four probe image types are grouped together for analysis of demographic differentials in performance, and for estimation of OIFR performance.

Figure 2: TPIR & FPIR by threshold – Custody/Oifr/Adhoc/Selfie face images



Face-match threshold	TPIR	FPIR
0.90	91.9 %	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 3: TPIR & FPIR by threshold – Oifr images

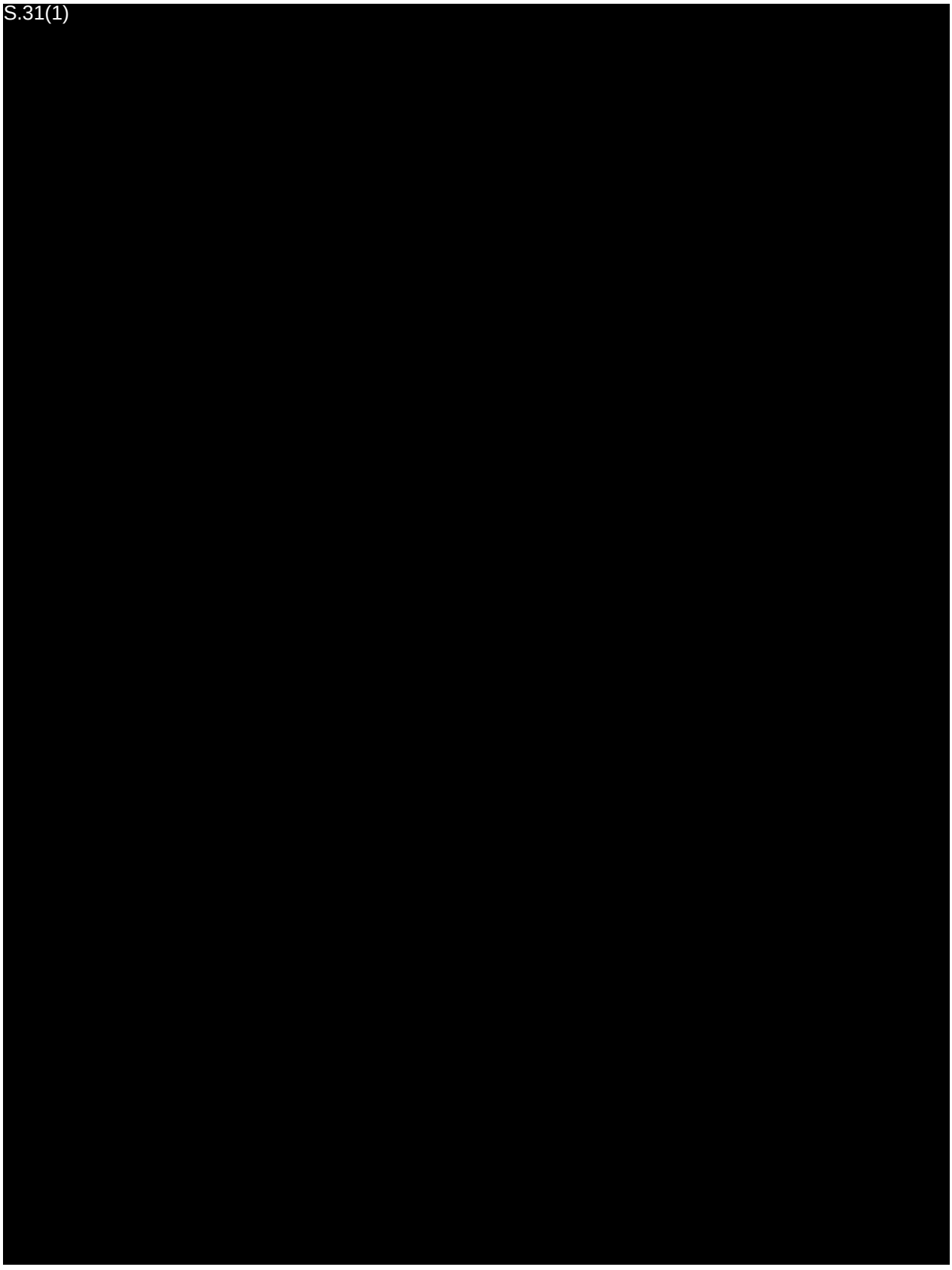


Face-match threshold	TPIR	FPIR
0.90	93.7 %	0 %
0.85	97.3 %	0.4 %
0.80	99.0 %	2.4 %
0.75	99.5 %	13.1 %
0.70	99.8 %	33.2 %

We observe that, for this portion of the test corpus, at a face-match threshold of 0.9, the recognition rate exceeds 90 % with no false positives.

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

S.31(1)



6 Variation in performance between demographics

Analysis of demographic variation in TPIR and FPIR has been conducted using results for the Custody, Oifr, Adhoc, Selfie images from the Equitability Study Dataset.

Table 5 shows TPIR and FPIR for this subset of face images at face-match thresholds ranging from 0.7 to 0.9 for and for differing demographic categories. 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 – Custody/Oifr/Adhoc/Selfie face images

	TPIR(180000, 2, *T*)					**FPIR**(180000, *T*)				
Demographic	**T = 0.9**	**T = 0.85**	**T = 0.8**	**T = 0.75**	**T = 0.7**	**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 %	0.0 %	0.2 %	2.7 %	13.1 %	32.6 %
Female	92.9 %	97.6 %	98.9 %	99.4 %	99.7 %	0.0 %	0.4 %	3.8 %	18.7 %	41.7 %
Male	90.8 %	97.3 %	99.0 %	99.6 %	99.9 %	0.0 %	0.1 %	1.5 %	7.3 %	23.2 %
Asian	97.9 %	99.2 %	99.4 %	100.0 %	100.0 %	0.0 %	0.2 %	4.0 %	18.6 %	43.7 %
Black	86.9 %	96.8 %	99.2 %	99.5 %	99.8 %	0.0 %	0.6 %	5.5 %	23.7 %	51.0 %
White	91.3 %	96.9 %	98.7 %	99.1 %	99.7 %	0.0 %	0.0 %	0.04 %	2.4 %	11.6 %
Asian/Female	96.9 %	98.8 %	99.2 %	100.0 %	100.0 %	0.0 %	0.0 %	3.0 %	20.9 %	48.0 %
Asian/Male	99.0 %	99.6 %	99.6 %	100.0 %	100.0 %	0.0 %	0.4 %	5.3 %	15.8 %	38.6 %
Black/Female	90.1 %	98.0 %	99.4 %	99.5 %	99.7 %	0.0 %	1.2 %	9.9 %	38.5 %	71.6 %
Black/Male	83.0 %	95.3 %	98.9 %	99.5 %	99.9 %	0.0 %	0.0 %	0.4 %	6.2 %	26.8 %
White/Female	92.1 %	96.8 %	98.6 %	98.9 %	99.5 %	0.0 %	0.0 %	0.1 %	3.2 %	14.9 %
White/Male	90.5 %	97.0 %	98.8 %	99.3 %	99.8 %	0.0 %	0.0 %	0.0 %	1.6 %	8.4 %
Age 12-20	90.1 %	95.6 %	97.9 %	98.9 %	99.5 %	0.0 %	0.3 %	3.3 %	13.5 %	33.8 %
Age 21-29	92.1 %	97.2 %	99.3 %	99.4 %	99.7 %	0.0 %	0.7 %	5.2 %	20.0 %	44.5 %
Age 30-41	93.9 %	99.1 %	99.4 %	99.8 %	100.0 %	0.0 %	0.0 %	2.2 %	15.2 %	36.9 %
Age 41-76	91.4 %	97.8 %	99.3 %	99.7 %	100.0 %	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 .05*
All cohort subjects		91.9 %	0.7 %		
Gender	Female	92.9 %	1.0 %	$p = .146$	
	Male	90.8 %	1.0 %		
Ethnicity	Asian	97.9 %	1.4 %	$p < .001$ (A v W) $p = .033$ (B v W)	*statistically significant*
	Black	86.9 %	1.4 %		
	White	91.3 %	1.1 %		
Ethnicity & Gender	Asian/Female	96.9 %	2.0 %	$p = .062$	
	Asian/Male	99.0 %	2.1 %		
	Black/Female	90.1 %	1.8 %	$p = .051$	
	Black/Male	83.0 %	2.0 %		
	White/Female	92.1 %	1.6 %	$p = .421$	
	White/Male	90.5 %	1.5 %		
Age	12-20 years	90.1 %	1.4 %	$p = .075$	
	21-29 years	92.1 %	1.5 %		
	30-41 years	93.9 %	1.4 %		
	42-76 years	91.4 %	1.4 %		

Demographic	T = 0.9	T = 0.85	T = 0.8	T = 0.75	T = 0.7	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 %	0.0 %	0.2 %	2.7 %	13.1 %	32.6 %
Female	92.9 %	97.6 %	98.9 %	99.4 %	99.7 %	0.0 %	0.4 %	3.8 %	18.7 %	41.7 %
Male	90.8 %	97.3 %	99.0 %	99.6 %	99.9 %	0.0 %	0.1 %	1.5 %	7.3 %	23.2 %
Asian	97.9 %	99.2 %	99.4 %	100.0 %	100.0 %	0.0 %	0.2 %	4.0 %	18.6 %	43.7 %
Black	86.9 %	96.8 %	99.2 %	99.5 %	99.8 %	0.0 %	0.6 %	5.5 %	23.7 %	51.0 %
White	91.3 %	96.9 %	98.7 %	99.1 %	99.7 %	0.0 %	0.0 %	0.04 %	2.4 %	11.6 %
Asian/Female	96.9 %	98.8 %	99.2 %	100.0 %	100.0 %	0.0 %	0.0 %	3.0 %	20.9 %	48.0 %
Asian/Male	99.0 %	99.6 %	99.6 %	100.0 %	100.0 %	0.0 %	0.4 %	5.3 %	15.8 %	38.6 %
Black/Female	90.1 %	98.0 %	99.4 %	99.5 %	99.7 %	0.0 %	1.2 %	9.9 %	38.5 %	71.6 %
Black/Male	83.0 %	95.3 %	98.9 %	99.5 %	99.9 %	0.0 %	0.0 %	0.4 %	6.2 %	26.8 %
White/Female	92.1 %	96.8 %	98.6 %	98.9 %	99.5 %	0.0 %	0.0 %	0.1 %	3.2 %	14.9 %
White/Male	90.5 %	97.0 %	98.8 %	99.3 %	99.8 %	0.0 %	0.0 %	0.0 %	1.6 %	8.4 %
Age 12-20	90.1 %	95.6 %	97.9 %	98.9 %	99.5 %	0.0 %	0.3 %	3.3 %	13.5 %	33.8 %
Age 21-29	92.1 %	97.2 %	99.3 %	99.4 %	99.7 %	0.0 %	0.7 %	5.2 %	20.0 %	44.5 %
Age 30-41	93.9 %	99.1 %	99.4 %	99.8 %	100.0 %	0.0 %	0.0 %	2.2 %	15.2 %	36.9 %
Age 41-76	91.4 %	97.8 %	99.3 %	99.7 %	100.0 %	0.0 %	0.0 %	0.1 %	3.9 %	15.7 %

Table 6 shows the variation in TPIR between genders, ethnicities and ages at a face-match threshold of 0.9. A *t*-test (Welch's unequal variance *t*-test) was used to determine whether the observed demographic variations in TPIR are statistically significant at the.05 significance level (*p*-value less than .05).

At a face-match threshold of 0.9, the statistically significant findings are:

- 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 .05*
All cohort subjects		2.7 %	0.5 %		
Gender	Female	3.8 %	0.9 %	p =.025	*statistically significant*
	Male	1.5 %	0.5 %		
Ethnicity	Asian	4.0 %	1.1 %	p < .001 (W vs A &B)	*statistically significant*
	Black	5.5 %	1.5 %		
	White	0.04%	0.04 %		
Ethnicity & Gender	Asian/Female	3.0 %	0.9 %	p = .339	
	Asian/Male	5.3 %	2.2 %		
	Black/Female	9.9 %	2.6 %	p < .001	*statistically significant*
	Black/Male	0.4 %	0.3 %		
	White/Female	0.1 %	0.1 %	p = .320	
	White/Male	0.0 %	0.0 %		
Age	12-20 years	3.3 %	1.2 %	p < .001 (42+ vs 41-)	*statistically significant*
	21-29 years	5.2 %	1.5 %		
	30-41 years	2.2 %	0.6 %		
	42-76 years	0.1 %	0.1 %		

Table 7 shows the variation in FPIR between genders, ethnicities and ages at a face-match threshold of 0.8. A *t*-test (Welch's unequal variance *t*-test) was used to determine whether the observed demographic variations in TPIR are statistically significant at the.05 significance level (*p*-value less than .05).

At face-match threshold 0.8, the statistically significant findings are:

- 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 (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 21-20, of age 21-29, and of age 30-41 are not significant.)

Figure 7: Distribution of rank 1 non-mated candidate scores by demographic

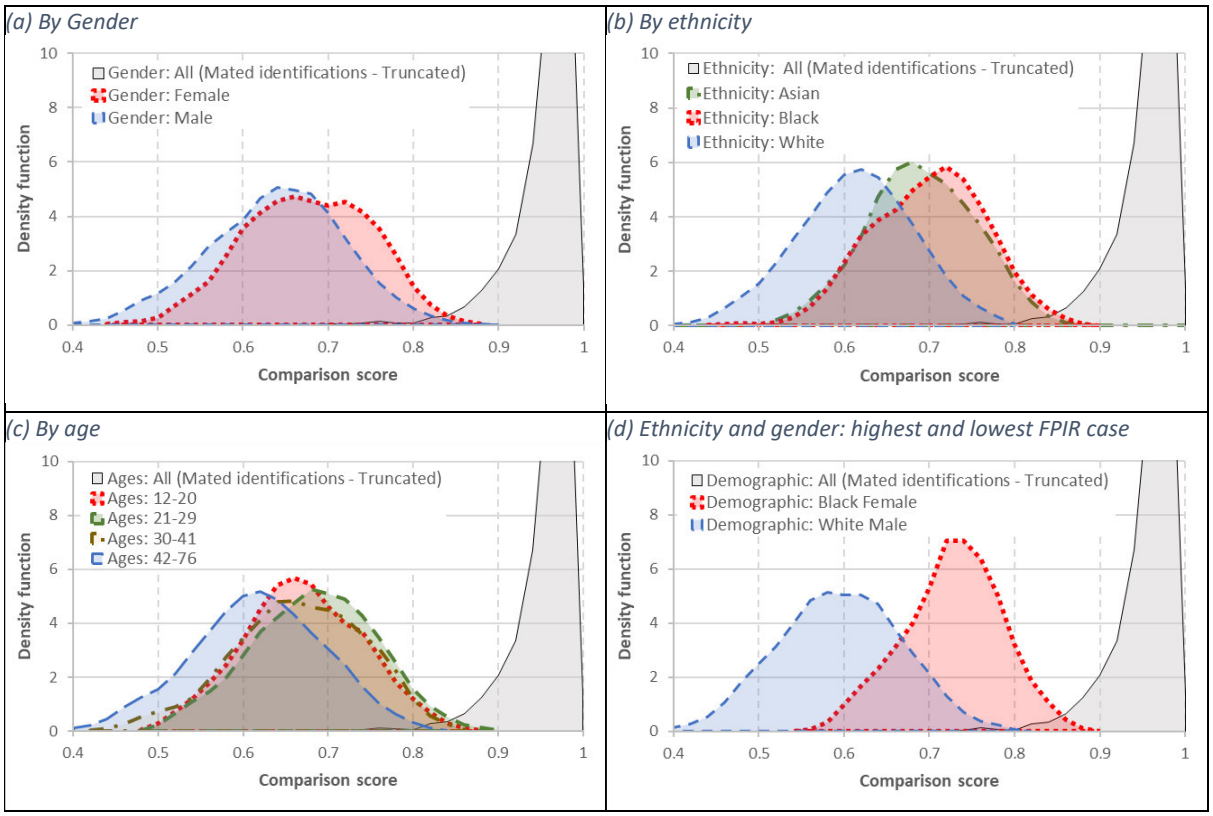


Figure 7 plots the distribution[4] of rank 1, non-mated candidate scores, for gender-, ethnicity-, and age-based subsets of the evaluation's identification searches. If there were no demographic variation FPIR these distributions would lie close together on the plots.

A false positive occurs when the rank 1 candidate score of a non-mated identification search exceeds the face-match threshold. When the distribution of non-mated scores spreads to higher comparison scores, a false positive becomes more likely. Also plotted is the mated comparison score distribution (over all demographics) to visualise the degree of overlap between mated and non-mated score distributions.

The plotted score distributions illustrate some of the findings regarding demographic variation in FPIR.

- In Figure 7(a), the 'Female' non-mated score distribution spans higher score values than the 'Male' distribution.
- In Figure 7(b), the 'White' non-mated score distribution spans a lower score range than the 'Asian' and 'Black' distributions. The 'Asian' and 'Black' distributions are closely overlapping.
- In Figure 7(c), the '42-76' non-mated score distribution spans a lower score range than the distributions for the other age categories. The '12-20', '21-29' and '30-41' mated score distributions are closely overlapping.

[4] The distributions have been generated from the comparison score data using Kernel Density Estimation using Gaussian basis functions with Silverman's rule of thumb for choice of bandwidth. [9].

7 Equitability

7.1 Retrospective Facial Recognition

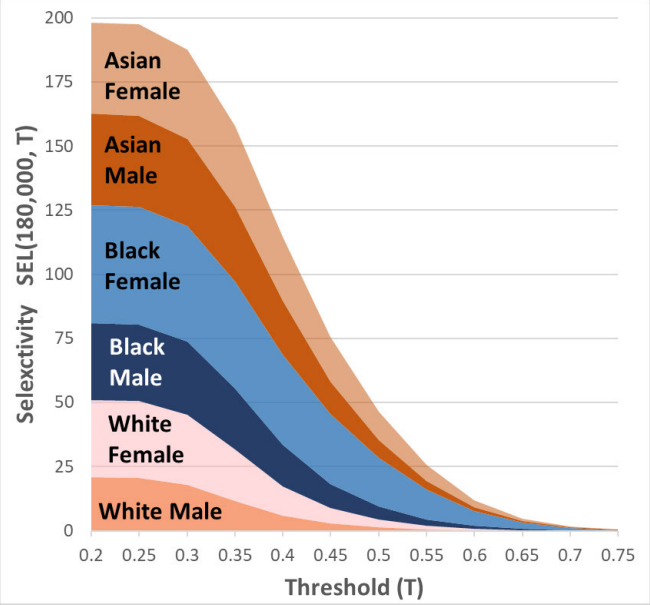
In the case of RFR, the data subjects that may be affected by algorithm performance are the individuals enrolled onto the reference database who may be added to a candidate list of an identification search as a result of a false positive. (A consequence, for example, could be being identified as a potential 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 8 plots, as a function of face-match threshold, the average number of candidates of each demographic returned in the non-mated identification searches (of Custody/Oifr/Adhoc/Selfie probe images). The figure does show an imbalance in the number of returned candidates of each demographic. E.g., 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 8: 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[5], or for Policing arrest numbers for 2021/22[6], 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 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 a 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 %

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

[6] (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

7.2 Officer Initiated Facial Recognition

In OIFR, it is the data subjects of the facial images taken and submitted for an identification who may be affected by a demographic variation in performance. A false positive will result in the officer further engaging with the subject photographed to resolve the disagreement on the subject's correct identity. From the subject's perspective a missed identification has the same outcome as not being in the reference database searched by OIFR. In most cases this will be of no consequence to the subject.

- In the case where OIFR is operated without thresholding of the candidate score, all but three of the OIFR relevant probe images returned a correct identification at rank 1. As the three exceptions are of a different gender/ethnicity, the demographic variation is not statistically significant.
- To limit the chances of a false positive in cases where the subject does not have their face image in the reference database (without reliance on officer adjudication) a face-match threshold may be used. At a face-match threshold of 0.9, there were no false positives, and no demographic variation in FPIR, TPIR was over 91 %, and lower than this for Black subjects. However, the missed identification is normally of no consequence.

Both cases may be considered equitable.

To evaluate demographic variation, the study used a uniformly balanced demographic for the reference database. In operational OIFR use cases the demographic composition of the reference database is more likely closer to that of the national population of England and Wales or the demographic profile on Policing arrests, and this will alter the demographic variation in FPIR. (TPIR is relatively unaffected by such a change).

It is possible to scale the evaluation results to provide an indicative FPIR performance using these notional operational profiles for the reference database demographic. See Table 11.

The evaluation results show the lowest FPIR values were for White ethnicity, so the change to a notional reference database with substantial White majority decreases FPIR and reduces the size of the demographic variation.

Table 11: Indicative FPIR(180000, T) by probe image demographic for different reference database demographic profiles

Reference database demographic profile	Face-match threshold (*T*)	Asian Female probe	Asian Male probe	Black Female probe	Black Male probe	White Female probe	White Male probe	All probes
Uniform	0.85	0.00 %	0.61 %	0.94 %	0.00 %	0.00 %	0.00 %	0.26 %
	0.80	2.88 %	5.31 %	8.82 %	0.46 %	0.12 %	0.00 %	2.93 %
	0.75	21.76 %	15.51 %	36.21 %	6.73 %	2.78%	1.27 %	14.04 %
Census	0.85	0.00 %	0.18 %	0.12 %	0.00 %	0.00 %	0.00 %	0.05 %
	0.80	0.69 %	1.56 %	1.57 %	0.06 %	0.03 %	0.00 %	0.65 %
	0.75	9.23 %	4.52 %	5.13 %	0.93 %	5.55 %	3.02 %	4.73 %
Arrest data	0.85	0.00 %	0.27 %	0.06 %	0.00 %	0.00 %	0.00 %	0.05 %
	0.80	0.13 %	2.30 %	0.76 %	0.13 %	0.00 %	0.00 %	0.56 %
	0.75	2.03 %	6.75 %	2.84 %	3.11 %	1.70 %	4.86 %	3.55 %

8 Discussion

8.1 Scalability

In operational RFR the reference database is typically much larger than of the evaluation, leading to a corresponding increase in FPIR from the evaluation's value.

Based on the evaluation results we can estimate the effect on FPIR of a 100-fold increase in reference database size FPIR. The estimate assumes similarity of the larger system with the evaluation system, i.e. (i) reference images are custody style images taken in accordance with the Police standard [7] (ii) probe images are consistent with the Custody/Oifr/Adhoc/Selfie image dataset (iii) subject demographics are those of the evaluation. We also assume that subject demographic of the reference dataset aligns with that of the 2021 Census [6].

We use the approximation $\text{FPIR}(c \times N, T) \approx c \times \text{FPIR}(N, T)$ which requires $\text{FPIR}(N, T) \ll 1/c$.

Taking the value $\text{FPIR}_{\text{census}}(180\,000, 0.85) = 0.05\% \ll 1\%$ from Table 11 in Section 7.2 we have

$\text{FPIR}_{\text{census}}(18\,000\,000, 0.85) \approx 100 \times 0.08\% \approx 5\%$. (This can be an indicative value only, as it is unlikely that a large operational system meets all our assumptions.)

There is scope to set a face-match threshold higher than 0.85 for a FPIR lower than 5% while maintaining a good TPIR. At threshold 0.90, TPIR was 91.9 % in the evaluation. (TPIR is relatively stable to changes in size of the reference database size)

The algorithm supports binning, (e.g., by ethnicity, gender, and date of birth) which could allow the identification transaction search only the subset of the reference relevant for the probe image.

8.2 Demographic performance variations for other thresholds and image types

Statistical analysis of demographic variation in TPIR and FPIR was conducted on the initial portion of the Equitability Study Dataset at face-match thresholds 0.9 and 0.8 respectively. Here we report which of the statistically significant observed variations in TPIR and FPIR (from Section 6), are also observed at other face-match thresholds, and for other image types.

For Custody/Oifr/Adhoc/Selfie image types, at face-match thresholds 0.85 – 0.98 the following demographic variation in TPIR was observed:

- The TPIR for Asian subjects is higher than that for White subjects.
- The TPIR for White subjects is higher than that for Black subjects.

For Custody/Oifr/Adhoc/Selfie and Outtake image types, at face-match thresholds 0.6, 0.65, 0.7, 0.75, 0.8 the following demographic variations in FPIR were observed:

- The FPIR for Male subjects was lower than that for Female subjects.
- The FPIR for White subjects was lower than that for Asian subjects and Black subjects.
- The FPIR for Black Males was lower than that for Black Females.
- The FPIR for subjects aged 42-76 was lower than that for subjects of younger age categories.

The general (but not universal) patten for ethnicity/gender variation in FPIR across image types and thresholds is that Asian Females, Asian Males, and Black Females have an FPIR above average, and Black Males, White Female, and White Males have an FPIR below average.

8.3 Utility and representativeness of the image type subsets

The combined the Custody, OIFR, Adhoc and Selfie image sets are useful, in providing several identification searches for each Cohort subject, providing good data for statistical investigation of demographic variation in performance. As a group the images are consistent with those that might be taken for OIFR but represent only a portion of the images used in RFR.

S.31(1)

9 Terminology and abbreviations

Binning: Partitioning of a reference database according to known on properties of the reference images (e.g., ethnicity, gender, year of birth). An identification transaction, based on the properties of the probe image, can then restrict the identification search to the relevant subset of reference images to speed up the search and potentially improve biometric performance.

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.

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.

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.

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

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

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).

10 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] Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman & Hall