

**REPORT FOR
NATIONAL BIOMETRIC FUNCTION AND
OFFICE OF THE POLICING CHIEF SCIENTIFIC ADVISER**

**ACCURACY AND EQUITABILITY EVALUATION OF
CORSIGHT APOLLO 4 LIVE FACIAL RECOGNITION**

MARCH 2026

Accuracy and Equitability Evaluation of Corsight Apollo 4 Live Facial Recognition

Data Science and AI Department

© NPL Management Limited, 2026

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

Extracts from this report must not be reproduced without prior approval from NPL.

CONTENTS

1	INTRODUCTION.....	1
1.1	About the evaluation	1
1.2	Key findings.....	2
2	BACKGROUND.....	4
2.1	Live Facial Recognition (LFR).....	4
2.2	Demographic categories	4
2.3	Assessing equitability.....	5
2.4	Statistical significance	5
3	TEST CORPUS.....	6
3.1	Cohort subjects	6
3.2	Filler images.....	6
3.3	LFR Video	7
4	METHODOLOGY.....	8
4.1	Watchlists.....	8
4.2	Offline running of LFR	8
4.3	Pre-analysis processing	8
4.4	Analyses	9
5	RECOGNITION ACCURACY	10
6	VARIATION IN PERFORMANCE BETWEEN DEMOGRAPHICS	12
6.1	Demographic variation in TPIR	12
6.2	Demographic variation in FPIR	13
6.3	Variation in TPIR and FPIR at threshold 55	15
6.4	Variation in TPIR and FPIR at threshold 40	16
6.5	Environmental variation in TPIR.....	17
7	EQUITABILITY	18
8	TERMINOLOGY AND ABBREVIATIONS	19
9	BIBLIOGRAPHY.....	21

1 INTRODUCTION

1.1 ABOUT THE EVALUATION

Policing is focused on enhancing practices and ensuring effective responses to community needs and the considered use of facial recognition (FR) technology can assist police in delivering efficiencies and effectiveness. Critical to use of the technology is ensuring it is implemented in a responsible, transparent, and ethical way: doing so requires an understanding of the accuracy and demographic equitability of the technology.

The objective of the evaluation described in this report is to assess the performance of a specific FR algorithm in an operational setting in terms of accuracy and equitability related to subject demographics. The FR technology and version evaluated is Corsight Live Facial Recognition, version Apollo 4.

NPL was commissioned by the National Biometric Function 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' [3] collected in 2022 for evaluation of the NEC NeoFace algorithm.

For the operational use case, the evaluation:

- determines the recognition accuracy of the Live Facial Recognition (LFR) algorithm for video and 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 LFR 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 case, the demographics addressed in the evaluation, and the assessment of equitability.
- Section 3 outlines the data of the Equitability Study Dataset used in the evaluation.
- Section 4 provides details regarding the evaluation methodology.
- Section 5 reports the recognition accuracy for the evaluation data.
- Section 6 examines the variation in performance between demographic categories.
- Section 7 addresses equitability of the FR algorithm on the operational use case.
- Section 8 provides a glossary of the key terms and abbreviations used in this report.

1.2 KEY FINDINGS

Live Facial Recognition (LFR) compares a live camera video feed of faces against a predetermined watchlist to find a possible match that generates an alert.

An accurate LFR system requires that, when a person passing through the zone of recognition has a reference facial image in the watchlist, this mated image and associated identity data should be returned by the identification search. Also, to avoid false matches, when a person passing through the zone of recognition does not have a mated image in the watchlist, the identification search should not return a candidate match for that person.

Recognition accuracy for LFR is stated in terms of the True Positive Identification Rate (TPIR), which is measured over mated recognition opportunities (where the person passing through the zone of recognition has a reference image in the watchlist) and the False Positive Identification Rate (FPIR) which is measured over non-mated recognition opportunities (where the person passing through the zone of recognition does not have a reference image in the watchlist).

- The True Positive Identification Rate, $TPIR(N, R, T)$, is the proportion of mated recognition opportunities where the mated reference is returned. TPIR depends on the number of facial images in the watchlist (N), the face-match threshold (T), and $R = 1$ as in LFR only a single candidate match is returned. TPIR is sometimes called the True Recognition Rate.
- The False Positive Identification Rate, $FPIR(N, T)$, is the proportion of non-mated recognition opportunities that return a candidate. FPIR depends on the number of facial images in the watchlist (N) and the face-match threshold (T). FPIR is sometimes called the False Alert Rate.

1.2.1 Recognition accuracy

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

For summarising operational performance, we use a face-match threshold of 55, which has been used in some operational deployments. We provide performance figures for two different watchlist sizes: (i) 18,000 reference images and (ii) 1,800 reference images.

Watchlist size 18,000

- $TPIR(18000, 1, 55) = 89 \%$
- $FPIR(18000, 55) \approx 0.017 \%$ (1 in 5,700)

Watchlist size 1,800

- $TPIR(1800, 1, 55) = 89 \%$
- $FPIR(1800, 55) \approx 0.002 \%$ (1 in 57,000)

Further details are provided in Section 5.

1.2.2 Demographic variation in TPIR

At face-match threshold 55, the Ethnicity/Gender group with the highest TPIR (94 %) was the Black/Male group, and the lowest TPIR (86 %) was for the White/Male group. However, the observed differences in TPIR by gender, by ethnicity, and by ethnicity-gender combined were not statistically significant at the 0.05 significance level. (Statistical significance quantifies whether the observed performance difference is likely due to chance, or due to some

underlying factor. Following convention, a 0.05 significance level was set prior to evaluation and analysis of results. Section 2.4 provides further detail on statistical significance.)

1.2.3 Demographic variation in FPIR

At face-match threshold 55, with a watchlist of 180,000 just three Cohort subjects had a false alert (1 Asian/Male, 1 Black/Female and 1 Black/Male). The number of false alerts is too small for this demographic variation to be statistically significant.

At face-match threshold 61 (and higher), there were no false alerts, and so no demographic variation in FPIR.

At face-match threshold 50 (and lower) the Gender/Ethnicity variation in FPIR is statistically significant at the 0.05 significance level.

1.2.4 Environmental variation in TPIR

Environmental conditions can affect TPIR, and we consider the variations in TPIR for the different locations and dates and of the Equitability Study Dataset video. At face-match threshold 55, the Piccadilly Circus 28/7/22 videos gave the highest TPIR (93 %), and the Oxford Street 7/7/22 videos gave the lowest TPIR (84 %). The range of variation is statistically significant at the 0.05 significance level and is slightly greater than that for variation due to demographics.

1.2.5 Equitability

Under the equitability criteria of the evaluation, the algorithm is considered equitable:

- for watchlist size 18,000, or 1,800 at face-match threshold 63
 - Demographic variation in TPIR is not statistically significant.
 - No variation in FPIR (no false positives observed at this threshold).
- for a watchlist size 18,000 or 1,800, at face-match threshold 55
 - Demographic variation in TPIR is not statistically significant.
 - Demographic variation in FPIR is not statistically significant.

Under the equitability criteria of the evaluation, the algorithm is considered not equitable:

- for watchlist size 18,000 at face-match threshold 50
 - Demographic variation in TPIR is not statistically significant.
 - Demographic variation in FPIR is statistically significant.
 - FPIR > 1 in 1000 for demographic with highest FPIR.

Further details are provided in Section 7.

2 BACKGROUND

2.1 LIVE FACIAL RECOGNITION (LFR)

Live Facial Recognition is an overt real-time deployment of facial recognition technology, which compares a live camera feed(s) of faces against a predetermined watchlist in order to locate persons of interest by generating an alert when a possible match is found.

In assessing equitability of LFR, classes of data subjects to consider are: (i) subjects not on the watchlist who pass through the zone of recognition of the LFR system, (ii) subjects on the watchlist who pass through the zone of recognition and (iii) subjects on the watchlist who do not pass through the zone of recognition. In the considerations of equitability in this report the focus is mainly on the first class of subjects, whose concerns may be false alerts by the system. The second class of subjects may have concerns about disparities in the True Recognition Rate between demographic groups. The third class of subjects may have concerns that false alerts will incorrectly log them as being present at the LFR location.

In LFR, the demographic composition of the watchlist can affect demographic variation in false positive identification 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 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 Gender¹

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.

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.)

2.3 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; the evaluation used a watchlist of 180,000 face images. However, equitability must be assessed at typical operational settings. We consider two watchlists of size closer to that of potential operational deployments containing 18,000 and 1,800 facial images. To scale results of larger watchlist (size $c \times N$) to a smaller watchlist (size N) we note that, provided $FPIR(N, T)$ is small:

$$\begin{aligned} TPIR(c \times N, 1, T) &\approx TPIR(N, 1, T) \text{ and} \\ FPIR(c \times N, T) &\approx c \times FPIR(N, T). \end{aligned}$$

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 the 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.

To be considered equitable any statistically significant demographic variation in performance should be inconsequential for the data subjects affected in the operational setting (i.e., with operational thresholds, composition of reference database, etc.).

Whether the impact on a data subject is negligible or of major concern is generally specific to the operational use case. In general, false alerts are of greater concern to the data subject than missed identifications. Sometimes there is a performance expectation. For example, the College of Policing Authorised professional practice on LFR [7] suggests that the FPIR should be less than 1 in 1000 in most operational scenarios. A demographic variation in performance would be of concern if the affected demographic group fails to meet the performance expectation.

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. To determine the statistical significance of demographic variations observed in the evaluation, when comparing TPIR or FPIR for two demographic groups a t-test (Welch's unequal variance t-test [8]) was used, and when comparing TPIR or FPIR of three or more demographic groups an analysis of variance (ANOVA) was used.

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. Correction for multiple testing is beyond the scope of this report.

3 TEST CORPUS

3.1 COHORT 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 number 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	

Custody-style images of the Equitability Study Dataset 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.

3.2 FILLER IMAGES

To provide a large demographically balanced reference database, Cohort data was supplemented by approximately 180,000 Filler images, provided from MPS holdings of custody image photos. The composition of the Filler set is shown in Table 2.

Table 2: Demographic composition of Filler Image Set ~180,000 faces from ~116,000 individuals

		Female	Male
Self-defined ethnicity ²	Asian (A1, A2, A3, A4, A9)	~30,000	~30,000
	Black (B1, B2, B9)	~30,000	~30,000
	White (W1, W2, W3, W9)	~30,000	~30,000
Age at date image taken	Age range ³	12-76 years	
	Lower quartile	22 years	
	Median	31 years	
	Upper quartile	40 years	

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

³ Age range of the Filler dataset taken as the 0.1th – 99.9th percentiles.

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.

3.3 LFR VIDEO

Cohort subjects were seeded into the Crowd flow over the course of 5 operational LFR deployments. These deployments were run with an operational watchlist that did not feature the Cohort. Two LFR cameras were used to span the zone of recognition.

At the street location of the LFR system, start and end points for the repeat walks by Cohort subjects through the zone of recognition were selected such that a round trip through the zone of recognition back to the start point should take at least one minute.

To count and log timings of the Cohort recognition opportunities, each Cohort subject was given a lanyard and badge showing their unique reference number (URN) as a number and a barcode. Scanning the barcodes logged the URN and time of scan to a spreadsheet.

At the LFR location Cohort subjects were briefed:

- To ensure that they walked through the zone of recognition of the LFR systems.
- To walk as they would normally; there was not a need to look directly at the camera, but not to be looking down at their mobile phone while walking towards the camera.⁴
- To avoid bunching into one large group. It was suggested that they might walk in pairs and allowed to converse.
- Cohort subjects were instructed to have their lanyard-badge barcode scanned for each of ten walks through the zone of recognition.

The wearing of caps, sunhats or sunglasses, glasses was not prohibited or suggested. All data collection deployment days were bright and sunny, and several Cohort subjects wore sunhats and sunglasses in keeping with the conditions, and typical of the non-Cohort public also passing the LFR system.

Video footage from the LFR was saved for use in the Equitability Study dataset. See Table 3 for details. The video includes Crowd subjects, giving an operationally realistic workload for LFR algorithms. No identifying data is held for Crowd subjects, and their alerts against the watchlist are discarded from analysis.

Table 3: Summary of LFR deployments with seeded Cohort subjects

Date	Location	Video duration	Number of Crowd subjects (estimate)	Number of Cohort subjects	Number of Cohort recognition opportunities
7 Jul 2022	London, Oxford Street	7 hours	24,000	71	706
14 Jul 2022	London, Oxford Street	8 hours	35,000	60	600
16 Jul 2022	London, Oxford Street	8 hours	38,000	89	890
28 Jul 2022	London, Piccadilly Circus	7 hours	28,000	136	1,350
13 Aug 2022	Cardiff, Queen Street	4.5 hours	7,000	45	442

⁴ While it is operationally realistic that some of the crowd intentionally or accidentally avoid showing their face to the LFR system, the focus of the evaluation was on the accuracy and equitability of the LFR for those faces processed by the system.

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.

4.1 WATCHLISTS

A Cohort watchlist was enrolled using custody-style image of each Cohort subject. We used images without glasses or facemask that were taken on the same day that subjects were seeded into the LFR video collected for the Equitability Study dataset.

A Filler watchlist was enrolled using all the images in the Filler dataset.

For determination of TPIR, the Equitability Study Dataset videos were run against the combined Cohort and Filler watchlists. (Mated recognition opportunities.)

For determination of FPIR, the Equitability Study Dataset videos were run against only the Filler watchlist. (Non-mated recognition opportunities.)

4.2 OFFLINE RUNNING OF LFR

The Equitability Study Dataset videos were processed using the Corsight algorithm. The algorithm was run at a lenient threshold of 35, allowing determination of TPIR and FPIR at thresholds above 35.

The Corsight algorithm recognitions are logged by “Appearances”, i.e. appearances of a subject in contiguous frames of video. Appearances correspond to Recognition Opportunities for the subject. Occasionally there may be multiple Appearances per Recognition Opportunity if a subject becomes occluded while in the field of view, or when the subject appears in the field of view of two cameras.

For each appearance the following information is saved:

- Approximate start and end times of the appearance (in terms of the time elapsed since start of video),
- Name of video being processed,
- The watchlists being used,
- Details of the highest scoring match between the subject and watchlist for the Appearance:
 - Image of face in video detected,
 - Image of matching face in watchlist,
 - Identity of matching face in watchlist (filename),
 - Comparison score.

4.3 PRE-ANALYSIS PROCESSING

Preprocessing of the raw results file included:

- Combine multiple appearances arising from a single recognition opportunity. For example, the LFR system used two cameras to cover the zone of recognition, resulting in subjects appearing in two videos.
- Account for the recognition opportunities without an appearance. The Equitability Study Dataset includes a log of the dates, times and counts of Cohort recognition opportunities.

- Inspect logs from non-mated recognition opportunities for false matches (i.e., Cohort subjects matching Filler subjects), providing correct detected face ID to the logged results. This process was facilitated by (i) the log of times of recognition opportunities (ii) the alignment between results files for mated recognition opportunities with non-mated recognition opportunities (iii) Cohort subjects wearing distinctive lanyards often visible in the Detected Face image.
- Removal of appearances of Crowd subjects (recognitions of Crowd subjects were not analysed).
- Removal of face images (this personally identifiable information is not required for performance analysis).
- Consolidation of results file to a single file for mated recognition opportunities, and a single file for non-mated recognition opportunities.

4.4 ANALYSES

- For each Cohort subject and for a selection of thresholds, count the number of recognition opportunities at or above the threshold. Divide by number of the subject's recognition opportunities to give a personal TPIR or FPIR.
- Filter subjects by demographic group to get mean and standard deviation of personal TPIR, FPIR for that demographic.
- Use t-test or ANOVA to conduct statistical analyses of significance in variations between demographic groups.

5 RECOGNITION ACCURACY

Figure 1 and Figure 2 plot the observed TPIR and FPIR as the face-match threshold varies from 35 to 70.⁵ Note that FPIR is plotted on a logarithmic axis, with threshold T on a linear scale (increasing the threshold by 3.5 approximately halves the FPIR).

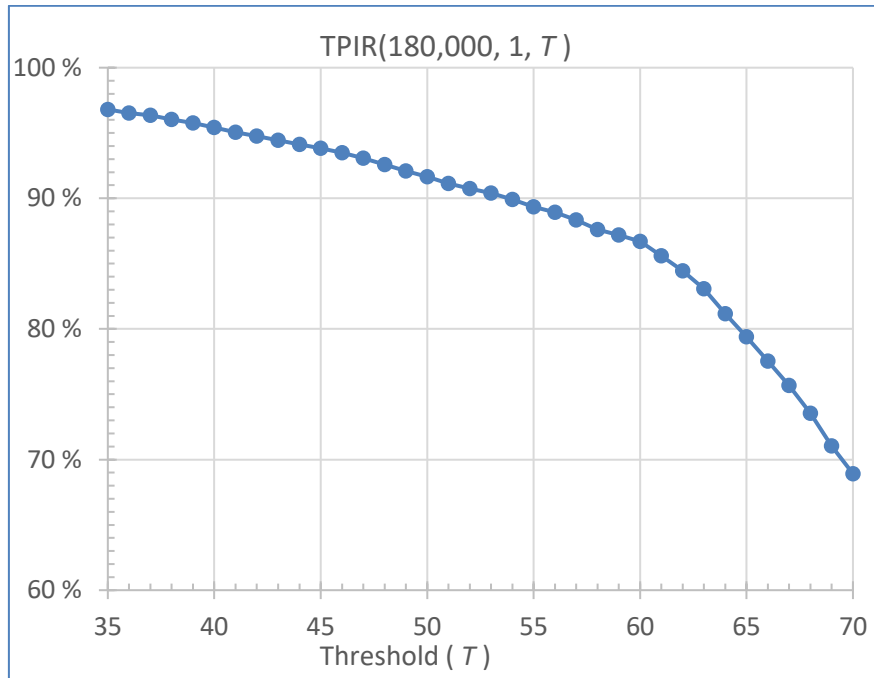


Figure 1: TPIR by threshold

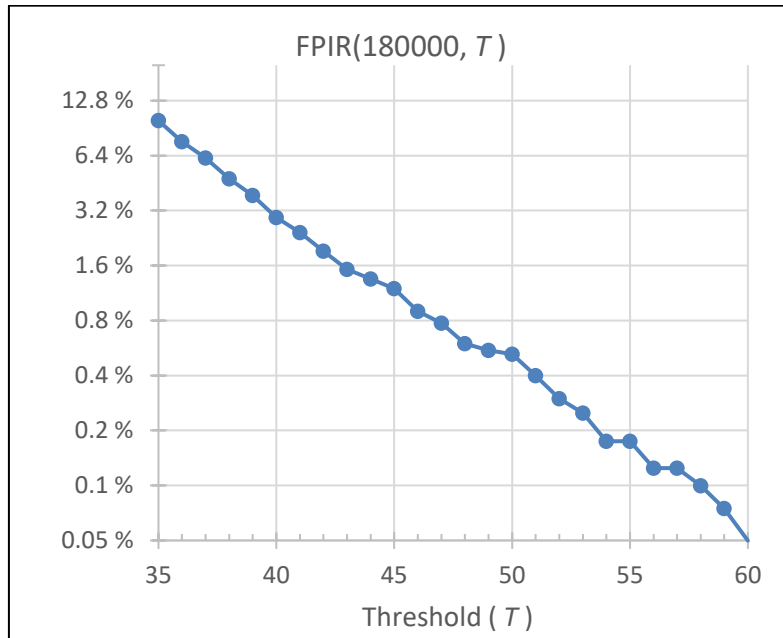


Figure 2: FPIR by threshold

⁵ Algorithm settings used for the evaluation were set to log face matches with best comparison score at 35 or above, which was deemed sufficient to cover normal operational settings. In the evaluation there were no false alerts scoring above 61.

Table 4 shows the observed TPIR and FPIR at face-match thresholds ranging from 45 to 63. Over this range TPIR reduces from 94 % to 83 %.

Table 4: TPIR and FPIR by face-match threshold

Face-match threshold T	Observed TPIR TPIR(180000, 1, T)	Observed FPIR FPIR(180000, T)	FPIR anticipated for watchlist size 18,000 FPIR(18000, T)	FPIR anticipated for watchlist size 1,800 FPIR(1800, T)
45	93.8 %	1.20 %	1 in 830	1 in 8,300
50	91.7 %	0.52 %	1 in 1,900	1 in 19,000
55	89.3 %	0.17 %	1 in 5,700	1 in 57,000
60	86.7 %	0.05 %	1 in 20,000	< 1 in 100,000
63	83.1 %	0.00 %	< 1 in 20,000	< 1 in 100,000

There were 48 false alerts at face-match threshold 45, 21 false alerts at threshold 50, seven false alerts at threshold 55, two false alerts at threshold 60, and no false alerts at threshold 63.

The evaluation was run with a watchlist of 180,000 facial images. This is much larger than watchlists in operational deployments to date, and we consider what FPIR would be anticipated for watchlists of one tenth or one hundredth of this size. FPIR should reduce by a factor of 10 for a watchlist of size 18,000 and by a factor of 100 for a watchlist of size 1,800. Thus, in an operational setting with a watchlist of 18,000 face images and a face-match threshold of 55, the anticipated False Alert Rate, FPIR(18000, 55) is 0.017 % (approximately 1 in 5700)⁶, and with a watchlist of size 1,800, the anticipated false alert rate FPIR(1800, 55) is approximately 1 in 57,000.

⁶ Sometimes it is easier to comprehend FPIR when expressed as a ratio e.g., "1 in 5700" than when expressed as a percentage, in this case "0.017%".

An FPIR percentage of $x\%$ converts to a ratio of = 1 false alert in $(100 \div x)$ recognition opportunities

6 VARIATION IN PERFORMANCE BETWEEN DEMOGRAPHICS

6.1 DEMOGRAPHIC VARIATION IN TPIR

Figure 3 shows the observed demographic variation in TPIR results over a range of face-match thresholds from 35 to 70.

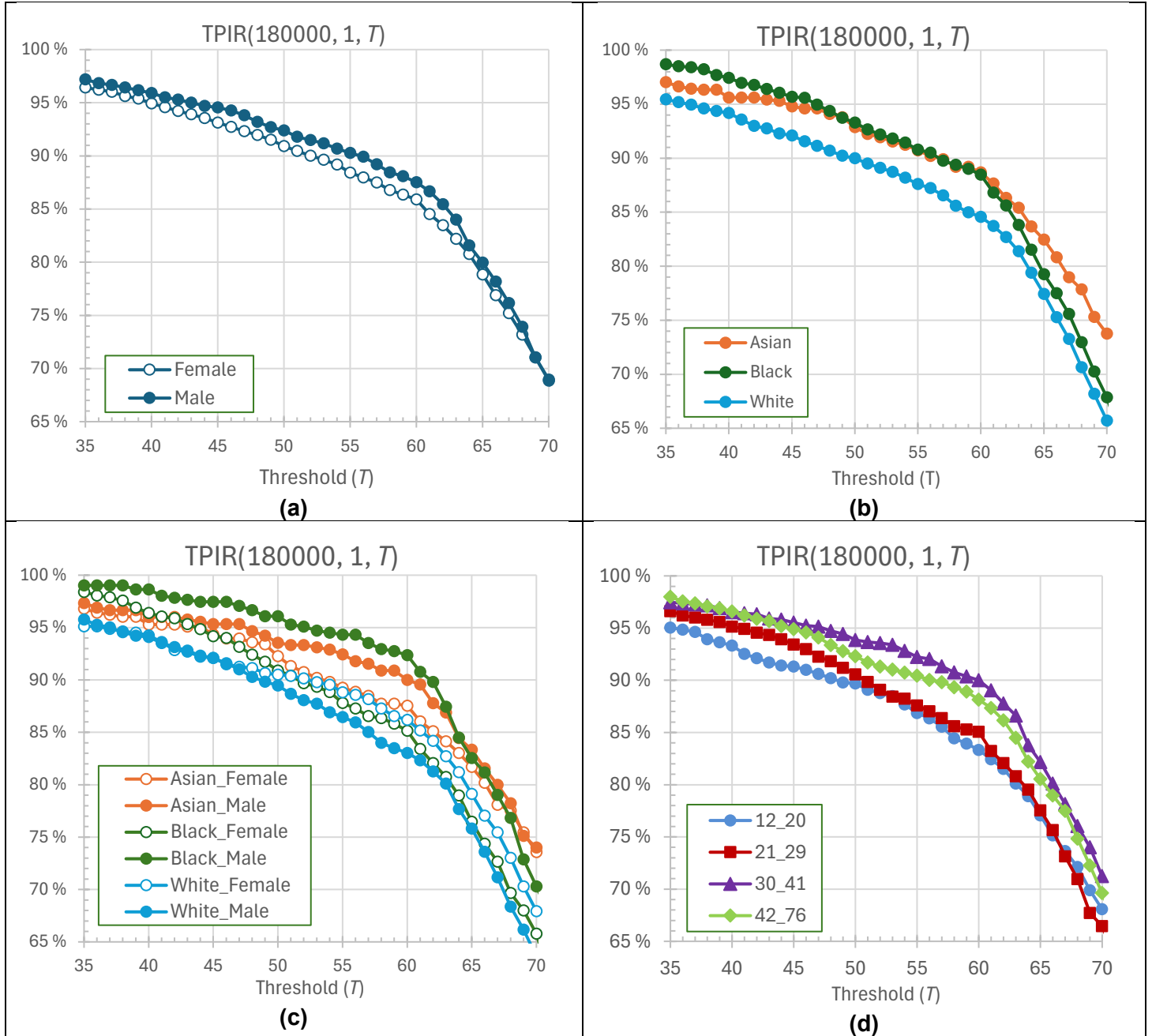


Figure 3: Demographic variation in TPIR
 (a) by Gender, (b) by Ethnicity, (c) by Ethnicity & Gender, (d) by Age

Table 5 shows the observed demographic variation in TPIR results over a range of face-match thresholds from 45 to 63.

Table 5: TPIR by demographic and face-match threshold

Demographic	TPIR(180000, 1, T)				
	T = 45	T = 50	T = 55	T = 60	T = 63
Full Cohort	94 %	92 %	89 %	87 %	83 %
Female	93 %	91 %	88 %	86 %	82 %
Male	95 %	92 %	90 %	88 %	84 %
Asian	95 %	93 %	91 %	89 %	85 %
Black	96 %	93 %	91 %	88 %	84 %
White	92 %	90 %	88 %	85 %	81 %
Asian/Female	94 %	92 %	89 %	88 %	84 %
Asian/Male	95 %	94 %	92 %	90 %	87 %
Black/Female	94 %	91 %	88 %	85 %	81 %
Black/Male	97 %	96 %	94 %	92 %	87 %
White/Female	92 %	91 %	89 %	86 %	83 %
White/Male	92 %	89 %	86 %	83 %	80 %
Age 12-20	91 %	90 %	87 %	83 %	80 %
Age 21-29	93 %	91 %	88 %	85 %	81 %
Age 30-41	96 %	94 %	92 %	90 %	87 %
Age 41-76	95 %	92%	90%	88 %	84 %

6.2 DEMOGRAPHIC VARIATION IN FPIR

Table 6 shows the observed demographic variation in FPIR results over a range of face-match thresholds from 45 to 63.

Table 6: FPIR by demographic and face-match threshold

Demographic	FPIR(180000, T)				
	T = 45	T = 50	T = 55	T = 60	T = 63
false alerts at threshold	48	21	7	2	0
Cohort with false alert	27	10	3	1	0
Full Cohort	1.20 %	0.52 %	0.17 %	0.05 %	0.00 %
Female	1.68 %	0.79 %	0.25 %	0.10 %	0.00 %
Male	0.71 %	0.25 %	0.10 %	0.00 %	0.00 %
Asian	0.71 %	0.20 %	0.10 %	0.00 %	0.00 %
Black	3.52 %	1.71 %	0.54 %	0.18 %	0.00 %
White	0.12 %	0.00 %	0.00 %	0.00 %	0.00 %
Asian/Female	0.75 %	0.00 %	0.00 %	0.00 %	0.00 %
Asian/Male	0.67 %	0.44 %	0.22 %	0.00 %	0.00 %
Black/Female	5.02 %	2.67 %	0.83 %	0.33 %	0.00 %
Black/Male	1.76 %	0.59 %	0.20 %	0.00 %	0.00 %
White/Female	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
White/Male	0.23 %	0.00 %	0.00 %	0.00 %	0.00 %
Age 12-20	0.30 %	0.00 %	0.00 %	0.00 %	0.00 %
Age 21-29	3.23 %	1.77 %	0.73 %	0.21 %	0.00 %
Age 30-41	1.16 %	0.38 %	0.00 %	0.00 %	0.00 %
Age 41-76	0.20 %	0.00 %	0.00 %	0.00 %	0.00 %

Figure 4 plots the observed demographic variation in FPIR over a range of face-match thresholds from 35 to 60. (FPIR has been plotted on a logarithmic scale.) At face match threshold 61 and above there were no false alerts.

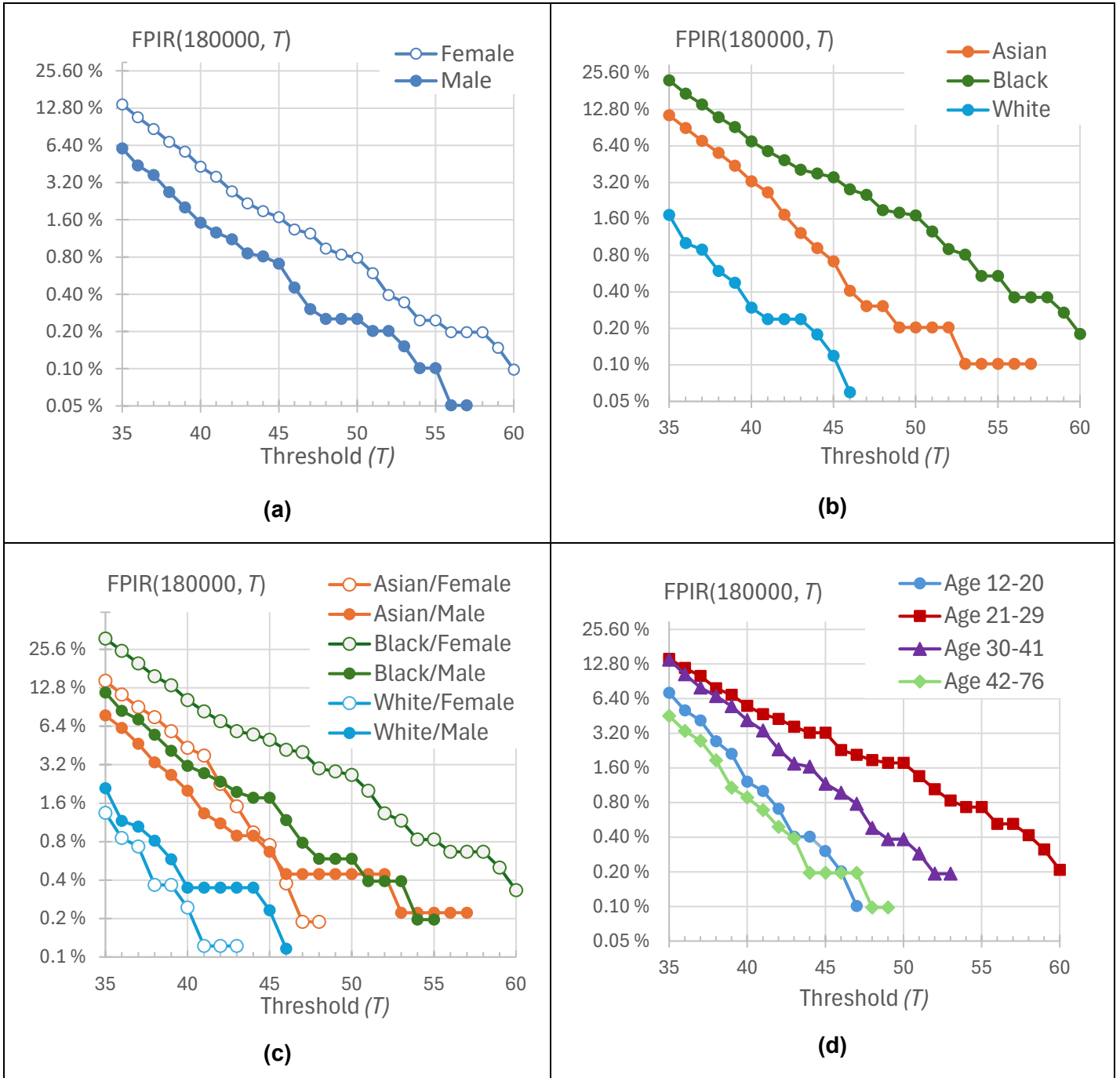


Figure 4: Demographic variation in FPIR
 (a) by Gender, (b) by Ethnicity, (c) by Ethnicity & Gender, (d) by Age

6.3 VARIATION IN TPIR AND FPIR AT THRESHOLD 55

Table 7 considers the demographic variation in TPIR at face-match threshold 55.

Across Ethnicity/Gender groups the TPIR ranged from 86 % (White/Male) to 94 % (Black/Male). At the 0.05 significance level, the demographic variations in TPIR by gender, by ethnicity, and by age are not statistically significant.

Table 7: Variation in TPIR by demographic at threshold 55

Demographic		TPIR(180000, 1, T = 55)		p-value	Significance of observed demographics variation at significance level 0.05
		Mean for demographic	Standard error of mean		
All Cohort subjects		89.3 %	0.9 %		
Gender	Female	88.4 %	1.4 %	$p = 0.30$	statistically not significant
	Male	90.3 %	1.2 %		
Ethnicity	Asian	90.7 %	1.7 %	$p = 0.23$	statistically not significant
	Black	90.8 %	1.5 %		
	White	87.6 %	1.5 %		
Ethnicity & Gender	Asian/Female	89.2 %	2.3 %	$p = 0.15$	statistically not significant
	Asian/Male	92.4 %	2.4 %		
	Black/Female	87.8 %	2.6 %		
	Black/Male	94.3 %	1.4 %		
	White/Female	88.8 %	2.2 %		
Age	White/Male	86.5 %	2.1 %	$p = 0.11$	statistically not significant
	12-20 years	86.9 %	2.1 %		
	21-29 years	87.6 %	2.0 %		
	30-41 years	92.2 %	1.4 %		
	42-76 years	90.4 %	1.6 %		

Table 8 considers the demographic variation in performance at face-match threshold 55. At this threshold only 3 Cohort subjects had false alerts, and together only 7 false alerts. The number of alerts is too low to reveal demographic variation in FPIR at this threshold.

Table 8: Variation in FPIR by demographic at threshold 55

Demographic		FPIR(180000, T = 55)		p-value	Significance of observed demographics variation at significance level 0.05
		Mean for demographic	Standard error of mean		
All Cohort subjects		0.17 %	0.13 %		
Gender	Female	0.25 %	0.25 %	$p = 0.57$	Statistically not significant
	Male	0.10 %	0.07 %		
Ethnicity	Asian	0.10 %	0.10 %	$p = 0.24$	Statistically not significant
	Black	0.54 %	0.46 %		
	White	0.00 %	0.00 %		
Ethnicity & Gender	Asian/Female	0.00 %	0.00 %	$p = 0.47$	Statistically not significant
	Asian/Male	0.22 %	0.22 %		
	Black/Female	0.83 %	0.83 %		
	Black/Male	0.20 %	0.20 %		
	White/Female	0.00 %	0.00 %		
Age	White/Male	0.00 %	0.00 %	$p = 0.12$	Statistically not significant
	12-20 years	0.00 %	0.00 %		
	21-29 years	0.73 %	0.54 %		
	30-41 years	0.00 %	0.00 %		
	42-76 years	0.00 %	0.00 %		

6.4 VARIATION IN TPIR AND FPIR AT THRESHOLD 40

Table 9 and Table 10 show the demographic variation in TPIR and FPIR at face-match threshold 40.

In contrast to the results at threshold 55, the observed demographic variation in FPIR by Gender, by Age, and by Ethnicity are all considered statistically significant. The demographic group with FPIR furthest from the overall in the Black/Female group.

The observed demographic variation in TPIR by Ethnicity is also statistically significant.

Table 9: Variation in TPIR by demographic at threshold 40

Demographic		TPIR(180000, 1, T = 40)		p-value	Significance of observed demographics variation at significance level 0.05
		Mean for demographic	Standard error of mean		
All Cohort subjects		95.4 %	0.5 %		
Gender	Female	94.9 %	0.9 %	p = 0.36	Statistically not significant
	Male	95.9 %	0.6 %		
Ethnicity	Asian	95.6 %	1.0 %	p = 0.043	Statistically significant
	Black	97.4 %	0.7 %		
	White	94.2 %	1.0 %		
Ethnicity & Gender	Asian/Female	95.3 %	1.3 %	p = 0.18	Statistically not significant
	Asian/Male	96.0 %	1.6 %		
	Black/Female	96.4 %	1.2 %		
	Black/Male	98.6 %	0.6 %		
	White/Female	94.2 %	1.6 %		
	White/Male	94.1 %	1.1 %		
Age	12-20 years	93.3 %	1.3 %	p = 0.11	Statistically not significant
	21-29 years	95.1 %	1.2 %		
	30-41 years	96.5 %	0.9 %		
	42-76 years	96.6 %	0.9 %		

Table 10: Variation in FPIR by demographic at threshold 40

Demographic		FPIR(180000, T = 40)		p-value	Significance of observed demographics variation at significance level 0.05
		Mean for demographic	Standard error of mean		
All Cohort subjects		2.93 %	0.48 %		
Gender	Female	4.31 %	0.82 %	p = 0.003	Statistically significant
	Male	1.52 %	0.47 %		
Ethnicity	Asian	3.27 %	0.80 %	p < 0.001	Statistically significant
	Black	6.98 %	1.47 %		
	White	0.30 %	0.18 %		
Ethnicity & Gender	Asian/Female	4.34 %	1.36 %	p < 0.001	Statistically significant
	Asian/Male	2.00 %	0.68 %		
	Black/Female	10.24 %	2.28 %		
	Black/Male	3.14 %	1.62 %		
	White/Female	0.24 %	0.24 %		
	White/Male	0.35 %	0.26 %		
Age	12-20 years	1.21 %	0.39 %	p = 0.001	Statistically significant
	21-29 years	5.54 %	1.57 %		
	30-41 years	4.16 %	0.92 %		
	42-76 years	0.88 %	0.46 %		

6.5 ENVIRONMENTAL VARIATION IN TPIR

Table 11 considers the variation in TPIR by LFR deployment location and date. Across deployments the TPIR ranged from 84 % (Oxford St, 7 July) to 93 % (Piccadilly Circus, 28 July). The variation is statistically significant. The demographic balance of the Cohort was not uniform over the deployments with, for example, more Black subjects attending on 28 July than on the other dates. This contributes to the observed demographic variations in TPIR.

Table 11: Environmental variation in TPIR between deployments at threshold 55

Deployment date	TPIR(180000, $T = 55$)		p -value	Significance of observed variation at significance level 0.05
	Mean for demographic	Standard error of mean		
7 Jul 2022	84.2 %	2.1 %	$p = 0.001$	Statistically significant
14 Jul 2022	91.0 %	2.3 %		
16 Jul 2022	85.4 %	2.3 %		
28 Jul 2022	93.3 %	1.3 %		
13 Aug 2022	91.2 %	2.2 %		

7 EQUITABILITY

To be considered equitable, any statistically significant demographic variation in performance should be inconsequential or negligible for the affected data subjects in the operational setting (i.e., with operational thresholds, composition of reference database, etc.).

At face-match threshold 61 or higher, there were no false positives in the evaluation, and so no demographic variation in FPIR. Moreover, at this threshold demographic variation in TPIR was not statistically significant. The algorithm is considered equitable at such a threshold.

At face-match thresholds 55 and 60, demographic variation in FPIR, and demographic variation in TPIR were not statistically significant, and the algorithm is considered equitable at these thresholds.

At face-match thresholds 50 and below, demographic variation in FPIR is statistically significant, and equitability depends on whether the impact on a data subject is negligible or major concern is generally specific to the operational use case. In general for LFR, false alerts, which may direct a police officer to engage with the data subject, are of greater impact on the data subject than missed identifications. Sometimes there is a performance expectation. For example, the College of Policing, Authorised professional practice on LFR [7] suggests that the FPIR should be less than 1 in 1000 in most operational scenarios. A demographic variation in performance would be of concern if the affected demographic group fails to meet the performance expectation.

Table 12 considers which combinations of watchlist sizes and face-match thresholds that give equitable performance under these criteria.

Table 12: FPIR for outlier demographic group at select thresholds and watchlist sizes

		FPIR Outlier	Black/Female		
		Threshold	50	45	40
180,000 watchlist	Observed FPIR for outlier demographic (Standard error)	2.67 % (1.73 %)	5.02 % (2.02 %)	10.24 % (2.28 %)	
	Variation statistically significant	Yes	Yes	Yes	
18,000 watchlist	Anticipated FPIR for outlier demographic (and Standard Error)	0.267 % (0.173 %)	0.502 % (0.202 %)	1.024 % (0.228 %)	
	Impact: FPIR > 1 in 1000	Yes	Yes	Yes	
	Equitable?	No	No	No	
1,800 watchlist	Anticipated FPIR for outlier demographic (and Standard Error)	0.027 % (0.018 %)	0.051 % (0.021 %)	0.103 % (0.023 %)	
	Impact: FPIR > 1 in 1000	No	No	Yes	
	Equitable?	Yes	Yes	No	

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.

Crowd: Members of the public passing through the zone of recognition of the LFR system.

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 that 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.

LFR: Live Facial Recognition.

Mated: A mated identification search is one in which the subject in probe image also has a reference image in the reference database. A mated recognition opportunity is one where the subject passing through the LFR zone of recognition has an image in the LFR watchlist. 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. A non-mated recognition opportunity is one where the subject passing through the LFR zone of recognition does not have a facial image in the LFR watchlist. Similarly, a non-mated comparison score is produced from comparison of face images of different individuals.

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

Recognition opportunity: The period when a subject moves through the zone of recognition of an LFR system facing towards to the LFR camera.

Reference image: A facial image in the reference database.

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

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

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).

Watchlist: A set of reference images (of individuals of interest to policing) against which a probe image is searched.

Zone of recognition: Three-dimensional space within the field of view of the Live Facial Recognition camera and in which the imaging conditions for robust facial recognition are met.

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] College of Policing, APP, Live Facial Recognition, Key performance metrics, 2022, www.college.police.uk/app/live-facial-recognition/key-performance-metrics
- [8] 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
- [9] NPIA, 'Police Standard for Still Digital Image Capture and Data Interchange of Facial/Mugshot and Scar, Mark & Tattoo Images', National Policing Improvement Agency, 2007
- [10] ISO/IEC 30137-1:2024 Use of biometrics in video surveillance systems - Part 1: System design and specification