NISTIR 8429

Face Recognition Vendor Test (FRVT)

Part 8: Summarizing Demographic Differentials

Patrick Grother
Information Access Division
Information Technology Laboratory

2023/03/24

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NISTIR 8429

Face Recognition Vendor Test (FRVT)

Part 8: Summarizing Demographic Differentials

Patrick Grother
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8429

July 2022

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie Locascio, Under Secretary of Commerce for Standards and Technology and Director

ACKNOWLEDGMENTS

The authors are grateful to the Department of Homeland Security's (DHS) Science & Technology Directorate (S&T) for their support of this work. We are grateful to staff at SAIC's Maryland Test Facility, Idemia and Amazon for contributions and discussions, and to Joyce Yang for multiple reviews.

Additionally we extend our appreciation to DHS' Office of Biometric Identity Management (OBIM) for their image datasets and ongoing support.

The authors are indebted to the staff in the NIST Biometrics Research Laboratory for infrastructure supporting rapid evaluation of algorithms.

DISCLAIMER

Specific hardware and software products identified in this report were used in order to perform the evaluations described in this document. In no case does identification of any commercial product, trade name, or vendor, imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

INSTITUTIONAL REVIEW BOARD

The National Institute of Standards and Technology's Research Protections Office reviewed the protocol for this project and determined it is not human subjects research as defined in Department of Commerce Regulations, 15 CFR 27, also known as the Common Rule for the Protection of Human Subjects (45 CFR 46, Subpart A).

RELEASE NOTES

2023-03-24: Removed text about this document being an open-for-comment draft. Removed incorrect subscript on left side of equation 35.

Abstract

In December 2019, NIST Interagency Report 8280 quantified and visualized demographic variations for many face recognition algorithms. The report also suggested various mitigations, one of which - the focus of this report - was to define summary inequity measures that developers can work to improve and which can guide algorithm selection. Since 2019, it has become apparent that false negative inequities are substantially due to poor photography of certain groups including under-exposure of dark-skinned individiuals, and that this can be addressed by using algorithms more tolerant of poor image quality or, better, by correcting the capture process with superior cameras, imaging environments and human-factors. At the same time, it is also clear that the much larger false positive variations, which occur even in high-quality photographs, must be mitigated by algorithm developers. To those ends, this report compiles and analyzes various demographic summary measures for how face recognition false positive and false negative error rates differ across age, sex, and race-based demographic groups. We exercise some of the proposed measures by tabulating them for many algorithms submitted to the one-to-one comparison track of the Face Recognition Vendor Test. Those results appear on a regularly updated public webpage.

1 Introduction

Components of biometric systems, whether AI-based or not, may have different outputs and performance for different demographic groups, and there is wide consensus that these differences should be minimized. Such components obviously include the core recognition algorithms but also quality algorithms used to adjudicate image suitability, presentation attack detection mechanisms checking for liveness, spoofing or evasion, and cameras or imaging environments that can produce systematically different images for different demographic groups [2]. In this latter respect, there are broadly two kinds of face capture: First with cameras that have no, or minimal, understanding of what they're looking at; and second those with some intelligent understanding of a good face photo[1] achieving that with adaptive exposure and head-orientation estimation. Note that for the first category, capture is guaranteed - some image *will* be captured - even if it has poor quality, and it often then falls on the downstream recognition algorithms to tolerate camera-caused or, more generally, photography-caused, variations. Indeed, in some applications, the owner or operator of the algorithm will have no control over how an image is collected. With the intelligent camera, it is possible that some proportion of capture attempts will yield no image at all - for example because the subject never looked at the camera directly, or because lighting was inadequate, or because the camera and or detection was slow or ineffective. That proportion - the failure to acquire rate - can exceed the core recognition error rates[2]. This report deals with summarizing performance differences of face recognition algorithms applied to image collections where failure to acquire rates were zero *by policy*.

In December 2019, we published NIST Interagency Report 8280: FRVT Part 3: Demographics [5] that quantified false negative and false positive error rates for demographic groups defined by the available sex, age, and race metadata. This was performed for hundreds of one-to-one verification and one-to-many identification algorithms submitted to the ongoing Face Recognition Vendor Test benchmarks. The main finding of the survey was that false positive differentials are widespread, occuring even in pristine images, and greatly exceed those for false negative effects - for example, the within-group false positives rates in Figure 1 vary by up a factor of 720[3] while the false negatives in Figure 2 vary by around a factor of 3. The report noted that a priori probability and impact of both types of error are highly application dependent. The report additionally provided guidance in three additional areas: 1. Context and definitions notably that face recognition deals with identity and is therefore done with different, specialized, machinery to that used for sex-classification[4] and age-estimation; 2. Metrics and reporting, particularly in differentiating false positive and false negative effects - to encourage more specificity than just saying face recognition is biased; and 3. ways to mitigate

[1] Correct exposure of faces has been addressed in mobile phones - for example this feature and a commercial for it - though not necessarily for biometric purposes.
[2] See, for example, results from DHS' comprehensive tests of rapid-capture solutions.
[3] Polish men aged 35-50 have FMR of 1 in 26000; Nigeria women aged 60 and over have FMR of 1 in 35.
[4] Demographic effects for this task were documented in the Gender Shades studies [1,10] and in prior NIST work [9].

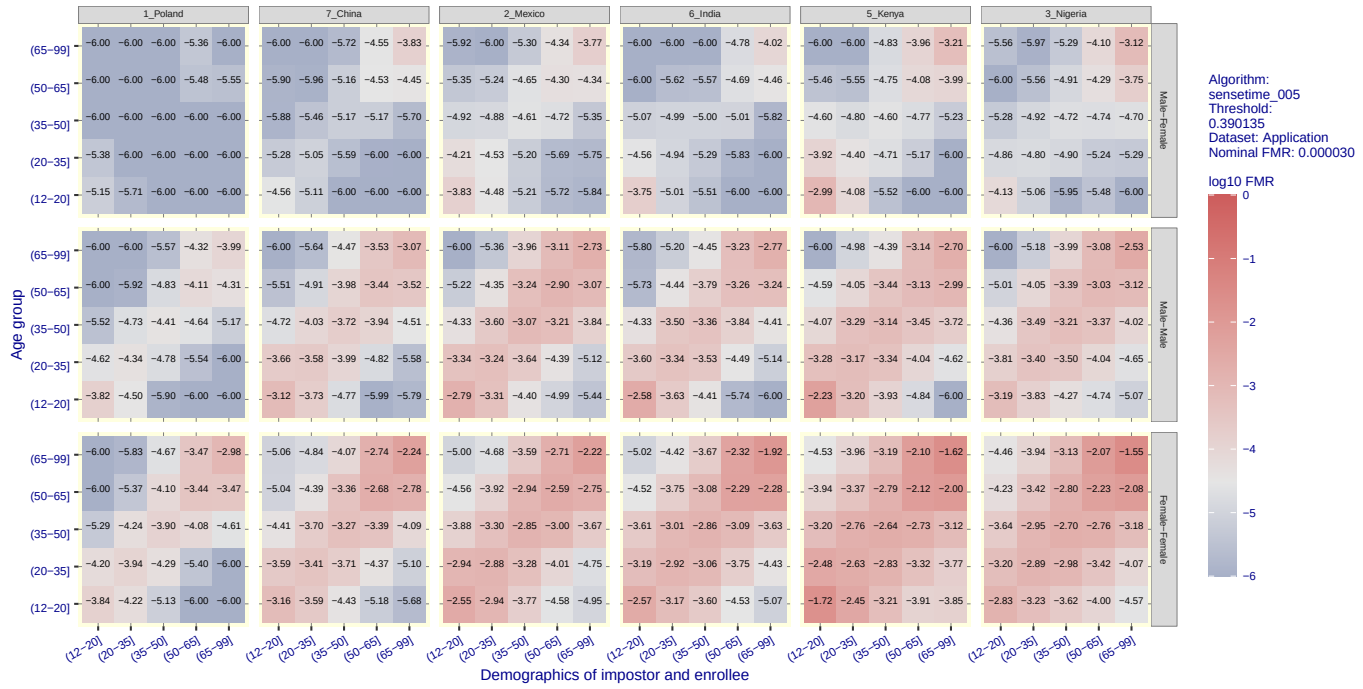

Figure 1: For six countries of birth and five age groups, the panels show false match rates at the single fixed threshold value given in the legend. The FMR estimates are measured over comparisons of photos of different people, in the top row different sex faces, then men, and finally women in the bottom row. The text in each cell and the color encode $\log_{10}$*FMR such that a difference of 2 between two cells corresponds to a factor of 100 excursion in false match rates.*

demographic effects. In the latter aspect, we advocated for the development of summary measures of recognition algorithm against which technologists could measure progress in reducing differences in recognition error rates across demographic groups, while continuing to reduce overall error rates.

This report documents various summary measures. It is intended to support developers, and to inform development of the ISO/IEC 19795-10 standard entitled *Quantifying biometric system performance across demographic groups*. The standard, which is expected to be published in 2023, includes equitability in its scope, and this relates directly to the topic of fairness currently being addressed in the broader AI community.

This report proceeds with a section on equity measures, including those proposed previously, then discussion of pertinent sub-topics (thresholds, weighting, uncertainty), and finally sections on results and discussion. Note that we apply our summary measures to hundreds of one-to-one face comparison algorithms and, of necessity, report results on an external web page, the demographics tab of the FRVT results page. These tables summarize both false negative and false positive effects. There is the erroneous belief that false positives have little effect in one-to-one comparison applications, because they effect impostors. But if a certain demographic group is associated with a high false positive rate, then applications of face comparison (such as automated border control, access control to a phone, authorization of a payment, and non-repudiation of the dispenser of a drug) would have security holes - the legitimate enrollee is vulnerable to impostors. False positives can be limited by adopting a higher threshold set globally to target a specific false match rate (FMR) in the worst-case (highest-FMR) demographic group. This will reduce FMR including in the non-problematic demographics, and will elevate false non-match rates (FNMR) generally. In a one-to-one setting it will be difficult for an impostor to exploit high FMR in particular groups: Specifically it will be difficult to arrange for a false match even if FMR were as high as 1 in 50. That value sounds un-realistic yet as Figure 1 shows, for a highly accurate algorithm, that Nigerian women aged 65 and over have an FMR of 1 in 35 when the threshold is set to achieve FMR of 1 in 25000 in Polish men aged 35 to 50. We include Annex B to discuss why results for one-to-one comparison algorithms are sometimes

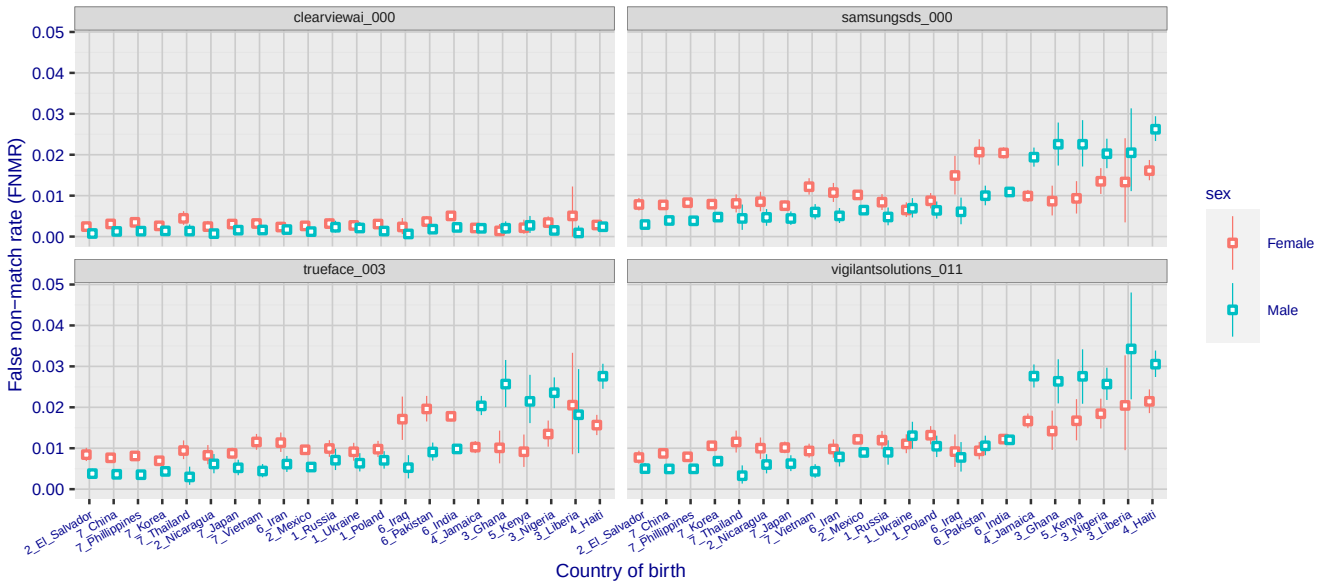

Figure 2: For four algorithms submitted to FRVT in the second half of 2021, the plots show false non-match rates by country-of-birth and sex. The error-bars cover 95% of bootstrap estimates of FNMR. Some error-bars are smaller than the plotted point. The x-axis is sorted in order of increasing FNMR. Note the range of FNMR across algorithms is larger than that across demographics.

pertinent to one-to-many search algorithms.

2 Equity measures

The next subsections detail equations for summarizing demographic differences in the basic biometric one-to-one comparison error rates, false match rate (FMR) and false non-match rate (FNMR). Throughout the paper these two quantities carry subscripts indexing a demographic group, so FNMR_i quantifies false non-match occurence for people in group i. Importantly FMR_i quantifies false matches between individuals *both* of whom are in group i. We do not consider cross-group FMR in any computation. Thus, even though such rates are depicted in some figures (e.g. the off-diagonal elements of Fig. 1), we do not use inter-group false match rates in what follows.

2.1 Differential error measures

In prior work Freitas et al. [3] of the Swiss Idiap Research Institute formulated a Fairness Discrepancy Rate from measurements of within-demographic $\text{FMR}(\tau)$ and $\text{FNMR}(\tau)$ at some threshold τ for each demographic group d in some set of demographics $\mathcal{D}$, and then finding the worst-case differentials from the maximum difference in FMR, and FNMR, across demographics

$$\text{A}(\tau) = \max_{d_i} \text{FMR}_{d_i}(\tau) - \min_{d_j} \text{FMR}_{d_j}(\tau) \quad (1) \qquad \text{B}(\tau) = \max_{d_i} \text{FNMR}_{d_i}(\tau) - \min_{d_j} \text{FNMR}_{d_j}(\tau) \quad (2)$$

These two measures were combined into a Fairness Discrepancy Rate (FDR), which we discuss later in section 2.6.

Note that the $\text{A}(\tau)$ difference approximates the maximum FMR when the minimum value is orders of magnitude below the maximum value, (as is often the case in face recognition) such that $\text{A}(\tau) \approx \max_{d_i} \text{FMR}$. For example, Figure 1 shows that for one highly accurate commercial prototype, FMR spans several orders of magnitude, in particular FMR

in Nigerian women over 65 is $10^{-1.55} = 0.03$, and $10^{-4.41} = 0.00004$ in Polish men aged 35-50, so equation 1 reduces to $\mathrm{A}(\tau) \approx \max_{d_i} \mathrm{FMR}_{d_i}(\tau)$. This matters when comparing algorithms: If algorithm P had $A = 10^{-2} - 10^{-5}$ and Q had $A = 10^{-2} - 10^{-4}$ then the A values are almost the same and this hides that Q produces a factor of ten more false matches in the best-case demographic than does P, and this could necessitate Q being configured with a high threshold to limit FMR to 10^{-5}.

2.2 Ratio of worst and best case error rates

NIST proposed an alternative to the Idiap difference measure by employing ratios

$$\mathrm{INEQUITY}_{nm}(\tau) = \frac{\mathrm{FMR}_{d_n}(\tau)}{\mathrm{FMR}_{d_m}(\tau)} \tag{3}$$

as this will accomodate the large range of variation in FMR and because it has a clear operational meaning, namely the number of times more likely it is to confuse two persons belonging to one demographic group versus another.

Specific ratios will be of interest to developers seeking to address specific inequities by altering a training procedure or by employing additional image data. Likewise, end-users may have interest in specific groups. In the results section later we state ratios specific to sex, age, and certain geographically-defined groups. For example, given that poor photography can lead to underexposure of faces and depressed mated similarity scores [2], most immediately those of dark-skinned subjects, the ratio of FNMR for say African vs. East European faces may be informative.

We then find worst-to-best case FMR and FNMR ratios,

$$\mathrm{A}(\tau) = \frac{\max_{d_i} \mathrm{FMR}_{d_i}(\tau)}{\min_{d_j} \mathrm{FMR}_{d_j}(\tau)} \quad \forall d_i, d_j \in \mathcal{D} \tag{4}$$

$$\mathrm{B}(\tau) = \frac{\max_{d_i} \mathrm{FNMR}_{d_i}(\tau)}{\min_{d_j} \mathrm{FNMR}_{d_j}(\tau)} \quad \forall d_i, d_j \in \mathcal{D} \tag{5}$$

where the $\max$ over $\min$ formulation expresses the worst-case to best-case error rates, and has lower-is-better semantics. One criticism of these measures is that the denominator values could be zero, a possibility that would be more likely in small tests and when the threshold is high (pushing FMR$\rightarrow 0$) or low (pushing FNMR$\rightarrow 0$). This concern could be addressed by changing the τ values, or by including an additive constant $\epsilon > 0$ in the denominator; this is discussed further in the next section.

The combination of these two quantities is discussed later, in section 2.6.

2.3 Ratios normalized by the mean

While worst-case error rate excursions are likely what algorithm designers should mitigate, worst-to-best case formulations (eqs. 4-21) are arguably non-robust because the maximum and minimum are potentially not robust. One alternative is to express the worst-case error rate relative to a mean (or a weighted mean - see section 2.7 below). For example,

$$\mathrm{A}(\tau) = \frac{\max_{d_i} \mathrm{FMR}_{d_i}(\tau)}{\mathrm{FMR}^{\diamond}} \quad \forall d_i \in \mathcal{D} \tag{6}$$

$$\mathrm{B}(\tau) = \frac{\max_{d_i} \mathrm{FNMR}_{d_i}(\tau)}{\mathrm{FNMR}^{\diamond}} \quad \forall d_i \in \mathcal{D} \tag{7}$$

where the $\diamond$ superscript connotes the arithmetic mean of $n = |\mathcal{D}|$ values:

$$x^{\diamond} = n^{-1} \sum_i x_i \tag{8}$$

A better variant[5] replaces the arithmetic mean with the geometric mean:

$$x^{\dagger} = \left(\prod_i x_i \right)^{1/n} \tag{9}$$

which gives the inequity measures

$$\mathrm{A}(\tau) = \frac{\max_{d_i} \mathrm{FMR}_{d_i}(\tau)}{\mathrm{FMR}^{\dagger}} \;\; \forall d_i \in \mathcal{D} \qquad (10) \qquad\qquad \mathrm{B}(\tau) = \frac{\max_{d_i} \mathrm{FNMR}_{d_i}(\tau)}{\mathrm{FNMR}^{\dagger}} \;\; \forall d_i \in \mathcal{D} \qquad (11)$$

with the advantage that it captures values spanning several decades. For example given $x = \{0.1, 0.0001\}$, the arithmetic mean $x^{\diamond} \approx 0.05$ does not reflect the variation captured by the geometric mean, $x^{\dagger} \approx 0.003$. This has a graphical interpretation: When looking at an error-tradeoff characteristic with FMR plotted on a log scale, the visual process of averaging FMR between two groups is actually an estimate of the geometric mean because the arithmetic mean of the log is the log of the geometric mean. Note that the geometric mean is tolerant of very large entries: for example the two sets $\{0.5, 0.75, 1, 1.333, 2, 40, 40\}$ and $\{0.5, 0.75, 1, 1.333, 2, 2, 800\}$ have the same geometric mean (2.87).

A possible problem with the method is that if any individual error rate is zero, the geometric mean will be zero also. A numeric remedy for this would be to disallow zero error rates via

$$x^{\dagger} = \left(\prod_i (x_i + \epsilon) \right)^{1/n} \tag{12}$$

which could be set to some "typical" low values (for FMR, $\epsilon = 10^{-7}$, for FNMR, $\epsilon = 10^{-5}$ say). However, the result of the computation sensitive to ϵ, especially for small n. The preferred approach, which we use throughout is to set $\epsilon = 0$ and replace x_i with the lowest value that is statistically sustainable given a finite number of trials - see section 4.2 .

2.4 Measures of error rate heterogeneity

Alternative approaches consider distribution of errors across demographic groups. One such, from researchers at SAIC and DHS Science and Technology [6], leverages the well-known Gini coefficient that has been used for many years as a summary of wealth and income disparity.

$$\mathrm{A}(\tau) = \frac{\sum_i \sum_j |\mathrm{FMR}_{d_i}(\tau) - \mathrm{FMR}_{d_j}(\tau)|}{2n^2 \, \mathrm{FMR}^{\diamond}} \tag{13} \qquad \mathrm{B}(\tau) = \frac{\sum_i \sum_j |\mathrm{FNMR}_{d_i}(\tau) - \mathrm{FNMR}_{d_j}(\tau)|}{2n^2 \, \mathrm{FNMR}^{\diamond}} \tag{14}$$

where the denominator includes the number of demographic groups $n = |\mathcal{D}|$, and the arithmetic mean (eq. 8). In this paper we modified this classic definition to use $n(n-1)$ in the denominator to render the estimator unbiased. This yields Gini values on $[0, 1]$, with higher values associated with unfair concentration of errors in a few demographics. The combination of these into an overall difference measure is discussed in section 2.6.

Another method, on intersectional weighted inequity[6], measures spread about the geometric mean:

$$\mathrm{A}(\tau) = \sum_{d \in \mathcal{D}} \left| \log_{10} \frac{\mathrm{FMR}_d(\tau)}{\mathrm{FMR}^{\dagger}(\tau)} \right| \tag{15} \qquad\qquad \mathrm{B}(\tau) = \sum_{d \in \mathcal{D}} \left| \log_{10} \frac{\mathrm{FNMR}_d(\tau)}{\mathrm{FNMR}^{\dagger}(\tau)} \right| \tag{16}$$

[5] Pierre Gacon (Idemia) contribution to ISO/IEC 19795-10 *Quantifying biometric system performance variation across demographic groups* in Working Group 5 of ISO/IEC JTC 1/SC 37.

[6] Greg Cannon, personal communication, 2021-08-11 and comment contributed to ISO/IEC 19795-10 *Quantifying biometric system performance variation across demographic groups* in Working Group 5 of ISO/IEC JTC 1/SC 37.

where the absolute value acts to treat ratios above and below one equally. The measure takes on a larger value when FMR and FNMR vary a lot. As in the last section, the geometric mean can go to zero - see the discussion in section 4.2. Additionally, if any error rate in the numerator is zero, the measure is undefined.

2.5 Ratios relative to a nominal error rate

A simple alternative[7] to the above ratios, one that solves the zero denominator problem, is to to normalize by some nominal target or reference value. For example, the worst-case error ratio (eqs. 10-11) would become

$$\mathrm{A}(\tau) = \frac{\max_{d_i} \mathrm{FMR}_{d_i}(\tau)}{\mathrm{FMR}_{\mathrm{REF}}(\tau)} \;\; \forall d_i \in \mathcal{D} \qquad (17) \qquad\qquad \mathrm{B}(\tau) = \frac{\max_{d_i} \mathrm{FNMR}_{d_i}(\tau)}{\mathrm{FNMR}_{\mathrm{REF}}(\tau)} \;\; \forall d_i \in \mathcal{D} \qquad (18)$$

Alternatively, the numerator could be Idiap's maximum minus minimum value (eqs. 1-2).

The reference value could be set in data-dependent way - perhaps as an empirical value from a general test or a test on one demographic group - or in a data-independent way - such as a value asserted by a manufacturer, or simply a value specified in an operational requirement. For example, if an automated border control gate is assumed to have a FMR of 0.0001, then empirical difference could be normalized by that value. This method avoids divide-by-zero issues (discussed later in sec. 4.2). We don't further analyze and implement this normalization.

2.6 Combining FMR and FNMR differentials

This section advances methods for combining the A and B meassure of the prior sections into one overall equity measure.

Idiap formed a Fairness Discrepancy Rate from the quantities in equations 1 and 2 using the weighted sum.

$$\mathrm{FDR}(\tau) = 1 - (\alpha A(\tau) + (1-\alpha)B(\tau)) \qquad (19)$$

where the leading $1 - x$ complement yields the larger-is-better semantics of an equity measure rather than an inequity one. As formulated, the value of FDR is limited because $\mathrm{A}(\tau)$ and $\mathrm{B}(\tau)$ are often orders of magnitude apart, an appropriate re-weighting would need to be incorporated in α requiring $\alpha \to 0$ or $\alpha \to 1$.

The Idiap team elected to define a Fairness Discrepancy Rate with higher is better semantics. The following summaries are lower-is-better inequity measures[8].

First, **NIST** suggested A and B values (eqs. 4-5) can be combined into a joint "number-of-times-more-errors" inequity measure using

$$\mathrm{INEQUITY}(\tau) = A(\tau)^{\alpha}\, B(\tau)^{\beta} \qquad (21)$$

where a low value for α or β can be used to unweight either FMR or FNMR differentials.

All of the contributors to ISO/IEC 19795-10 suggested combination via a weighted sum

$$\mathrm{INEQUITY}(\tau) = \alpha\mathrm{A}(\tau) + \beta\mathrm{B}(\tau) \qquad (22)$$

SAIC termed their weighted sum of Gini estimates the "Gini Aggregation Rate for Biometric Equitability (GARBE)" [6].

[7] Due to Yevgeniy Sirotin, personal communication 2022-02-08.

[8] All the lower-is-better inequity measures can trivially be recast as higher-is-better equity measures via simple inversions. For example, the NIST product could be changed to a higher-is-better equity measure by negating logarithms,

$$\mathrm{EQUITY}(\tau) = -\alpha \log_{10} A(\tau) - \beta \log_{10} B(\tau) \qquad (20)$$

which would also reduce its numerical range.

The parameters α and β can be used to appropriately weight the relative importance of false matches and false non-matches. This is essential because the false positive and false negative errors have markedly different impacts in most biometric applications. In addition, their frequency will depend also on the prior probabilities, respectively, of non-mate (impostor) and mate (genuine) comparisons. For example, in unlocking a mobile phone, almost all transactions are mated, whereas searches in a casino watchlist application would include many non-mate comparisons. In this latter example, setting $\beta = 0$ would quantify FMR differentials and entirely disregard FNMR differentials. It may be useful but is not necessary to set $\beta = 1 - \alpha$. The use of the $()^{\alpha}.()^{1-\alpha}$ product in equation 21 allows FMR and FNMR to exist on different ranges. If, as is typical, FMR spans several orders of magnitude and FNMR spans much less than one, and FNMR is more critical operationally because impostors are very rare, α can be set to a small value.

We do not investigate further the combined summary measures (eqs. 19, and 21 or 20) as combination will always be application-specific: For example, the relative importance of FMR and FNMR are vastly different in access control where impostors are rare, and in soccer-stadium watchlist application where most searches are not enrolled in the system. In any case, the combination step is not neccessary for our purposes, and additionally we consider the combinations of $\mathrm{A}(\tau)$ and $\mathrm{B}(\tau)$ are more abstract and thereby detract from the intuitive value of the two parts alone. The individual quantities are themselves meaningful, informative and more actionable.

2.7 Weighting demographic groups

The methods could be extended to allow weighting of each demographic group. For example eq. 15 would be

$$\mathrm{A}(\tau) = \sum_{d \in \mathcal{D}} \left| \log_{10} \frac{u_d \, \mathrm{FMR}_d(\tau)}{\mathrm{FMR}^{\dagger}(\tau)} \right| \tag{23}$$

with the geometric means likewise being weighted

$$x^{\dagger} = \exp\left(\frac{\sum_i u_i \log x_i}{\sum_i u_i} \right) \tag{24}$$

An immediate candidate for a weighting policy would be to assign low u and v values when the FMR or FNMR estimates have high uncertainty due to low sample size. This may be injurious to an under-represented demographic that had a large error ratio (as is typical with imbalanced training sets) as discounting it solely on the basis of limited amount of data is likely retrograde. We don't further analyze or advocate for particular demographic group weighting strategies.

3 Analysis of the measures

This section discusses the advantages and disadvantages of the summary methods given in section 2. To do that we compare bahaviors on various elemental datasets appearing in the rows of Table 2. We also consider desirable properties of the measures. The first three identifed by Howard et al. [6] apply to overall equity measures (combinations of A and B) and are termed Functional Fairness Measure Criteria (FFMC). Paraphrasing, these are:

- ▷ **FFMC.1** - The net contributions of FMR and FNMR differentials to the overall fairness measure should be intuitive when using a normal range of risk parameter weights and operationally relevant error rates.
- ▷ **FFMC.2** - There should be recognizable points of reference in the domain of the fairness measure, e.g. one bounded by known minimum and maximum possible values.

Criterion	Max-Min	Max/Min	Vary/GeoMean	Max/GeoMean	Gini
FFMC.1	N	Y	Y	Y	Y
FFMC.2	2	1	1	1	2
FFMC.3	Y	N	N	N	Y
FFMC.4	N	N	N	N	N

Table 1: Compliance of the proposed error-rate differentials against the candidate functional fairness measure criteria. For FFMC.2, the number indicates the number of reference points - for example, for a measure confined to [0,1] the value is 2; for a ratio whose ideal value is 1, the number is 1. No measure succeeds on FFMC.4 and this is discussed separately. FFMC.5 is not amenable to tabulation and is discussed in the text.

- ▷ **FFMC.3** - The fairness measure should be calculable when no errors are observed for a demographic group. Given a finite image dataset partitioned into intersectional demographic groups, the likelihood that one group has zero FNMR rises with the number of groups.

We add to these two criteria to support algorithm comparison:

- ▷ **FFMC.4** - The measure should reward more accurate algorithms if they distribute errors uniformly or in the same way as less accurate ones.
- ▷ **FFMC.5** - The measure should rank algorithms intuitively, correctly penalizing algorithms with the most non-uniform error rates.

Table 1 compares the inequity measures against the FFMC criteria. FFMC.1 holds for all measures except, the Idiap difference, which is not of itself interpretable. FFMC.2 is fully implemented by the [0,1] bounded Max-Min and Gini measures, and partially otherwise with 1 being a reference point for a ratio. FFMC.3 compliance is not achieved for those ratio methods where a zero denominator is possible (but avoided per sec. 4.2). FFMC.4 is never met: None of the measures automatically reward more accurate algorithms; this drawback motivates the mechanisms to visualize *both* accuracy and inequity discussed in section 4.4. The intent of FFMC.5 is addressed next.

Refering to Table 2a, consider algorithms A and B which achieve a target FMR of 0.001 for all demographic groups except one: namely algorithm A makes 10 times fewer false matching errors than the target, while algorithm B gives 10 times more on Group 1. The fairest algorithm is the one that achieves uniform FMR for all groups. The summary measure should favor algorithm A, which makes no more false matches than the target for all groups and is actually more accurate for group 6. Algorithm B in inferior in its high false match rate on group 1. The proposed measures don't always prefer A over B. Particularly the Max/Min measure (eq. 4) and the variance around the geometric mean measure (eq. 15) assign the same value to algorithms A and B, i.e. 10 and 0.278 respectively. The other measures correctly favor algorithm A over B.

Now compare algorithms C and D. We consider C to be fairer than D because D has more high values. The Max-Min (eq. 1) essentially doesn't quantify this because the minimum is too small. The Max/Min (eq. 4) behaves correctly but given the A vs. B result, the other measures are more interesting. We note Max/Mean (eq. 6) fails to order correctly, but GeoVary (eq. 15), Max/GeoMean (eq. 10) do. Gini (eq. 13) is also effective, but the coefficient values differ only in the third decimal place.

Rows E and F exhibit the problem of using the mean in the denominator for a variable that can span several order of magnitude. Algorithms E and F are identical except in group 6, where FMR is three and five orders of magnitude lower than for all other groups. This is typical in actual data - see Figure 6. In rows E and F, the two measures with the arithmetic mean in the denominator - Max/Mean (eq. 6) and Gini (eq. 13) - essentially do not differentiate between E and F.

The last paragraphs deal with FMR. The following discussion considers FNMR which is usually distributed over fewer orders of magnitude - for example, see Figure 2. Table 2b includes various FNMR distributions across six demographic

(a) Comparing FMR Inequity Measures. For six algorithms giving FMR on six demographic groups, the table shows how the proposed summary measures quantify inequity

Alg	FMR1	FMR2	FMR3	FMR4	FMR5	FMR6	Max-Min	Max/Min	GeoVary	Max/Mean	Max/GeoMean	Gini
A	0.001	0.001	0.001	0.001	0.001	1e-04	0.001	1e+01	0.278	1.176	1.468	0.176
B	0.010	0.001	0.001	0.001	0.001	1e-03	0.009	1e+01	0.278	4.000	6.813	0.600
C	0.010	0.001	0.001	0.001	0.001	1e-04	0.010	1e+02	0.333	4.255	10.000	0.702
D	0.010	0.002	0.001	0.001	0.001	1e-05	0.010	1e+03	0.628	3.997	13.077	0.706
E	0.001	0.001	0.001	0.001	0.001	1e-06	0.001	1e+03	0.833	1.200	3.162	0.200
F	0.001	0.001	0.001	0.001	0.001	1e-08	0.001	1e+05	1.389	1.200	6.813	0.200

(b) Comparing FNMR Inequity Measures. For five algorithms giving FNMR on six demographic groups, the table shows how the proposed summary measures quantify inequity

Alg	FNMR1	FNMR2	FNMR3	FNMR4	FNMR5	FNMR6	Max-Min	Max/Min	GeoVary	Max/Mean	Max/GeoMean	Gini
a	0.03	0.03	0.03	0.03	0.03	3e-02	0.00	1e+00	0.000	1.000	1.000	0.000
b	0.01	0.02	0.03	0.04	0.05	6e-02	0.05	6e+00	0.217	1.714	2.004	0.333
c	0.01	0.01	0.01	0.06	0.06	6e-02	0.05	6e+00	0.389	1.714	2.449	0.429
d	0.03	0.03	0.03	0.03	0.03	1e-04	0.03	3e+02	0.688	1.199	2.587	0.199
e	0.03	0.03	0.03	0.03	0.03	1e-08	0.03	3e+06	1.799	1.200	12.009	0.200

Table 2: Example cases supporting comparison of the inequity measures

groups. The first algorithm, "a", gives uniform FNMR, so all measures take on their ideal value. Comparing algorithms "b" and "c", we favor algorithm "b" as the FNMR for most of the groups is closer to the mean (which is 3). Metrics Max/Min and Max/Mean fail to acknowledge this, while the variance around the geometric mean, the Max/GeoMean and the Gini coefficient agree.

Examples 'd' and 'e' show that if an algorithm achieves a very good FNMR for one group, and is identical for all other groups, then the new metric would give a bad fairness evaluation. This should probably not be the case. Hence, for the FNMR as well, the metric should be protected against very low values of FNMR.

4 Results and Discussion

We report demographic summary measures for algorithms submitted to the 1:1 track of the Face Recognition Vendor Test. The results are reported in tables in the False Positive Demographics and False Negative Demographics tabs on this webpage. Tables with just the max-over-geometric mean appear on this page.

The tables are documented in brief there, and more completely in this document. For algorithm comparison, and for comparison of demographic groups, we set a threshold for each algorithm to give a nominal FMR on a fixed dataset. We use one threshold for FMR tables, and a somewhat higher threshold for FNMR tables.

4.1 Datasets

The following describes datasets used in the tabulated results.

- The **false negative** results are computed from 1 442 511 genuine scores produced by each algorithm when comparing high quality, visa-like, frontal portraits with medium quality primary inbound airport immigration line photos captured with a flexible-mount web camera. There are 827 550 genuine scores from women, 602 613 from men. The individuals have a place of birth listed as one of 22 countries[9] which we assign to 7 regions. We have age metadata for the two photos, and we associate a comparison with the mean of the age in the two photos, and bin this in to one of five age groups: $(12, 20], (20, 35], (35, 50], (50, 65], (65, 99])$.

[9] E. Europe (Poland, Russia, Ukraine), C. America (Mexico, El Salvador, Nicaragua), W. Africa(Nigeria, Liberia, Ghana), Caribbean (Haiti, Jamaica), E. Africa(Kenya), S. Asia(Iran, Iraq, India, Pakistan), and E. Asia(Vietnam, Phillippines, Korea, Japan, Thailand, and China). Other region assignments are possible.

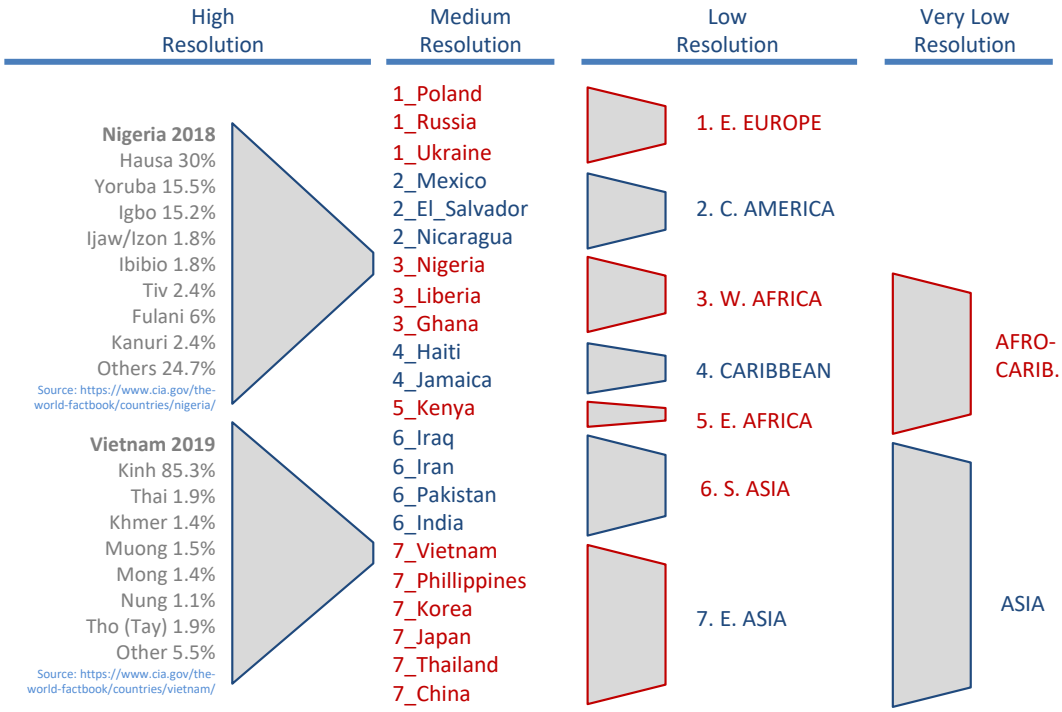

Figure 3: Countries and regions used in quantifying demographic dependence on race. The fine-grained or local ethnicities shown at left are not available to us. The rightmost grouping is possible but not useful.

- ▷ The **false positive** results are computed from a subset of 195 billion impostor scores produced by comparing disjoint sets of 442 019 and 441 517 high quality, visa-like, frontal portraits. We have country-of-birth (22 countries), sex ("M", "F") and age-group information ($(12, 20], (20, 35], (35, 50], (50, 65], (65, 99]$). We have false match rate estimates for comparison of individuals from all possible pairs of demographic groups (e.g. Japanese males, 20-25 with Kenyan females over 65) - these are depicted as a heatmap plotting $\log_{10}$FMR for each algorithm (see, for example, the large dimension PDF files: NTechLab-11 and Megvii-4). Figures 1 and, in Annex A, Figures 6-8, are interesting extracts from that larger matrix. The matrix itself has a role in modeling one-to-many performance - see Annex B.

The analysis of race in this report is enabled by using country-of-birth as a proxy. The countries are listed in Figure 3. This proxy is imperfect in two ways: First, it ignores local ethnic variations, shown in the figure for Nigeria and Vietnam, which may be germane to face recognition, but metadata for which is unavailable to us. Second, some part of the population will have trans-national ancestry. That is clearly true of the USA, UK and France for example, which is why those countries are not considered.

We group the 22 countries into 7 regions, giving a "lower resolution" label of race, as shown in Figure 3. We do this because there are considerable false matches between these countries, as shown by the block diagonal structure of Figures 6 and 7. This analysis demonstrates that within-country false match rates are not significantly influenced by undetected identity fraud. This could apply cross-border too but we note that many high false match rates exist between countries that have no border and no common language. The result is that we work with false match rates like those depicted in Figure 8.

All FMR inequity measures in the report are generated from same-sex, same-age-group and same-region individuals.

We compute regional-FMR as follows. Given the number of false positives NFP_{xyas} produced for comparing subjects from countries x and y in region r, and in age group $a \in \mathcal{A}$, with sex $s \in \mathcal{S}$, and similarly the number of impostor

comparisons NIMP_{xyas}, the regional rate is simply the FMR estimate as if we didn't have country information:

$$\text{FMR}_{ras} = \frac{\sum_{x\in r}\sum_{y\in r}\text{NFP}_{xyas}}{\sum_{x\in r}\sum_{y\in r}\text{NIMP}_{xyas}} \tag{25}$$

Thus, we estimate the FMR within Eastern Europe from intra- and inter-country comparisons of Polish, Russian, and Ukranian photos, and we do that separately for each sex and age group. With these estimates we do not consider individual countries further, but note that obviously some countries have higher FMR than this mean. We drop the Caribbean region from the tables because the population is relatively small. We also drop E. Africa because it only includes one country (Kenya).

We compute regional-FNMR as follows. Given the number of false negatives NFN_{cas} produced for subjects from country, c, in age group $a \in \mathcal{A}$, with sex $s \in \mathcal{S}$, and similarly the number of genuine comparisons NGEN_{cas}, the region r FNMR estimate is computed as if we didn't have country information:

$$\text{FMR}_{ras} = \frac{\sum_{c\in r}\text{NFN}_{cas}}{\sum_{c\in r}\text{NGEN}_{cas}} \tag{26}$$

4.2 Restating error rates at or near zero

Some of the inequity measures are sensitive to low error rate values (in their denominators). This section advances a mitigation method.

Given a finite number of comparisons, the estimated error rates are uncertain. This is of concern in tests of demographic effects because while the total available population may be large, the number of individuals for any intersectional demographic can be much smaller. For N people, and K demographic factors (for example, $K = 3$ for age, race and sex), and b_k partitions for each group, the expected number of individuals in any one group would be $N/\prod b_k$ with a minimum below that when, inevitably, the population is not balanced.

When population sizes are small, the number of observed errors will be small, particularly for accurate recognition algorithms. We are concerned with an upper bound on the error rate. We seek, for confidence level α, an error rate that is sustained by the observation of x errors in the n comparisons executed. From Binomial theory[10] we can invert the Binomial's cumulative distribution function,

$$P(X \le x) = 1 - I_p(x+1, n-x) = 1 - \alpha \tag{27}$$

via the incomplete beta function, I_p, to give an upper bound p_u above the point estimate $p = x/n$. This is the quantile of the beta function that we compute numerically. For x = 0, p_u = 3/n which is the *Rule-of-Three* special case: Given 0 errors in n trials, the lowest sustainable error rate claim, at 95% confidence[11], is 3/n. For $x > 0$, the formula increases p by a factor that decreases to 1 as $x \to n$.

In all our results we do the following: In the numerators we use $p = x/n$ which is the best estimate of error rate. In the denominators, we use p as the inverse of equation 27 which is the lowest sustainable rate given n. This has the effect of removing zeroes in the denominators and of slightly decreasing the various inequities. We do this for the geometric means, and also for the arithmetic means of eqs. 6, 7 including Gini eqs. 13, 14, and in the minimums of eqs. 1, 2.

This technique has the disadvantage that we inherit a sample size dependence: smaller samples lead to larger corrections, larger denominators, and reduced inequities. This does not compromise comparison of algorithms within a test but could undermine comparison of tests across laboratories.

[10] See section 2.2 in Scholz [11]

[11] The rule is 4.6/n for 99% confidence, 6.9/n for 99.9% confidence - numerator is $-\ln(1-\alpha)$ generally

Note that in all our trials FMR and FNMR are rarely 0, given the thresholds we use, and that we accumulate errors across countries into their respective regions.

4.3 Ratios of specific demographic interest

From the point FMR estimates of eq. 26 we formulate an overall female-to-male FMR differential as a geometric mean of ratios

$$\mathrm{A}_{sex} = \left(\prod_{r\in\mathcal{R},a\in\mathcal{A}} \frac{\mathrm{FMR}_{raF}}{\mathrm{FMR}_{raM}} \right)^{1/n} \tag{28}$$

where $n = |\mathcal{R}|\,|\mathcal{A}|$ is 25 (5 regions, 5 age groups).

Similarly, we express older-to-younger FMR differentials as a ratio

$$\mathrm{A}_{age} = \left(\prod_{r\in\mathcal{R},s\in\mathcal{S}} \frac{\mathrm{FMR}_{rOs}}{\mathrm{FMR}_{rYs}} \right)^{1/n} \tag{29}$$

where $n = |\mathcal{R}|\,|\mathcal{S}|$ is 10 (5 regions, 2 sexes), and the the subscripts "O" and "Y" denote older $(65-99]$ and $(20-35]$ respectively.

We also produce three region-of-birth FMR ratios by dividing their FMR by that for East Europe as follows

$$\mathrm{A}_{\eta\rho} = \left(\prod_{a\in\mathcal{A},s\in\mathcal{S}} \frac{\mathrm{FMR}_{\eta as}}{\mathrm{FMR}_{\rho as}} \right)^{1/n} \tag{30}$$

where subscript ρ indicates Eastern Europe, and η is one of West Africa, South Asia and East Asia, and $n = |\mathcal{A}|\,|\mathcal{S}|$ is 10 (5 age groups, 2 sexes). We denominate these with Eastern Europe recognizing that many algorithms are assumed to be trained on imbalanced, majority white, databases. Note that these four ratios are not adjusted by the Binomial uncertainty correction of section 4.2.

4.4 Visualization

We tabulate the various measures on the demographic tabs of the main FRVT results page. There are two tabs, one each for false positive and negative effects. False positive effects are much larger than false negative effects - see the Figures in Annex A vs. Figure 2. False negatives are in large part due to one or both photographs being of poor quality[12], something that can be coupled with demographics. For example, tall individuals may not be captured with a frontal view, and dark skin has more challenging dynamic range capture requirements, with underexposed facial regions resulting in reduced information available to the algorithm.

The tables include, and can be sorted by, an overall FNMR value, so that the inequity measures can be visualized for the more accurate algorithms. But we recognize that some algorithms, as prototypes, are less accurate so would be less useful unless they offered some other advantage (e.g. speed, or reduced demographic differentials). We therefore plot Figure 4 as a mechanism to show those algorithms that have two desirable properties: Low FNMR *and* better more even FMR across demographics. This Figure, suggested by Idemia, is useful in showing toward the bottom left those algorithms with both properties. Howard et al. [6] include plots that include a Pareto frontier to elicit similar information.

[12]False negatives are caused by any significant change in appearance of a subject between two photos - this can arise to due long-run ageing, acute injury, and most commonly poor image quality.

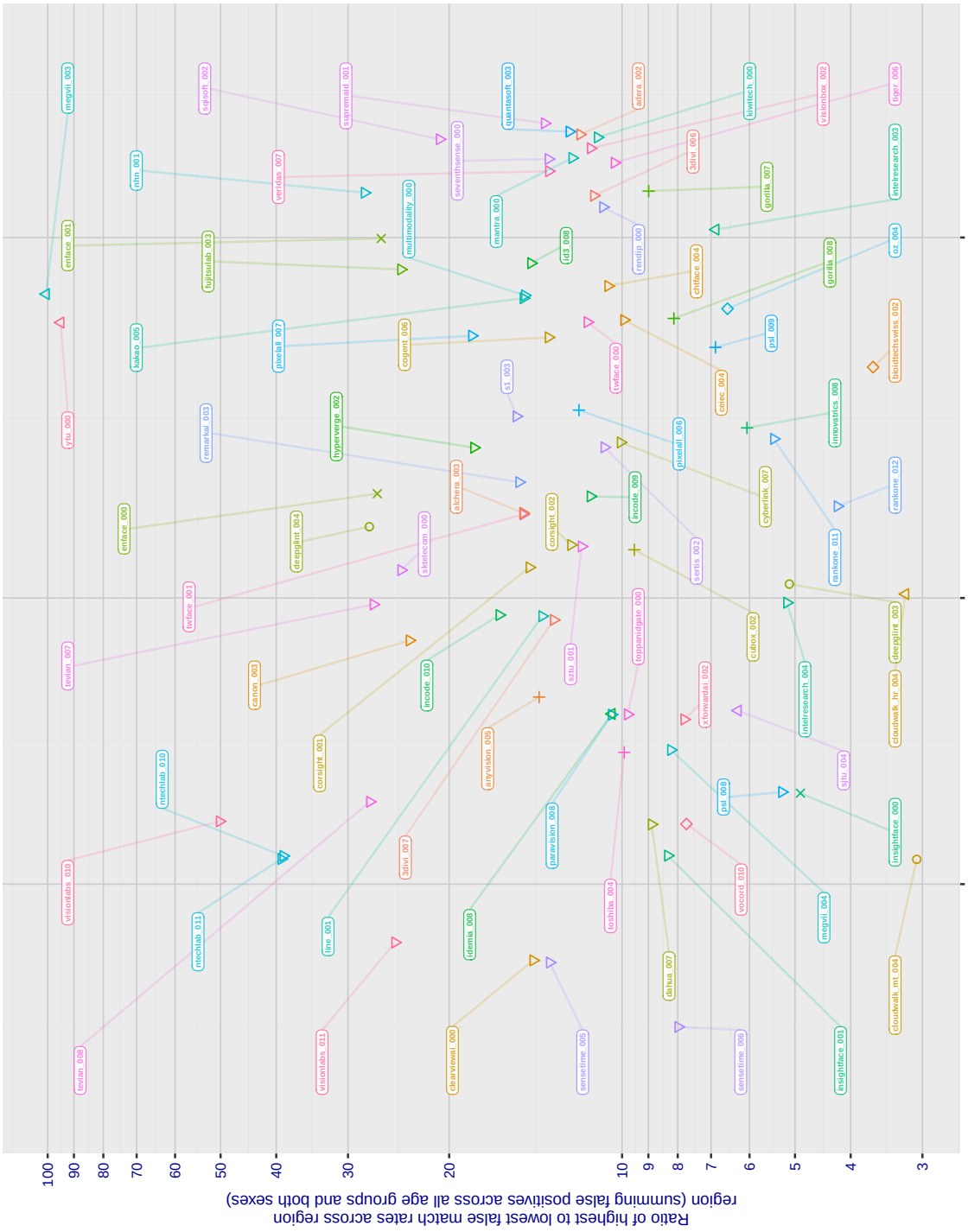

Figure 4: For algorithms submitted to FRVT in 2021, the figure shows FMR-Ratio against a generic FNMR value that is obtained as a mean of FNMR values from four separate FRVT sets identified in the x-axis label. This FNMR value differs somewhat from that used in the generation of FNMR inequity measures, which derives from a different partition of visa-like to border-crossing comparisons.

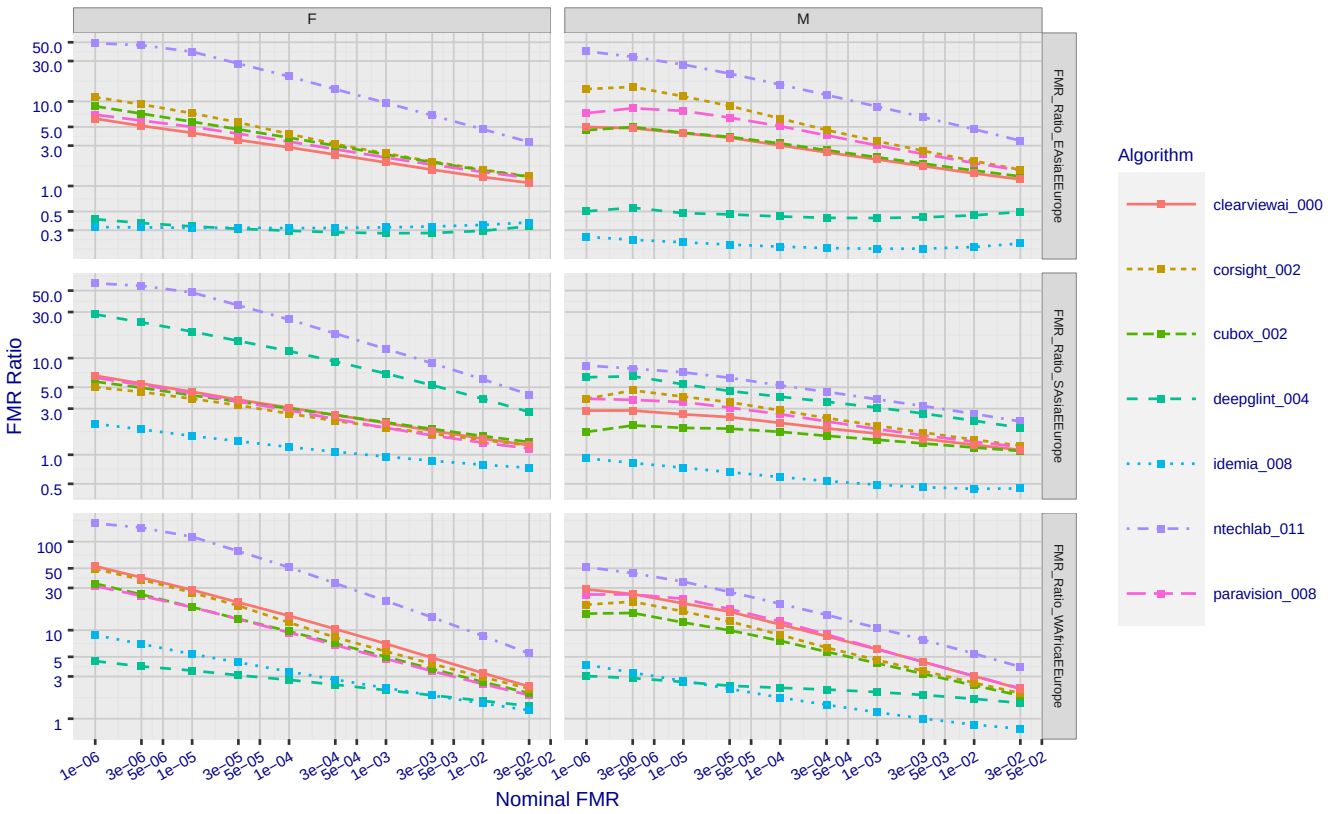

Figure 5: For eight algorithms and both sexes, the panels plot the FMR for three global groups divided by that for E. Europeans against an overall FMR achieved by setting 10 different threshold values. Higher thresholds are on the left side. The ideal values are 1.0. FMR ratios can be below 1 - for a European algorithm (idemia-008) and a Chinese one (deepglint-004) - indicating a lower FMR in an East Asian population than in East Europeans.

4.5 Recognition thresholds

Figures 1, 2. and 8 show false match and non-match rates across demographics at a single threshold value. The magnitude of the demographic differentials change if the threshold is changed. As shown in Figure 5, FMR ratios decrease as threshold is reduced. They necessarily converge toward 1 as FMR in all groups will become 1. We advise setting a fixed threshold for each recognition algorithm to allow for comparison across demographic groups. The threshold affects the absolute magnitude of the ratios, and that information would be needed for a given application.

One means to avoid threshold-dependence is to consider measures of how two whole distributions differ. We suggest such threshold-independent measures in Annex C but do not implement them here.

5 Summary

False negative differentials will yield inequities in those applications where false negatives have material impact. These include access control or more general authorization for access to a resource, in binding some event to a person (e.g. time-and-attendance), and in verification of identity claims. While the results show some variation across algorithms, the more accurate algorithms give lower differentials - low FNMR implies low differences in FNMR. But we also emphasized the importance of false positive differentials, particularly, because their remediation is the role of the algorithm developer, while false negatives can be remediated by better photography, and by using more accurate algorithms.

The various measures have strengths and weaknesses; some are less interpretable than others; some do aggregation that can hide individual effects; likewise averaging measures can obscure large individual demographic effects; and, some will be more sensitive to sample size and will not be statistically tractible if narrow uncertainty bounds are needed. We quote the Max/GeoMean measure (eq. 10) as the leading candidate measure.

We expect to implement similar measures for 1:N identification, particularly to run empirical trials to show how FPIR and FNIR varies across demographic groups.

A Cross-country and cross-region false positive rates

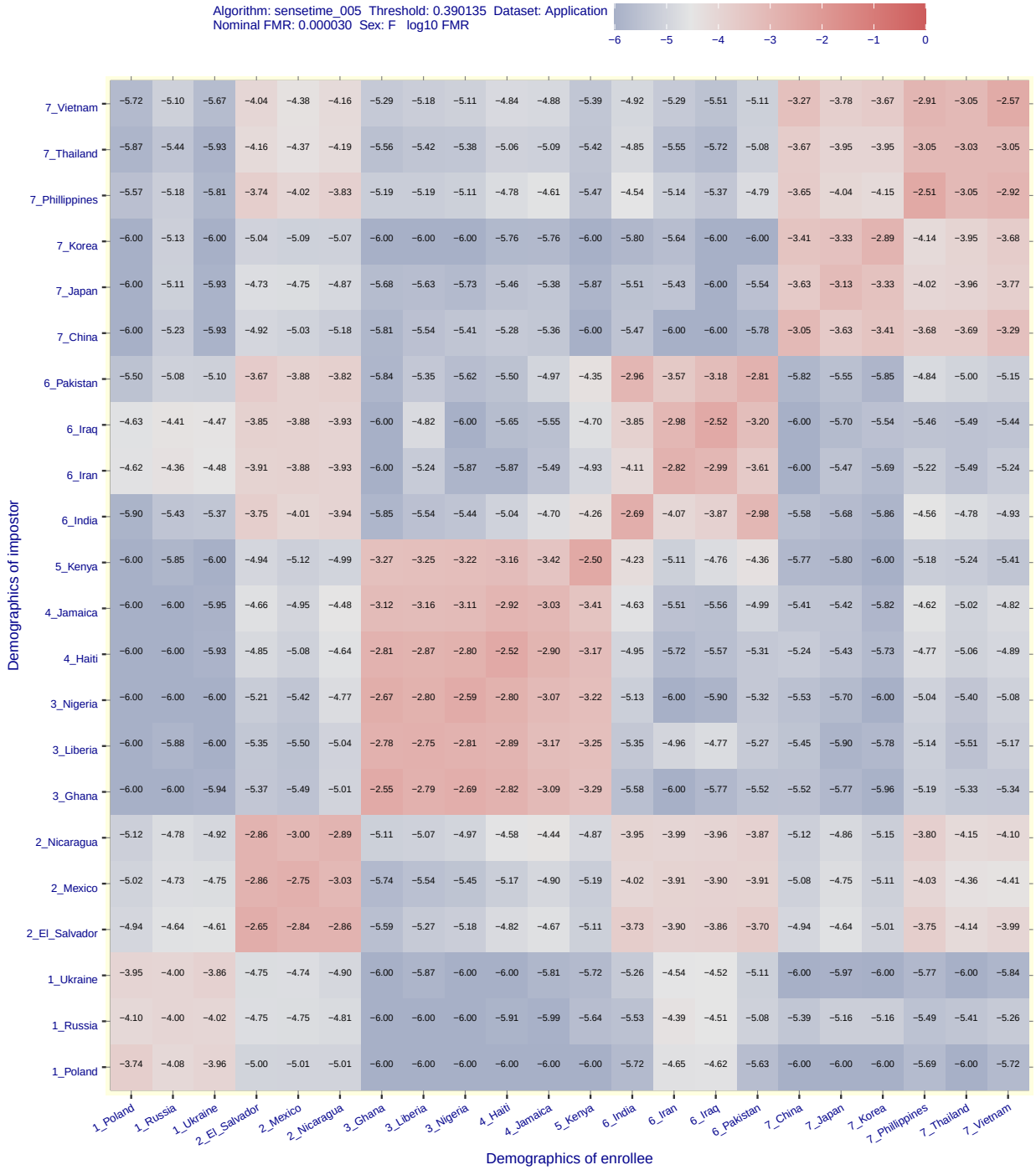

Figure 6: For 22 countries-of-birth the heatmap and its text entries encode base 10 logarithms of false match rates measured when comparing high quality immigration application portraits of different women of the same age group from the two countries given in the axis labels. The algorithm is identified in the legend - similar figures exist in the reports hyperlinked from the algorithm names on the main FRVT results page. The threshold is the same across all cells. Note higher within-country and *within-region FMR, and variation across regions.*

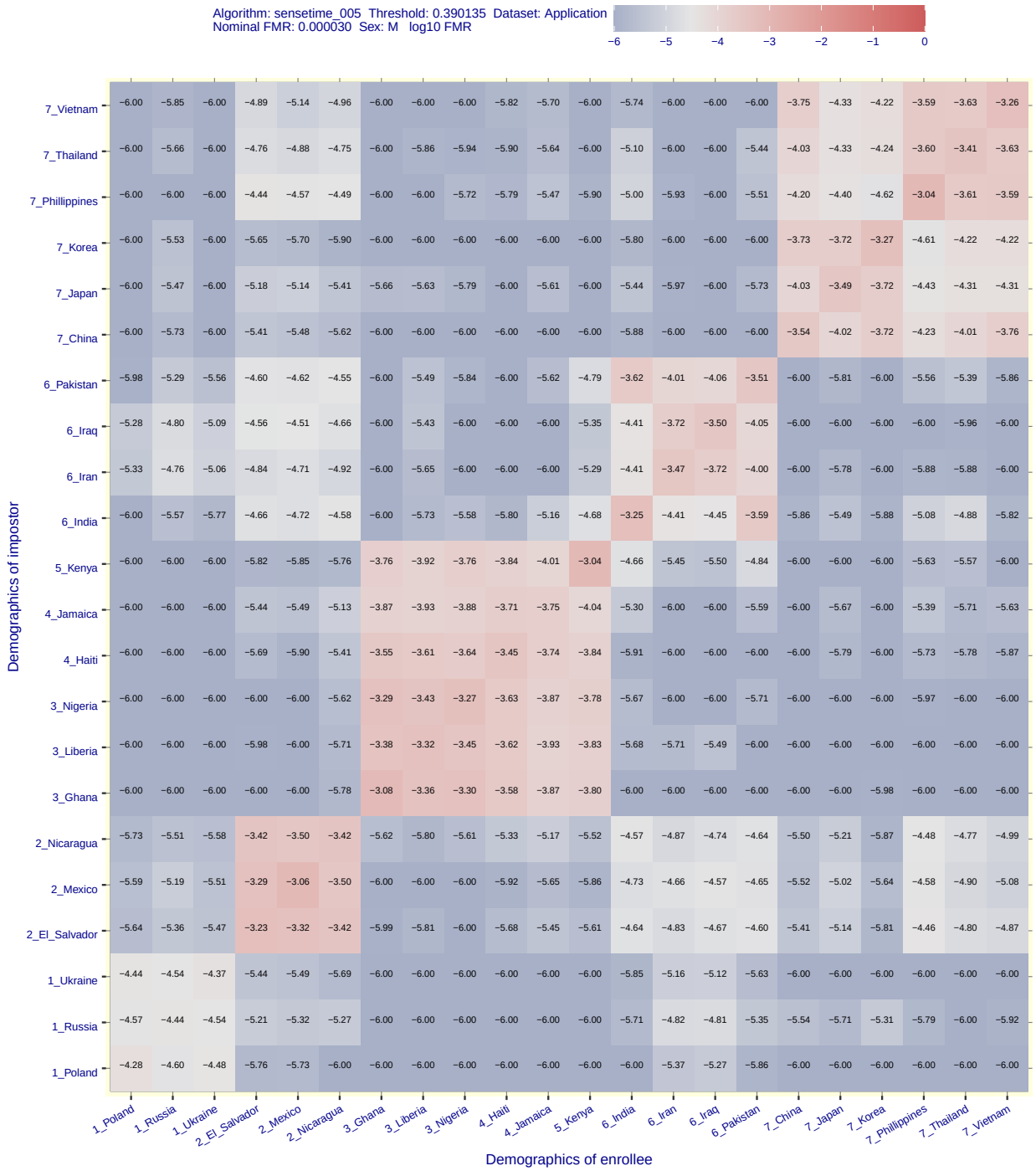

Figure 7: For 22 countries-of-birth the heatmap and its text entries encode base 10 logarithms of false match rates measured when comparing high quality immigration application portraits of different men of the same age group from the two countries given in the axis labels. The algorithm is identified in the legend - similar figures exist in the reports hyperlinked from the algorithm names on the main FRVT results page. The threshold is the same across all cells. Note higher within-country and *within-region FMR, and variation across regions.*

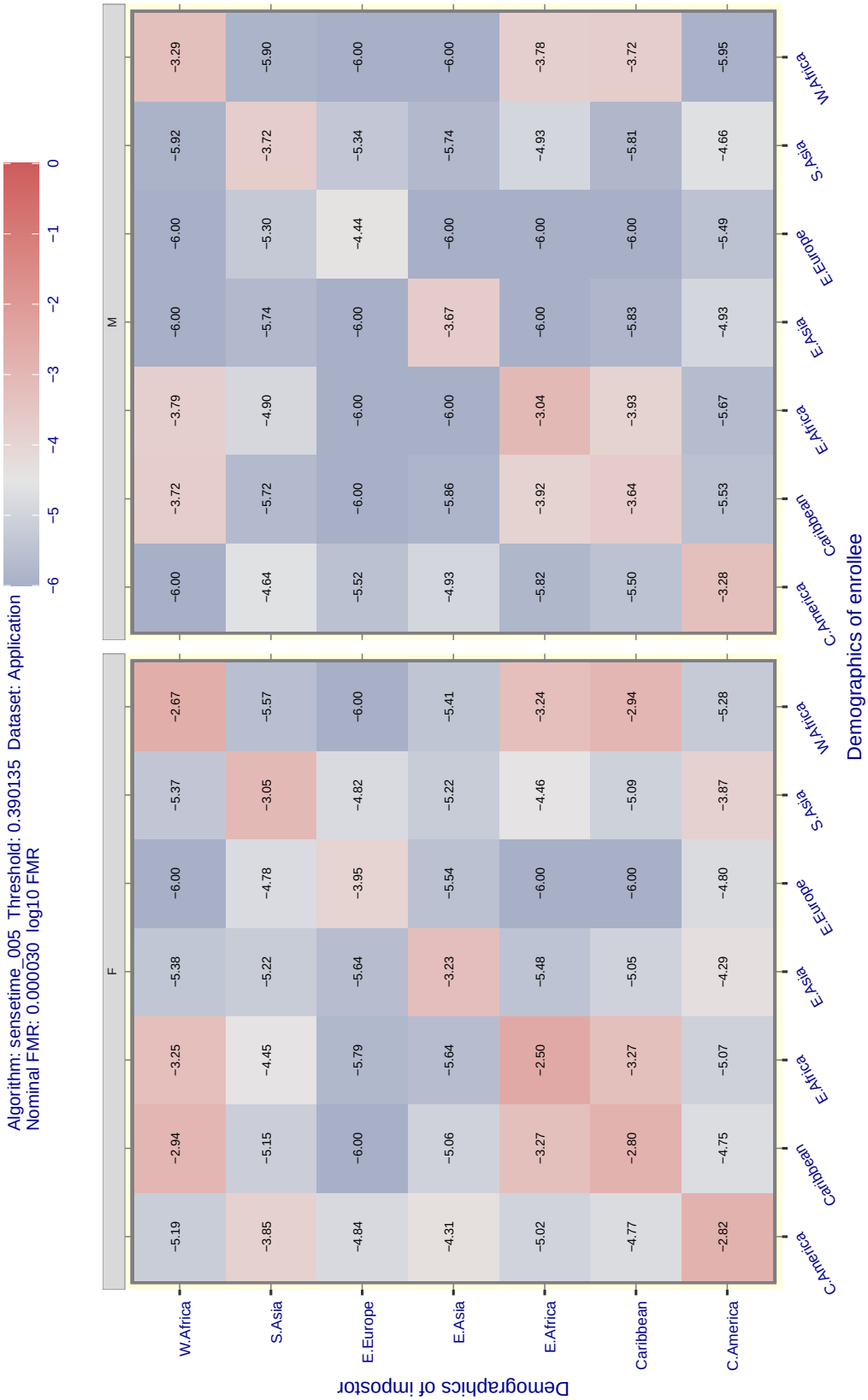

Figure 8: For 8 regions, the heatmap and text entries encode base 10 logarithms of false match rates measured when comparing high quality immigration application portraits of same age group subjects from the regions given in the axis labels. The figure is a variation of the prior two figures, but with FMR now computed over regions instead of countries. As such, this is a lower-resolution consideration of race as an influential demographic variable.

B Relating 1:1 results to 1:N applications

This report includes tabulation of error rates, differentials and summaries for 1:1 face comparison algorithms. This will be important also to that subset of 1:N identification search algorithms that implement search by computing N 1:1 scores, sorting them, and then returning candidate hits if the scores exceed a recognition threshold.

Note that a majority of 1:N algorithms operate in this way. A significant minority however do not[13], such that the binomial model of recognition given below does not apply. In such cases, demographic effects can only be measured empirically by running one-to-many trials - this was done in NIST Interagency Report 8280.

Using the the N 1:1 comparison construct, the following extends the well known Binomial model of false postive identification rate in an N-person gallery, namely that a false positive occurs unless *all* comparisons are below threshold:

$$\text{FPIR}(\tau) = 1 - (1 - \text{FMR}(\tau))^N \tag{31}$$

which is approximately

$$\text{FPIR}(\tau) = N\,\text{FMR}(\tau) \tag{32}$$

at high thresholds for which FMR $\ll N^{-1}$.

The following adapts points made in a presentation by Sirotin et al. at the March 2021 EAB Demographics Conference and then openly published [7] and then re-iterated by others [4]. Others have previously considered heterogeneous false match rates in identification systems [12].

Given demographic groups i and j and estimates for false match rate, $\text{FMR}_{ij}(\tau)$, for comparison of samples from those groups, at threshold τ, we estimate one-to-many false positive identification rate for group i for a enrollment database comprised of n_j samples from demographic groups $1 \leq j \leq J$

$$\text{FPIR}_i(\tau) = 1 - \prod_j (1 - \text{FMR}_{ij})^{n_j} \tag{33}$$

where the matrix FMR_{ij} expresses cross-demographic false match rates[14]. If all $\text{FMR}_{ij} \ll 1/n_j$ this simplifies to

$$\text{FPIR}_i(\tau) = \sum_j \text{FMR}_{ij}(\tau)\, n_j \tag{34}$$

which has the convenient vector-matrix notation:

$$\mathbf{FPIR}(\tau) = \mathbf{FMR}(\tau)\ \mathbf{n} \tag{35}$$

where $\mathbf{n}$ is the database composition vector, whose i-th element is the integer count of people in demographic i. Note that the matrix notation is an elegant device made possible by the approximation used for eq. 32 but is not necessary: We could exhaustively re-write with the full Binomial from eq. 31.

Further, if this database is later searched with p_i probes from each demographic group $1 \leq i \leq I$ then the expected number of false positives (NFP) for that group is

$$\text{NFP}_i(\tau) = p_i\ \text{FPIR}_i(\tau) \tag{36}$$

[13] See Figure J in algorithm-specific report cards for a one-to-many algorithm that has FPIR(T) scaling linearly as predicted by binomial models, and an example of one that does not.

[14] See Figures 6 for women only, and the much larger PDF file for all combinations of age, sex, and country-of-birth.

and the total number would be

$$\text{NFP}(\tau) = \mathbf{p}^T\ \mathbf{FMR}(\tau)\ \mathbf{n} \tag{37}$$

where $\mathbf{p}$ is the probe search count vector. An overall FPIR is available from its definition as the number of false positives divided by the number of searches:

$$\text{FPIR}(\tau) = \frac{\text{NFP}(\tau)}{\sum_i p_i} \tag{38}$$

Special case 1: Worth considering are two special forms for **FMR**. First is the case of broadly homogeneous [8] false match rates in which $\mathbf{FMR} = f\mathbf{1}\mathbf{1}^T$ (with $\mathbf{1}^T = (1, 1, \ldots 1)$) meaning that false match rates don't depend on these demographics at all. In that case the number of false positives is

$$\text{NFP}(\tau) = f(\tau) \sum_i n_i \ \sum_i p_i \tag{39}$$

and the false positive identification rate is

$$\text{FPIR}(\tau) = f(\tau) \sum_i n_i \quad = N\,\text{FMR}(\tau) \tag{40}$$

which is equation 32. This is widely considered to hold for the features extracted from fingerprint and iris characteristics, and yields the situation where demographic false positive counts are driven simply by representation of the groups in the enrollee population, with $f(\tau)$ being a pan-demographic FMR scalar value.

Special case 2: A second case is of narrow homogeneity, $\mathbf{FMR} = f\mathrm{I}$, meaning that false matches only occur within-demographic and all groups have the same rate, f.

$$\text{NFP}(\tau) = f(\tau)\mathbf{p}^T\mathrm{I}\mathbf{n} = f(\tau)\mathbf{p}^T\mathbf{n} = f(\tau) \sum_i n_i p_i \tag{41}$$

$$\text{FPIR}(\tau) = f(\tau) \frac{\sum_i n_i p_i}{\sum_i p_i} \tag{42}$$

This means that false positive outcomes depend now on the demographic structure of the searches, in addition to the enrollments. This point was made by Howard et al. [7].

For a given f, equation 39 gives a higher value than 41 but a biometric modality or algorithm that offered broad homogeneity could be configured with a different threshold τ to give lower f.

In summary, the expected number of false positives for a demographic will depend on

- ▷ **Gallery presence:** How commonly members of the particular demographic are present in the gallery.
- ▷ **False match rates within demographic:** The FMR_{ii} values govern how often individuals false match against people with the same demographics.
- ▷ **False match rates against other demographics:** As is evident in, for example, Fig. 6, false matches with other demographic groups are not insignificant, and must be accounted for. The full matrix shows, for example, significant male-female false match rates in the young, $(12 - 20]$.
- ▷ **Search volumes:** Once an FR system is deployed, the frequency with which individuals from a particular group are searched will increase the number of false positives for that group. This is separate to their presence in the enrollment database and their propensity to match within and across demographic groups.

Important: An important subset of 1:N search algorithms do not implement search as N 1:1 comparisons, and the Binomial formulation above does not apply. In particular, as noted in NIST Interagency Report 8280: FRVT Part 3: Demographics, some algorithms, specifically stabilize the right tail of the impostor distribution so that gallery size does not affect FPIR (FPIR is constant vs. linear in N) and they thereby reduce demographic variations in FPIR. This caveat is not present in the cited publications.

C Threshold-independent measures

The error rate changes summarized by the measures introduced above, are often underlied by changes in the underlying score distributions - often a relative shift of the distributions between two demographics. That is, FNMR and FMR may not just vary because of effects in, respectively, the left and right tails of the score distributions, they would vary because the entire distributions are different also. We think this can be quantified as follows.

▷ **EMD**: The Earth Mover's Distance (EMD) for two one-dimensional distribution functions is

$$\text{EMD}_{nm} = \int_{x=-\infty}^{\infty} |F_{d_m}(x) - F_{d_n}(x)| \; dx \tag{43}$$

The EMD quantifies how different the impostor distributions are by integrating across their range. This gives us a measure of inequity that is threshold independent. For false match rates, the distribution functions are related to FMR via $\text{FMR}(x) = 1 - F(x)$ so

$$\text{EMD}_{nm} = \int_{x=-\infty}^{\infty} |\text{FMR}_{d_n}(x) - \text{FMR}_{d_m}(x)| \; dx \tag{44}$$

This quantity can be computed from empirical cumulative distribution functions using numerical integration.

▷ **KLD**: A second formulation is the Kullback-Leibler divergence, measuring the "surprise" of a second distribution relative to the first.

$$\text{KLD}_{nm} = \int_{x=-\infty}^{\infty} f_{d_m}(x) \log \frac{f_{d_m}(x)}{f_{d_n}(x)} \; dx \tag{45}$$

where $f_{d_m}(x)$ and $f_{d_n}(x)$ are density functions, for example of impostor scores from two demographics. The measure is asymmetric.

Both of these measures are threshold-independent, summarizing full distributional differences. That aspect is at once attractive because it removes the need to set a threshold, and weak in that the measures are arguably less human-interpretable.

References

[1] Joy Buolamwini. Gender shades: Intersectional phenotypic and demographic evaluation of face datasets and gender classifiers. Technical report, MIT Media Lab, 01 2017.

[2] Cynthia M. Cook, John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science (IEEE T-BIOM)*, 1(1):32–41, February 2019.

[3] Tiago de Freitas Pereira and Sébastien Marcel. Fairness in biometrics: a figure of merit to assess biometric verification systems. *CoRR*, abs/2011.02395, 2020.

[4] Pawel Drozdowski, Christian Rathgeb, and Christoph Busch. The watchlist imbalance effect in biometric face identification: Comparing theoretical estimates and empiric measurements. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3750–3758, 2021.

[5] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (frvt) - performance of automated gender classification algorithms. Nist interagency report 8280, National Institute of Standard and Technology (NIST), December 2019. https://doi.org/10.6028/NIST.IR.8280.

[6] John J. Howard, Eli J. Laird, Yevgeniy B. Sirotin, Rebecca E. Rubin, Jerry L. Tipton, and Arun R. Vemury. Evaluating proposed fairness models for face recognition algorithms. *arXiv*, March 2022. https://arxiv.org/abs/2203.05051.

[7] John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury. Quantifying the extent to which race and gender features determine identity in commercial face recognition algorithms. Technical paper series, DHS Science and Technology Directorate, May 2021. https://www.dhs.gov/sites/default/files/publications/21_0922_st_quantifying-commercial-face-recognition-gender-and-race_updated.pdf.

[8] John J. Howard, Yevgeniy B. Sirotin, and Arun R. Vemury. The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. In *Proc. 10th International Conference on Biometric Theory, Applications, and Systems*. IEEE, 2019. https://mdtf.org/publications/broad-and-specific-homogeneity.pdf.

[9] Mei Ngan and Patrick Grother. Face recognition vendor test (frvt) - performance of automated gender classification algorithms. Nist interagency report 8052, National Institute of Standard and Technology (NIST), April 2015.

[10] Inioluwa Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Conference on AI, Ethics and Society*, pages 429–435, 01 2019.

[11] Fritz Scholz. Confidence bounds and intervals for parameters relating to the binomial, negative binomial, poisson and hypergeometric distributions with applications to rare events. Technical report, University of Washington, November 2019. https://faculty.washington.edu/fscholz/DATAFILES498B2008/ConfidenceBounds.pdf.

[12] J. Sherrah. False alarm rate: a critical performance measure for face recognition. In *Proc. Sixth IEEE International Conference on Automatic Face and Gesture Recognition.*, pages 189–194, 2004.