NISTIR XXXX Draft

Ongoing Face Recognition Vendor Test (FRVT)

Part 1: Verification

Patrick Grother
Mei Ngan
Kayee Hanaoka
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing

National Institute of
Standards and Technology
U.S. Department of Commerce

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.

ABOUT THIS REPORT

This report is a draft NIST Interagency Report, and is open for comment. It documents the verification-track of the ongoing Face Recognition Vendor Test. The report will be updated continuously as new algorithms are evaluated, as new datasets are added, and as new analyses are included. Comments and suggestions should be directed to frvt@nist.gov.

Contents

List of Tables

List of Figures

Developer	Short	Seq.	Validation	Config[1]	Template		GPU	FTE[3]	
Name	Name	Num.	date	data (KB)	size (B)	time (ms)[2]		Visa	Mugshot
3DiVi	3divi	000	2017-03-16	169360	511 ± 16	276 ± 58	Yes	0.0008	0.0019
Dermalog	dermalog	001	2017-02-22	0	1041 ± 47	107 ± 4	No	0.0013	0.0045
Dermalog	dermalog	002	2017-02-22	0	1041 ± 47	83 ± 7	No	0.0013	0.0045
Neurotechnology	neurotech	000	2017-03-22	62129	7148 ± 0	609 ± 47	No	0.0000	0.0000
N-Tech Lab	ntech	000	2017-03-13	191530	2906 ± 1	277 ± 12	No	0.0016	0.0015
Rank One Computing	rankone	000	2017-03-21	0	144 ± 0	83 ± 10	No	0.0003	0.0005
Rank One Computing	rankone	001	2017-04-12	0	208 ± 0	61 ± 7	No	0.0000	0.0001
TongYi Transportation Techology	tongyi	001	2017-04-01	625339	2048 ± 145	285 ± 42	No	0.0040	0.0068
VCognition	vcog	001	2017-03-28	86103	4126 ± 0	223 ± 60	Yes	0.0018	-
Vigilant Solutions	vigilant	000	2017-03-30	352218	31540 ± 0	851 ± 92	No	0.0007	0.0018
Vocord	vocord	001	2017-04-21	616989	6107 ± 728	950 ± 89	No	0.0038	0.0158

Notes	
1	The size of configuration data does not capture static data included in the libraries. We do not include the size of the libraries because some algorithms include common ancilliary libraries for image processing (e.g. openCV) or numerical computation (e.g. blas).
2	The template creation times are measured on Intel®Xeon®CPU E5-2630 v4 @ 2.20GHz processors. or, in the case of GPU-enabled implementations, NVidia Tesla K40
3	FTE is the proportion of template generation function calls that give non-zero error code OR that give a small template, i.e. one whose size is less than 0.3 times the median template size. This second rule is needed because some algorithms do not return a non-zero error code when template generation fails The effects of FTE are included in the accuracy results later in this report by regarding any template comparison that involves an failed template is taken to produce a low similarity score.

Table 1: Algorithms included in this report

1 Metrics

1.1 Core accuracy

Given a vector of N genuine scores, u, the false non-match rate (FNMR) is computed as the proportion at or above some threshold, T:

$$\text{FNMR}(T) = 1 - \frac{1}{N}\sum_{i}^{N} H(u_i - T) \tag{1}$$

where $H(x)$ is the unit step function, and $H(0)$ taken to be 1.

Similarly, given a vector of N impostor scores, v.

$$\text{FMR}(T) = \frac{1}{N}\sum_{i}^{N} H(v_i - T) \tag{2}$$

The threshold, T, can take on any value. We typically generate a set of thresholds from quantiles of the observed impostor scores, v, as follows. Given some interesting false match rate range, $[\text{FMR}_L, \text{FMR}_U]$, we form a vector of K thresholds corresponding to FMR measurements evenly spaced on a logarithmic scale

$$T_k = Q_v(1 - \text{FMR}_k) \tag{3}$$

where Q is the quantile function, and FMR_k comes from

$$\log_{10} \text{FMR}_k = \log_{10} \text{FMR}_L + \frac{k}{K}\ [\log_{10} \text{FMR}_U - \log_{10} \text{FMR}_L] \tag{4}$$

Error tradeoff characteristics are plots of FNMR(T) vs. FMR(T). These are plotted with $\text{FMR}_U \to 1$ and FMR_L as low as is sustained by the number of impostor comparisons, N, This is somewhat higher than the "rule of three" limit $3/N$ because samples are not independent, due to re-use of images.

1.2 Cross age impostor distribution stability

In later sections this report documents how FMR varies with individuals age and place of birth. Such variation is often unwelcome because certain low FMR values are targeted by system owners. For example, an eGate might be set up to target FMR of 0.001.

To summarize an algorithm's sensitivity to age we define a new metric as follows. Given a N x N matrix of false match rates, FMR, where element FMR_{ij} is the FMR measured when comparing samples from age groups i and j the age sensitivity metric, β_a is taken to be the standard deviation of the on-diagonal within-age-group FMR

$$\beta_a^2 = \frac{1}{N-1}\sum_{i}^{N}(\text{FMR}_{ii} - f)^2 \tag{5}$$

where mean FMR is $f = (1/N)\sum \text{FMR}_{ii}$. This is one measure of how unstable FMR is with age. In an immmigration context, it models the case where impostors attempt to use the passport of someone of the same approximate age. Readers might suggest an improved metric.

1.3 Cross birth country impostor distribution stability

Similarly for sensitivity to national origin, FMR varies with the birth places of the subjects in the two images.

FNMR(T)	"False non-match rate"
FMR(T)	"False match rate"

Given the N x N matrix of false match rates, FMR, where element FMR_{ij} is the FMR measured when comparing samples from countries (or regions) C_i and C_j the birth place sensitivity metric, β_c is taken to be the standard deviation of the on-diagonal within-country FMR

$$\beta_c^2 = \frac{1}{N-1}\sum_i^N(\mathrm{FMR}_{ii} - f)^2 \tag{6}$$

where mean FMR is $f = (1/N)\sum \mathrm{FMR}_{ii}$. This is one measure of how unstable FMR is with country of birth. In an immmigration context, it models the case where impostors attempt to use the passport of someone from the same country. Readers might suggest an improved metric.

2 Datasets

2.1 Visa images

▷ The number of images is O(10^5).

▷ The number of subjects is O(10^5).

▷ The number of subjects with two images O(10^4).

▷ The images have geometry in reasonable conformance with the ISO/IEC 19794-5 Full Frontal image type. Pose is generally excellent.

▷ The images are of size 252x300 pixels. The mean IOD is 69 pixels.

▷ The images are of subjects from greater than 100 countries, with significant imbalance due to visa issuance patterns.

▷ The images are of subjects of all ages, including children, again with imbalance due to visa issuance demand.

▷ Many of the images are live capture. A substantial number of the images are photographs of paper photographs.

2.2 Mugshot images

▷ The number of images is O(10^6).

▷ The number of subjects is O(10^5).

▷ The number of subjects with two images O(10^5).

▷ The images have geometry in reasonable conformance with the ISO/IEC 19794-5 Full Frontal image type.

▷ The images are of variable sizes. The median IOD is 104 pixels. The mean IOD is 123 pixels.

▷ The images are of subjects from the United States.

▷ The images are of adults.

▷ The images are all live capture.

2.3 Selfie images

▷ The number of images is below 500.

▷ The number of subjects is below 500.

▷ All subjects have a selfie image, and a portrait image.

▷ The portrait images are in reasonable conformance with the ISO/IEC 19794-5 Full Frontal image type.

▷ The selfie images vary: taken with camera above and below eye level, with one hand or two hands. Pitch angles vary more than yaw angles, which are frontal. Some perspective distortion is evident.

▷ The images have mean IOD of 140 pixels.

▷ The images are of subjects from the United States.

▷ The images are of adults.

▷ The images are all live capture.

3 Results

3.1 Overall accuracy

Goals:

- ▷ To state overall accuracy.
- ▷ To compare algorithms.

Method: For visa images:

- ▷ The comparisons are of visa photos against visa photos.
- ▷ The number of genuine comparisons is O(10^4).
- ▷ The number of impostor comparisons is O(10^{10}).
- ▷ The comparisons are fully zero-effort, meaning impostors are paired without attention to sex, age or other covariates. However, later analysis is conducted on subsets.
- ▷ The number of persons is O(10^5).
- ▷ The number of images used to make a template is 1.
- ▷ The number of templates used to make each comparison score is two corresponding to simple one-to-one verification.

For mugshot images:

- ▷ The comparisons are of mugshot photos against mugshot photos.
- ▷ The number of genuine comparisons is O(10^5).
- ▷ The number of impostor comparisons is O(10^7).
- ▷ The comparisons are fully zero-effort, meaning impostors are paired without attention to sex, age or other covariates.
- ▷ The number of persons is O(10^6).
- ▷ The number of images used to make a template is 1.
- ▷ The number of templates used to make each comparison score is two corresponding to simple one-to-one verification.

For selfie images:

- ▷ The comparisons are of selfie photos against portrait photos.
- ▷ The number of genuine comparisons is O(10^2).
- ▷ The number of impostor comparisons is O(10^8) selfies are compared with portraits of O(10^6) other subjects.
- ▷ The comparisons are fully zero-effort, meaning impostors are paired without attention to sex, age or other covariates.
- ▷ The number of persons is O(10^6).

- ▷ The number of images used to make a template is 1.
- ▷ The number of templates used to make each comparison score is two corresponding to simple one-to-one verification.

Summary: Core algorithm accuracy is stated via the error tradeoff characteristics of Figures 1 and 3.

Figure 4 shows dependence of false match rate on algorithm score threshold. This allows a deployer to set a threshold to target a particular false match rate appropriate to the security objectives of the application.

Figure 5 likewise shows FMR(T) but for mugshots, and specially four subset of the population.

Note that in both the mugshot and visa sets false match rates vary with the ethnicity, age, and sex, of the enrollee and impostor - see section 3.3.

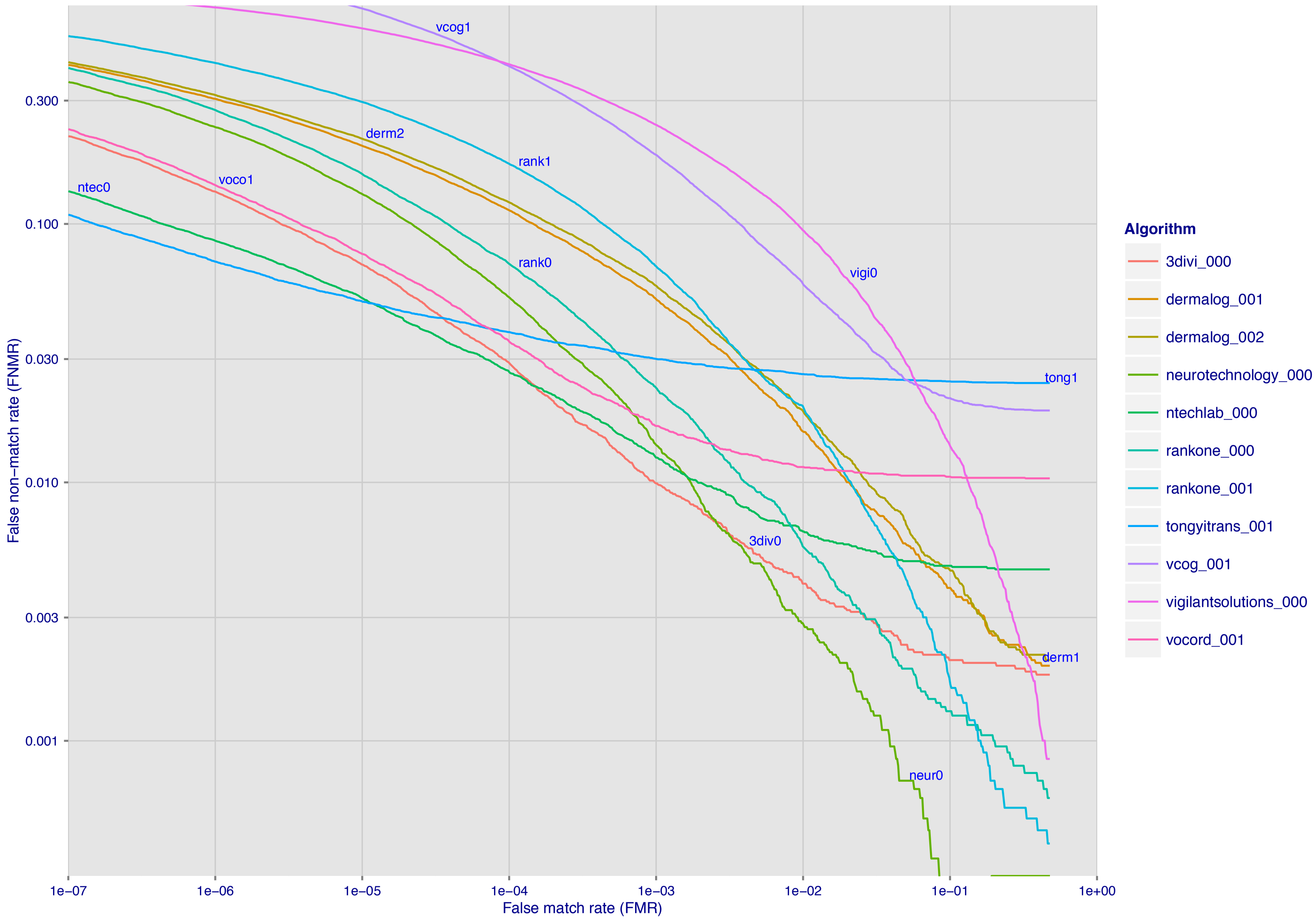


Figure 1: For the visa images, detection error tradeoff (DET) *characteristics showing false non-match rate vs. false match rate plotted paratemetrically on threshold, T. The scales are logarithmic in order to show many decades of* FMR.

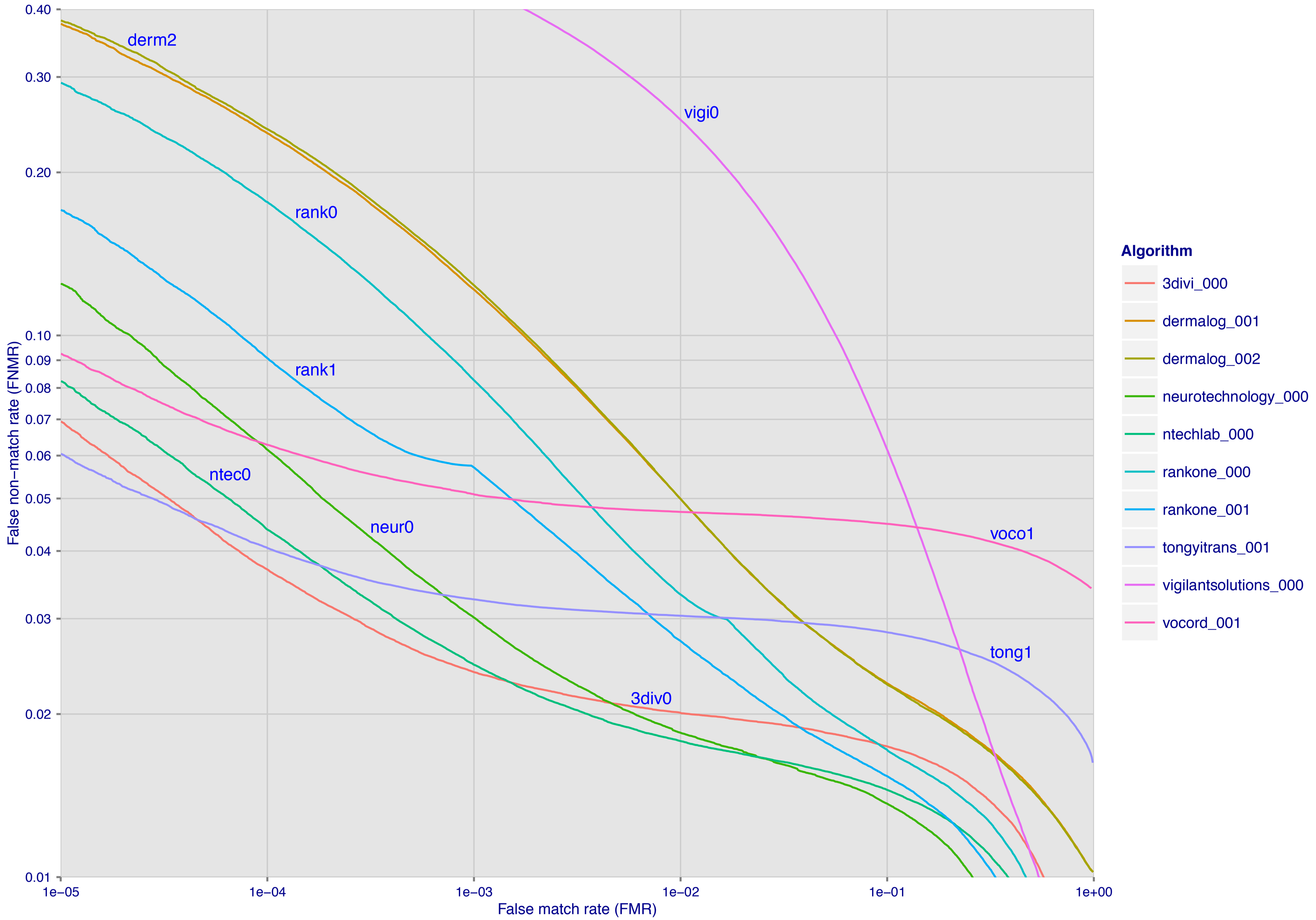


Figure 2: For the mugshot images, detection error tradeoff (DET) *characteristics showing false non-match rate vs. false match rate plotted paratemetrically on threshold, T. The scales are logarithmic in order to show decades of* FMR.

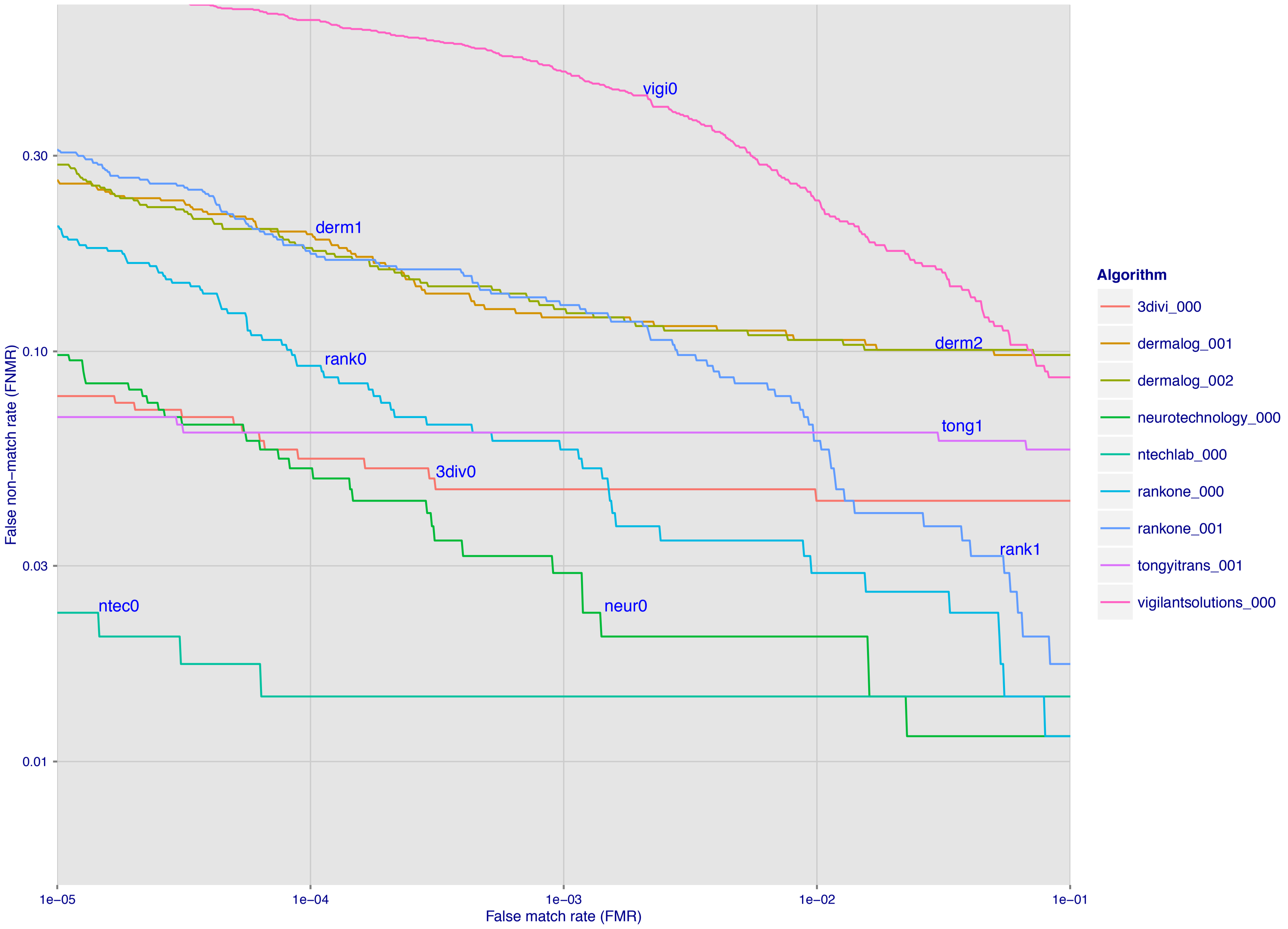


Figure 3: For the selfie-to-portrait comparisons, detection error tradeoff (DET) characteristics showing false non-match rate vs. false match rate plotted paratemetrically on threshold, T. The scales are logarithmic in order to show decades of FMR. Caution: The FNMR values here are optimistic statements of accuracy because the image pairs were collected on the same day. This is known across biometrics to give better accuracy, and is operationally relevant only rarely.

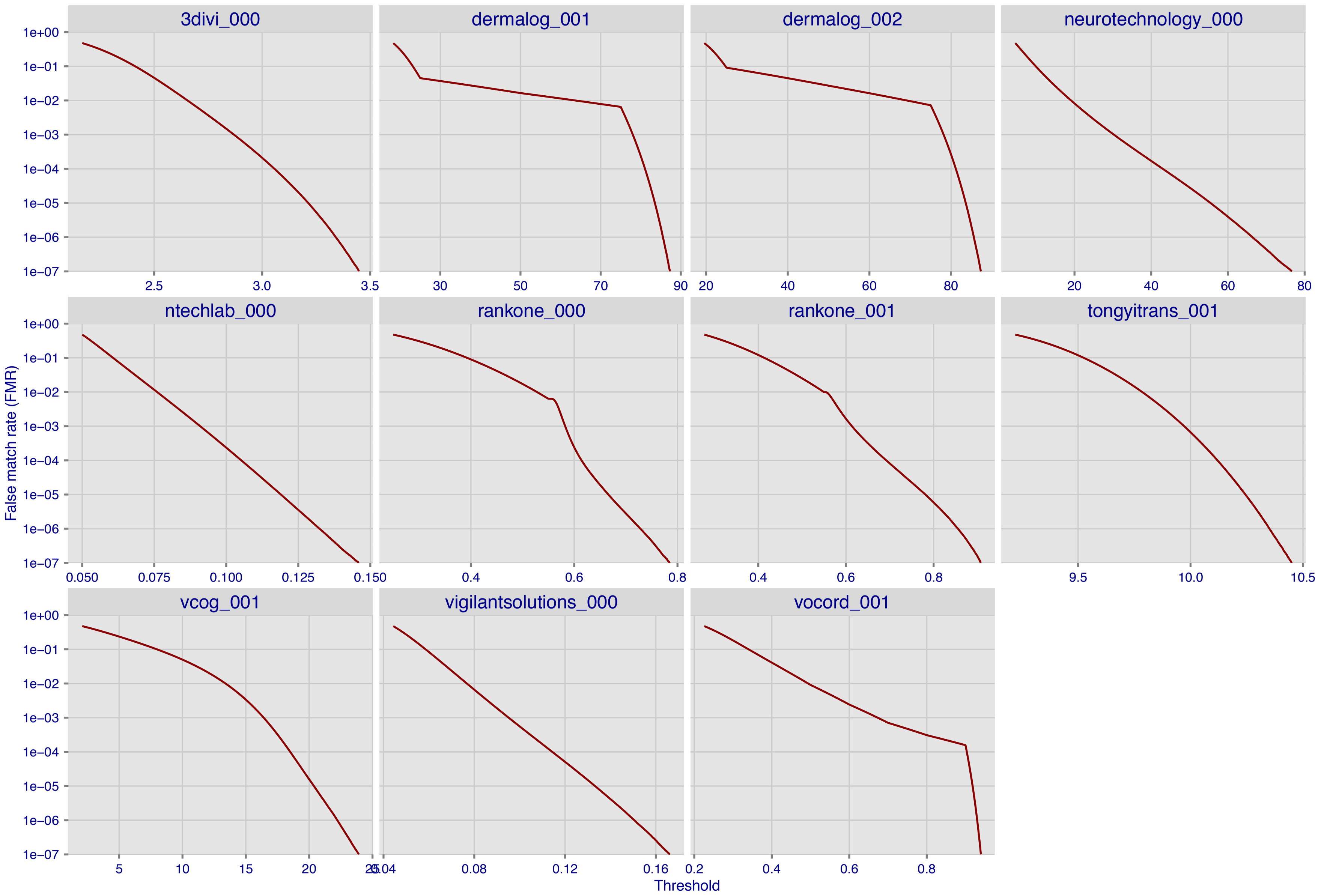


Figure 4: For the visa images, the false match calibration curves show false match rate vs. threshold. These curves apply to zero-effort impostors. As shown later (sec. 3.3), FMR is higher for demographic-matched impostors.

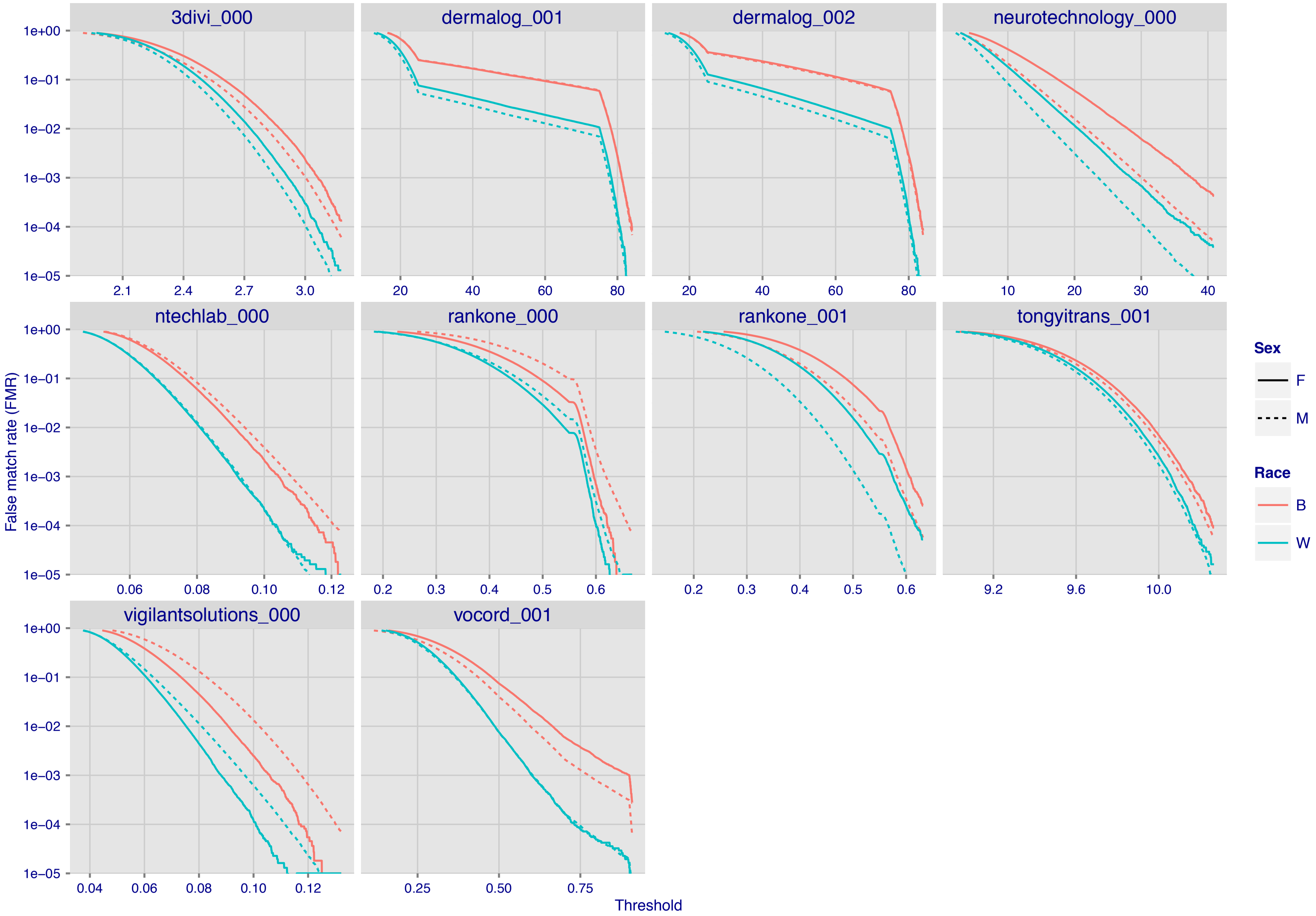


Figure 5: For the mugshot images, the false match calibration curves show false match rate vs. threshold. Separate curves appear for white females, black females, black males and white males.

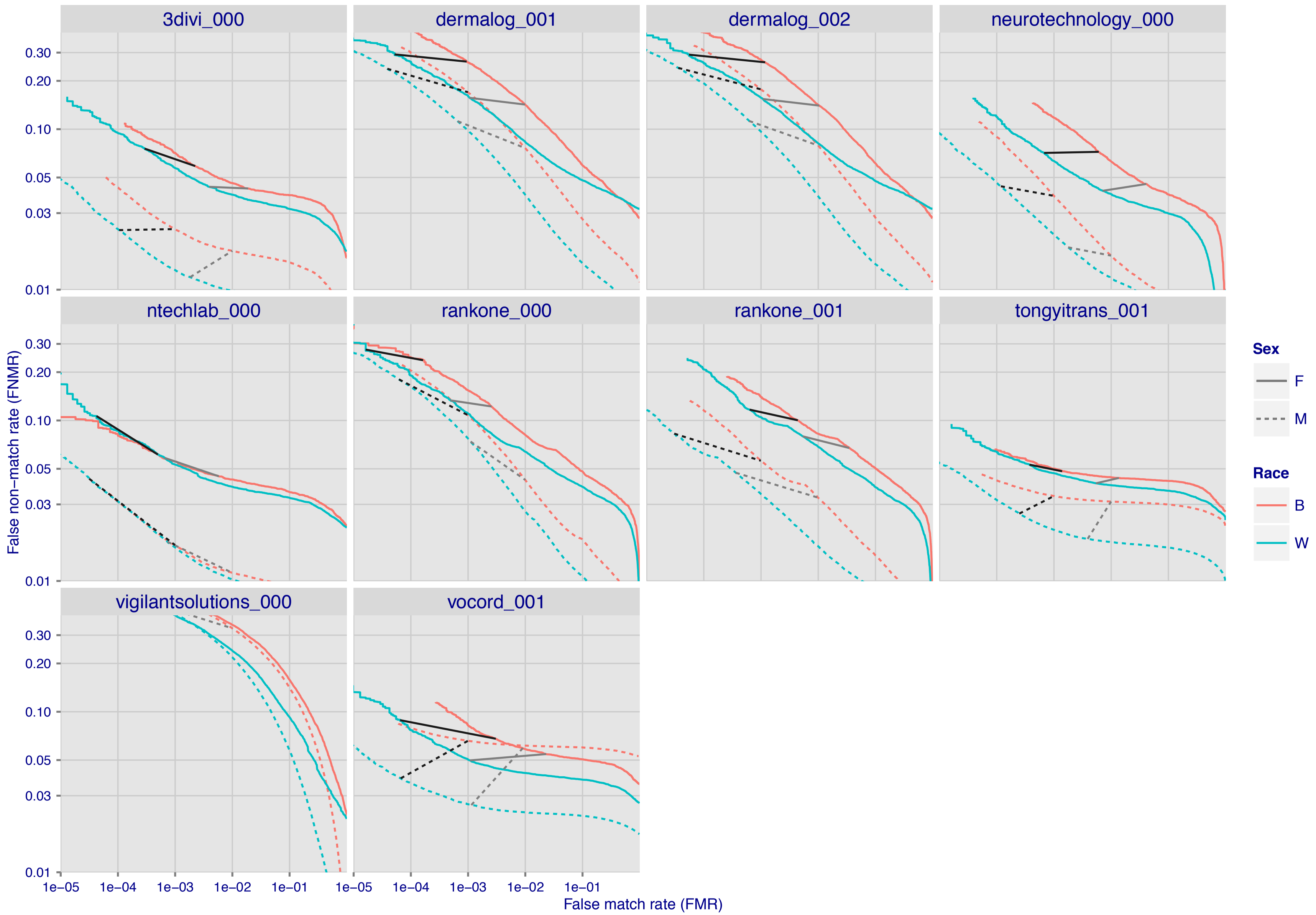


Figure 6: For the mugshot images, error tradeoff characteristics for white females, black females, black males and white males. The grey lines correspond to fixed thresholds, showing how both *FNMR and FMR vary at one operating threshold. Important: Many of the plots will naively be read as saying whites gives lower error rates than blacks because the blue traces lie beneath the red ones. However, this is misleading and incomplete: The grey lines show the traces are generally shifted horizontally. Thus for the dermalog-001 algorithm FNMR for whites is* higher than for blacks *at a fixed threshold* but, at the same time, *FMR is higher for blacks - see Figure 5. As access control systems almost always operate at a fixed threshold, the naive interpretation is incorrect.*

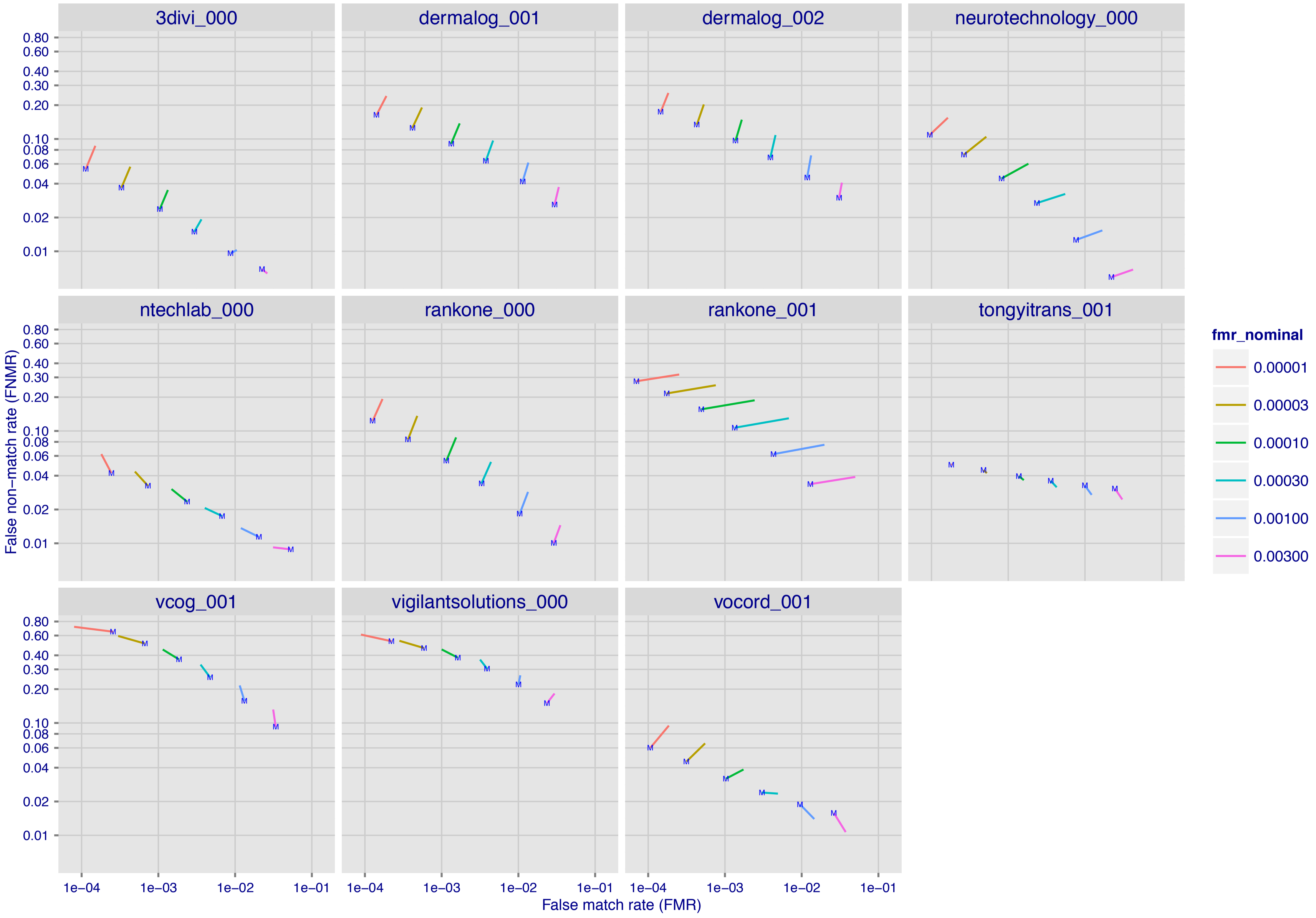


Figure 7: For the visa images, FNMR and FMR at six operating points along the DET characteristic. At each point a line is drawn between (FMR,FNMR)$_{\text{MALE}}$ and (FMR,FNMR)$_{\text{FEMALE}}$ showing how which sex has lower FMR and/or FNMR. The ”M” label denotes male, the other end of the line corresponds to female. The six operating thresholds are selected to give the nominal false match rates given in the legend, and are computed over all impostor pairs regardless of age, sex, and place of birth. The plotted FMR values are broadly an order of magnitude larger than the nominal rates because FMR is computed over demographically-matched impostor pairs i.e indviduals of the same sex, from the same geographic region (see section 3.3.1), and the same age group (see section 3.3.2).

3.2 Genuine distribution stability

3.2.1 Effect of birth place on the genuine distribution

Background: Both skin tone and bone structure vary geographically. Prior studies have reported variations in FNMR and FMR.

Goal: To measure false non-match rate (FNMR) variation with country of birth.

Methods: Thresholds are determined that give FMR = $\{0.001, 0.0001\}$ over the entire impostor set. Then FNMR is measured over 1000 bootstrap replications of the genuine scores. Only those countries with at least 140 individuals are included in the analysis.

Results: Figure 8 shows FNMR by country of birth for the two thresholds.

Caveats: The results may not relate to subject-specific properties. Instead they could reflect image-specific quality differences, which could occur due to collection protocol or software processing variations.

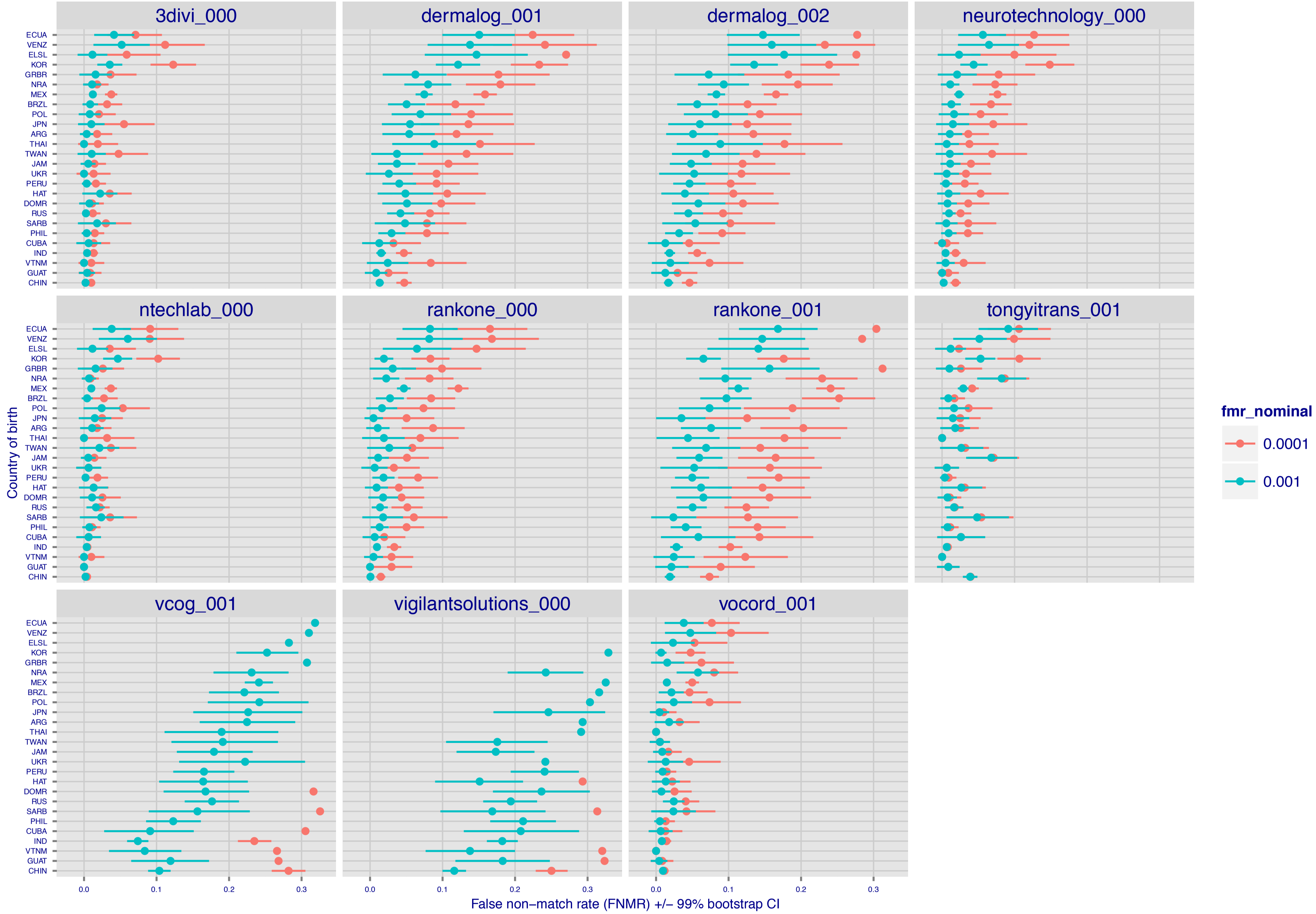


Figure 8: For the visa images, the dots show FNMR by country of birth for two operating thresholds corresponding to FMR = $\{0.001, 0.0001\}$ computed over all $O(10^{10})$ impostor scores. The figures shows an order of magnitude variation in FNMR across country of birth; these effects are due to quality variations. The least accurate countries vary by algorithm.

3.2.2 Effect of age on genuine subjects

Background: Faces change appearance throughout life. Face recognition algorithms have previously beeen reported to give better accuracy on older individuals (See NIST IR 8009).

Goal: To quantify false non-match rates (FNMR) as a function of age. We do not aim to quantify ageing effects here as the separation between two samples is limited to just a few years.

Methods: Using the visa images, thresholds are determined that give FMR = 0.001 and 0.0001 over the entire impostor set. Then FNMR is measured over 1000 bootstrap replications of the genuine scores. Only those countries with at least 30 individuals are included in the analysis.

Results: For the visa images, Figure 9 shows how false non-match rates for genuine users, as a function of age group.

The notable aspects are:

- ▷ Younger subjects give considerably higher FNMR. This is likely due rapid growth and change in facial appearance.
- ▷ FNMR trends down throughout life. The last bin, AGE $>$ 72, contains fewer than 140 mated pairs, and may be affected by small sample size.

Caveats: None.

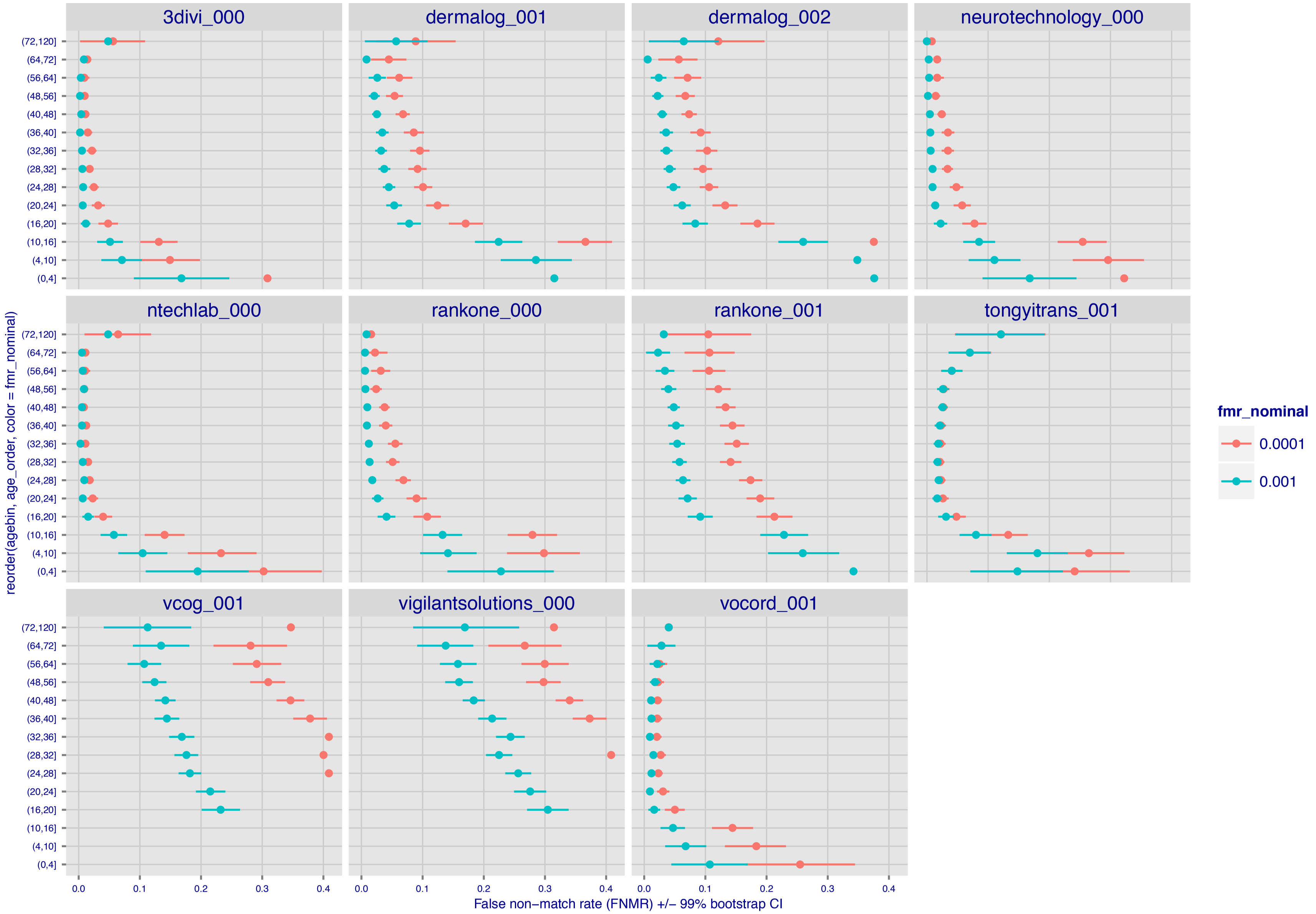


Figure 9: For the visa images, the dots show FNMR by age group for two operating thresholds corresponding to FMR = $\{0.001, 0.0001\}$ *computed over all* $O(10^{10})$ *impostor scores. Given a pair of face images taken at different times, we assign a false non-match to the bin that is the arithmetic average of the subject's ages. This plot shows only the effect of age, not ageing. The number of comparisons in each bin is generally in the thousands. However the FNMR for the first and last bins are each computed over fewer than 150 comparisons.*

3.3 Impostor distribution stability

3.3.1 Effect of birth place on the impostor distribution

Background: Facial appearance varies geographically, both in terms of skin tone, cranio-facial structure and size. This section addresses whether false match rates vary intra- and inter-regionally.

Goals:

- ▷ To show the effect of birth region of the impostor and enrollee on false match rates.
- ▷ To determine whether some algorithms give better impostor distribution stability.

Methods:

- ▷ For the visa images, NIST defined 10 regions: Sub-Saharan Africa, South Asia, Polynesia, North Africa, Middle East, Europe, East Asia, Central and South America, Central Asia, and the Caribbean.
- ▷ For the visa images, NIST mapped each country of birth to a region. There is some arbitrariness to this. For example, Egypt could reasonably be assigned to the Middle East instead of North Africa. An alternative methodology could, for example, assign the Philippines to *both* Polynesia and East Asia.
- ▷ FMR is computed for cases where all face images of impostors born in region r_2 are compared with enrolled face images of persons born in region r_1.

$$\text{FMR}(r_1, r_2, T) = \frac{\sum_{i=1}^{N_{r_1,r_2}} H(s_i - T)}{N_{r_1,r_2}} \tag{7}$$

  where the same threshold, T, is used in all cells, and H is the unit step function. The threshold is set to give FMR(T) = 0.001 over the entire set of visa image impostor comparisons.
- ▷ This analysis is then repeated by country-pair, but only for those country pairs where both have at least 1000 images available. The countries[1] appear in the axes of graphs that follow.
- ▷ The mean number of impostor scores in any cross-region bin is 33 million. The smallest number of impostor scores in any bin is 135000, for Central Asia - North Africa. While these counts are large enough to support reasonable significance, the number of individual faces is much smaller, O($N^{0.5}$).
- ▷ The numbers of impostor scores in any cross-country bin is shown in Figure 31.

Results: Subsequent figures show heatmaps that use color to represent the base-10 logarithm of the false match rate. Red colors indicate high (bad) false match rates. Dark colors indicate benign false match rates. There are two series of graphs corresponding to aggregated geographical regions, and to countries. The notable observations are:

- ▷ The on-diagonal elements correspond to within-region impostors. FMR is generally above the nominal value of FMR = 0.001. Particularly there is usually higher FMR in, Sub-Saharan Africa, South Asia, and the Caribbean. Europe and Central Asia, on the other hand, usually give FMR closer to the nominal value.
- ▷ The off-diagonal elements correspond to across-region impostors. The highest FMR is produced between the Caribbean and Sub-Saharan Africa.
- ▷ Algorithms vary.

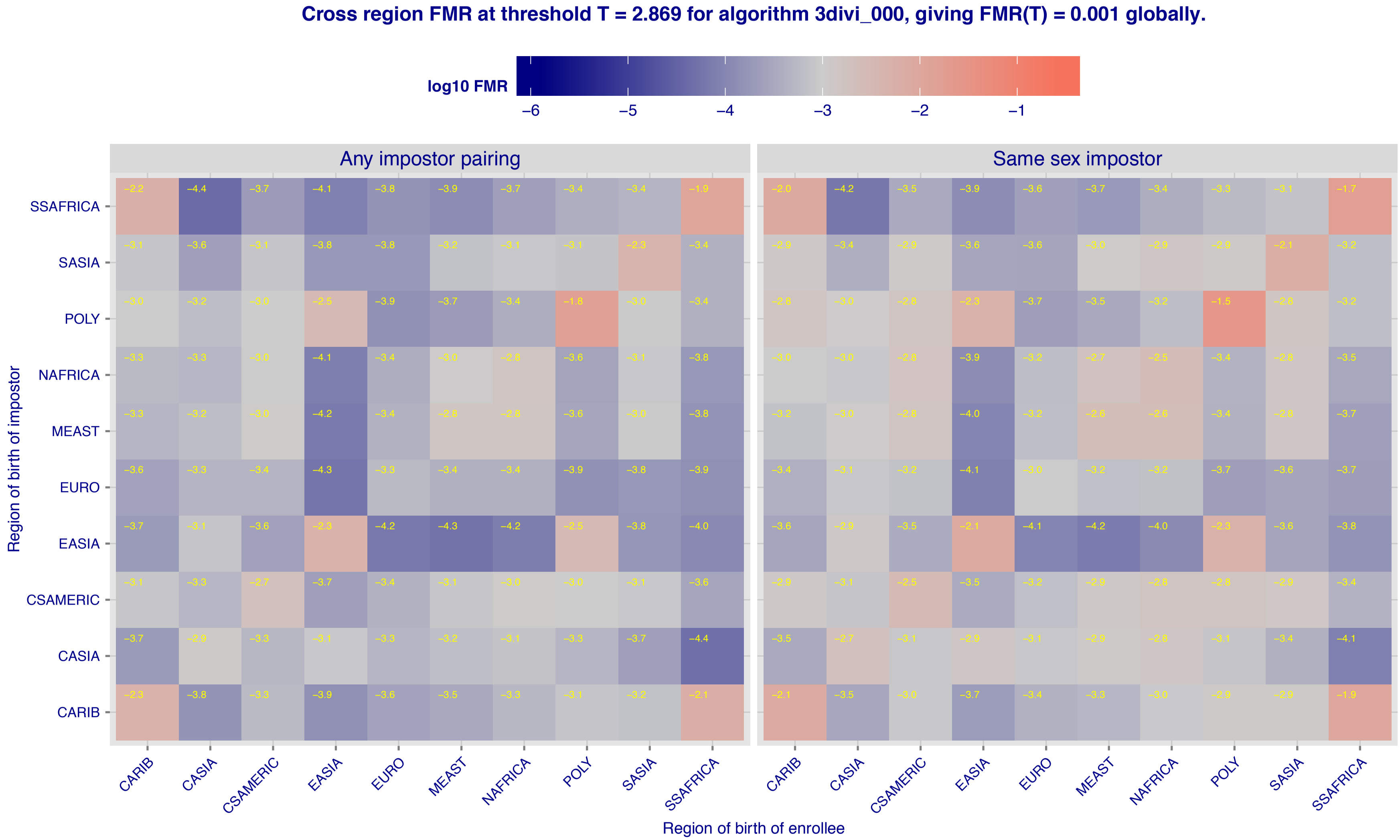
[1] These are Argentina, Australia, Brazil, Chile, China, Costa Rica, Cuba, Czech Republic, Dominican Republic, Ecuador, Egypt, El Salvador, Germany, Ghana, Great Britain, Greece, Guatamala, Haiti, Hong Kong, Honduras, Indonesia, India, Israel, Jamaica, Japan, Kenya, Korea, Lebanon, Mexico, Malaysia, Nepal, Nigeria, Peru, Philippines, Pakistan, Poland, Romania, Russia, South Africa, Saudi Arabia, Thailand, Trinidad, Turkey, Taiwan, Ukraine, Venezuela, and Vietnam.

▷ We computed the same quantities for a global FMR = 0.0001. The effects are similar.

Caveats:

▷ The effects of variable impostor rates on one-to-many identification systems may well differ from what's implied by these one-to-one verification results. Two reasons for this are a) the enrollment galleries are usually imbalanced across countries of birth, age and sex; b) one-to-many identification algorithms often implement techniques aimed at stabilizing the impostor distribution. Further research is necessary.

▷ In principle, the effects seen in this subsection could be due to differences in the image capture process. We consider this unlikely since the effects are maintained across geography - e.g. Caribbean vs. Africa, or Japan vs. China.



Any impostor pairing

	CARIB	CASIA	CSAMERIC	EASIA	EURO	MEAST	NAFRICA	POLY	SASIA	SSAFRICA
SSAFRICA	−2.2	−4.4	−3.7	−4.1	−3.8	−3.9	−3.7	−3.4	−3.4	−1.9
SASIA	−3.1	−3.6	−3.1	−3.8	−3.8	−3.2	−3.1	−3.1	−2.3	−3.4
POLY	−3.0	−3.2	−3.0	−2.5	−3.9	−3.7	−3.4	−1.8	−3.0	−3.4
NAFRICA	−3.3	−3.3	−3.0	−4.1	−3.4	−3.0	−2.8	−3.6	−3.1	−3.8
MEAST	−3.3	−3.2	−3.0	−4.2	−3.4	−2.8	−2.8	−3.6	−3.0	−3.8
EURO	−3.6	−3.3	−3.4	−4.3	−3.3	−3.4	−3.4	−3.9	−3.8	−3.9
EASIA	−3.7	−3.1	−3.6	−2.3	−4.2	−4.3	−4.2	−2.5	−3.8	−4.0
CSAMERIC	−3.1	−3.3	−2.7	−3.7	−3.4	−3.1	−3.0	−3.0	−3.1	−3.6
CASIA	−3.7	−2.9	−3.3	−3.1	−3.3	−3.2	−3.1	−3.3	−3.7	−4.4
CARIB	−2.3	−3.8	−3.3	−3.9	−3.6	−3.5	−3.3	−3.1	−3.2	−2.1

Same sex impostor

	CARIB	CASIA	CSAMERIC	EASIA	EURO	MEAST	NAFRICA	POLY	SASIA	SSAFRICA
SSAFRICA	−2.0	−4.2	−3.5	−3.9	−3.6	−3.7	−3.4	−3.3	−3.1	−1.7
SASIA	−2.9	−3.4	−2.9	−3.6	−3.6	−3.0	−2.9	−2.9	−2.1	−3.2
POLY	−2.8	−3.0	−2.8	−2.3	−3.7	−3.5	−3.2	−1.5	−2.8	−3.2
NAFRICA	−3.0	−3.0	−2.8	−3.9	−3.2	−2.7	−2.5	−3.4	−2.8	−3.5
MEAST	−3.2	−3.0	−2.8	−4.0	−3.2	−2.6	−2.6	−3.4	−2.8	−3.7
EURO	−3.4	−3.1	−3.2	−4.1	−3.0	−3.2	−3.2	−3.7	−3.6	−3.7
EASIA	−3.6	−2.9	−3.5	−2.1	−4.1	−4.2	−4.0	−2.3	−3.6	−3.8
CSAMERIC	−2.9	−3.1	−2.5	−3.5	−3.2	−2.9	−2.8	−2.8	−2.9	−3.4
CASIA	−3.5	−2.7	−3.1	−2.9	−3.1	−2.9	−2.8	−3.1	−3.4	−4.1
CARIB	−2.1	−3.5	−3.0	−3.7	−3.4	−3.3	−3.0	−2.9	−2.9	−1.9

Figure 10: For algorithm 3divi-000 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given region pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all $O(10^{10})$ impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

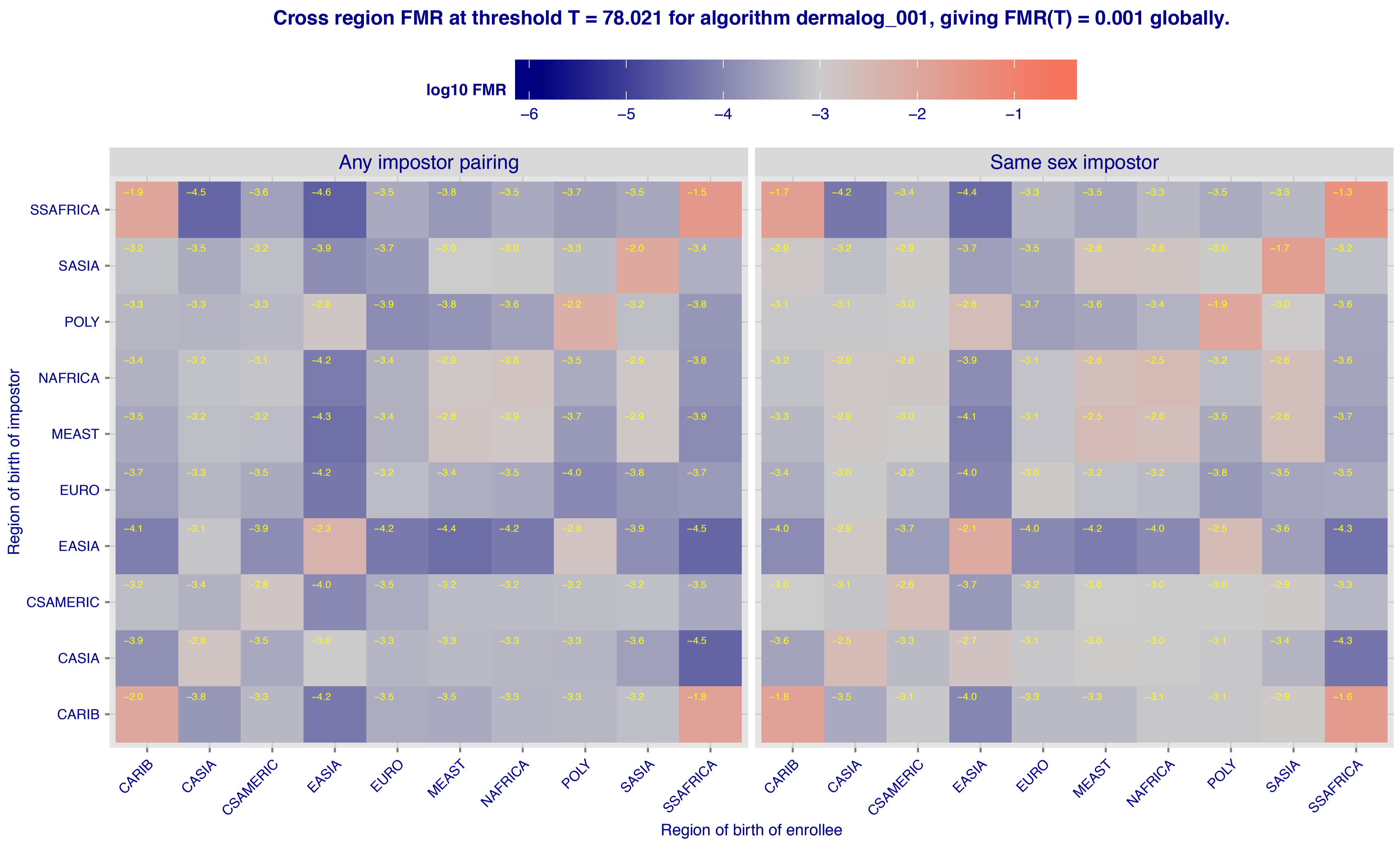


Figure 11: For algorithm dermalog-001 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given region pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all $O(10^{10})$ impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

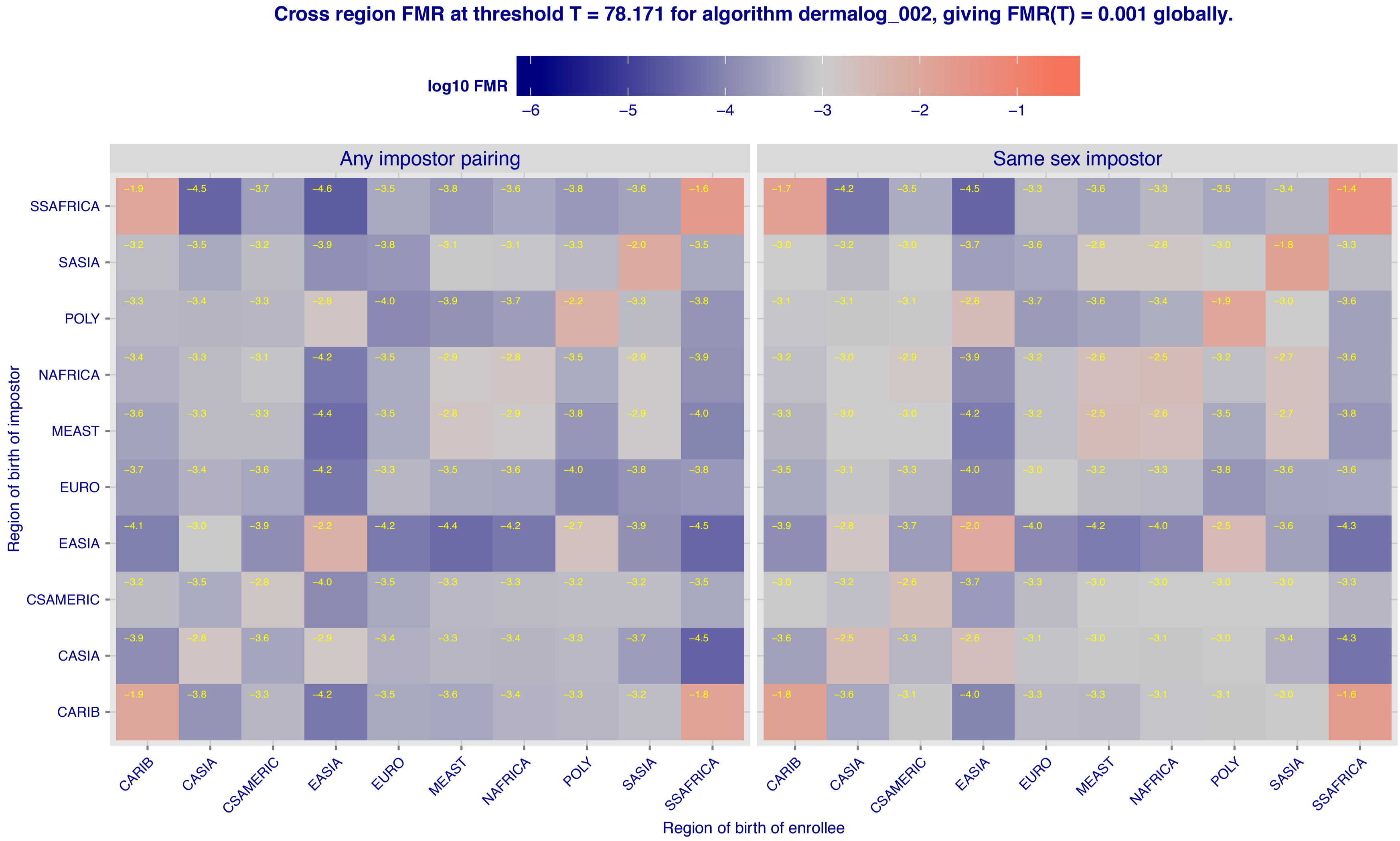


Figure 12: For algorithm dermalog-002 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given region pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

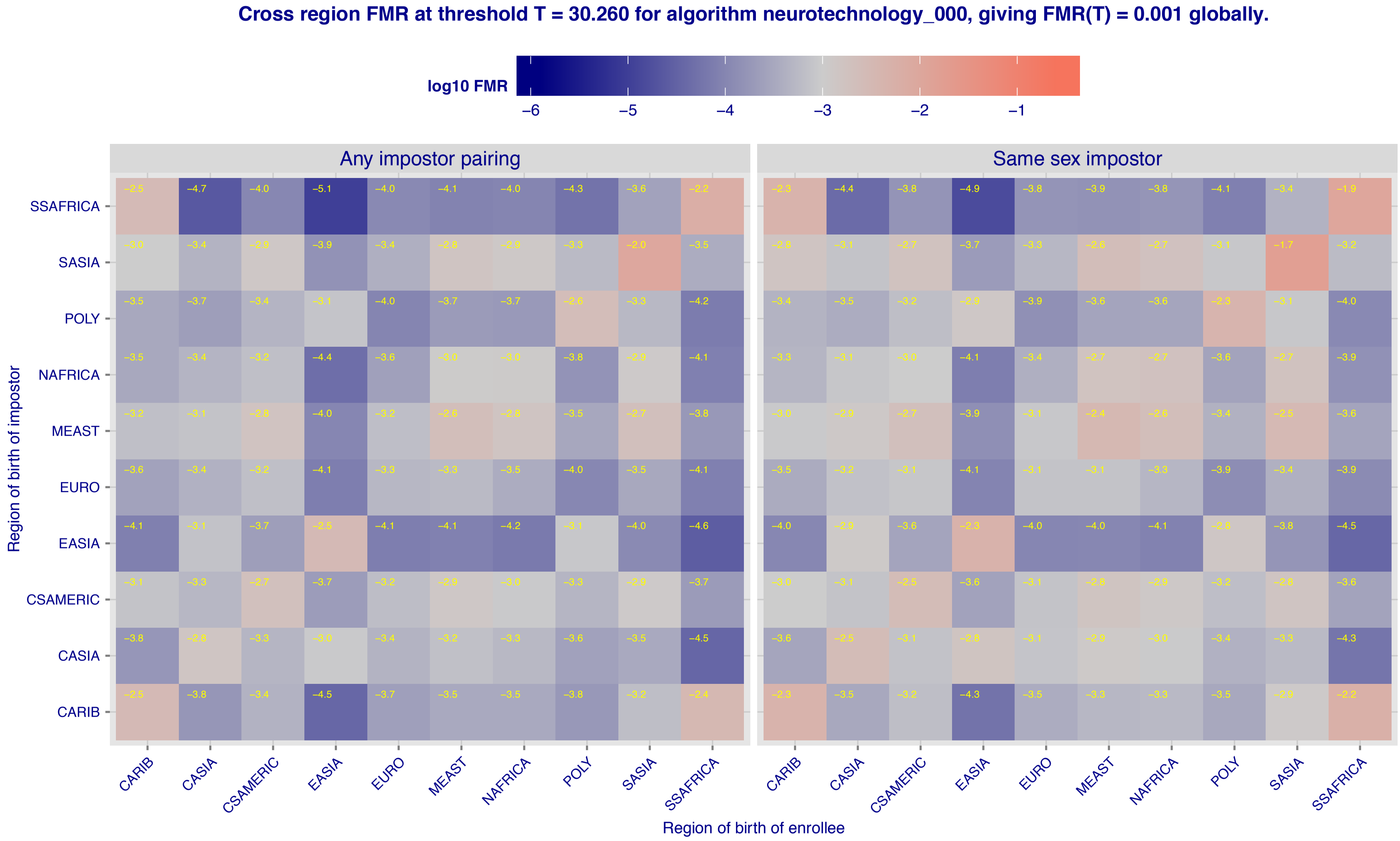


Figure 13: For algorithm neurotechnology-000 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given region pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

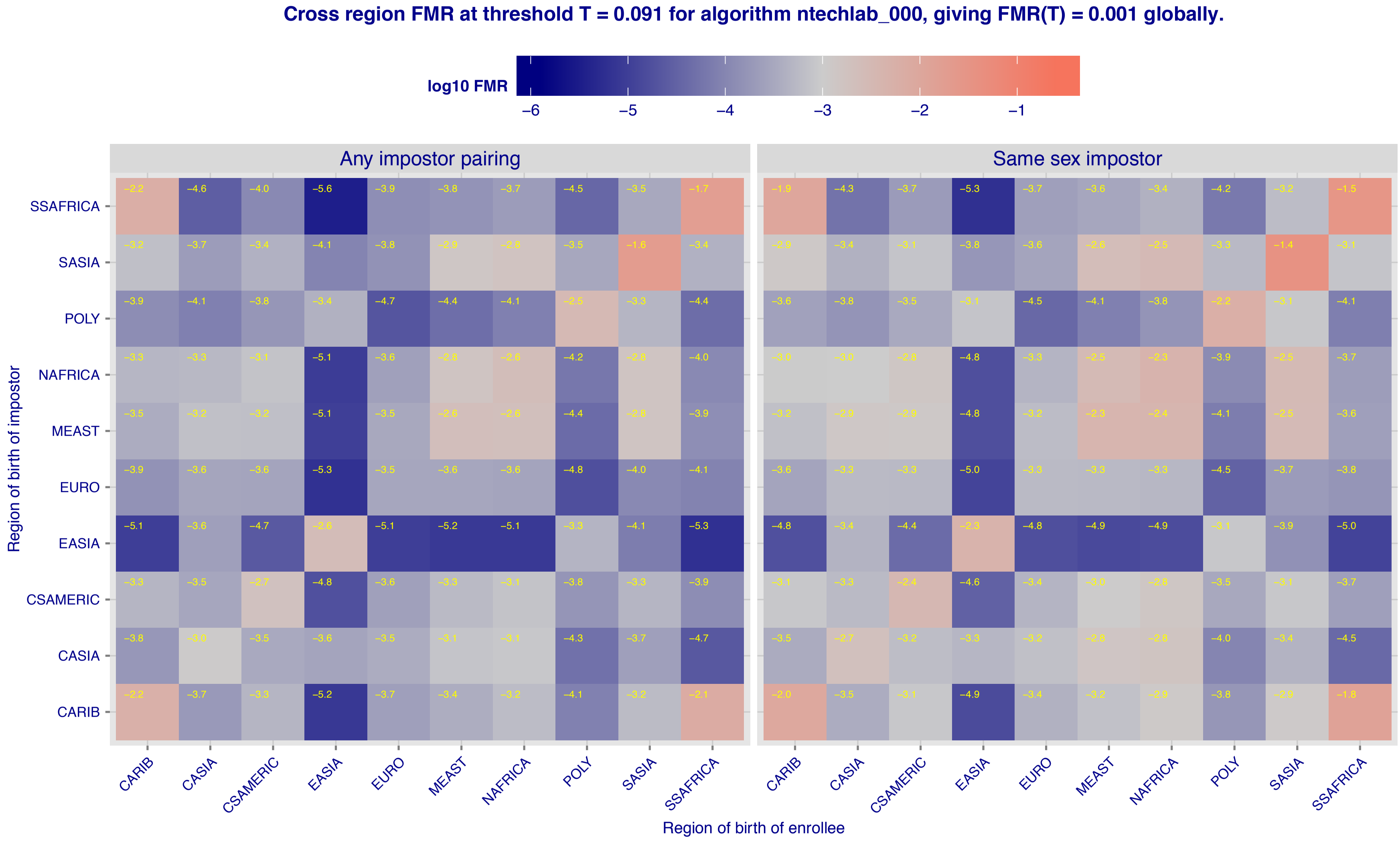


Figure 14: For algorithm ntechlab-000 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given region pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

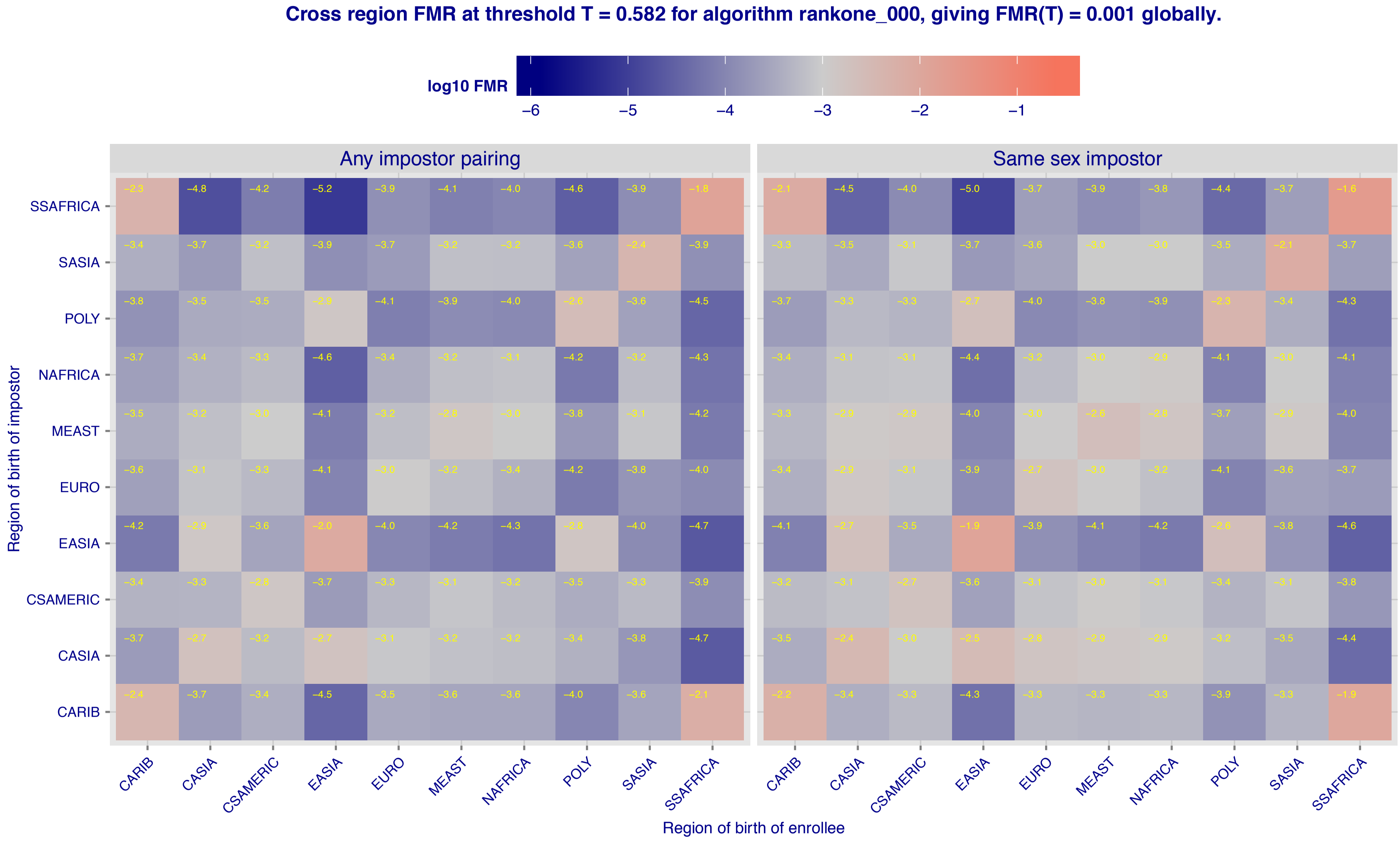


Figure 15: For algorithm rankone-000 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given region pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all $O(10^{10})$ impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

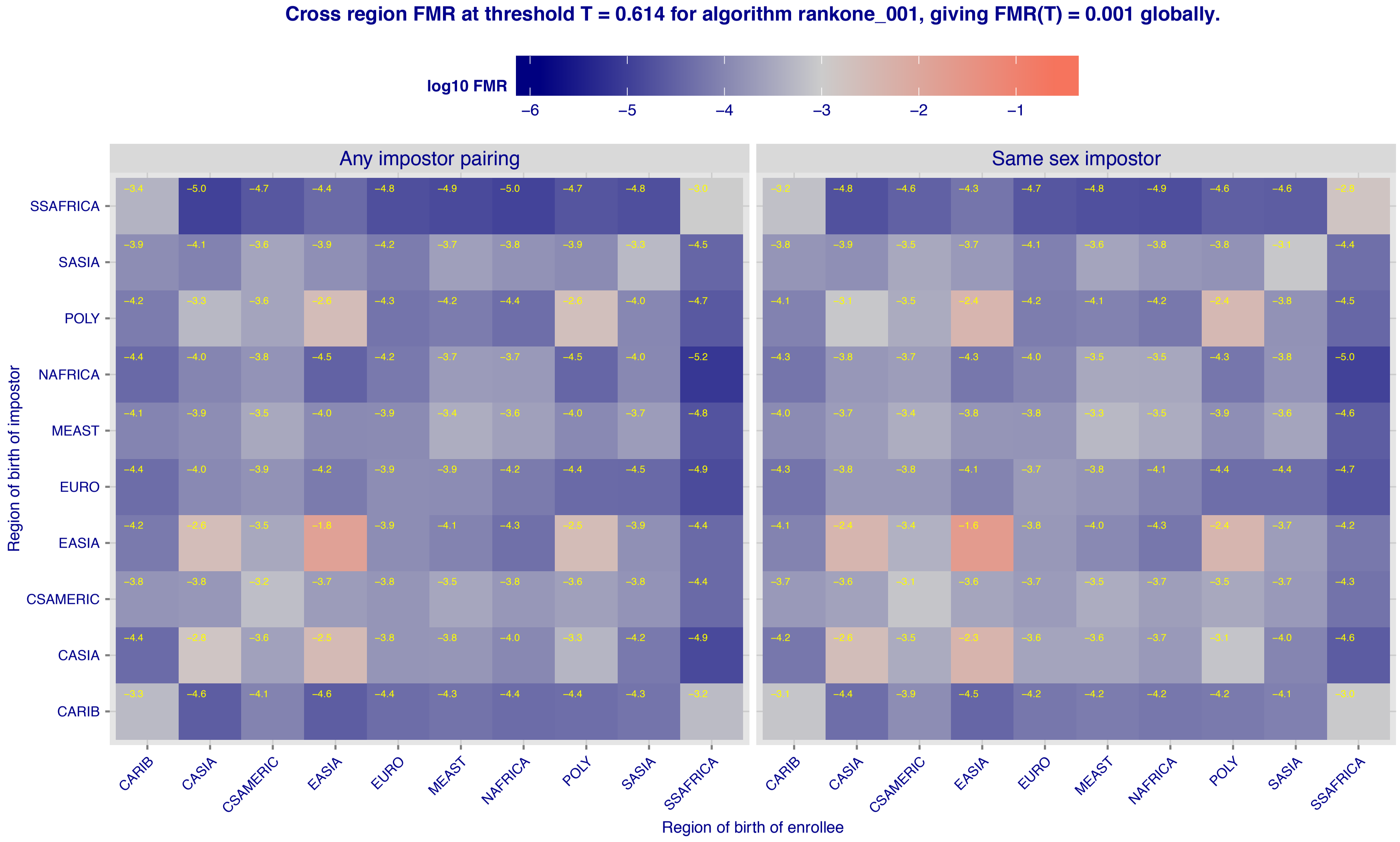


Figure 16: For algorithm rankone-001 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given region pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all $O(10^{10})$ impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

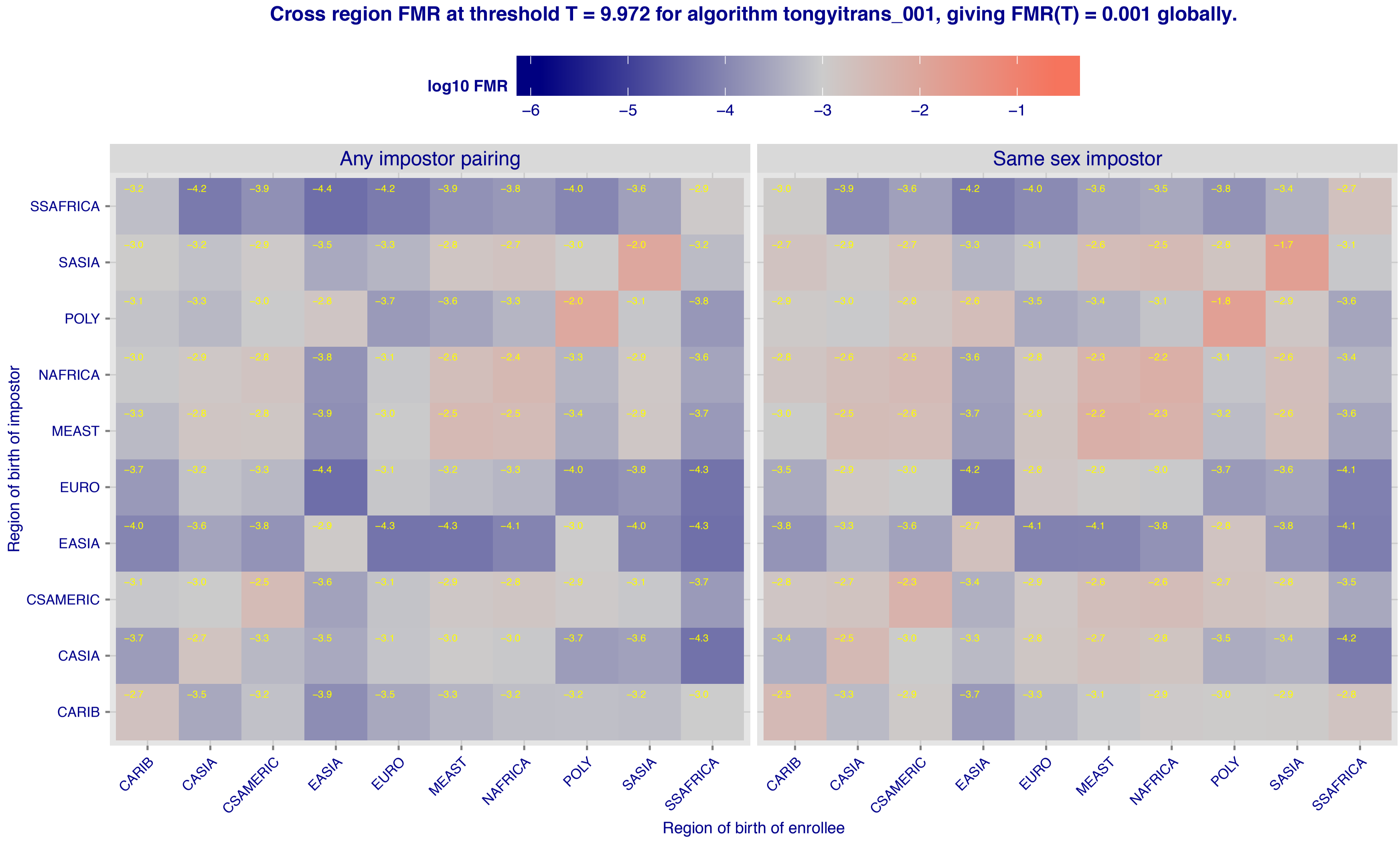


Figure 17: For algorithm tongyitrans-001 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given region pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all $O(10^{10})$ impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

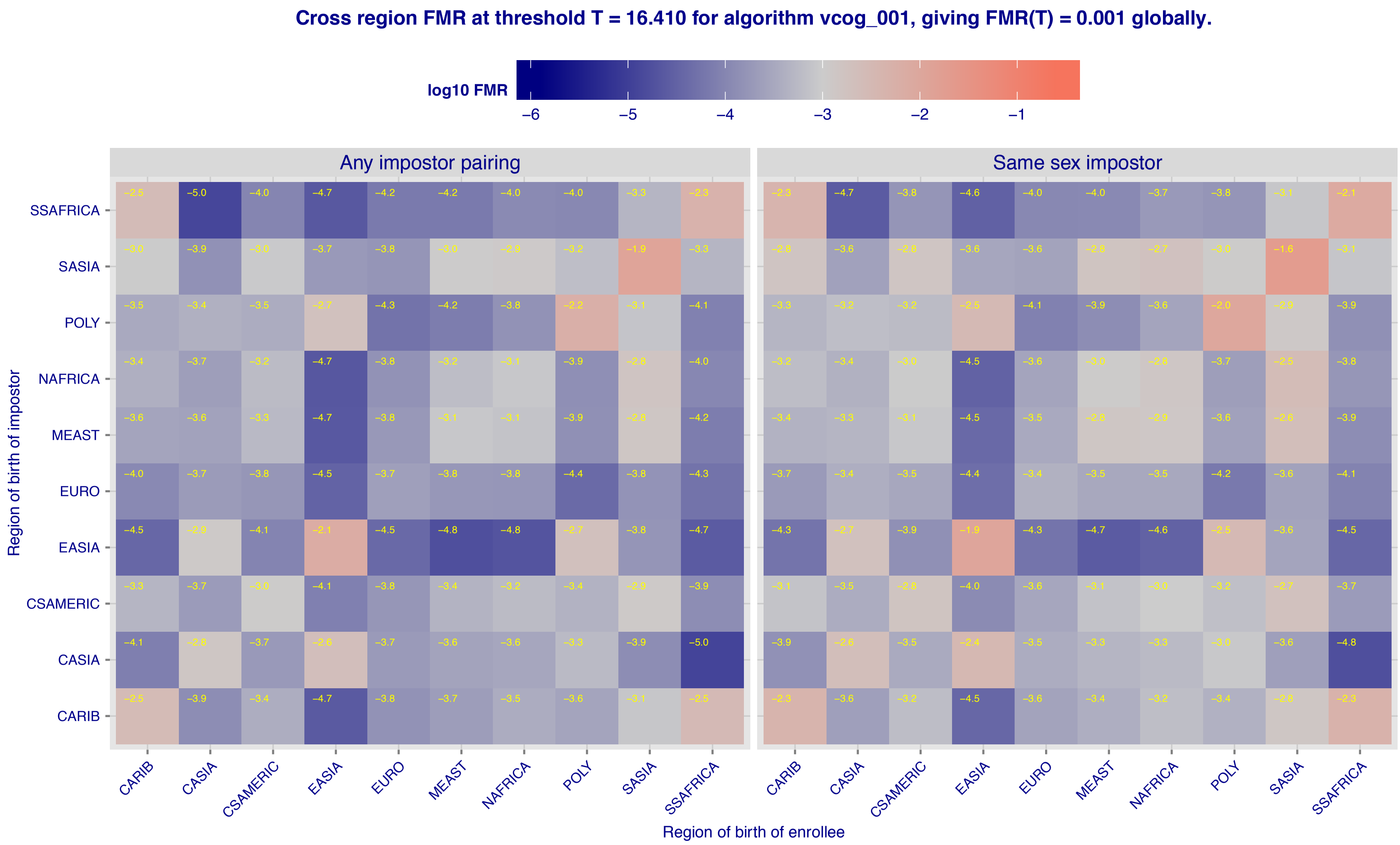


Figure 18: For algorithm vcog-001 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given region pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

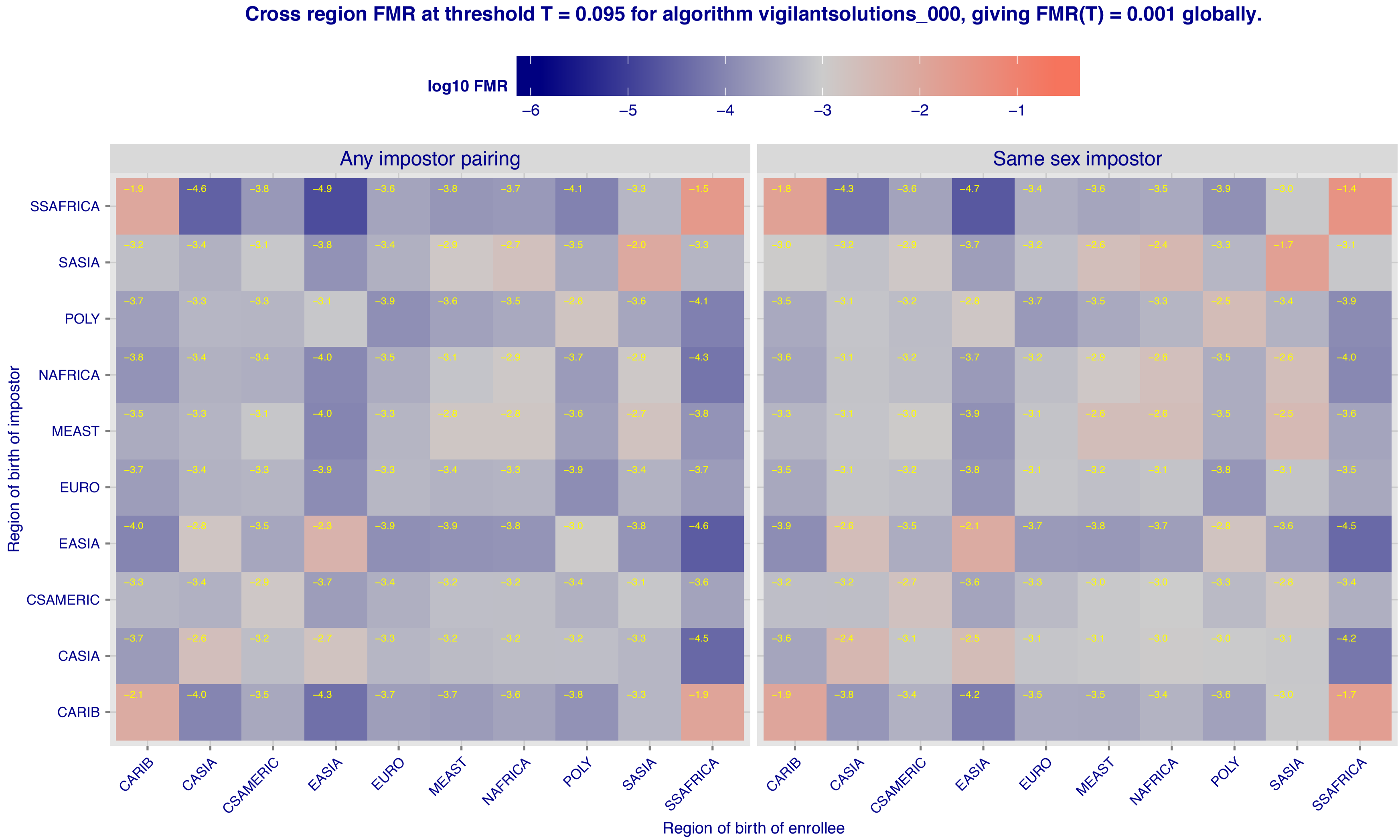


Figure 19: For algorithm vigilantsolutions-000 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given region pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

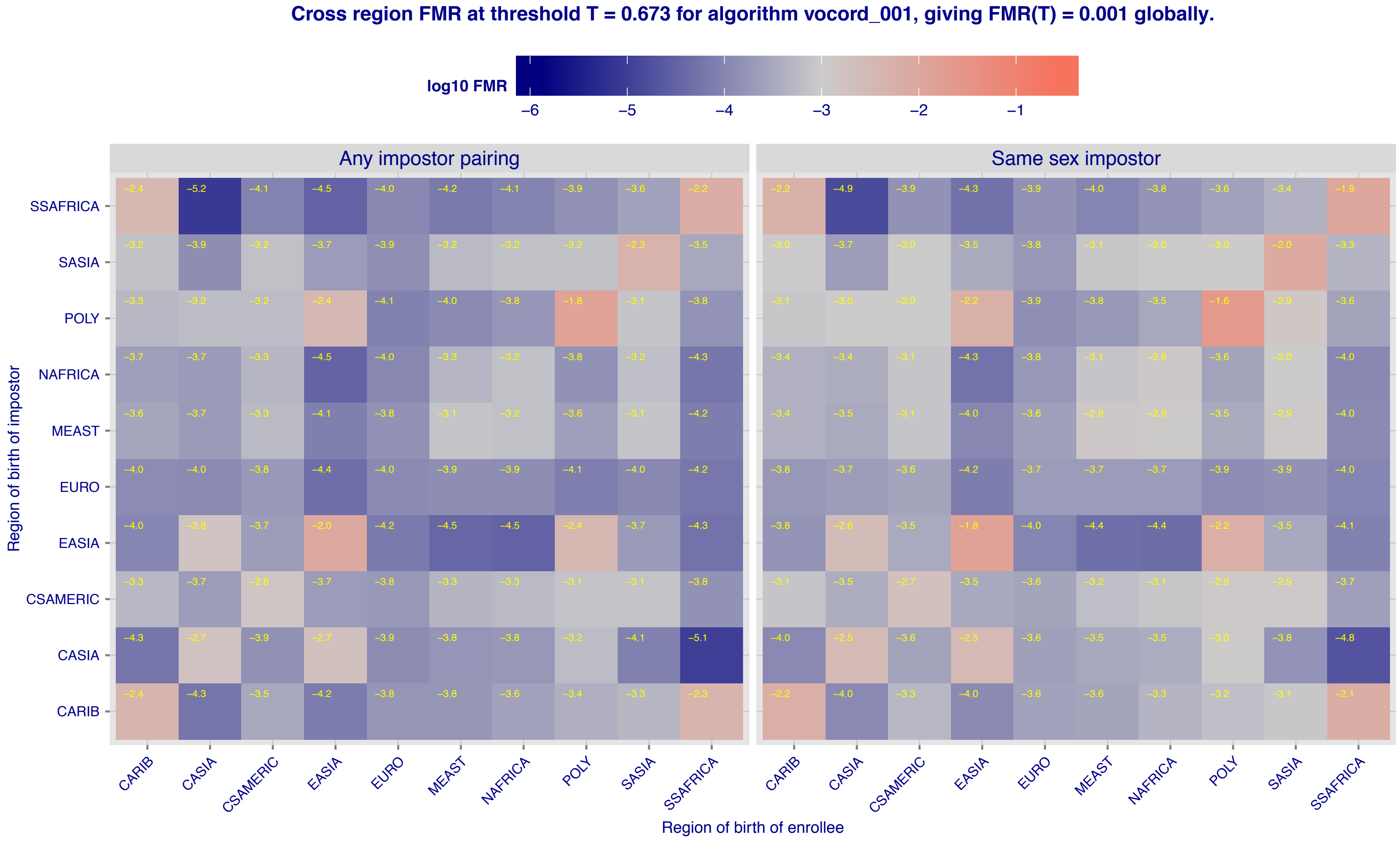


Figure 20: For algorithm vocord-001 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given region pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all $O(10^{10})$ impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

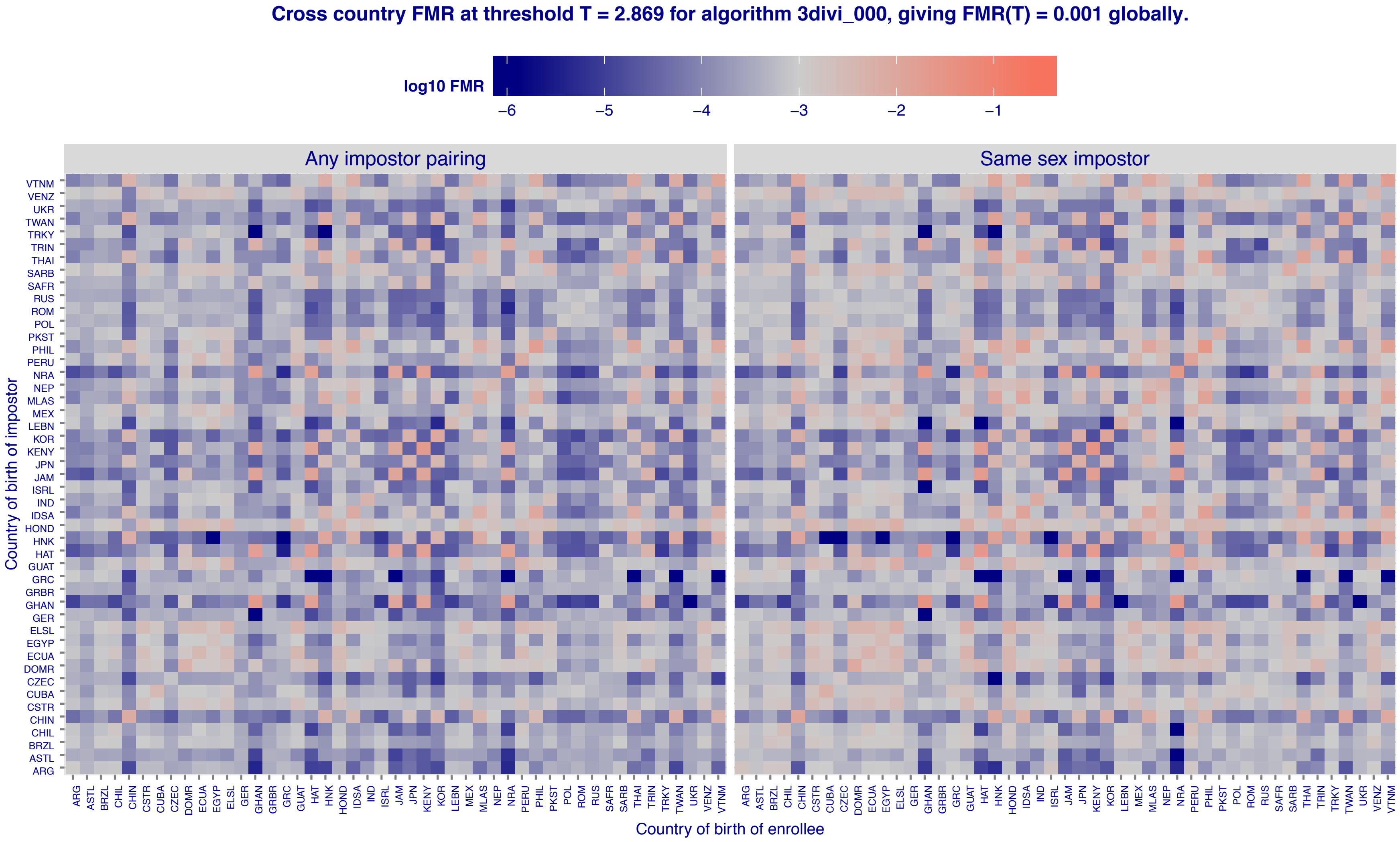


Figure 21: For algorithm 3divi-000 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given country pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all $O(10^{10})$ impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

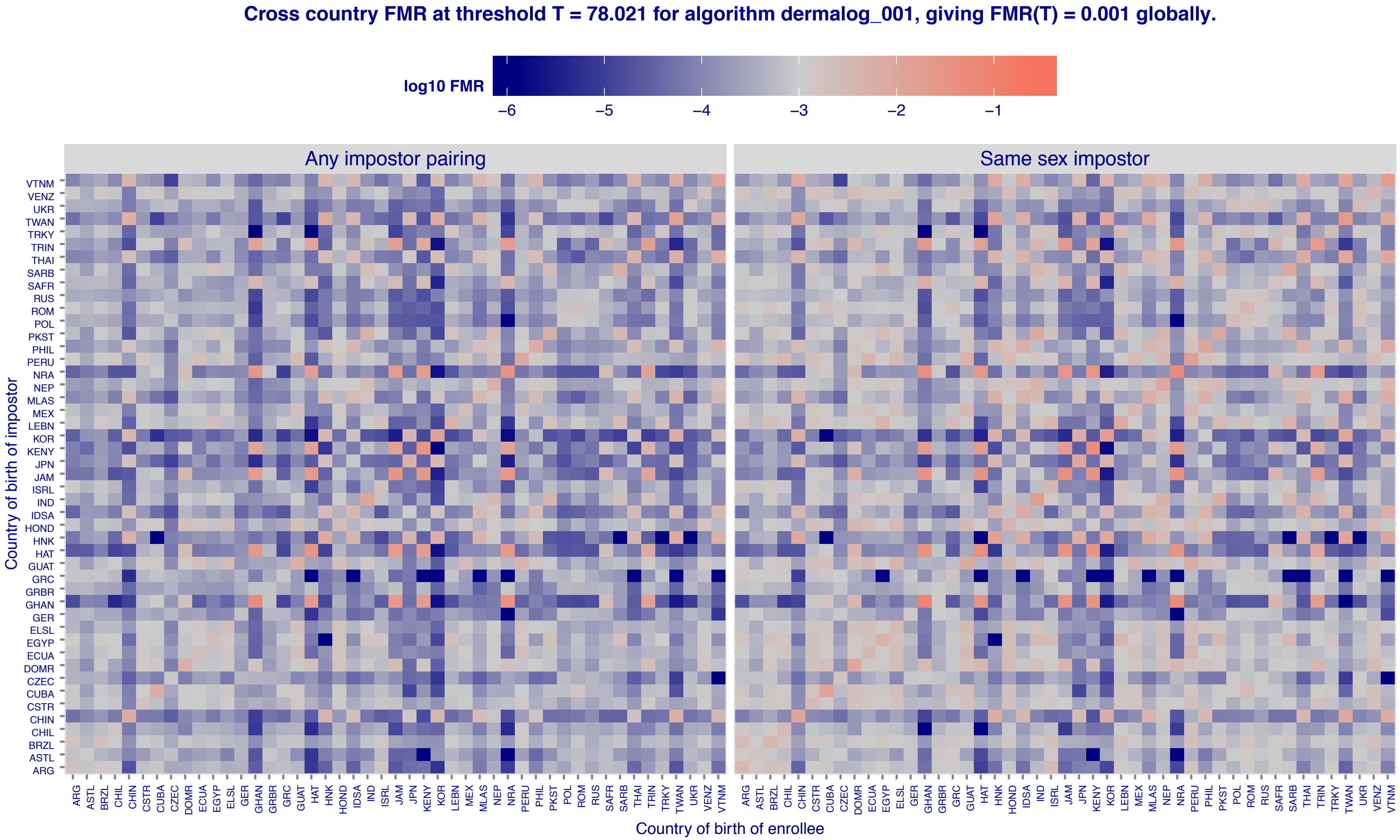


Figure 22: For algorithm dermalog-001 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given country pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

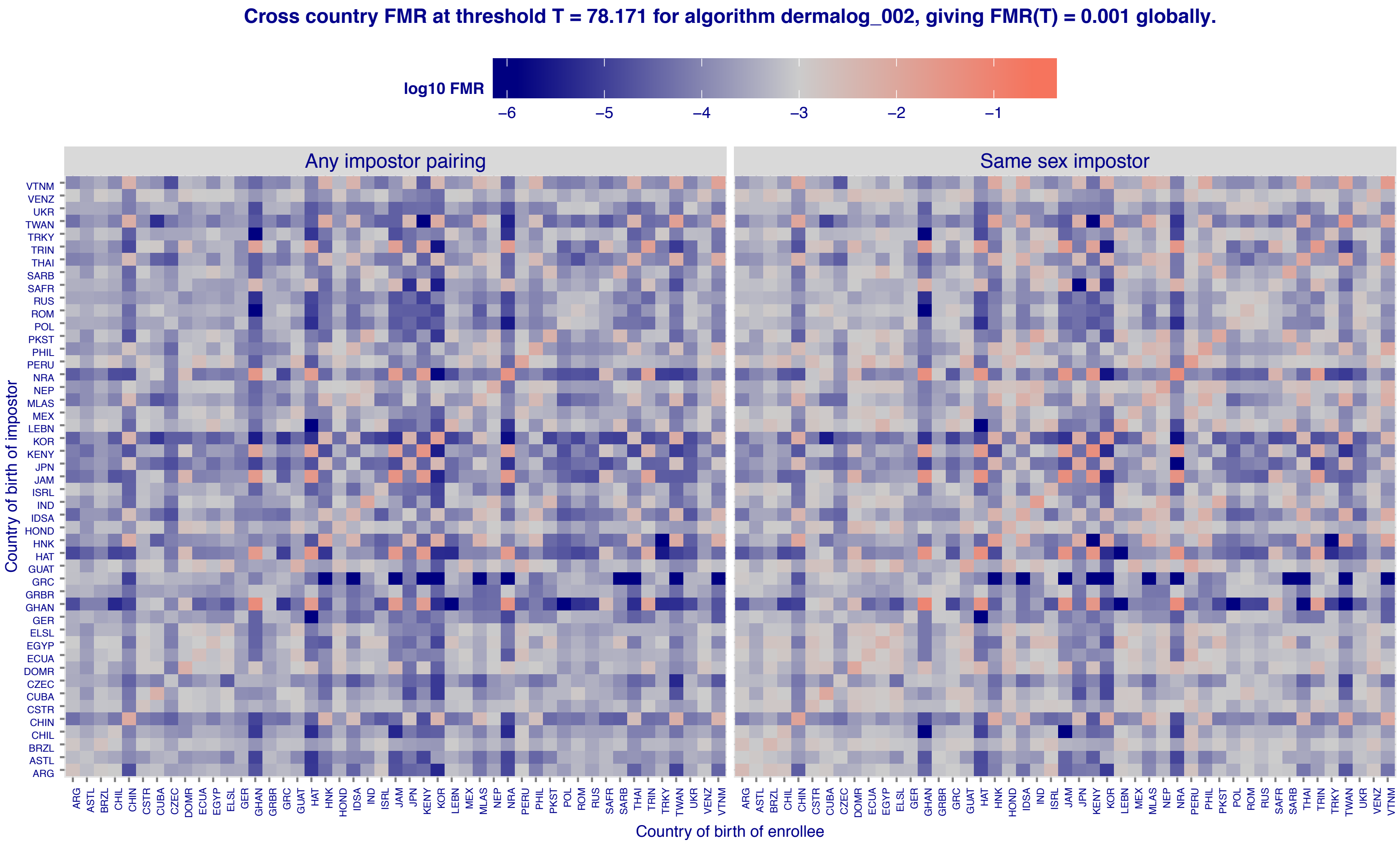


Figure 23: For algorithm dermalog-002 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given country pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all $O(10^{10})$ impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

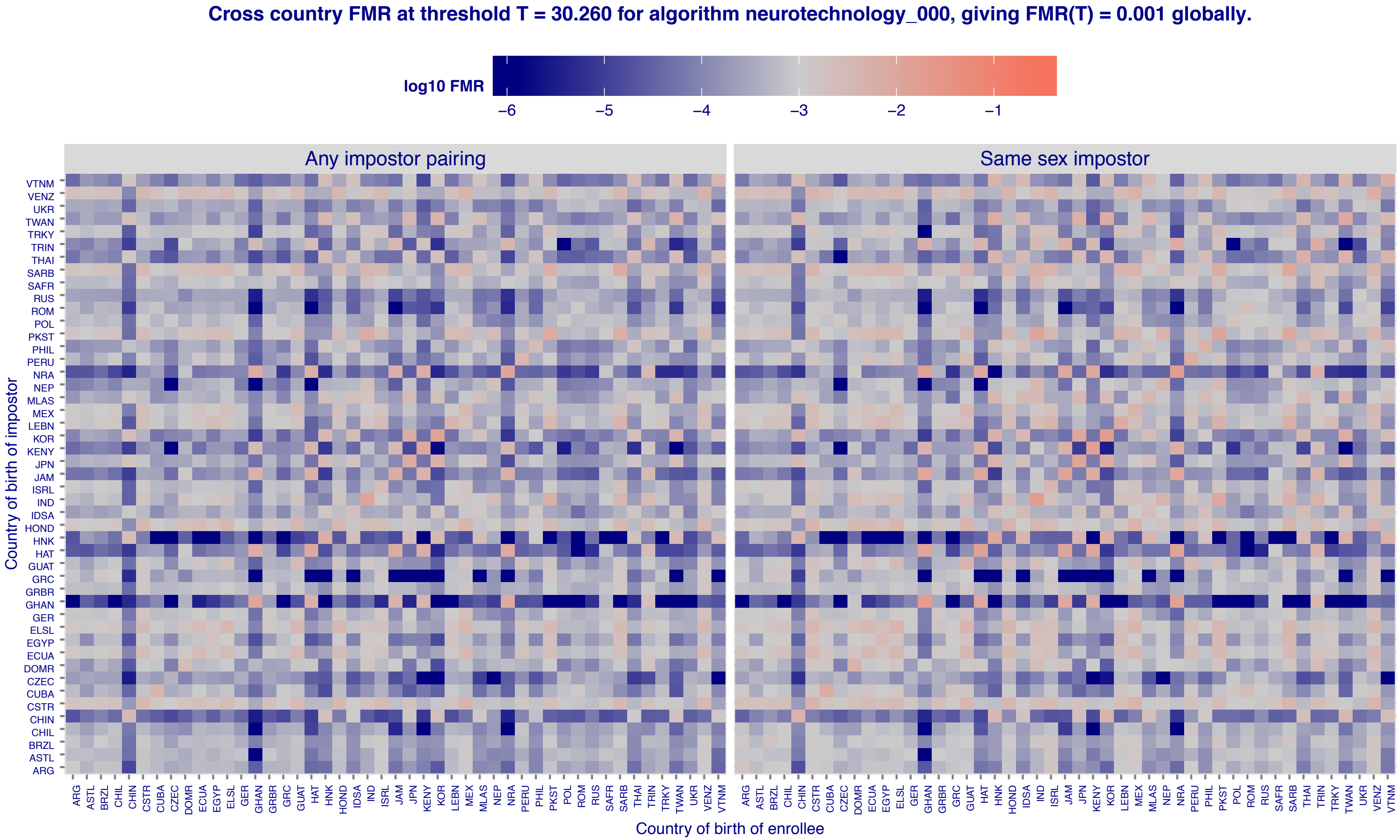


Figure 24: For algorithm neurotechnology-000 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given country pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all $O(10^{10})$ impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

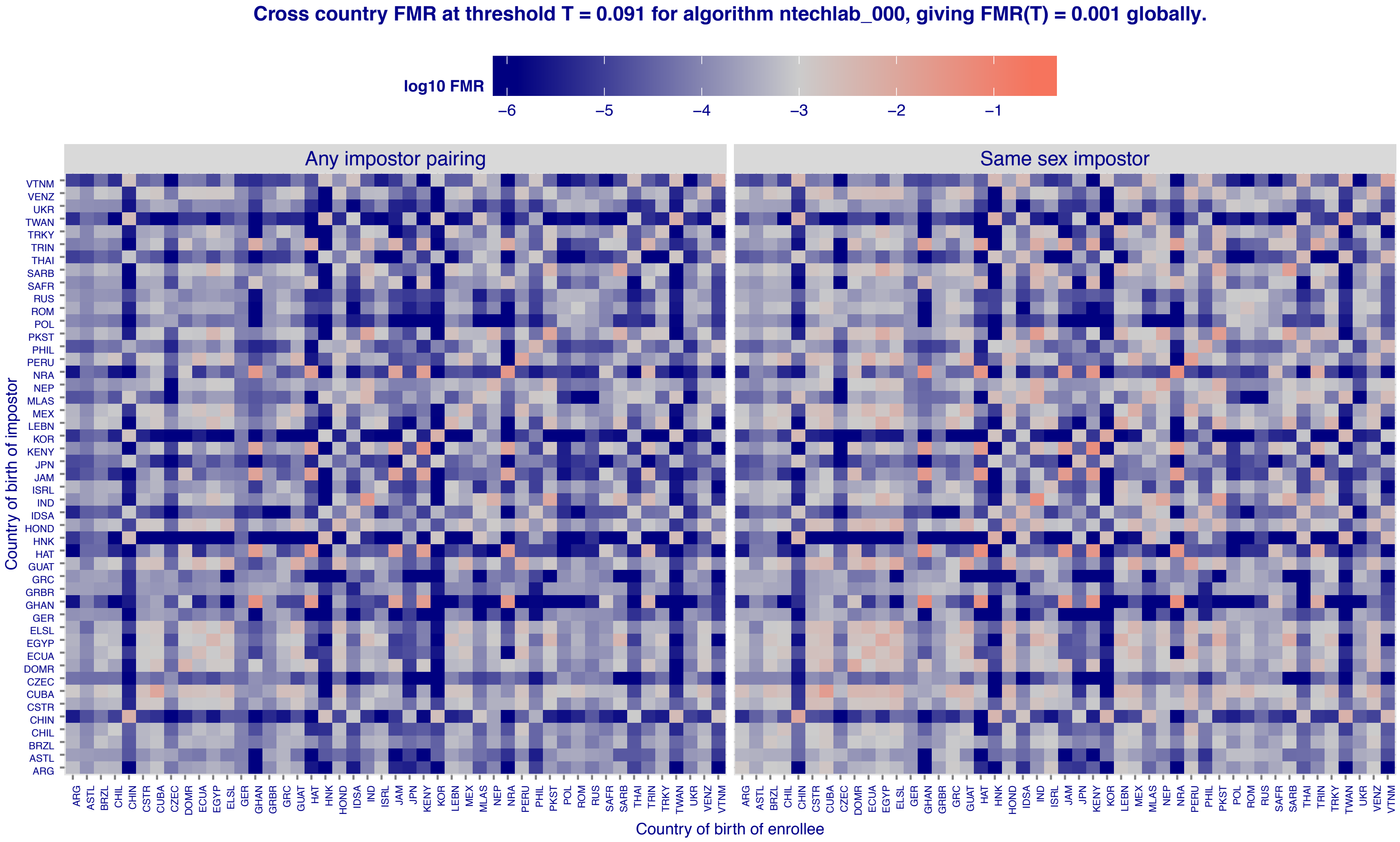


Figure 25: For algorithm ntechlab-000 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given country pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all $O(10^{10})$ impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

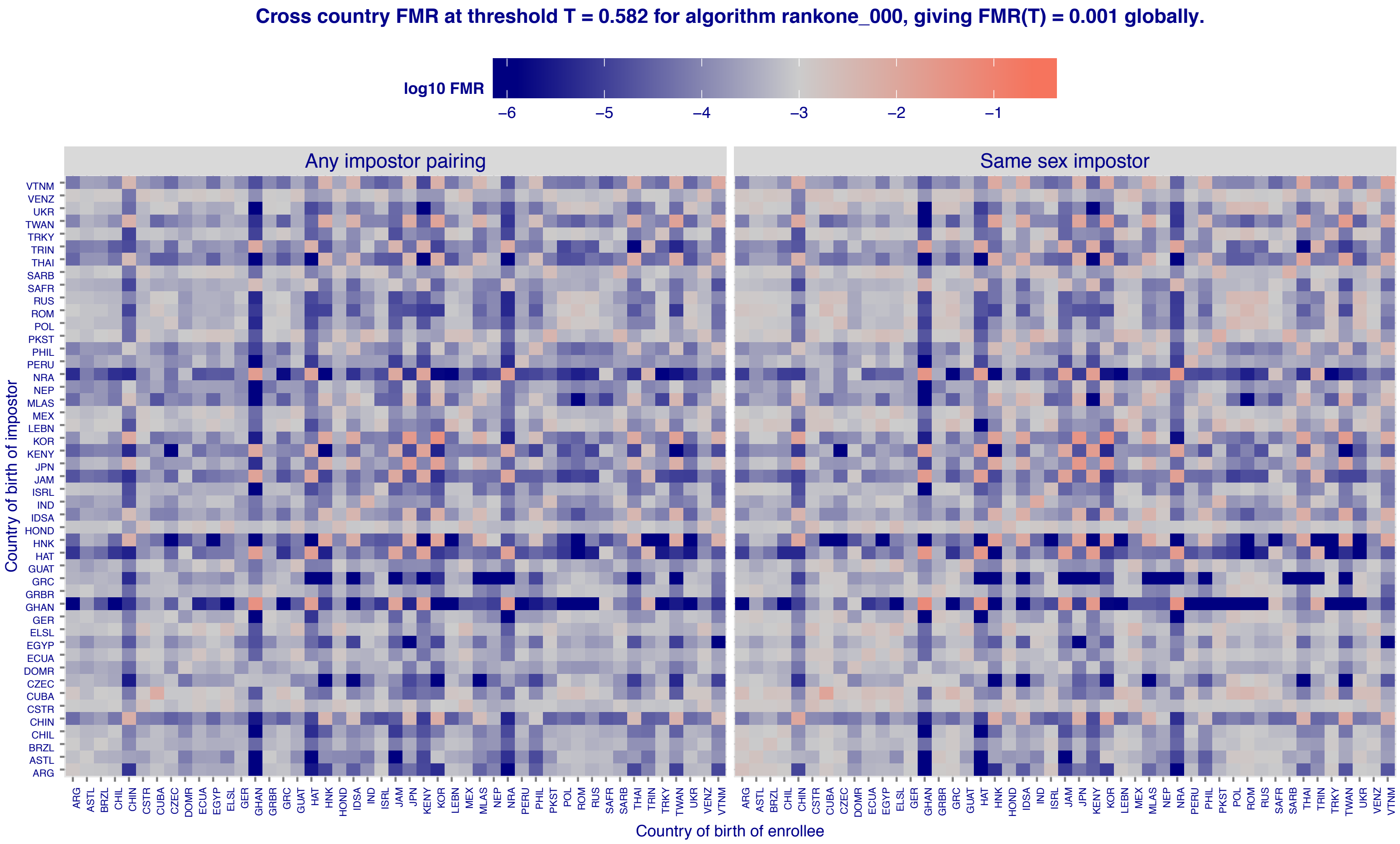


Figure 26: For algorithm rankone-000 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given country pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all $O(10^{10})$ impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

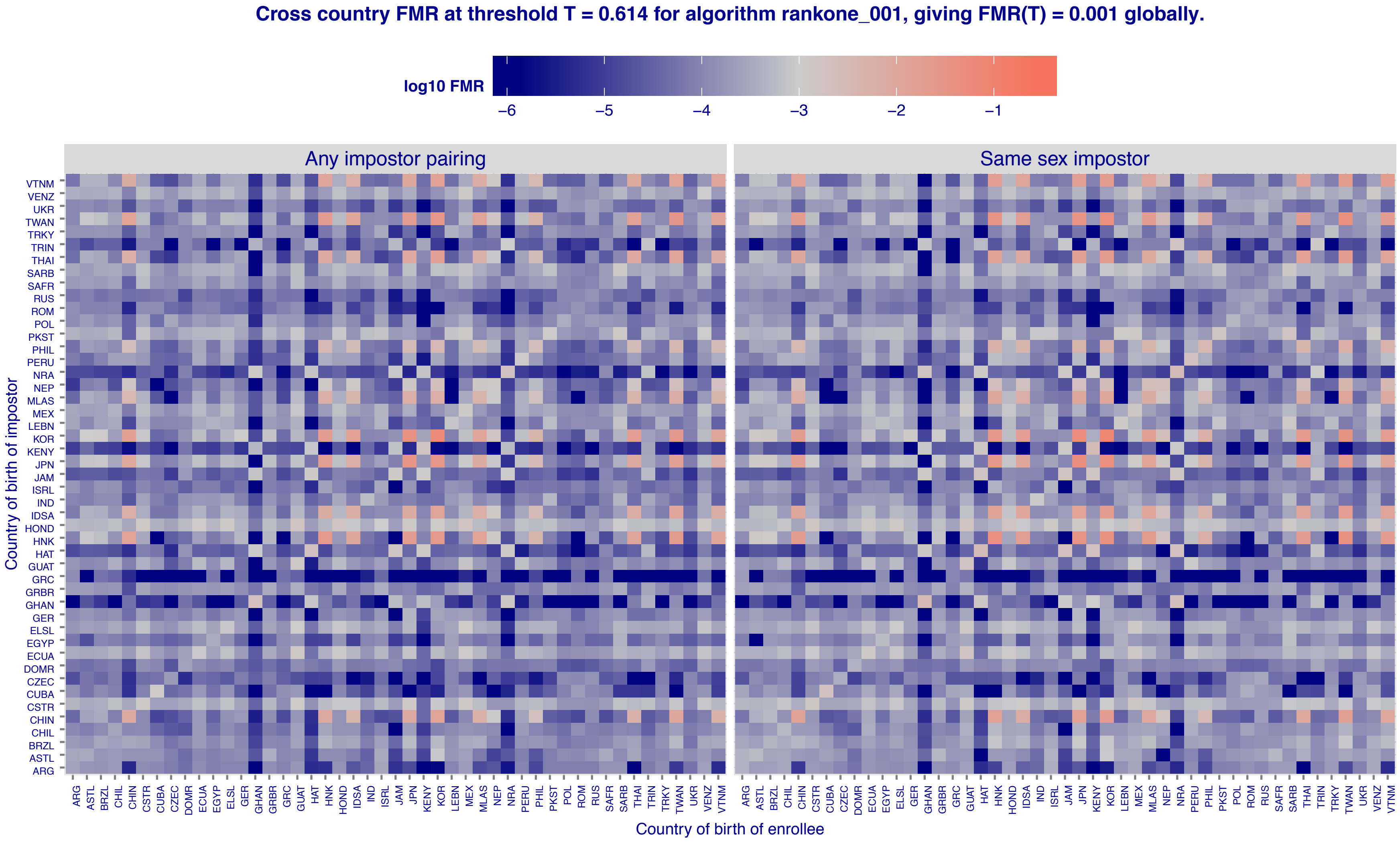


Figure 27: For algorithm rankone-001 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given country pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

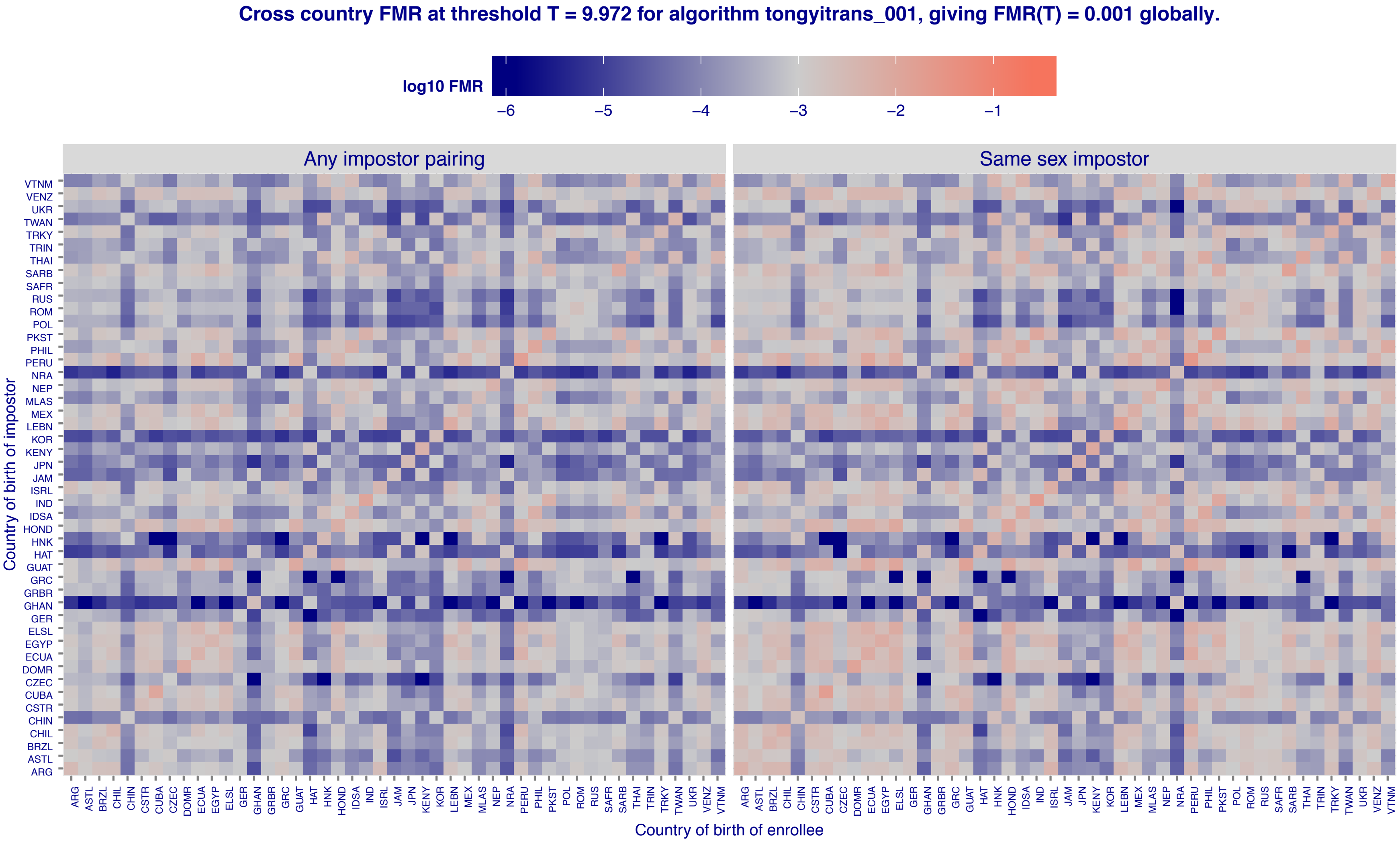


Figure 28: For algorithm tongyitrans-001 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given country pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all $O(10^{10})$ impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

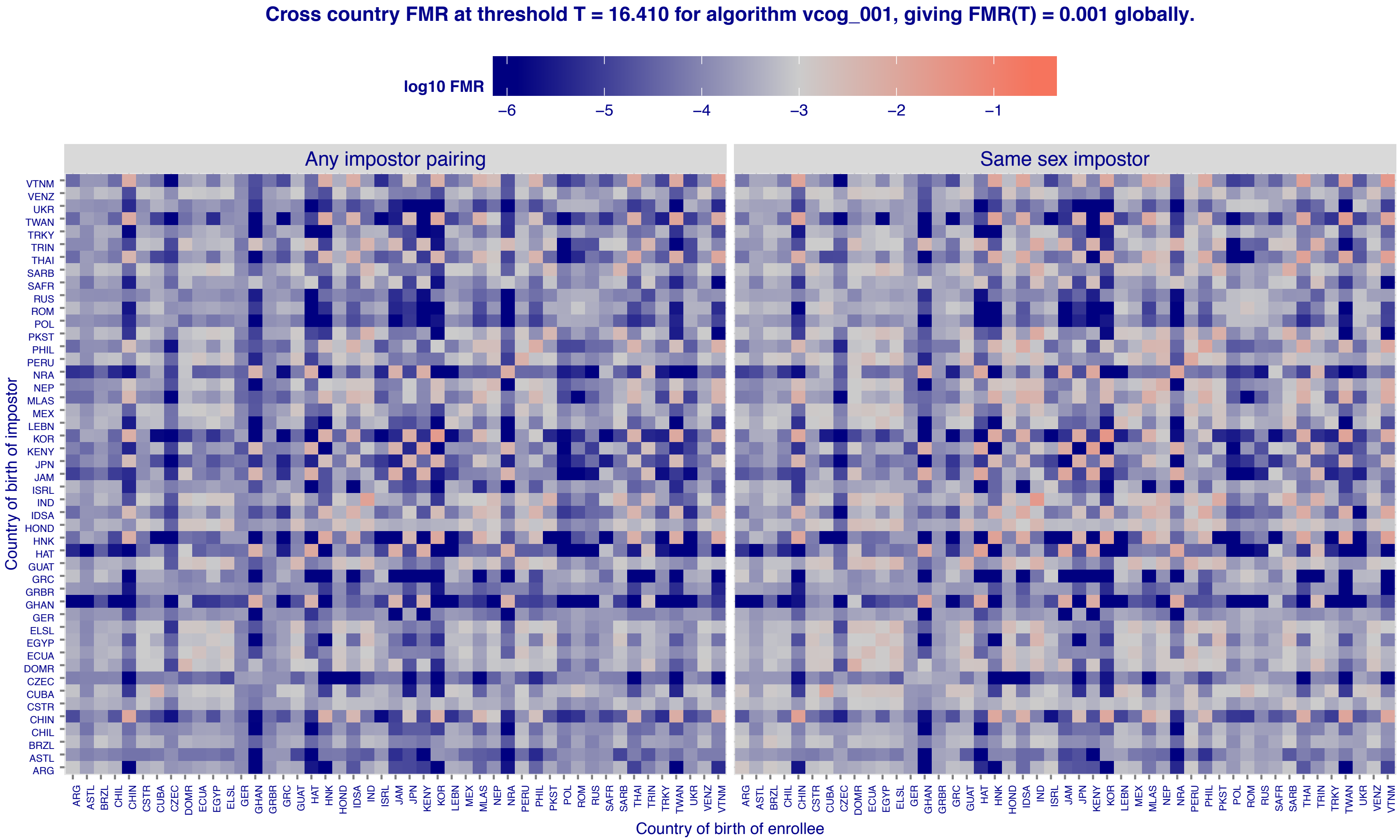


Figure 29: For algorithm vcog-001 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given country pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

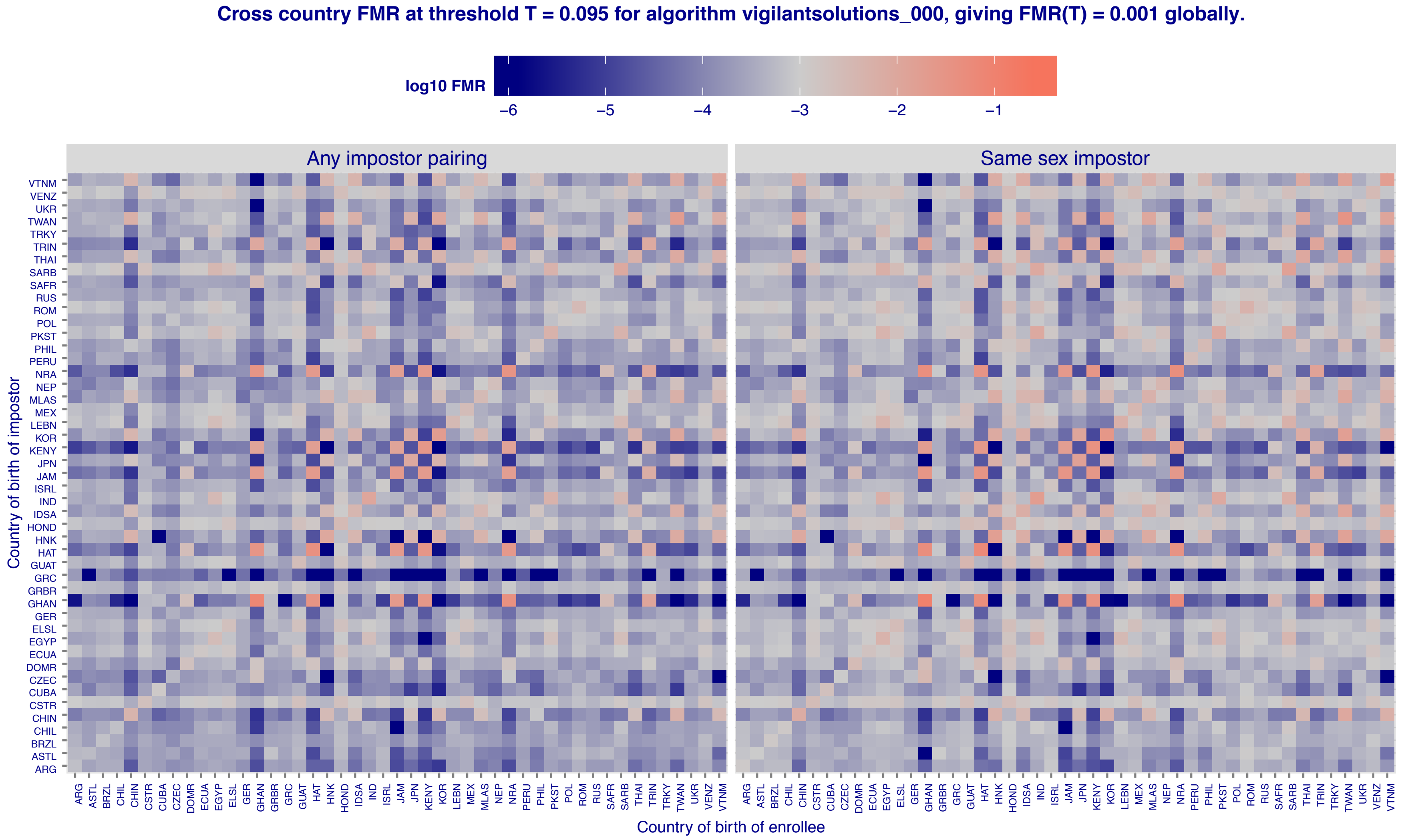


Figure 30: For algorithm vigilantsolutions-000 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who were born in the given country pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all $O(10^{10})$ impostor comparisons. If text appears in each box it give the same quantity as that coded by the color. Light red colors present a security vulnerability to, for example, a passport gate.

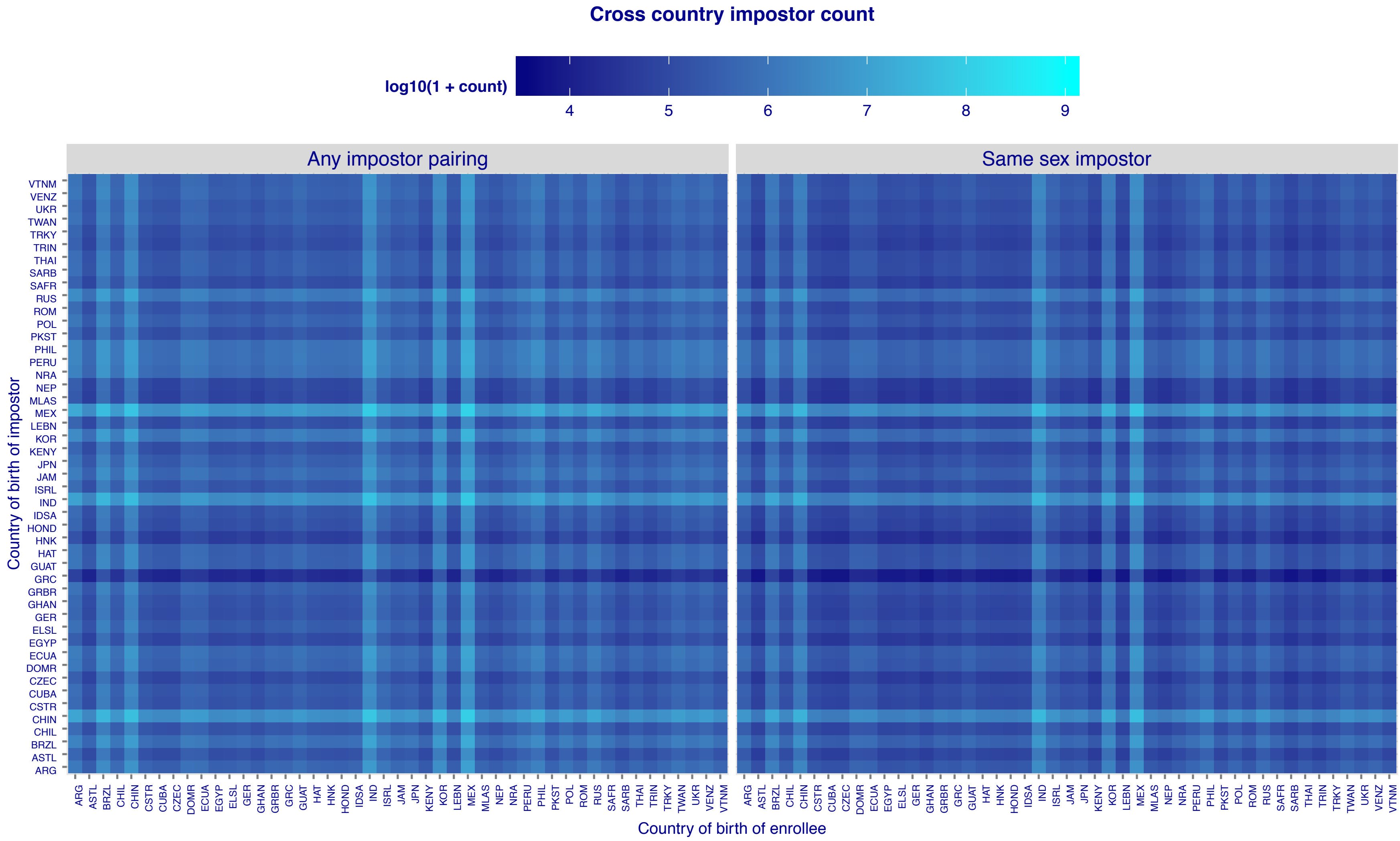


Figure 31: For visa images, the heatmap shows The count of impostor comparisons of faces from different individuals who were born in the given country pair.

3.3.2 Effect of age on impostors

Background: This section shows the effect of age on the impostor distribution. The ideal behaviour is that the age of the enrollee and the impostor would not affect impostor scores. This would support FMR stability over sub-populations.

Goals:

- ▷ To show the effect of relative ages of the impostor and enrollee on false match rates.
- ▷ To determine whether some algorithms have better impostor distribution stability.

Methods:

- ▷ Define 14 age group bins, spanning 0 to over 100 years old.
- ▷ Compute FMR over all impostor comparisons for which the subjects in the enrollee and impostor images have ages in two bins.
- ▷ Compute FMR over all impostor comparisons for which the subjects are additionally of the same sex, and born in the same geographic region.

Results:

The notable aspects are:

- ▷ Diagonal dominance: Impostors are more likely to be matched against their same age group.
- ▷ Same sex and same region impostors are more successful. On the diagonal, an impostor is more likely to succeed by posing as someone of the same sex. If $\Delta \log_{10}$ FMR $= 0.2$, then same-sex same-region FMR exceeds the all-pairs FMR by factor of $10^{0.2} = 1.6$.
- ▷ Young children impostors give elevated FMR against young children. Older adult impostor give elevated FMR against older adults. These effects are quite large, for example if $\Delta \log_{10}$ FMR $= 1.0$ larger than a 32 year old, then these groups have higher FMR by a factor of $10^1 = 10$. This would imply an FMR above 0.01 for a nominal (global) FMR = 0.001.
- ▷ Algorithms vary.
- ▷ We computed the same quantities for a global FMR = 0.0001. The effects are similar.

Note the calculations in this section include impostors paired across all countries of birth.

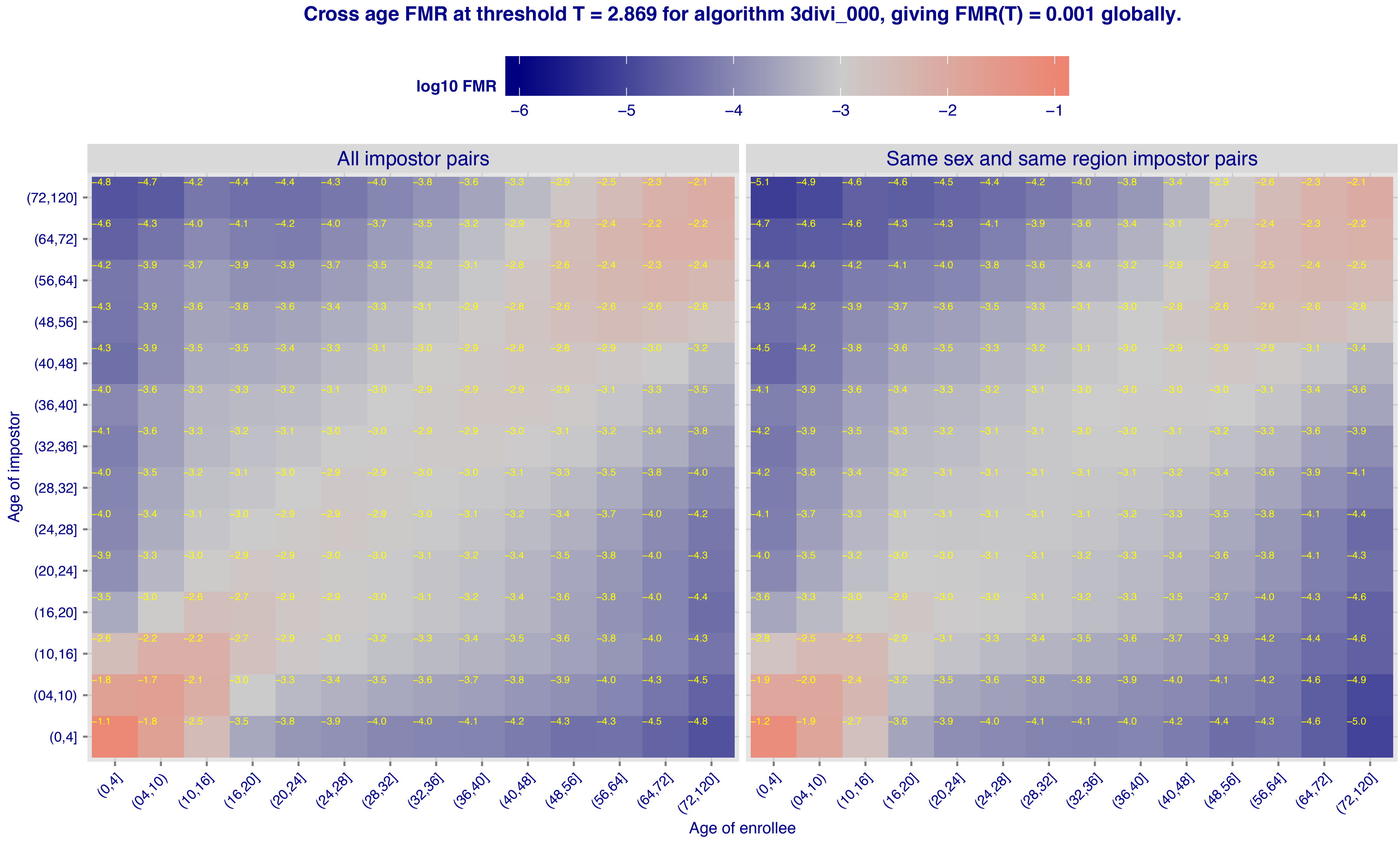


Figure 32: For algorithm 3divi-000 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who have the given age pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. The text in each box it give the same quantity as that coded by the color. Light colors present a security vulnerability to, for example, a passport gate.

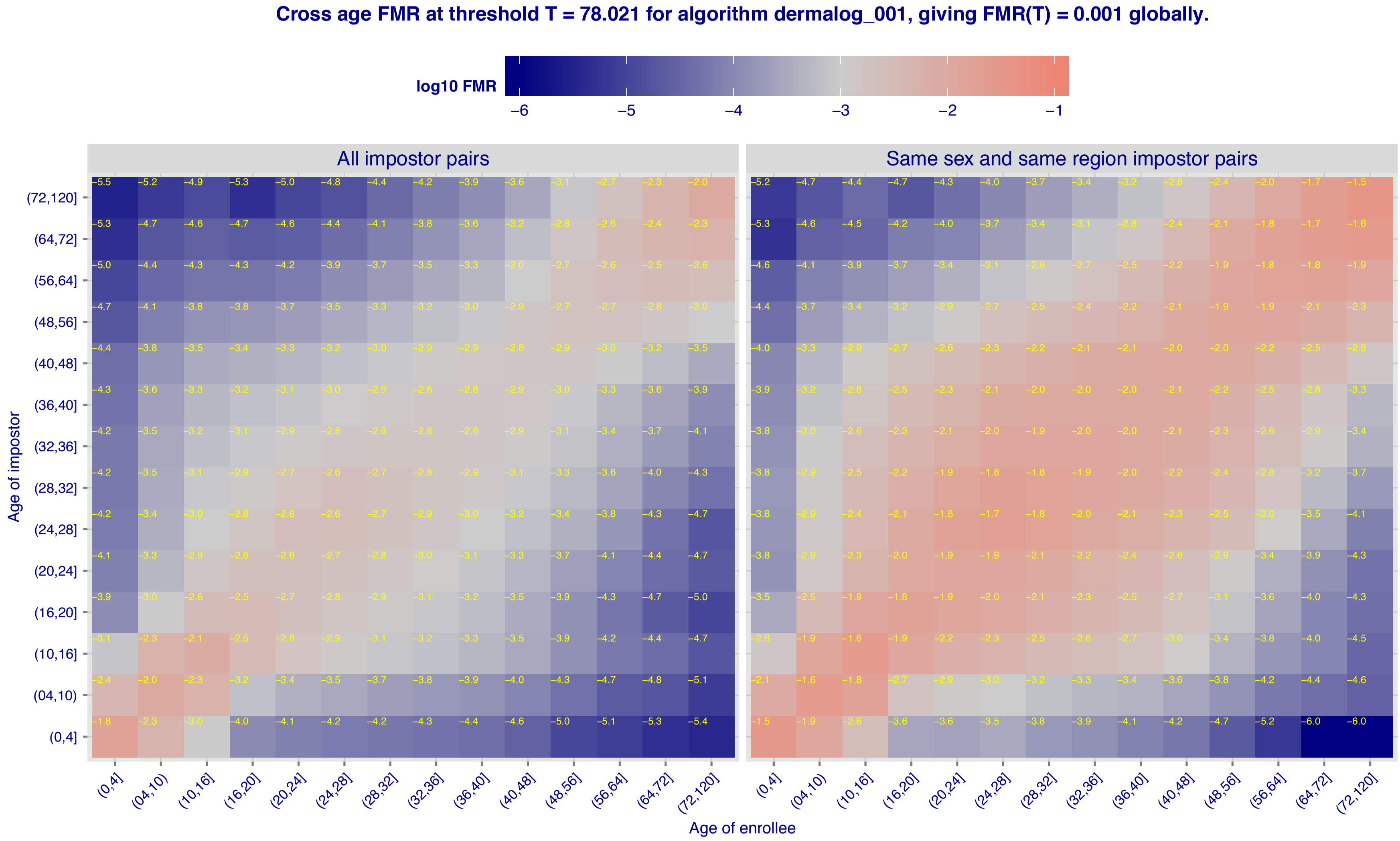


Figure 33: For algorithm dermalog-001 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who have the given age pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. The text in each box it give the same quantity as that coded by the color. Light colors present a security vulnerability to, for example, a passport gate.

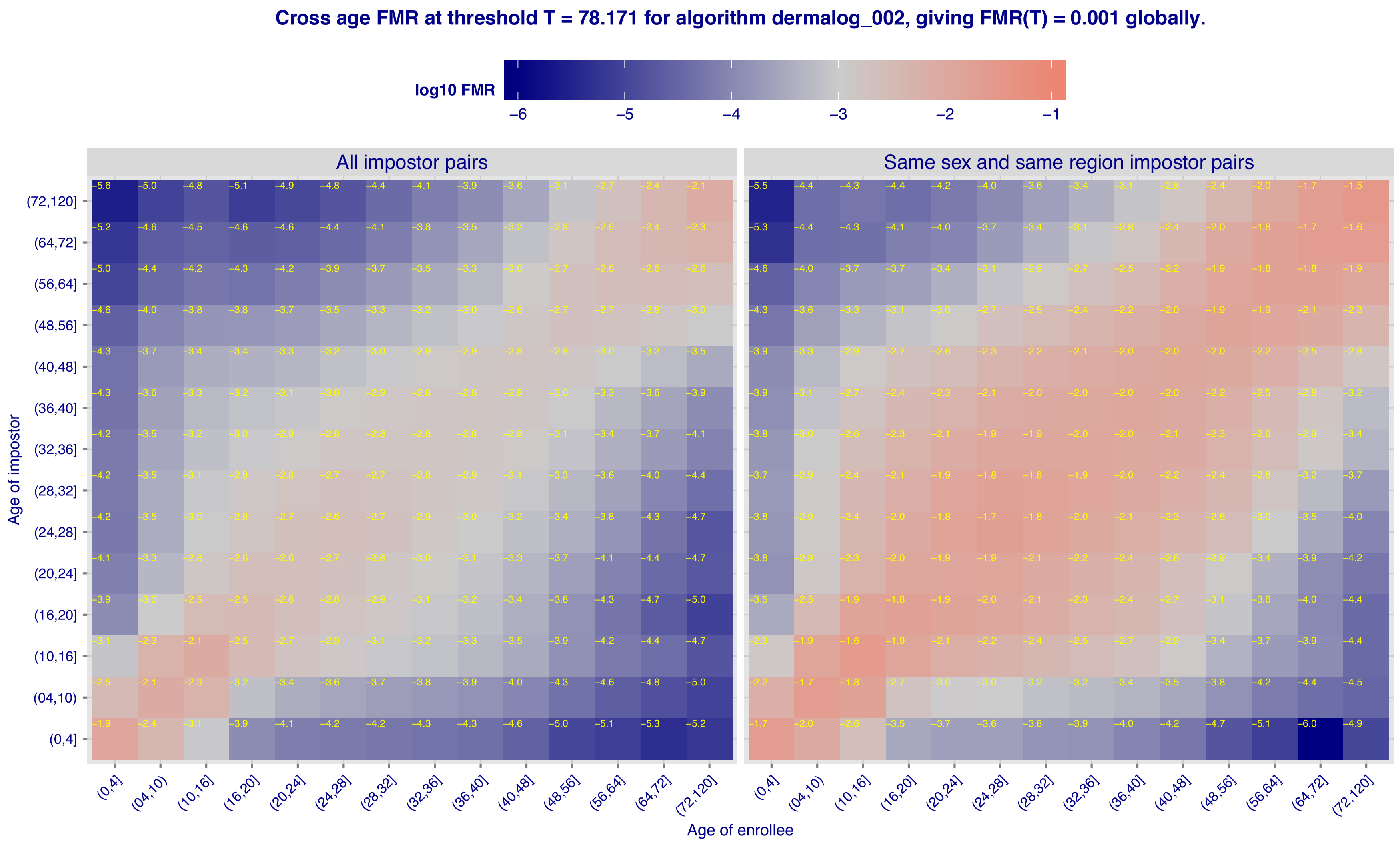


Figure 34: For algorithm dermalog-002 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who have the given age pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. The text in each box it give the same quantity as that coded by the color. Light colors present a security vulnerability to, for example, a passport gate.

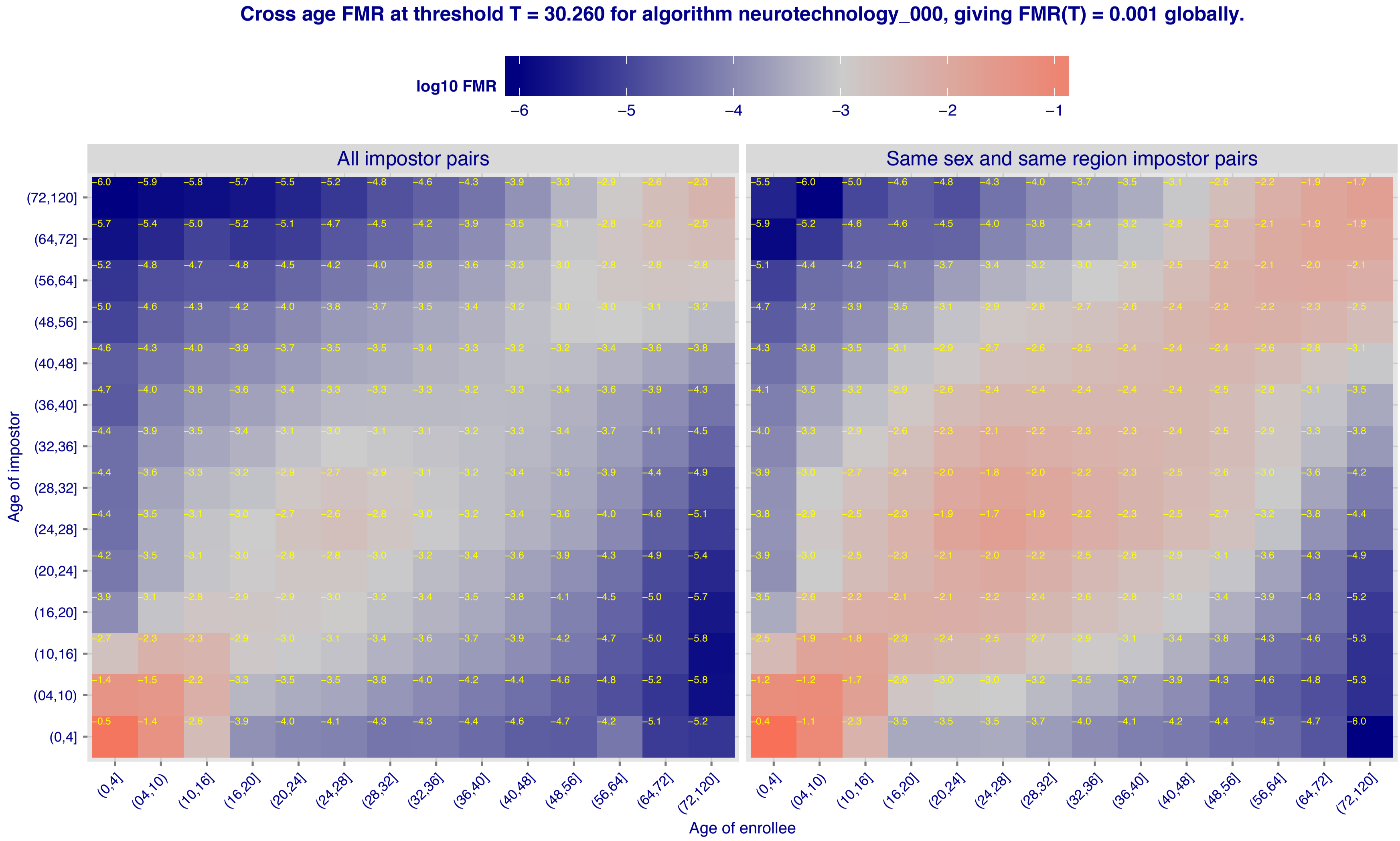


Figure 35: For algorithm neurotechnology-000 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who have the given age pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. The text in each box it give the same quantity as that coded by the color. Light colors present a security vulnerability to, for example, a passport gate.

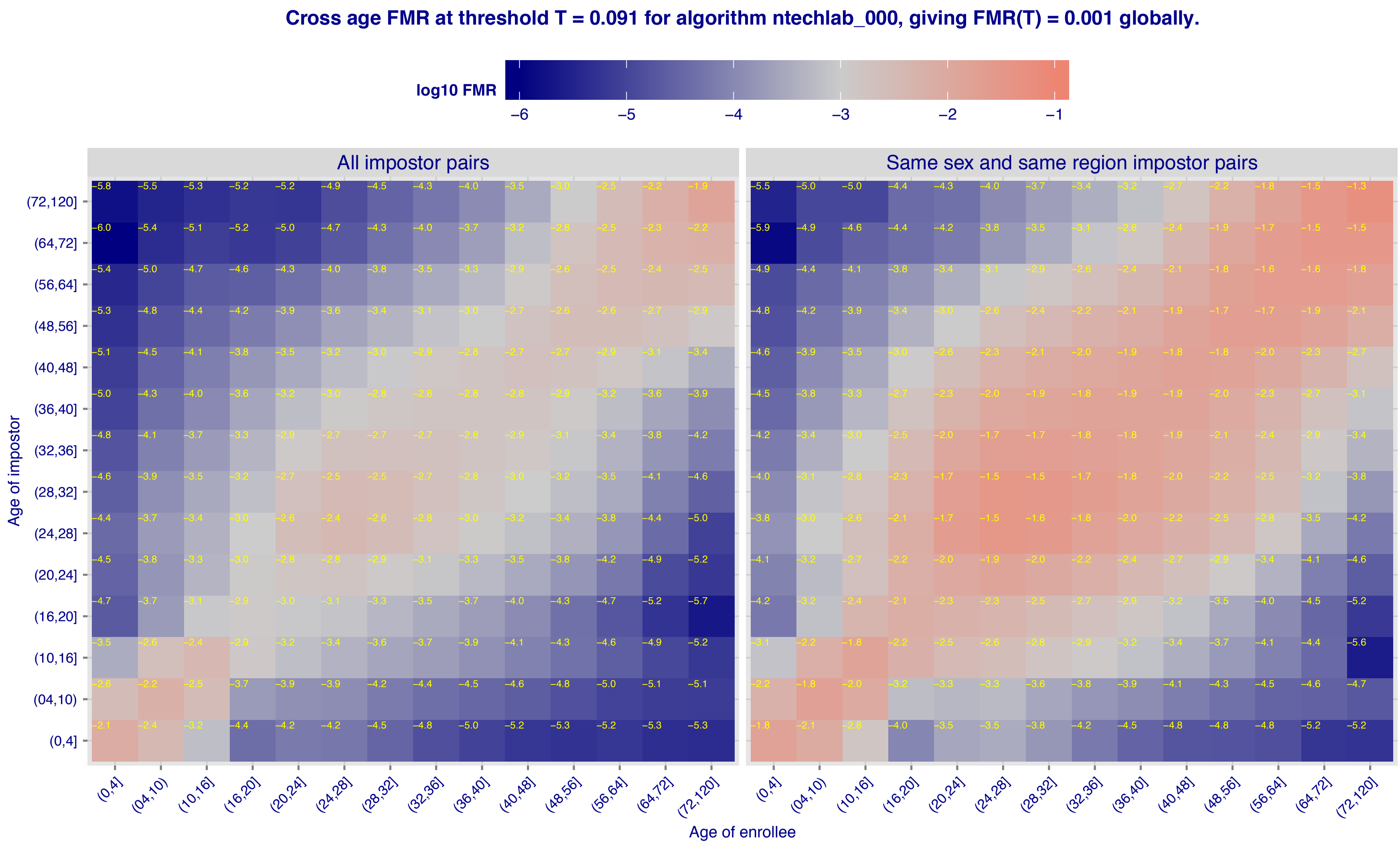


Figure 36: For algorithm ntechlab-000 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who have the given age pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. The text in each box it give the same quantity as that coded by the color. Light colors present a security vulnerability to, for example, a passport gate.

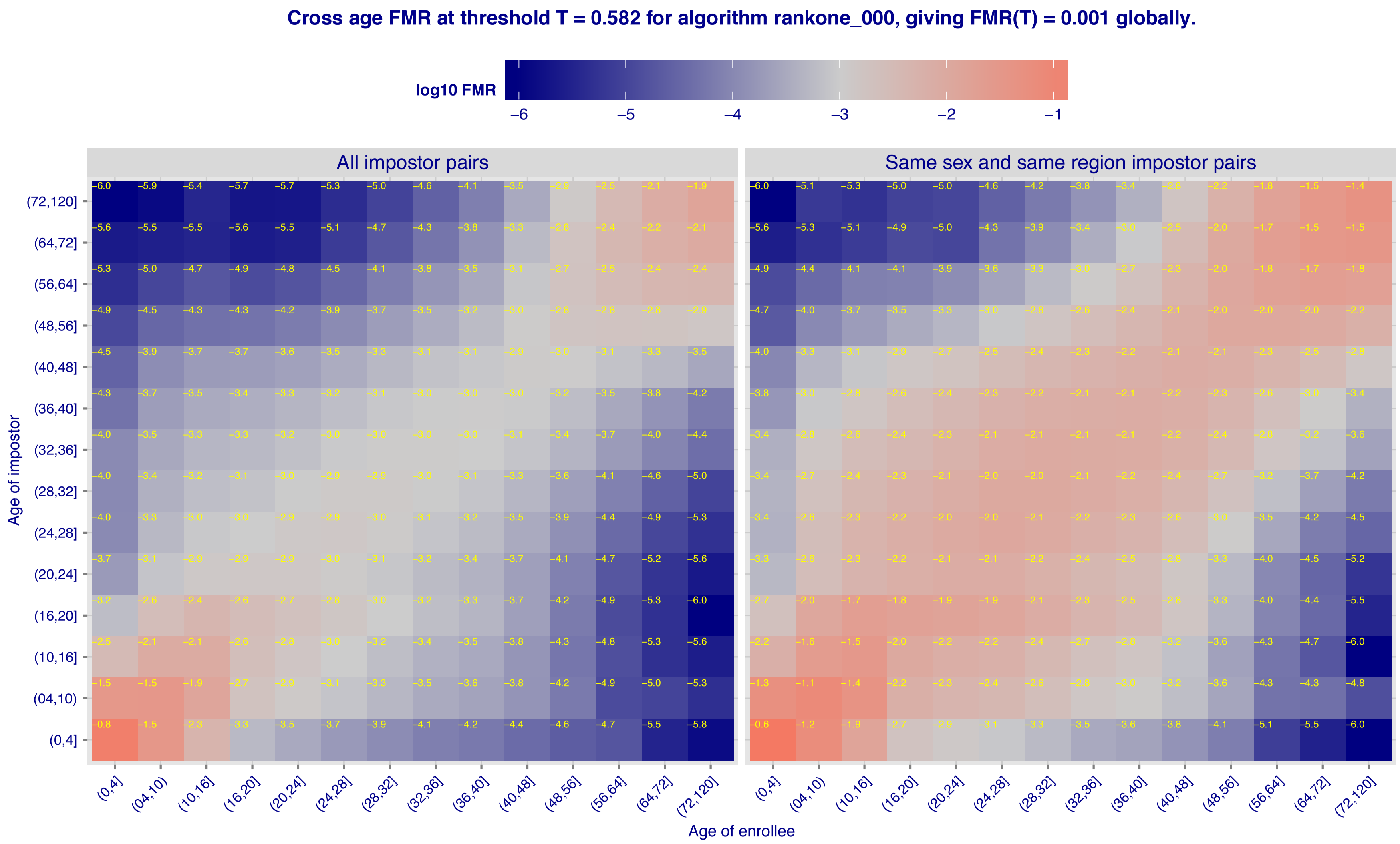


Figure 37: For algorithm rankone-000 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who have the given age pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. The text in each box it give the same quantity as that coded by the color. Light colors present a security vulnerability to, for example, a passport gate.

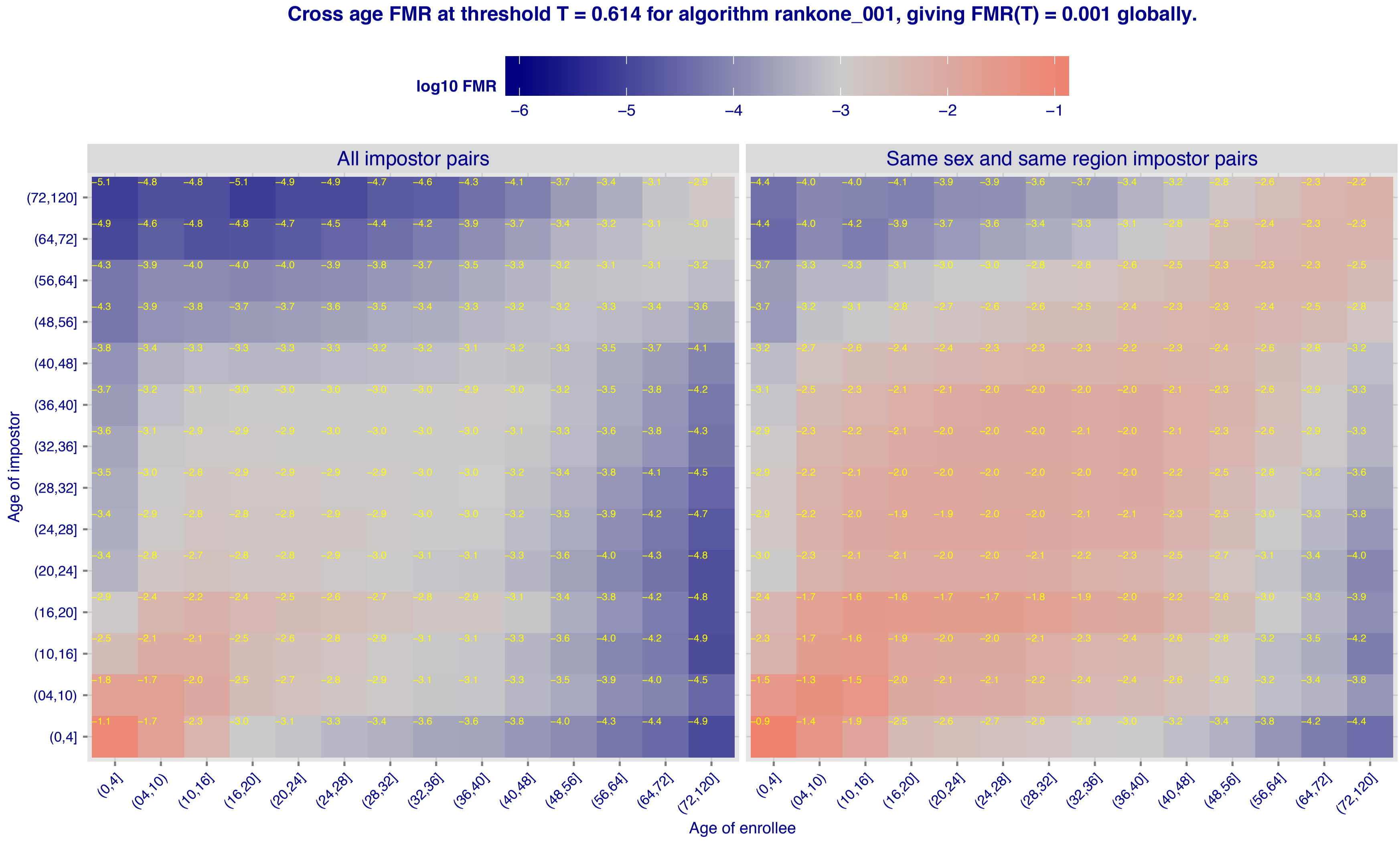


Figure 38: For algorithm rankone-001 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who have the given age pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. The text in each box it give the same quantity as that coded by the color. Light colors present a security vulnerability to, for example, a passport gate.

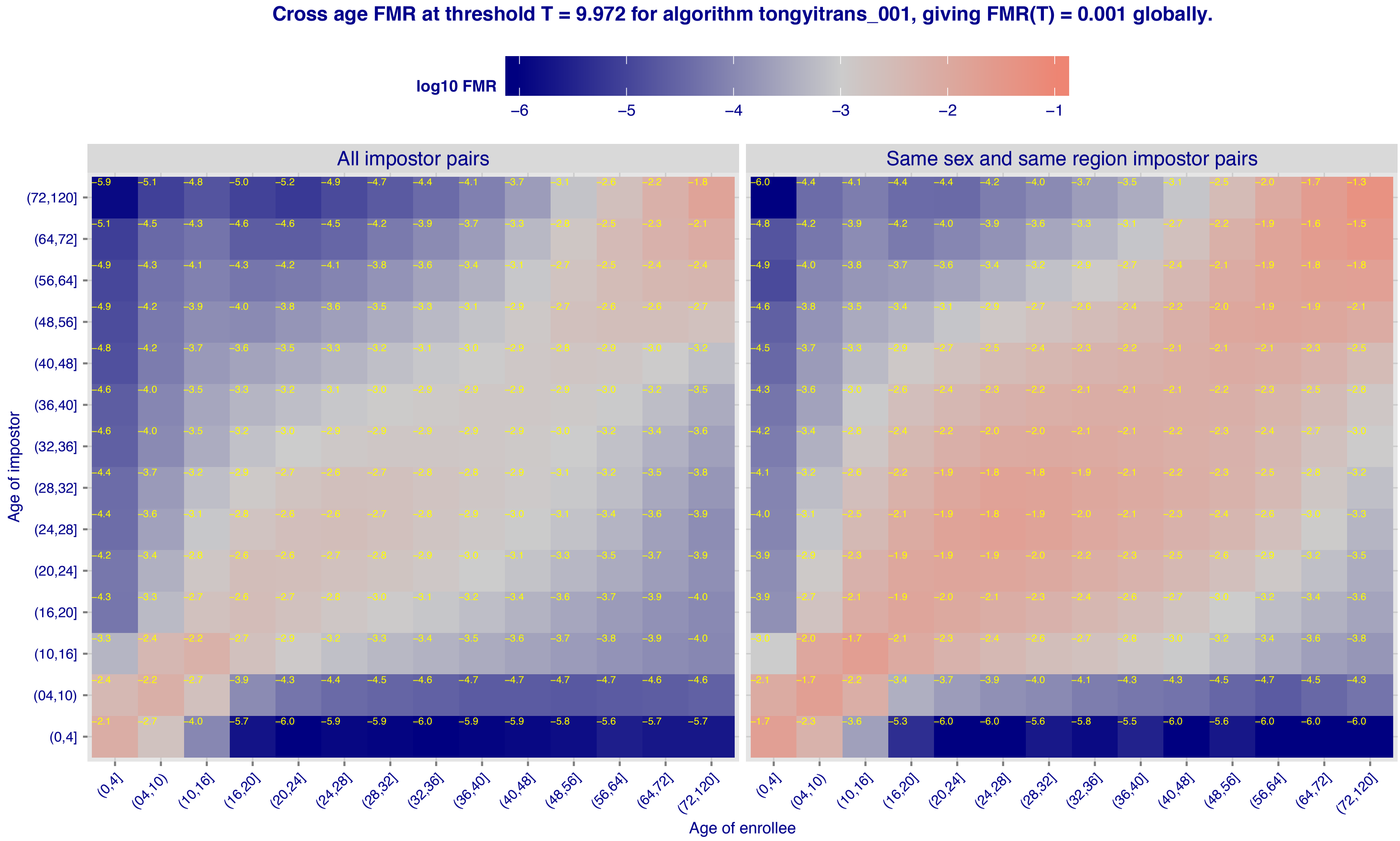


Figure 39: For algorithm tongyitrans-001 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who have the given age pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. The text in each box it give the same quantity as that coded by the color. Light colors present a security vulnerability to, for example, a passport gate.

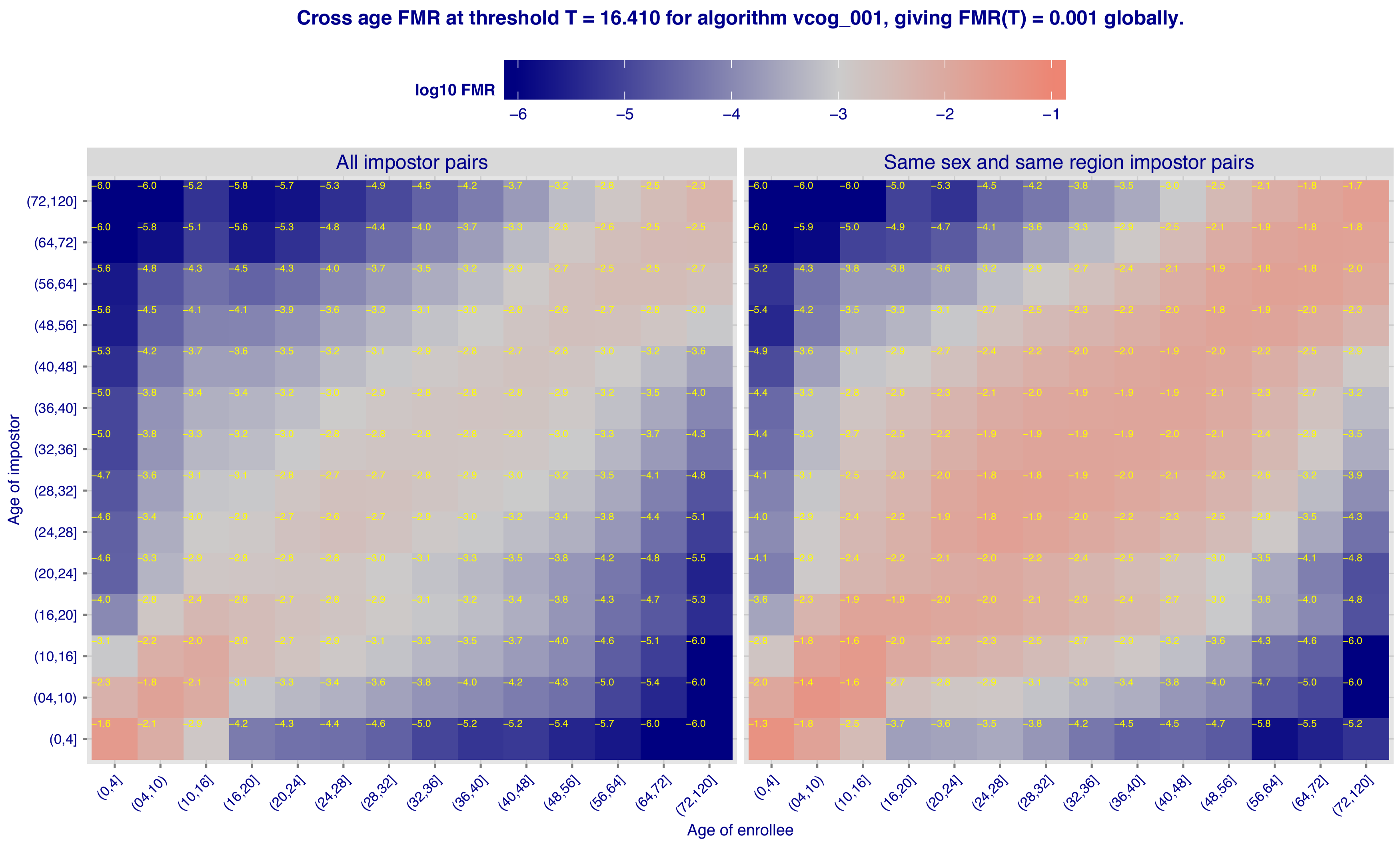


Figure 40: For algorithm vcog-001 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who have the given age pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. The text in each box it give the same quantity as that coded by the color. Light colors present a security vulnerability to, for example, a passport gate.

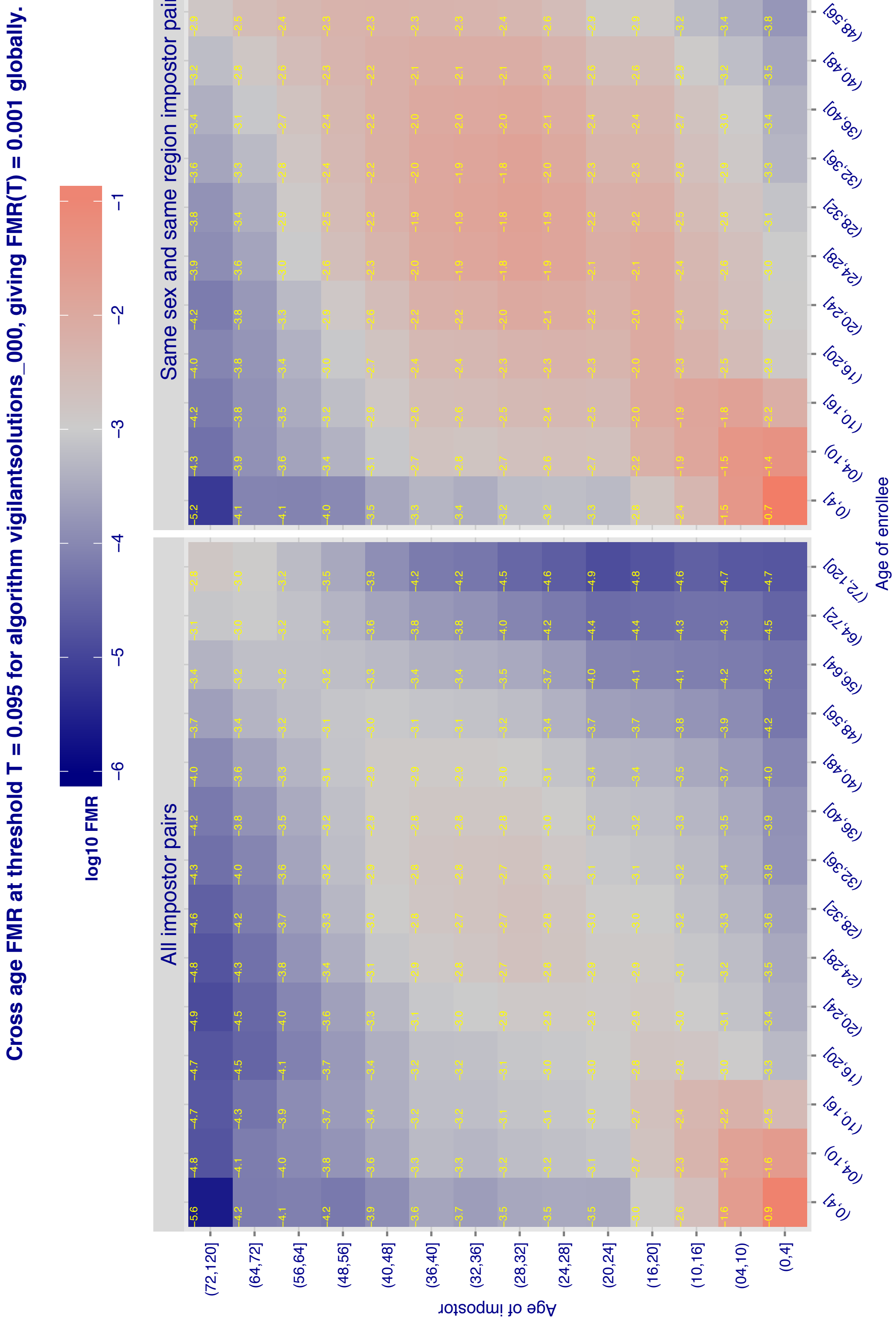


Figure 41: For algorithm vigilantsolutions-000 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who have the given age pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. The text in each box it give the same quantity as that coded by the color. Light colors present a security vulnerability to, for example, a passport gate.

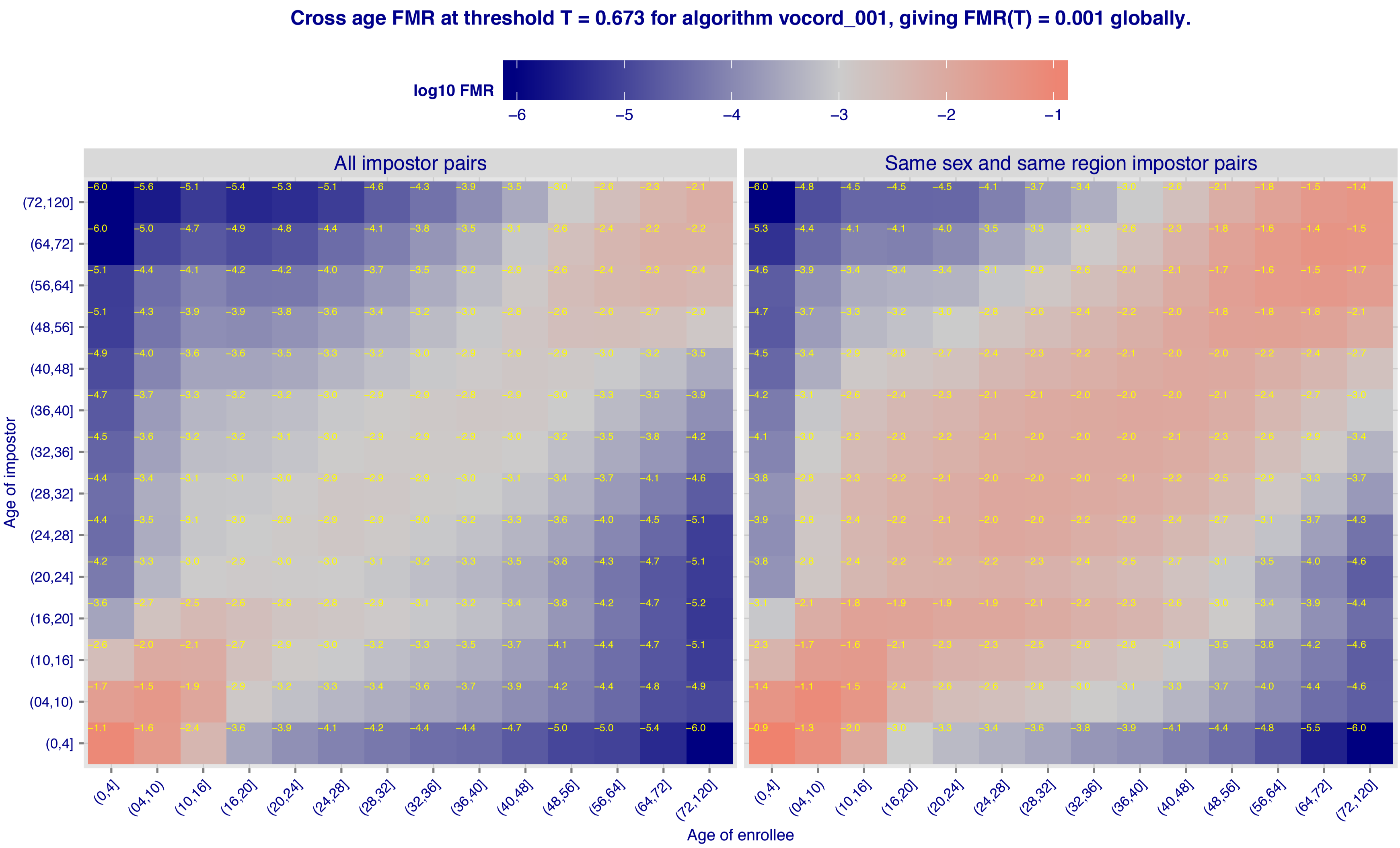
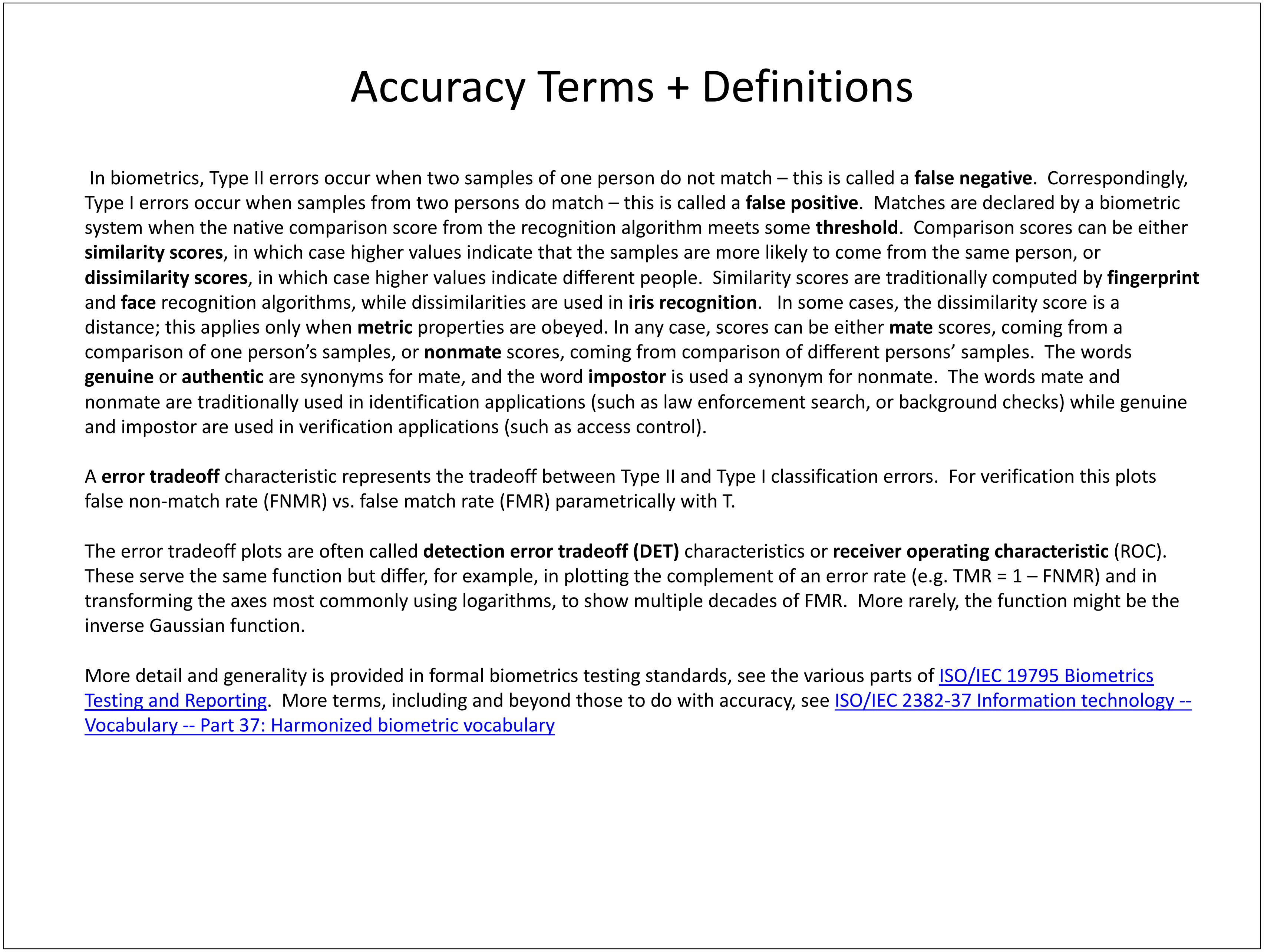


Figure 42: For algorithm vocord-001 operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who have the given age pair. False matches are counted against a recognition threshold fixed globally to give FMR = 0.001 over all O(10^{10}) impostor comparisons. The text in each box it give the same quantity as that coded by the color. Light colors present a security vulnerability to, for example, a passport gate.



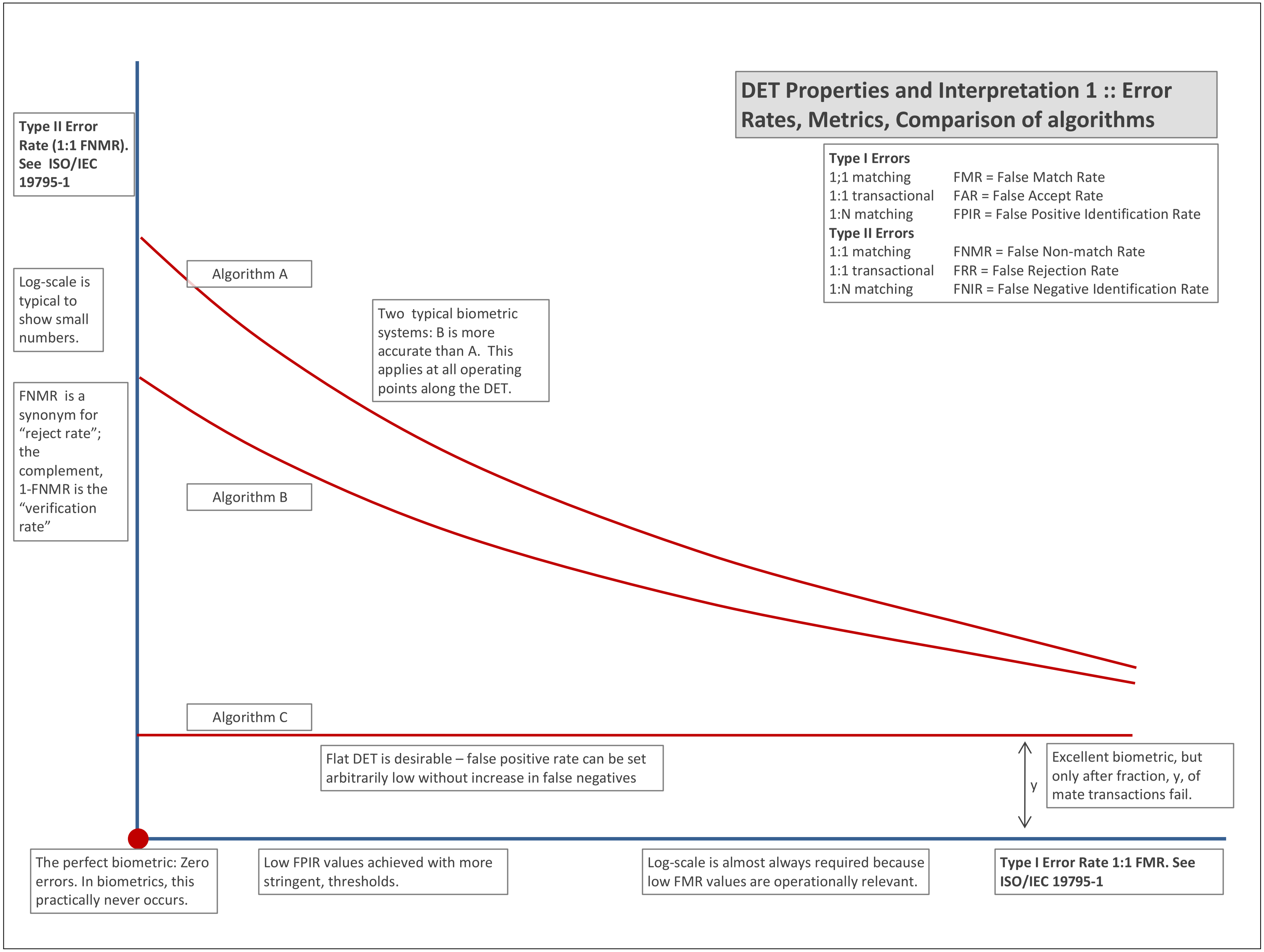
Accuracy Terms + Definitions

In biometrics, Type II errors occur when two samples of one person do not match – this is called a **false negative**. Correspondingly, Type I errors occur when samples from two persons do match – this is called a **false positive**. Matches are declared by a biometric system when the native comparison score from the recognition algorithm meets some **threshold**. Comparison scores can be either **similarity scores**, in which case higher values indicate that the samples are more likely to come from the same person, or **dissimilarity scores**, in which case higher values indicate different people. Similarity scores are traditionally computed by **fingerprint** and **face** recognition algorithms, while dissimilarities are used in **iris recognition**. In some cases, the dissimilarity score is a distance; this applies only when **metric** properties are obeyed. In any case, scores can be either **mate** scores, coming from a comparison of one person's samples, or **nonmate** scores, coming from comparison of different persons' samples. The words **genuine** or **authentic** are synonyms for mate, and the word **impostor** is used a synonym for nonmate. The words mate and nonmate are traditionally used in identification applications (such as law enforcement search, or background checks) while genuine and impostor are used in verification applications (such as access control).

A **error tradeoff** characteristic represents the tradeoff between Type II and Type I classification errors. For verification this plots false non-match rate (FNMR) vs. false match rate (FMR) parametrically with T.

The error tradeoff plots are often called **detection error tradeoff (DET)** characteristics or **receiver operating characteristic** (ROC). These serve the same function but differ, for example, in plotting the complement of an error rate (e.g. TMR = 1 – FNMR) and in transforming the axes most commonly using logarithms, to show multiple decades of FMR. More rarely, the function might be the inverse Gaussian function.

More detail and generality is provided in formal biometrics testing standards, see the various parts of ISO/IEC 19795 Biometrics Testing and Reporting. More terms, including and beyond those to do with accuracy, see ISO/IEC 2382-37 Information technology -- Vocabulary -- Part 37: Harmonized biometric vocabulary



DET Properties and Interpretation 1 :: Error Rates, Metrics, Comparison of algorithms

Type I Errors	
1;1 matching	FMR = False Match Rate
1:1 transactional	FAR = False Accept Rate
1:N matching	FPIR = False Positive Identification Rate
Type II Errors	
1:1 matching	FNMR = False Non-match Rate
1:1 transactional	FRR = False Rejection Rate
1:N matching	FNIR = False Negative Identification Rate

Type II Error Rate (1:1 FNMR). See ISO/IEC 19795-1

Log-scale is typical to show small numbers.

FNMR is a synonym for “reject rate”; the complement, 1-FNMR is the “verification rate”

Two typical biometric systems: B is more accurate than A. This applies at all operating points along the DET.

Flat DET is desirable – false positive rate can be set arbitrarily low without increase in false negatives

Excellent biometric, but only after fraction, y, of mate transactions fail.

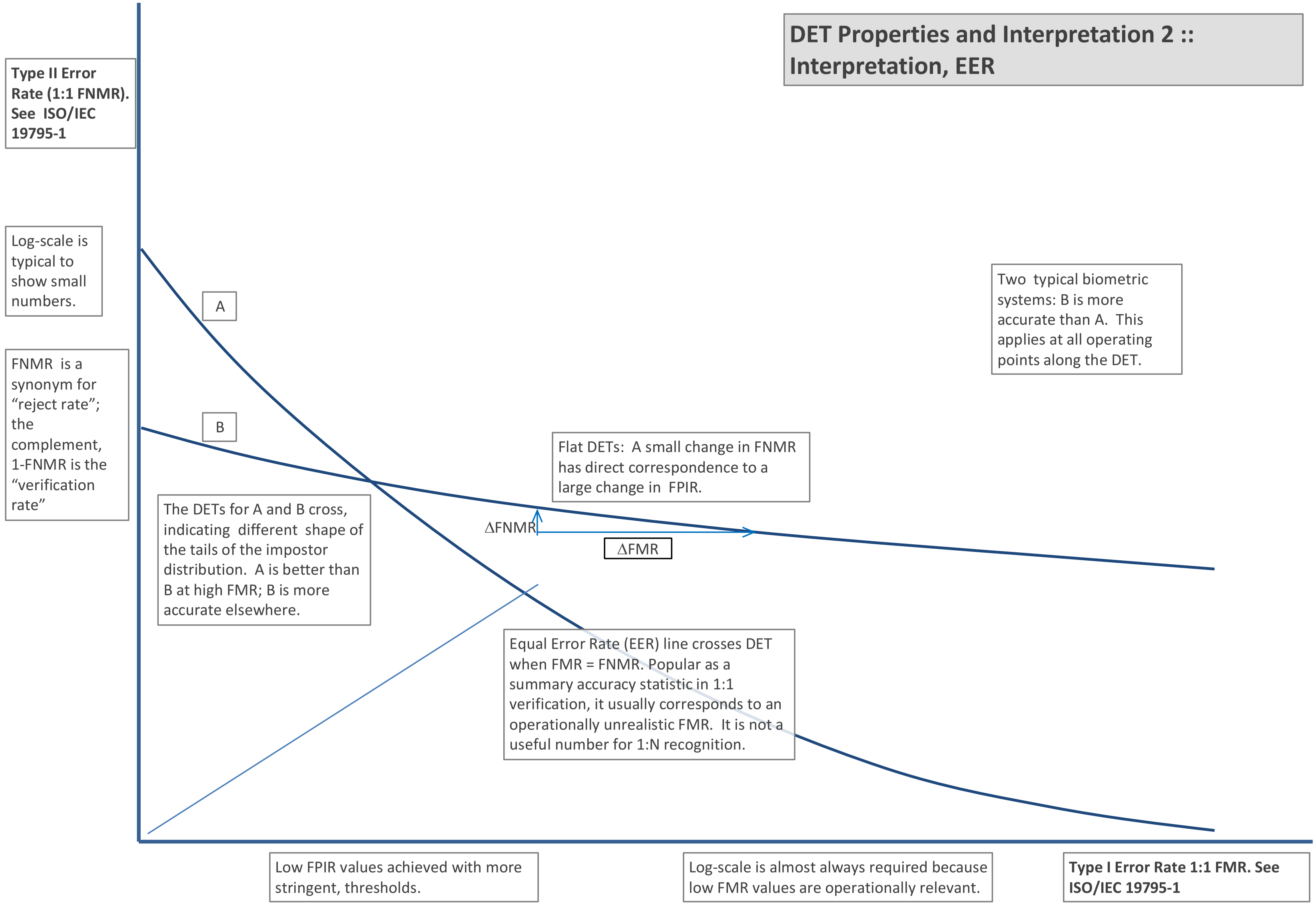
The perfect biometric: Zero errors. In biometrics, this practically never occurs.

Low FPIR values achieved with more stringent, thresholds.

Log-scale is almost always required because low FMR values are operationally relevant.

Type I Error Rate 1:1 FMR. See ISO/IEC 19795-1

DET Properties and Interpretation 2 :: Interpretation, EER



Type II Error Rate (1:1 FNMR). See ISO/IEC 19795-1

Log-scale is typical to show small numbers.

FNMR is a synonym for “reject rate”; the complement, 1-FNMR is the “verification rate”

The DETs for A and B cross, indicating different shape of the tails of the impostor distribution. A is better than B at high FMR; B is more accurate elsewhere.

Flat DETs: A small change in FNMR has direct correspondence to a large change in FPIR.

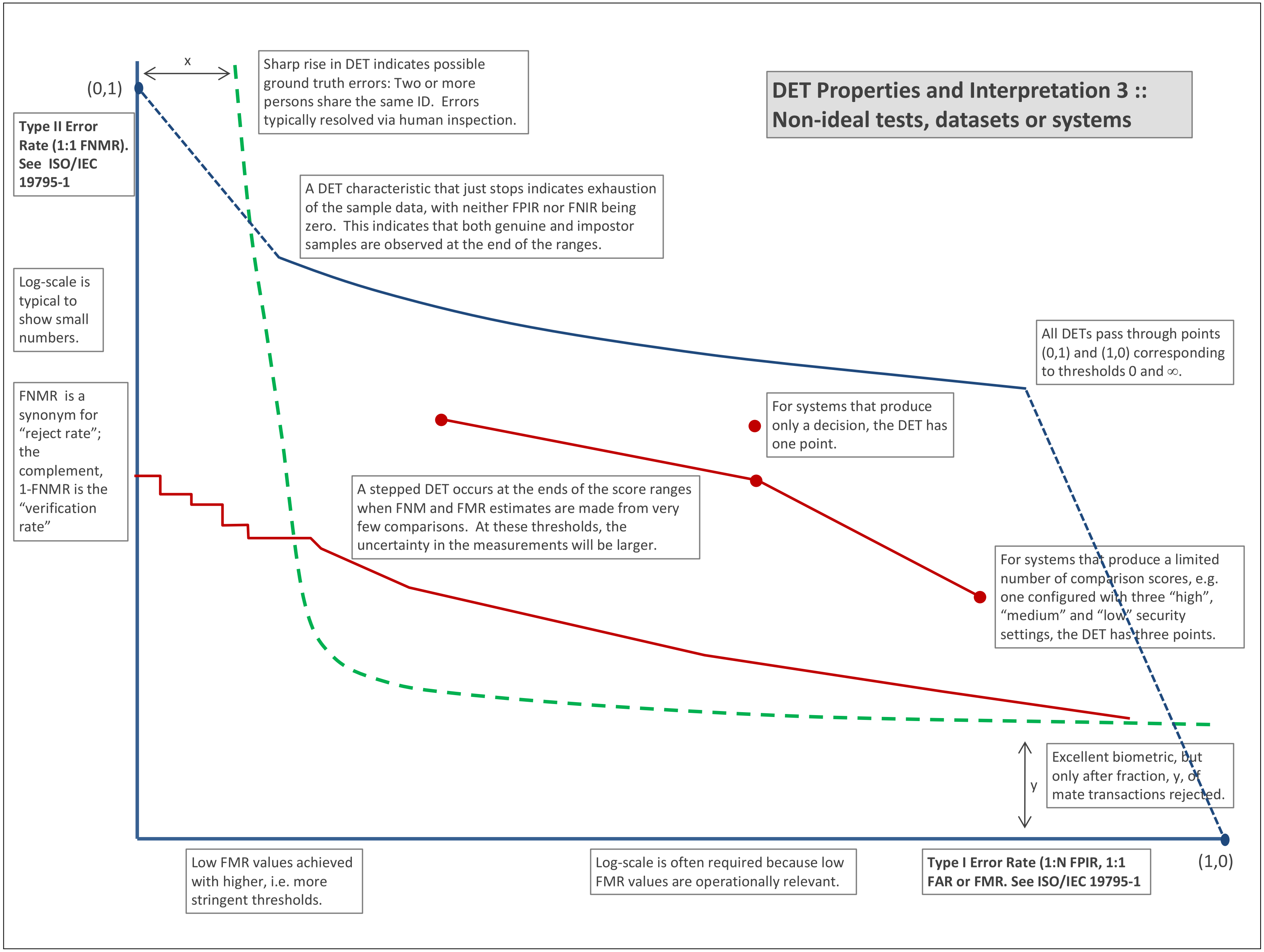
Two typical biometric systems: B is more accurate than A. This applies at all operating points along the DET.

Equal Error Rate (EER) line crosses DET when FMR = FNMR. Popular as a summary accuracy statistic in 1:1 verification, it usually corresponds to an operationally unrealistic FMR. It is not a useful number for 1:N recognition.

Low FPIR values achieved with more stringent, thresholds.

Log-scale is almost always required because low FMR values are operationally relevant.

Type I Error Rate 1:1 FMR. See ISO/IEC 19795-1



DET Properties and Interpretation 3 :: Non-ideal tests, datasets or systems

Type II Error Rate (1:1 FNMR). See ISO/IEC 19795-1

Log-scale is typical to show small numbers.

FNMR is a synonym for "reject rate"; the complement, 1-FNMR is the "verification rate"

Sharp rise in DET indicates possible ground truth errors: Two or more persons share the same ID. Errors typically resolved via human inspection.

A DET characteristic that just stops indicates exhaustion of the sample data, with neither FPIR nor FNIR being zero. This indicates that both genuine and impostor samples are observed at the end of the ranges.

All DETs pass through points (0,1) and (1,0) corresponding to thresholds 0 and ∞.

For systems that produce only a decision, the DET has one point.

A stepped DET occurs at the ends of the score ranges when FNM and FMR estimates are made from very few comparisons. At these thresholds, the uncertainty in the measurements will be larger.

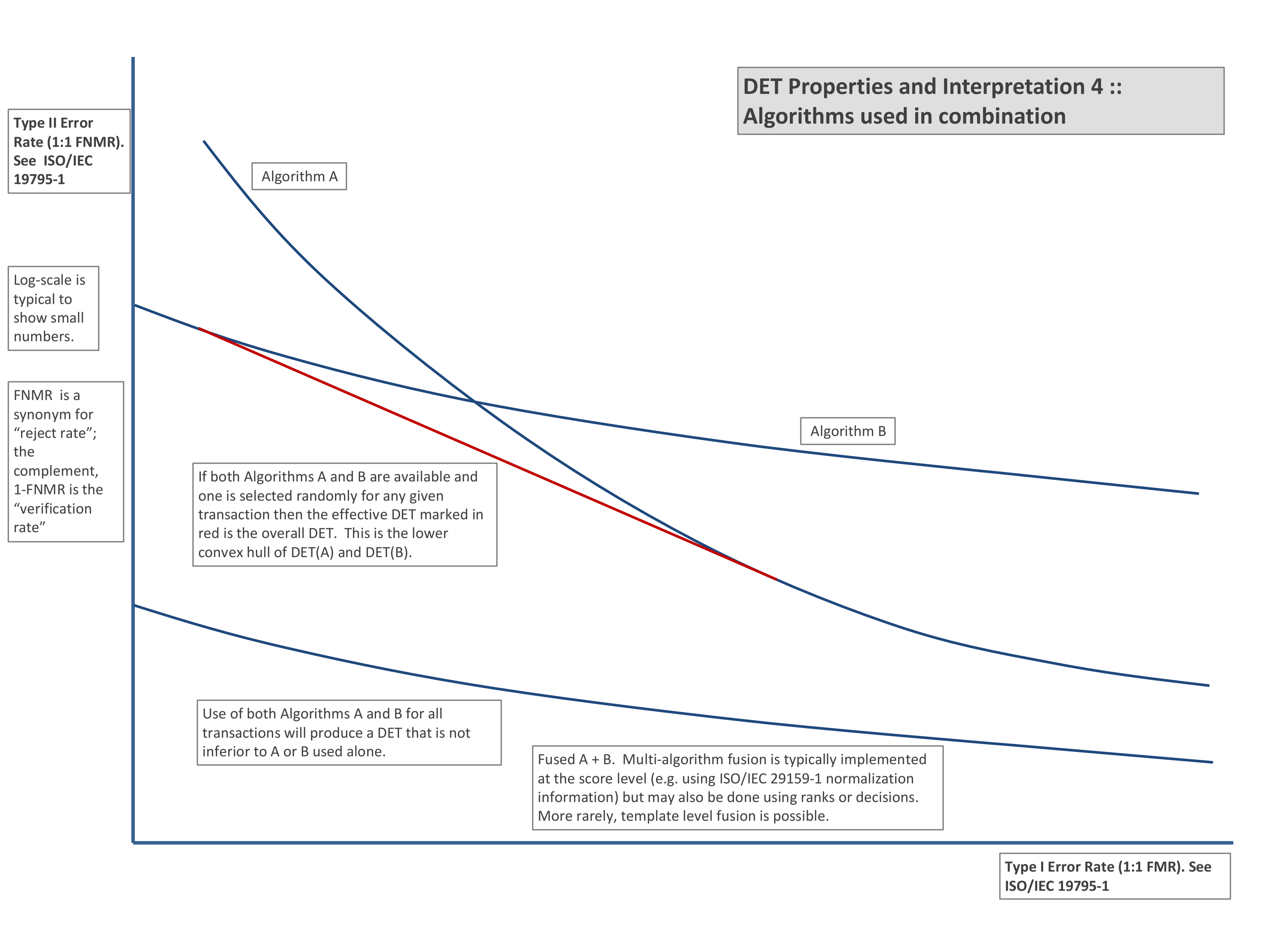
For systems that produce a limited number of comparison scores, e.g. one configured with three "high", "medium" and "low" security settings, the DET has three points.

Excellent biometric, but only after fraction, y, of mate transactions rejected.

Low FMR values achieved with higher, i.e. more stringent thresholds.

Log-scale is often required because low FMR values are operationally relevant.

Type I Error Rate (1:N FPIR, 1:1 FAR or FMR. See ISO/IEC 19795-1



DET Properties and Interpretation 4 :: Algorithms used in combination

Type II Error Rate (1:1 FNMR). See ISO/IEC 19795-1

Log-scale is typical to show small numbers.

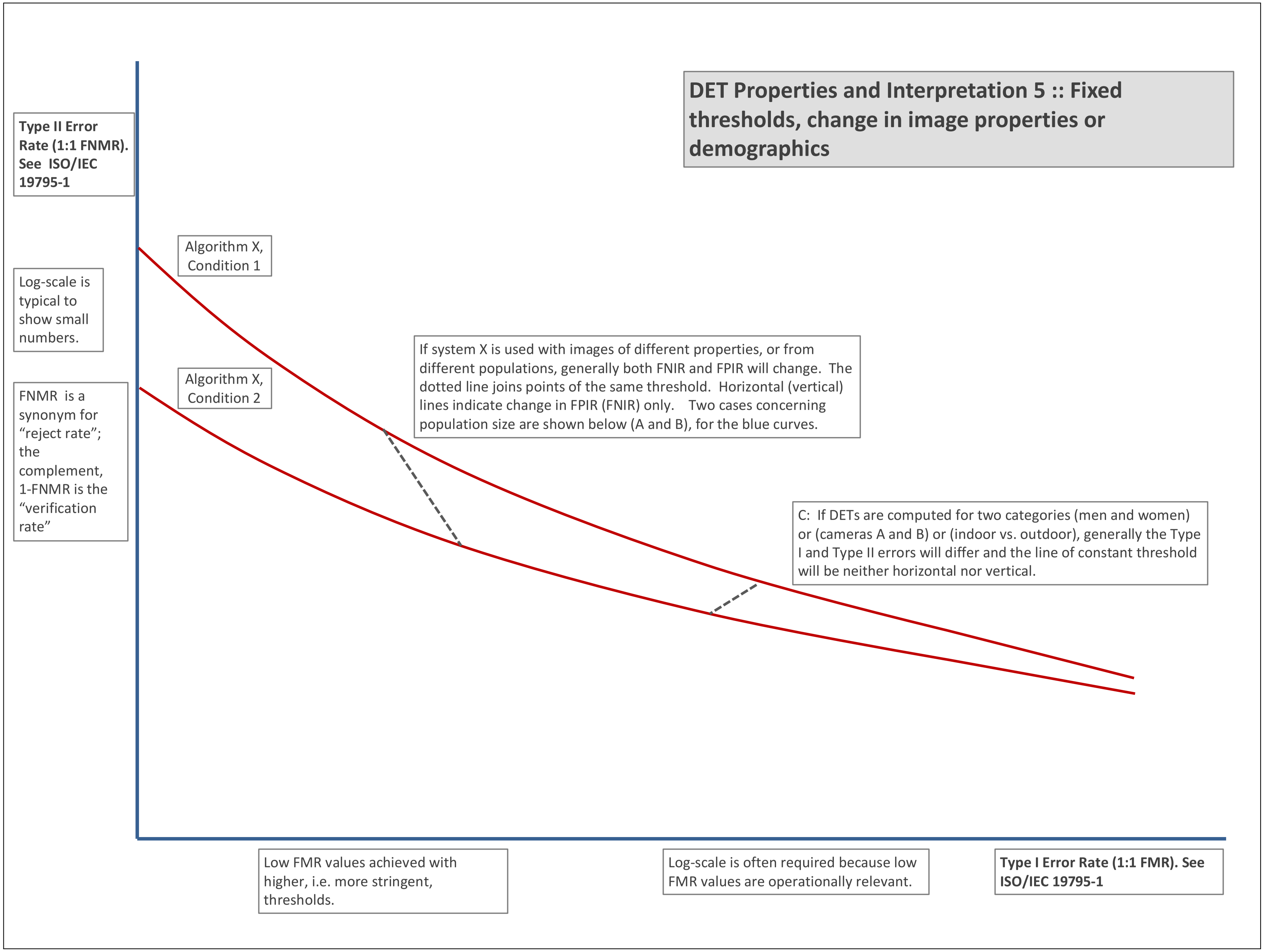
FNMR is a synonym for “reject rate”; the complement, 1-FNMR is the “verification rate”

If both Algorithms A and B are available and one is selected randomly for any given transaction then the effective DET marked in red is the overall DET. This is the lower convex hull of DET(A) and DET(B).

Use of both Algorithms A and B for all transactions will produce a DET that is not inferior to A or B used alone.

Fused A + B. Multi-algorithm fusion is typically implemented at the score level (e.g. using ISO/IEC 29159-1 normalization information) but may also be done using ranks or decisions. More rarely, template level fusion is possible.

Type I Error Rate (1:1 FMR). See ISO/IEC 19795-1



DET Properties and Interpretation 5 :: Fixed thresholds, change in image properties or demographics

Type II Error Rate (1:1 FNMR). See ISO/IEC 19795-1

Log-scale is typical to show small numbers.

FNMR is a synonym for "reject rate"; the complement, 1-FNMR is the "verification rate"

Algorithm X, Condition 1

Algorithm X, Condition 2

If system X is used with images of different properties, or from different populations, generally both FNIR and FPIR will change. The dotted line joins points of the same threshold. Horizontal (vertical) lines indicate change in FPIR (FNIR) only. Two cases concerning population size are shown below (A and B), for the blue curves.

C: If DETs are computed for two categories (men and women) or (cameras A and B) or (indoor vs. outdoor), generally the Type I and Type II errors will differ and the line of constant threshold will be neither horizontal nor vertical.

Low FMR values achieved with higher, i.e. more stringent, thresholds.

Log-scale is often required because low FMR values are operationally relevant.

Type I Error Rate (1:1 FMR). See ISO/IEC 19795-1