Appeared in IEEE Transactions on Biometrics, Behavior, and Identity Science (IEEE T-BIOM)
February 2019, DOI: 10.1109/TBIOM.2019.2897801

Demographic Effects in Facial Recognition and their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems

Cynthia M. Cook, John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury

Abstract—We examined the effect of demographic factors on the performance of the eleven commercial face biometric systems tested as part of the 2018 United States Department of Homeland Security, Science and Technology Directorate (DHS S&T) Biometric Technology Rally. Each system that participated in this evaluation was tasked with acquiring face images from a diverse population of 363 subjects in a controlled environment. Biometric performance was assessed using a systematic, repeatable test process measuring both efficiency (transaction times) and accuracy (mated similarity scores using a leading commercial algorithm). Prior works have documented differences in biometric algorithm performance across demographic categories and proposed that skin phenotypes offer a superior explanation for these differences. To test this concept, we developed an automatic method for measuring relative facial skin reflectance using subjects' enrollment images and quantified the effect of this metric and other demographic covariates on performance using linear modeling. Both the efficiency and accuracy of the tested acquisition systems were significantly affected by multiple demographic covariates including skin reflectance, gender, age, eyewear, and height. Skin reflectance had the strongest net linear effect on performance. Linear modeling showed that lower (darker) skin reflectance was associated with lower efficiency (higher transaction times) and accuracy (lower mated similarity scores). Skin reflectance was also a statistically better predictor of these effects than self-identified race labels. Unlike other covariates, the degree to which skin reflectance altered accuracy varied between systems. We show that the size of this skin reflectance effect was inversely related to the overall accuracy of the system such that the effect was almost negligible for the system with the highest overall accuracy. These results suggest that, in evaluations of biometric accuracy, the magnitude of measured demographic effects depends on image acquisition.

Index Terms—Face Recognition, Demographics, Skin Reflectance, Scenario Testing, Commercial Systems, Acquisition Systems

1 Introduction

The performance of computer systems in general [1] and biometric systems in particular [2] has long been known to be affected by user demographics. Recent advances in facial recognition algorithms and increases in the adoption of facial recognition systems, especially in the public sphere, have renewed the need to understand how this evolving technology impacts diverse users [3].

Since 2014, the United States Department of Homeland Security, Science and Technology Directorate (DHS S&T) has sponsored biometric research and scenario testing at the Maryland Test Facility (MdTF). As part of this ongoing work, we have observed that biometric performance can vary based on subject demographics and behavior [4] [5]. Recently, we carried out a large-scale scenario test, the 2018 Biometric Technology Rally ("Rally"), designed to simulate a high-throughput security screening/inspection process. The Rally tested eleven facial recognition systems from different commercial organizations and measured the efficiency and effectiveness with which participating systems acquired and matched face images from a population of diverse volunteer subjects. Official results of this evaluation have been presented elsewhere [6]. Here we extend these results by investigating the effects of collected demographic covariates and calculated phenotypic measures of skin reflectance on the biometric performance of the evaluated systems.

Prior work showed that demographic factors, such as gender, race, age, and ageing could influence the performance of facial recognition and classification algorithms [7] [8] [9] [10] [11]. Recently, [12] found lower face gender classification algorithm accuracy for images of women and people with a darker skin phenotype measured by manually assigning skin types to face photos. Further, [12] set out a clear rationale for using phenotypic measures over demographic labels. However, the relationship between skin phenotypes and the performance of face recognition algorithms has not been investigated and quantitative evidence for the superiority of phenotypes as predictors of biometric performance is lacking. Finally, whether demographic effects vary across different biometric face acquisition systems has not been examined. For example, are all face acquisition systems equal in the magnitude of performance variations across demographic groups? For which demographic factors does performance differ more and for which less? Is the difference in performance between different demographic groups greater than or less than the difference in the performance of different systems? This research investigates these questions.

- *C. Cook, J. Howard, Y. Sirotin, and J. Tipton work at the Maryland Test Facility in Upper Malboro, Maryland.*
- *A. Vemury works at the United States Department of Homeland Security, Science and Technology Directorate in Washington, DC.*
- *Authors listed alphabetically. E-mail correspondence should be sent to rally@mdtf.org*

To understand the factors affecting the performance of modern commercial face acquisition systems, we examined images and associated meta data gathered during the Rally (more details in Sections 2.2 - 2.3). For each acquired sample, we used a leading commercial algorithm [8] to find the rank-one mated similarity score (Section 2.6) against two separate galleries, a gallery of same-day images and a gallery of up to 4 year old historic images (Section 2.4). We also developed a new methodology for calculating the relative skin reflectance of each subject using their same-day gallery images (Section 2.5). Finally, we modeled the statistical relationship between the eleven tested biometric acquisition systems, mated similarity scores, and the collected/calculated demographic covariates, namely reflectance, gender, race, eyewear, age, height, and weight (Section 2.7). Our results show that biometric system performance is strongly affected by demographic factors, notably skin reflectance, and that the degree to which these factors affect mated similarity scores varies across systems.

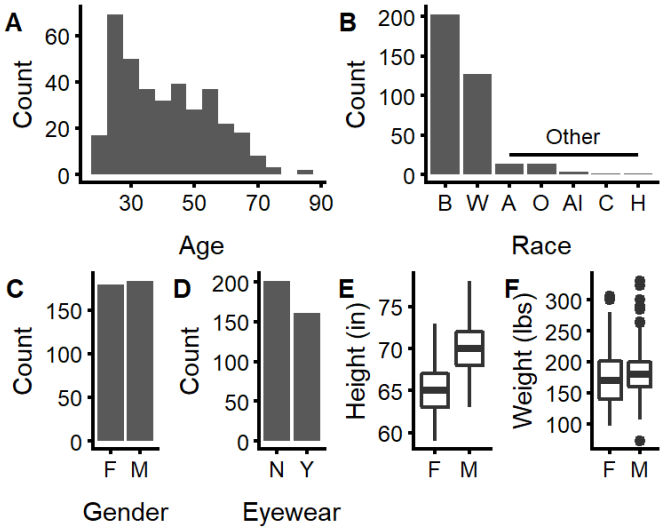


Fig. 1. Distributions of the demographic variables self-reported by test subjects. **A.** Distribution of test subject ages. **B.** Counts of subjects identifying with each racial category: (B) Black or African-American; (W) White; (A) "Asian"; (O) "Other Race"; (AI) "American Indian or Alaska Native"; (C) "Aboriginal peoples of Canada"; (H) "Native Hawaiian or other Pacific Islander". Groups A, O, AI, C, and H are grouped in to a general "Other" category during analysis. **C.** Distribution of subject gender: (F) Female; (M) Male. **D.** Subject response to whether or not they wear eyewear: (N) No; (Y) Yes. **E-F.** Boxplots of subject height and weight by gender.

2 Methods

2.1 Systems

Eleven commercial organizations participated in the Rally. To ensure a fair test protocol, Rally systems were required to fit within a 7 by 8 foot space ("station") and be capable of capturing face images. All systems were unstaffed and directed all aspects of subject interaction automatically. Systems were also required to capture and submit images before each subject left the station. Outside of these restrictions, participating organizations were free to use any combination of hardware, software, and human machine interaction methods they thought would meet the goals of the Rally (99% true identification rate with an average transaction time of 5 seconds or less per test subject). Consequently, systems tested in this evaluation were notably different in terms of work-flow, configuration, face camera technology, and the cybernetics of interaction with the subjects. Goals for the Rally were set based on previous tests performed at the MdTF that established baseline industry performance in similar use cases.

2.2 Subjects and Sample Size

Rally systems were tested with a demographically diverse population of 363 volunteer subjects (Fig. 1). All test subjects consented to participate in the study under an established Institutional Review Board (IRB) protocol, and most had volunteered for past test activities at the MdTF. Race, age, gender, eyewear, height, and weight were self-reported during study enrollment. Race was defined in accordance to the U.S. Census categories [13].

2.3 Test Process

The test process and evaluation was designed to provide a systematic, repeatable framework for evaluating the acquisition speed and matching performance of arbitrary biometric systems. Test subjects were briefed as to the purpose of the scenario test and told that the facial acquisition systems were intended to perform biometric identifications. They were asked to comply with all instructions presented by the systems, but were not specifically instructed regarding the mechanistic details of the individual systems.

Subjects were organized into treatment groups of ∼15, which queued at a station. Test subjects entered a station one-at-a-time after their ground truth identity was recorded by test staff. The order in which each treatment group used each system was counterbalanced so every system was used in each serial position and every system followed every other system an equal number of times. All systems operated autonomously and were completely unstaffed. Image submissions were made by each station in real time via a common web based API (Fig. 2).

2.4 Face Image Galleries

At the start of each test session, subjects were enrolled into a "same-day" face image gallery by staff trained in biometric collection. Subjects stood in front of a 18% neutral gray background. Diffuse illumination was measured at 600-650 lux. Enrollment staff collected a single face image using a Logitech C920 camera at a 1 meter standoff (resolution: 1920x1080). Staff asked subjects to remove any hats or glasses and assume a neutral expression. Staff assessed any image quality issues and re-acquired images when necessary. This resulted in a same-day face image gallery of 363 face samples from 363 unique people.

A "historic" face image gallery of 1,848 samples from 525 unique people (average of 3.5 images per person) was assembled. These samples were acquired over the course of four years using a variety of cameras including digital single lens reflex (DSLR) cameras, web-cameras, and cameras embedded in biometric devices. The historic gallery contained images for 326 of the 363 test subjects that participated in

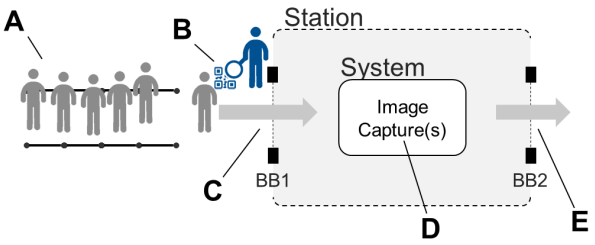


Fig. 2. The test process performed at each test station during the Rally. Commercial face capture systems (System) were installed within a dedicated test station (Station). **A.** Test subjects queued at each station. **B.** Test staff established the ground-truth identity of each subject by scanning a QR code printed on the subject's wristband. **C.** Subjects entered the test station, triggering a beam break (BB1). **D.** Subjects interacted with the face capture system, which submitted images (biometric samples) for storage. **E.** Subjects exited the test station, triggering a beam break (BB2). The duration of each subject's interaction with the system was measured as the difference in time between BB2 and BB1. The images submitted by each system were used to analyze biometric performance.

the evaluation as well as 199 "distractors" or out of gallery subjects.

2.5 Relative Skin Reflectance

At a basic level, the operation of any facial recognition algorithm is dependent on the pixel intensities of the provided facial images. The intensity of a pixel of skin in a digital image of the face is affected by three factors: physical properties of the skin and underlying tissue (layers, absorbers, and scattering coefficients), physical properties of the skin surface (specular reflectance), and imaging artifacts (incident light intensity, camera gain). Any method seeking to relate physical properties of the skin to the performance of facial recognition algorithms must remove the confounding effects of imaging artifacts.

Our method for achieving this involved obtaining normalized (R, G, B) color values for skin pixels according to the process in Fig. 3. This method is distinct from using non-linear color space transformations like CIELab color or YCbCr that have been previously examined in their utility for segmenting skin in images [14]. The primary difference is that the goal is not to measure "skin color", a non-linear perceptual phenomenon captured by color spaces optimized for human perception, but the physical skin properties, which ideally rely on light intensity measurements at specific wavelengths. Skin pixels were selected by face finding, circular masking, and outlier removal using methods adapted from [15]. Because some image artifacts are multiplicative in nature and we had a constant reference region in the neutral 18% gray background, the average (R, G, B) color values from these skin pixels could be corrected for artifacts by divisive normalization using background regions selected from gray areas around the face. This operation corrects for camera exposure and incident light intensity, but not for variation in shadows across the face or specular reflection. Additionally, it does not correct for camera color balance.

After background correction and outlier removal, the resulting (R, G, B) values are dependent primarily on the physical properties of the skin and are proportional to the amount of incident light reflected from the face. This is consistent with the technical definition of reflectance. The methodology for calculating this metric relied on the specific collection conditions used in this study, namely, the consistent lighting, the same acquisition camera, and the constant neutral gray background. This method provided an estimate of the physical properties of each subject's skin obtained on independent samples.

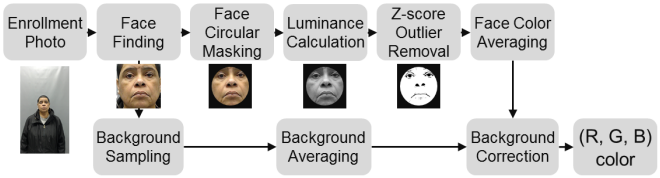


Fig. 3. Process for extracting skin (R, G, B) color values from same-day enrollment images.

To quantify the variation in skin reflectance across subjects, we performed Principal Components Analysis (PCA) on the (R, G, B) color values. The first two principal components (PC1 and PC2) explained 96.1% and 3.4% of the variance in (R, G, B) color values, respectively, collectively explaining 99.5% of the total color variance. This may be related to the fact that melanin and hemoglobin are the two main absorbers of light in skin, with most of the variation in reflectance across skin types due to melanin [16]. We call this final metric, namely the position of each test volunteer along PC1, a measure of their *relative* skin reflectance[1] across all face enrollment images in our study.

Visualizing the face images in the (PC1, PC2) space showed that PC1 is strongly related to net skin reflectance. Fig. 4A shows the space formed by the first two magnitude normalized PCs (faces of some test volunteers are obscured). Fig. 4B shows average faces for men and women in our sample and average faces for people in each skin reflectance quartile Fig. 4C.

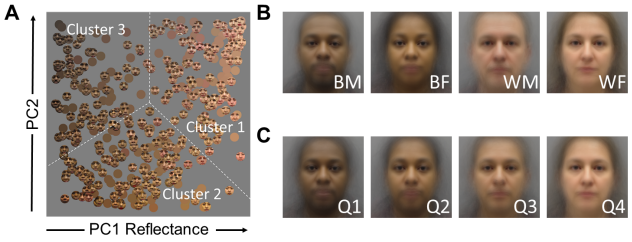


Fig. 4. **A.** Faces of test subjects plotted within the normalized color space obtained by PCA of extracted skin (R, G, B) color values. Dashed-lines separate color clusters identified using k-means clustering in the PC1-PC2 space (k=3). Note: faces of some subjects have been obscured and plotted only as color swatches. **B.** Eye-aligned average faces for black males (BM), black females (BF), white males (WM) and white females (WM). **C.** Eye-aligned average faces for each skin reflectance quartile (Q1-Q4).

1. For brevity, we will often use the term "Reflectance" in tables, figures, equations, and captions while the term "Skin Reflectance" is used in the main text. Both terms denote the relative skin reflectance measure as calculated per Section 2.5.

2.6 Mated Similarity Score Analysis

Face images acquired by each system were matched using a leading commercial biometric algorithm [8] against both the same-day and historic galleries. The set of mated similarity scores against the same-day gallery used samples from all 363 test subjects, while the corresponding set for the historic gallery used only samples from the 326 test subjects who had corresponding images in the historic gallery. For systems $s \in S$ returning multiple face images, $i \in I$, mated similarity scores were examined for the last image acquired, $(i) \in I$, within the 20 second time interval following the entry beam-break. The same-day mated similarity score for subject $j \in J$ is denoted $\Phi_{j,s}^{sd} = \phi_{(i)}$. For the historic gallery, which contained $i \in I$ gallery images for each subject $j \in J$, the top mated similarity score $\Phi_{j,s}^{hist} = \max_{i \in I} \phi_i$ was used for statistical analysis. Some systems occasionally had technical issues or acquired images for individuals in the background causing them to submit a photo for the next or previous subject in the queue. We therefore removed any images for which the rank-one similarity score was higher than the mated similarity score. This occurred in fewer than 30 transactions across all systems (i.e. less than 1% of collected data was affected/removed) and manual review indicated that most were artifacts introduced during testing.

2.7 Statistical Modeling

2.7.1 All-System Average Models

To estimate the overall average demographic effects, we applied linear regression to the subject's all-system average historic and same-day mated similarity scores $\bar{\Phi}$ as well as average transaction times $\bar{\Psi}$. Specifically, for each subject j using all $N = 11$ systems, s, we computed the average mated similarity score $\bar{\Phi}$ to the historic gallery as $\bar{\Phi}_j = \frac{1}{N}\sum_{s=1}^{s=N} \Phi_{j,s}^{hist}$ and to the same-day gallery as $\bar{\Phi}_j = \frac{1}{N}\sum_{s=1}^{s=N} \Phi_{j,s}^{sd}$. In constructing the linear model, we considered eleven demographic covariates including three categorical variables: gender, eyewear, and race. Race was grouped as "White", "Black", and "Other" (see Section 2.2). We normalized the continuous variables age, height, weight, and skin reflectance (see Section 2.5) prior to fitting according to $z = (x - \mu_x)/\sigma_x$ and included their squared transformations in the full model for each response variable $\Theta_j \in \{\bar{\Phi}_j, \bar{\Psi}_j\}$. The inclusion of interaction terms, which could lead to over-fitting, was not considered in this analysis.

$$\begin{aligned}\Theta_j = \beta_0 + \beta_1 gender_j + \beta_2 eyewear_j + \beta_3 race_j + \\ \beta_4 age_j + \beta_5 age_j^2 + \beta_6 height_j + \beta_7 height_j^2 + \\ \beta_8 weight_j + \beta_9 weight_j^2 + \beta_{10} reflectance_j + \\ \beta_{11} reflectance_j^2 + \epsilon_j\end{aligned} \quad (1)$$

We estimated model parameters β, using ordinary least squares (OLS) fitting. We defined the optimal all-system model as one that minimizes the Akaike Information Criteria, $AIC = 2k - 2ln(\hat{L})$, where k represents the number of estimated parameters in the model and $\hat{L}$ represents the maximum value of the model's fitted likelihood. AIC measures the goodness of fit of the model while discouraging over-fitting with a penalty for increasing the number of model parameters k. To find the optimal models, we first fit the full model with all eleven covariates. We then applied a step wise procedure in both directions using the `stepAIC()` function in the R package MASS. We applied this procedure to both the historic and same-day average mated similarity scores and average transaction times. Equation 2, describes a final optimal model with $k - 1$ covariates selected, for the jth subject.

$$\begin{aligned}\mathbf{x}_j &= [x_{1,j}, x_{2,j}, ...x_{k-1,j}] \\ \boldsymbol{\beta} &= [\beta_1 \ , \beta_2 \ , ...\beta_{k-1} \ ] \\ \bar{\Theta}_j &= \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_j + \epsilon_j\end{aligned} \quad (2)$$

2.7.2 Bootstrapping for Estimating Confidence Intervals

We assessed the accuracy of model fits through residual analysis. For all three optimal all-system models, we found the residuals deviated from normality, with noticeable deviations present in the QQ plots of our response variables Θ_j (data not shown). We, therefore, obtained confidence intervals for model parameter estimates using a bootstrapping technique instead of relying on the standard error. We generated 1000 bootstrap samples and calculated the bias corrected bootstrapped confidence intervals or the BC_α for each of the fitted coefficients in the optimal model [17].

2.7.3 Cross-Validation of Optimal Model Parameters

Our model selection approach showed that some covariates did not improve model fit sufficiently as judged using AIC and are therefore excluded from the optimal model (Section 2.7.1). We used the non-parametric technique of cross-validation to independently confirm the optimality of select covariates included in our optimal model. We used ten-fold cross-validation and compared the cross-validated R^2 of the optimal model to a model where a covariate present in the optimal model is replaced by an alternate covariate not present in the optimal model. Since the exact fold compositions, and therefore the cross-validated R^2 values are dependent on a random seed, this procedure was executed with 100 randomly drawn starting seeds to compute the mean and 95% confidence intervals for the cross-validated R^2 values.

2.7.4 Mixture Models for Cross System Effects

The results from the average linear regression models explain the effects of demographics on all-system average mated similarity scores and average transaction times. However, because all 363 subjects interacted with each of the eleven Rally systems, we can examine if mated similarity scores for images acquired on different systems had distinct demographic covariate effects. To model these effects, we applied linear mixture modeling with system s as the random effect. To start, we used all demographic covariates retained in the optimal model from Equation 2 as fixed effects. This approach allowed us to model our response variable by estimating both the variance across all systems (fixed effects: β_0 and $\boldsymbol{\beta}^T$) and the variance between different systems (random effects: $\beta_{0,s}$ and $\boldsymbol{\beta}_s^T$) according to Equation 3 where $\mathbf{y}$ is the set of m selected system-specific slope covariates and $\boldsymbol{\beta}_s$ are the corresponding parameters.

$$\mathbf{y}_j = [y_{1,j}, y_{2,j}, ...y_{m,j}]$$
$$\boldsymbol{\beta}_s = [\beta_{1,s}, \beta_{2,s}, ...\beta_{m,s}] \quad (3)$$
$$\Theta_{j,s} = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_j + \beta_{0,s} + \boldsymbol{\beta}_s^T \mathbf{y}_j + \epsilon_j + \gamma_s$$

Starting with only the fixed effects model, we added a system-specific slope $\beta_{0,s}$. If this reduced AIC, it signified that there are statistical performance differences between systems. Then, given the intercept model that includes $\beta_{0,s}$, we used a forward model selection approach to identify the mixed individual effects that continue to minimize AIC, adding each demographic covariate ($\mathbf{y}_j$) one at a time. A reduction in AIC for a given demographic covariate signifies the inclusion of a system-specific coefficient *for this variable* improves model fitness and thus, there are notable performance differences *between stations* for this demographic factor. We performed this procedure for the historic gallery similarity scores. Since the goal of this analysis was to estimate the system-specific effects, we estimated all model parameters β, by maximizing the restricted likelihood (REML) [18].

3 Results

3.1 Self-reported Race and Skin Reflectance

In the test population, skin reflectance quartiles showed that the proportion of subjects identifying as Black or African-American is inversely associated with skin reflectance values (Fig. 5A). The test subject group was composed largely of people identifying as Black or African-American (56%) or as White (35%), and for all race categories in our sample, skin reflectance for women was higher than for men (Fig. 5B). Using a technique, known as the "gap-statistic" [19], we identified $k = 3$ as the optimal number of clusters in the normalized (R, G, B) color space. K-means clustering identified the cluster boundaries, shown as dashed lines in Fig. 4A. Cluster 1 was largely composed of subjects identifying as White whereas both Clusters 2 and 3 were largely composed of those identifying as Black or African American (Fig. 5C). Skin reflectance for each self-reported race varied by cluster (Fig. 5D).

3.2 Demographic Effects on All-System Average Similarity Scores

We measured the effects of skin reflectance and other demographic factors on similarity scores using linear modeling (Section 2.7.1). To identify the overall effect of subject demographics on similarity scores, we computed an average similarity score for each subject on all eleven tested systems. To examine longitudinal changes in appearance (e.g. changes in attire, self styling, and ageing), we fit separate models to average similarity scores obtained from matching images to the historic and to the same-day galleries. Starting with a full model including all eleven demographic covariates (Equation 1), we used an AIC-based model selection approach to find an optimal model including only those demographic covariates that improved model fit while minimizing the number of model parameters. Following model selection, we computed the 95% bootstrapped, bias corrected confidence intervals, (BC_α) for each parameter.

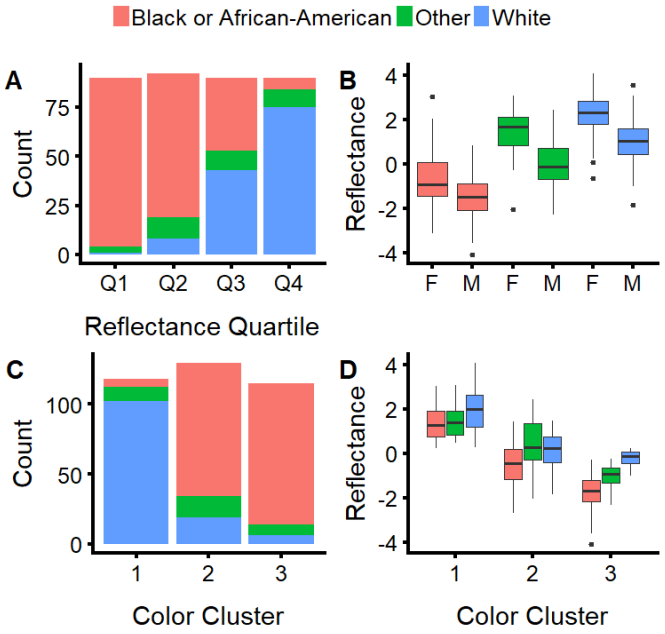


Fig. 5. Relationship between relative skin reflectance (Reflectance) and self-reported race and gender. **A.** Race composition of each reflectance quartile. **B.** The distribution of reflectance values for each race and gender: (M) Male; (F) Female. **C.** Race composition of each skin color cluster. **D.** Distribution of reflectance values for each race in each skin color cluster. Note: Reflectance values for each race vary greatly across color clusters.

	Covariate	Estimate	BC_α	Range	Net Effect
Optimal Historic Model					
$\hat{\beta}_0$	Intercept	0.830	$(0.817, 0.841)$	NA	NA
$\hat{\beta}_1$	Gender	0.034	$(0.022, 0.048)$	$\{0, 1\}$	0.034
$\hat{\beta}_2$	Eyewear	-0.025	$(-0.039, -0.012)$	$\{0, 1\}$	0.025
$\hat{\beta}_4$	Age	0.007	$(0.001, 0.013)$	$(-1.47, 3.09)$	0.032
$\hat{\beta}_7$	Height^2	-0.006	$(-0.011, 0.0002)$	$(0.0004, 7.43)$	0.041
$\hat{\beta}_{10}$	Reflectance	0.016	$(0.009, 0.024)$	$(-2.41, 2.39)$	0.075
Optimal Same-Day Model					
$\hat{\beta}_0$	Intercept	0.894	$(0.885, 0.900)$	NA	NA
$\hat{\beta}_2$	Eyewear	-0.019	$(-0.029, -0.009)$	$\{0, 1\}$	0.019
$\hat{\beta}_6$	Height	0.005	$(0.0001, 0.011)$	$(-2.33, 2.73)$	0.027
$\hat{\beta}_7$	Height^2	-0.004	$(-0.008, 0.001)$	$(0.0004, 7.43)$	0.027
$\hat{\beta}_8$	Weight	-0.005	$(-0.011, 0.001)$	$(-2.61, 3.50)$	0.031
$\hat{\beta}_{10}$	Reflectance	0.010	$(0.006, 0.016)$	$(-2.41, 2.39)$	0.050

TABLE 1
Parameter estimates for the optimal models, fitting all-system average same-day and historic similarity scores as discussed in Section 2.7.1. Parameters not included in the optimal model are not shown (see Equation 1). 95% confidence intervals BC_α are estimated using bootstrap (Section 2.7.2.). Net effect of the covariate is estimated as the product of $\hat{\beta}$ and the magnitude of the observed range of values for the covariate $|max - min|$. The units of similarity scores are arbitrary.

Plotting average mated similarity scores as a function of skin reflectance, age bins, and gender (Fig. 6) showed that scores tended to be lower for subjects with lower reflectance values for both the historic and same-day galleries. For the historic gallery, scores for male subjects were notably higher than for female subjects. For the same-day gallery, however, male and female subjects tended to have similar score distributions. Further, scores for younger subjects tended to be lower for the historic gallery. The linear fits depicted in Fig. 6 do not include the effect of eyewear; those who reported wearing some form of eyewear had lower scores

for both same-day and historic galleries.

Table 1 shows the estimates, 95% confidence intervals (BC_α), and the net effect of each coefficient (β) in the optimal historic and same-day gallery model of mated similarity scores. We estimated the net effect of a covariate as the product of $\hat{\beta}$ and the magnitude of the observed range of values for the covariate $|max - min|$. By this metric, reflectance was the covariate with the single greatest net effect on mated similarity scores with an net effect equal to roughly 10% of the intercept value of the historic similarity scores and roughly 6% of the intercept of the same-day similarity scores.

Consistent with visual impressions from Fig. 6, fitted parameter estimates for the historic gallery model indicated that average mated similarity scores decreased significantly for younger subjects, those who identified as female, those with lower (darker) skin reflectance and those who reported eyewear. The effect of Height^2 on mated similarity scores was negative, suggesting deviations from average height decreased the scores, but BC_α for this factor overlapped 0 in both the same-day and historic model. The effects of height, eyewear, and skin reflectance also appeared in both models. Notably, age and gender only appeared as covariates in the historic model, indicating that these covariates did not influence same-day similarity scores. Weight was present in the same-day optimal model, but BC_α for this factor overlapped 0.

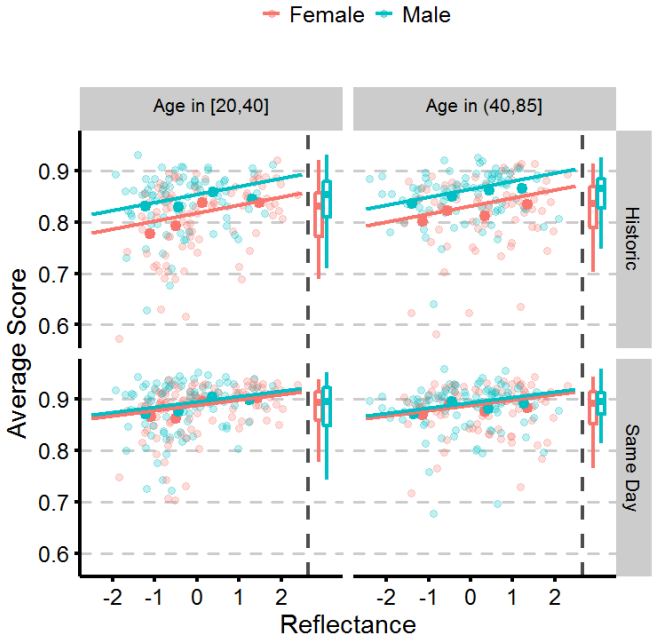


Fig. 6. Average mated similarity scores (Average Scores) variation with relative skin reflectance (Reflectance) and gender faceted by gallery and age. Points show average mated similarity scores for female and male subjects. Lighter points show average mated similarity scores for individual subjects. Darker points denote grand average of scores across subjects binned by reflectance quartile. Lines indicate optimal age and gender model fits, fixing other factors constant at the average value of the subject population in each facet. Box plots within each facet show marginal distributions of similarity scores by gender.

3.3 Demographic Effects on All-System Average Transaction Times

We also measured the effects of demographic factors on the time it took subjects to complete biometric transactions, and found large effects for subjects with eyewear and lower (darker) skin reflectance. Transaction times were measured at each station as the time interval between subjects crossing the entry and exit beam breaks, which included all interactions with the biometric face capture system (Fig. 2). We again used linear modeling (Section 2.7) to measure the net effect of demographic covariates on all-system average transaction times. Table 2 shows the estimated coefficients along with their 95% confidence intervals (BC_α) and net effect for each covariate in the optimal transaction time model. Transaction times increased significantly for subjects who reported eyewear, those with lower (darker) skin reflectance, and had a complex relationship with subject height and age (both linear and quadratic factors included). Again, skin reflectance was the factor with the greatest net effect, this time on transaction time, with a net effect of 20% on the intercept transaction time of 6.2 seconds. These results indicate that demographic factors significantly affected not only biometric matching effectiveness, but also the efficiency of biometric acquisitions.

	Covariate	Estimate	(BC_α)	Range	Net Effect
	Optimal Time Model				
$\hat{\beta}_0$	Intercept	6.161	(5.933, 6.417)	NA	NA
$\hat{\beta}_0$	Gender	−0.238	(−0.473, 0.029)	{0, 1}	0.238
$\hat{\beta}_2$	Eyewear	0.317	(0.050, 0.557)	{0, 1}	0.317
$\hat{\beta}_4$	Age	0.250	(0.114, 0.387)	(−1.47, 3.09)	1.139
$\hat{\beta}_4$	Age^2	0.134	(0.018, 0.270)	(0.0001, 9.53)	1.257
$\hat{\beta}_7$	Height^2	0.116	(0.021, 0.223)	(0.0004, 7.43)	0.860
$\hat{\beta}_{10}$	Reflectance	−0.258	(−0.384, −0.144)	(−2.41, 2.39)	1.235

TABLE 2
Parameter estimates for optimal models, fitting all-system average transaction times as discussed in Section 2.7.1. Parameters not included in the optimal model are not shown (see Equation 1). 95% confidence intervals BC_α are estimated using bootstrap (Section 2.7.2.). Net effect of the covariate is estimated as the product of $\hat{\beta}$ and the magnitude of the observed range of values for the covariate $|max - min|$. The unit of transaction time is seconds.

3.4 Relative Skin Reflectance is a Better Performance Predictor

The optimal models for average historic and same-day similarity scores as well as average transaction times presented in Tables 1 and 2 all retained skin reflectance, but not race as an explanatory variable. This suggested that our new phenotypic metric of relative skin reflectance is a better predictor of similarity score and transaction time than demographic race categories. To confirm this finding, we compared cross-validated R^2 for optimal models that include reflectance and non-optimal models replacing skin reflectance with race (Section 2.7.3). For each outcome variable, the AIC values were higher for models replacing skin reflectance with race. This is expected given our model selection approach (Section 2.7.1). Notably, however, cross-validated R^2 values were also marginally, but consistently lower for all three models replacing skin reflectance with race, providing strong evidence that skin reflectance scores perform better than self-reported race labels at predicting

mated similarity scores and transaction times. These results are shown in Table 3.

Model	R^2	R^2 CI	AIC
Historic			
Reflectance	0.1676	(0.1674, 0.1678)	−923.0524
Race	0.1605	(0.1603, 0.1607)	−919.8711
Same Day			
Reflectance	0.1004	(0.1002, 0.1006)	−1182.4279
Race	0.0969	(0.0967, 0.0971)	−1181.3800
Time			
Reflectance	0.1341	(0.1339, 0.1343)	1113.2365
Race	0.1149	(0.1147, 0.1151)	1122.2368

TABLE 3
R^2 and AIC model fitness estimates for the optimal models described in Tables 1 and 2 ("Reflectance" models). The "Race" models shows the R^2 and AIC values when relative skin reflectance is removed from the optimal model and replaced with race.

3.5 Demographic Effects Across Systems

We found that average similarity scores as well as the net effect of demographic covariates on historic similarity scores could vary between systems (Fig. 7A). To compare demographic effects across systems, we modeled the mated similarity scores of probes to historic gallery images across all tested systems using mixed effects models (Section 2.7.4). To identify those demographic covariates that varied between systems, we performed model selection using AIC, starting with the baseline optimal model selected for explaining all-system average similarity scores. Table 4 shows the AIC values for the optimal regression model (Optimal), developed based on average mated similarity scores and discussed in Section 3.2. Table 4 also shows several mixed effects models, namely the random intercept model (Optimal + $\hat{\beta}_{0,s}$) and models with random slopes included (Optimal + $\hat{\beta}_{0,s} + \beta_{n,s}covariate_j + \gamma_s$)). From Table 4, we can see AIC was reduced (i.e. the model was improved) with the addition of the random intercept $\hat{\beta}_{0,s}$, indicating that there were performance differences between systems. AIC was further reduced only with the addition of a random slope parameter on reflectance $\hat{\beta}_{10,s}$, indicating that reflectance, but not other covariates, had different effects on the performance of different systems.

Historic Model	AIC
Optimal	-6219.808
Optimal + $\hat{\beta}_{0,s} + \gamma_s$	-7069.695
Optimal + $\hat{\beta}_{0,s} + \beta_{1,s}gender_j + \gamma_s$	-7065.695
Optimal + $\hat{\beta}_{0,s} + \beta_{2,s}eyewear_j + \gamma_s$	-7065.723
Optimal + $\hat{\beta}_{0,s} + \beta_{4,s}age_j + \gamma_s$	-7069.181
Optimal + $\hat{\beta}_{0,s} + \beta_{7,s}height_j^2 + \gamma_s$	-7067.857
Optimal + $\hat{\beta}_{0,s} + \beta_{10,s}reflectance_j + \gamma_s$	-7074.124

TABLE 4
AIC values for fixed and mixed effects models fitted to historic mated similarity scores. The optimal historic model is as in Table 1. $\hat{\beta}_{0,s}$ is the random system intercept and $\beta n, s$ are the random system slopes for each named covariate as indicated.

The mixed effect modeling approach shows that a model which includes 1) the original fixed effects, 2) the system-specific intercept, and 3) a system-specific slope associated with reflectance, minimized the AIC. The coefficients of this optimal model are shown in Table 5. We note the fixed effect coefficients of the selected mixed effect model are approximately equal to the fixed effect coefficients of the selected average model (compare Tables 5 and 1). This demonstrates a consistency in modeling and that the average model in Table 1 is not unduly affected by system-specific outliers.

In the selected model, the difference in overall performance is captured by the system-specific intercept $\hat{\beta}_{0,s}$ and system-specific variation from reflectance is captured by the reflectance slope $\beta_{10,s}$. Plotting the net effect of system-specific reflectance ($(\beta_{10} + \beta_{10,s}) * \Delta_{Reflectance}$, where $\Delta_{Reflectance}$ is the span of observed reflectance values) on mated similarity scores versus the system-specifc intercept of each system showed that systems with lower overall levels of performance also showed a greater net effect of reflectance on mated similarity scores (Fig. 7B). In other words, better *overall* quality acquisition systems can maintain high performance across the full range of skin reflectance values. Systems 3 and 7 show the largest difference

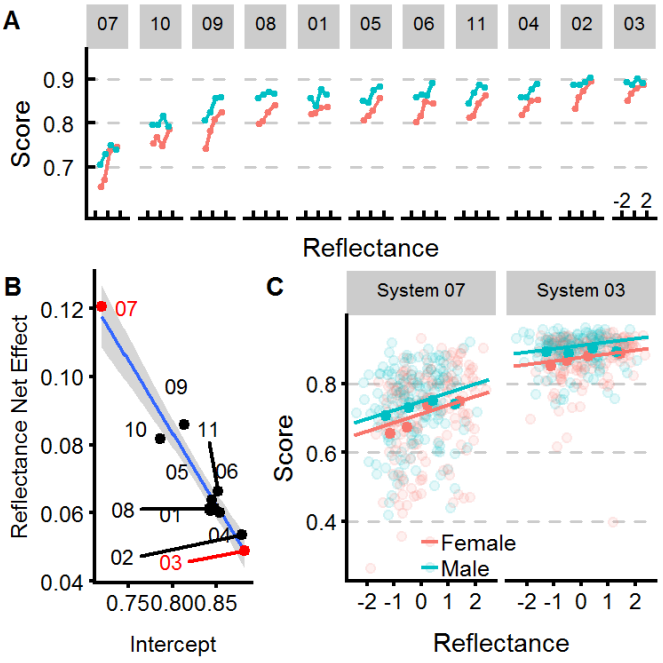


Fig. 7. The net effect of relative skin reflectance (Reflectance) is greater for systems with lower historic mated similarity scores. **A.** Mated similarity scores faceted by acquisition system. In each facet, colored points plot average scores across subjects binned by reflectance quartile. Facets are arranged based on the average similarity score produced by the acquisition system. Note variation in average scores across acquisition systems, consistently lower scores for women (see color key in panel C), and variation of scores with reflectance. **B.** Linear modeling estimates of the reflectance net effect on mated similarity scores $(\beta_{10} + \beta_{10,s}) * \Delta_{Reflectance}$ (where $\Delta_{Reflectance}$ is the span of observed reflectance values) plotted as a function of system-specific intercept $(\beta_0 + \beta_{0,s})$. Note decreasing net effect of reflectance with increasing level of performance (intercept). Points marked in red correspond to stations with the highest and lowest intercepts detailed in B. **C.** Distribution of mated similarity scores for systems marked red in B. Light points denote individual subject mated similarity scores. Dark colored points denote average scores across subjects binned by reflectance quartile. Lines indicate optimal age and gender model fits, fixing other factors constant at the average value of the subject population in each facet. Note higher net effect of reflectance on mated similarity scores for System 07 as well as lower overall mated similarity scores

	Covariate	Estimate	CI	Range
$\hat{\beta}_0$	Intercept	0.833	$(0.806, 0.860)$	NA
$\hat{\beta}_0$	Height^2	-0.006	$(-0.008, -0.003)$	$(0.0004, 7.43)$
$\hat{\beta}_1$	Gender	0.034	$(0.028, 0.039)$	$\{0, 1\}$
$\hat{\beta}_2$	Eyewear	-0.023	$(-0.029, -0.017)$	$\{0, 1\}$
$\hat{\beta}_4$	Age	0.007	$(0.005, 0.010)$	$(-1.47, 3.09)$
$\hat{\beta}_{10}$	Reflectance	0.014	$(0.011, 0.018)$	$(-2.41, 2.39)$
$\hat{\beta}_{0,s}$	System Intercept	*	*	$(0.71, 0.88)$
$\hat{\beta}_{10,s}$	System Reflectance	*	*	$(0.01, 0.03)$

TABLE 5
Parameter estimates and confidence intervals of the mixed-effects model fitted to mated similarity scores across systems. $\hat{\beta}_{1,s}$ represents the random intercept parameter and $\hat{\beta}_{2,s}$ represents the random slope parameter, which vary by system s. Mixed effects parameter estimates and associated 95% confidence intervals (marked with *) are plotted in Fig. 7

between system-specific intercepts, with a 0.16 difference in mated similarity scores. This value is comparable to the largest net effect of reflectance – a difference of 0.12 in mated similarity score between the highest and lowest reflectance observed on System 7. These differences are illustrated in Fig. 7C which allows visual inspection of the net effect of reflectance and acquisition system on mated similarity scores.

4 Discussion

Our analyses show that demographic factors influenced both the speed and accuracy of all eleven commercial biometric systems evaluated. For example, modeling showed that mated similarity scores were higher for men versus women, for older versus younger people, for those without eyewear, and those with relatively lighter skin. Of the different demographic covariates examined, our calculated measure of skin reflectance had the greatest net effect on average biometric performance (Tables 1 and 2). For mated similarity scores, the fixed effects of gender, eyewear, and age were constant across the tested systems while the magnitude of the random effect of skin reflectance varied between systems in a manner inversely correlated with overall system accuracy (Fig. 7).

The inverse relationship between the net effect of skin reflectance observed for a system and that system's overall performance (Fig. 7B) has important implications. Our data shows that systems with better overall performance had improved performance *most* for individuals with lower skin reflectance. Thus, a woman with darker skin using a superior system was more likely to match her mated gallery images than a man with lighter skin using an inferior system. It did not have to be this way. Performance for superior systems could have improved only for subjects with lighter skin while performance for those with darker skin stayed the same, or even decreased. Fortunately, this was not the outcome for the sample of commercial biometric syste ms we tested.

Another consequence of these results is that deploying a superior biometric acquisition system may significantly reduce or eliminate performance differences between some demographic groups. Indeed, in our data set, image quality varied between acquisition systems. As an example of this, Fig. 8 shows same-day enrollment images and face images

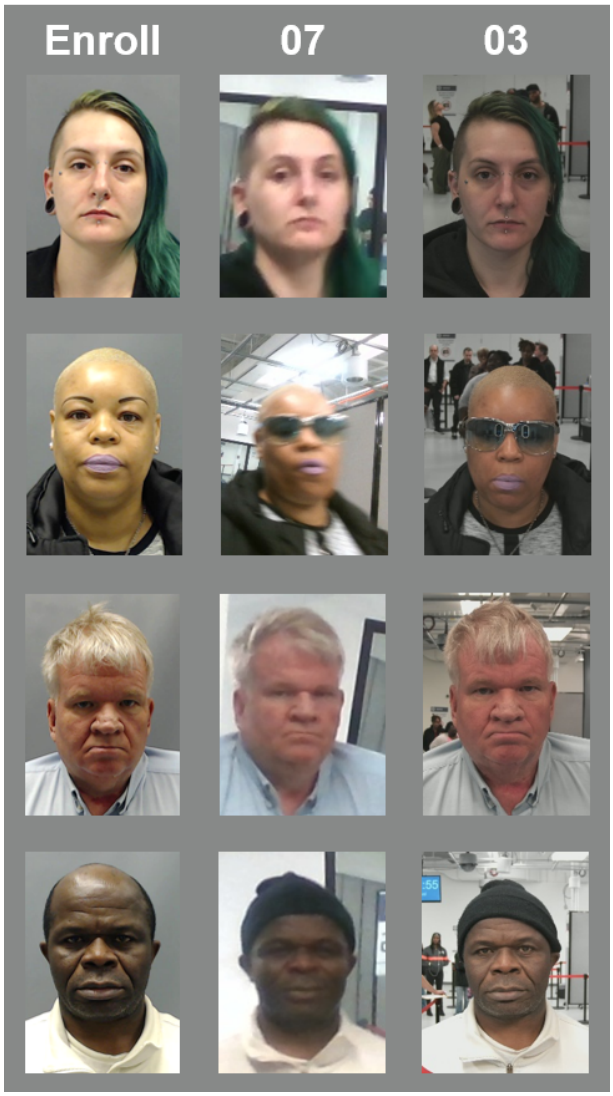


Fig. 8. Example imagery from enrollment, System 03, and System 07. Subjects shown are (in order from top), self-identified as White with highest skin reflectance, self-identified as Black or African-American with highest skin reflectance, self-identified as White with lowest skin reflectance, and self-identified as Black or African-American with lowest skin reflectance.

acquired by the two systems detailed in Fig. 7C for select subjects. This figure illustrates the strong variation in skin tone observed for individuals within each self-identified race group. As expected from Fig. 5B the two highest reflectance subjects in each self-reported race group are female and the two lowest reflectance subjects in each race group are male. Notably, face images acquired by System 07 show a stronger variation in pose, stronger motion blur, and lower contrast relative to System 03. This suggests that System 03 included a superior camera and stricter pose control. No company participating in the rally knew which algorithm we would use to evaluate their performance, precluding the possibility that any system was specifically tuned to the matching algorithm. While we do not attempt to establish a causal relationship, our data shows that acquisition system differences can strongly affect (magnify or eliminate) measured differences in algorithm accuracy across demographic categories.

Though emphasized less in this paper, user demographics also had strong effects on transaction times (Table 2) and, in addition to effects of reflectance, gender, and age, biometric performance also varied with volunteer height and weight. For high-throughput biometric systems, small changes in transaction times can lead to large changes in system throughput (e.g. reducing a five second transaction time by one second increases throughput by 25%). Effects of height and weight on transaction times can be expected for the tested systems since some systems adjusted camera position to each volunteer. Volunteer anthropometry may alter the speed with which systems adjusted to each volunteer and the speed with which the volunteers used the systems. It is possible that cameras took longer to find and focus on darker faces, explaining the longer transaction times observed for darker volunteers. Differences in face angle associated with volunteer height may explain the effect of this covariate on similarity scores. Taken together, our findings indicate that the acquisition system, independent of the matching algorithm, contributes to total biometric system performance across different demographic groups.

Notional arguments for why phenotypes may be superior to race category labels were laid out in [12]. In our data modeling, the skin reflectance phenotype was indeed a better predictor of performance than demographic race categories for all three independent measures of biometric performance investigated in this study, namely, historic similarity scores, same-day similarity scores, and transaction times (Section 2.7.3). Despite the fact that race and skin reflectance are highly correlated in our population, reflectance values could vary widely even within a single race category label (Fig. 5), indicating that reflectance carries distinct information. Importantly, skin reflectance can easily be calculated through our automated procedure; whereas race must be obtained manually via self-report or expert analysis, which may be subjective.

Our measure of reflectance is an estimate of the physical properties of skin and is not a measure of the intensity of skin pixels in the acquired probe images. Recent work found no effect of skin pixel intensity on gender classification accuracy [20] suggesting that other factors underlie differences in classification performance. However, face skin pixel intensity alone does not reflect the interaction between the incident light, the skin tissue, and finally the camera sensor, which can result not only in changes to face pixel intensity, but in changes to the discriminative information contained in the face image (e.g. spatial frequency content or under-saturation). Nonetheless, the space of possible facial phenotypes related to categories like race or nationality likely has many relevant dimensions rooted in population genetics. A more complete description of biometric performance will likely include additional phenotypes and would benefit from tethering to our understanding of the role of genes and environment in shaping these face phenotypes.

Two images of the same person taken on different occasions will differ systematically with the passage of time due to face aging [9] [10]. Here we compared demographic effects on mated scores for images taken on the same-day versus on different days (1 month to 4 years). We found that gender and age covariates were notably excluded from optimal models of same-day similarity scores. This finding suggests that faces of older people in our sample may be more stable in their appearance over time relative to younger people. On the other hand, faces of women in our sample are more variable over time relative to men, possibly due to differences in hair-styling and makeup. That skin reflectance was preserved as a covariate in optimal models of same-day scores argues that this effect, as expected, is a fixed trait of the subject and varies little over time.

In our study as well as in [7], mated similarity scores were lower for those identifying as Black or African-American. Mated similarity scores for Black or African-American subjects have previously been reported as higher than for White subjects [9] [10]. Such differences may arise due to differences in biometric algorithm, differences in race labels (our race labels were self-reported), differences in the quality of images, or in the traits of individuals included the datasets. In our sample, subjects identifying as Black or African-American varied significantly in skin reflectance, and reflectance within each race further varied by gender (Fig. 5). Given our demonstrated effect of skin reflectance on biometric performance, we suggest that differences in skin reflectance should be considered when comparing biometric performance for race labeled datasets [9].

Our work comes with important caveats which we hope can be addressed with further experimentation and analysis. First, while we examined similarity scores for images acquired on eleven systems, we used only a single commercial matching algorithm. Since demographic effects can vary significantly between algorithms [8] it will be important to consider how the choice of matching algorithm affects biometric system performance across diverse demographics. Second, we examined only mated similarity scores and our modeling only accounted for a fraction ($< 20\%$) of the total score variance. Operational biometric system performance will depend on thresholds set relative to both the mated and non-mated score *distributions*. Future work will need to clarify how true accept rates and false accept rates are affected by skin reflectance and other demographic variables. Future work will also need to evaluate these effects across different commercial matching algorithms. Finally, future work should compare our measure of relative skin reflectance, derived from photos, against both common survey instruments such as the Fitzpatrick scale [12] and objective instruments [21].

As biometrics continue to be integrated into everyday processes, it is important to understand the underlying cause of any observed demographic effects. This allows system designers, algorithm researchers, and operational users to precisely pinpoint how to make the technology equitable to all demographic groups.

ACKNOWLEDGMENTS

This research was funded by the Department of Homeland Security, Science and Technology Directorate on contract number W911NF-13-D-0006-0003. The views presented here are those of the authors and do not represent those of the Department of Homeland Security, the U.S. Government, or their employers. The biometric dataset utilized in this work is not publicly available at this time.

The authors thank the staff of the SAIC Identity and Data Sciences Laboratory: Andrew Blanchard and Kirsten Huttar for providing software engineering support; Laura Rabbitt and Nelson Jaimes for human factors support; Frederick Clauss and Jeffrey Chudik for providing integration engineering support; Jacob Hasselgren for directing rally execution; Rebecca Rubin for technical document support and editing; as well as Rebecca Duncan, Colette Bryant, Kevin Slocum, and Robert Wilson-Cruze for support in Rally execution. The authors thank Patrick Grother and James Matey of the National Institute of Standards and Technology for helpful comments and suggestions on early versions of the manuscript.

The paper authors acknowledge the following author contributions: Cynthia Cook performed linear modeling and wrote the paper; John Howard conceived the work, advised on statistical analysis, and wrote the paper; Yevgeniy Sirotin conceived the work, developed skin reflectance analyses, directed statistical analysis, and wrote the paper; Jerry Tipton and Arun Vemury conceived the work and edited the paper.

References

[1] B. Friedman, E. Brok, S. K. Roth, and J. Thomas, "Minimizing bias in computer systems," *ACM SIGCHI Bulletin*, vol. 28, no. 1, pp. 48–51, 1996.

[2] J. N. Pato and L. I. Millet, "Biometric recognition: Challenges and opportunities," National Academies Press, Washington, D.C., Tech. Rep., 2010.

[3] C. Garvie, A. M. Bedoya, and J. Frankle, "The perpetual lineup: Unregulated police face recognition in america," Georgetown Law Center on Privacy and Technology, Tech. Rep., Oct 2016, http://www.perpetuallineup.org/, last accessed on 06/07/18.

[4] Y. B. Sirotin, J. A. Hasselgren, and A. Vemury, "Usability of biometric iris-capture methods in self-service applications," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 60, no. 1, pp. 2019–2023, 2016. [Online]. Available: https://doi.org/10.1177/1541931213601459

[5] Y. B. Sirotin, "Usability and user perceptions of self-service biometric technologies." International Biometric Performance Conference, Gathersburg, MD, 2016, https://www.nist.gov/sites/default/files/documents/2016/12/06/07_ibpc_usability_20160414.pdf, last accessed on 06/07/18.

[6] J. J. Howard, A. A. Blanchard, Y. B. Sirotin, J. A. Hasselgren, and A. Vemury, "An investigation of high-throughput biometric systems: Results of the 2018 department of homeland security biometric technology rally," in *2018 Nineth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*. IEEE, 2018.

[7] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1789–1801, Dec 2012.

[8] P. Grother, M. Ngan, and K. Hanaoka, "Ongoing face recognition vendor test (frvt) part 1: Verification," National Institute of Standards and Technology, Tech. Rep., Apr 2018, https://www.nist.gov/sites/default/files/documents/2018/04/03/frvt_report_2018_04_03.pdf, last accessed on 06/07/18.

[9] L. Best-Rowden and A. K. Jain, "Longitudinal study of automatic face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 148–162, Jan 2018.

[10] D. Deb, L. Best-Rowden, and A. K. Jain, "Face recognition performance under aging," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 548–556.

[11] U. Park, Y. Tong, and A. K. Jain, "Age-invariant face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 5, pp. 947–954, 2010.

[12] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. New York, NY, USA: PMLR, 23–24 Feb 2018, pp. 77–91. [Online]. Available: http://proceedings.mlr.press/v81/buolamwini18a.html

[13] U. Census, "Race and ethnicity," United States Census Bureau, Tech. Rep., Jan 2017, https://www.census.gov/mso/www/training/pdf/race-ethnicity-onepager.pdf, last accessed on 06/07/18.

[14] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A survey on pixel-based skin color detection techniques," in *IN PROC. GRAPHICON-2003*, 2003, pp. 85–92.

[15] T. M. Michael J. Taylor, "Adaptive skin segmentation via feature-based face detection," in *Proc.SPIE*, vol. 9139, 2014, pp. 9139 – 9139 – 12. [Online]. Available: https://doi.org/10.1117/12.2052003

[16] G. Zonios, J. Bykowski, and N. Killias, "Skin melanin, hemoglobin, and light scattering properties can be quantitatively assessed in vivo using diffuse reflectance spectroscopy," *Journal of Investigative Dermatology*, vol. 117, pp. 1452 – 1457.

[17] B. Efron, "Better bootstrap confidence intervals," *Journal of the American Statistical Association*, vol. 82, no. 397, pp. 171–185, 1987.

[18] K. W. Brady West and A. Galecki, *Linear Mixture Models: A Practical Guide Using Statistical Software*, 2nd ed. Boca Raton: Chapman-HAll/CRC, 2014.

[19] R. Tibshrani, G. Walther, and T. Hastie, "Estimating the number of clusters in a dataset via the gap statistic," *Journal of the Royal Statistical Society Series B (Statistical Methodologies)*, vol. 63, no. 2, pp. 441–423, 2001.

[20] V. Muthukumar, T. Pedapati, N. Ratha, P. Sattigeri, C.-W. Wu, B. Kingsbury, A. Kumar, S. Thomas, A. Mojsilovic, and K. R. Varshney, "Understanding unequal gender classification accuracy from face images," *arXiv preprint arXiv:1812.00099*, 2018.

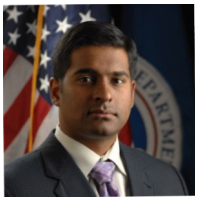
[21] M. van der Wal, M. Bloemen, P. Verhaegen, W. Tuinebreijer, H. de Vet, P. van Zuijlen, and E. Middelkoop, "Objective color measurements: clinimetric performance of three devices on normal skin and scar tissue," *Journal of Burn Care & Research*, vol. 34, no. 3, pp. e187–e194, 2013.

Cynthia Cook Cynthia Cook earned her M.S. in Statistics from American University. Her research interests include the areas of social data analytics, statistical networks, relational data structures, and survey sampling. She currently works as a Senior Data Scientist at the Identity and Data Sciences Laboratory at SAIC, which supports applied research in biometric identity technologies at the Maryland Test Facility.

John Howard Dr. Howard earned his Ph.D. in Computer Science from Southern Methodist University. His thesis was on pattern recognition models for identifying subject specific match probability. His current research interests include biometrics, computer vision, machine learning, testing human machine interfaces, pattern recognition, and statistics. He has served as the principal investigator on numerous R&D efforts across the intelligence community, Department of Defense, and other United States Government agencies. He is currently the Principal Data Scientist at the Identity and Data Sciences Laboratory at SAIC where he conducts biometrics related research at the Maryland Test Facility.

Yevgeniy Sirotin Dr. Sirotin holds a Ph.D. in Neurobiology and Behavior from Columbia University and has diverse research interests in behavior and human computer interaction. His past research spans mathematical psychology (cognitive modeling), neurophysiology (multi-spectral imaging of the brain), psychometrics (mechanisms of visual and olfactory perception), biometrics (design and testing of identity systems), and human factors (usability). He currently works as Principal Investigator and Manager of the Identity and Data Sciences Laboratory at SAIC which supports applied research in biometric identity technologies at the Maryland Test Facility.

Jerry Tipton Jerry Tipton is the Program Manager at SAIC's Identity and Data Sciences Lab. He has over 20 years experience in the biometric industry with over 15 years managing research portfolios in support of various United States Government agencies. He currently supports the Department of Homeland Security, Science and Technology Directorate at the Maryland Test Facility.



Arun Vemury Arun Vemury received his M.S. in Computer Engineering from George Washington University. His current research interests include biometrics, pattern recognition, machine learning, and operations research. He serves as the Director of the Biometrics and Identity Technology Center for the United States Department of Homeland Security Science and Technology Directorate.