DHS S&T Technical Paper Series

Demographic Effects Across 158 Facial Recognition Systems

Cynthia M. Cook
John J. Howard
Yevgeniy B. Sirotin
Jerry L. Tipton

The Maryland Test Facility,
Identity and Data Sciences Laboratory

Arun R. Vemury

The U.S. Department of Homeland Security
Science and Technology Directorate
Biometric and Identity Technology Center

Keywords: Face Recognition, Demographic, Skin Reflectance, Scenario Testing, Commercial Matching Systems, Commercial Acquisition Systems

August 2023

Homeland Security
Science and Technology

Executive Summary

BACKGROUND: This study was sponsored by the U.S. Department of Homeland Security (DHS) and conducted at the Maryland Test Facility (MdTF) as part of ongoing evaluations of biometric performance across demographics groups. Using data gathered at the MdTF, we extend the demographic analysis previously published following the 2018 Biometric Technology Rally. The 2018 study analyzed data from 11 acquisition systems and one matching algorithm to draw conclusions regarding the impact of race, skin lightness, gender, and other demographic variables on biometric performance.

MOTIVATION: Analyses conducted following the 2018 Rally were meant to explain the demographic effects on rank one mated similarity scores. Since, they have been widely cited as evidence of demographic differentials in face recognition technology more broadly. However, the 2018 study included a relatively limited sample of face recognition systems (11 acquisition and one matching system). It remained unclear if those results were generalizable to face recognition as a technology. Here, we examine 158 combinations of acquisition systems and matching algorithms with the goal of extending the 2018 Rally analysis and gaining a better understanding of demographic effects on mated similarity scores using a significantly larger set of biometric system combinations. The presence of demographic effects in mated similarity scores may or may not cause increased error rates in these systems, depending on system configuration.

WHAT WE FOUND: We find the conclusions from the 2018 Rally remain consistent in most system combinations tested since. We show that both self-reported demographic variables, as well as measured skin lightness, affect rank one mated similarity scores across a wide variety of system combinations. The majority of system combinations tested showed a statistically significant effect related to eyewear, gender, and skin lightness on rank one mated similarity scores. When modeled using regression techniques, mated similarity scores averaged across acquisition system were lower for people wearing eyewear on 96% of models. For 74% of models, women tended to have lower mated similarity scores when matched to a gallery of historic images. This gender effect disappears when matching the same probe images to a gallery of face images taken on the same day. For 57% of models, those with darker skin had lower mated similarity scores. We further show that, for models where skin lightness is found to be significant, skin lightness is a better predictor of average mated similarity scores than self-reported race.

Demographic Effects Across 158 Facial Recognition Systems

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

1 Introduction

Biometric face recognition systems have gained widespread adoption in a variety of use-cases, including in the U.S. Department of Homeland Security (DHS) [1], [2]. As part of its efforts to identify technologies that perform well in such DHS use-cases, the DHS Science and Technology Directorate (DHS S&T) sponsors biometric research and scenario testing at the Maryland Test Facility (MdTF).

In 2018, DHS S&T carried out the first Biometric Technology Rally: a large scale scenario test designed to measure the performance of commercial biometric technology within a simulated, high-throughput, unattended border control process. A particular focus of this test was to ascertain if the performance of tested commercial face recognition systems varied for people belonging to different demographic groups [3]. DHS S&T found that demographic factors influenced the similarity score output of all tested face recognition systems. More specifically, 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 factors examined, measures of skin lightness had the greatest net effect on average biometric performance [4]. Since 2018, DHS S&T has continued to carry out yearly Biometric Technology Rallies. This technical paper extends the original analysis to the 2019, 2020, and 2021 Biometric Technology Rallies [5], finding that the rank one mated similarity scores returned by face recognition systems continue to be influenced by eyewear, gender, and skin lightness.

2 Background

Biometric recognition systems are made up of multiple components. At a minimum, a biometric recognition system consists of a biometric data subject whose identity is to be ascertained or confirmed (“subject”), a sensor which captures samples (“probe” images) of subject’s biometric characteristics (“acquisition system”), and an algorithm that processes and compares information across different biometric samples to compute a similarity score (“matching system”). Biometric identification systems specifically also utilize a database of biometric samples with known identities (“gallery”). Biometric scenario testing, as defined by ISO/IEC 19795-2 [6], is the process by which multiple system components are combined to measure the performance of simulated full systems. Each component can independently alter the performance of the full biometric system. This report uses statistical modeling to examine the demographic performance across biometric face recognition systems, which are composed of a commercial acquisition and matching system.

The first published analysis of demographic effects in biometric facial acquisition systems was based on the 2018 DHS S&T Biometric Technology Rally [4] and has been widely cited as evidence of demographic differentials or “bias” in face recognition technology more broadly [7] [8] [9] [10] [11] [12] [13] [14] [15]. However, [4] tested just one face recognition algorithm in combination with eleven acquisition systems. Since 2018, three additional Rallies have been conducted at the MdTF. Each test was conducted in the same facility with the objective to measure the effectiveness of commercial face recognition technology in an unstaffed, high-throughput scenario. Each Rally tested a subset of the acquisition systems and matching algorithms available on the market in that year and each recruited a new subset of test volunteers from the local area to serve as subjects. Each test was designed to follow the same process, with the exception of introduction of masking and social distancing in the 2020 and 2021 Rallies (this manuscript examines the performance of face recognition systems identifying subjects asked to remove their face masks prior to image acquisition). Table 1 shows, for each Rally, the number of participating volunteers, face acquisition systems, face matching systems, and the resulting number of system combinations tested.

This technical paper analyzes performance for a total of 158 system combinations tested from 2019 through 2021 with a cumulative sample of 1,590 volunteers, with 949 unique individuals (Table 1). Some systems that were tested as part of each Rally were not included in this report. For example, this report focuses on visible light face recognition and one acquisition system was not included in analysis because it acquired samples in the near infra-red wavelength range. A total of five matching systems were also

- *C. Cook, J. Howard, Y. Sirotin, and J. Tipton staff the Identity and Data Sciences Laboratory at the Maryland Test Facility.*
- *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 info@mdtf.org*

Rally	Acquisition Systems	Matching Systems	System Combinations	Volunteers
	Tested (Analyzed)			
2018	11	1	11	363
2019	11 (10)	8 (7)	88 (70)	430 (428)
2020	6 (6)	10 (8)	60 (48)	582 (570)
2021	5 (5)	10 (8)	50 (40)	601 (592)
Total (19-21)	22 (21)	28 (23)	198 **(158)**	1613 (1590)

TABLE 1
Counts of the acquisition systems, matching systems, system combinations, and volunteers participating in the 2018-2020 Biometric Technology Rallies (Tested) and those examined in this report (Analyzed; see text for details).

excluded from analysis because they were judged to be non-representative of state-of-industry systems, due to persistent technical issues encountered during testing. Finally, 23 volunteers (cumulative) were excluded from analysis in this report because they had missing data, declined to provide requisite demographic information, or did not participate in measurement of skin tone.

Using data from the analyzed system combinations and volunteers, this technical paper replicates and extends the analyses performed in 2018 (described in Section 3). Section 3 provides the methods employed in data acquisition and analysis. Section 4 presents the overall results these analyses and Section 5 discusses the significance of the results.

3 Methods

3.1 Process and Data

Each Biometric Technology Rally was carried out at the MdTF in Upper Marlboro, MD. The test process and evaluation for each of the four Rallies were designed to provide a systematic, repeatable framework for evaluating the effectiveness of biometric systems. Briefly, acquisition systems providers were given a three month period to design and implement a biometric system capable of capturing samples from subjects as they interacted with the system. Acquisition systems were required to be both “high-throughput,” defined as having transaction times under five seconds, and highly effective, defined as true identification rates above 99%. Both in-gallery and out-of-gallery (i.e. “distractor”) subjects interacted with acquisition systems. Although not discussed here, true identification rate in the original Rally concept was the combination of: 1) correct identifiers for in-gallery subjects returned at rank-one above threshold and 2) no identifier returned at any rank above threshold for out-of-gallery subjects. Acquisition systems were also required to operate in a confined 6x8 foot space and be entirely unstaffed.

All test volunteers 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 options provided volunteers included the five U.S. Census categories from 2017 [16] and “Other”. For analysis, the categories American Indian or Alaskan Native, Native Hawaiian or Other Pacific Islander, and Asian are combined with the Other category to create three categories: “White” (W), “Black or African-American” (B), and “Other Race” (O). Subjects were selected such that each test contained a demographically diverse group in terms of gender, race, and age (Figure 1).

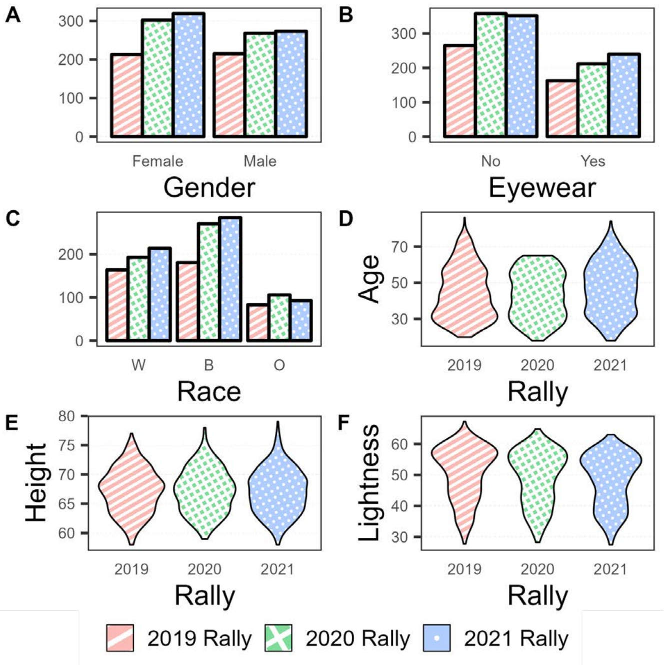

Fig. 1. Demographic factor distributions for volunteers included in each Rally test: 2019 (Pink with Strips); 2020 (Green with Crosses); 2021 (Blue with Dots). **A.** Counts of volunteer self-reported gender: Female (F); Male (M). **B.** Counts of volunteer self-reported use of eyewear: No (N); Yes (Y). **C.** Counts of volunteers self-reported race: Black or African-American (B); White (W); “Other Race” (O). **D.** Distribution of volunteer age. **E.** Distribution of volunteer height. **F.** Distribution of volunteer face skin lightness (L*).

For the 2019 - 2021 Rallies, skin lightness was electronically quantified using a calibrated dermatological color device (cyberDERM, DSM III Colormeter [17]). This sensor measures skin color in the CIELAB color space by using an RGB sensor to image a 7 mm^2 patch of skin under standard illumination provided by two white light emitting diodes. The device can accurately measure the color as well as erythema and melanin indices for skin [18] [19]. Skin lightness measures were collected by an attendant, using this device, from each hand and each temple. Additional details are available in [20]. The L* component of CIELAB was used as the measure of skin lightness during analysis. This use of skin lightness values from an in person dermatological sensor is an important distinction of this work, as skin lightness has been shown to be unreliable when assessed from images [21] [22] [23]. Figure 2 shows example enrollment images and the corresponding L* values for a select number of subjects in this study.

3.1.1 *Face Image Galleries*

The gallery is an important part of a face recognition system. This analysis examined the performance of biometric systems with two galleries: “same-day” and “historic”.

The images in the “same-day” gallery were collected from the volunteers at the start of each test session. This

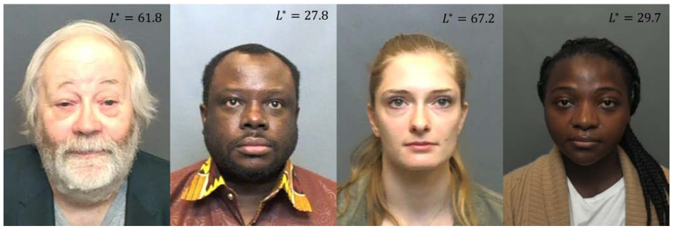

Fig. 2. Enrollment image for select subjects and their measured skin lightness value (L*).

gallery was designed to house high-quality reference samples taken on the same day as the probe samples acquired by the tested acquisition systems. Subjects were enrolled into the "same-day" gallery by staff trained in biometric collection. Subjects stood in front of a 18% neutral gray background with diffuse illumination ($600 - 650$ lux). Enrollment staff collected a single face image using a Logitech C920 camera at a 1 meter standoff (resolution: 1920x1080). Staff asked volunteers 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 430, 582, and 601 face samples from the 2019, 2020, and 2021 Rallies, respectively.

The "historic" face image gallery consisted of biometric samples acquired from prior test events, dating between 2014 and at least one month prior to each Rally test in question. These samples were acquired using a variety of cameras including digital single lens reflex (DSLR) cameras, web-cameras, and cameras embedded in biometric systems. This gallery was designed to be broadly representative of identification galleries used during airline boarding and contained images of varying quality. Table 2 describes each Rally's "historic" gallery, which consisted of a number of samples or gallery images from 500 unique people where some people are Rally test volunteers and others are "distractors" or out-of-gallery subjects for the Rally.

Rally	Gallery Images	Gallery Participants	Samples per Participant	Test Volunteers	Distractors
2019	1958	500	3.92	430	78
2020	1479	500	2.96	582	93
2021	1925	500	3.85	601	113

TABLE 2
Description of each Rally "historic" gallery.

3.1.2 Rank One Mated Similarity Scores

For each Rally test, probe face images acquired by each acquisition system were compared, using each matching system, against both the same-day and historic galleries producing a set of similarity scores. The set of similarity scores against the same-day gallery used samples from all 430, 582, 601 test volunteers, while the corresponding set for the historic gallery used only samples from the 352, 489, 488 test volunteers who had corresponding images in the historic gallery (in-gallery subjects). Every Rally $r \in \{2019, 2020, 2021\}$ had different acquisition systems a, matching systems m, and volunteers s which were matched to each gallery $g \in \{\textit{historic}, \textit{same-day}\}$. Since all probe subjects were in-gallery, the rank one mated similarity score $\Phi^{g,m}_{a,s}$ is defined as the maximum score obtained for the probe image against a gallery. Probe images for which the rank one similarity score was higher than the mated similarity score were removed from this analysis. This was done because some systems occasionally had technical issues, such as sending a probe photo late, or to aquiring images for individuals in the background. This subsequently caused a ground-truth labeling error (i.e. the ground-truth identifier of the probe image was incorrect). Such instances occurred in less than 1% of collected data. Matching system providers did not have the opportunity to normalize galleries g across template space or otherwise "finalize" individual galleries. Thus each "identification operation" in this sense can be viewed as a set of sorted 1:1 comparisons. This is representative of how some, if not the majority, of biometric systems perform identification [24].

3.2 Statistical Modeling

Statistical modeling was performed to explain the variation in $\Phi^{g,m}_{a,s}$ of each system combination and volunteer demographics. Modeling was performed for each system combination in each Rally. Given 158 system combinations (see Table 1) and two separate identification galleries (g), there were 316 rank one similarity score distributions to model: 158 system combinations evaluated against the historic gallery ("historic" system combinations) and 158 system combinations evaluated against the same-day gallery ("same-day" system combinations). These final modeled system combinations included 23 unique matching systems and 21 unique acquisition systems, though not in full combination (e.g. $21 \times 23 \neq 158$, see Table 1). Statistical modeling results are presented below.

3.2.1 System Combination Models

We used linear regression models to explain the variation in rank one mated similarity scores produced by different system combinations across volunteers based on volunteer demographics [4]. Modeling was performed using the R statistical programming language. For the rank one mated similarity scores produced by each system combination $\Phi^{g,m}_{a,s}$, we constructed a "full" linear model using nine demographic covariates according to Equation 1. We included three categorical variables: gender, eyewear, and race as well as three continuous variables: age, height, and skin lightness. We normalized the continuous variables age, height, and skin lightness prior to fitting according to $z = (x - \mu_x)/\sigma_x$. For each continuous variable, we also included their squared transformations, as deviation from the mean could produce a significant effect (i.e. very short or very tall subjects can see changes in score due to pitch angle to the camera) . The inclusion of interaction terms, which could lead to over-fitting, was not considered in this analysis.

$$\begin{aligned} \mathbf{\Phi} = \Phi^{g,m}_{a,s} = \beta_{0,a} + \beta_{1,a} gender_s + \beta_{2,a} eyewear_s + \\ \beta_{3,a} race_s + \beta_{4,a} age_s + \beta_{5,a} age^2_s + \beta_{6,a} height_s + \quad (1) \\ \beta_{7,a} height^2_s + \beta_{8,a} lightness_s + \beta_{9,a} lightness^2_s + \epsilon_{a,s} \end{aligned}$$

We estimated model parameters $\boldsymbol{\beta}$, using ordinary least squares (OLS) fitting. We then found the "optimal" system combination model by down-selecting demographic covariates to minimize 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 nine demographic covariates. We then applied a step wise covariate selection procedure in both directions using the `stepAIC()` function in the R package MASS. We applied this procedure to select optimal historic system combination models and, separately, select optimal same-day system combination models. Equation 2 describes the optimal system combination model with $k-1$ covariates selected, for the sth volunteer and ath acquisition system. There was one optimal average model for each unique system combination for a total of 158 historic and 158 same-day models.

$$\begin{aligned}\mathbf{x} = x_{a,s}^{g,m} &= [x_{1,s}, x_{2,s}, ...x_{k-1,s}]\\ \boldsymbol{\beta} = \beta_{a}^{g,m} &= [\beta_1, \beta_2, ..., \beta_{k-1}]\\ \boldsymbol{\Phi} = \Phi_{a,s}^{g,m} &= \beta_{0,a} + \boldsymbol{\beta}^T\mathbf{x} + \epsilon_{a,s}\end{aligned} \tag{2}$$

3.2.2 Monte Carlo Parameter Selection

We performed Monte Carlo Simulations to estimate the likelihood of selecting a random covariate by chance in our optimal models as described in Section 3.2.1. For each system combination model and each of the nine covariates, we ran 1,000 simulations. Each simulation started by creating a new randomized covariate by re-sampling from the original covariate distribution, with replacement. This sampling removes any link between the demographic covariate and the response variable $\Phi_{a,s}^{g,m}$ while preserving other statistical properties. We then replaced the true covariate with the randomized covariate, re-fit the model, and applied stepwise regression to minimize AIC. We then calculated the proportion of simulations in which the randomized covariate was selected in the optimal model. We find that for all simulations, the likelihood of selecting a random covariate by chance is approximately 16%. Given our degrees of freedom and number of observations, this is in line with expectations when identifying noise variables [25].

3.2.3 Average Models

We used linear regression models to examine the relationship between volunteer demographics and average rank one mated similarity scores returned by each matching system across all acquisition systems included in each Rally test. For each volunteer s, we first averaged the rank one mated similarity scores across each acquisition system according to $\bar{\Phi}_s^{g,m} = \frac{1}{N_r}\sum_a \Phi_{a,s}^{g,m}$ where $N_r \in \{10, 6, 5\}$ is the number of acquisition systems analyzed in each Rally r (Table 1), m is the matching system and g corresponds to the reference gallery (either historic or same-day). We then created a full "average" linear model for $\bar{\Phi}_s^{g,m}$ using the same nine demographic covariates as described in Section 3.2.1 according to Equation 3.

$$\begin{aligned}\bar{\boldsymbol{\Phi}} = \bar{\Phi}_s^{g,m} = \beta_0 + \beta_1 gender_s + \beta_2 eyewear_s +\\ \beta_3 race_s + \beta_4 age_s + \beta_5 age_s^2 + \beta_6 height_s + \beta_7 height_s^2 +\\ \beta_8 lightness_s + \beta_9 lightness_s^2 + \epsilon_s\end{aligned} \tag{3}$$

Next, we estimated model parameters β using the same procedure as in Section 3.2.1. Equation 4 describes a final "optimal average" model with $k-1$ covariates selected. There was one optimal average model for each unique matching system for a total of 158 historic and 158 same-day models.

$$\begin{aligned}\bar{\boldsymbol{x}} = \bar{x}_s^{g,m} &= [\bar{x}_{1,s}, \bar{x}_{2,s}, ..., \bar{x}_{k-1,s}]\\ \boldsymbol{\beta} = \beta^{g,m} &= [\beta_1, \beta_2, ...\beta_{k-1}]\\ \bar{\boldsymbol{\Phi}} = \bar{\Phi}_s^{g,m} &= \beta_0 + \boldsymbol{\beta}^T\bar{\boldsymbol{x}} + \epsilon_s\end{aligned} \tag{4}$$

3.2.4 Bootstrapping for Estimated Confidence Intervals

We assessed the accuracy of model fits through residual analysis. For the majority of both the system combination and the average models, we found the residuals deviated from normality, with noticeable deviations present in the QQ plots of our response variables $\{\Phi, \bar{\Phi}\}$ (data not shown), likely due to the presence of outliers. We therefore obtained confidence intervals for model parameter estimates using a bootstrapping technique instead of relying on the standard error for each average optimal model. We generated 3,000 bootstrap samples and calculated the bias corrected bootstrapped confidence intervals or the BC_α, by sampling from the rank one scores, for each of the fitted coefficients in the optimal models [26]. Covariates with 95% BC_α confidence intervals that contain 0 were removed from the optimal system combination and optimal average models.

3.2.5 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 via examination of bootstrapped confidence intervals. These covariates are therefore excluded from the optimal models (Equation 3). To independently confirm the optimal selection of covariates included in our optimal model, we used the non-parametric technique of cross-validation. Specifically, of interest for further examination was the inclusion of either race or skin lightness. For all models, we specify a base model (B), which includes all optimal terms except for any race or lightness terms. We then consider three models built from the base model, namely the base model including race (B+R), the base model including lightness (B+L), and the base model including both lightness terms (B+L+L^2). For each model, we performed ten-fold cross-validation and compared the cross-validated R^2 of the base model to all other models. 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. Comparing the cross validated model with the highest R^2 to the optimal model, we can confirm the inclusion of the race and lightness terms in the optimally fit model.

3.2.6 Mixture Models for Cross System Effects

The average models explain the effects of demographic covariates on average mated similarity scores (Section 3.2.3). We examined whether rank one mated similarity scores for samples acquired on different acquisition systems were associated with distinct demographic covariate effects in combination with a given matching system. This analysis was possible because each volunteer in each Rally interacted with each of the acquisition systems.

To model these effects, we applied linear mixture modeling with system a_r as the random effect. To start, we used all demographic covariates retained in the optimal model from Equation 4 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 effect: $\beta_{0,a}$ and $\boldsymbol{\beta}_a^T$) as described in Equation 5 where $\mathbf{y}$ is the set of k selected system-specific slope covariates and $\boldsymbol{\beta}_a$ are the corresponding parameters.

$$\begin{aligned} \mathbf{y} = y_s &= [x_{1,s}, x_{2,s}, ...x_{k,s}] \\ \boldsymbol{\beta}_a &= [\beta_{1,a}, \beta_{2,a}, ...\beta_{k,a}] \\ \Theta_{a,s} = \beta_0 + \boldsymbol{\beta}^T\mathbf{x} &+ \beta_{0,a} + \boldsymbol{\beta}_a^T\mathbf{y} + \epsilon_{a,s} \end{aligned} \tag{5}$$

Starting with only the fixed effects model, we added an acquisition system-specific slope $\beta_{0,a}$. If this reduced AIC, it signified that there are statistical performance differences between systems. Then, given the intercept model that includes $\beta_{0,a}$, we used a forward model selection approach to identify the mixed individual effects that continue to minimize AIC, adding each demographic covariate ($\mathbf{y}$) 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 acquisition systems* for this demographic factor. We performed this procedure for the historic gallery similarity scores. Since the goal of this analysis was to estimate the acquisition system-specific effects, we estimated all model parameters β, by maximizing the restricted likelihood (REML) [27].

4 Results

4.1 Demographic Effects Observed in Optimal Historic System Combination Models

We first examined the effects of volunteer demographics on each system combination selected for modeling (Section 3.2). We used linear modeling (Section 3.2.1) to determine whether a relationship between each demographic covariate and rank one mated similarity score was present and to estimate the direction (positive or negative) of each relationship. We do not report the size of the relationship between a demographic covariate and score because each matching system produces scores that scale along different units.

We first modeled similarity scores for 158 historic system combination models using rank one mated scores against a historic gallery of samples gathered during prior test events (Section 3.1.1). Starting with a full model including nine demographic covariates (Equation 3), we used an AIC-based model selection approach to find the 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 and removed those covariates for which the confidence interval of the parameter estimate included 0.

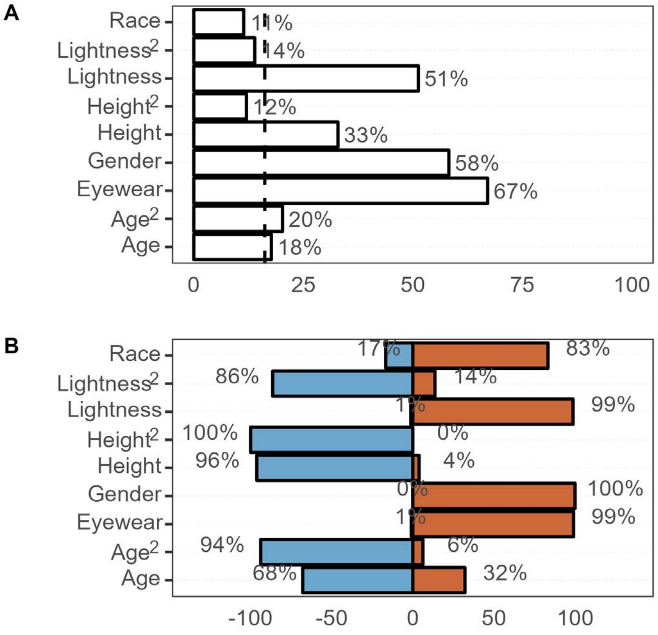

Fig. 3. **A.** Percentage of optimal historic system combination models ($n = 158$) retaining each listed demographic covariate. Vertical dashed line represents percentage of optimal models expected to retain the covariate by chance (Section 3.2.2). **B.** Percentage of optimal historic system combination models that retain each covariate with a positive (Red) versus a negative (Blue) relationship determined as the sign of the lower bound of the parameter estimate.

Figure 3A plots the percentage of the 158 optimal historic system combination models retaining each demographic covariate. It shows that eyewear, gender, and lightness were the only demographic covariates retained in the majority of models; 67%, 58%, 51%, respectively. Recall that each covariate has a 16% probability of appearing in an optimal model by chance (see Section 3.2.2). Interestingly, race and age were retained in just 11% and 18% of the models, respectively; approximately chance levels. This analysis shows that eyewear, skin lightness, and gender are reliable covariates of rank one mated similarity scores across this sample of face recognition systems and test population whereas age and race are not. Height was selected in 33% of the models, better than chance, suggesting that some face recognition systems were also affected by volunteer height. A nonlinear effect of age^2 was selected in 20% of the models, better than expected by chance, suggesting that similarity scores were lower for individuals that were older or younger than average (average age $= 44.79$ years).

Figure 3B plots the direction of the relationship between each demographic covariate and similarity score. This analysis shows that all models retaining gender had a positive relationship with similarity score such that rank one mated similarity scores were higher for volunteers who self-identified as male. Further, 99% of models that retained lightness and eyewear had a positive relationship between score and lightness, i.e., for volunteers reporting having no eyewear and for volunteers with lighter skin, scores were higher. On the other hand, when included, height had a negative relationship with score such that scores were lower for taller individuals in these system combinations. Similar to [4], this is likely due to larger face pitch angles associated with taller subjects. In a finding that diverges from [4], the same effect was not present for shorter individuals as evidenced by the low percentage of models where the height^2 was retained (13%, Figure 3). This perhaps indicates progress in the acquisition industry in capturing shorter subjects and subjects in wheel-chairs.

4.2 Demographic Effects Observed in Optimal Historic Average Models

The demographic effects observed for each combination of acquisition and matching systems are influenced by the interaction between the biometric samples supplied by a particular acquisition system and the way they are processed by a particular matching system. The notion of the "quality" of a biometric sample generally refers to the degree to which that sample can contribute to robust matching performance. We wanted to examine the relationship between subject demographics and the performance of matching systems more generally, independent of the quality of a specific probe sample returned by a given acquisition system.

Following [4], we modeled the relationship between volunteer demographics and the average rank one mated similarity score across all acquisition systems for that volunteer (Section 3.2.3). This effectively averages out any variation due to quality differences across samples while retaining demographic effects shared across acquisition systems. We did this for the 158 historic system combinations to generate 23 historic average models, one for each matching system (see Table 1). We then used our model selection process to find optimal historic average models.

Figure 4 plots the percentage of the 23 optimal historic average models retaining each demographic covariate and the sign of the relationship. The pattern of results generally mimicked those observed for models of individual system combinations (Section 4.1). As for individual models, eyewear, lightness, gender, and height were retained more often than expected by chance whereas race was not. Interestingly, more optimal historic average models retained a nonlinear effect of lightness^2 than expected by chance, suggesting that similarity scores were lower for individuals that were darker or lighter than average; as well as retaining a more than expected number of models with a linear effect of age, suggesting that similarity scores were lower for individuals who were older.

This meta-analysis shows that face matching systems tested in 2019, 2020, and 2021 show the same gender, eyewear, and lightness effects as the one matching system tested in 2018 [4]. Figure 5 visualizes the demographic covariates retained in the 23 optimal historic average models as well as those retained in the matching system tested in 2018 [4] as reference.

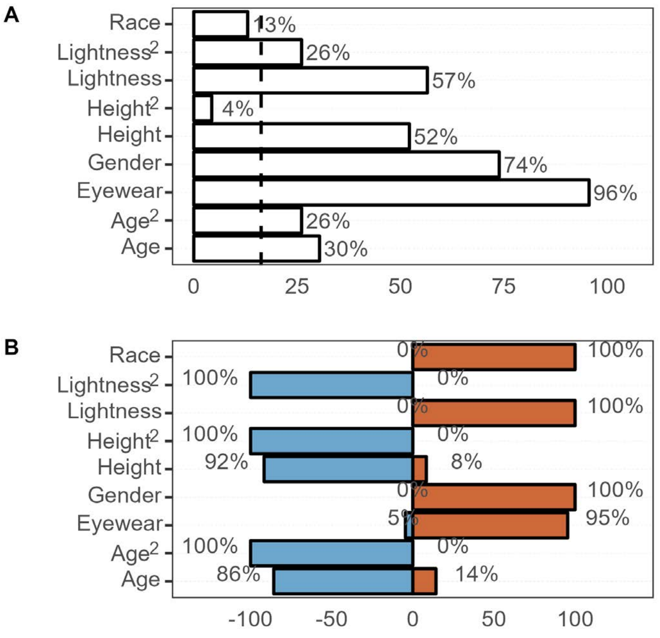

Fig. 4. **A.** Percentage of optimal historic average models ($n = 23$) retaining each listed demographic covariate. Vertical dashed line represents the percentage of optimal models expected to retain the covariate by chance (Section 3.2.2). **B.** Percentage of optimal historic average models that retain each covariate with a positive (Red) versus a negative (Blue) relationship determined as the sign of the lower bound of the parameter estimate.

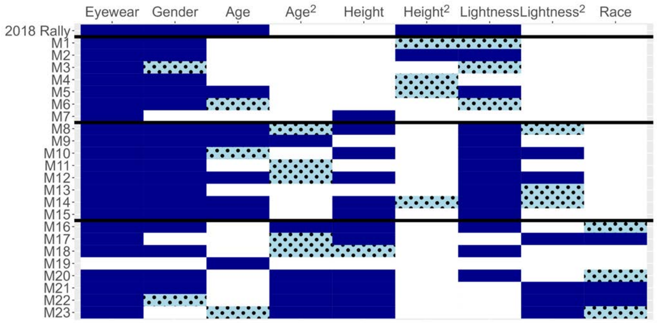

Fig. 5. Demographic covariates retained in each optimal historic average model analyzed (M1-M23). The first row depicts results for optimal historic average model examined in the 2018 Rally. Dark blue: Covariates included in the optimal average models. Light blue filled with dots: Covariates removed from the optimal model because the 95% (BC_{α}) boot strapped confidence intervals of their parameter estimate overlapped 0.

4.3 Self-reported Race and Skin Lightness

One surprising finding of [4] is that skin lightness ("reflectance" in the original work) was a better explanatory covariate of rank one mated similarity scores than self-reported racial categories (see Section 3.1). We further examined the relationship between skin lightness, race, and face recognition performance using data from the 2019, 2020, and the 2021 Rallies.

Figure 6 plots the normalized distributions of skin lightness values for volunteers self-identifying as different race categories. The figure shows that lightness was not only correlated with race such that lightness for volunteers identifying as Black or African-American was lower than lightness for volunteers identifying as White ($\mu_B = 42.0, \mu_O = 52.4, \mu_W = 56.8$ on average across Rallies), but also that the distribution of measured lightness for volunteers identifying as Black was broader ($\sigma_B = 6.40, \sigma_O = 5.18, \sigma_W = 3.16$) on average across Rallies.

Here we found that this finding continued to hold for this broader sample of biometric systems. Only 12% of optimal historic system combination models (Figure 4A) and 13% of optimal historic average models retained race (Figure 5A). On the other hand, 54% of the optimal historic models and 57% of the optimal historic average models retained lightness.

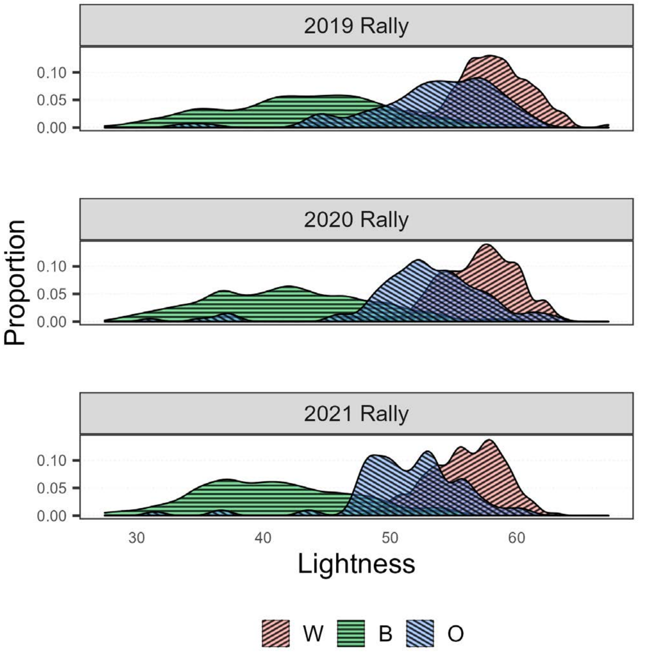

Fig. 6. Distributions of lightness values measured using a calibrated instrument. Colors and patterns correspond to self-reported race: Black or African-American (B), White (W), and Other (O). Facets correspond to distributions observed in each Rally test.

Using the present dataset, we replicated the original finding that lightness is a better explanatory covariate of face recog nition scores compared with race. For the set of 23 optimal average models, we independently confirmed whether race or lightness best explains face recognition scores. For each optimal model, we first created a baseline model (B) by removing any lightness or race covariates and computing adjusted R^2 for this model. We next computed adjusted R^2 values for models where we systematically added race (B+R), lightness (B+L), lightness^2 (B+L$_2$), or both lightness terms (B+L+L^2) using cross-validation (Section 3.2.5). We then selected the model permutation (B, B+R, B+L, or B+L+L^2) that maximized adjusted R^2.

Figure 7 shows the percentage of historic average models for which race, lightness, lightness^2, or both lightness and lightness^2 optimized the model in terms of adjusted R^2. This independent model selection approach shows that for 77% of the models the fit was optimized with the addition of a lightness covariate whereas race optimized the fit for only 9% of the models. This analysis confirmed that skin lightness, not race, best explains rank one mated similarity scores across a wide sample of face recognition systems.

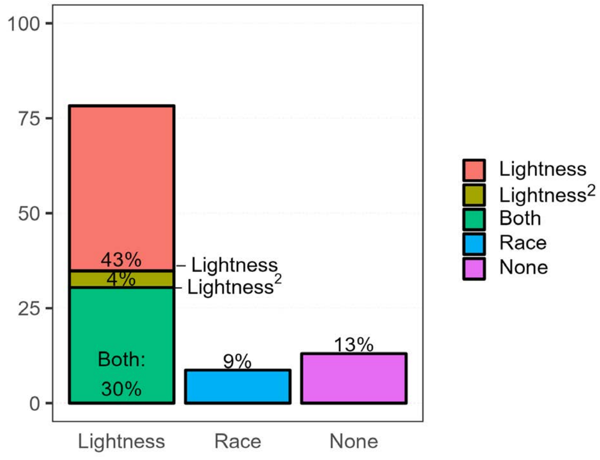

Fig. 7. Proportion of historic average models ($n = 23$) for which adjusted R^2 was maximized by Lightness (made up of Lightness in pink, Lightness^2 in olive and "Both" Lightness and Lightness^2 in green), Race in blue, or None in purple (adjusted R^2 was maximized without the inclusion of Race or Lightness).

Further, we examined whether skin lightness could explain score variation in average historic models when separately considering volunteers within each race category (e.g., if similarity scores for volunteers identifying as Black were higher if their skin was lighter as compared with Black volunteers with darker skin). We did not examine effects of lightness this way for volunteers that self-identified as "Other Race" due to a limited sample size.

For each Rally, we subsetted our data to include a cohort of volunteers from only one self-identified race group (Black or African-American or White). We then performed linear modeling as before (Section 3.2.3) to select optimal models that best explain rank one mated similarity scores observed separately for each cohort.

Figure 8 shows the percentage of optimal historic average models that included each demographic covariate (A) and the direction of its relationship with scores (B) for fits to each cohort. We observed several differences between optimal models fit to each race cohort. For the cohort of Black volunteers, results generally followed findings for modeling the full population: lightness, gender, eyewear, height, age, age^2, and lightness^2 were included in a percentage of models substantially above levels expected by chance. For the White volunteer cohort, however, there were differences relative to full population. For this cohort, only eyewear, age, and height were retained at levels above chance whereas lightness and gender were generally not retained.

There are two reasons why lightness was not retained in most models for volunteers self-identifying as White. First, the distribution of lightness values observed for this cohort is relatively narrow. Second, lightness effects may be disproportionally more present for darker skin tones below $L^* = 50$. Notwithstanding, this analysis shows that face recognition scores were influenced by volunteer lightness for 70% of the examined matching systems, even when considering only the cohort of volunteers self-identifying as

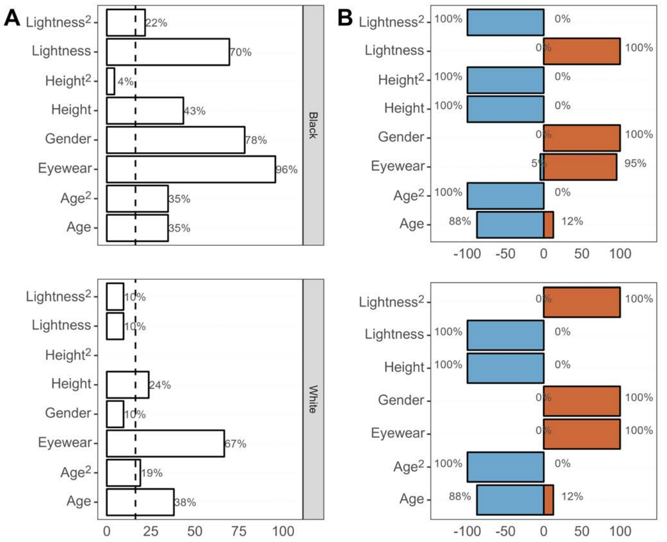

Fig. 8. Optimal historic average models fit to separate volunteer cohorts by race. **A.** Percentage of optimal historic average models retaining each listed demographic covariate. Top: models fit using the Black or African-American volunteer cohort. Bottom: models fit using the White volunteer cohort. **B.** Percentage of optimal historic average models that retain each covariate with a positive (Red) versus a negative (Blue) relationship. Top: models fit using the Black or African-American volunteer cohort. Bottom: models fit using the White volunteer cohort. Note that race was not included as a covariate since all modeled volunteers in each cohort were of the same race.

Black or African-American, which have a broadest range of measured skin lightness values (Figure 6).

4.4 Same Day vs. Different Day Gallery Effects

Earlier analysis of the 2018 Rally showed that rank one mated similarity scores can be lower for females relative to males when matching probe samples against reference samples collected on a different day, but not when matching the same probe samples against reference samples collected on the same day [4], suggesting that gender effects in face recognition may be mediated by changes in self-styling and personal appearance over time.

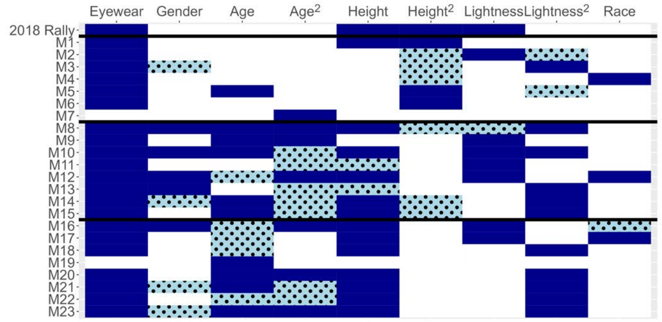

Fig. 9. Demographic covariates retained in each optimal same-day average model analyzed (M1-M23). The first row depicts results for optimal historic average model examined in the 2018 Rally. Dark blue: Covariates included in the optimal average models. Light blue filled with dots: Covariates removed from the optimal model because the 95% (BC_α) boot strapped confidence intervals of their parameter estimate overlapped 0.

To examine whether this finding holds for a broader sample of face recognition systems, we repeated our analysis of average models, but examined similarity scores against a same-day gallery (Section 3.1.1) to generate 23 optimal same-day average models. Figure 9 visualizes the demographic covariates retained in each optimal same-day models examine here as well as covariates retained in the earlier analysis of the 2018 Rally as reference. Results presented for optimal historic average models in Figure 5 shows that just five same-day models retained gender as compared with seventeen of the historic models, replicating prior results. Additionally, more optimal same-day average models retained age and lightness^2, but fewer retained the linear effect of lightness.

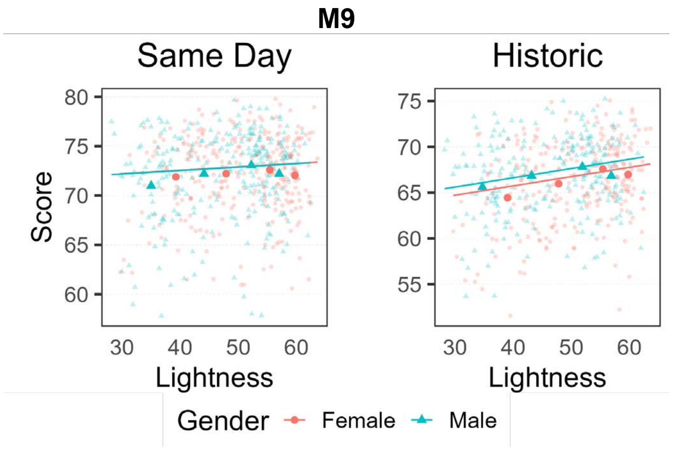

Fig. 10. Average mated similarity scores (Scores) variation with lightness and gender faceted by gallery. Lighter circles (Female) and lighter triangles (Male) show average mated similarity scores for individual volunteers. Darker circles (Female) and darker triangles (Male) denote grand average of scores across volunteers binned by lightness quartile. Lines indicate optimal fits to gender and lightness, fixing other factors constant at the average value of the subject population in each facet. Note y-axis is truncated for readability.

Figure 10 shows the variation in average rank one mated similarity score with lightness and gender for an example matching system, M9. Figure 10 shows a positive effect of lightness with both male and female lines having a positive slope for both same-day and historic models. The gender effect is present in the historic model, i.e. higher scores for males relative to females, but not in the same-day model. This study did not control for self-styling conditions (i.e. hair and make-up). However, this finding is generally supported by other studies that did [28] [29].

4.5 Demographic Effects Across Acquisitions Systems

To examine if demographic effects observed for a given matching system varied across acquisition systems (i.e., if the covariate had a greater or lesser effect for samples from some acquisition systems relative to others), we modeled historic gallery scores returned by each analyzed matching system across multiple acquisition systems using mixed effects modeling (Section 3.2.6). For this analysis, we considered every matching system because all 23 had optimal historic average models that included at least one demographic covariate (see Figure 5). We therefore generated 23 mixed effects models which started with the optimal historic average model (fixed effects) and then included random intercepts for each acquisition system and random slopes for each demographic covariate in a stepwise procedure (minimizing AIC) to identify the optimal model that explains similarity score variation across volunteers and acquisition systems. For this analysis, we only considered random effects for demographic covariates retained in each optimal historic average model.

Following this procedure, we confirmed that the new mixed effects models retained the original fixed effect coefficients from each starting optimal historic average model, indicating consistency in modeling and confirming that the average models are not unduly affected by any acquisition system specific outliers. All 23 models were improved as assessed by the AIC values with the inclusion of a random intercept.

Table 3 tallies the number of models for which the fit was improved with the addition of specific mixed effect terms. Six of the models were not improved by the addition of mixed effect terms (mixed effect: none) meaning no difference in demographic covariate effects across acquisition systems. Eight of the 23 models were improved by adding a mixed effect of lightness, indicating that the degree to which lightness influenced similarity scores varied depending on acquisition system for those eight matching systems. Interestingly, six of the 23 models included a random effect of eyewear, suggesting that effects of eyewear may vary depending on acquisition system used (e.g. if an acquisition system asks subjects to remove their glasses). Rally vendors could install signage requesting volunteers to remove glasses while interacting with their acquisition system. Any signage was unique to each system and was not controlled during the test.

Mixed Effect	Count
None	6 of 23
Lightness	8 of 23
Gender	7 of 23
Eyewear	6 of 23
Height	3 of 23

TABLE 3
Count of the optimal historic mixed-effects models ($n = 23$) that include specific demographic mixed effects. Note that each model could contain more than one demographic effect.

5 Discussion

This study examined whether demographic factors reliably explained face recognition performance across a large sample of commercial face recognition systems (21 acquisition systems, 23 matching algorithms, 158 acquisition-matching system combinations) tested across a three year period from 2019 to 2021 as part of the DHS S&T Biometric Technology Rallies. The analyzed sample of the 158 acquisition-matching system combinations indicated some demographic factors influence the accuracy of the majority of system combinations. For example, in 81 of 158 system combinations (51%), optimal historic models relating rank one mated score to demographics retained a skin lightness term. In 99% of these models, scores were higher for those with lighter skin (Figure 3). This effect also persisted when created a matching system specific model by averaging across acquisition system, in 13 of 23 optimal models (57%, Figure 4). Gender also impacted the majority of system combination models (92 of 158 or 58%) and average system models (17 of 23; 74%). In 100% of these models, rank one mated scores were higher for males.

However, the relationship between gender, skin lightness, and biometric performance was notably different in one regard. Unlike skin tone effects, gender effects were notably reduced or eliminated when matching between two images gathered on the same day (Section 4.4). This suggests self-styling elections may be contributing to these gender effects, an outcome which is consistent with prior work [29]. Hence, reducing the relationship between self-styling decisions and mated match score may be a viable avenue to reduce gender effects in commercial face recognition.

By far the most ubiquitous demographic effect detected in this study is that of eyewear, with higher rank one mated similarity scores for individuals without eyewear. Eyewear effects were detected in 67% of the 158 tested system combinations and from all but one of the 23 tested matching algorithms when scores were averages across acquisition system (Figures 3 and 4, respectively). This effect is somewhat expected, as glasses create a distractor in an image, which could lower mated similarity scores on average due to occlusions and distortions of face features around the eyes. For this reason, subject enrollment into the Rally requires subjects remove their eyeglasses, as do many other applications, such as passports.

Next, this study builds on work showing that lightness is a better predictor than race. While skin lightness is only one of many phenotypic measures that can be quantified from the human face, it appears to be a salient one as it relates to automated face recognition performance. In our matching system models, 77% of the models kept an effect of lightness and only 9% kept an effect of race (Figure 7). Similar effects were shown in [4]. However, here we also show that individuals identifying as Black or African-American have a larger variation in skin lightness as compared to individuals that self-identify as White (Figure 6). Critically, skin lightness only influenced rank one mated scores at levels above chance when considering volunteers identifying as Black or African-American but not those who identified as White (Figure 8). This suggests that mated scores are reduced specifically for those individuals with skin lightness below a certain value and not necessarily based on race categories. This is further evidence that race labels are problematic when discussing the causality of face recognition performance variation. Improving imaging for darker skin tones, so that the quality of images is comparable across the full color gamut of human skin tone, is a viable avenue to reduce skin tone effects in commercial face recognition

Finally, we close with two notes. First, at a high level, this study largely replicates our prior study in 2018, which tested one matching algorithm averaged across eleven acquisition systems using a historic gallery [4]. However, unlike the effects of eyewear, gender, and skin lightness, the effect of age was not found reliably in the present analysis. This difference may be due to the different biometric systems tested in this report. Indeed, recent work suggests that modern face recognition algorithms have become better able to maintain matching performance for individuals with varying ages [30].

Second, this study examined demographic differentials in rank one mated similarity scores. Whether any of the observed score differentials manifest in *actual* biometric false non-match error rates would depend on the thresholds used in deployed systems. The presence of score differentials and their direction suggests that error rate differentials may be observed in some applications of the technology but not others [31]. It is therefore important to test for demographic differentials in biometric error rates in specific applications of face recognition technology. We hope these findings can assist the developers and procurers of face recognition technologies improving face recognition system performance across demographic groups.

6 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 data were acquired using the IRB protocol "Development and Evaluation of Enhanced Screening" number 120180237, approved by New England IRB.

The paper authors acknowledge the following author contributions: Cynthia M. Cook, John J. Howard and Yevgeniy B. Sirotin conceived the work, generated biometric templates and comparisons, developed and performed analyses, and wrote the paper; Jerry Tipton and Arun Vemury conceived the work and edited the paper.

The authors thank Juan Kehoe for performing initial data aggregation and analyses and Rebecca Rubin for her editorial input on the final draft. The authors also thank Lars Ericson, Patrick Grother, Alicia Locker, and James Wayman for their expert review of this manuscript prior to its publication. This research was also supported by the multi-disciplinary staff of the Identity and Data Sciences Laboratory at the Maryland Test Facility and the Biometric and Identity Technology Center who planned and executed the Biometric Technology Rallies as well as the biometric technology vendors that provided systems for evaluation.

References

[1] "DHS/CBP/PIA 056-2018 privacy impact assessment for the traveler verification service," PIA, 2018.

[2] "DHS/CBP/PIA 046(b)-2020 privacy impact assessment for the travel document checker automation using facial verification," PIA, 2020.

[3] J. J. Howard, A. J. Blanchard, Y. B. Sirotin, J. A. Hasselgren, and A. R. Vemury, "An investigation of high-throughput biometric systems: Results of the 2018 department of homeland security biometric technology rally," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2018, pp. 1–7.

[4] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems," *IEEE T-BIOM*, vol. 1, no. 1, p. null, 02 2018, link.

[5] J. A. Hasselgren, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "A Scenario Evaluation of High-Throughput Face Biometric Systems: Select Results from the 2019 Department of Homeland Security Biometric Technology Rally," in *The DHS S&T Technical Paper Series*. The U.S. Department of Homeland Security, 2020, pp. 1–13.

[6] "ISO/IEC 19795-2:2007 Information technology–biometric performance testing and reporting–part 2: Testing methodologies for technology and scenario evaluations," Standard, 2007.

[7] S. Gong, X. Liu, and A. K. Jain, "Jointly de-biasing face recognition and demographic attribute estimation," in *European conference on computer vision*. Springer, 2020, pp. 330–347.

[8] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole, "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?" *IEEE transactions on biometrics, behavior, and identity science*, vol. 3, no. 1, pp. 101–111, 2020.

[9] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic bias in biometrics: A survey on an emerging challenge," *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89–103, 2020.

[10] P. Grother, M. Ngan, and K. Hanaoka, *Face Recognition Vendor Test (FRVT): Part 3: Demographic Effects*. National Institute of Standards and Technology Gaithersburg, MD, 2019.

[11] P. Grother, *Face Recognition Vendor Test (FRVT) Part 8: Summarizing Demographic Differentials*. National Institute of Standards and Technology Gaithersburg, MD, 2022.

[12] K. Bowyer and M. King, "Why face recognition accuracy varies due to race," *Biometric Technology Today*, vol. 2019, no. 8, pp. 8–11, 2019.

[13] S. Barocas, A. Guo, E. Kamar, J. Krones, M. R. Morris, J. W. Vaughan, W. D. Wadsworth, and H. Wallach, "Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 368–378.

[14] J. J. Howard, Y. B. Sirotin, and A. R. Vemury, "The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance." 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2019.

[15] J. J. Howard, E. J. Laird, and Y. B. Sirotin, "Disparate impact in facial recognition stems from the broad homogeneity effect: A case study and method to resolve," *2022 International Conference on Pattern Recognition*, 2022.

[16] U. Census, "Race and ethnicity," United States Census Bureau, Tech. Rep., Jan 2017, last accessed on 06/07/18.

[17] "DSM III - Skin Colormeter," 2021, link.

[18] P. Clarys, K. Alewaeters, R. Lambrecht, and A. Barel, "Skin color measurements: comparison between three instruments: the chromameter®, the dermaspectrometer® and the mexameter®," *Skin research and technology*, vol. 6, no. 4, pp. 230–238, 2000.

[19] B. Diffey, R. Oliver, and P. Farr, "A portable instrument for quantifying erythema induced by ultraviolet radiation," *British Journal of Dermatology*, vol. 111, no. 6, pp. 663–672, 1984.

[20] J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Quantifying the Extent to Which Race and Gender Features Determine Identity in Commercial Face Recognition Algorithms," in *The DHS S&T Technical Paper Series*. The U.S. Department of Homeland Security, 2021, pp. 1–14.

[21] J. J. Howard, Y. B. Sirotin, and J. L. Tipton, "Reliability and validity of image-based and self-reported skin phenotype metrics," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 4, pp. 550–560, 2021.

[22] A. R. Vemury, Y. B. Sirotin, C. M. Cook, J. J. Howard, and J. L. Tipton, "Skin reflectance image correction in biometric image capture," 2021, US Patent App. 17/185,588.

[23] Y. B. Sirotin, A. R. Vemury, C. M. Cook, J. J. Howard, and J. L. Tipton, "Detection of skin reflectance in biometric image capture," 2021, US Patent App. 17/185,487.

[24] P. Grother, M. Ngan, and K. Hanaoka, *Face Recognition Vendor Test (FRVT) Part 2: Identification*. National Institute of Standards and Technology Gaithersburg, MD, 2019.

[25] V. F. Flack and P. C. Chang, "Frequency of selecting noise variables in subset regression analysis: A simulation study," *The American Statistician*, vol. 41, no. 1, pp. 84–86, 1987.

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

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

[28] A. Bhatta, V. Albiero, K. W. Bowyer, and M. C. King, "The gender gap in face recognition accuracy is a hairy problem," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 303–312.

[29] V. Albiero, K. Zhang, M. C. King, and K. W. Bowyer, "Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 127–137, 2021.

[30] P. Grother, M. Ngan, K. Hanaoka, J. C. Yang, and A. Hom, "Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification," National Institute of Standards and Technology, Tech. Rep., Sep 2022, last accessed on 10/12/22.

[31] J. J. Howard, E. J. Laird, Y. B. Sirotin, R. E. Rubin, J. L. Tipton, and A. R. Vemury, "Evaluating proposed fairness models for face recognition algorithms," *2022 International Conference on Pattern Recognition*, 2022.