Longitudinal Study of Fingerprint Recognition

Soweon Yoon and Anil K. Jain
Department of Computer Science and Engineering,
Michigan State University, East Lansing, MI 48824

Abstract

Human identification by fingerprints is based on the fundamental premise that (i) a ridge pattern of a finger is distinct from the ridge patterns from any other fingers (uniqueness) and (ii) a fingerprint pattern does not change over time (persistence). While the uniqueness of fingerprints has been investigated by developing statistical models to estimate the probability of error in comparing two random samples of fingerprints, the persistence of fingerprints has remained a general belief based on only a few case studies. In this study, the fingerprint match (similarity) scores of genuine and impostor pairs are analyzed by multilevel statistical model based on an operational fingerprint database containing longitudinal records of 15,597 subjects with a maximum time span of 12 years. The longitudinal analysis of this dataset shows that (i) the genuine match scores tend to significantly decrease when time interval between two fingerprints being compared and subject's age increase and fingerprint image quality decreases, while the decrease in impostor match scores is negligible, (ii) the fingerprint recognition accuracy, nevertheless, remains stable as the time interval between fingerprints in comparison increases up to 12 years, the maximum time span in the database, and (iii) the fingerprint image quality, rather than time interval and subject's age, is the best covariate to explain the variation in genuine match scores. Furthermore, the analysis of the fingerprint match scores obtained by fusing all ten fingers of each subject in this database reinforces that the fingerprint recognition accuracy does not change over time.

1 Introduction

Friction ridge skin on fingers and palms has been purportedly known to be a physical characteristic of an individual that does not change over time (i.e., persistence or permanence of friction ridge pattern) and can be used as a person's "seal" or "signature" (i.e., uniqueness or individuality of ridge pattern). Starting with the first known case where the latent fingerprints found at a crime scene in Argentina in 1893 were officially accepted as evidence to convict a suspect [1], friction ridge analysis has become one of the most crucial methods in crime scene investigations worldwide. The decision made in *Frye v. United States* in 1923 [2] is widely cited as the basis for the admissibility of forensic evidence, including friction ridge pattern; *Frye* standard states that a scientific principle or discovery which has gained a *general acceptance* in the relevant field is admissible in the courts.

In *Daubert v. Merrell Dow Pharmaceuticals, Inc.* in 1993 [3], however, the general acceptance test of *Frye* was superseded by the Federal Rules of Evidence. The *Daubert* ruling established a guideline for admitting forensic evidence which consists of the following factors: (i) empirical testing, (ii) peer review and publication, (iii) known or potential error rate, (iv) standards controlling the operation, and (v) the *Frye* standard of general acceptance. The *Daubert* standard provoked challenges to admissibility of friction ridge evidence in the courts. Although all of about 40 such challenges resulted in a decision that friction ridge analysis is acceptable as forensic evidence, the *Daubert* case highlighted a lack of scientific basis of persistence and uniqueness and standards that can be universally referred to in friction ridge analysis.

Along with the development of standards and guidelines for friction ridge analysis [4] and retraining of latent examiners [5] as a result of the *Daubert* ruling, a body of research to demonstrate uniqueness and persistence of friction ridge patterns has emerged. While the uniqueness of fingerprints has been studied by (i) estimating the probability of a random correspondence (i.e., two different fingerprints selected at random

will be sufficiently similar to be claimed as genuine mates) [6, 7, 8] or (ii) measuring the evidential value (see Supporting Information S1) of latent fingerprint comparisons [9], the persistence of fingerprints has been generally accepted based on anecdotal evidence, including case studies conducted by Herschel [28] and Galton [11] (see Supporting Information S3), and the anatomical structure of friction ridge skin—the ridge pattern formed in the inner (dermal) layer during gestation remains unchanged with the protection of the outer (epidermal) layer [12].

The persistence of fingerprints typically refers to the invariance of friction ridge pattern itself. However, the pertinent question of interest is whether the fingerprint recognition methodology (see Supporting Information S2) maintains high recognition accuracy as the time interval between two fingerprints being compared increases. The 2009 National Research Council report on "Strengthening Forensic Science in the United States: A Path Forward" [13] pointed out that *"Uniqueness and persistence are necessary conditions for friction ridge identification to be feasible, but those conditions do not imply that anyone can reliably discern whether or not two friction ridge impressions were made by the same person."* Fingerprint recognition exhibits two types of comparison errors: (i) false rejection: two impressions of the same finger (a genuine fingerprint pair) are declared as a non-match due to large *intra-finger* variability, and (ii) false acceptance: impressions from two distinct fingers (an impostor fingerprint pair) are declared as a match due to large *inter-finger* similarity. The intra-finger variability is observed due to changes in intrinsic skin condition (e.g., finger skin dryness, cuts, and abrasions) and extrinsic acquisition process (e.g., finger pressure and placement), and sensing technology (known as the *interoperability* problem [14]). The inter-finger similarity is observed when the partial fingerprint impressions from two distinct fingers coincide.

In the biometric recognition literature, a phenomenon called *template aging* has been reported, which refers to an increase in the error rate in biometrics recognition with respect to the time gap between the query and the template (or reference) [15]. A study comparing groups of fingerprint pairs with respect to time gap reported that the fingerprint comparisons with less than 5-year time gap show lower error rate than comparisons with a larger time gap [16]. Similar studies on face and iris recognition observed a decrease in matching accuracy as the time interval between two acquisitions of a person's face or iris increases. However, the conclusions drawn in these studies cannot be trusted since their statistical analysis was not suitable for the datasets used in the studies (see Supporting Information S3).

In order to determine the trend of fingerprint recognition accuracy with respect to time interval between fingerprint acquisitions, we need to (i) collect longitudinal data (see Supporting Information S1) consisting of multiple acquisitions of fingerprints from a sampled population over a reasonably long period of time, and (ii) conduct an appropriate statistical analysis, considering the characteristics of the longitudinal data. If the longitudinal dataset is *balanced* and *time structured* (see Supporting Information S1), cross-sectional analysis can be applied by grouping the longitudinal data according to cohort (for example, short-term and long-term fingerprint comparison groups) under the assumption of compound symmetry (see Supporting Information S1). In reality, however, it is not feasible to collect longitudinal fingerprint data by following an identical measurement schedule over a large number of subjects in the sample satisfying the compound symmetry. To handle the unbalanced and/or time-unstructured longitudinal data, several statistical models have been developed, including multilevel statistical models [17, 18].

In this study, we obtained a longitudinal database of fingerprints collected from 15,597 subjects booked by the Michigan State Police multiple times (at least five different time points) over at least a 5-year time span. A multilevel statistical model is used to analyze this longitudinal dataset which is unbalanced and time unstructured. This paper addresses the following specific issues pertaining to the longitudinal study of fingerprint recognition:

- Trend of fingerprint match scores of genuine and impostor pairs with respect to various covariates, including time interval between fingerprints in comparison, subject's demographic factors (age, gender, and race), and fingerprint image quality
- Assessment and comparison of the multilevel models with various combinations of the covariates
- Correlations and interactions among covariates

June 2001 July 2002 April 2003 September 2007 March 2008 October 2008

Figure 1: Six different impressions of the right index finger of a subject in the longitudinal fingerprint database used in this study

- Temporal trend of fingerprint recognition accuracy in terms of probabilities of true acceptance and false acceptance
- Trend of fingerprint match scores and recognition accuracy when a subject's all ten fingers are used for recognition, a prevailing practice in law enforcement and forensics.

2 Longitudinal Fingerprint Database

A longitudinal database of fingerprints was collected from the records of repeat offenders booked by the Michigan State Police. Fig. 1 shows an example of six fingerprint impressions of the right index finger of a subject in the database acquired between June 2001 and October 2008. A total of 15,597 subjects were randomly selected who had at least 5 fingerprint acquisitions from all ten fingers on a formatted fingerprint card (called tenprint card) over a minimum of 5-year time span. The tenprint impressions of each subject are ordered according to the time sequence; a set of tenprints of subject i ($i = 1, \ldots, N$; N is the total number of subjects in the database) is labeled as follows: $\mathcal{F}_i = \{F_{i,1}, \cdots, F_{i,n_i}\}$, such that $T_{i,1} < \ldots < T_{i,n_i}$, where $T_{i,j}$ is the time stamp of the j-th tenprint impression of subject i, and n_i denotes the number of tenprints of subject i.

A summary of the database is as follows:

- Each of the 15,597 subjects has at least 5 tenprint cards, providing 122,685 tenprint cards in total. The average number of tenprints per subject in the database is 8, with the maximum of 26 cards for one of the subjects.
- The tenprint impressions of a subject have a minimum of 5-year time span (the time difference between the first and the last fingerprint acquisitions of a subject); that is, $\triangle T_{i,1n_i} \geq 5$ years for $i = 1, \ldots, N$. The average time span is 9 years, and the maximum time span in the database is 12 years.
- Any two consecutive tenprint impressions of a subject are obtained with at least a 2-month time gap; $(T_{i,j+1} - T_{i,j}) \geq 2$ months.
- Along with tenprint images, the following demographic information is also available for each subject:
  - Gender: Male or female
  - Race: White/Hispanic, Black, American Indian/Eskimo, or Asian/Pacific Islander
  - Age at the time of tenprint acquisition: The youngest subject's age at the time of the first impression is 8 years; the oldest subject's age at the time of the last impression is 78 years.

Two commercial off-the-shelf (COTS) fingerprint matchers (denoted as COTS-1 and COTS-2) are used to compute match scores. For subject i with n_i fingerprint impressions, we conduct all pairwise comparisons[1];

[1]The pairwise comparisons of the fingerprint records of a subject result in correlations among the genuine match scores of the subject.

Table 1: Multilevel models with different combinations of covariates

Model	Level-1 Model	Level-2 Model
Model A	$y_{ijk} = \varphi_{0i} + \varepsilon_{ijk}$	$\varphi_{0i} = \beta_{00} + b_{0i}$
Model B_T	$y_{ijk} = \varphi_{0i} + \varphi_{1i}\triangle T_{ijk} + \varepsilon_{ijk}$	$\varphi_{0i} = \beta_{00} + b_{0i}$, $\varphi_{1i} = \beta_{10} + b_{1i}$
Model B_A	$y_{ijk} = \varphi_{0i} + \varphi_{1i}AGE_{ijk} + \varepsilon_{ijk}$	$\varphi_{0i} = \beta_{00} + b_{0i}$, $\varphi_{1i} = \beta_{10} + b_{1i}$
Model B_Q	$y_{ijk} = \varphi_{0i} + \varphi_{1i}Q_{ijk} + \varepsilon_{ijk}$	$\varphi_{0i} = \beta_{00} + b_{0i}$, $\varphi_{1i} = \beta_{10} + b_{1i}$
Model C_G	$y_{ijk} = \varphi_{0i} + \varphi_{1i}\triangle T_{ijk} + \varepsilon_{ijk}$	$\varphi_{0i} = \beta_{00} + \beta_{01}bM_i + b_{0i}$, $\varphi_{1i} = \beta_{10} + \beta_{11}bM_i + b_{1i}$
Model C_R	$y_{ijk} = \varphi_{0i} + \varphi_{1i}\triangle T_{ijk} + \varepsilon_{ijk}$	$\varphi_{0i} = \beta_{00} + \beta_{01}bW_i + b_{0i}$, $\varphi_{1i} = \beta_{10} + \beta_{11}bW_i + b_{1i}$
Model D	$y_{ijk} = \varphi_{0i} + \varphi_{1i}\triangle T_{ijk} + \varphi_{2i}AGE_{ijk} + \varphi_{3i}Q_{ijk} + \varepsilon_{ijk}$	$\varphi_{0i} = \beta_{00} + b_{0i}$, $\varphi_{1i} = \beta_{10} + b_{1i}$, $\varphi_{2i} = \beta_{20} + b_{2i}$, $\varphi_{3i} = \beta_{30} + b_{3i}$
Model E	$y_{ijk} = \varphi_{0i} + \varphi_{1i}\triangle T_{ijk} + \varphi_{2i}AGE_{ijk} + \varphi_{3i}Q_{ijk} + \varphi_{4i}\triangle T_{ijk}Q_{ijk} + \varphi_{5i}AGE_{ijk}Q_{ijk} + \varepsilon_{ijk}$	$\varphi_{0i} = \beta_{00} + b_{0i}$, $\varphi_{1i} = \beta_{10} + b_{1i}$, $\varphi_{2i} = \beta_{20} + b_{2i}$, $\varphi_{3i} = \beta_{30} + b_{3i}$, $\varphi_{4i} = \beta_{40}$, $\varphi_{5i} = \beta_{50}$

that is, ${}_{n_i}C_2$ genuine match scores are generated from each matcher. This is because law enforcement agencies often store all the tenprint records for every booked subject and compare a query fingerprint to all the records in the database. To obtain impostor match scores of a subject, 10 tenprint cards from 10 different subjects are randomly selected and compared to each of the tenprint cards of the subject. The analysis utilizes 481,181 genuine match scores and 1,226,850 impostor match scores obtained by each of the COTS matchers.

3 Multilevel Statistical Model

With multilevel modeling for longitudinal data analysis [17, 18], this study aims at analyzing the following observed responses (y_{ijk}):

- Case I: A single finger[2] is used for recognition
  - Normalized genuine match score obtained by:

$$\tilde{s}_{i,jk} = \frac{s_{i,jk} - \mu}{\sigma}, \tag{1}$$

    where $s_{i,jk}$ is the genuine match score between the j-th and k-th fingerprint impressions of the right index finger of subject i, and μ and σ are the mean and standard deviation of $\{s_{i,jk}\}$, respectively
  - Binary identification decision made on a genuine pair with match score of $s_{i,jk}$ by applying a decision threshold (Th) corresponding to a false acceptance rate of 0.01%:

$$s^*_{i,jk} = \begin{cases} 1, & \text{if } s_{i,jk} > Th \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

  - Normalized impostor match score ($\tilde{s}_{ij,k}$) between the k-th fingerprint impression of the right index finger of subject i and the right index finger impression in the first tenprint of subject j, for $i \neq j$ and $k = 1, \ldots, n_i$

[2]We carry out the analysis on the right index finger of the subjects that is typically chosen as the primary finger in the single-finger based recognition systems.

  - Binary identification decision ($s^*_{ij,k}$) made on an impostor pair with match score of $s_{ij,k}$ by applying the decision threshold Th

- Case II: All ten fingers are used for recognition

  - Normalized genuine fusion score obtained by a sum rule as follows:

$$S_{i,jk} = \sum_{m=1}^{10} s^{(m)}_{i,jk}, \tag{3}$$

    where $s^{(m)}_{i,jk}$ is the genuine match score between the impressions from finger m in the j-th and k-th tenprint cards of subject i
  - Binary identification decision made on a pair of genuine tenprint cards with fusion score of $S_{i,jk}$ by applying a decision threshold ($Th^\dagger$) corresponding to a false acceptance rate of 0.01%:

$$S^*_{i,jk} = \begin{cases} 1, & \text{if } S_{i,jk} > Th^\dagger \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

  - Normalized impostor fusion score ($\tilde{S}_{ij,k}$) between the k-th tenprint of subject i and the first tenprint of subject j
  - Binary identification decision ($S^*_{ij,k}$) made on an impostor pair of tenprints with fusion match score of $S_{ij,k}$ by applying the decision threshold $Th^\dagger$.

The covariates (x_{ijk})[3] investigated in the study are:

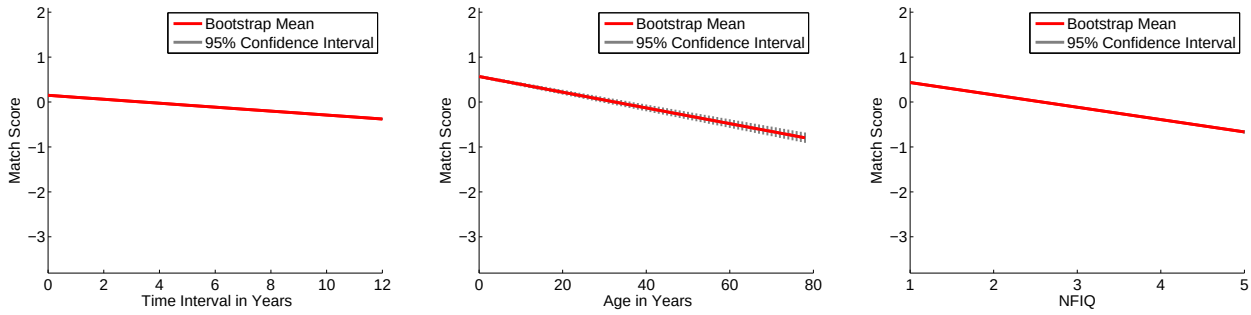
- $\triangle T_{i,jk}$ for genuine fingerprint comparisons: Time interval between the j-th and k-th fingerprint impressions of subject i; $\triangle T_{i,jk} = T_{i,k} - T_{i,j}$, for $k > j$
- $\triangle T_{ij,k}$ for impostor fingerprint comparisons: Time elapsed after the first tenprint of subject i is obtained; $\triangle T_{ij,k} = T_{i,k} - T_{i,1}$, $k = 1, \ldots, n_i$
- $AGE_{i,jk}$ for genuine fingerprint comparisons: Age of subject i when the latter of the j-the and k-th tenprint impressions was made, where $T_{i,j} < T_{i,k}$
- $AGE_{ij,k}$ for impostor fingerprint comparisons: Age of subject i when the k-th tenprint impression was made; the age of impostor subject j is not considered
- $Q_{i,jk}$: The value corresponding to the lower of the qualities of the j-th and k-th fingerprint impressions of subject i. In this study, NIST Fingerprint Image Quality (NFIQ) measure [19] is used, which assigns one of the five discrete values ranging from 1 (the highest quality) to 5 (the lowest quality), to define fingerprint image quality. According to the definition of NFIQ, $Q_{i,jk} = \max(Q_{i,j}, Q_{i,k})$, where $Q_{i,j}$ is the NFIQ value of fingerprint impression j of subject i
- bM_i: A binary indicator of gender of subject i; 1 for male, and 0 for female
- bW_i: A binary indicator of race of subject i; 1 for whites, and otherwise 0.

As a fingerprint comparison essentially involves two fingerprint impressions to generate a single match score, a simple linear 2-level model with a single covariate for continuous match scores can be represented by:

Level-1 Model (Intra-subject variability):

$$y_{ijk} = \varphi_{0i} + \varphi_{1i} x_{ijk} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2_\varepsilon), \tag{5}$$

[3] $Q_{i,jk}$, bM_i, and bW_i are used only for genuine fingerprint comparisons.



(a) Model B_T (b) Model B_A (c) Model B_Q

Figure 2: Population-mean trends of genuine match scores obtained by COTS-1 matcher along with 95% confidence intervals with respect to (a) $\triangle T_{i,jk}$, (b) $AGE_{i,jk}$, and (c) $Q_{i,jk}$, when a single finger is used for recognition. The confidence intervals for models B_T and B_Q are too tight along the mean to be visible.

Level-2 Model (Inter-subject variability):

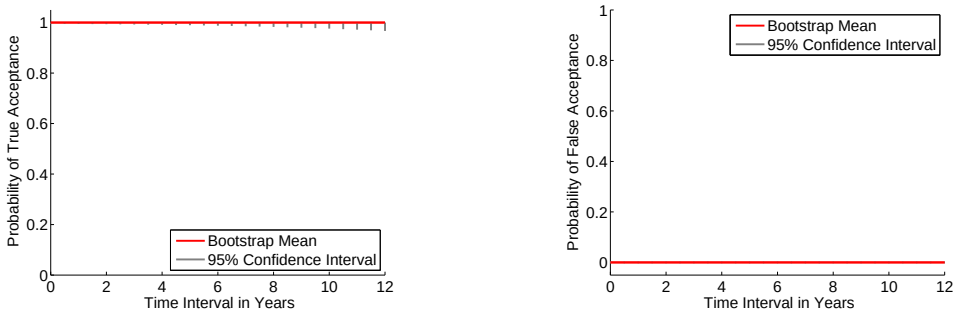
$$\varphi_{0i} = \beta_{00} + b_{0i}, \quad \varphi_{1i} = \beta_{10} + b_{1i},$$
$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right). \tag{6}$$

The level-1 model in Eq. (5) is regressed to the repeated measurements taken from each subject, and accounts for the intra-subject variability. The variables and parameters in the level-1 model are defined as follows: y_{ijk} is the subject i's observed response of match score when two fingerprints (j and k) are compared, x_{ijk} is the explanatory variable, φ_{0i} and φ_{1i} are the true parameters representing the intercept and slope of the linear model for subject i, and ε_{ijk} is the error in the observed response y_{ijk} from the model fit. The error is assumed to be normally distributed with a zero mean and a variance of σ_ε^2.

In the level-2 model (Eq. (6)) where the population-averaged tendency and deviations of subjects from the mean trend are modeled to account for the inter-subject variability, the true parameters for subject i (φ_{0i} and φ_{1i}) can be modeled by a mixture of fixed and random effects: fixed-effects parameters β_{00} and β_{10} represent the grand means of intercept and slope across all N subjects in the data, and random-effects parameters b_{0i} and b_{1i} represent the deviations of subject i's intercept and slope from β_{00} and β_{10}. The random effects are assumed to follow a Gaussian distribution.

In order to determine whether two fingerprint impressions are from the same finger, a binary decision for a fingerprint pair is made by applying a predetermined decision threshold to the match score. If the match score of a fingerprint pair is greater than the threshold, the two fingerprints are determined to be a genuine match; otherwise, they are determined to be an impostor match. If a fingerprint pair is determined to be a genuine pair and they are indeed from the same finger, the binary decision is a *true acceptance*. If a genuine-match decision is made on a fingerprint pair which are in fact from two different fingers, the decision is a *false acceptance*. In multilevel model, a binary response is viewed as a Bernoulli trial with the probability of true (or false) acceptance π_{ijk}, and the expected π_{ijk} is modeled after being transformed by a logit link function.

$$\begin{aligned} \text{Level-1 Model: } & g(\pi_{ijk}) = \varphi_{0i} + \varphi_{1i} x_{ijk} + \varepsilon_{ijk}, \\ & y^*_{ijk} \sim Bin(1, \pi_{ijk}), \\ \text{Level-2 Model: } & \varphi_{0i} = \beta_{00} + b_{0i}, \\ & \varphi_{1i} = \beta_{10} + b_{1i}, \end{aligned} \tag{7}$$



(a) (b)

Figure 3: Population-mean trend of fingerprint matching accuracy along with 95% confidence interval with respect to $\triangle T_{i,jk}$. (a) Probability of true acceptance and (b) probability of false acceptance with respect to $\triangle T_{i,jk}$. Match scores are obtained by COTS-1 matcher when a single (right index) finger is used for recognition. The confidence interval in (b) is too tight along the mean to be visible.

where $g(\cdot)$ is a logit link function; $g(\pi_{ijk}) = \log\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right)$. The 2-level linear models investigated in this study are listed in Table 1.

The maximum likelihood (ML) and generalized least-squares (GLS) estimations are widely used to estimate parameters in the multilevel model [18]. Under the assumption that the residuals are normally distributed, the ML estimates of the parameters are typically obtained by iterative GLS [17].

4 Results

4.1 Population-mean Trend of Genuine Fingerprint Match Scores

Given that the normality assumptions of the residuals and random effects in the multilevel model fit to the data are violated (see Supporting Information S5), the parameters in the multilevel models are estimated by a fully nonparametric bootstrap [17]. We generate 1,000 bootstrap samples, where each bootstrap sample is obtained by a cluster bootstrap—N subjects with replacement are resampled at level 1 and all the level-2 data belonging to those subjects are included in the sample—to preserve the hierarchy in the longitudinal data. The mean of the parameter estimates of the bootstrap samples and the percentile confidence intervals are reported in Tables S3 and S4.

The population-mean trends of Models B_T, B_A, and B_Q based on the fixed-effects parameter estimates (β_{00} and β_{10}) show that the genuine match scores tend to decrease when $\triangle T_{i,jk}$, $AGE_{i,jk}$, and $Q_{i,jk}$ increase (see Figs. 2 and S5). The null hypothesis—$\beta_{10} = 0$ in Models B_T, B_A, and B_Q (i.e., the slope of the linear model is zero)—is rejected for all three models at the significance level of 0.05 since the 95% confidence interval for β_{10} does not contain zero.

Models D and E incorporate all three covariates ($\triangle T_{i,jk}$, $AGE_{i,jk}$, and $Q_{i,jk}$) into the model; Model E includes interaction terms ($\triangle T_{i,jk}Q_{i,jk}$ and $AGE_{i,jk}Q_{i,jk}$) while Model D does not have any interaction terms. The covariance matrix in Model D shows that the correlations between (i) $\triangle T_{i,jk}$ and $Q_{i,jk}$ and (ii) $AGE_{i,jk}$ and $Q_{i,jk}$ are very small (see Supporting Information S6). Also, the population-mean trends of Models D and E and their 95% confidence intervals indicate that the impact of the interactions between (i) $\triangle T_{i,jk}$ and $Q_{i,jk}$ and (ii) $AGE_{i,jk}$ and $Q_{i,jk}$ on genuine match scores is not significant (see Supporting Information S6).

4.2 Outlier Subjects in Model B_T

The random effects (b_{ri}, $r = 0, 1, 2, 3$) at level 2 in the multilevel model represent the deviation of subject i from the population-mean trend (β_{ri}). The parameter estimates of $(\varphi_{0i}, \varphi_{1i})$ for each subject in Model B_T are shown in Fig. S9 in addition to the population-mean trend (β_{00}, β_{11}). The parameter estimates associated with several outlier subjects whose trend for genuine match scores markedly deviate from the population-mean trend are also indicated. Figs. S10–S14 show the individual trends of the outlier subjects and their fingerprint impressions.

- Outlier case 1 (Fig. S10): The estimated intercept of this subject is very small. The subject consistently gives low genuine match scores since his fingerprints are severely scarred. This subject can be called a "goat" in the Doddington's biometric zoo nomenclature [20] which refers to subjects that are susceptible to false rejections.
- Outlier case 2 (Fig. S11): The intercept of the fitted model for this subject is rather large while the slope is negative. This subject consistently gives high genuine match scores because his fingerprint impressions are of good quality. This subject can be viewed as a "sheep" in the Doddington's biometric zoo who is easy to successfully verify.
- Outlier case 3 (Fig. S12): This subject shows a very sharp decrease in genuine match scores as a function of time interval. In Fig. S12(a), the genuine match scores involving the first fingerprint impression are very low. This fingerprint impression is indeed an impostor fingerprint (see Fig. S12(b)) since it is of tented arch type while the actual pattern of this finger is a right loop. This shows that the operational fingerprint data can be mislabeled.
- Outlier case 4 (Fig. S13): This subject also has a steep slope. It turns out that the fingerprint impressions of this subject were collected during his adolescence (starting at the age of 11 until the age of 21). This explains the sharp decrease in genuine match scores due to growth in finger size [21].
- Outlier case 5 (Fig. S14): A positive slope is observed for this subject since the comparisons involving a lower quality fingerprint were made in shorter time interval than the comparisons with higher quality fingerprints. This example illustrates that the fingerprint image quality is not necessarily variable with respect to time elapsed.

4.3 Model Assessment and Comparison

Goodness-of-fit of a model evaluates how well the model fits the data. Furthermore, the impact of covariates on the observed responses can be assessed by comparing the goodness-of-fit of different models. The following three criteria are used to measure the goodness-of-fit: (i) Deviance, (ii) Akaike Information Criterion (AIC), and (iii) Bayesian Information Criterion (BIC). While the deviance measure is used to compare nested models, AIC and BIC add a constant term to the deviance for the sake of comparing non-nested models (see Supporting Information S4). The smaller the deviance (AIC or BIC), the better the model fit.

Table S2 shows the goodness-of-fit measures of the multilevel models fit to genuine match scores obtained by the two COTS matchers. The model comparisons based on the goodness-of-fit lead to the following observations:

- A decrease in deviance is observed when Models B_T, B_A, and B_Q are compared to Model A. This means that each individual covariate used in Model B ($\triangle T_{i,jk}$, $AGE_{i,jk}$, and $Q_{i,jk}$) can explain some of the variation in genuine match scores.
- Model B_Q provides a better fit to the data than Models B_T and B_A. This implies that fingerprint quality ($Q_{i,jk}$) is the best covariate to explain the variation in genuine match scores among the three covariates used in Model B.
- Gender and race are not important factors to explain the variation in genuine match scores since the deviance barely decreases from Model B_T to Models C_G or C_R.

- Models D and E show significantly smaller goodness-of-fit values than the other models. In other words, including all the three covariates ($\triangle T_{i,jk}$, $AGE_{i,jk}$, and $Q_{i,jk}$) in the multilevel model better explains the trend in genuine match scores compared to including only a single covariate. The additional interaction terms in Model E further improve the model fit.

4.4 Population-mean Trend of Impostor Fingerprint Match Scores

The impact of $\triangle T_{ij,k}$ and $AGE_{ij,k}$ on impostor match scores is evaluated in Models B_T and B_A. Although the hypothesis test ($H_0 : \beta_{10} = 0$) is rejected in both Models B_T and B_A (see Table S5), the population-mean trends of impostor match scores with respect to $\triangle T_{i,jk}$ and $AGE_{i,jk}$ show that the impostor match scores remain almost the same even as $\triangle T_{i,jk}$ and $AGE_{i,jk}$ increase (see Fig. S15).

4.5 Population-mean Trend of Probability of True Acceptance

The binary decisions are made on genuine match scores according to Eq. (2), and the probability of making a correct decision on a genuine fingerprint pair (true acceptance) is modeled by the multilevel model as shown in Eq. (7). The population-mean trends of the probability of true acceptance ($\pi_{i,jk}$) of the two COTS matchers with respect to $\triangle T_{i,jk}$ (Figs. 3(a) and S16(a)) indicate that the probability of true acceptance tends to remain close to 1 even though the time interval between the two fingerprints in comparison increases up to 12 years, the maximum time span in the longitudinal fingerprint dataset used in this study. This demonstrates that the genuine fingerprint pairs can be correctly recognized despite the increase in time interval between the fingerprints.

4.6 Population-mean Trend of Probability of False Acceptance

The probability of making an incorrect genuine-match decision on an impostor fingerprint pair (false acceptance) is also investigated. Figs. 3(b) and S16(b) indicate that the probability of false acceptance tends to remain close to 0 regardless of the time interval between the two fingerprints in comparison (within the 12-year time gap).

4.7 Results When Using All Ten Fingers for Recognition

Models B_T and B_A are fit to the genuine match scores from subject's all ten fingers fused by a sum rule according to Eq. (3). We observe a negative relationship between genuine match scores and the two covariates ($\triangle T_{i,jk}$ and $AGE_{i,jk}$); the null hypothesis ($\beta_{10} = 0$ in Model B) is rejected for both the models (see Fig. S17 and Table S6).

The population-mean trends of impostor match scores with respect to $\triangle T_{i,jk}$ and $AGE_{i,jk}$ (Fig. S18 and Table S7), probability of true acceptance (Figs. S19(a) and S20(a)), and probability of false acceptance (Figs. S19(b) and S20(b)) show the same behavior as the single-finger experiments. The 95% confidence intervals of the ten-finger fusion case become negligibly small, compared to using a single finger.

5 Conclusions

Since ancient times, fingerprints have been accepted as persistent and unique to an individual. Early scientific studies on fingerprint recognition in the late 19th century claimed that there is no significant change in the friction ridge structure over time by examining small sets of genuine fingerprint pairs captured over a large time interval. Although fingerprint recognition is now prevalent in distinguishing an extremely large number of individuals (for example, India's Aadhar program [22] involving over 1 billion residents), acceptance of the persistence of fingerprints has been only solely based on anecdotal evidence.

To understand the temporal behavior of fingerprint recognition accuracy, multiple fingerprint records of 15,597 subjects booked by the Michigan State Police over a duration of 5–12 years were collected. The

genuine and impostor match scores obtained by two COTS fingerprint matchers were analyzed by linear multilevel statistical models with various covariates, including time interval between the two fingerprints being compared, subject's age, and fingerprint image quality. Our longitudinal study of fingerprint recognition led to the following conclusions:

- The hypothesis test for the slope of a linear model indicates that the genuine and impostor match scores tend to decrease as the time interval between two fingerprints being compared increases. While a significant decrease in genuine match scores is observed, the decrease in impostor match scores is negligible.
- The genuine match scores also tend to decrease as the subject's age increases or when the fingerprint image quality decreases. However, the impostor match scores for the two COTS matchers show inconsistent tendencies over subject's age. Nevertheless, the change in impostor match scores with respect to subject's age is small.
- A comparison among the models with different covariates fit to the genuine match scores shows that:
  - Time interval, subject's age, and fingerprint image quality best explain the variation in genuine match scores; subject's gender and race are not significant covariates.
  - Among the three significant covariates (time interval, subject's age, and fingerprint image quality), fingerprint image quality is the most influential covariate.
  - The correlations (i) between time interval and fingerprint image quality and (ii) between subject's age and fingerprint image quality are negligibly small.
  - The impact of the interactions (i) between time interval and fingerprint image quality and (ii) between subject's age and fingerprint image quality on genuine match scores is not significant.
- It is observed that several subjects in the database do not conform to the population-mean trend as determined by model fit. These outlier subjects illustrate (i) subjects that follow the nomenclature in the Doddington's biometric zoo, (ii) a degradation in genuine match scores when a juvenile fingerprint is compared to the corresponding adult fingerprint, and (iii) presence of labeling errors in the operational fingerprint database.
- Despite the downward trend in genuine match scores over time, the probability of true acceptance, at a predetermined decision threshold, remains close to 1 (up to 12 years, the maximum time span in our database). On the other hand, the probability of false acceptance at the same decision threshold remains at 0 regardless of the time interval between two fingerprints. This demonstrates that fingerprint recognition accuracy tends to be stable even though the time interval between a fingerprint pair being compared increases.
- The inference made with a single finger (impressions from the right index finger) applies to the inference from ten-finger score fusion results.
- The results from two different COTS fingerprint matchers used in the study coincide, except for the tendency of impostor match scores with respect to subject's age.

The future work includes: (i) given that we make all pairwise comparisons of the fingerprint impressions from each subject, the correlation among the genuine match scores of a subject needs to be reflected in the model, and (ii) nonlinear multilevel models will be investigated and compared to the linear models presented in this study.

Acknowledgments

This research was supported by a grant from the NSF Center for Identification Technology Research (CITeR). We thank Professor Joseph Gardiner in the Department of Epidemiology and Biostatistics, Michigan State

University, and Karthik Nandakumar at the Institute for Infocomm Research, Singapore, for helpful discussions and suggestions. We also acknowledge the help of Capt. Greg Michaud, Director of Forensic Science Division at the Michigan State Police, in providing us the fingerprint database used in this study.

References

[1] Hawthorne MR (2009) *Fingerprints: Analysis and Understanding*, CRC Press.

[2] Frye v. United States (1923) *54 App. D.C. 46, 293 F. 1013.*

[3] Daubert V. Merrell Dow Pharmaceuticals, Inc. (1993) *509 U.S. 579, 113 S.Ct. 2786.*

[4] Science Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST) (2013) *Standards for Examining Friction Ridge Impressions and Resulting Conclusions (Latent/Tenprint).*

[5] Vanderkolk JR (2011) Chapter 9: Examination process. *Fingerprint Sourcebook*, National Institute of Justice/NCJRS.

[6] Pankanti S, Prabhakar S, Jain AK (2002) On the individuality of fingerprints. *IEEE Trans Pattern Anal Mach Intell*, 24(8):1010–1025.

[7] Zhu Y, Dass SC, Jain AK (2007) Statistical models for assessing the individuality of fingerprints. *IEEE Transactions on Information Forensics and Security*, 2(3):391–401.

[8] Chen Y, Jain AK (2009) Beyond minutiae: a fingerprint individuality model with pattern, ridge and pore features. *Proceedings of International Conference on Biometrics*, pp. 523-533.

[9] Neumann C, et. al. (2007) Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *J Forensic Sci*, 52(1):54–64.

[10] Herschel WJ (1916) *The Origins of Finger-Printing*, Oxford University Press.

[11] Galton F (1892) *Finger Prints*, Macmillan.

[12] Cummins H, Midlo M (1961) *Finger Prints, Palms and Soles: An Introduction to Dermatoglyphics*, Dover Publications.

[13] National Research Council (2009) *Strengthening Forensic Science in the United States: A Path Forward*, The National Academies Press.

[14] NIST Minutiae Interoperability Exchange Test (MINEX) (2004) `http://www.nist.gov/itl/iad/ig/minex04.cfm` (Date of access: June 3, 2014)

[15] Mansfield AJ, Wayman JL (2002) *Best Practices in Testing and Reporting Performance*, Center for Mathematics and Scientific Computing, National Physical Laboratory Teddington, U.K.

[16] Federal Office for Information Security (BSI) (2004) Evaluation of fingerprint recognition technology–bioFinger. `https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/Studies/BioFinger/BioFinger_pdf.pdf?__blob=publicationFile` (Date of access: June 3, 2014)

[17] Goldstein H (2010) *Multilevel Statistical Models* (Fourth Edition), Wiley.

[18] Singer JD, Willett JB (2003) *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*, Oxford University Press.

[19] Tabassi E, Wilson C, Watson C (2004) Fingerprint image quality. *National Institute of Standards and Technology Internal Report 7151.*

[20] Doddington G, Liggett W, Martin A, Przybocki N, Reynolds D (1998) Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. *Proceedings International Conference on Spoken Language Processing*.

[21] Gottschlich C, et. al. (2011) Modeling the growth of fingerprints improves matching for adolescents. *IEEE Transactions on Information Forensics and Security*, 6(3):1165–1169.

[22] Unique Identification Authority of India. `http://uidai.gov.in/` (Date of access: June 3, 2014)

Supporting Information

S1 Definitions

The terminologies used in this article are defined as follows.

S1.1 Evidential value

Evidential value of a comparison of two fingerprints refers to the strength of the fingerprint comparison as evidence to claim whether or not they come from the same finger [9, 11].

S1.2 Longitudinal data

Longitudinal data refers to repeated measurements on a collection of individuals sampled from a population over time. This is in contrast to cross-sectional data, where a single measurement is made on each individual [17].

S1.3 Balanced and time-structured data

A longitudinal dataset is characterized by (i) the number of measurements per individual and (ii) the time schedule used to make the measurements [18]. *Balanced* dataset means that every subject has the same number of measurements. *Time-structured* dataset consists of the repeated measurements following an identical time schedule across individuals. The sequence of measurements for each individual can be spaced either regularly or irregularly.

S1.4 Compound symmetry

The compound symmetry requires (i) homoscedasticity of variance: the variance of the measurements at a time instance across all subjects is the same as that of the measurements at another time instance, and (ii) constant covariance: the correlation between the measurements at the first and second time instances, for example, is the same as that between the measurements at the first and third time instances, and so on.

S2 Fingerprint Recognition

A fingerprint pattern consists of intervening ridge lines that are equidistantly spaced. Fingerprint features used for matching, both by forensic experts and machines (i.e., Automated Fingerprint Identification Systems (AFIS)), are typically represented at three different levels: (i) level-1 features (orientation field and singular points) describe ridge flow and pattern type (e.g., arch, loop, and whorl), (ii) level-2 features (minutiae) represent ridge details such as ridge ending and bifurcation points, and (iii) level-3 features (pore, incipient ridges, etc.) represent the finest details in fingerprints [23].

A comparison between two fingerprints is primarily based on the spatial configurations of minutiae in the corresponding impressions. If two fingerprint impressions show a high degree of agreement in minutiae configurations (resulting in high match score), the fingerprints are deemed to be a genuine pair, originating from the same finger (Fig. S1(a)). Otherwise, they are deemed to be an impostor pair (Fig. S1(b)).

Starting around 1900, the Scotland Yard included fingerprints in anthropometric identification cards which recorded measurements of various physical attributes of criminals [24]. Since then, the use of fingerprints has spread rapidly worldwide primarily for the purpose of tracking habitual criminals (repeat offenders) and identifying suspects based on partial fingerprints (latent fingerprints) found at crime scenes. With a phenomenal and continual increase in the size of fingerprint databases held by various law enforcement agencies, fingerprint recognition technology has made great strides both in terms of matching accuracy and matching speed (throughput). The Federal Bureau of Investigation (FBI) alone currently holds tenprint

(a)

(b)

Figure S1: Fingerprint comparison using minutiae configuration. Minutiae correspondences are shown for (a) a genuine fingerprint pair and (b) an impostor fingerprint pair. The fingerprint match (similarity) scores obtained by the COTS-2 matcher are (a) 389 and (b) 11 (note that the match score corresponding to false acceptance rate of 0.01% is 32).

records of over 75 million apprehended criminals and 39 million civilian government job applicants as of November 2013 [25]. The FBI's Integrated Automated Fingerprint Identification System (IAFIS) responds to a tenprint record of arrests and prosecutions (RAP) sheet request in 1 minute and 9 seconds, on average (97% of the requests are completed within 15 seconds) [25]. In the 2003 Fingerprint Vendor Technology Evaluation (FpVTE), the best performing commercial matcher achieved a 99.4% verification rate in searching against a database with 10,000 fingerprints [26]. When latent fingerprints are analyzed, human experts are inevitably involved in the latent search procedure to compensate for the limitations of state-of-the-art AFIS in reliable feature extraction and identification [27].

Improvements in fingerprint acquisition technology have led to the prevalent use of fingerprint recognition in various applications beyond law enforcement and forensics. Fingerprint impressions that were traditionally obtained by smearing fingers with ink and pressing them on paper are now acquired by optical sensors (e.g., at immigration counters in U.S. airports) and solid state sensors (e.g., in iPhone 5S), and these digital images of fingerprints can be readily processed by AFIS.

(a) (b)

Figure S2: Fingerprint pairs with minutiae correspondences labeled by Galton in his study on fingerprint persistence (image excerpt from [11]). (a) A pair of fingerprints with 13-year time interval showing perfect minutiae correspondences, and (b) a pair of fingerprints from another finger of the same subject with one minutia missing in the later age impression (denoted as 'A').

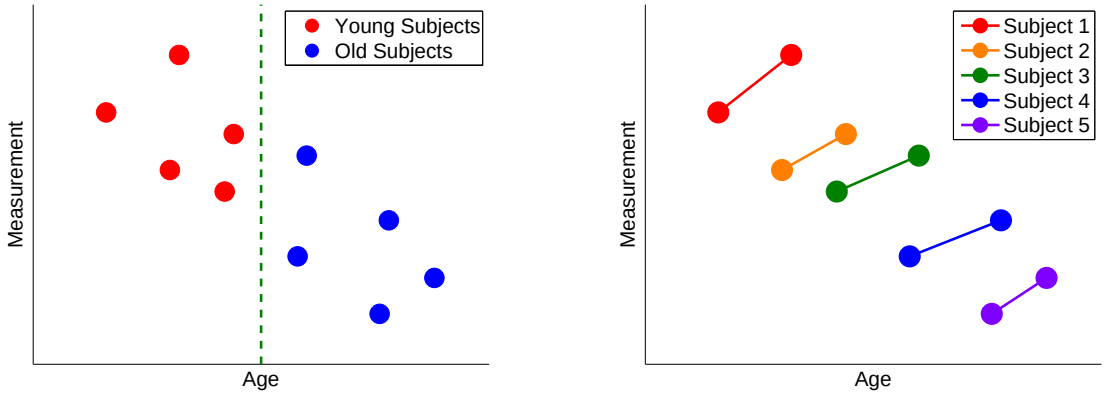
S3 Persistence Study of Biometrics Traits

Early studies on persistence of fingerprints focused on demonstrating the invariance of ridge structure in fingerprints with respect to time. Herschel collected three fingerprints of his son when he was 7, 17, and 40 years old and verified that all ridge details in the three fingerprints did not change over time [28]. Galton collected 11 pairs of fingerprints from six different individuals at two different time instances [11]. The time interval between a pair of fingerprints in Galton's collection ranged from 11 years to 31 years. The six subjects in his study were selected from different age groups; the age of the subject at the second impression was as young as 15 years and as old as 79 years. Among the 389 minutiae pairs that were manually labeled by Galton, only a single minutia was missing in a fingerprint pair (see Fig. S2).

More recently, a number of published studies have claimed template aging—an increase in the error rate in biometrics recognition with respect to the time gap between the query and the template [15]—for major biometrics modalities, including fingerprint [16], iris [29, 30], and face [31]. The biometric template is a compact representation of a subject's biometric data that is captured at the time of his initial enrollment in the system. A template then becomes the reference against which subsequent acquisitions of the subject are compared for authentication. The question these studies raised is essentially the following: "does the stored biometric template remain adequate for person authentication over time or should the template be updated to account for possible changes in a person's biometric trait?"

These prior studies used cross-sectional analysis by grouping the longitudinal data according to time interval between two acquisitions of a biometric trait and comparing the groups. However, cross-sectional analysis is not suitable for the longitudinal datasets used in studies which are unbalanced and time unstructured. Fig. S3 shows a hypothetical example which illustrates that if the dataset is unbalanced and/or time unstructured, cross-sectional analysis makes incorrect inference against the actual longitudinal behavior.

The longitudinal study on iris recognition by the National Institute of Standards and Technology (NIST) [32] properly used a nonlinear mixed-effects model to show the relationship between genuine iris match scores and covariates such as time elapsed after enrollment and the difference in iris dilation. However, the NIST study suffers from the following drawbacks: (i) the dataset used was truncated in the sense that the iris match scores from falsely rejected genuine comparisons were not included, and (ii) the validity of some of the assumptions in the mixed-effects model (i.e., normality of residuals and random effects) was



(a) (b)

Figure S3: Cross-sectional analysis versus longitudinal analysis of balanced but time-unstructured longitudinal data (adapted from [33]). For this dataset (two measurements for each of 5 subjects), (a) the cross-sectional analysis that discards subject labels on data makes an inference that *the measurement values tend to decrease with respect to subject's age*, while (b) the longitudinal analysis interprets the data as *the measurement values tend to increase with respect to age*.

not reported. The truncated data is problematic because the truncated portion of the data—erroneous identification decisions—is the very target of the analysis to determine the tendency of error rate with respect to time. Also, unless all model assumptions are satisfied, the analysis and inferences from the data cannot be trusted.

S4 Assessment of Goodness-of-Fit

The details of the goodness-of-fit measures used in this study are as follows.

- Deviance (D): Deviance can be used to compare the goodness-of-fit of nested models. The nested property is easily determined by checking if one model becomes equivalent to the other by setting the coefficients for some of the covariates to zero. For example, whereas Models A and B_T are nested and Models A and B_Q are nested, Models B_T and B_Q are not nested. The deviance is defined as:

$$D = -2\log(L), \tag{S1}$$

  where L is the maximum value of the likelihood function for the model.

- AIC: AIC can be used for any model comparison task (models do not need to be nested). AIC is defined as:

$$AIC = 2k - 2\log(L), \tag{S2}$$

  where k is the number of parameters in the model, and L is the maximum value of the likelihood function for the model.

- BIC: Under the assumption that the data distribution is in the exponential family, BIC is defined as:

$$BIC = k\log(n) - 2\log(L), \tag{S3}$$

  where k is the number of parameters in the model, n is the number of data points, and L is the maximum value of the likelihood function for the model. BIC also can be used for comparisons of non-nested models.

Table S2: Goodness-of-fit of the models shown in Table 1

Model	COTS-1			COTS-2		
	Deviance	AIC	BIC	Deviance	AIC	BIC
Model A	1,114,948	1,114,954	1,114,988	1,142,532	1,142,538	1,142,571
Model B_T	1,099,980	1,099,992	1,100,058	1,115,191	1,115,203	1,115,269
Model B_A	1,100,979	1,100,991	1,101,057	1,120,911	1,120,923	1,120,990
Model B_Q	1,028,899	1,028,911	1,028,978	1,060,037	1,060,049	1,060,115
Model C_G	1,099,969	1,099,985	1,100,074	1,115,117	1,115,133	1,115,222
Model C_R	1,099,817	1,099,833	1,099,921	1,114,378	1,114,394	1,114,483
Model D	1,003,908	1,003,938	1,004,105	1,019,412	1,019,442	1,019,608
Model E	1,003,839	1,003,873	1,004,062	1,018,986	1,019,020	1,019,209

S5 Validation of Normality Assumptions in Multilevel Model

The multilevel model assumes that the residuals ($\varepsilon_{i,jk}$) and random effects (b_{ri}) follow normal distributions. The inference made based on the model fitting is valid only if the underlying assumptions of the multilevel model are satisfied. The normal probability plot is a way to visually verify the normality of the data. If the normal probability plot is linear, one can ascertain that the data is from a normal distribution. Fig. S4 shows the normal probability plots of $\varepsilon_{i,jk}$, b_{0i}, and b_{1i} when Model B_T is fit to the genuine match scores obtained from the two COTS matchers.

While the residuals generally follow normal distributions, significant departures from normality are observed at the tails for the scores output by both the matchers. A possible cause of non-normality at the tails is that the scores from the COTS fingerprint matchers are typically censored, i.e., very low and high match scores are trimmed so that the output scores are in a finite range.

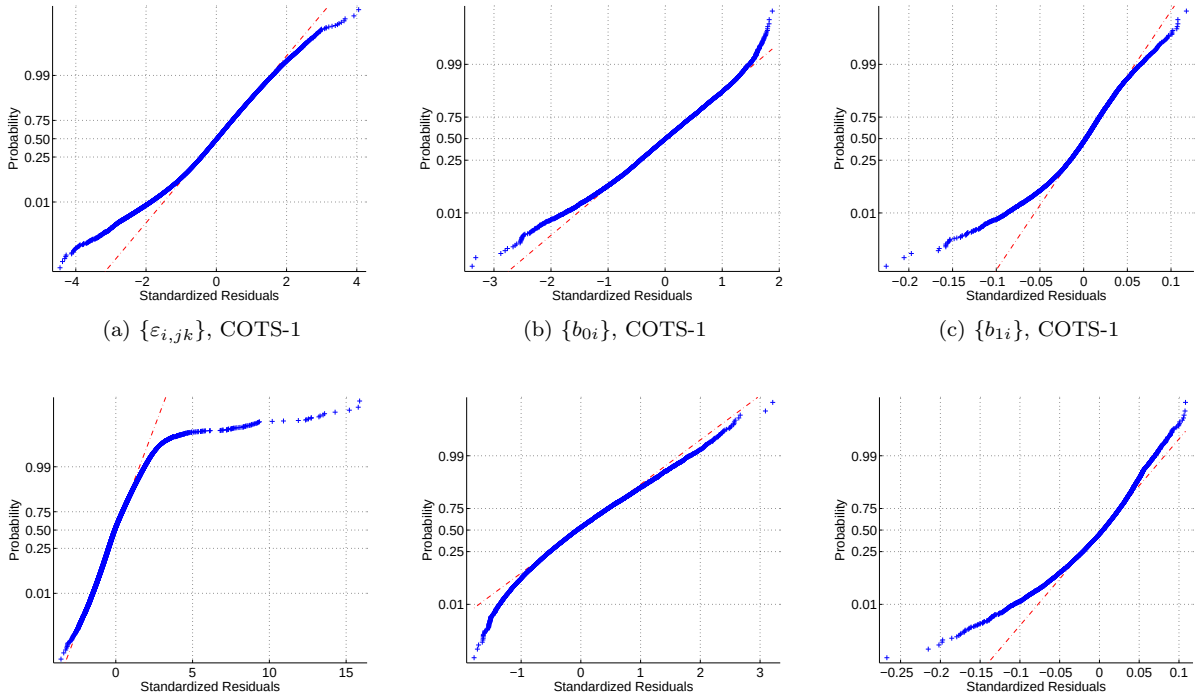
When the model assumptions are violated, the parameter estimates for fixed and random effects tend to be reliable while the standard errors (consequently, the confidence intervals) tend to be underestimated [34]. In this case, bootstrapping is a useful way to estimate parameters and confidence intervals [35].

S6 Parameter Estimates and Hypothesis Tests

In Models D and E, the fixed-effects parameter estimates for $\triangle T_{i,jk}$ (β_{10}), $AGE_{i,jk}$ (β_{20}), and $Q_{i,jk}$ (β_{30}) remain negative, similar to Models B_T, B_A, and B_Q. The correlations between any two covariates can be calculated from the estimated covariance matrices in Models D and E. In particular, we are interested in (i) σ_{13} which gives the correlation between $\triangle T_{i,jk}$ and $Q_{i,jk}$ and (ii) σ_{23} which gives the correlation between $AGE_{i,jk}$ and $Q_{i,jk}$. Although the estimated values for σ_{13} and σ_{23} are negative, the correlations among the covariates are very small—in Model D, the correlation coefficients for σ_{13} and σ_{23} based on COTS-1 match scores are -0.0324 and -0.0464; for COTS-2 matcher, they are -0.0174 and -0.1035. Moreover, σ_{13} in Models D and E with COTS-2 matcher cannot be claimed to be significantly different from 0 since the null hypothesis $\sigma_{13} = 0$ is not rejected at the 0.05 significance level.

The impact of the interactions between (i) $\triangle T_{i,jk}$ and $Q_{i,jk}$ and (ii) $AGE_{i,jk}$ and $Q_{i,jk}$ on genuine match scores is assessed by comparing the population-mean trends of genuine match scores with respect to $\triangle T_{i,jk}$ at different values of $AGE_{i,jk}$ and $Q_{i,jk}$ in Models D and E (see Fig. S6). As the 95% confidence intervals in Models D and E are overlapped, it cannot be said that these interactions significantly affect the variation in genuine match scores with respect to $\triangle T_{i,jk}$.

The temporal trend of genuine match scores is analyzed by fixing one of the covariates $AGE_{i,jk}$ and $Q_{i,jk}$ in Model E (see Figs. S7 and S8). In Fig. S7, the population-mean trends of genuine match scores with respect to $\triangle T_{i,jk}$ for each subject's age group ($AGE_{i,jk}$ is (a) 20, (b) 40, (c) 60, and (d) 78) are shown. For the age group of 20, the NFIQ reliably predicts the genuine match scores since the the population-mean trends at different values of $Q_{i,jk}$ are separable at the 0.05 significance level. However, for subjects at older



(a) $\{\varepsilon_{i,jk}\}$, COTS-1 (b) $\{b_{0i}\}$, COTS-1 (c) $\{b_{1i}\}$, COTS-1

(d) $\{\varepsilon_{i,jk}\}$, COTS-2 (e) $\{b_{0i}\}$, COTS-2 (f) $\{b_{1i}\}$, COTS-2

Figure S4: Normal probability plots of (a) and (d) residuals at level 1 ($\varepsilon_{i,jk}$), (b) and (e) random-effects for intercept at level 2 (b_{0i}), and (c) and (f) random-effects for slope at level 2 (b_{1i}) of Model B_T fit to the genuine match scores obtained from the two COTS matchers.

ages, the reliability of NFIQ in predicting genuine match score reduces. On the other hand, the population-mean trends of genuine match scores with respect to $\triangle T_{i,jk}$ for each fingerprint quality group ($Q_{i,jk}$ is (a) 1, (b) 3, and (c) 5) are shown in Fig. S8). At any level of fingerprint quality, the impact of subject's age is not significant on genuine match scores since the 95% confidence intervals of all age groups are completely overlapped.

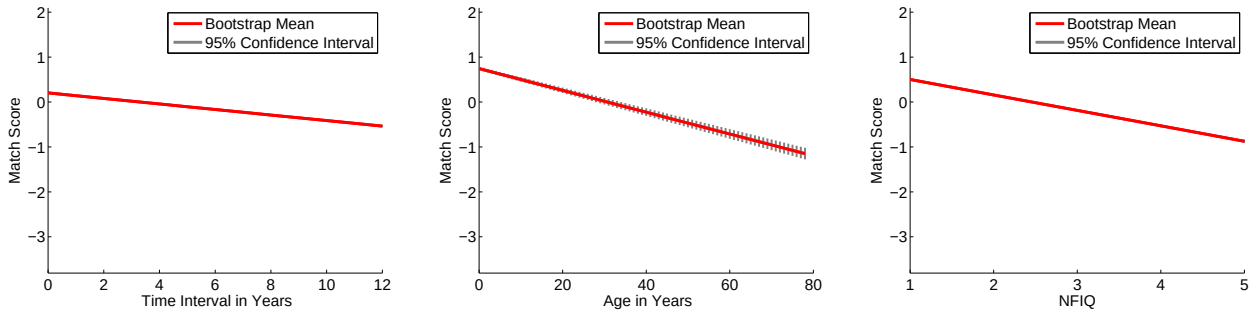
References

[23] Maltoni D, Maio D, Jain AK, Prabhakar S (2009) *Handbook of Fingerprint Recognition* (Second Edition), Springer-Verlag.

[24] Cole SA (2001) *Suspect Identities: A History of Fingerprinting and Criminal Identification*, Harvard University Press.

[25] The Federal Bureau of Investigation (FBI), Integrated Automated Fingerprint Identification System (IAFIS). `http://www.fbi.gov/about-us/cjis/fingerprints_biometrics/iafis/iafis` (Date of access: June 3, 2014)

[26] Wilson C, et. al. (2004) Fingerprint vendor technology evaluation 2003: summary of results and analysis report. *National Institute of Standards and Technology Internal Report 7123.*



(a) Model B_T (b) Model B_A (c) Model B_Q

Figure S5: Population-mean trends of genuine match scores obtained by COTS-2 matcher and 95% confidence intervals with respect to (a) $\triangle T_{i,jk}$, (b) $AGE_{i,jk}$, and (c) $Q_{i,jk}$, when a single finger is used for recognition.

[27] Indovina M, Dvornychenko V, Hicklin R, Kiebuzinski GI (2012) Evaluation of latent fingerprint technologies: extended feature sets [evaluation #2]. *National Institute of Standards and Technology Internal Report 7859*.

[28] Herschel WJ (1916) *The Origins of Finger-Printing*, Oxford University Press.

[29] Fenker SP, Bowyer KW (2011) Experimental evidence of a template aging effect in iris biometrics. *IEEE Workshop on Applications of Computer Vision*, pp. 232–239.

[30] Fenker SP, Bowyer KW (2012) Analysis of template aging in iris biometrics. *Proceedings of Computer Vision and Pattern Recognition Workshops*, pp. 45–51.

[31] Lanitis A (2010) A Survey of the effects of aging on biometric identity verification. *International Journal of Biometrics*, 2(1):34–52.

[32] Grother P, Matey JR, Tabassi E, Quinn GW, Chumakov M (2013) IREX VI: temporal stability of iris recognition accuracy. *National Institute of Standards and Technology Internal Report 7948*.

[33] Diggle PJ, Heagerty P, Liang KY, Zeger S (2003) *Analysis of Longitudinal Data*, Oxford University Press.

[34] Maas CJM, Hox JJ (2004) The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis*, 46(3):427–440.

[35] Leeden R, Busing F, Meijer E (1997) Bootstrap methods for two-level models. *Multilevel Conference*, Amsterdam, April 1–2.

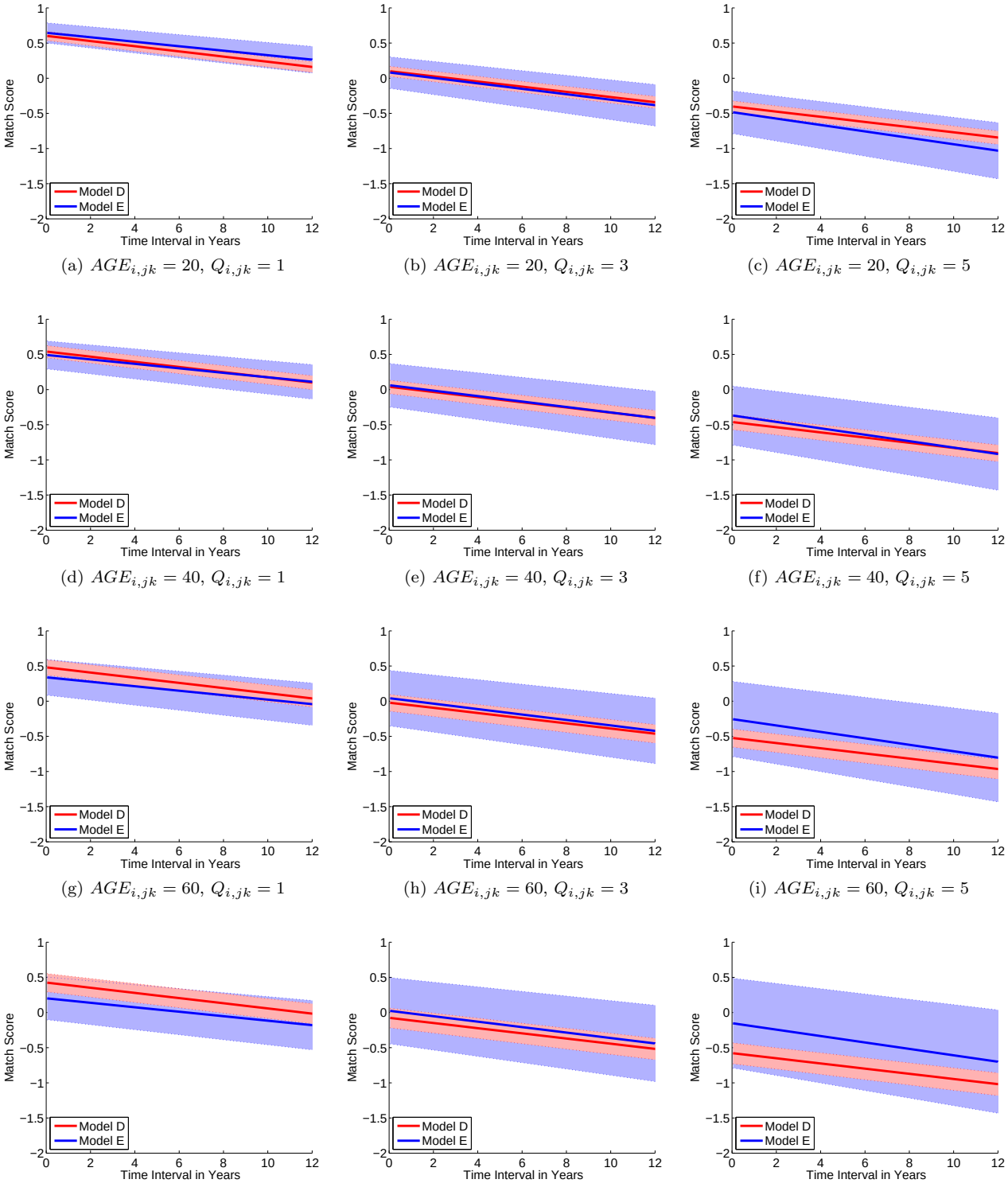
Table S3: Parameter estimates and 95% confidence intervals of genuine match scores obtained by COTS-1 matcher when a single finger is used for recognition

	Parameters	Model B_T	Model B_A	Model B_Q	Model D	Model E
Fixed Effects	β_{00}	0.1496 (0.1406; 0.1590)	0.5682 (0.5335; 0.6020)	0.7087 (0.6954; 0.7221)	0.9137 (0.8765; 0.9471)	1.1472 (1.0828; 1.2100)
	β_{10}	-0.0440 (-0.0450; -0.0430)	-0.0175 (-0.0185; -0.0164)	-0.2750 (-0.2798; -0.2702)	-0.0368 (-0.0378; -0.0358)	-0.0283 (-0.0313; -0.0255)
	β_{20}				-0.0030 (-0.0042, -0.0019)	-0.0110 (-0.0130; -0.0090)
	β_{30}				-0.2509 (-0.2558, -0.2463)	-0.3486 (-0.3739; -0.3242)
	β_{40}				-0.2509 (-0.2558, -0.2463)	-0.0035 (-0.0045; -0.0024)
	β_{50}				-0.2509 (-0.2558, -0.2463)	0.0033 (0.0026; 0.0041)
Variance Components	σ_{ε}^2	0.7057	0.6998	0.6489	0.6033	0.6032
	σ_0^2	0.5298	5.6003	0.9096	6.6328	6.6570
	σ_1^2	0.0034	0.0050	0.1163	0.0041	0.0041
	σ_{01}	-0.0134	-0.1574	-0.2543	0.0944	0.0950
	σ_2^2				0.0068	0.0068
	σ_{02}				-0.1941	-0.1945
	σ_{12}				-0.0036	-0.0036
	σ_3^2				0.1165	0.1181
	σ_{03}				-0.2092	-0.2207
	σ_{13}				-0.0007	-0.0011
	σ_{23}				-0.0013	-0.0010

Table S4: Parameter estimates and 95% confidence intervals of genuine match scores obtained by COTS-2 matcher when a single finger is used for recognition

	Parameters	Model B_T	Model B_A	Model B_Q	Model D	Model E
Fixed Effects	β_{00}	0.2032	0.7447	0.8456	1.0353	1.4399
		(0.1939; 0.2127)	(0.7072; 0.7843)	(0.8316; 0.8595)	(0.9947; 1.0750)	(1.3706; 1.5103)
	β_{10}	-0.0616	-0.0243	-0.3439	-0.0533	-0.0654
		(-0.0625; -0.0606)	(-0.0254; -0.0231)	(-0.3489; -0.3385)	(-0.0543; -0.0522)	(-0.0679; -0.0629)
	β_{20}				-0.0024	-0.0130
					(-0.0036, -0.0011)	(-0.0152; -0.0107)
	β_{30}				-0.3064	-0.4694
					(-0.3112, -0.3015)	(-0.4925; -0.4466)
	β_{40}				-0.3064	0.0048
					(-0.3112, -0.3015)	(0.0039; 0.0057)
	β_{50}				-0.3064	0.0043
					(-0.3112, -0.3015)	(0.0036; 0.0050)
Variance Components	σ^2_ε	0.7185	0.7120	0.6738	0.6127	0.6125
	σ^2_0	0.5744	7.5575	0.9105	7.8996	7.8362
	σ^2_1	0.0039	0.0066	0.1027	0.0040	0.0040
	σ_{01}	-0.0277	-0.2136	-0.2473	0.0800	0.0825
	σ^2_2				0.0082	0.0081
	σ_{02}				-0.2335	-0.2314
	σ_{12}				-0.0033	-0.0033
	σ^2_3				0.1039	0.1077
	σ_{03}				-0.1466	-0.1646
	σ_{13}				-0.0004 *	-0.0005*
	σ_{23}				-0.0030	-0.0027

* The hypothesis test gives that the null hypothesis that the parameter is zero is not rejected at a significance level of 0.05.



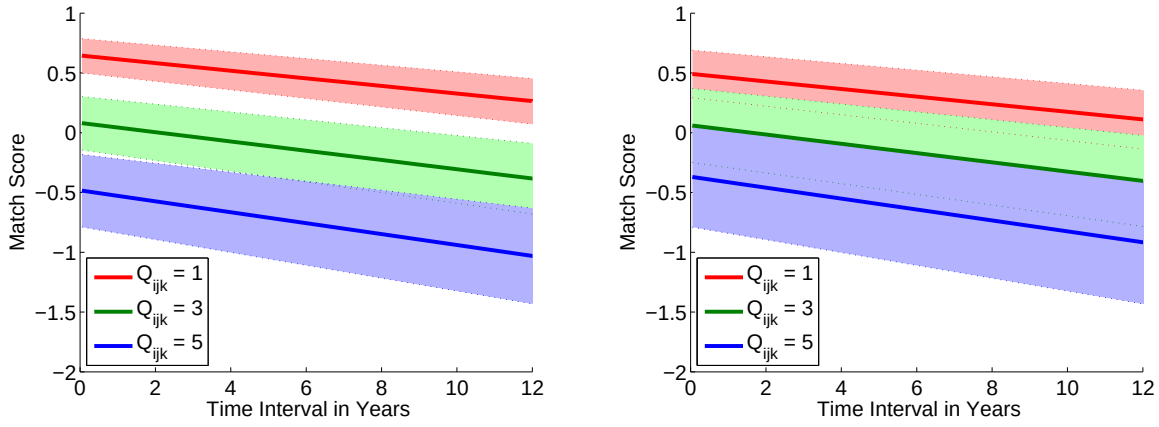
(a) $AGE_{i,jk} = 20$, $Q_{i,jk} = 1$ (b) $AGE_{i,jk} = 20$, $Q_{i,jk} = 3$ (c) $AGE_{i,jk} = 20$, $Q_{i,jk} = 5$

(d) $AGE_{i,jk} = 40$, $Q_{i,jk} = 1$ (e) $AGE_{i,jk} = 40$, $Q_{i,jk} = 3$ (f) $AGE_{i,jk} = 40$, $Q_{i,jk} = 5$

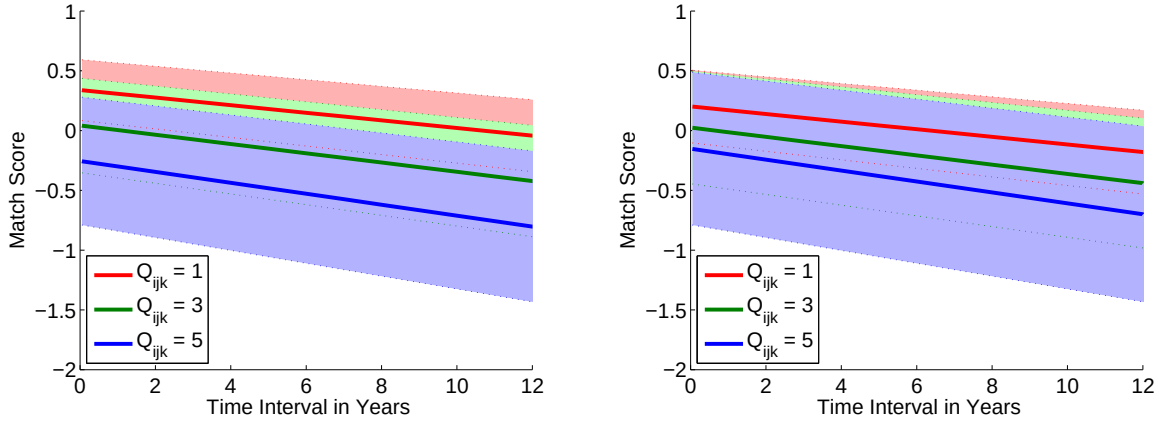
(g) $AGE_{i,jk} = 60$, $Q_{i,jk} = 1$ (h) $AGE_{i,jk} = 60$, $Q_{i,jk} = 3$ (i) $AGE_{i,jk} = 60$, $Q_{i,jk} = 5$

(j) $AGE_{i,jk} = 78$, $Q_{i,jk} = 1$ (k) $AGE_{i,jk} = 78$, $Q_{i,jk} = 3$ (l) $AGE_{i,jk} = 78$, $Q_{i,jk} = 5$

Figure S6: Comparison between Models D and E. Population-mean trends of genuine match scores with respect to $\triangle T_{i,jk}$ are shown when $AGE_{i,jk}$ varies from 20 to 78 and $Q_{i,jk}$ varies from 1 to 5 in Models D and E. Solid lines are the bootstrap means, and the shaded areas represent the 95% confidence intervals. A single finger is used for recognition and match scores are obtained from COTS-1 matcher.

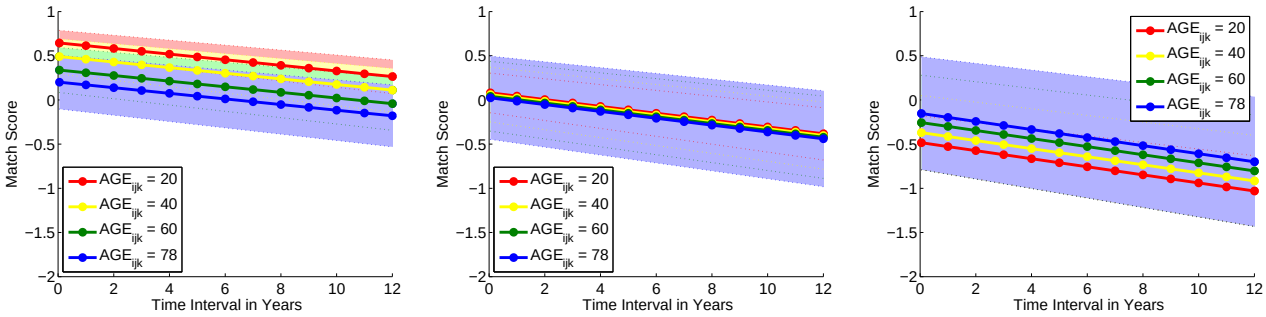


(a) $AGE_{i,jk} = 20$ (b) $AGE_{i,jk} = 40$



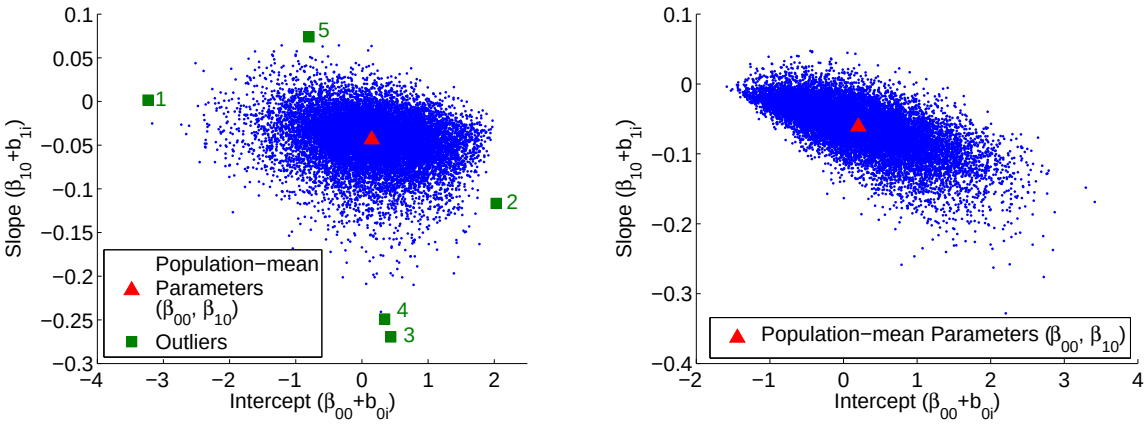
(c) $AGE_{i,jk} = 60$ (d) $AGE_{i,jk} = 78$

Figure S7: Population-mean trends of genuine match scores with respect to $\triangle T_{i,jk}$ when $AGE_{i,jk}$ is fixed and $Q_{i,jk}$ varies from 1 to 5 in Model E. Solid lines are the bootstrap means, and the shaded areas represent the 95% confidence intervals. A single finger is used for recognition and match scores are obtained from COTS-1 matcher.



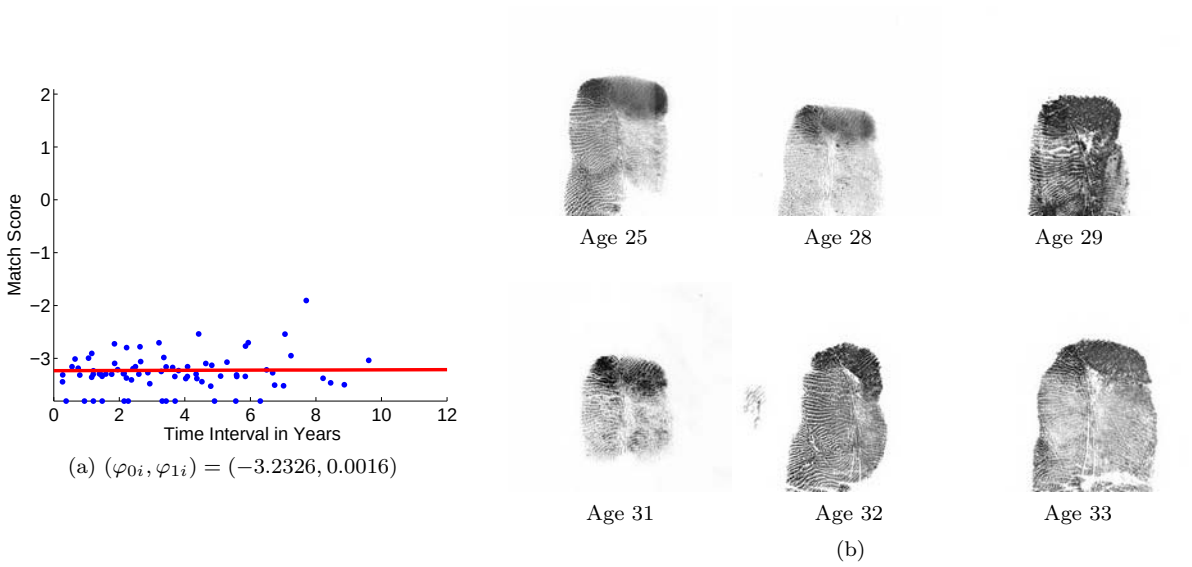
(a) $Q_{i,jk} = 1$ (b) $Q_{i,jk} = 3$ (c) $Q_{i,jk} = 5$

Figure S8: Population-mean trends of genuine match scores with respect to $\triangle T_{i,jk}$ when $Q_{i,jk}$ is fixed and $AGE_{i,jk}$ varies from 20 to 78 in Model E. Solid lines are the bootstrap means, and the shaded areas represent the 95% confidence intervals. A single finger is used for recognition and match scores are obtained from COTS-1 matcher.



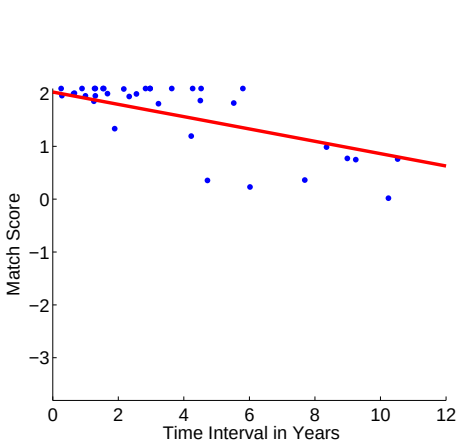
(a) COTS-1 (b) COTS-2

Figure S9: Parameter estimates of Model B_T with genuine match scores provided by two COTS matchers. The estimates for the population-mean parameters (β_{00}, β_{10}) and the parameters for each subject $(\varphi_{0i}, \varphi_{1i})$ are represented as red triangle and blue dots, respectively. The parameters associated with five outlier subjects are marked as green squares in (a).

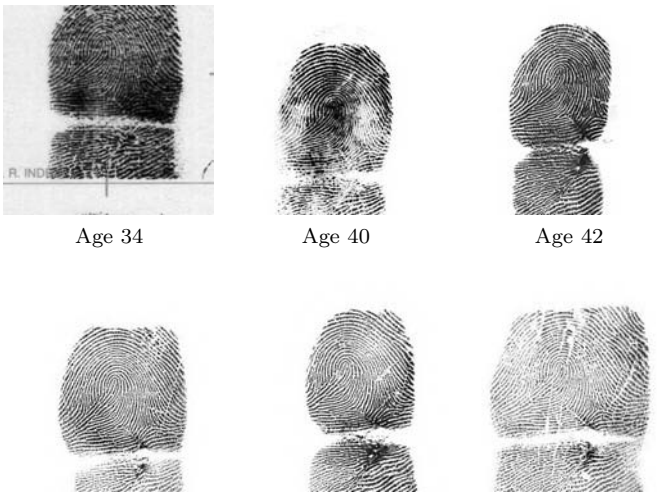


(a) $(\varphi_{0i}, \varphi_{1i}) = (-3.2326, 0.0016)$ (b)

Figure S10: A subject whose intercept in Model B_T is very small due to the severe alteration (i.e., scarring) of the fingerprint pattern (outlier case 1). (a) The observed responses and fitting result of the subject, and (b) fingerprint impressions of the subject at different ages.



(a) $(\varphi_{0i}, \varphi_{1i}) = (2.0281, -0.1167)$

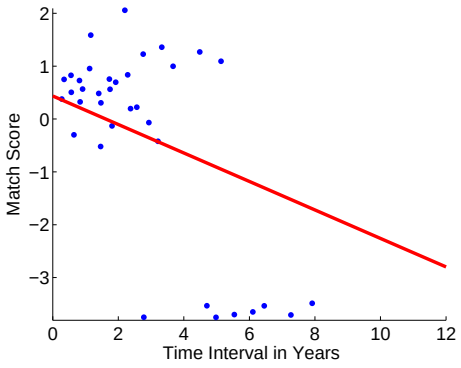


Age 34 Age 40 Age 42

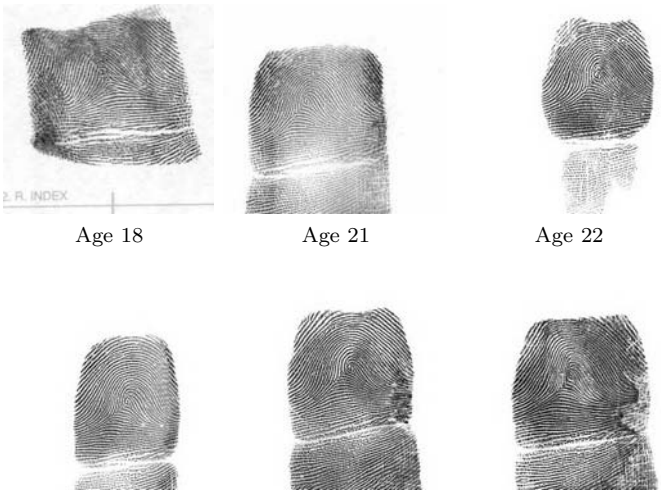
Age 43 Age 44 Age 45

(b)

Figure S11: A subject with high quality ridge pattern resulting in the large intercept in Model B_T (outlier case 2). (a) The observed responses and fitting result of the subject, and (b) fingerprint impressions of the subject at different ages.



(a) $(\varphi_{0i}, \varphi_{1i}) = (0.4353, -0.2695)$

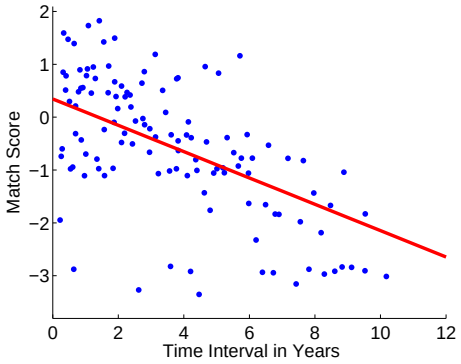


Age 18 Age 21 Age 22

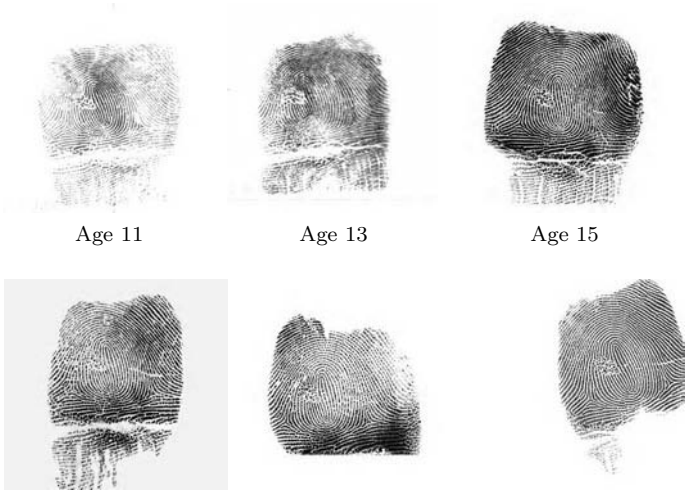
Age 23 Age 24 Age 26

(b)

Figure S12: A subject with steep negative slope (outlier case 3) resulting from a mislabeled fingerprint (the first impression). (a) The observed responses and fitting result of the subject, and (b) fingerprint impressions of the subject at different ages.



(a) $(\varphi_{0i}, \varphi_{1i}) = (0.3443, -0.2491)$

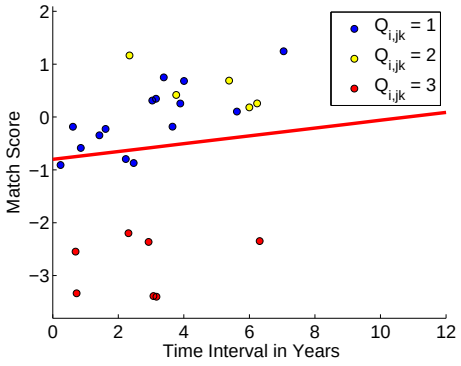


Age 11 Age 13 Age 15

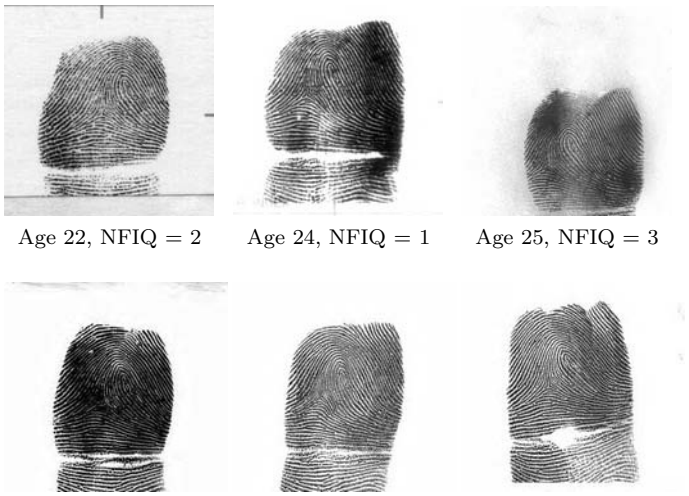
Age 17 Age 19 Age 21

(b)

Figure S13: A subject with steep negative slope due to fingerprint impressions made during his adolescence (outlier case 4). (a) The observed responses and fitting result of the subject, and (b) fingerprint impressions of the subject at different ages.



(a) $(\varphi_{0i}, \varphi_{1i}) = (-0.8006, 0.0740)$

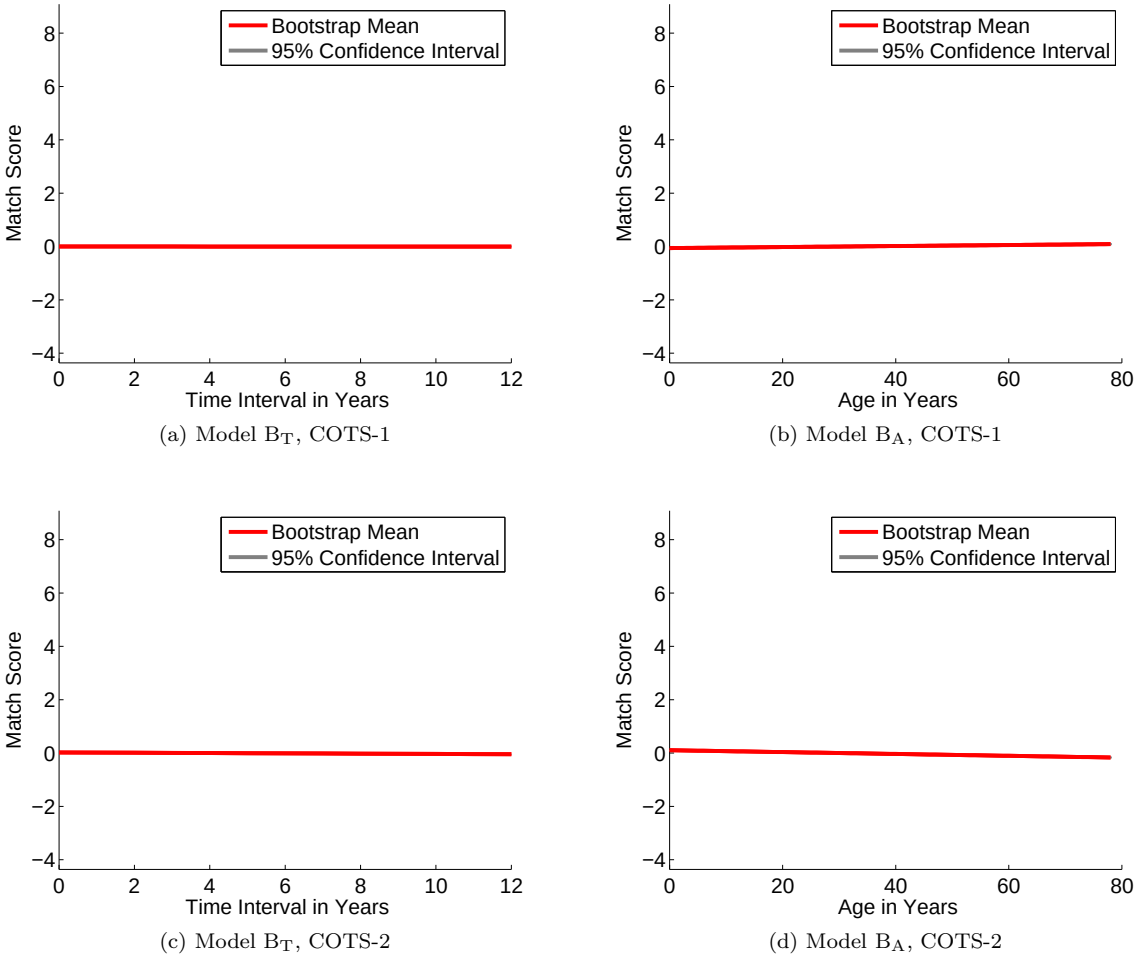


Age 22, NFIQ = 2 Age 24, NFIQ = 1 Age 25, NFIQ = 3

Age 27, NFIQ = 1 Age 28, NFIQ = 1 Age 32, NFIQ = 1

(b)

Figure S14: A subject with positive slope (outlier case 5) where the comparisons involving a lower quality fingerprint (at age 25) have short time intervals. (a) The observed responses and fitting result of the subject, and (b) fingerprint impressions of the subject at different ages.



(a) Model B_T, COTS-1

(b) Model B_A, COTS-1

(c) Model B_T, COTS-2

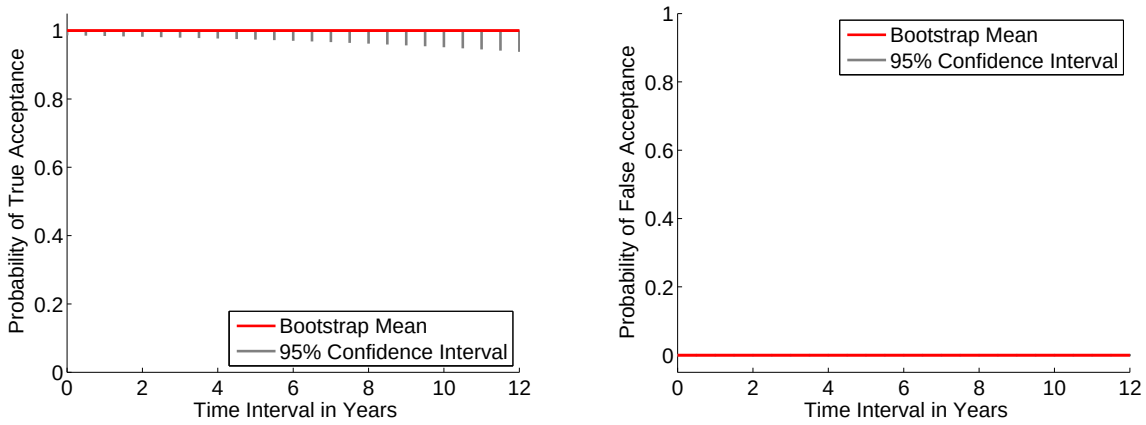
(d) Model B_A, COTS-2

Figure S15: Population-mean trends of impostor match scores obtained by two COTS matchers and 95% confidence intervals with respect to (a) and (c) $\triangle T_{i,jk}$, and (b) and (d) $AGE_{i,jk}$, when a single finger is used for recognition.

Table S5: Parameter estimates and 95% confidence intervals of impostor match scores obtained by two COTS matchers when a single finger is used for recognition

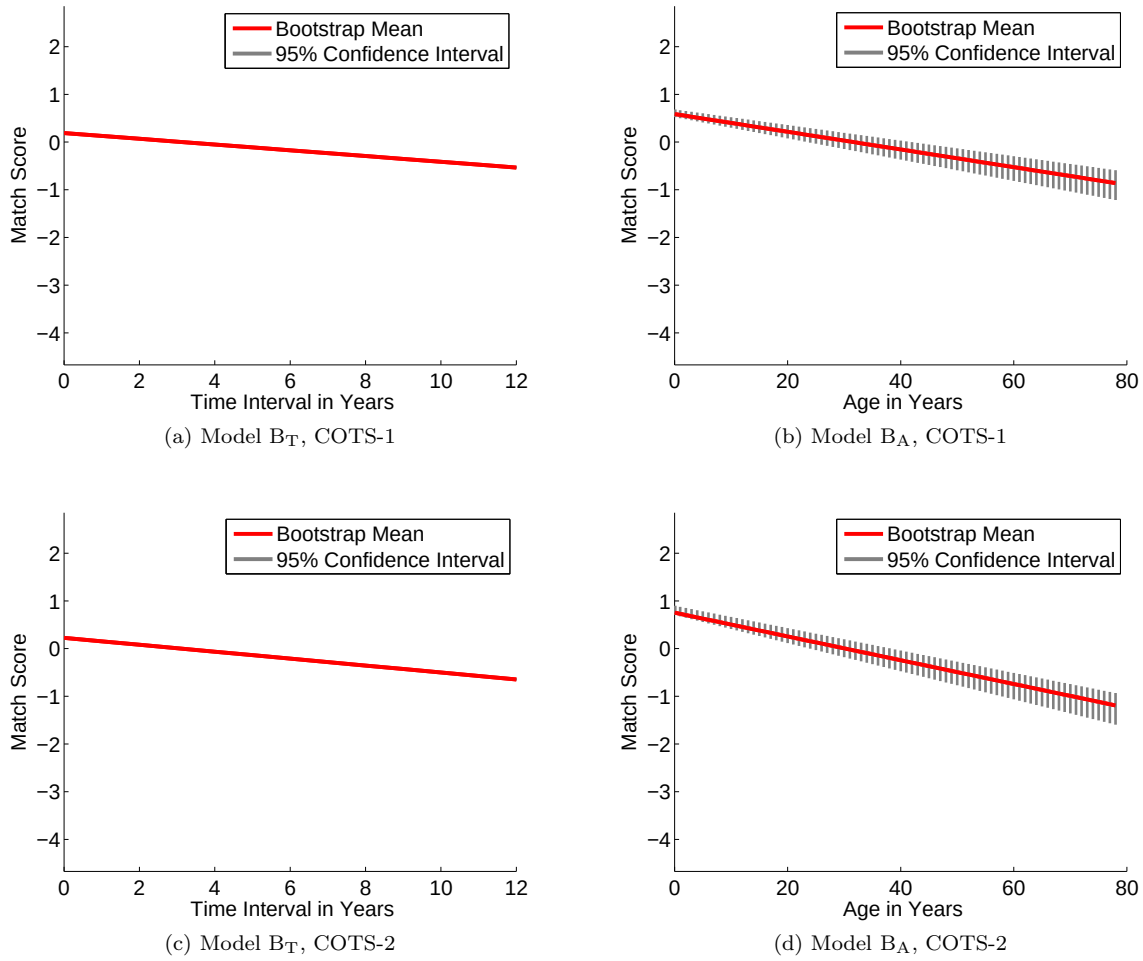
	Parameters	COTS-1		COTS-2	
		Model B_T	Model B_A	Model B_T	Model B_A
Fixed Effects	β_{00}	0.0004 (-0.0044; 5.4390e-03)*	-0.0589 (-0.0711; -0.0475)	0.0220 (0.0183; 0.0257)	0.1057 (0.0946; 0.1162)
	β_{10}	-0.0005 (-0.0009; -6.3655e-07)	0.0019 (0.0016; 0.0023)	-0.0056 (-0.0061; -0.0051)	-0.0035 (-0.0038; -0.0032)
Variance Components	σ_ε^2	0.9292	0.9299	0.9673	0.9690
	σ_0^2	0.1479	0.4347	0.0720	0.2949
	σ_1^2	0.0004	0.0002	0.0008	0.0002
	σ_{01}	-0.0028	-0.0088	-0.0036	-0.0076

* The hypothesis test gives that the null hypothesis that the parameter is zero is not rejected at a significance level of 0.05.



(a) (b)

Figure S16: Population-mean trend of fingerprint matching accuracy and 95% confidence interval with respect to $\triangle T_{i,jk}$. (a) Probability of true acceptance and (b) probability of false acceptance with respect to $\triangle T_{i,jk}$. Match scores are obtained by COTS-2 matcher when a single (right index) finger is used for recognition.



(a) Model B_T, COTS-1

(b) Model B_A, COTS-1

(c) Model B_T, COTS-2

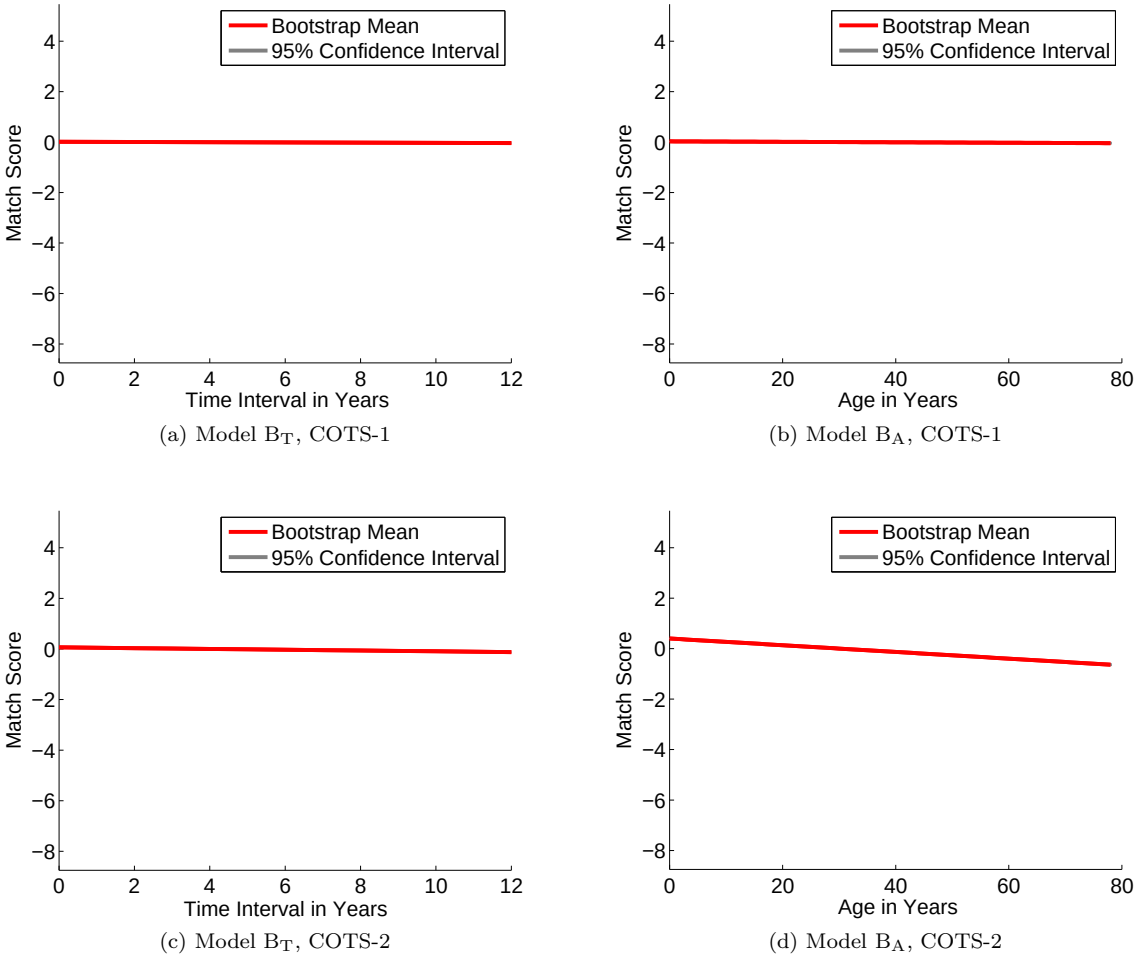
(d) Model B_A, COTS-2

Figure S17: Population-mean trends of genuine match scores obtained by two COTS matchers and 95% confidence intervals with respect to (a) and (c) $\triangle T_{i,jk}$ and (b) and (d) $AGE_{i,jk}$, when the scores from ten fingers are fused.

Table S6: Parameter estimates and 95% confidence intervals when the genuine match scores from ten fingers obtained by two COTS matchers are fused by a sum rule.

	Parameters	COTS-1		COTS-2	
		Model B_T	Model B_A	Model B_T	Model B_A
Fixed Effects	β_{00}	0.1896 (0.1800; 0.1995)	0.5867 (0.5231; 0.6841)	0.2258 (0.2159; 0.2360)	0.7537 (0.7056; 0.8996)
	β_{10}	-0.0603 (-0.0612; -0.0594)	-0.0185 (-0.0223; -0.0163)	-0.0726 (-0.0736; -0.0717)	-0.0249 (-0.0295; -0.0235)
Variance Components	σ^2_ε	0.6562	0.6651	0.6615	0.6602
	σ^2_0	0.5986	5.0636*	0.6599	8.6870
	σ^2_1	0.0037	0.0046	0.0040	0.0074
	σ_{01}	-0.0144	-0.1409*	-0.0304	-0.2423

* The hypothesis test gives that the null hypothesis that the parameter is zero is not rejected at a significance level of 0.05.



(a) Model B_T, COTS-1

(b) Model B_A, COTS-1

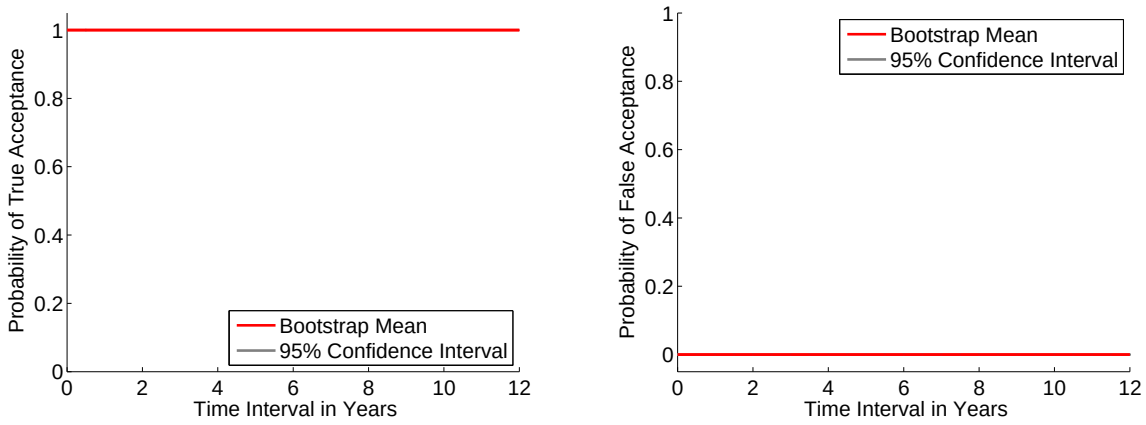
(c) Model B_T, COTS-2

(d) Model B_A, COTS-2

Figure S18: Population-mean trends of impostor match scores obtained by two COTS matchers and 95% confidence intervals with respect to (a) and (c) $\triangle T_{i,jk}$ and (b) and (d) $AGE_{i,jk}$, when the scores from ten fingers are fused.

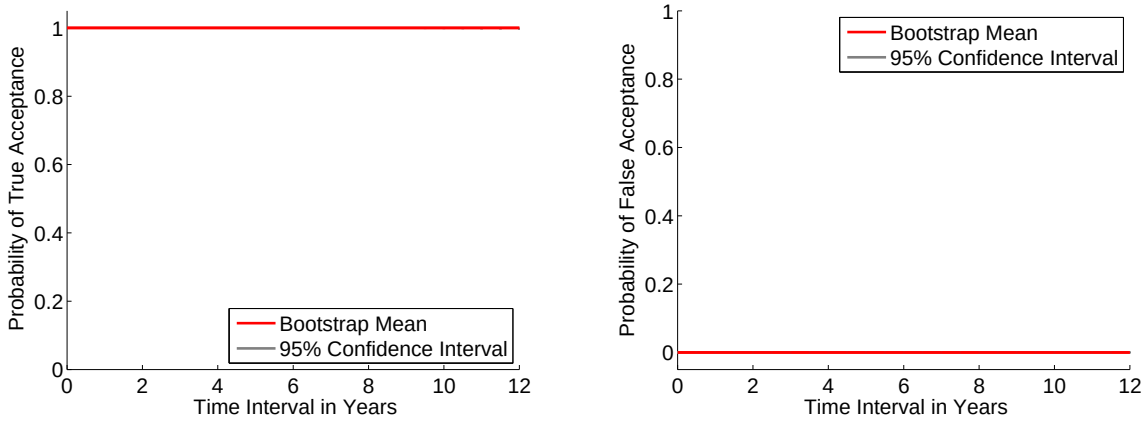
Table S7: Parameter estimates and 95% confidence intervals when the impostor match scores from ten fingers obtained by two COTS matchers are fused by a sum rule.

		COTS-1		COTS-2	
	Parameters	Model B_T	Model B_A	Model B_T	Model B_A
Fixed Effects	β_{00}	0.0133 (0.0065; 0.0202)	0.0290 (0.0134; 0.0456)	0.0633 (0.0575; 0.0689)	0.4013 (0.3850; 0.4176)
	β_{10}	-0.0044 (-0.0050; -0.0039)	-0.0009 (-0.0014; -0.0004)	-0.0155 (-0.0161; -0.0149)	-0.0133 (-0.0138; -0.0128)
Variance Components	σ^2_ε	0.8520	0.8527	0.9037	0.9049
	σ^2_0	0.3004	0.9797	0.1966	1.0748
	σ^2_1	0.0009	0.0007	0.0016	0.0010
	σ_{01}	-0.0058	-0.0216	-0.0071	-0.0298



(a) (b)

Figure S19: Fingerprint matching accuracy with respect to $\triangle T_{i,jk}$. (a) Probability of true acceptance and (b) probability of false acceptance with respect to $\triangle T_{i,jk}$. Match scores are obtained by COTS-1 matcher when the match scores from all ten fingers are fused by the sum rule.



(a) (b)

Figure S20: Fingerprint matching accuracy with respect to $\triangle T_{i,jk}$. (a) Probability of true acceptance and (b) probability of false acceptance with respect to $\triangle T_{i,jk}$. Match scores are obtained by COTS-2 matcher when the match scores from all ten fingers are fused by the sum rule.