An Empirical Study of Sample Size in ROC-Curve Analysis of Fingerprint Data

Jin Chu Wu* and Charles L. Wilson
Image Group, Information Access Division, Information Technology Laboratory
National Institute of Standards and Technology, Gaithersburg, MD 20899

Abstract

The fingerprint datasets in many cases may exceed millions of samples. Thus, the needed size of a biometric evaluation test sample is an important issue in terms of both accuracy and efficiency. In this article, an empirical study, namely, using Chebyshev's inequality in combination with simple random sampling, is applied to determine the sample size for biometric applications. No parametric model is assumed, since the underlying distribution functions of the similarity scores are unknown. The performance of fingerprint-image matcher is measured by a Receiver Operating Characteristic (ROC) curve. Both the area under an ROC curve and the True Accept Rate (TAR) at an operational False Accept Rate (FAR) are employed. The Chebyshev's greater-than-95% intervals of using these two criteria based on 500 Monte Carlo iterations are computed for different sample sizes as well as for both high- and low-quality fingerprint-image matchers. The stability of such Monte Carlo calculations with respect to the number of iterations is also explored. The choice of sample size depends on matchers' qualities as well as on which performance criterion is invoked. In general, for 6,000 match similarity scores, 50,000 to 70,000 scores randomly selected from 35,994,000 non-match similarity scores can ensure the accuracy with greater-than-95% probability.

Keywords: Empirical Study, Chebyshev's Inequality, Simple Random Sampling, Sample Size, Receiver Operating Characteristic (ROC) Curve, Data Analysis, Stability Metric, Monte Carlo Calculation, Biometrics, Fingerprint Matching

1. INTRODUCTION

The fingerprint datasets in many cases may exceed millions of samples. Thus, the needed size of a biometric evaluation test sample is an important issue in terms of both accuracy and efficiency. Over the past two years, the National Institute of Standards and Technology (NIST) has used large samples of fingerprint data from a wide range of government sources to evaluate the fingerprint-image matchers from different vendors[1] [1,2]. In the SDK tests [2], 6,000 subjects' fingerprint images are used as a probe, and 6,000 second fingerprint images of the same subjects are used as a gallery. The probe is matched against the gallery. It creates 6,000 match similarity scores from the

* Corresponding author. Tel: + 301-975-6996; fax: + 301-975-5287. E-mail address: jinchu.wu@nist.gov (J.C. Wu).

1 These tests were performed for the Department of Homeland Security in accordance with section 303 of the Border Security Act, codified at 8 U.S.C. 1732. Specific hardware and software products identified in this report were used in order to adequately support the development of technology to conduct the performance evaluations described in this document. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

Biometric Technology for Human Identification III, edited by Patrick J. Flynn, Sharath Pankanti,
Proc. of SPIE Vol. 6202, 620207, (2006) · 0277-786X/06/$15 · doi: 10.1117/12.665601

same subjects' different fingerprint-image comparisons, and 35,994,000 non-match similarity scores from different subjects' fingerprint-image comparisons.

For such fingerprint data, there is usually no underlying parametric distribution function for similarity scores. Thus, the nonparametric approach must be employed [3]. The True Accept Rate (TAR) and the False Accept Rate (FAR) are defined, respectively, as the cumulative probability of the match and non-match similarity scores at a specified similarity score (i.e., threshold) from the highest match and non-match similarity score. A Receiver Operating Characteristic (ROC) curve is constructed based on TAR and FAR by moving the threshold, one similarity score at a time, from the highest similarity score to the lowest similarity score [3]. An ROC curve can be measured by using either the area under the ROC curve [3] or the TAR value at an operational FAR value [1,2]. In this article, the fingerprint-image matcher is evaluated by an ROC curve using both of these two criteria. The size of our fingerprint datasets is very large in comparison to the applications of ROC curves in other areas [4,5,6,7, and references therein].

How much fingerprint data should be selected from a large dataset to obtain both efficiency and accuracy in biometric evaluation? Different sizes of samples generate different ROC curves. Hence, the sample size can be determined by the accepted deviations of ROC curves for samples with reduced sizes from the ROC curve in the baseline, i.e., *Δ (ROC curve)*, in terms of both or either of the above two criteria, at a specified probability (e.g., 95%). The baseline can be generated from the largest dataset that the available computer power can handle from the largest consolidated dataset.

If the current SDK evaluation [2] is set to be a baseline, then with respect to 6,000 match similarity scores, out of 35,994,000 non-match similarity scores, how many scores are needed to achieve the same performance? In other words, the issue of determining the sample size for SDK turns out to be: 1) reduce the number of non-match similarity scores, 2) take *one* trial, 3) the result must be close to the baseline result within an accepted tolerance at a specified probability. In this article, an empirical study, namely, using Chebyshev's inequality in combination with simple random sampling, was applied to determine the sample size for biometric applications. Although, for simplicity, in this article the number of match similarity scores was fixed and the number of non-match similarity scores varied, the same methodology can be applied to other scenarios.

As specified above, one of our requirements is that the test be performed only once. To satisfy this objective, Chebyshev's inequality is invoked. Using Chebyshev's inequality, an interval in which a percentage of population resides can be determined, provided that the lower bound on the probability is specified. If an interval can contain the baseline result as well as, for example, greater than 95% of the test results in a specified circumstance, then the one-trial test result will have greater-than-95% probability to fall in that interval and its deviation from the baseline result will not exceed the length of the interval. This is consistent with the above statement of *Δ (ROC curve)*.

Further, as stated above, the number of match similarity scores is fixed as 6,000. Thus, to ensure that the ROC curves from the test results are close to the ROC curve in the baseline using the above criteria, the distributions of non-match similarity scores with reduced sizes must be "very similar" to the distribution of 35,994,000 non-match similarity scores in the baseline. To serve this purpose, the simple random sampling without replacement is applied. A simple random sample selected from

35,994,000 non-match similarity scores constitutes a new set of non-match similarity scores, and its distribution is used with the distribution of 6,000 match similarity scores to generate an ROC curve.

A Chebyshev's greater-than-95% interval can be obtained using a Monte Carlo calculation. Different sizes of simple random samples are selected from 35,994,000 non-match similarity scores in the baseline. 500 Monte Carlo iterations are carried out for different sample sizes. Thereafter, the sample size of non-match similarity scores can be determined according to whether the Chebyshev's interval is within an accepted tolerance. In addition, the stability of the Monte Carlo calculation with respect to the number of iterations is also explored in this article. It is quantified by the worst deviation of the test result from the baseline result within the Chebyshev's interval.

The methods, i.e., Chebyshev's inequality and Chebyshev's greater-than-95% interval, the simple random sampling, and the stability metric, are presented in Section 2. The results of their applications to determining sample sizes in biometric evaluation of fingerprint-image matchers are provided in Section 3. Discussion of the sampling error of the sample mean and other issues can be found in Section 4.

2. METHODS

An empirical study, i.e., using Chebyshev's inequality along with simple random sampling, is used to determine the sample size for biometric applications. The stability of the calculation with respect to the number of Monte Carlo iterations will be addressed as well.

2.1 CHEBYSHEV'S INEQUALITY [8] AND CHEBYSHEV'S GREATER-THAN-95% INTERVAL

If ξ is a random variable and its mean and variance exist, i.e., $M(\xi) = \mu < \infty$ and $V(\xi) = \sigma^2 < \infty$, then Chebyshev's inequality

$$P\{|\xi - \mu| \geq k\sigma\} \leq \frac{1}{k^2} \tag{1}$$

is valid for any $k > 1$. A variation of Chebyshev's inequality can be expressed as

$$P\{|\xi - \mu| < k\sigma\} > 1 - \frac{1}{k^2} \tag{2}$$

It states that greater than $(1 - 1 / k^2)$ of population falls within k ($k > 1$) standard deviations, i.e., $k\sigma$, from the population mean μ.

The proof of Chebyshev's inequality is trivial. However, its concept is profound. First of all, it is important that Chebyshev's inequality holds good without any assumption regarding the shape of the distribution of population as long as the mean and variance exist. This nonparametric characteristic is just the one that was encountered and dealt with for fingerprint data distributed with respect to similarity scores generated by fingerprint-image matchers [3]. Second, Chebyshev's inequality provides a way to compute a quantitative relationship between an interval, which is greater than one standard deviation from the population mean, and the lower bound on the probability at which observed values of a random variable fall into that interval. In fact, there are many other implications of Chebyshev's inequality, which are out of the scope of this article.

On the other hand, Chebyshev's inequality cannot offer the lower bound of the proportion of the population that lies within one standard deviation or less from the population mean. Furthermore, for distributions that have a special shape, such as normal distribution, etc., the probability at which the population falls into an interval that is greater than one standard deviation from the population mean is much larger than the lower bound on the probability calculated using Chebyshev's inequality for the same size of interval. In other words, for instance, for the normal distribution, 95% of the population is within 1.96σ from the population mean. However, if the lower bound on the probability for any type of distribution is also set to be 95%, then the required interval provided by Chebyshev's inequality is 4.48σ from the population mean, which is about 2.29 times larger.

Such an interval that is 4.48σ from the population mean is defined as Chebyshev's greater-than-95% interval in this article. Chebyshev's greater-than-95% interval is different from a 95% confidence interval for an estimate of the population mean, which is a consequence of the Central Limit Theorem. Chebyshev's interval only describes the fact that the probability at which the population falls into an interval that is within 4.48σ from the population mean is greater than 95%, in spite of the shape of the population distribution. Therefore, no inferential statistics, such as hypothesis tests, etc., can be carried out based on Chebyshev's interval.

Chebyshev's greater-than-95% interval serves the objective that the result of taking *one* trial must be close to the baseline result within a desired tolerance at a specified probability. Since greater than 95% of population lies in Chebyshev's interval, assuming the interval contains the baseline result, the probability at which the one-trial test result falls in that interval and satisfies the requirement is greater than 95%. In addition, 95% for Chebyshev's interval is the lower bound on the probability. Thus, it provides conservative estimates in its applications.

In Chebyshev's greater-than-95% interval, the population mean μ and the population standard deviation σ are used. The sample mean $\hat{\mu} = \sum_{i=1}^{n} x_i / n$ and the sample standard deviation $\hat{\sigma} = \sqrt{\sum_{i=1}^{n} (x_i - \hat{\mu})^2 / (n-1)}$, where x_i's are independent observed values of a random variable ξ and n is the number of observations, are unbiased point estimators of μ and σ, respectively. However, according to the Law of Large Numbers, $\hat{\mu}$ and $\hat{\sigma}$ converge to μ and σ, respectively, as the number of observations increases [9]. While comparing the sample mean with the baseline result, the sampling error of the sample mean, i.e., the absolute value of the difference between $\hat{\mu}$ and μ, might not be needed to be taken into account in our application. This will be discussed in Section 4.

$4.48\hat{\sigma}$ is not a small quantity in many applications, and it could happen that $4.48\hat{\sigma}$ went beyond the allowed range of random variables. However, in our applications, thanks to the simple random sampling and the large size of fingerprint data, the sample standard deviation is very small (see Section 3.2). Therefore, the unbiased point estimator $4.48\hat{\sigma}$ can be used as a criterion to determine the sample size in biometric evaluation of fingerprint data.

2.2 SIMPLE RANDOM SAMPLING

Both match and non-match similarity scores will be referred to as similarity scores in this section. In order to test how far the number of similarity scores can be reduced with respect to the baseline, a simple random sample of similarity scores is selected from the finite set of similarity scores in the baseline. The simple random sampling (SRS) applied here is carried out under three assumptions: 1) the population is finite, 2) each member in the population has the same probability of being selected, and 3) it is a sampling without replacement (WOR) for members in the population.

The similarity scores can be represented as integers within different ranges, depending on different fingerprint-image matchers [3]. Let the integral score set be $\{s\} = \{s_{min}, s_{min}+1, \ldots, s_{max}\}$, consecutively from s_{min} to s_{max}, where s_{min} and s_{max} are the minimum and maximum similarity scores, respectively. Thus, the similarity score set is a set of integral scores,

$$S = \{ s_i \mid \forall\, i \in \{1, \ldots, N\}\} \qquad (3)$$

where $s_i \in \{s\}$ and N is the total number of similarity scores. The similarity scores s_i may not exhaust all members in the integral score set $\{s\}$. Moreover, some of the fingerprint-image comparisons may very well share the same integral value. Therefore, the similarity score set S can be partitioned into pairwise-disjoint subsets $\{S_s\}$. In each of the subsets, S_s, the members share the same integer $s \in \{s\}$. The similarity score set S is the union of all these subsets $\{S_s\}$.

The frequency f(s) of the similarity score s, which appears in the similarity score set S, is the size of the subset S_s. By applying the empirical distribution, i.e., putting probability 1/N on each of the observed scores in the similarity score set S, the corresponding probability p(s) equals the frequency f(s) divided by the total number of similarity scores, i.e., p(s) = f(s) / N. Thus, in the baseline, the empirical discrete probability distribution function of the similarity scores, by including zero probability caused by some similarity scores that appear in the integral score set $\{s\}$ but not in the similarity score set S, can be represented as

$$P = \{ p(s) \mid \forall\, s \in \{s\} \text{ and } \sum_{\tau=s\min}^{s\max} p(\tau) = 1 \} \qquad (4)$$

According to the SRS as stated above, each member in the similarity score set S with size N in the baseline has the same probability of being selected. Hence, the probability of being chosen for such a member is 1/N. Furthermore, the SRS is assumed to be WOR for scores in the similarity score set S. Thus, for any similarity score s in the integral score set $\{s\}$, whose frequency of appearing in the similarity score set S is f(s), the probability of being selected is f(s)/N. As a result, after sufficiently large amount of such selections, the discrete probability distribution function of the selected similarity scores will approach to the empirical discrete probability distribution function of the similarity scores in the baseline as expressed in Equation (4).

The size of the similarity scores is relatively large. Therefore, for a large amount of simple random samples, the variance of area under an ROC curve and even the variance of the TAR value at an operational FAR value, caused by the discrepancy between the distribution of the selected similarity scores and the empirical distribution in the baseline, are quite small. It follows that Chebyshev's greater-than-95% interval can be applied as a criterion to determine the sample size and is suitable

to serve our objectives. As for using the Kolmogorov-Smirnov Test to see the difference between such two distributions, it depends on how to deal with the ties of similarity scores in these two discrete probability distribution functions (see Section 3.1).

2.3 THE STABILITY METRIC

Chebyshev's greater-than-95% interval can be obtained using Monte Carlo calculation, i.e., by running a number of Monte Carlo iterations. How many iterations of the Monte Carlo calculation based upon SRS are needed to determine the sample size of similarity scores in the biometric evaluation of fingerprint data for a fingerprint-image matcher? In other words, how stable is the Monte Carlo calculation with respect to the number of iterations? The Monte Carlo stability is related to how much the sample size of similarity scores is, which fingerprint-image matcher is dealt with, and which criterion of evaluation of ROC curve is involved.

The discrete probability distribution functions of the selected similarity scores for the amount of sample sizes discussed in this article do not deviate very much from the empirical discrete probability distribution function in the baseline. Thus, Chebyshev's intervals always contain the baseline result as observed in our tests (see Section 3.2). Thereafter, the maximum of two distances between the baseline result and two end points of Chebyshev's greater-than-95% interval, respectively, can be chosen as a metric to measure the stability of our Monte Carlo calculation.

Such a stability metric can be expressed as

$$M_n = \max[b - (\hat{\mu}_n - 4.48\hat{\sigma}_n), (\hat{\mu}_n + 4.48\hat{\sigma}_n) - b] \qquad (5)$$

where M_n is the stability metric for n Monte Carlo iterations, b is the baseline result (either the area under an ROC curve or the TAR value at an operational FAR value in the baseline), and $\hat{\mu}_n$ and $\hat{\sigma}_n$ are the unbiased point estimators of the population mean μ and the population standard deviation σ after n iterations, respectively. And the population is determined by the sample size of similarity scores and a chosen fingerprint-image matcher. This stability metric describes the worst deviation of the one-trial test result from the baseline result inside Chebyshev's greater-than-95% interval. That is, with greater-than-95% probability, a one-trial test result will not deviate from the baseline result by more than the stability metric in a specified circumstance. Once the variation of the stability metric over the number of iterations is within an accepted tolerance, the stability of the Monte Carlo calculation is achieved.

3. RESULTS

Chebyshev's greater-than-95% intervals vary depending upon 1) the quality of the fingerprint-image matcher, 2) the criterion to evaluate ROC curves, 3) the sample size of SRS, and 4) the number of Monte Carlo iterations. Two matchers were taken as examples, among which Matcher 1 was high quality and Matcher 2 was low quality. Both matchers were executed on the same fingerprint dataset. Two criteria were employed, i.e., the area under an ROC curve (AUROC) and the TAR value at an operational FAR value that is set to be 0.001 (abbreviated as TVAFV).

3.1 THE DISCRETE PROBABILITY DISTRIBUTION FUNCTIONS [3]

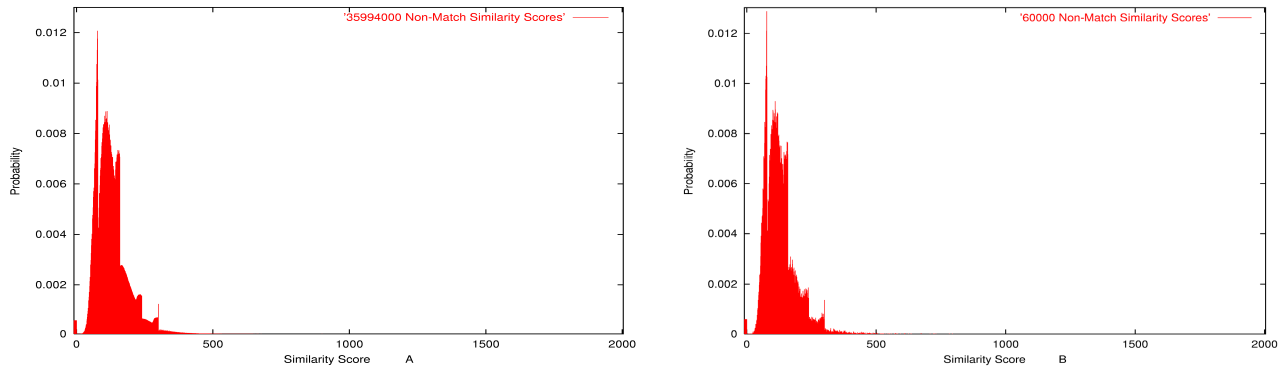

Figure 1 Matcher 1's discrete probability distribution functions of the 35,994,000 non-match similarity scores (A) and a simple random sample with 60,000 non-match similarity scores (B), respectively.

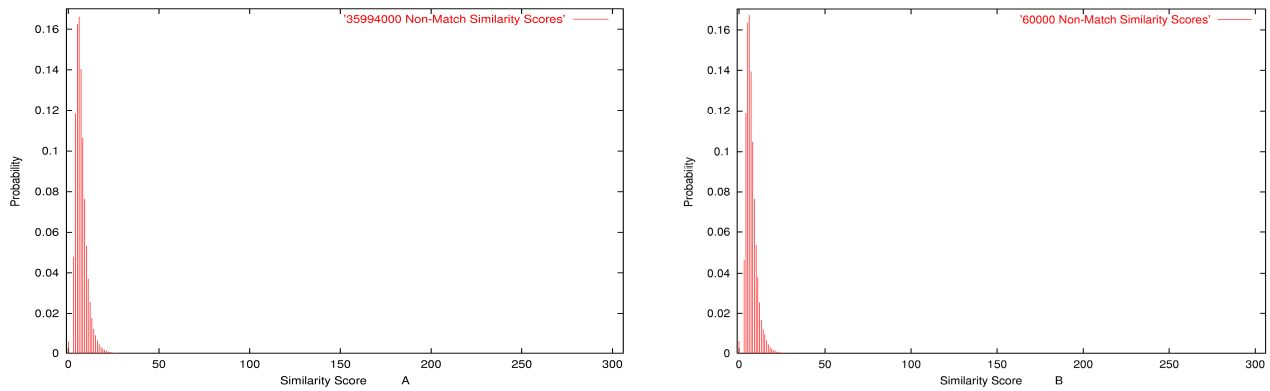

Figure 2 Matcher 2's discrete probability distribution functions of the 35,994,000 non-match similarity scores (A) and a simple random sample with 60,000 non-match similarity scores (B), respectively.

It is always a good thing to take a look at the distribution function first. For Matcher 1, the discrete probability distribution functions of the 35,994,000 non-match similarity scores in the baseline and the 60,000 non-match similarity scores that were a simple random sample selected from 35,994,000 non-match similarity scores are shown in Figure 1 A and B, respectively. Its integral similarity scores run from 0 through 2000. And for Matcher 2, the corresponding discrete probability distribution functions are presented in Figure 2 A and B, respectively. Its integral similarity scores run from 0 to 306. For Matcher 1 and 2, the distribution functions of 35,994,000 non-match similarity scores are completely different, as presented in Figure 1 A and Figure 2 A, respectively. Indeed, in many cases there is no parametric model to fit those distribution functions [3].

To explore the relationship between the distribution of simple random samples of non-match similarity scores and the distribution of 35,994,000 non-match similarity scores, the Kolmogorov-Smirnov Test was invoked. If a tie of non-match similarity scores in a discrete distribution was treated as separated individual scores that shared the same value while comparing two cumulative distribution functions [10], it was found that the one-tailed p-values of two distribution functions (i.e., 35,994,000 against 60,000 non-match similarity scores) for Matcher 1 and 2, respectively, were much less than 1%. This indicates that these two distributions are likely to be different.

However, if a tie of non-match similarity scores was dealt with as a single bar at the shared value of these scores, the one-tailed p-values were much larger than 5%. In this sense, these two distribution functions are unlikely to be different. This is why it is hard to see the difference between them visually. Indeed, such a treatment of ties matches the way of formation of ROC curve, which is generated by moving the threshold, one similarity score at a time, from the highest similarity score to the lowest similarity score [3]. Thus, the SRS has little impact on ROC curves, even when the sample size gets as small as 60,000. In other words, the variances of AUROC and TVAFV, caused by SRS, i.e., by the discrepancy between the distribution of the selected similarity scores and the distribution in the baseline, are so small that Chebyshev's greater-than-95% interval can be invoked.

3.2 CHEBYSHEV'S GREATER-THAN-95% INTERVAL AND SAMPLE SIZE

The results of fingerprint-image Matcher 1 and 2 are presented. The quality of Matcher 1 is higher than that of Matcher 2. Two criteria, AUROC and TVAFV, were used to evaluate the qualities of matchers. AUROC has a standard error associated with [3], but TVAFV does not. To be consistent between these two criteria as well as for simplicity, the standard error of AUROC is not used in this article. The baseline values of AUROC and TVAFV for Matcher 1 and 2 are shown in Table 1.

Matcher	AUROC	TVAFV
1	0.997170	0.991167
2	0.983862	0.892333

Table 1 The baseline results of Matcher 1 and 2.

Generally speaking, the smaller the sample size, the wider the Chebyshev's greater-than-95% interval. The error bars, i.e., $4.48\hat{\sigma}$, are relatively small for sample sizes greater than 100,000, and relatively large for those less than 10,000. Thus, results are presented with the sample sizes decreasing from 100,000 down to 10,000 by every 10,000 for both Matcher 1 and 2. The Monte Carlo calculation was run for 500 iterations in each case. Chebyshev's greater-than-95% interval is expressed in terms of sample mean, the error bar, the upper bound (sample mean plus error bar), and the lower bound (sample mean minus error bar). All numerical results are shown from Table 2 to Table 5. And the trend of variations of Chebyshev's greater-than-95% intervals and their relationship with the baseline results are depicted in Figure 3 and Figure 4.

As illustrated in the tables, if using the criterion of AUROC, the error bars monotonically increase from 0.000024 to 0.000070 while the sample sizes decrease from 100,000 down to 10,000 for high-quality Matcher 1, but from 0.000230 to 0.000716 for low-quality Matcher 2. If using the criterion of TVAFV, the error bars vary between 0.000301 and 0.001778 for Matcher 1, but between 0.012728 and 0.030589 for Matcher 2. As shown in the figures, all Chebyshev's greater-than-95% intervals contain the corresponding baseline results, and no upper bound exceeds one.

The error bars as well as the deviations between the sample means and the corresponding baseline results all depend on matchers' qualities and the criteria invoked for evaluation of ROC curve. For high-quality matcher as opposed to low-quality matcher, the error bars are smaller and the sample

means are closer to the corresponding baseline results. The higher the matcher's quality is, the more convergent the outcome is, therefore the less the variance will be. To reach the same error bar, the higher-quality matcher needs much smaller number of non-match similarity scores than the lower-quality matcher. The same relationship exists between AUROC and TVAFV. AUROC is taking the whole ROC curve into account [3], but TVAFV is only picking a TAR value on an ROC curve at an operational FAR value. Hence, TVAFV is more sensitive to SRS than AUROC.

Sample Size	100,000	90,000	80,000	70,000	60,000	50,000	40,000	30,000	20,000	10,000
Mean	0.997171	0.997170	0.997171	0.997170	0.997171	0.997171	0.997170	0.997170	0.997171	0.997171
Error Bar	0.000024	0.000025	0.000028	0.000030	0.000031	0.000033	0.000038	0.000042	0.000052	0.000070
Upper Bound	0.997194	0.997195	0.997198	0.997200	0.997201	0.997204	0.997207	0.997212	0.997223	0.997241
Lower Bound	0.997147	0.997146	0.997143	0.997140	0.997140	0.997137	0.997132	0.997129	0.997119	0.997101

Table 2 Matcher 1's Chebyshev's greater-than-95% intervals of 500 Monte Carlo iterations using AUROC.

Sample Size	100,000	90,000	80,000	70,000	60,000	50,000	40,000	30,000	20,000	10,000
Mean	0.991168	0.991171	0.991176	0.991178	0.991172	0.991193	0.991196	0.991204	0.991224	0.991257
Error Bar	0.000301	0.000325	0.000414	0.000434	0.000453	0.000609	0.000720	0.000898	0.001273	0.001778
Upper Bound	0.991469	0.991496	0.991590	0.991612	0.991625	0.991802	0.991916	0.992102	0.992498	0.993035
Lower Bound	0.990868	0.990846	0.990761	0.990744	0.990719	0.990584	0.990476	0.990305	0.989951	0.989478

Table 3 Matcher 1's Chebyshev's greater-than-95% intervals of 500 Monte Carlo iterations using TVAFV.

Sample Size	100,000	90,000	80,000	70,000	60,000	50,000	40,000	30,000	20,000	10,000
Mean	0.983857	0.983862	0.983862	0.983864	0.983863	0.983862	0.983871	0.983863	0.983857	0.983875
Error Bar	0.000230	0.000233	0.000242	0.000255	0.000281	0.000317	0.000344	0.000401	0.000509	0.000716
Upper Bound	0.984087	0.984095	0.984104	0.984119	0.984144	0.984179	0.984215	0.984263	0.984366	0.984592
Lower Bound	0.983627	0.983630	0.983621	0.983608	0.983582	0.983545	0.983527	0.983462	0.983348	0.983159

Table 4 Matcher 2's Chebyshev's greater-than-95% intervals of 500 Monte Carlo iterations using AUROC.

Sample Size	100,000	90,000	80,000	70,000	60,000	50,000	40,000	30,000	20,000	10,000
Mean	0.890305	0.890031	0.890005	0.889935	0.889876	0.889819	0.889999	0.889649	0.889493	0.889071
Error Bar	0.012728	0.013214	0.013112	0.013309	0.013486	0.014527	0.014990	0.016486	0.020722	0.030589
Upper Bound	0.903033	0.903245	0.903117	0.903243	0.903362	0.904346	0.904988	0.906135	0.910215	0.919660
Lower Bound	0.877578	0.876817	0.876894	0.876626	0.876390	0.875292	0.875009	0.873163	0.868771	0.858482

Table 5 Matcher 2's Chebyshev's greater-than-95% intervals of 500 Monte Carlo iterations using TVAFV.

The tolerances used to determine the sample size for high-quality matchers must be smaller than the ones for low-quality matchers, since the values of AUROC and TVAFV of high-quality matchers are very close to 1. Therefore, if invoking AUROC, 10,000 non-match similarity scores are enough for both Matcher 1 and 2, once the tolerances for Matcher 1 and 2 are set to be 0.0001 and 0.001, respectively. If using TVAFV, 30,000 non-match similarity scores are enough for both Matcher 1 and 2, while the tolerances for Matcher 1 and 2 are set to be 0.001 and 0.02, respectively.

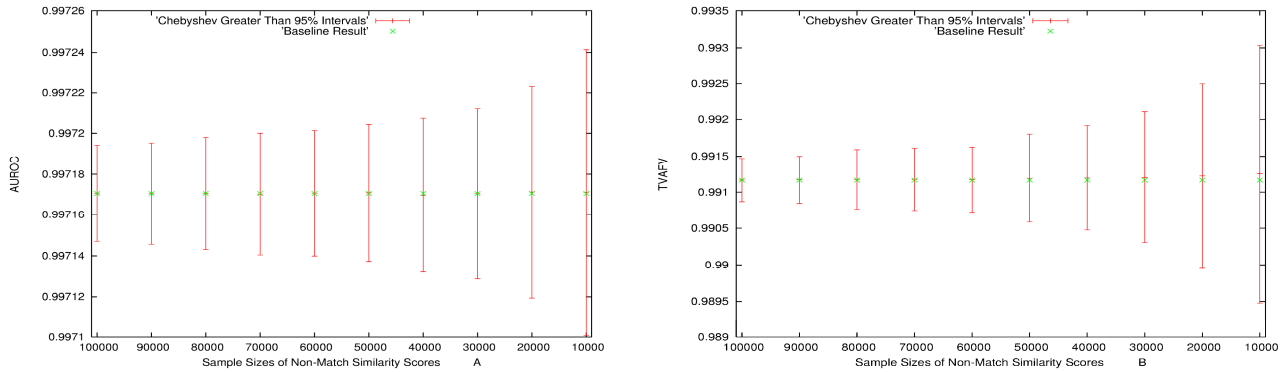

Figure 3 Matcher 1's Chebyshev's greater-than-95% intervals of 500 Monte Carlo iterations for different sample sizes and the baseline results using AUROC (A) and TVAFV (B), respectively.

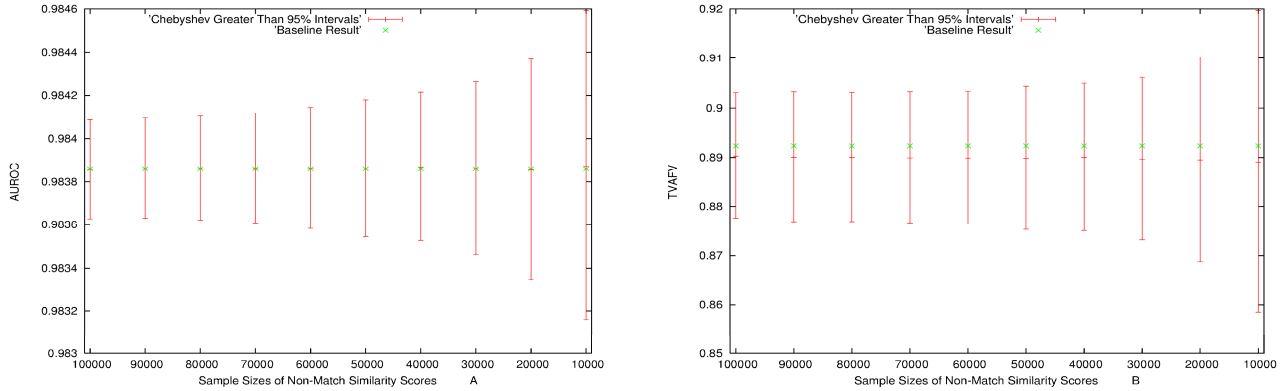

Figure 4 Matcher 2's Chebyshev's greater-than-95% intervals of 500 Monte Carlo iterations for different sample sizes and the baseline results using AUROC (A) and TVAFV (B), respectively.

As a result, the choice of sample size depends on the qualities of fingerprint-image matchers as well as on which criterion is invoked. To be more conservative, as well as to get balance among different qualities of matchers and between two different criteria of AUROC and TVAFV, in general, it seems that for 6,000 match similarity scores, 50,000 to 70,000 non-match similarity scores randomly selected from 35,994,000 non-match similarity scores are enough to ensure that the error bars of Chebyshev's intervals are within the accepted tolerance range with greater-than-95% probability.

3.3 THE STABILITY OF MONTE CARLO CALCULATION

The above results were derived from 500 iterations of Monte Carlo calculations. How stable are the results with respect to the number of Monte Carlo iterations? The smaller the sizes of simple random samples selected from 35,994,000 non-match similarity scores are, the larger deviation the distributions of selected non-match similarity scores have from the distribution in the baseline,

therefore the less stable the outcome is. Hence, the case of 10,000 non-match similarity scores is chosen to show the stability.

The stability metrics from 100 to 500 Monte Carlo iterations for sample size of 10,000 are presented in Table 6, for Matcher 1 and 2 as well as for two different criteria of AUROC and TVAFV, respectively. As expected, the stability metric of Matcher 1 is smaller than the one of Matcher 2 for a fixed criterion, and the stability metric of AUROC is smaller than the one of TVAFV for a specified matcher. This indicates again that the higher-quality matcher has less variance, and AUROC criterion is more convergent than TVAFV criterion.

Criterion	Matcher	Monte Carlo Iterations				
		100	200	300	400	500
AUROC	1	0.000083	0.000072	0.000068	0.000073	0.000071
	2	0.000689	0.000721	0.000669	0.000673	0.000730
TVAFV	1	0.001686	0.001738	0.001764	0.001910	0.001868
	2	0.035012	0.032547	0.035804	0.032913	0.033851

Table 6 The stability metrics of 10,000 non-match similarity scores.

In Table 6, it shows that the outcome of Monte Carlo calculation for 10,000 non-match similarity scores is very stable from 100 iterations up to 500 iterations with respect to specified matcher and criterion. The worst deviations of the one-trial test result from the baseline result with greater-than-95% probability vary by no-larger-than 0.000015, 0.000061, 0.000224, and 0.003257 (the maximum minus the minimum in each row of Table 6) from 100 to 500 iterations for Matcher 1 and 2 and for two different criteria, respectively, even when the sample size is chosen down to only 10,000 non-match similarity scores. As a consequence, the results presented above out of 500 Monte Carlo iterations are reliable.

4. DISCUSSION

The empirical approach of invoking Chebyshev's greater-than-95% interval in combination with simple random sampling serves our objective well. The result of taking one trial with reduced number of non-match similarity scores must be close to the baseline result. In terms of evaluation of ROC curves, that is, *Δ (ROC curve)* must be within an accepted tolerance with greater-than-95% probability. The half of Chebyshev's greater-than-95% interval is $4.48\hat{\sigma}$. And the margin of error of 95% confidence interval estimate of the population mean is $1.96\ \hat{\sigma}/\sqrt{n}$, where n is 500 in our case. The former is about 51 times larger than the latter. Therefore, the sampling error of the sample mean, namely, the absolute value of the difference between the unbiased point estimator of the population mean (i.e., the sample mean) and the population mean, is relatively negligible in each case, while using Chebyshev's greater-than-95% interval.

In this article, only the case is explored, in which the number of non-match similarity scores needs to be reduced. Indeed, the same technique can be applied to other scenarios of the biometric evaluation of fingerprint data, as long as the standard deviation of the population is small and the

objective is only taking one trial instead of taking average of many trials. However, the requirement that the standard deviation be small is the disadvantage of employing Chebyshev's inequality.

As has been demonstrated, the outcome is very much dependent on the qualities of fingerprint-image matchers. The higher-quality matchers are more convergent, thus have less variance than the lower-quality matchers. Hence, the higher-quality matchers need fewer number of non-match similarity scores than the lower-quality matchers in our application. Accordingly, the accepted tolerance is also dependent on the quality of matchers. Presented in this article are only two matchers, namely, Matcher 1 and 2. And Matcher 1's quality is higher than Matcher 2's. In our tests for this article, four fingerprint-image matchers were used, two of which were high-quality matchers, and the other two were low-quality matchers. They exhibited the similar behavior.

In biometric evaluation of fingerprint data, the sample sizes are also determined by other factors. For instance, if using the TVAFV criterion to evaluate ROC curve and setting the operational FAR value to be 0.001, for very high-quality fingerprint-image matchers, the TAR value could reach as high as 0.999. If there are only 6,000 match similarity scores, then the number of failures related to Type I error is only about 6, which is very much less significant. For such quality of matchers, in order to increase the significance of test, the number of match similarity scores must increase.

In conclusion, in the current framework of SDK tests, with respect to 6,000 match similarity scores, it seems that 35,994,000 non-match similarity scores are much more than what is needed. The number of non-match similarity scores can be dramatically reduced down to 50,000 to 70,000, as long as that amount of non-match similarity scores is a simple random sample of 35,994,000 non-match similarity scores. It holds good for different qualities of fingerprint-image matchers as well as for criteria of both AUROC and TVAFV. And it is valid with greater-than-95% probability.

REFERENCES

1. C.L. Wilson, *et al*., Fingerprint vendor technology evaluation 2003: summary of results and analysis report, NISTIR 7123, National Institute of Standards and Technology, June 2004.
2. C. Watson, C. Wilson, K. Marshall, M. Indovina, R. Snelick, Studies of one-to-one fingerprint matching with vendor SDK matchers, NISTIR 7119, National Institute of Standards and Technology, May 2004.
3. J.C. Wu, C.L. Wilson, Nonparametric Analysis of Fingerprint Data, NISTIR 7226, National Institute of Standards and Technology, May 2005.
4. S.D. Walter, The partial area under the summary ROC curve, Statist. Med. 24 (2005) 2025-2040.
5. G.S. Gazelle, P.M. McMahon, U. Siebert, M.T. Beinfeld, Cost-effectiveness analysis in the assessment of diagnostic imaging technologies, Radiology 235 (2005) 361-370.
6. J.A. Hanley, B.J. McNeil, A method of comparing the area under two ROC curves derived from the same cases, Radiology 148 (1983) 839-843.
7. J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29-36.
8. B. Ostle, L.C. Malone, Statistics in research: basic concepts and techniques for research workers, fourth ed., Iowa State University Press, Ames, 1988.
9. P.J. Bickel, K.A. Doksum, Mathematical statistics: basic ideas and selected topics, Holden-Day, Inc., San Francisco, 1977.
10. W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical recipes in C++: the art of scientific computing, second ed., Cambridge University Press, New York, 2002, pp. 647-648.