- **biometric identification: device specification and actual performance considered for the operations of the** UIDAI

○ *hans varghese mathews* *The Centre for Internet and Society, Bangalore*

○ **summary of findings**

A biometric device intended for the identification of individuals is supplied with a specified error rate: an experimental esimate of the probability that the biometrics of distinct individuals will match. When more than one device is used, and a suite of biometrics is to identify an individual, the chance of such identification errors can be derived from the specified error rates of the individual devices: and for the matching procedure the UIDAI is following we compute $0.115512486 \cdot 10^{-11}$ as the specified *identification error*.[1] Manufacturers estimate error rates under laborotary conditions: when they are used for the rapid identification of a large population, as in our case, their performance in the field might fall short of what their specified errors promise: and we find that

- **identification error in the field exceeds by a factor of 6 , almost, the specified identification error for the UIDAI's matching procedure.**

We are able to draw our conclusions by examining the result of an experiment performed by the UIDAI when 84 million citizens had been registered in their biometric database. The process of obtaining and storing biometrics is termed "enrollment" usually; and the stored suites of biometrics are called templates. The experiment estimated the chance of a false positive match: which occurs when the suite of biometrics of a new individual, one who is not actually enrolled, happens to match *some or other* stored template. The chance of a false positive match is the conditional probability, therefore, of a match occuring *given* that the individual is not enrolled: and it is called 'the false reject rate' usually. The rate depends on the number of individuals already enrolled. Write $\phi(n)$ for the false reject rate when n individuals have been enrolled: the identification error is the chance, now, that the biometrics of a new individual will match *any one* given template: and we have $\phi(n) = 1 - (1-\xi)^n$ when ξ is the identification error. Now for a specified identification error of $(0.1155001) \cdot 10^{-11}$ and an enrolled base of 84 million the false reject rate should be $(0.97020084) \cdot 10^{-4}$ at most: but the UIDAI got an estimate of $(0.57725) \cdot 10^{-3}$ from its experiment. The bound on this rate comes from the relation $1 - (1-\xi)^n \leq n \cdot \xi$, which holds for $0 < \xi < 1$ generally; from this, and from the relation of $\phi(n)$ to ξ just given, one can get the bounds

$$\frac{\phi(n)}{n} \quad \leq \quad \xi \quad \leq \quad \frac{-\log[1-\phi(n)]}{n} \tag{1}$$

We have estimated the identification error in the field by using the UIDAI's experimental value as a reliable operational estimate of $\phi(n)$ for $n = 84$ million; and $(0.687400801) \cdot 10^{-11}$ is our estimate of what ξ must be in the field.

[1] The UIDAI is the Unique Identity Authority of India. It is using iris scanners and fingerprint scanners: and has made their specified error rates available to researchers at the Takshashila Institute, who have made them public. The specified error for their make of iris scanner is repored as $1/13100$; the specified error for the fingerprint scanner as $1/500$. The UIDAI has not published its matching procedure: but our investigations have led us to conclude the following: a match is taken to occur if both irises match and *any one* digit also does.

The false reject rate is one measure of the operational accuracy, in the field, of a suite of biometric devices. A equally important measure is its converse: the conditional probability that an individual is not enrolled, actually, given a match between his or her biometrics and some or other stored template. We shall term this *mistaken identification*: and our principal finding is that

- **the probability of mistaken identification rises considerably between the initial and final stages of enrollment: by a factor of 10 almost between the first and last tenths of the population enrolled.**

We have proceeded here by estimating the total number of matches expected, and the number of false matches among these, for successive millions of individuals enrolled: for which we have used the lower of the bounds on $\phi(n)$ given by

$$n \cdot \xi \cdot \left[\frac{1-\xi}{1-\xi+n\cdot\xi} \right] \quad \leq \quad 1-(1-\xi)^n \quad \leq \quad n \cdot \xi \tag{2}$$

The actual numbers are not negligible. For example: the UIDAI should expect $534,010$ matches to occur for the first 100 million enrolled, out of which $34,180$ will be false matches; but a total of $1,280,208$ matches are expected for the last 100 million enrolled, and of these fully $780,382$ would be false matches.[2]

When a match occurs the UIDAI must decide whether or not the individual is already enrolled: for which the templates matching that person's suite of biometrics must be examined. The amount of work here depends on how many templates will match a given suite of biometrics, generally, when a match does occcur. We get an upper bound of $10,922,437$ on the total number of matches when the entire population of 1.2 billion has been enrolled: of which $4,924,539$ would be false ones. But we estimate that only $11,267,203$ matching templates will have to be examined, at most, to decide which of these matches are false: and our last finding is that

- **only rarely will more than one matching template have to be examined, when a match occurs, to see if the match is a false one.**

Suppose n many are enrolled and ξ is the identification error: the computation requires an upper bound on the probability $\psi_q(n)$ of finding q or more matching templates, for any $q > 0$ now, should a false positive match occur: and we use

$$\psi_q(n) \quad = \quad 1 - \sum_{r=0}^{q-1} \binom{n}{r} \xi^r (1-\xi)^{n-r} \quad \leq \quad \frac{\xi^q}{(q-1)!} \prod_{r=0}^{q-1} (n-r) \tag{3}$$

We have $\psi_1(n) = \phi(n)$ of course; and the standard identification of $\psi_q(n)$ with the value $I_\xi(q, n-q+1)$ of the Incomplete Beta Function yields this bound.

---

[2] The discrepancy is even more extreme for small initial and final subsets: we estimate $50,325$ matches for the first 10 million, of which only 341 would be false ones; but $131,050$ matches are expected for the last 10 million, out of which $8,1607$ would be false matches. The bounds on $\phi(n)$ come from Professor Nico Temme of the CWI in The Netherlands: whose freely given help we gratefully acknowledge. To counts matches and mistaken matches one needs, besides identification error, the probability that enrolled individuals will try to register again; and one needs, as well, the chance of a match for an already enrolled person. The UIDAI has conducted an experiment which allows one to estimate the latter; and it has estimated to its satisfaction the former probability as well.

1 A biometric is a numerized representation of some generic physical or physiological feature of an organism: for precision and brevity we shall call such a feature an *organic object*: using the word "object" in a less than daily way. The numerized representation is typically a real or binary vector: which must be of suitably large dimension if the biometric is to be used for *the identification of organic individuals.* We shall only be considering such identificatory biometrics. A device or arrangement for obtaining these numerized representations will *scan* the organic object in some way: and it is important to keep in mind that *a biometric is always the output that a particular device produces* when it is given as input an organic object of the sort it was designed to scan. The crucial circumstances now are:

a the numerized representations obtained from any two scans of the same organic object are almost never *precisely* the same

b specifying *in advance* how the numerized representations of different organic objects will themselves differ, with any exactitude, seems impossible

These inconvenient imprecisions arise in two ways: from unavoidable variations in the physical process of scanning, first, and then from the peculiarities of whatever algorithm converts the 'signal' produced by the scanning into a numerized representation. The circumstance (**a**) allows the 'identification error' already noted: because the numerized representations of two distinct organic objects may differ only as much as the representations produced by different scannings of one or other of those objects. The output of a particular scanning of an object may sometimes appreciably differ, as well, from some *identifying representation* produced by a prior scanning of that object: which may occasion what one might call *verification error*: and we shall specify both identification and verification error more precisely in a moment. So (**a**) necessitates the measurement of similarity and difference: one has to decide how similar two biometrics must be in order to be accounted numerized representations of the same organic object: and how different they must be, conversely, to be accounted representations of different organic objects. The cirumstance (**b**) ensures that these difficulties can be met in an empirical way only: measures or means to decide similarity and difference for the biometrics produced by a particular device can be obtained only by sampling its output.

One usually proceeds here by deciding on some suitable distance between the numerized representations that a device produces: upon which certain distributions of these distances are experimentally estimated. Let $X^{\mathfrak{d}}$ denote the random variable whose values are distances between numerized representations of distinct organic objects: of which a random sample of $\binom{n}{2}$ values may be obtained from n objects by scanning each once. Let $X^{\mathfrak{s}}$ denote the variable whose values are distances between different numerized representations of the same organic object: of which n values may be obtained by scanning each object twice, though special care may have to be taken to ensure that the sample is random. The primary requirement on the design of the device may now be stated thus: one must be able to find a number τ such that both $p[X^{\mathfrak{d}} < \tau]$ and $p[X^{\mathfrak{s}} > \tau]$ are miniscule: the chance that a value of $X^{\mathfrak{d}}$ lies below τ, on the one hand, and the chance that a value of $X^{\mathfrak{s}}$ lies above τ, on the other, must both be miniscule.

We emphasize that the variables $X^{\mathfrak{d}}$ and $X^{\mathfrak{s}}$ depend on the device that produces the biometrics. Let $f^{\mathfrak{d}}$ and $f^{\mathfrak{s}}$ denote their distribution functions. The decisive number τ above is called an *error threshold*: and experimental estimates of $f^{\mathfrak{d}}$

and $f^{\mathfrak{s}}$ must allow a ready choice of threshold. These estimates must allow the secure estimation of the probabilities $p[X^{\mathfrak{d}} < \tau]$ and $p[X^{\mathfrak{s}} > \tau]$ as well: the first of these is what we had called the *specified identification error* for the device: and *specified verification error* seems a good name for the second. Biometrics produced by a device are said to *match* if the distance between them falls below a chosen threshold: and *falsely match* if they are numerized representations of distinct organic objects. We shall sometimes say that the objects themselves match or falsely match. We shall say "identification error" for the chance of a false match, misusing words for the sake of brevity: and the specified identification error for a device is its manufacturer's estimate of the probability of a false match: for a specifed threshold of course. Verification error will not figure a very great deal in what follows: but we note that it is usually called the probability of a *false non-match.*

For numerized representations that are real vectors the most common measure of difference would be Euclidean distance: or the *Mahalanobis distance* if the vectors may be taken for values of a multivariate normal distribution, whose covariance matrix can be reliably estimated. When the biometrics are binary vectors a common measure would be the Hamming distance: provided the vectors are not sparse: which would most likely be the case.

One might ask how licit it is to treat such distances as values of a random variable: arising as they do from a particular device. The justification presumably is that the design of the device is detailed enough to produce others which are operationally identical to it: and so the device may regarded as a paticular 'black box' among many such, each of which will produce like outputs for like inputs. Write $p^{\mathfrak{s}}_{\tau}$ for $p[X^{\mathfrak{s}} > \tau]$ and $p^{\mathfrak{d}}_{\tau}$ for $p[X^{\mathfrak{d}} < \tau]$ now; taking $f^{\mathfrak{s}}$ and $f^{\mathfrak{d}}$ for integrable functions we would have $p^{\mathfrak{s}}_{\tau} = \int_{\tau}^{\infty} f^{\mathfrak{s}}(x)\,dx$ and $p^{\mathfrak{d}}_{\tau} = \int_{-\infty}^{\tau} f^{\mathfrak{d}}(x)\,dx$. We note that $X^{\mathfrak{s}}$ is kin to the *genuine* distribution of [1]; and $X^{\mathfrak{d}}$ akin to the *inter-template* distribution there. The graphs $f^{\mathfrak{s}}$ and $f^{\mathfrak{d}}$ will look peaked, ideally, around means or primary modes that are widely separated, given their respective variances: look at Slide 49 in [2]. One expects that the mass of $f^{\mathfrak{s}}$ will be concentrated around a mode far to the left of where the mass of $f^{\mathfrak{d}}$ piles likewise: and the threshold τ would lie many standard deviations to the right and to the left, respectively, of the primary mode of $f^{\mathfrak{s}}$ and the primary mode of $f^{\mathfrak{d}}$. The choice of a threshold τ, and the determining of the associated identification and verification errors, is complicated affair in practice: see [1]. Moving τ to the left will decrease identification error while increasing verification error: and moving it right will increase the first and decrease the second. Plotting estimates of $p^{\mathfrak{d}}_{\tau}$ against estimates of $p^{\mathfrak{s}}_{\tau}$, for varying values of the threshold τ, gives a curve called the *receiver operating characteristic* for the device: which would under ideal circumstances display how changes in either identification error or verification error will inversely affect the other.

We must note that "identification error" and "verification error" are not standard ways of referring to the probabilities they denote. They are often referred to as 'error crossover rates': and the first, which we shall most be concerned with, is sometimes called the chance of a *false positive match* for the device in question: or the *false reject rate* for the device. But these latter terms find their proper use in the context of creating and maintiaining biometric databases: as we shall shortly see.

We confine ourselves from here on to the biometric identification of human individuals. Suppose now that more than one physical or physiological feature will used to do so. Let $\{\mathbf{F}_1, \mathbf{F}_2, \ldots \mathbf{F}_K\}$ be the features whose numerized representations are to provide a suite of identificatory biometrics for an individual; and suppose further that, for some considerable population of individuals, one such suite of identifying representations is to be stored, for each individual, in some biometric database. The process of gathering and storing identificatory biometrics is termed *enrollment* usually: and these stored representations are called *templates.* So in the situation we envisage there are *keepers of identity*: let us, melodramatically enough, style so the administrators of the biometric database: and they are charged with obtain-

ing and storing, for each individual in the population, a suite $\{\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_K\}$ of templates, where each $\mathbf{b}_k$ is a numerized representation of the feature $\mathbf{F}_k$ now, for $k \in \{1, 2, \ldots K\}$. There will be a device $\mathbf{M}_k$ to scan $\mathbf{F}_k$: whose manufacturer will have supplied the keepers of identity a crossover threshold τ_k for the device, as well as the specified errors ρ_k and ν_k of identification and verification, respectively, which are associated with that threshold. The threshold and associated errors will have been obtained by estimating the distributions of random variables $X_k^{\mathfrak{d}}$ and $X_k^{\mathfrak{s}}$ which are the counterparts of $X^{\mathfrak{d}}$ and $X^{\mathfrak{s}}$ above: and ρ_k is the manufacturer's estimate of the probability $p[X_k^{\mathfrak{d}} < \tau_k]$ now, while ν_k is the like estimate of the probability $p[X_k^{\mathfrak{s}} > \tau_k]$.

In the typical situation of enrollment the keepers of identity are facing a putatively new enrollee: a person whose biometrics are not yet in their database. Let us call this individual S. Numerized representations $\{\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_K\}$ of the features $\{\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_K\}$ are obtained from S now: and before they are entered in the database as his or her suite of biometric templates they must be compared with *every other* suite of templates already in the database, of course, because each suite is meant to identify one and only one individual. So suppose that some n many persons have already been enrolled; and let $\{P_1, P_2, \ldots P_n\}$ be some listing of these. For each index i in $\{1, 2, \ldots, n\}$ let $\{\mathbf{b}_{1,i}, \mathbf{b}_{2,i}, \ldots, \mathbf{b}_{K,i}\}$ be the templates of the person P_i. We shall suppose now that each suite of biometrics in the database does correspond to a distinct individual: we suppose that there are no duplications, that is, in the database. For each k in $\{1, 2, \ldots, K\}$ let $x_{k,i}$ be the distance between $\mathbf{v}_k$ and $\mathbf{b}_{k,i}$ now: assuming that S is not already enrolled, the number $x_{k,i}$ can be taken for a value of the random variable $X_k^{\mathfrak{d}}$: and it does seem licit, assuming this, to regard the collection $\{x_{k,1}, x_{k,2}, \ldots, x_{k,i}, \ldots, x_{k,n}\}$ as a random sample of values drawn from $X_k^{\mathfrak{d}}$.

As ρ_k estimates the chance that any given value of $X_k^{\mathfrak{d}}$ will fall below τ_k we may expect that $\rho_k \cdot n$ values from the sample will fall below this threshold: assuming, of course, that S is not actually enrolled. For each k in $\{1, 2, \ldots K\}$ let $\Gamma_k[S]$ denote the subset of those enrolled persons for whom such falling below the corresponding threshold happens: P_i is in $\Gamma_k[S]$ when $x_{k,i} < \tau_k$ that is to say. So the chance of finding any given enrolled person in $\Gamma_k[S]$ is ρ_k now; and when n is suitably large — large enough for each quantity $\rho_k \cdot n$ to appreciably exceed 1, for instance — we must expect that none of the sets $\Gamma_k[S]$ will be empty. That is no reason to suppose, of course, that any one person will be found in each of these sets: it may be that their intersection $\Gamma[S] \equiv \bigcap_k \Gamma_k[S]$ is empty. But if we assume that each variable $X_j^{\mathfrak{d}}$ is independent of every other $X_k^{\mathfrak{d}}$ now, then for any enrolled person P_i we can estimate the probability that he or she will be found in each and every $\Gamma_k[S]$: that will happen only if $x_{k,i} < \tau_k$ for every index k in $\{1, 2, \ldots, K\}$ of course, so the product $\rho_1 \rho_2 \cdots \rho_K$ estimates that probability.

Suppose next that finding any P_i at all in $\Gamma[S]$ provides grounds enough for the keepers of identity to doubt that S is not already enrolled: the circumstance that $\Gamma[S]$ is not empty, then, when S is actually not enrolled, is usually termed a *false postive match.* Let us say that S and an enrolled P_i are *matched at* the feature $\mathbf{F}_k$ if $x_{k,i} < \tau_k$: the condition of finding of an enrolled P_i in $\Gamma[S]$ specfies the *matching procedure* followed by the keepers of identity: who take a match to occur between an enrollee and any given enrolled person just in case they match at every numerized feature. Set $\xi = \rho_1 \rho_2 \cdots \rho_K$ now: we shall call this the *specified*

identification error for the matching proedure: which, to make it entirely explicit, is the estimated probability that a match will occur between an unenrolled individual and *any given* enrolled person. ξ is a prior estimate of that probability, of course, got from the errors specified by their manufacturers for the devices being used: the conditions of their use in the process of enrollment, *in the field* as it were, may differ from the laborotary conditions of their testing. We shall return to the question.

On our terms here the chance of a false positive match is the probability that $\Gamma[S]$ does not turn out empty even though S is not enrolled: and we can make a prior estimate of this latter chance in terms of the identification error ξ for the matching procedure. As ξ estimates the probability that any given enrolled person is found in the set $\Gamma[S]$, we may take $1-\xi$ for the complementary probability that a given enrolled person *is not found* in $\Gamma[S]$. It seems licit to assume the following: whether or not any one enrolled person will be found in $\Gamma[S]$ is independent of whether or not any other enrolled person will be found there: so when n many distinct persons have already been enrolled we have $(1-\xi)^n$ estimating the chance that $\Gamma[S]$ is empty: and so the probability that $\Gamma[S]$ is not empty, which is the chance of a false positive match here, is the quantity $1-(1-\xi)^n$.

Since $0<\xi<1$ we have $0<1-\xi<1$ as well, so $(1-\xi)^n$ decreases as n increases: and so the false reject rate will rise as enrollment proceeds. The increase in the false reject rate keeps pace with enrollment: we have

$$(i) \qquad 1-(1-\xi)^n \;=\; n\cdot\int_0^{\xi}(1-t)^{n-1}\,dt \;\leq\; n\cdot\int_0^{\xi}dt \;=\; n\cdot\xi$$

actually, as $0<(1-t)\leq 1$ for $0\leq t\leq\xi<1$. Since it depends thus on the number enrolled we shall write $\phi(n)\equiv 1-(1-\xi)^n$ for what the false reject rate comes to when n many have been enrolled. To save writing in what follows we shall simply say *a match occurs* when the biometrics of an enrollee match some stored template. So $\phi(n)$ is the probabililty that a false positive match will occur for the next individual S who is to be enrolled: the chance that, even though he or she is not enrolled, the biometrics of S will match the templates of *some or other* enrolled person P.

That the chance of a false postive match remain low is an evident desideratum. That that enrolled persons trying to enroll again should be detected would be equally important: and we must consider next the chance that an enrolled indvidual will succeed in enrolling again. The probability of such an event is called the *false accept rate* usually. Let us write α simply for the false accept rate: for the keepers of identity may expect this rate to depend on the specified verifications errors ν_k only: or on what the corresponding errors will be in the field, more properly: but let us suppose for the moment that their devices perform as they are expected to.

Let $\mathcal{D}$ denote the biometric database, for brevity, and suppose next that S is an already enrolled person seeking to enroll again: using an alias presumably. The templates taken at his or her prior enrollment will have created an *avatar* $P[S]$ for S in $\mathcal{D}$ let us say: a different virtual or spectral person, from the perspective of the keepers of identity: and ν_k is their estimate, provided by the manufacturer of $\mathbf{M}_k$, of the chance that S and this spectral $P[S]$ are *not matched at* the feature $\mathbf{F}_k$. The probability that enrolled persons and their avatars do match at $\mathbf{F}_k$ should be estimated as $1-\nu_k$ then. Write ν now for the probability that *no match occurs* between an enrolled person P and his or her own avatar $P[S]$. Given the

matching procedure they are following, and assuming again that matching at one feature is independent of matching at any other feature, the keepers of identity should estimate

$$(ii) \qquad 1 - \nu \;=\; \prod_{k=1}^{K} [1 - \nu_k]$$

as the chance that a match *will occur* between an already enrolled person and his or her own avatar in $\mathcal{D}$: because a match is taken to occur only when matches occur at every feature. Note that whether or not matchings at distinct features are independent, this complementary probability — which is the probability that *an attempt to enroll again will be detected* — actually decreases when more than one device is used: but if K is a small count and if each ν_k is a miniscule quantity the decrease should be negligible. In the mean we have

$$(iii) \qquad \nu \;=\; 1 - \textstyle\prod_k [1 - \nu_k]$$

$$(iv) \qquad 1 - \alpha \;\geq\; 1 - \nu$$

The inequality (iii) only restates (ii) of course. The inequality (iv) obtains because the probability of a match occurring between an already enrolled applicant S and *some or other* enrolled P is at least as great as the chance of a match occurring between S and his or her own avatar $P[S]$: and so we finally get the upper bound

$$(v) \qquad \alpha \;\leq\; \nu \;=\; 1 - \textstyle\prod_k [1 - \nu_k]$$

We record once again that our bounds on the false reject and false accept rates derive from assuming that matching at any one feature $\mathbf{F}_k$ is independent of matching at any other feature $\mathbf{F}_j$: a circumstance we shall call *the independence of the metrized features.* Assuming that such independence obtains amounts to assuming that $\{X_k^{\mathfrak{d}}\}_{k=1}^K$ is a collection of independent random variables; and $\{X_k^{\mathfrak{s}}\}_{k=1}^K$ likewise; and we note that such assumptions are generally made.

Consider next the converse of a false postive match: the circumstance that an individual is not enrolled, in fact, though a match does occur for him or her. We shall cast matters in the usual language of probability now. Suppose that a number of persons have already been enrolled, and let S be the next individual presented to the keepers of identity. Let A now denote the circumstance that S is already enrolled, and A^c the complementary circumstance that he or she is not. Let B denote the circumstance that a match occurs for S, and B^c the complementary circumstance of no match occurring. The false reject rate is the conditional probability $p[B|A^c]$ then: the chance that a match occurs given that S is not actually enrolled. The false accept rate is the conditional probability $p[B^c|A]$: the probability that no match occurs given that S is already enrolled.

Now $p[A^c|B]$ is the probability that S is not enrolled, actually, given that a match occurs: the chance of what was termed *mistaken identification* in the introductory summary of findings. Write $p(E)$ for the probability of any event or circumstance E ; write $p(E\,\&\,F)$ for the probability of the conjunction of events or circumstances E and F ; the formula

$$(vi) \qquad p[A^c|B] \cdot p(B) \;=\; p(A^c\,\&\,B) \;=\; p[B|A^c] \cdot p(A^c)$$

relates the probability of mistaken identification to the false reject rate: so we could estimate the former from the latter if we could estimate $p(A^c)$ and $p(B)$. Let us suppose now that S is a randomly selected individual. The probabilities $p(A)$ and $p(A^c) = 1 - p(A)$ may be taken generally, now, to depend on the compulsions of individuals in the population that is being enrolled. Estimating them would, nonetheless, be very risky. But suppose the population is very large: as it is in our case, where UIDAI is engaged upon biometrically identifying every resident of India. Suppose as well that, when a substantial number have been enrolled, the keepers of identity may reasonably take themselves to have sampled the population randomly: they may estimate $p(A)$ then: provided that they have been very successful in detecting attempts to enroll more than once. The personnel of the UIDAI have been bold enough to do so in fact: as we shall momentarily see. Now $p(B)$ is the probability of a match occurring regardless of whether or not S is already enrolled: and the formula

$$(vii) \qquad \begin{aligned} p(B) &= p[B|A^c] \cdot p(A^c) + p[B|A] \cdot p(A) \\ &= p[B|A^c] \cdot p(A^c) + (1 - p[B^c|A]) \cdot p(A) \end{aligned}$$

computes the chance of a match occurring for S from the false reject rate and the false accept rate, and from the general probability that an enrolled individual will try to enroll again. The specified verification errors yield the upper bound on the false accept rate given in (iv) above: from which the keepers of identity may expect that these probabilities will not change appreciably as enrollment proceeds. So when a substantial number have been enrolled they could try to estimate, with a suitable experiment, what the false accept rate is for their matching procedure: and the personnel of the UIDAI have done so in fact.

In the following section, where we consider the operations of the UIDAI in detail, we shall be using their experimentally obtained estimates of $p(A)$ and $p[B^c|A]$ to estimate the chance of mistaken identification. To do so accurately we require an estimate of what the false reject rate will be in the field: which depends the identification error, in the field, for the matching procedure they are following. As we mentioned our introductory summary, after enrolling a substantial number the UIDAI had conducted an experiment to estimate the false reject rate. We use that to estimate identification error in the field; and we end this preparatory section by deriving the bounds given by the formulae (1) and (2) and (3) in the summary of findings.

The inequality in (i) provides the lower bound in (1) and the upper bound in (2) already. To obtain the upper bound in (1) note first that $1 - \phi(n) = (1 - \xi)^n$ by definition, which gives $\log[1 - \phi(n)] = n \cdot \log(1 - \xi)$; so we must relate ξ and $\log(1 - \xi)$ to proceed. For $0 < x < 1$ generally we have

$$\log(1 - x) = -\int \frac{dx}{1 - x} = -\int (1 + x + x^2 + \ldots)\, dx = -\left(x + \frac{x^2}{2} + \frac{x^3}{3} + \ldots\right)$$

since the series $\sum_{r=0}^{\infty} x^r$ converges absolutely. It is immediate that $\log(1 - x) < -x$ then: and hence $x < -\log(1 - x)$. So we have

$$\xi \leq \log(1 - \xi) = \frac{\log[1 - \phi(n)]}{n}$$

yielding the upper bound for (1). We get a lower bound on $\log(1-x)$ from its expression as the series above: simply note that

$$(x + x^2/2 + x^3/3 + \ ...\ ) \ < \ (x + x^2 + x^3 + \ ..\ ) \ = \ x(1 + x + x^2\ ...\ ) = x/(1-x)$$

which gives us $-x/(1-x) < \log(1-x)$: and the bounds $x < -\log(1-x) < x/(1-x)$ will help obtain the lower bound in (2). For $0 < x < 1$ and positive integers n we now have

$$-n\log(1-x) \ < \ nx/(1-x)\ ; \ \ 1 - n\log(1-x) \ < \ (1 - x + nx)/(1-x)$$
$$nx \ < \ -n\log(1-x)$$

from these bounds on $-\log(1-x)$; and these together yield

$$nx \cdot \left[\frac{1-x}{1-x+nx}\right] \ < \ \frac{-n\log(1-x)}{1-n\log(1-x)} \tag{1.1}$$

Let $y > 0$ next; to proceed we must take a detour and note that from $1 + y < e^y$ we get

$$e^{-y} + ye^{-y} \ < \ e^y e^{-y} \ = \ 1$$
$$0 \ < \ 1 - e^{-y} - y \cdot e^{-y}$$
$$y \ < \ 1 + y - e^{-y} - y \cdot e^{-y} \ = \ (1+y) \cdot (1 - e^{-y})$$
$$\frac{y}{1+y} \ < \ 1 - e^{-y} \tag{1.2}$$

For $0 < x < 1$ set $t = -\log(1-x)$; we have $x = 1 - e^{-t}$ and $e^{-t} = 1 - x$ then, whence $e^{-nt} = (1-x)^n$ and $nt = -n\log(1-x)$; then for positive integers n we get

$$nx \cdot \left[\frac{1-x}{1-x+nx}\right] \ < \ \frac{-n\log(1-x)}{1-n\log(1-x)} \ = \ \frac{nt}{1+nt} \ < \ 1 - e^{-nt} \ = \ 1 - (1-x)^n$$

from (1.1) and (1.2) just above, as we need for the lower bound in (2). For the upper bound we have in (3) we must consider the *Incomplete Beta Function.* Set

$$\mathcal{B}_x(a,b) \ \equiv \ \int_0^x t^{a-1}(1-t)^{b-1}dt$$

for arguments a, b and any $0 < x \leq 1$ first; then $\mathcal{I}_x(a,b) \equiv \mathcal{B}_x(a,b)/\mathcal{B}_1(a,b)$ defines the Incomplete Beta Function for these arguments. It is usual to write $\mathcal{B}_1(a,b)$ as $\mathcal{B}(a,b)$ simply. Elementary integration, by parts, will give us

$$\int_0^x t^k(1-t)^{n-k-1}dt \ = \ \frac{-t^k(1-t)^{n-k}}{n-k}\Big|_0^x \ + \ \left[\frac{k}{n-k}\right] \cdot \int_0^x t^{k-1}(1-t)^{n-k}dt \tag{1.3}$$

Setting $x = 1$ here yields the relation

$$\mathcal{B}(k+1, n-k) \ = \ [k/(n-k)] \cdot \mathcal{B}(k, n-k+1)\,; \tag{1.4}$$

we have $\mathcal{B}(1,n) = \int_0^1 (1-t)^{n-1}dt \ = \ \frac{-(1-t)^n}{n}\Big|_0^1 \ = \ \frac{1}{n}$ to begin with; so iterating (1.4) gives

$$\mathcal{B}(k+1, n-k) \ = \ \frac{k \cdot (k-1) \cdot \ ... \ \cdot 2 \cdot 1}{(n-k) \cdot (n-(k-1)) \cdot \ ... \ \cdot (n-1) \cdot n} \ = \ \left[(n-k) \cdot \binom{n}{k}\right]^{-1} \tag{1.5}$$

Write $\mathcal{B}(a,b)$ as $\mathcal{B}_a^b$ for the moment, to save space; upon evaluating the first term on its right the equation (1.3) gives us

$$\mathcal{B}_{k+1}^{n-k} \cdot \mathcal{I}_x(k+1, n-k) \ = \ \frac{-x^k(1-x)^{n-k}}{n-k} \ + \ \left[\frac{k \cdot \mathcal{B}_k^{n-k+1}}{n-k}\right] \cdot \mathcal{I}_x(k, n-k+1)$$
$$= \ \mathcal{B}_{k+1}^{n-k} \cdot \left[-\binom{n}{k}x^k(1-x)^{n-k} \ + \ \mathcal{I}_x(k, n-k+1)\right]$$
$$\mathcal{I}_x(k+1, n-k) \ = \ \mathcal{I}_x(k, n-k+1) \ - \ \binom{n}{k}x^k(1-x)^{n-k} \tag{1.6}$$

since $1/(n-k) = \binom{n}{k} \cdot \mathcal{B}_{k+1}^{n-k}$ and $[k/(n-k)] \cdot \mathcal{B}_k^{n-k+1} = \mathcal{B}_{k+1}^{n-k}$ from (1.4) and (1.5) respectively. It is immediate from the computation of $\mathcal{B}(1,n)$ that $\mathcal{I}_x(1,n) = 1 - [1-x]^n$; and (1.6) yields

$$\mathcal{I}_x(q, n-q+1) \;=\; 1 \;-\; \sum_{j\,=\,0}^{q-1} \binom{n}{j} \xi^j \cdot (1-\xi)^{n-j}$$

by induction then. Now to obtain the upper bound in (3) we need only note that

$$\mathcal{B}_q^{n-q+1} \cdot \mathcal{I}_x(q, n-q+1) \;=\; \int_0^x t^{q-1}(1-t)^{n-q}\, dt \;\;\leq\;\; x^{q-1} \int_0^x (1-t)^{n-q}\, dt$$

generally; then, since $(1-t) \leq 1$ when $0 \leq t \leq \xi < 1$, as we have here, we finally get

$$\begin{aligned}
\mathcal{B}_q^{n-q+1} \cdot \mathcal{I}_\xi(q, n-q+1) \;&\leq\; \xi^{q-1} \int_0^\xi dt \;\leq\; \xi^q \\
\mathcal{I}_\xi(q, n-q+1) \;&\leq\; \xi^q / \mathcal{B}_q^{n-q+1} \\
&=\; \frac{\xi^q \cdot [n-(q-1)] \cdot [n-(q-2)] \cdots [n-1] \cdot n}{(q-1) \cdot (q-2) \cdots 2 \cdot 1} \qquad \textit{cf.}\ (1.5) \\
&=\; \frac{\xi^q}{(q-1)!} \prod_{r\,=\,0}^{q-1} (n-r)
\end{aligned}$$

We do not need lower bounds on $\mathcal{I}_\xi(q, n-q+1)$ when $q > 1$: but we give them for completeness. Note that $(1-\xi)^k \geq 1 - k \cdot \xi$ by (2) first; as $1 - t \geq 1 - \xi$ for $t < \xi$ we then have

$$\begin{aligned}
\mathcal{B}_q^{n-q+1} \cdot \mathcal{I}_\xi(q, n-q+1) \;&\geq\; (1-\xi)^{n-q} \int_0^\xi t^{q-1} dt \;=\; (1-\xi)^{n-q} \cdot \xi^q / q \\
\mathcal{I}_\xi(q, n-q+1) \;&\geq\; \frac{(1-\xi)^{n-q} \cdot \xi^q}{q \cdot \mathcal{B}_q^{n-q+1}} \;=\; \left[\frac{(1-\xi)^{n-q}}{q}\right] \cdot \left[\frac{\xi^q}{(q-1)!} \prod_{r\,=\,0}^{q-1} (n-r)\right] \\
&\geq\; \frac{(1 \,-\, n \cdot \xi \,+\, q \cdot \xi) \cdot \xi^q}{q!} \prod_{r\,=\,0}^{q-1} (n-r)
\end{aligned}$$

The upper bound in (3) agrees with the upper bound in (2) when $q = 1$: the latter bounds are a special case of the former. But the lower bound in (2) will exceed the lower in (3) unless

$$\begin{aligned}
1 - n \cdot \xi + \xi \;&\geq\; (1-\xi)/(1-\xi+n\cdot\xi) \\
[1 \,-\, (n \cdot \xi \,-\, \xi)] \cdot [1 + n \cdot \xi \,-\, \xi] \;=\; 1 - (n-1)^2\xi^2 \;&\geq\; 1 - \xi \\
(n-1)^2 \xi^2 \;&\leq\; \xi
\end{aligned}$$

or $(n-1)^2 \xi \leq 1$; and since this will hold less often than not for the range of n and the particular ξ we shall have in what follows, we shall be using the lower bound in (2) always.

---

2 In a note appearing in the journal *Pragati,* pertaining to the workings of the UIDAI and authored by a senior research associate at the Takshashila Institute, it is asserted that "**the error crossover rate for fingerprinting and iris scans are 1 in 500 and 1 in 131000, respectively.**" [3] These are error rates supplied by their manufacturers, presumably, for the biometric devices used by the UIDAI. The author does not say if these are rates for errors in identification, particularly, or for errors in verfication. But a standard way of proceeding is to choose the error

[3] The note is titled "**Securing the Identity**"; it appeared in the 'roundup' section of the journal, on 06.01.2012; the author is R. Srikanth.

threshold for a device in such a way that identification error and verfication error are identical: so we shall suppose that the specified identification errors are such.

2.1 The UIDAI itself has recently circulated a paper, titled **The Role of Biometric Technology in Aadhar Enrollment**, which reports an experimental estimate for the chance of a false positive match. The estimate was made when 84 million individuals had been enrolled. The experiment consisted of 4 million trials of the following description: in each trial a template is picked from the database, and compared against the remainder to see if a match occurs. Assuming that the database contains no duplicates of the sampled templates, the number of times a matc occurs should allow us to reliably estimate the chance of a false positive match: which, remember, is the probability that a match occurs given that the person is not enrolled.[4]

The paper reports that a match occurred 2309 times out of 4 million. There is no mention there of error crossover rates for the devices being used: and the UIDAI proceeds on the basis of its experiment to take $2309/4 \cdot 10^6$ for what the probability of a false positive match was when 4 million had been enrolled. The matching procedure they are following is not specified either: so one must look about a little to see if the experimental estimate accords with the specified identification errors for their devices. But taking $r = 2309/4 \cdot 10^6 = 0.00057725$ as a reliable estimate of $\phi(n)$ for $n = 84 \cdot 10^6$ we may estimate the identification error in the field: writing ξ for identification error in the field now, and using the formula (1) from the summary of findings, we have

$$\frac{r}{n} \leq \xi \leq \frac{-\log(1-r)}{n} = \frac{r}{n} + \frac{r^2}{2 \cdot n} + \frac{r^3}{3 \cdot n} + \ ...$$

for bounds on ξ now: and since $r^4/4 \cdot n < 10^{-21}$ already for the values of r and n here, and as we have $(1.2) \cdot 10^9$ for an upper limit on the number to be enrolled, it should suffice to set

$$(\dagger) \qquad \xi = \frac{r}{n} + \frac{r^2}{2 \cdot n} + \frac{r^3}{3 \cdot n} = \frac{0.687400801}{10^{11}}$$

with $r = 2309/4 \cdot 10^6$ and $n = 84 \cdot 10^6$. We must now contrive to discern the matching procedure: to do which we must consider various possibilities and compute, from the identification errors specified for the devices, the specified identification for these matching schemes.

But we need consider a very few variants only, as it happens: for one can make a good guess, from the paper itself, at what the UIDAI's matching procedure must be. The physical features being metrized here are two irises and the insides of fingers and thumbs. The UIDAI's stated reason for using irises iss that "**people working in jobs that require repeated use of fingers — for example, in fireworks factories or in areca nut plantations — often find their fingerprints degraded, which makes iris useful in ensuring uniqueness**": and further remarks made apropos of the need for using irises suggest how the decision might be made. The 'proof of concept' exercise that the UIDAI had conducted,

[4] I am construing in the natural way the contents of the item numbered 2 in section 4.1 of the UIDAI's paper.

before enrollment was begun *en masse,* is said to have "**clearly demonstrated that iris capture was indeed necessary, and along with fingerprint, it was sufficient to de-duplicate and uniquely identify the entire population**" : and "**the accuracy of the combined system is an order of magnitude better**," the paper goes on to say, "**than fingerprints alone or iris alone**." One 'order of magnitude' is a factor of 10 here: and the second remark tempts one to guess that deciding on matches is done as follows: a match is taken to occur between an individual S presented for enrollment and an enrolled P if S and P are matched at *both irises* and at *any one* digit. The immediate motive for this guess is that the chance of a false match occurring at some or other digit, between an unenrolled S and an enrolled P, is roughly $1/50$ now: given a specified identification error of $1/500$ for fingerprint scanners: and assuming, again, the independence of the metrized features. But to make good our guess we should compute specified identification errors for close variants of this matching scheme. A few preliminaries are in order. Set

$$\begin{aligned} \rho_\iota &\equiv \textit{specified identification error for the iris scanner} \\ \rho_\delta &\equiv \textit{specified identification error for the fingerprint scanner} \\ \rho_{\delta,k} &\equiv \sum_{s\,=\,k}^{10} \binom{10}{s} \rho_\delta^{\,s} \, (1-\rho_\delta)^{10-s} \end{aligned}$$

For $1 \leq k \leq 10$ the quantity $\rho_{\delta,k}$ is the probability that, when S is not enrolled, comparing the numerized representations of his or her fingers to the representations of P's fingers, digit to corresponding digit, will result in at least k matches: $\rho_{\delta,k}$ is the probability, in short, that an unenrolled S and an enrolled P are matched at k digits at least.

Let ξ_0 denote the specified identification error for any matching procedure: we have $\xi_0 = \rho_\iota^{\,2} \cdot \rho_{\delta,1}$ now for the matching scheme that is our guess: assuming, to note it once more, the independence of the metrized features. We summarize below that scheme and its closest variants, and their specified identification errors, with the $\rho_\delta = 1/500$ and $\rho_\iota = 1/131000$; the first two columns list required matches.

irises	*digits*	ξ_0
2		$\rho_\iota^{\,2} = 0.582779560 \cdot 10^{-10}$
2	≥ 1	$\rho_\iota^{\,2} \cdot \rho_{\delta,1} = 0.115512486 \cdot 10^{-11}$
2	≥ 2	$\rho_\iota^{\,2} \cdot \rho_{\delta,2} = 0.103776040 \cdot 10^{-13}$

Our conjectured matching procedure is the second one here; and given the estimate in (†) for identification error in the field, the guess seems a good one, for the difference between specified error and field error is the least for that choice. Note also that, in going from the first to the second of the variants listed above, specified identification error does decrease by somewhat more than "**one order of magnitude**". We shall take our conjectured one to be the matching procedure followed by the UIDAI then; and for this scheme the ratio of field error to specified error is $0.687400801/0.115512486 = 5.950878773 \approx 6$.

We are assuming that, whatever the matching procedure of the UIDAI is, both irises will be scanned: and we had just noted the 'one order of magnitude' decrease in the specified identification error when going from using only irises to our conjectured matching procedure. But we shall summarize, for completeness, a few other possible matching schemes. As a preliminary set

$$\rho_{\iota,1} \;\equiv\; 1 - (1-\rho_\iota)^2$$

$\rho_{\iota,1}$ is the probability that an unenrolled S and an enrolled P will match at *one or other* iris. With $\rho_{\delta,k}$ as it was defined above, and with the values of ρ_ι and ρ_δ just given, the following table lists matching schemes and specified identification errors. The first two columns list required matches, as before; for the first four of the listed schemes only one iris is scanned, while for the last three both are.

irises	*digits*	ξ_0
1	$\geq$ 1	$\rho_\iota \cdot \rho_{\delta,1} = 0.151313186 \cdot 10^{-6}$
1	$\geq$ 2	$\rho_\iota \cdot \rho_{\delta,2} = 0.135953906 \cdot 10^{-8}$
1	$\geq$ 3	$\rho_\iota \cdot \rho_{\delta,3} = 0.725230000 \cdot 10^{-11}$
1	$\geq$ 4	$\rho_\iota \cdot \rho_{\delta,4} = 0.256474002 \cdot 10^{-13}$
$\geq$ 1	$\geq$ 2	$\rho_{\iota,1} \cdot \rho_{\delta,2} = 0.271890000 \cdot 10^{-8}$
$\geq$ 1	$\geq$ 3	$\rho_{\iota,1} \cdot \rho_{\delta,3} = 0.145036500 \cdot 10^{-10}$
$\geq$ 1	$\geq$ 4	$\rho_{\iota,1} \cdot \rho_{\delta,4} = 0.512914410 \cdot 10^{-13}$

Consider the third matching scheme summarized here. Only one iris is scanned now, and all ten digits: and a match will be taken to occur if a match occurs at the iris and at *any three* corresponding digits. The specified identification error here is nearly an order of magnitude less than the scheme which uses only the two irises: so this scheme is also a candidate for the matching procedure the UIDAI is actually using: and in that case the specified identification error would be very close to the estimate in (†) for identification error in the field.

But if the use of fingerprints is risky because "**people working in jobs that require repeated use of fingers ... often find their fingerprints degraded**," then it would be prudent to use as few digits as possible for the matching procedure: and the specified identification error for our conjectured scheme is considerably less besides, by more than a factor of 6 now, than the error for candidate we have just looked at: and that should have decided UIDAI to choose the former scheme over the latter, surely, were both being considered. Moreover, if using three digits poses little risk, then using two digits poses even less; and in that case the best strategy would have been to use the stricter variant of our conjectured scheme, which was summarized in the third line of the previous table, where a match would be taken to occur if matches occurred at both irises and at *any two* digits. The specified identification error of $0.103776040 \cdot 10^{-13}$ would have been three orders of magnitude less, now, than the error of $0.582779560 \cdot 10^{-10}$ for the scheme using matches only at both irises; so the UIDAI should surely have gone with that, rather, if they were willing to go with the candidate looked at here. So we shall persist with our conjectured scheme.

2.2 As we are now considering the operations of the UIDAI in particular, we shall refer to individuals presented for enrollment as *applicants* now: for a UID or 'unique identity' precisely. Following the report of the experiment to estimate the false match rate $p(B|A^c)$ for the next unenrolled applicant after 84 million had been enrolled, the paper put out by the UIDAI describes an experiment designed to estimate the false accept rate: which is the conditional probability $p(B^c|A)$ that a match will not occur given that, on the contrary, the applicant is already enrolled. We have already noted that the false accept rate does not depend on the number of individuals already enrolled: and in (v) of the previous section we had derived an upper bound for this probability from the specified verification errors

of the devices being used. The second experiment consisted in 31,399 trials of the following description. In each trial an enrolled person is selected at random from the database, and a fresh suite of biometrics is obtained from him: and that suite is compared against each suite of templates in the database, then, to see if a match occurs. One expects matches now: particularly because the fresh suite is being compared against the template of the enrolled individual himself, among others. In 31,388 of these trials a match did indeed occur: so $11/31399$ is the UIDAI's experimental estimate of the false accept rate $p(B^c|A)$.[5] From the relation

$$p(B|A) + p(B^c|A) \,=\, [p(B\ \&A) \,+\, p(B^c\ \&A)]/p(A) \,=\, p(A)/p(A) \,=\, 1\,.$$

we get $p(B|A) \,=\, 1 - p(B^c|A)$: and $31388/31399$ is the estimate we have of this probability now: which we shall regard as fixed because, as we have noted, the complementary false accept rate does not change as enrollment proceeds.

One might wonder about the propriety of estimating both $p(B|A^c)$ and $p(B^c|A)$ with just such experiments: by computing suites of distances $\{x_1, x_2, \,...\,, x_K\}$ between the templates of enrolled persons, or between the latter and freshly taken biometrics of such persons. But complications would arise only if appeal were made to the distributional properties an x_k would have when it is regarded as a value of $X_k^{\mathfrak{d}}$: and that cannot be done, obviously, when $p(B^c|A)$ is being estimated. No such appeal is being made, however, in the calculations. We have suites of distances obtained in two distinct ways: by template-to-template comparisons first, where the elements of a compared pair derive from distinct persons, and then by comparisions of fresh biometrics to templates. In the latter case the elements of a compared pair will not always derive from distinct persons. We may identify the conditional circumstances $(B|A^c)$ and $(B^c|A)$ with concatenations of certain numerical events, now, which are determined by how each x_k lies to the threshold τ_k only, without regarding the number as a value of either $X_k^{\mathfrak{d}}$ or $X_k^{\mathfrak{s}}$: or as a value of any distribution, for that matter: and our *interpretations* of these numerical circumstances as recognizable eventualities are secured by how these suites of distances are got.

Our ultimate interest here is in the proability $p(A^c|B)$ of mistaken identification: the computing of which requires the converse $p(B|A^c)$ and as the simple probabilities $p(B)$ and $p(A^c)$ as well, as we noted in (vi) above: and from (vii) we see that an estimate of either $p(A^c)$ or its complement $p(A) = 1 - p(A^c)$ together with estimates of $p(B|A^c)$ and $p(B|A)$ will allow us to estimate $p(B)$.

Now following the report of the second experiment one finds that the UIDAI has determined to its satisfaction a "**current 0.5% rate of duplicate submissions**" : which seems to say that 0.5% of the applicants turn out to have been enrolled already, by the time 84 million had been enrolled. Assuming that the entire population to be enrolled had been randomly sampled by then, we may take $p(A) \,=\, 0.005$ and $p(A^c) \,=\, 0.995$ for stable estimates of these probabilities: and the UIDAI is prepared to assume, we note, that the 'rate of duplicate submission' will neither decrease nor increase as enrollment proceeds.[6]

---

[5] The results of the latter experiment are reported in the item numbered 3 in Section 4.1 of **The Role of Biometric Technology in Aadhar Enrollment**. The phrase "**false negative identification rate**" is used there for the false accept rate: which is, of course, the chance that an already enrolled person will succeed in getting enrolled again: presumably under an alias. The paper approximates $p(B^c|A)$ as 0.000352 , while $0.0003503 < 11/31399 < 0.0003504$; a Bayesian estimate must have been made, one supposes, with some likely prior.

[6] I refer the reader again to the item numbered 3 in Section 4.1 of **The Role of Biometric Technology in Aadhar Enrollment**.

2.3 We can now proceed with the calculations for our other two findings. Our first object is to compute the probability of mistaken identification for various subsets of the population: which is estimated by the ratio of false matches to all matches. We shall begin by estimating the number of matches that will have occurred by the time enrollment is complete, and estimate the number of false matches among these. The population of India is said to be 1.2 billion; so for each n between 1 and $(1.2)\cdot 10^9$ we must consider what the false reject rate $\phi(n)$ becomes by the time n individuals have been enrolled; and the lower bound in (2) gives the safe approximation

$$\phi(n) \;\approx\; n\cdot\left[\frac{\xi\cdot(1\,-\,\xi)}{1\,-\,\xi\,+\,n\cdot\xi}\right] \tag{2.1}$$

of $p(B|A^c)$ for the next individual to be enrolled. Using ξ from (†) now, the easiest way to go would be to obtain a linear estimate for last term: otherwise we shall have to make 1.2 billion estimates. We shall come back to the question.

Write $\beta(n)$ for the probabililty that a match will occur for the next individual to be enrolled: we get $\beta(n) = \phi(n)\cdot p(A^c) + p(B|A)\cdot p(A)$ then from the formula (vii) that we had for $p(B)$ above. The estimates of the UIDAI give us

$$\begin{aligned} p(B|A)\cdot p(A) \;&\approx\; (31388\cdot(0.005))/1399 \;=\; 0.004998248\\ p(A^c) \;&\approx\; 0.995\end{aligned}$$

Set $\gamma \equiv (0.004998248)$ for convenience. Suppose a total of Q many individuals are to be enrolled; let $m(Q)$ denote the total number of matches we expect. Assuming that the occurence of matches is independent, as we may, we get

$$m(Q) \;=\; \sum_{n\,=\,1}^{Q-1}\beta(n) \;\approx\; p(A^c)\cdot\left[\sum_{n\,=\,1}^{Q-1}\phi(n)\right] \;+\; \gamma\cdot(Q-1) \tag{2.2}$$

For quick count of $m(Q)$ note first that $\phi(n) \geq [n\cdot\xi\cdot(1-\xi)]/[1-\xi+Q\cdot\xi]$ always. Setting $\lambda = (0.687400801)$ we have $\xi = \lambda\cdot 10^{-11}$ from (†) now; and with and $Q = (1.2)\cdot 10^9$ we will, with a little calculation, obtain the approximations

$$\frac{\xi\cdot(1-\xi)}{1-\xi+Q\cdot\xi} \;\approx\; \frac{\lambda}{10^9\cdot(10^2+(1.2)\cdot\lambda)} \;=\; \frac{0.681777}{10^{11}} \tag{2.3}$$

$$p(A^c)\cdot\left[\frac{\xi\cdot(1-\xi)}{1-\xi+Q\cdot\xi}\right] \;\approx\; \frac{(0.995)\cdot(0.681777)}{10^{11}} \;=\; \frac{0.678368115}{10^{11}} \tag{2.4}$$

Set $\eta \equiv (0.678368115)/10^{11}$ now: from (1) and (2) and (4) just above we have

$$m(Q) \;\geq\; \eta\cdot\left[\sum_{n\,=\,1}^{Q-1} n\right] \;+\; \gamma\cdot(Q-1) \;=\; [\eta\cdot Q\cdot(Q-1)]/2 \;+\; \gamma\cdot(Q-1)$$

then. Note next that $(Q-1) \;=\; (1.2)\cdot(10^9-1)+(0.2)$ when $Q = (1.2)\cdot 10^9$: since $(a\cdot 10^q - 1) \;-\; a\cdot(10^q-1) \;=\; a-1$. So the approximations

$$\begin{aligned} Q-1 \;&\approx\; (1.2)\cdot(10^9-1)\\ Q\cdot(Q-1) \;&\approx\; (1.2)\cdot 10^9\cdot(1.2)\cdot(10^9-1) \;=\; (1.44)\cdot 10^9\cdot(10^9-1)\end{aligned}$$

may safely be used for estimation, given the values declared for η and γ here; and from the inset calculation below we see that $m(Q)$ will equal or exceed $10,882,148$ or $10^7 + 882,148$ now.

$$\begin{aligned} m(Q) &\geq [\eta \cdot Q \cdot (Q-1)]/2 + \gamma \cdot (Q-1) \\ &\approx (0.678368115) \cdot (1.44) \cdot 10^9 \cdot (10^9 - 1)/(2 \cdot 10^{11}) \quad + \quad (1.2) \cdot (10^9 - 1) \cdot (0.004998248) \\ &= (0.488425043) \cdot (10^9/10^2) \cdot ((10^9 - 1)/10^9) \quad + \quad (0.005997898) \cdot (10^9 - 1) \\ &\approx (4884250.43) \cdot (1 - (1/10^9)) \quad + \quad 5997898 \quad - \quad 0.005 \\ &\approx 4884250 \; + 5997898 \; + \; (0.43 \; - \; 0.005 \; - \; 0.005) \; \approx \; 10882148 \end{aligned}$$

To obtain the approximation in (2.3) we do as follows with λ and Q as they were above:

$$\begin{aligned} \xi &= \lambda \cdot 10^{-11} = (\lambda \cdot 10^{-6})/10^5 \\ 1 - \xi &= (10^5 - \lambda \cdot 10^{-6})/10^5 \\ Q \cdot \xi &= (1.2) \cdot \lambda \cdot 10^{-2} = [(1.2) \cdot \lambda \cdot 10^3]/10^5 \\ 1 - \xi + Q \cdot \xi &= [10^5 - \lambda \cdot 10^{-6} + (1.2) \cdot \lambda \cdot 10^3]/10^5 \\ \frac{1-\xi}{1-\xi+Q\cdot\xi} &= \frac{10^5 - \lambda \cdot 10^{-6}}{10^5 - \lambda \cdot 10^{-6} + (1.2) \cdot \lambda \cdot 10^3} \approx \frac{10^5}{10^5 + (1.2) \cdot \lambda \cdot 10^3} = \frac{10^2}{10^2 + (1.2) \cdot \lambda} \\ \frac{\xi \cdot (1-\xi)}{1-\xi+Q\cdot\xi} &\approx \frac{\lambda \cdot 10^2}{10^{11} \cdot (10^2 + (1.2) \cdot \lambda)} = \frac{\lambda}{10^9 \cdot (10^2 + (1.2) \cdot \lambda)} \end{aligned}$$

We underestimate $m(Q)$ here by using a uniform lower bound for $\phi(n)$: and we shall provide a more accurate estimate momentarily. But one should expect more than 10.88 million matches to have occurred, at any rate, by the time enrollment is complete.[7] Let us now count the false ones among all the matches. The probability $p(B \,\&\, A^c)$ is what we must attend to: so let $\varepsilon(n)$ denote the chance that, once n individuals have been enrolled, the next applicant is not enrolled and that, nonetheless, a match does occur. As $p(B \,\&\, A^c) = p(B|A^c) \cdot p(A^c)$ we get

$$\varepsilon(n) \; = \; \phi(n) \cdot p(A^c) \; \approx \; \eta \cdot n \; = \; (0.678368115) \cdot n/10^{11}$$

from (2.1) and (2.2) and (2.4) again. Let $d(Q)$ denote the total number of unenrolled applicants for whom a match will have occurred by the time enrollment is complete: the total number of false matches, that is to say. Assuming that such matches are independent of each other as well, we get $\sum_{n=1}^{Q-1} \varepsilon(n)$ as our estimate of this number: whence

$$d(Q) \; = \; \sum_{n=1}^{Q-1} \varepsilon(n) \; \approx \; \eta \cdot \left[\sum_{n=1}^{Q-1} n \right] \; = \; \eta \cdot Q \cdot (Q-1)/2$$

now; and from the computation we had just now performed for $m(Q)$ we see that $\eta \cdot Q \cdot (Q-1)/2$ will equal or exceed $4,884,250$. To summarize: by the time enrollment is complete the UIDAI should expect to have adjudicated matches for more than 10.88 million applicants: and more than 4.88 million among these will have been false matches.

---

[7] In its paper the UIDAI computes, mistakenly, that only about 570 matches will occur for every 1 million enrollments: and the total number of matches they are expecting is only about $570 \cdot (1.2) \cdot 10^3 = 684,000$.

To save writing, let us call an applicant whose suite of biometrics matches that of someone already enrolled a *matched applicant*: who will be falsely matched if he or she happens not to be enrolled. Now the numbers 10.88 million and 4.88 million may seem trifling when they are set beside 1.2 billion; and perhaps they are, in calculations that the functionaries of the Indian state are accustomed to perform. But there is an asymmetry here which one should note: as we shall see momentarily, the number of matches, and the number of false matches among these, will vary considerably between the initial and final stages of enrollment.

To estimate the difference here we need to refine our estimate of the factor η above: by using small values of Q now. Repeating the calculation of (2.3) with $\lambda = 0.687400801$ and $Q = 10^6$ we get

$$\begin{aligned} \frac{\xi \cdot (1-\xi)}{1-\xi \,+\, 10^6 \cdot \xi} &\approx \frac{\lambda \cdot 10^5}{10^{11} \cdot (10^5 + (1.2) \cdot \lambda)} = \frac{0.687395131}{10^{11}} \\ p(A^c) \cdot \left[\frac{\xi \cdot (1-\xi)}{1-\xi \,+\, 10^6 \cdot \xi} \right] &\approx \frac{(0.995) \cdot (0.687395131)}{10^{11}} = \frac{0.683958155}{10^{11}} \end{aligned}$$

Setting $\eta_1 = (0.683958155)/10^{11}$ and with γ as above, and using η_1 just as we used η, the matches expected for the first million enrolled is

$$\sum_{n=1}^{10^6-1} \beta(n) \;\approx\; \eta_1 \cdot 10^6 \cdot (10^6-1)/2 \;+\; \gamma \cdot (10^6-1) \;:$$

which comes to $(0.341979078) \cdot [10^6/10^5] \cdot [(10^6-1)/10^6] + (0.004998248) \cdot (10^6-1)$ or $3 + 4998 = 5001$ approximately: and only 3 among these will be falsely matched applicants. But the number of matches expected for the last million enrolled, and the falsely matched applicants among them, will be somewhat larger. To estimate these numbers set $Q = (1.2) \cdot 10^9$ again and $M = (1.2) \cdot 10^9 - 10^6$; the number of matches in the last million is $\sum_{n=M}^{Q-1} \beta(n)$ now, which comes to 13135 approximately with η and γ as they are above; see the inset calculation.

$$\begin{aligned} \sum_{n=M}^{Q-1} \beta(n) &\approx \eta \cdot \left(\sum_{n=M}^{Q-1} n \right) \;+\; \gamma \cdot (Q-M) \\ &= \eta \cdot [(M-1) \cdot (Q-M) \;+\; (Q-M+1) \cdot (Q-M)/2] + \gamma \cdot (Q-M) \\ &= (\eta/2) \cdot 10^6 \cdot \left[\, (2.4) \cdot 10^9 \;-\; 10^6 \;-\; 1 \, \right] \;+\; \gamma \cdot 10^6 \\ &= (0.339184057) \cdot \frac{10^6}{10^2} \cdot \left[\frac{(2.4) \cdot 10^9}{10^9} \;-\; \frac{10^6}{10^9} \;-\; \frac{1}{10^9} \right] \;+\; (0.004998248) \cdot 10^6 \\ &\approx (0.339184057) \cdot 10^4 \cdot (2.39899) \;+\; (0.004998248) \cdot 10^6 \\ &\approx 8137 \;+\; 4998 \;=\; 13135 \end{aligned}$$

The numbers 5001 and 13135 are comparable enough. But let us count among the latter matched applicants those who are falsely matched: for which we get

$$\sum_{n=M}^{Q-1} \varepsilon(n) \;\approx\; \eta \cdot \left[\sum_{n=M}^{Q-1} n \right] \;=\; (0.339184057) \cdot \frac{10^6}{10^2} \cdot \left[\frac{(2.4) \cdot 10^9}{10^9} \;-\; \frac{10^6}{10^9} \;-\; \frac{1}{10^9} \right]$$

or 8137 approximately: the calculation was just made in the inset passage above. To summarize: one expects about 5001 matches for the first one million enrolled, and 3 among the matched applicants should be falsely matched: in the last one million enrolled one expects around 13135 matches, and among these matched applicants some 8137 should be falsely matched.

These exercises can be repeated for the sets or batches formed by successive millions of applicants. Let N_k denote the expected number of matches among the k-th million applicants, and let F_k denote the expected number of falsely matched applicants among these; we must now estimate a factor like η for different values of k. Setting $Q_k = k \cdot 10^6$ and repeating the calculations for (2.3) with Q_k in place of Q we get

$$\frac{\xi \cdot (1-\xi)}{1-\xi \,+\, Q_k \cdot \xi} \;\approx\; \frac{\lambda}{10^6 \cdot (10^5 + \, k \cdot \lambda)}$$

and we set $\eta_k = p(A^c) \cdot \lambda / [10^6 \cdot (10^5 + \, k \cdot \lambda)]$ now; then (2.1) allows us to use η_k just as we have used η so far; and with the set values of $p(A^c)$ and λ and γ above we finally obtain

$$\begin{aligned} N_k \;&=\; \sum_{n\,=\,Q_{k-1}}^{Q_k - 1} \beta(n) \;&\approx\; &\sum_{n\,=\,(k-1)\cdot 10^6}^{k\,\cdot 10^6 - 1} [p(A^c) \cdot \phi(n) \,+\, \gamma] \\ &&\approx\; &\eta_k \cdot 10^6 \cdot [(2k-1) \cdot 10^6 \,-\, 1]/2 \;+\; \gamma \cdot 10^6 \\ &&=\; &\frac{p(A^c) \cdot \lambda \cdot [(2k-1) \cdot 10^6 \,-\, 1]}{2 \cdot (10^5 + \, k \cdot \lambda)} \;+\; \gamma \cdot 10^6 \\ F_k \;&=\; \sum_{n\,=\,Q_{k-1}}^{Q_k - 1} \varepsilon(n) \;&\approx\; &N_k \;-\; \gamma \cdot 10^6 \end{aligned}$$

with $\sum_{i\,=\,r}^{s} i \,=\, (s-r+1) \cdot (s+r)/2$ giving the second line. Using these formulae we can estimate and compare matches and false matches for the first and last m million applicants, respectively, for varying values of m. Write $N_\alpha(m)$ for the matches among the first m million, and $F_\alpha(m)$ for the false matches among these; write $N_\omega(m)$ for the matches among the last m million and $F_\omega(m)$ for the false matches among these. Here are some estimates of these counts, for a few values of m, among first and last aggregated millions.

m	N_α	F_α	N_ω	F_ω
1	5001	3	13135	8137
2	10011	13	26264	16267
5	25077	85	65609	40168
10	50325	341	131050	81067
100	534010	34180	1280208	780382
200	1136318	136665	2493042	1493392
250	1463057	213490	3074167	1824605

In its paper the UIDAI claims that "**the system will be able to scale to handle the entire population without significant drop in accuracy**." The

ratio of false matches to total matches expected, for any substantial subset of the population, seems an appropriate measure of accuracy here: the lower that ratio — the lower the chance of mistaken identification, that is to say — the more accurate the system will be. So the ratios $F_\alpha(m)/N_\alpha(m)$ and $F_\omega(m)/N_\omega(m)$ are what we must attend to now: and here they are for the values of m with which we had computed above:

m	F_α/N_α	F_ω/N_ω
1	0.00059988	0.619489912
2	0.012858556	0.61936491
5	0.00338956	0.619091893
10	0.006775956	0.618595956
100	0.064006292	0.609574382
200	0.12027003	0.599024004
250	0.145920494	0.593528263

The expected ratio of false matches to total matches rises considerably between the initial and final stages of enrollment. One or ten million may not count as substantial subsets of the population: but a hundred million surely does, in a population exceeding one billion: and so we see that, as recorded in the summary of findings, the chance of mistaken identification will increase almost ten-fold between the first and the last 100 millions enrolled.

Matters improve somewhat between the first 250 million enrolled and the last 250 million, for the drop in accuracy is less than five-fold now; but note how dramatically accuracy drops as enrollment rises. The expected number of matches in the second 100 million is $1136318 - 534010 = 602308$; the expected number of false matches among these is $136665 - 34180 = 102485$; and the ratio between these numbers becomes 0.170153808 for the second 100 million enrolled, while it was only 0.064006292 for the first 100 million. This last manouevre suggests the appropriate way to summarize the drop in accuracy: estimate and display the number of matches, and the false matches among these, for each successive million enrolled. Let us call these *false match ratios* for brevity. The graph in *figure 1* shows the results: it plots both the cumulative and successive false match ratios as enrollment proceeds. The lower curve charts the cumulative ratios: by which is meant the ratio of false matches among *all* matches, of course, when a given number have been enrolled. The horizontal axis counts the population in millions, while the vertical axis records the respective ratios; the numerical labels display actual counts at each hundredth million.

Computing and summing with the formulae for N_k and F_k above gives us our revised estimate of $10,895,510$ for the total count $m(Q)$ of matches expected: among which one may expect $4,897,609$ to be false matches. These are lower bounds for these counts: we have underestimated them by using the lower bound in (2) for the probabilities $\phi(n)$. But the using the upper bound $n \cdot \xi$ will not change things too much. Write $m^+(Q)$ for upper bound on the total number matches; we have

$$m^+(Q) \;=\; \sum_{n=1}^{Q-1} [p(A^c) \cdot n \cdot \xi \,+\, \gamma] \;\;=\;\; p(A^c) \cdot \xi \cdot \sum_{n=1}^{Q-1} n \;\;+\;\; \gamma \cdot (Q-1)$$

then. Repeating the calculation for $m(Q)$ above, with $Q = 1.2$ billion again and with $p(A^c)\cdot\xi$ in place of η there, yields $10,922,437$ as an upper bound on matches, and $4,924,539$ as an upper bound on false matches. The differences between the upper and lower bounds here are 0.002471385 and 0.005498601 as fractions of the respective lower bounds: and as these are marginal for the largest of our counts, we may safely go with the numbers listed above. The graph in *figure 2* plots the differences, for successive millions, between false matches counted with the upper and lower bounds for $\phi(n)$ respectively; and *figure 3* plots these differences for all matches. Notice that these differences grow in the same way as enrollment proceeds: so using the upper bound for $\phi(n)$ should make neither more pronounced, nor less, the differences between initial false match ratios and final false match ratios.

2.4 When a match occurs for a putatively new applicant, his or her suite of biometrics may match more than one suite of templates in the database: let us call such a suite of matching templates a *matched record*: and each matched record will have to be examined, now, to determine whether or not the match is a false one. To measure how efficient the process of enrollment is we shall have to estimate the total number of matched records that will have been found, by the time enrollment is complete, for all the matched applicants the UIDAI should expect.

The process of estimation here is somewhat tortuous. Let $q < n$ and suppose n individuals have already been enrolled. Let $R_q(n)$ abbreviate the following circumstance: that *at least* q suites of templates in database will match the biometrics of the next applicant. We must estimate its probability: and, keeping with the notation we have been using, to do that we require the probability of the conditional circumstance $[R_q(n)|A^c]$: the circumstance that at least q suites of templates in database will match the biometrics of the next applicant S *given* that he or she is not in fact enrolled. The probability of a match occuring for S is $1-(1-\xi)^n$ recall: the chance that his or her suite of biometrics will match the templates of *some or other* enrolled individual: which is the probability $p[R_1(n)|A^c]$ also, then, the chance of finding at least one matched record for an unenrolled applicant. The probability that *exactly one* suite of templates in the database will match the biometrics of S is $n\cdot\xi\cdot(1-\xi)^{n-1}$ now: assuming, again, that matching with any one suite is independent of matching with any other. With this assumption the probability that *exactly* j suites of templates will thus match is $\binom{n}{j}\cdot\xi^j\cdot(1-\xi)^{n-j}$ now; and we have

$$p[R_q(n)|A^c] \;=\; \sum_{r\,=\,q}^{n}\binom{n}{r}\xi^r\cdot(1-\xi)^{n-r} \;=\; 1-\sum_{j\,=\,0}^{q-1}\binom{n}{j}\xi^j\cdot(1-\xi)^{n-j}$$

then. The quantity on the right is the term $\psi_q(n)$ of formula (3) in the summary of findings: which gives us the upper bound on $p[R_q(n)|A^c] \equiv \psi_q(n)$ that we shall be using. It seems reasonable to suppose that $p[R_{m+1}(n)|A] \approx p[R_m(n)|A^c]$: the chance of finding at least $m+1$ matched records when the $(n+1)$-st applicant is already enrolled should approximate, surely, the chance of finding at least m matched records were he or she not enrolled. So we have

$$\begin{aligned} p[R_q(n)] \;&=\; p[R_q(n)|A]\cdot p(A) \;+\; p[R_q(n)|A^c]\cdot p(A^c) \\ &=\; \psi_{q-1}(n)\cdot p(A) \;+\; \psi_q(n)\cdot p(A^c) \end{aligned}$$

for $q > 1$. We may assume independence of matches for different applicants: the chance that at least q matched records will be found for one applicant is not affected by whether or not that happens for any other. Write N for the total number of individuals to be enrolled, and for $q > 1$ set

$$T_q \;=\; \sum_{n=1}^{N-1} p[R_q(n)] \;=\; p(A) \cdot \sum_{n=1}^{N-1} \psi_{q-1}(n) \;+\; p(A^c) \cdot \sum_{n=1}^{N-1} \psi_q(n)$$

T_q will equal or exceed the total number of matched records found for those matched applicants whose suite of biometrics matches q suites of templates, at least, in the database. Write S_q for this latter set of matched applicants: we have $S_q \supseteq S_{q+1}$ of course: and for the total count T of matched records we have

$$T \;=\; T_1 - T_2 \,+\, 2 \cdot (T_2 - T_3) \,+\, \ldots \,+\, q \cdot (T_q - T_{q+1}) \,+\, \ldots \tag{2.5}$$

then. We expect to have $T_q - T_{q+1} \approx 0$ before q gets at all large: because

$$\psi_{q+1}(n) \;=\; \psi_q(n) \;-\; (1-\xi)^n \cdot \binom{n}{q} \cdot \left[\frac{\xi}{1-\xi}\right]^q$$

and the sum $\sum_n \binom{n}{q} \cdot \xi^q/(1-\xi)^q \;<\; N \cdot [N^q/q!] \cdot [\xi^q/(1-\xi)^q]$ will become negligibly small quite soon for ξ and the population size N as they are here. We shall use the bound in (3) as our approximation for $\psi_q(n)$ now: given the tight bounds that (1) gives for $\psi_1(n)$ already, and as $\psi_{q+1}(n) < \psi_q(n)$, doing so should be quite safe. We have T_1 already: that will equal the total number of matched applicants, of course, and at the end of the last section we had computed this count as $10,922,437$ using $\psi_1(n) = 1 - [1-\xi]^n \approx n \cdot \xi$ as our approximation. Now for $N = (1.2) \cdot 10^9$ we obtain

$$\begin{aligned}
T_2 \;&=\; p(A) \cdot \sum_{n=1}^{N-1} \psi_1(n) \;+\; p(A^c) \cdot \sum_{n=1}^{N-1} \psi_2(n) \;\approx\; 274,545 \,+\, \tfrac{4}{10} \\
T_3 \;&=\; p(A) \cdot \sum_{n=1}^{N-1} \psi_2(n) \;+\; p(A^c) \cdot \sum_{n=1}^{N-1} \psi_3(n) \;\approx\; 219 \,+\, \tfrac{9}{10} \\
T_4 \;&=\; p(A) \cdot \sum_{n=1}^{N-1} \psi_3(n) \;+\; p(A^c) \cdot \sum_{n=1}^{N-1} \psi_4(n) \;\approx\; \tfrac{7}{10}
\end{aligned}$$

Set $\lambda = 0.687400801$ for convenience, as before, so that $\xi = \lambda \cdot 10^{-11}$ again; and write a for $p(A) = 0.005$ now. With $\psi_q(n) \le [\xi^q/(q-1)!] \cdot \prod_{r=0}^{q-1}(n-r)$ from (3) we get

$$\begin{aligned}
T_2 \;&=\; p(A) \cdot \sum_{n=1}^{N-1} \psi_1(n) \;+\; p(A^c) \cdot \sum_{n=1}^{N-1} \psi_2(n) \\
&\approx\; \frac{a \cdot \lambda}{10^{11}} \cdot \left[\sum_{n=1}^{N-1} n\right] \;+\; \frac{(1-a) \cdot \lambda^2}{10^{22}} \cdot \left[\sum_{n=1}^{N-1} n \cdot (n-1)\right] \\
&=\; \frac{a \cdot \lambda}{10^{11}} \cdot \left[\sum_{n=1}^{N-1} n\right] \;+\; \frac{(1-a) \cdot \lambda^2}{10^{22}} \cdot \left[\sum_{n=1}^{N-1} n^2 \;-\; \sum_{n=1}^{N-1} n\right] \\
&\approx\; \frac{a \cdot \lambda \cdot N \cdot (N-1)}{10^{11} \cdot 2} \;+\; \frac{(1-a) \cdot \lambda^2 \cdot N \cdot (N-1) \cdot (2N-1)}{10^{22} \cdot 6} \\
&\approx\; \frac{a \cdot \lambda \cdot N^2}{2 \cdot 10^{11}} \;+\; \frac{(1-a) \cdot \lambda^2 \cdot 2 \cdot N^3}{6 \cdot 10^{22}} \\
&=\; \frac{(0.247464288) \cdot 10^{18}}{10^{13}} \;+\; \frac{(0.270810583) \cdot 10^{27}}{10^{22}} \;=\; 274545.3463
\end{aligned}$$

In going from line 3 to line 4 above we discard $(1-a)\lambda^2 10^{-22}\cdot\sum_n n \;<\; N^2/10^{22} < 10^{-3}$; in going from line 4 to line 5 we discard $(a\lambda/2)N\cdot 10^{-11} < 10^{-4}$ from the first term, and from the second term all the summands where N has power at most 2, doing as we did in going from line 3 to line 4. Continuing, we have

$$\begin{aligned}
T_3 &= p(A)\cdot\sum_{n=1}^{N-1}\psi_2(n) \;+\; p(A^c)\cdot\sum_{n=1}^{N-1}\psi_3(n)\\
&\approx \frac{a\cdot\lambda^2}{10^{22}}\cdot\left[\sum_{n=1}^{N-1} n\cdot(n-1)\right] \;+\; \frac{(1-a)\cdot\lambda^3}{10^{33}\cdot 2}\cdot\left[\sum_{n=1}^{N-1} n\cdot(n-1)\cdot(n-2)\right]\\
&= \frac{a\cdot\lambda^2}{10^{22}}\cdot\left[\sum_{n=1}^{N-1} n^2 - \sum_{n=1}^{N-1} n\right] \;+\; \frac{(1-a)\cdot\lambda^3}{2\cdot 10^{33}}\cdot\left[\sum_{n=1}^{N-1} n^3 - 3\cdot\sum_{n=1}^{N-1} n^2 + 2\cdot\sum_{n=1}^{N-1} n\right]\\
&\approx \frac{a\cdot\lambda^2}{10^{22}}\cdot\left[\sum_{n=1}^{N-1} n^2\right] \;+\; \frac{(1-a)\cdot\lambda^3}{2\cdot 10^{33}}\cdot\left[\sum_{n=1}^{N-1} n^3\right]\\
&= \frac{a\cdot\lambda^2\cdot N\cdot(N-1)\cdot(2N-1)}{10^{22}\cdot 6} \;+\; \frac{(1-a)\cdot\lambda^3}{2\cdot 10^{33}}\cdot\left[\sum_{n=1}^{N-1} n\right]^2\\
&\approx \frac{a\cdot\lambda^2\cdot 2\cdot N^3}{6\cdot 10^{22}} \;+\; \frac{(1-a)\cdot\lambda^3}{2\cdot 10^{33}}\cdot\left[\frac{N^4 - 2N^3 + N^2}{4}\right]\\
&\approx \frac{a\cdot\lambda^2\cdot 2\cdot N^3}{6\cdot 10^{22}} \;+\; \frac{(1-a)\cdot\lambda^3\cdot N^4}{8\cdot 10^{33}}\\
&= \frac{(0.001360857)10^{27}}{10^{22}} \;+\; \frac{(0.083769935)\cdot 10^{36}}{10^{33}} \;=\; 219.855635
\end{aligned}$$

In going from line 3 to line 4 we have done with the first term just as we did in computing T_2, which was to discard summands where the power of N is at most 2; and in the second term we discard summands where the power of N is 3 or less, for these will not exceed 10^{-5} in absolute value. We discard terms for the same reasons in going from line 5 to line 7. Going on we get

$$\begin{aligned}
T_4 &= p(A)\cdot\sum_{n=1}^{N-1}\psi_3(n) \;+\; p(A^c)\cdot\sum_{n=1}^{N-1}\psi_4(n)\\
&\approx \frac{a\cdot\lambda^3}{10^{33}\cdot 2}\cdot\left[\sum_{n=1}^{N-1} n\cdot(n-1)\cdot(n-2)\right] + \frac{(1-a)\cdot\lambda^4}{10^{44}\cdot 3!}\cdot\left[\sum_{n=1}^{N-1} n\cdot(n-1)\cdot(n-2)\cdot(n-3)\right]\\
&\approx \frac{a\cdot\lambda^3}{10^{33}\cdot 2}\cdot\left[\sum_{n=1}^{N-1} n^3\right] + \frac{(1-a)\cdot\lambda^4}{10^{44}\cdot 3!}\cdot\left[\sum_{n=1}^{N-1} n^4\right]\\
&\approx \frac{a\cdot\lambda^3}{2\cdot 10^{33}}\cdot\left[\frac{N^4}{4}\right] + \frac{(1-a)\cdot\lambda^4}{10^{44}\cdot 3!}\cdot\left[\frac{N^5}{5} - 10\cdot\left(\sum_{n=1}^{N-1}(n^3+n^2)\right) - 5\cdot\left(\sum_{n=1}^{N-1} n\right) - N\right]\\
&\approx \frac{a\cdot\lambda^3\cdot N^4}{8\cdot 10^{33}} \;+\; \frac{(1-a)\cdot\lambda^4\cdot N^5}{5\cdot 3!\cdot 10^{44}}\\
&\approx \frac{(0.000420954)\cdot 10^{36}}{10^{33}} \;+\; \frac{(0.018426727)\cdot 10^{45}}{10^{44}} \;=\; 0.60522127
\end{aligned}$$

and the rationale for discarding terms above should be clear now from what has already been said. Proceeding in this fashion will yield

$$T_5 \;\approx\; \frac{a\cdot\lambda^4\cdot N^5}{5\cdot 3!\cdot 10^{44}} \;+\; \frac{(1-a)\cdot\lambda^5\cdot N^6}{6\cdot 4!\cdot 10^{55}} \;\approx\; 0.000092597 \;+\; 0.0003166637$$

and the terms T_q will be even smaller for $q \geq 6$: so we may well assume that $T_q \approx 0 \approx T_{q+1}$ for $q \geq 5$ now.

Suppose q is the least value for which $T_{q+1} = T_q$; then from the relation (2.5) we had above we get

$$T \;=\; \left[\sum_{r=1}^{q-1} r \cdot (T_r - T_{r+1}) \right] \;+\; q \cdot T_q \;=\; T_1 + T_2 \;+\; \ldots \;+\; T_q$$

The inset calculation just above lets us take $q = 5$ for such a least value, and with $T_5 \approx 0$ besides; so, since we have computed with upper bounds for our probabilities $\psi_n(q)$, we shall hazard that

$$T \;\approx\; 10,922,437 \;+\; 274,545 \;+\; 220 \;+\; 1 \;=\; 11,267,203$$

is what the count of matched records will come to, at most, by the time the entire population has been enrolled. At the end of section **2.3** we had computed an upper limit of 10,922,437 on total matches, and an upper limit of 4,924,539 on false matches among these: and with 11,267,203 for an upper limit on matched records we see that, as asserted in the summary of findings, to determine which matches are false only rarely will more than one matched record have to be examined: even though the number of matched records examined for each match will increase in the final stages of enrollment.

3 In the course of discussing specified errors we referred, in the inset passage in section **1** above, to the documents referenced as [1] and [2] below. The collection [3] would be useful to readers who want to learn more about the complexities of biometric identification. These documents are freely available online, on the World Wide Web. The paper we considered in some detail, **The Role of Biometric Technology in Aadhar Enrollment**, is available on the website of the UIDAI; and our source for the specified errors of their devices, the note titled **Securing the Identity** which appeared online in the journal *Pragati* early in 2012, presumably remains available there. We close by registering a very minor complication. We have taken a biometric for a real or binary vector: which it usually is. The 'numerized representations' of fingerprints may sometimes be two-dimensional patterns, though, of certain 'minutiae' that are characteristic of them: see [4] for a discussion. But no grave consequences follow for the calculations done here: the specified probabilities of distances falling below or above a threshold will have to be replaced, only, with the chances of spatial patterns matching either too closely, on the one hand, or not matching closely enough, on the other.

---

hans varghese mathews, Bangalore, 10/2012

references

[1] *Technical Testing and Evaluation of Biometric Identification Devices,* James L. Wayman
Technical Report, for the US National Biometric Test Center, San Jose State University,
San Jose, California, U.S.A.

[2] *Evaluation of Biometric Identification Systems,* William Barrett
Summary, for the US National Biometric Test Center, San Jose State University,
San Jose, California, U.S.A.

[3] *National Biometric Test Center Collected Works, 1997-2000,* edited by James L. Wayman
US National Biometric Test Center, San Jose State University, San Jose, California, U.S.A.

[4] *Biometrics: A Tool for Information Security,* Anil K. Jain, Arun Ross, Sharath Pankanti
IEEE Transactions on Information Forensics and Security, Volume 1, N0. 2, June 2006.