Preprint of article subsequently published:

Ulery BT, Hicklin RA, Kiebuzinski GI, Roberts MA, Buscaglia J (2013) Forensic Science International. 226(1):106-117. (http://dx.doi.org/10.1016/j.forsciint.2013.01.012)

Understanding the sufficiency of information for fingerprint value determinations

Bradford T. Ulery,[a] R. Austin Hicklin,[a] George I. Kiebuzinski,[a] Maria Antonia Roberts,[b] and JoAnn Buscaglia[c] *

[a] Noblis, 3150 Fairview Park Drive, Falls Church VA 22042

[b] Latent Print Support Unit, Federal Bureau of Investigation Laboratory Division, 2501 Investigation Parkway, Quantico VA 22135

[c] Counterterrorism and Forensic Science Research Unit, Federal Bureau of Investigation Laboratory Division, 2501 Investigation Parkway, Quantico VA 22135

* Corresponding author: joann.buscaglia@ic.fbi.gov, 703-632-4553, 703-632-4557 (fax)

Abstract

A latent print examiner's assessment of the value, or suitability, of a latent impression is the process of determining whether the impression has sufficient information to make a comparison. A "no value" determination preemptively states that no individualization or exclusion determination could be made using the impression, regardless of quality of the comparison prints. Factors contributing to a value determination include clarity and the types, quantity, and relationships of features. These assessments are made subjectively by individual examiners and may vary among examiners. We modeled the relationships between value determinations and feature annotations made by 21 certified latent print examiners on 1850 latent impressions. Minutia count was strongly associated with value determinations. None of the models resulted in a stronger intraexaminer association with "value for individualization" determinations than minutia count alone. The association between examiner annotation and value determinations is greatly limited by the lack of reproducibility of both annotation and value determinations.

Keywords: latent fingerprint | fingermark | biometrics | friction ridge examination | fingerprint value | fingerprint analysis

Introduction

Assessment of the value, or suitability, of a latent fingerprint* is the process by which a latent print examiner determines if the impression has sufficient information to make a comparison. In the Analysis, Comparison, Evaluation, and Verification (ACE-V) methodology [5,6], value is assessed prior to the Analysis stage (to determine if the impression is suitable for collection), and during the Analysis stage [7]. "The assessment is made based on the quality of features (clarity of the observed features), the quantity of features (amount of features and area), the specificity of features, and their relationships" [8,9]. Operationally, the key role of the latent value assessment is to decide whether to proceed with a comparison: a "no value" (NV) determination is a preemptive assertion that no individualization or exclusion determination could be made using the impression, regardless of the quality of the

* *Regarding the use of terminology — "latent print" is the preferred term in North America for a friction ridge impression from an unknown source, and "print" is used to refer generically to known or unknown impressions [1]. We recognize that outside of North America, the preferred term for an impression from an unknown source is "mark" or "trace," and that "print" is used to refer only to known impressions. We are using the North American standard terminology to maintain consistency with our previous and future papers in this series [2,3,4].*

comparison prints. If an inappropriate NV determination is made, then the opportunity to make an individualization or exclusion conclusion is lost (a *missed conclusion*); an inappropriate determination that an impression is "of value" wastes examiner time on fruitless comparisons. An inappropriate NV determination results in the failure to bring evidence to light, but does not result in an erroneous individualization or exclusion conclusion.

The objective of this study is to describe how image clarity and feature content are associated with the assessment of latent value by latent print examiners. Our motivations for studying the associations between latent annotation and value determinations are to understand the variability in latent print examiners' value determinations, to determine if there is a basis for defining value based on the clarity and quantity of features, and to develop a basis for understanding sufficiency for comparison determinations.

In making a value determination, a latent print examiner assesses whether an impression is of value as evidence, thereby containing sufficient information to make an individualization or exclusion determination, assuming a suitable exemplar may be available. Prints considered suitable for individualization are referred to as *value for individualization* (VID). Prints with information insufficient for individualization but sufficient for exclusion are considered of *value for exclusion only* (VEO). NV indicates that there is insufficient information for either individualization or exclusion. Agency policy often reduces these three categories into two, either by combining VID and VEO into a *value for comparison* (VCMP) category, or by combining VEO with NV into a *not of value for individualization* (Not VID) category [8, survey in 2].

There are no formal criteria for making value, individualization, or exclusion determinations in the U.S.: latent print examiners use their knowledge and experience rather than a quantitative standard to make these determinations. Current guidelines published by the Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST) specify that "agency policy should define what constitutes a latent print 'of value'" [10]. A value determination is an assessment based on the information content of the latent print alone, whereas individualization or exclusion determinations are based on the information in both the latent and exemplar prints. The sufficiency of information for value determinations corresponds to that for comparison determinations: a value determination assumes an idealized case in which sufficient corresponding (or non-corresponding) information may be found in the exemplar to support an individualization or exclusion conclusion.

Some countries use a minimum minutia count standard as a requirement for individualization determinations to be used in court. A 2011 survey of 73 countries by INTERPOL found that 44 countries use a minimum point standard for identification[†] determinations: although minutia thresholds vary from 4 to 16 points, 24 of those countries have a 12-minutiae minimum [12]. The rules for counting points and the operational and legal implications of point thresholds vary by country: for example, in Spain, the point thresholds are adjusted when "unusual" points are present [13]; in the Netherlands, a higher point threshold is required to testify to an identification in court than for an identification within the agency [14]. The U.S. and the U.K. previously used point count standards, but abandoned them in favor of a non-numeric, holistic approach [15, 16]. In 1973, the International Association for Identification (IAI) resolved that "No valid basis exists at this time for requiring that a pre-determined minimum number of friction ridge characteristics must be present in two impressions in order to establish positive identification." [17] The abandonment of numeric thresholds in the U.S. was recognition that minutia counts alone are inadequate to determine sufficiency for a comparison, without inclusion of other features, clarity, and relationships among features. In the U.S., point counts still may be used in an informal manner: for example, Budowle et al. report that some examiners require a minimum of seven points before proceeding with a comparison [18]. This practice may be a result of common training methods or, for older latent print examiners, memory of when there was a standard numeric threshold. When there are no formal criteria for making value determinations, the only way to assess whether or not a specific value determination is appropriate is by agreement among latent print examiners.

Some agencies use an additional value determination, *"of value for AFIS"*, to indicate that the latent print has sufficient information to be searched in an Automated Fingerprint Identification System (AFIS). For agencies that use this designation, prints designated of value for AFIS must have a greater amount of information than would be required

[†] *Although "individualization" is synonymous with "identification" for latent print conclusions in the U.S. [11], when referring to international standards and usage we use the term "identification."*

for a manual individualization. The NIST ELFT-EFS evaluation [19] analyzed the relationship between examiners' value determinations and the accuracy of latent fingerprint matchers. The results showed that several of the matchers tested successfully matched many latent prints that examiners evaluated as VEO or NV: the three most accurate matchers successfully matched 8-20% of NV latent prints at rank 1 (n=25), and 28-35% of VEO latent prints at rank 1 (n=113); the fingerprints in that evaluation are a subset of those used in this study. These results do not support the need for a distinct "of value for AFIS" determination that is a subset of VID, for current state-of-the-art latent matchers.

Despite the importance of value assessments, there is little published literature to establish a scientific basis for value determinations. A recurring observation is that the results of the Analysis stage differ among latent print examiners: prior studies have shown substantial inter- and intraexaminer variation in minutia counts [e.g. 15, 20, 21, 22], as well as value determinations [4]. Some prints are associated with lower repeatability and reproducibility of minutia counts [22] and value determinations [2, 4]. To bring greater objectivity to the analysis of latent prints, Langenburg et al. assessed the use of latent image quality annotations using multiple latent print examiners to achieve consensus on annotated minutiae. While they found that the use of consensus and an image quality annotation tool provided greater consistency in minutia selection, they reported that these measures produced limited influence on the comparison determinations [20].

Our current study builds on the "Black Box" study [2, 4], which analyzed the repeatability (intraexaminer) and reproducibility (interexaminer) of value assessments. Figure 1 shows that value determinations were only unanimous on some of the impressions, and that much of the lack of reproducibility was associated with those prints on which individual examiners did not repeat their own decisions. The extent of agreement is a function of data selection: for example, a greater proportion of unusable or pristine prints would increase the proportion of unanimous determinations. In that study, the proportion of value determinations that examiners repeated (using the three categories VID, VEO, NV) was 0.846 after a period of months; the proportion reproduced by a different examiner was 0.757; very few value determinations changed between VID and NV. The Black Box study provided an indication of the contribution of value determinations to the rates of missed conclusions: when one examiner made an NV determination the proportion of times that a second examiner made an exclusion or individualization determination using the same latent print was 0.055; for VEO determinations the proportion of individualization determinations was 0.089. Although it was demonstrated that some prints are associated with lower repeatability and reproducibility of value determinations, that previous work did not address how the characteristics of those prints are associated with value determinations, and thereby did not provide a means to identify which prints are less likely to have reproducible value determinations.

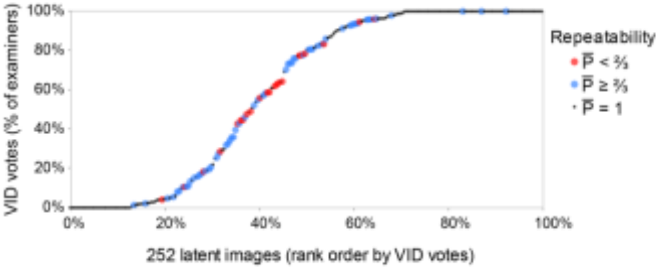


Figure 1. Repeatability and reproducibility of VID determinations (VID vs. Not VID). Proportion of examiners rating each latent VID (y-axis), in rank order (x-axis), color-coded by repeatability. (previously published result [19])

The objective of the current study is to relate examiners' latent value determinations to the clarity and quantity of annotated features, specifically, to determine how well value determinations can be explained by measures of clarity and quantity.

Materials and Methods

This study analyzed extended friction ridge feature annotation and value determinations for 1850 latent fingerprints; the impressions, annotations and value determinations were the products of previous research studies [2, 3, 19, 23]. The mix of latent prints is a heterogeneous composite from six sources: 58% were drawn from casework, the remainder was provided by volunteers in laboratory settings. Substrates and processing methods varied by source.

When the impressions were originally selected for annotation, latent print examiners rated each latent print using an informal quality scale of "Excellent," "Good," "Bad," "Ugly" and "Unusable" ("GBU" scale) in order to balance the proportions of Good, Bad, and Ugly prints; few Excellent or Unusable prints were included (Appendix, Table S3). All of the latent prints were digitized as 8-bit grayscale, 1000 pixels per inch, uncompressed.

International Association for Identification (IAI) Certified Latent Print Examiners annotated extended friction ridge features and made latent value determinations in compliance with the extended feature set (EFS) specification from the ANSI/NIST 2011 standard [24].[‡] The examiners used the Universal Latent Workstation (ULW) [25] to manually annotate the images and record their value determinations. Twenty-one examiners performed the annotations; six of the examiners annotated 53% of the impressions. Since most of the annotations were produced by a small number of examiners, our results may not be broadly representative of all examiners. In general, individual examiners performed the annotations; however, for 17% of the impressions, groups of four examiners collaborated in an attempt to produce the best possible annotation of each print.

The value of latent prints was assessed using EFS categories: "Value" (VID), "Limited" (VEO), and "No value" (NV) [see EFS Field 9.353, Examiner analysis assessment, 24]. Examiners were instructed to annotate all features they observed in the latent prints in accordance with specific guidelines [26]: the features annotated included minutiae, cores, deltas, dots, incipient ridges, ridge edge protrusions, pores, creases, scars, and dysplasia; examiners also indicated pattern classification and orientation. Image clarity was annotated according to the EFS definitions shown in Figure 2 (EFS Field 9.308, Ridge quality/confidence map). EFS and the annotation guidelines do not use AFIS vendor-specific rules for feature annotation (e.g. some AFIS annotation rules instruct examiners to ignore minutiae on short ridges or close to cores). Latent prints were annotated without reference to exemplars. Several steps were taken to ensure the quality of the annotations, including quality assurance (QA) review of the annotations; early QA findings were used to tighten the guidance to ensure uniformity across examiners. Example annotations are shown in Figure 3.

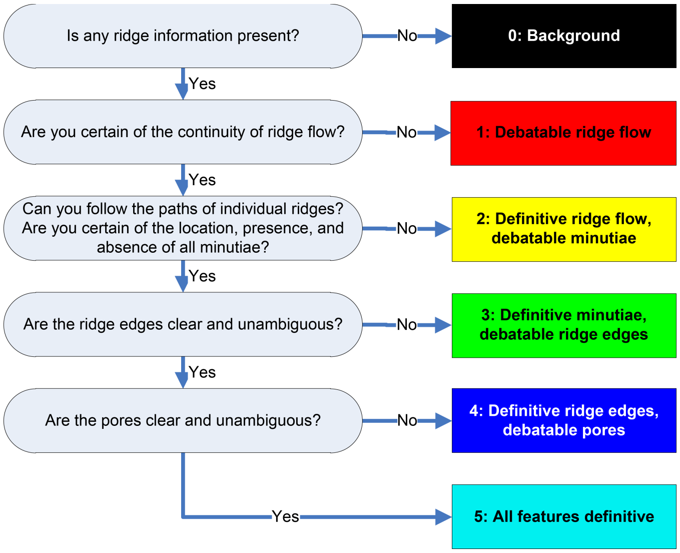

Is any ridge information present?
No
0: Background
Yes
Are you certain of the continuity of ridge flow?
No
1: Debatable ridge flow
Yes
Can you follow the paths of individual ridges? Are you certain of the location, presence, and absence of all minutiae?
No
2: Definitive ridge flow, debatable minutiae
Yes
Are the ridge edges clear and unambiguous?
No
3: Definitive minutiae, debatable ridge edges
Yes
Are the pores clear and unambiguous?
No
4: Definitive ridge edges, debatable pores
Yes
5: All features definitive

Figure 2. EFS annotation decision process for friction ridge clarity [21]

[‡] *Annotation was based on the 15 Sept 2008 draft of the EFS specification, which has since been incorporated into the ANSI/NIST 2011 standard.*

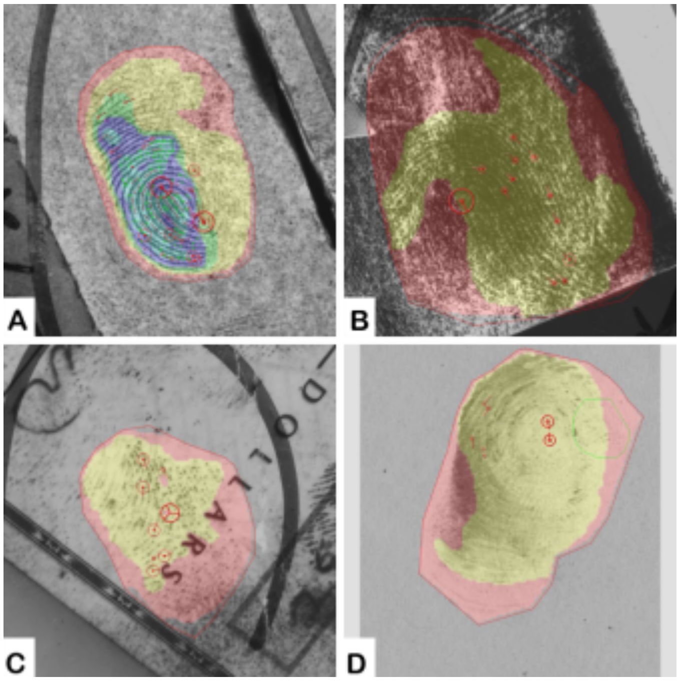


Figure 3. Examples of images with latent print examiner annotation. For each print, value determinations were available from multiple examiners and the annotation was available from one examiner. Examiners made the following value determinations for these four prints: A (100% VID; 5 examiners); B (57% VID, 14% VEO, 29% NV; 7 examiners); C (14% VID, 14% VEO, 71% NV; 7 examiners); D (2% VID, 47% VEO, 51% NV; 59 examiners).[§] For these four latent prints, the specific clarity and feature annotations shown here were performed by examiners who assessed each latent as VID. All four prints are from the DS1850 dataset; print D is also in the BB166 dataset.

For each of the 1850 latent prints, the value determination was provided by the same examiner(s) who provided the annotations; this dataset (*DS1850*) was used to analyze intraexaminer associations between annotations and value assessments. Among the 1850 latent prints are a subset of 421 prints on which a second examiner independently provided annotations and value assessments, following the same procedures and guidelines; this dataset (*DS421*) was used to analyze interexaminer agreement on latent annotations and value assessments. Also among the 1850 latent prints are 166 prints with a mean (also median) of 56 assessments each from the Black Box Study [2]; this dataset (*BB166*) was used to analyze interexaminer associations between a single examiner's annotation and multiple examiners' value assessments. The fingerprints in DS421 and BB166 are not representative of the latent prints in DS1850 because they were selected from specific sources. For further description of the DS1850 dataset, see Appendix, Table S1 through Table S3.

We derived numerous metrics from the annotations of the latent prints, including counts of the various feature types, and various area measures of image clarity [23]. We explored associations between these metrics (independent variables) and value determinations (dependent variables) using conventional modeling and analysis techniques such as logistic regression and recursive partitioning; graphical visualization techniques were used to explore the data for non-linear associations among the variables, including patterns associated with subsets of the data.

[§] *Examiners' value determinations for prints A-C were from our Latent Quality Survey [**Error! Bookmark not defined.**]. These specific images were selected because they are publicly releasable.*

Results

Figure 4 shows intraexaminer associations between the number of minutiae annotated and value assessments. Notice that there are no sharp decision thresholds (specific minutia counts that divide one type of determination from another). There are unexpectedly high-count VEO determinations (ranging up to 27 minutiae), unexpectedly low-count VIDs (down to 0 minutiae), and unexpectedly high-count NVs (up to 12 minutiae); reproducibility of these is discussed below.

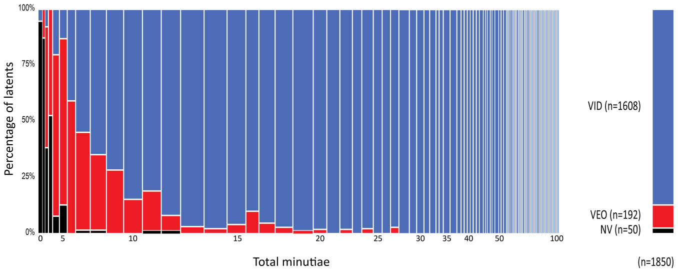


Figure 4. Intraexaminer association of minutia count and value assessments. Mosaic plot represents the percentage of each type of determination associated with each count (DS1850 dataset, n=1850 latent prints)

Figure 5 shows the strength of association between an examiner's minutia count and that examiner's VID determinations using a receiver operating characteristic (ROC) curve. This curve describes the data presented in Figure 4, but in terms of the error trade-offs that would have resulted from using an examiner's minutia count to predict that examiner's VID determinations. For example, at a threshold of 12 or more minutiae (the national standard in many countries), 84% of the examiners' VID determinations would have been successfully predicted, but 12% of the NV and VEO determinations would have been predicted to be VID. The ROCs presented in this paper provide meaningful comparisons of models on our data, but should not be extrapolated beyond this data: a different distribution of data could substantively change the rates shown in these graphs.

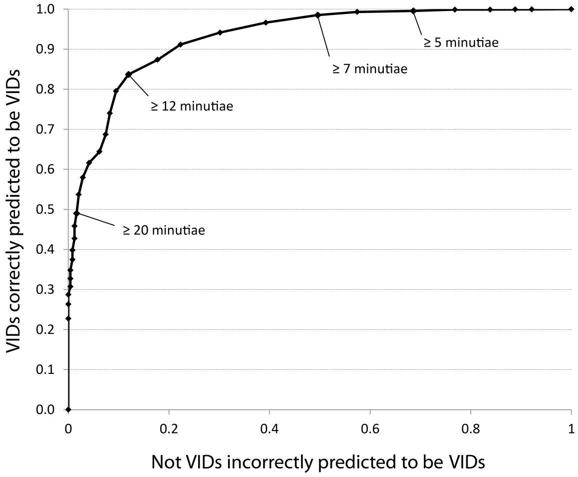


Figure 5. Receiver operating characteristic (ROC) curve showing the accuracy of a simple model using an examiner's minutia count as a predictor of that examiner's VID determination. The x-axis shows the false positive rate (1-specificity); the y-axis shows the true positive rate (sensitivity). Minutia counts are indicated as dots; four counts are labeled as examples. (DS1850 dataset, n=1850)

Budowle et al. [18] discussed an informal threshold of seven or more minutiae used by some examiners to "proceed with an analysis." In our data, while there is clearly no sharp decision threshold at seven minutiae, 1% of VID (and 50% of Not VID) determinations were made on latent prints with fewer than seven minutiae.

We considered the possibility that value assessments might be strongly associated with minutia counts for individual examiners, but that decision thresholds might vary by examiner. Although the number of minutiae an examiner annotated is strongly associated with that examiner's value determination, our analyses show no evidence that these examiners' value determinations were based solely on minutia count.

Figure 6 shows results from several intraexaminer models predicting VID determinations from metrics derived from latent annotations (summary statistics for additional models are listed in Appendix, Table S4 and Table S5). When comparing alternative models using ROCs, stronger associations result in operating points closer to the upper left corner; a lack of any association would result in a diagonal line from the top upper right corner to the bottom lower left corner. While most of the metrics had some predictive capability, none of the individual metrics approached minutia count as a predictor of VID determinations. We used logistic regression and recursive partitioning to explore each pairwise combination of metrics with respect to VID determinations; none of these combinations provided significant discriminatory power beyond minutia count alone. For example, adding terms such as total area annotated as green or higher clarity, or counts of cores and deltas, did not improve the model; likewise separately weighting debatable minutiae (those in areas of yellow clarity) and definitive minutiae (those in areas of green or higher clarity) did not improve upon the minutia count model (shown in Figure 6 as an example; see Appendix, Table S5 for other combinations). We examined scatterplots for non-linear relations, and explored various logical subsets of the data, all with similar negative results. One important factor contributing to these results is the lack of statistical independence among the annotation metrics: minutia count was strongly associated with value determinations and most of the other annotation metrics were strongly associated with minutia count.

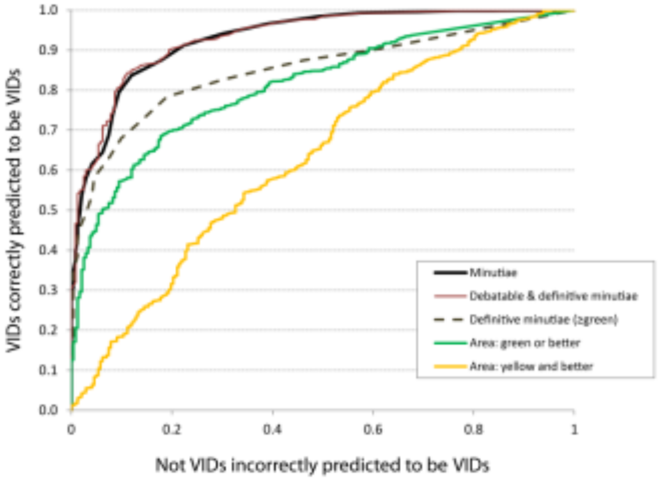


Figure 6. Comparison of intraexaminer logistic regression models predicting VID determinations from examiner annotation. (DS1850 dataset; n=1850 latent prints, 242 of which were rated Not VID)

Figure 7 shows results from several intraexaminer models predicting value for comparison (VCMP) determinations (summary statistics for additional models are listed in Appendix, Table S4 and Table S5). The results are similar to those for VID determinations, but the effective sample size is much smaller (only 50 latent prints were rated NV), hence the resulting ROC is more jagged. The minutia count model corresponds to the separation of NV from VCMP (VEO and VID) in Figure 4. Since a VCMP determination is based on the ability to exclude as well as individualize potential comparisons, factors such as pattern classification, cores, and deltas provide a greater level of discrimination for VCMP than for VID determinations.

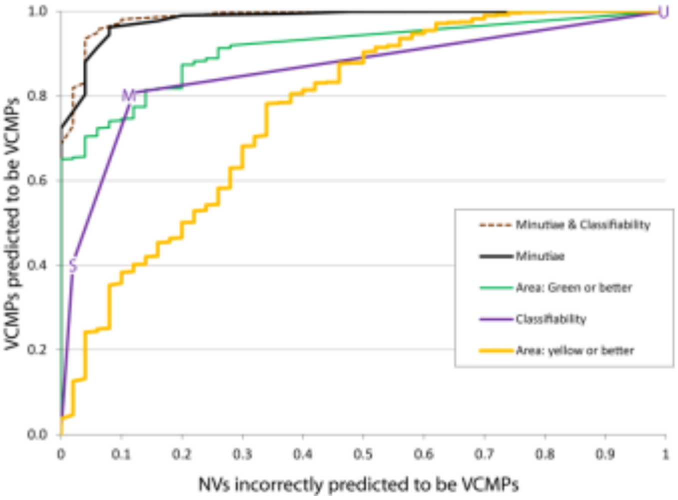


Figure 7. Comparison of intraexaminer logistic regression models predicting VCMP determinations from examiner annotations. The x-axis shows the false positive rate (1-specificity); the y-axis shows the true positive rate (sensitivity). Classifiability (labeled as U: unclassifiable; M: multiple possible pattern classes; S: single class) is the extent to which examiners were able to specify an impression's pattern classification, out of the eight subpattern types. (DS1850 dataset; n=1850 latent prints, 50 of which were rated NV)

As seen in the results of the Black Box study [2], one examiner's value determinations are not always reproduced by other examiners. The BB166 dataset provides an opportunity to model the association between one examiner's annotation and the value determinations of other examiners. Figure 8 shows the extent to which agreement on value decisions is associated with minutia count. No prints with one or more minutiae annotated were unanimously rated NV; only one print with more than nine minutiae annotated was rated NV by a majority of examiners. No prints were unanimously rated VEO; only one print was rated VEO by more than 75% of examiners. No prints with fewer than ten minutiae annotated were unanimously rated VID; only two prints with fewer than ten minutiae annotated were rated VID by a majority of examiners.

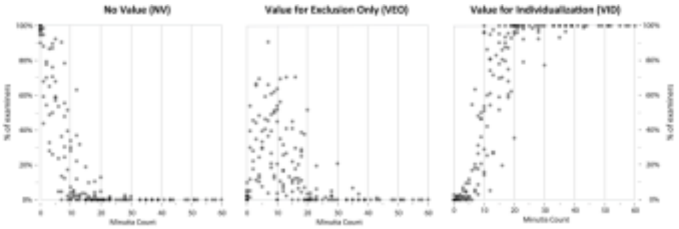
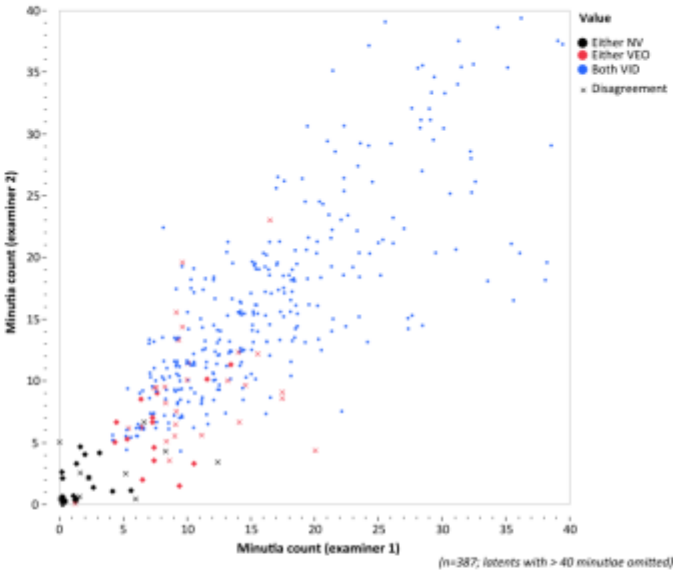


Figure 8. Reproducibility of value determinations (mean of 56 examiners) by minutia count (single examiner). Y-axis indicates the percentage of examiners who assigned each latent to the category indicated in the chart title (charts omit 4 latent prints with > 60 minutiae). (BB166 dataset)

Previous research has shown that examiners vary substantially in their minutia counts. We expect that this variability accounts for much of the dispersion in Figure 8. To investigate this further, we analyzed reproducibility in the DS421 dataset. Figure 9 shows associations between the value determinations and minutia counts made by pairs of examiners on the DS421 dataset. Examiners often disagreed substantially in their minutia counts: the standard deviation for the difference in minutia counts is 4.0 among latent prints with a mean minutia count of 5 to 15. The dispersion (disagreement on minutia counts) increases as the number of minutiae increases, roughly in proportion to the square root of the mean minutia count. Dispersion is substantially higher among those latent prints on which examiners disagreed on the latent value than among those where examiners agreed. When examiners disagreed on latent value, the examiner making the higher value assessment usually counted more minutiae (n=37; p=0.001, one-sided): the mean difference in minutia count was 2.5 (std. dev. 4.7); the mean minutia count was 8.1 (std. dev. 5.2).



(n=387; latents with > 40 minutiae omitted)

Figure 9. Variation in value determination between two examiners assessing the same latent (DS421 dataset; n=421 latent prints – 387 shown due to truncating axes at 40 minutiae)

To better understand the wide spread in minutia counts associated with value decisions in the DS1850 dataset, a 3-examiner panel reviewed the annotations and value determinations for about 5% of the data flagged as potential anomalies, including NV or VEO prints with high minutia counts, VID prints with low minutia counts, and NV prints with pattern class, core, or deltas. The panel was asked to perform a review of the annotations and value determinations, and to give the benefit of the doubt to the initial examiner. The results from this review were consistent with our expectations, based on the limited reproducibility observed in the Black Box study. The panel considered the value or annotation to be incorrect for 66 of the 98 prints, debatable for an additional 18 prints, and concurred with both value and annotation for only 14 of the 98 prints reviewed. For the 24 VID latent prints with fewer than 7 minutiae, the panel considered the value or annotation to be incorrect or debatable for all 24; the panel considered the value or annotation to be incorrect or debatable for 40 of the 41 VEO latent prints with more than 10 minutiae.

Figure 10 summarizes the strength of association of one examiner's annotation and the value determinations of other examiners. The "value" ROC shows the effectiveness of using examiners' value determinations to predict the value determinations of other examiners; for example, the annotating examiners' VID determinations successfully predicted 95% of the Black Box VID determinations, but also predicted 31% of the Black Box Not VID determinations. The informal GBU scale is as effective as minutia count in predicting VID determinations. The interexaminer limit curve (dashed red) describes a logistic regression model with 166 parameters, one for each latent print. This model accounts for all of the variability in value determinations that can be attributed to the impressions themselves; the remaining variability arises from examiner disagreements on their value determinations. This model represents an upper limit to what might be achieved in any intraexaminer model derived from the annotations of these prints. Therefore, even models that account for the specificity or relationships of features, or additional feature or clarity-based metrics would not exceed this limit. The minutia model has an equal error rate (where FPR equals 1-TPR) of 15%, whereas the red curve shows that the equal error rate cannot be less than 10% for any model of this data that is based on latent print characteristics alone: two-thirds of the residual error resulting from the minutia model is explained by examiner variability on value determinations. The latent prints in this dataset, although a subset of those in DS1850, are not representative of those in DS1850 as they are derived entirely from laboratory-collected prints. For this reason, the strength of associations for the ROCs in Figure 10 are not directly comparable to those in Figure 6.

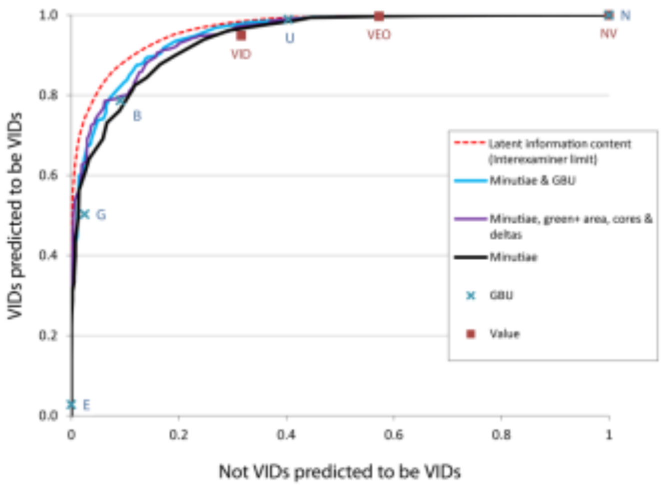


Figure 10. Predicting independent value assessments for VID determinations. These ROCs compare models for predicting a second (independent) value assessment {Not VID, VID} from an initial annotation and value assessment (n=166 latent prints with annotations and value assessments, predicting 9,322 independent value assessments). The "interexaminer limit" describes a logistic model having one parameter for each of the 166 latent prints; this shows the limit imposed by interexaminer variation on predicting value determinations. The “value” ROC uses the initial examiner’s value determination to predict the independent value assessments. See Figure S2 for corresponding VCMP results. (BB166 dataset)

Discussion

This study presents an analysis of the associations between the value determinations made by latent print examiners and examiners’ annotation of minutiae, other features, and image clarity. The results show a strong association between minutia count and value determinations. Most of the variability that could not be explained by minutia count can be attributed to the lack of reproducibility of determinations among examiners. Value determinations vary significantly among examiners, in confirmation of previous studies: one examiner’s NV may be another’s “Ugly” (poor-quality VID), or VEO; all VEO prints are likely to have low reproducibility, as are VID prints with low minutia count and NV prints with high minutia count.

There was no evidence that examiners made value determinations based solely on fixed minutia count thresholds. A variety of metrics were analyzed with respect to value determinations, including counts of various features and image clarity metrics. Examiner disagreements accounted for two-thirds of the residual error resulting from the minutia-only model. Having accounted for those two factors (minutia count and examiner disagreements), the theoretical potential for other metrics to contribute is relatively small. None of the metrics other than minutia count provided significant additional discriminatory power for VID determinations. For discriminating between NV and VEO determinations, pattern classifiability and counts of cores and deltas were effective in combination with minutia count. Some features were less effective than minutia count because they were associated with high minutia count prints in our data (such as incipient ridges), or were found in prints regardless of value (such as creases).

A surprising result of the study was the failure of image clarity metrics to improve substantially on the minutia count model. The decision space for analysis and comparison determinations has been described as having two dimensions: quality and quantity of features [8, 27]. We would therefore expect latent value determinations to depend on both dimensions, so that VID determinations could be associated with high clarity, low minutia count prints as well as low clarity, high minutia count prints. In our data, we found a strong association between clarity (quality) and minutia count (quantity): although print clarity was strongly associated with value determinations, clarity metrics provided little additional discriminatory power beyond minutia count alone. For example, areas of clear level-3 detail (e.g. ridge edge details) were only present in prints with high minutia count. Several factors might account for the failure of clarity metrics to substantially complement minutia count: clarity and quantity of features were highly associated in our data; the metrics used may not fully capture clarity in the ways that examiners use that information; examiners may not have used the clarity categories consistently in their annotations; our data included a relatively small sample size of NV and VEO decisions, limiting our ability to observe small effects; and the general lack of repeatability and reproducibility of both annotations and value determinations limits the potential for improving models beyond minutiae alone.

Our study does not address whether examiners' value determinations are correct: it only reveals patterns of association. In theory, the correctness of value determinations could be based on whether examiners could subsequently make correct comparison conclusions given a suitable exemplar; however, this is likely to be impractical given the variability in examiners' comparison determinations [2, 4], and the impracticality of determining the correctness of the comparison conclusions. Instead of correctness, we evaluated the appropriateness of individual value determinations based on consensus among latent print examiners. There may be situations where examiner subjectivity in value determinations is acceptable: for example, a skilled examiner may make a VID determination for an impression that would be far beyond the expertise of a junior examiner to compare. Should examiners make value judgments based on their own skill levels, or based on their expectation of other examiners' skills? If a forensic laboratory intends to report reproducible value determinations regardless of the examiner assigned to the case, then that examiner would have to predict the value determinations that would be made by other examiners.

Our results indicate that the value of latent prints is a continuum that is not well described by binary (value vs. no value) determinations. Additional means of describing value (such as an indication that an impression is "complex" [8], the GBU scale, or a clarity/quality metric [e.g. 23]) may be useful in flagging prints whose value determinations are likely to be debatable. Such means of expressing value could be used in establishing business processes to manage risk and optimize workload based on value: for example, a quality assurance process could require review of value determinations for ugly or complex prints, direct such prints to highly qualified examiners, or require rigorous verification when such prints are used in comparison.

One should not expect that value determinations on complex prints will be highly reproducible. Frequently, NV determinations are not verified; although inappropriate VID determinations will often be detected by verification of the subsequent comparison determinations, inappropriate NV determinations will not be detected, potentially resulting in missed conclusions.

To further our understanding of the accuracy and reliability of the latent print examination process, we are developing fingerprint quality and quantity metrics; exploring how complexity of background, substrate and processing are related to comparison determinations; and extending our analyses to include detailed examiner annotation of feature correspondence to analyze the relationships of the quality and quantity of features to comparison determinations. The results and lessons learned from this study were used in the design of and data selection for our ongoing study of the relationship between comparison conclusions and the quality and quantity of features annotated by examiners.

References

1 Scientific Working Group on Friction Ridge Analysis, Study and Technology (2011) Standard terminology of friction ridge examination, ver.3.0. (http://swgfast.org/documents/terminology/110323_Standard-Terminology_3.0.pdf)

2 Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2011) Accuracy and reliability of forensic latent fingerprint decisions. Proc Natl Acad Sci USA 108(19): 7733-7738. (http://www.pnas.org/content/108/19/7733.full.pdf)

3 Hicklin RA, et al (2011) Latent fingerprint quality: a survey of examiners; J. Forensic Identification, 61(4): 385-418.

4 Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2012), Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners. PLoS ONE 7:3. (http://www.plosone.org/article/info:doi/10.1371/journal.pone.0032800)

5 Huber RA (1959) Expert witness. *Criminal Law Quarterly* 2:276–296.

6 Ashbaugh D (1999) Quantitative-Qualitative Friction Ridge Analysis: An Introduction to Basic and Advanced Ridgeology (CRC Press, New York).

7 Expert Working Group on Human Factors in Latent Print Analysis (2012) Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach; NIST Interagency Report 7842. (http://www.nist.gov/customcf/get_pdf.cfm?pub_id=910745)

8 SWGFAST (2011) Standards for Examining Friction Ridge Impressions and Resulting Conclusions, ver 1.0. (http://swgfast.org/documents/examinations-conclusions/110913_Examinations-Conclusions_1.0.pdf)

9 Locard E (1920) L'enquête criminelle et les méthodes scientifiques, Ernest Flammarion, Paris, pp. 129-130. (http://babel.hathitrust.org/cgi/pt?id=njp.32101068784956)

10 SWGFAST (2010) Standard for the Documentation of Analysis, Comparison, Evaluation, and Verification (ACE-V) (Latent), ver 1.0.

11 SWGFAST (2012) Individualization / Identification Position Statement, ver 1.0. (http://swgfast.org/Comments-Positions/120306_Individualization-Identification.pdf)

12 Farelo A (2012) "Fingerprints Survey 2011" [Conference presentation, 7th International Symposium on Fingerprints]

13 Exposito Marquez N (2012) La Acreditacion del Metodo de Identificacion Lofoscopico en la Guarda Civil. 7th Intl Symp on Fingerprints (conference presentation)

14 Riemen (2012) "Netherlands Case Study: Auto encoding of latents". 7th Intl Symp on Fingerprints (conference presentation)

15 Evett IW, Williams RL (1996) A Review of the Sixteen Point Fingerprint Standard in England and Wales, *Journal of Forensic Identification*, 46 (1), January/February, 1996 [Also published in *Fingerprint Whorld, 21 (82), October, 1995*]

16 Campbell (2011) The Fingerprint Inquiry; Part 6, The Law and Practice of Fingerprints - Chapter 32 The Sixteen Point Standard; 14 December 2011. (http://www.thefingerprintinquiryscotland.org.uk/inquiry/3127.html)

17 Report of the Standardization Committee of the International Association for Identification; *Identification News*; Aug. 1973; p 13. (http://www.latent-prints.com/images/IAI%201973%20Resolution.pdf)

18 Budowle B, Buscaglia J, Schwartz Perlman R (2006) Review of the Scientific Basis for Friction Ridge Comparisons as a Means of Identification: Committee Findings and Recommendations, *Forensic Science Communications*, Vol. 8. (http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/jan2006/research/2006_01_research02.htm)

19 Indovina M, Dvornychenko V, Hicklin RA, Kiebuzinski GI (2012) ELFT-EFS Evaluation of Latent Fingerprint Technologies: Extended Feature Sets, Evaluation #2; NIST Interagency Report 7859. (http://dx.doi.org/10.6028/NIST.IR.7859)

20 Langenburg G, Champod C, Genessay T (2012) Informing the Judgments of Fingerprint Analysts Using Quality Metric and Statistical Assessment Tools, *Forensic Science International* 219:183–198.

21 Schiffer B, Champod C (2007) The potential (negative) influence of observational biases at the analysis stage of finger mark individualization, *Forensic Science International* 167:116–120.

22 Dror IE, et al. (2011) Cognitive issues in fingerprint analysis: Inter- and intra-expert consistency and the effect of a 'target' comparison; *Forensic Science International* 208:10-17.

23 Hicklin RA, Buscaglia J, Roberts MA (2012) Assessing the Clarity of Friction Ridge Impressions (submitted to *Forensic Science International*)

24 ANSI/NIST-ITL 1-2011, NIST Special Publication 500-290 Data Format for the Interchange of Fingerprint, Facial & Other Biometric Information. (www.nist.gov/itl/iad/ig/ansi_standard.cfm)

25 Federal Bureau of Investigation, Universal Latent Workstation (ULW). (www.fbibiospecs.org/Latent/LatentPrintServices.aspx)

26 Hicklin RA (2009) Guidelines for Extended Feature Set Markup of Friction Ridge Images; Working Draft Version 0.3, 12 June 2009. *[Note: This document has been formalized in "Markup Instructions for Extended Friction Ridge Features", version 1.0, March 2012. (http://www.noblis.org/interop)]*

27 Vanderkolk J (2009) Forensic Comparative Science: Qualitative Quantitative Source Determination of Unique Impressions, Images, and Objects. Academic Press.

Appendix – Supplemental Information

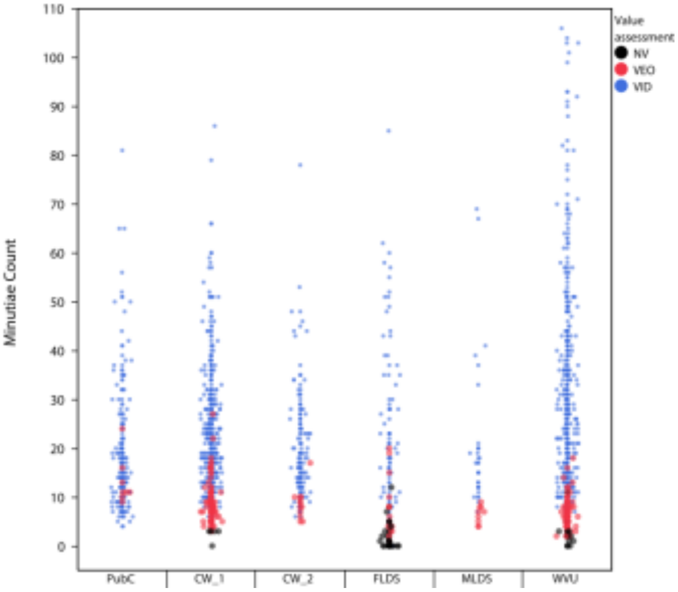


Figure S1: Multiple source datasets with distinct characteristics. 58% of the latent prints are from casework (PublicChallenge[5] [1,2], Casework1, Casework2); 42% laboratory-collected from volunteers (MLDS, FLDS, WVU). Substrates and processing methods varied by source. (DS1850 dataset, n=1850)

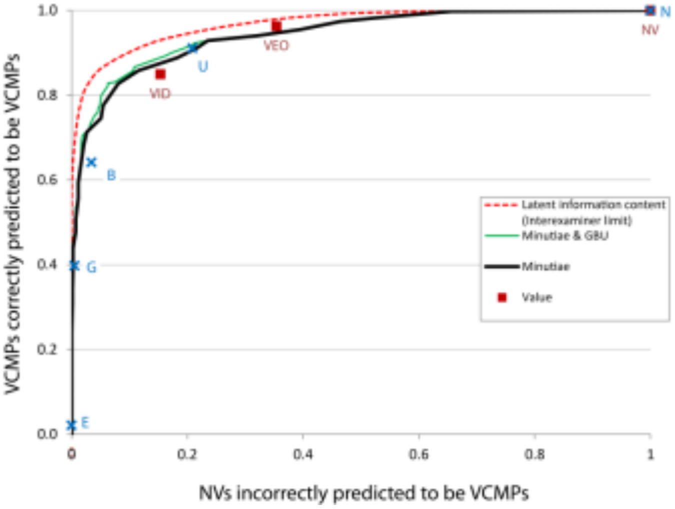


Figure S2. Predicting independent value assessments for VCMP determinations. These ROCs compare models for predicting a second (independent) value assessment {NV, VCMP} from an initial annotation and value assessment (n=166 latent prints with annotations and value assessments, predicting 9,322 independent value assessments). The "interexaminer limit" describes a logistic model having one parameter for each of the 166 latent prints. See Figure 10 for corresponding VID results.

[5] *The ELFT-EFS Public Challenge Dataset includes 255 latent prints that were newer, higher-resolution scanned images from the same set of photographs previously used in the NIST SD27 dataset; most but not all of the public challenge latent prints were included in SD27.*

	% of	1608 VID latent prints						192 VEO latent prints						50 NV latent prints					
	all prints	% of VID prints	Min	Q1	Med	Q3	Max	% of VEO prints	Min	Q1	Med	Q3	Max	% of NV prints	Min	Q1	Med	Q3	Max
Minutiae	98.9%	99.9%	-	13	19	29	106	100.0%	1	5	7	10	27	62.0%	-	-	1	3	12
Minutiae (debatable)	96.6%	97.8%	-	5	9	13	72	96.9%	-	3	5	7	20	58.0%	-	-	1	3	12
Minutiae (definite)	86.5%	90.3%	-	4	9	17	99	72.4%	-	-	2	3	23	18.0%	-	-	-	-	2
Cores	70.5%	74.4%	-	-	1	1	3	53.1%	-	-	1	1	2	12.0%	-	-	-	-	2
Deltas	37.8%	41.2%	-	-	-	1	2	18.2%	-	-	-	-	2	2.0%	-	-	-	-	1
Dots	21.0%	23.2%	-	-	-	-	52	6.3%	-	-	-	-	6	6.0%	-	-	-	-	1
Incipient ridges	17.6%	19.7%	-	-	-	-	96	4.2%	-	-	-	-	6	2.0%	-	-	-	-	1
Creases	28.3%	29.5%	-	-	-	1	39	19.8%	-	-	-	-	21	22.0%	-	-	-	-	16
Protrusions	3.0%	3.4%	-	-	-	-	43	0.5%	-	-	-	-	1	0.0%	-	-	-	-	-
Pores	60.0%	62.6%	-	-	24	100	1267	50.0%	-	-	1	34	396	16.0%	-	-	-	-	398

Table S1: Distribution of feature counts by value determination in the DS1850 dataset. "% of prints" indicates the percentage of latent prints with non-zero counts of each feature; the other columns describe the distribution of feature counts across all latent prints (minimum, lower quartile, median, upper quartile, maximum). Debatable minutiae are those found in areas of yellow clarity; definite minutiae are those found in areas of green or higher clarity.

	% of	1608 VID latent prints						192 VEO latent prints						50 NV latent prints					
	all prints	% of VID prints	Min	Q1	Med	Q3	Max	% of VEO prints	Min	Q1	Med	Q3	Max	% of NV prints	Min	Q1	Med	Q3	Max
Red+ area	100.0%	100.0%	43	243	334	442	1513	100.0%	60	229	327	446	1300	100.0%	99	281	384	528	914
Yellow+ area	99.7%	100.0%	28	184	257	349	1286	100.0%	25	148	223	298	659	90.0%	-	55	134	235	537
Green+ area	90.3%	93.7%	-	31	67	114	476	77.6%	-	5	21	38	166	28.0%	-	-	-	7	38
Blue+ area	10.1%	11.4%	-	-	-	-	382	1.6%	-	-	-	-	14	0.0%	-	-	-	-	-
Aqua area	5.5%	6.3%	-	-	-	-	76	0.5%	-	-	-	-	0	0.0%	-	-	-	-	-

Table S2: Distribution of areas of image clarity in the DS1850 dataset, in mm^2. Areas are cumulative: for example, "green+" (green or higher clarity) includes green, blue, and aqua areas. "Red+" corresponds to the area of the region of interest selected by the examiner. "% of prints" indicates the percentage of latent prints with non-zero areas for each clarity type; the other columns describe the distribution of feature counts across all latent prints (minimum, lower quartile, median, upper quartile, maximum). See Figure 2 for definitions of clarity levels.

	% of all prints	% of VID prints	% of VEO prints	% of NV prints
Scars	1.9%	2.1%	0.5%	2.0%
Dysplasia	0.3%	0.3%	0.0%	0.0%
Unknown orientation	7.8%	6.2%	15.1%	32.0%
Classifiability				
Specific class	39.7%	43.2%	20.3%	2.0%
Multiple classes	39.2%	39.1%	47.9%	10.0%
Unable to class	21.1%	17.7%	31.8%	88.0%
GBU				
Excellent	0.4%	0.5%	0.0%	0.0%
Good	31.2%	34.9%	7.8%	2.0%
Bad	31.6%	34.1%	17.7%	6.0%
Ugly	31.8%	27.7%	67.2%	26.0%
Unusable	2.4%	0.2%	5.7%	60.0%
Not specified	2.5%	2.5%	1.6%	6.0%

Table S3: Distribution of other attributes of the latent prints in the DS1850 dataset. Numbers represent the percentage of prints within each value category having the specified attribute. The GBU assessments were done at the time of data selection, by different examiners than the examiners who did annotation; the GBU categories were not always used consistently; the Excellent and Unusable prints came from the laboratory-collected datasets. The classifiability of an impression is based on the number of EFS pattern classes indicated by the examiner, out of eight possible classes.

	VID		VCMP	
	R^2	AUC	R^2	AUC
Total minutiae	0.469	0.930	0.661	0.982
Definitive minutiae (green+)	0.267	0.849	0.356	0.925
Green+ area	0.196	0.812	0.297	0.911
GBU	0.215	0.785	0.462	0.900
Debatable minutiae (yellow)	0.125	0.756	0.266	0.891
Total cores and deltas	0.089	0.707	0.218	0.850
Classifiability	0.069	0.682	0.239	0.860
Total cores	0.056	0.662	0.165	0.806
Pores	0.034	0.635	0.049	0.731
Yellow+ area	0.032	0.633	0.138	0.768
Total deltas	0.049	0.633	0.086	0.684
Dots	0.029	0.587	0.026	0.580
Incipient ridges	0.034	0.581	0.031	0.581
Unknown orientation	0.025	0.562	0.056	0.624
Blue+ area	0.027	0.551	0.024	0.552
Creases	0.001	0.543	0.002	0.538
Aqua area	0.019	0.530	0.013	0.529
Ridge protrusions	0.007	0.515	0.007	0.515
Scars or dysplasia	0.002	0.508	0.000	0.501

Table S4: Intraexaminer associations between annotation variables and value determinations. The table evaluates performance using the uncertainty coefficient (entropy R^2), and the area under the ROC curve (AUC) for nominal logistic fits. Note that AUC nominally ranges from 0.5 to 1.0. Various metrics derived from image clarity annotation were analyzed, including area, area of good flow, largest contiguous area, and entropy/consistency measures as described in [*3*]; none were substantively different from area in predicting value determinations.

	VID		VCMP	
	R^2	AUC	R^2	AUC
Total minutiae, total cores & deltas, classifiability	0.478	0.933	0.713	0.988
Total minutiae, yellow+ area	0.478	0.933	0.664	0.982
Definitive minutiae, debatable minutiae	0.473	0.932	0.682	0.987
Total minutiae, green+ area	0.472	0.932	0.672	0.985
Total minutiae, classifiability	0.474	0.931	0.704	0.985
Total minutiae, total cores & deltas	0.474	0.931	0.694	0.987
Yellow+ area, green+ area	0.196	0.812	0.333	0.919

Table S5: Intraexaminer associations between example combinations of annotation variables and value determinations. We modeled each pairwise combination of annotation variables with respect to VID determinations; none provided significant discriminatory power beyond minutia count alone. Adding predictors to a model *always* results in a higher R^2 and AUC.

Appendix References

1 Indovina M, Hicklin RA, Kiebuzinski GI (2011), ELFT-EFS Evaluation of Latent Fingerprint Technologies: Extended Feature Sets, Evaluation #1; NIST Interagency Report 7775. Available: http://biometrics.nist.gov/cs_links/latent/elft-efs/NISTIR_7775.pdf

2 National Institute of Standards and Technology, Fingerprint Minutiae from Latent and Matching Tenprint Images, Special Database 27. Available: http://www.nist.gov/itl/iad/ig/sd27a.cfm

3 Hicklin RA, Buscaglia J, Roberts MA (2012) Assessing the Clarity of Friction Ridge Impressions (submitted to *Forensic Science International*)