NISTIR 8207

IREX IX Part One
Performance of Iris Recognition Algorithms

George W. Quinn
Patrick Grother
James Matey

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8207

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 8207

IREX IX Part One
Performance of Iris Recognition Algorithms

George W. Quinn
Patrick Grother
James Matey
Information Technology Laboratory, NIST

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8207

April 2018

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

Executive Summary

Introduction

Iris Exchange (IREX) IX is an evaluation of automated iris recognition algorithms. The first part of the evaluation is a performance test of both verification (one-to-one) and identification (one-to-many) recognition algorithms over operational test data. The results are summarized in this report. Thirteen developers submitted recognition algorithms for testing, more than any previous IREX evaluation. Performance was measured for 46 matching algorithms over a set of approximately 700K field-collected iris images. This report is very similar to *IREX IV: Part 1, Evaluation of Iris Identification Algorithms* [1] in both format and scope.

Key Results

The key results of this part of the IREX IX evaluation are described below.

- **Core Accuracy:** Accuracy is reported for two-eye matching since most iris cameras acquire samples of both irides simultaneously. The most accurate *one-to-many* matcher yields an FNIR (False Negative Identification Rate) of 0.0067 (about 1 in 150) at an FPIR (False Positive Identification Rate) of 10^{-3} (1 in 1000) when searching against an enrolled population of 160 thousand people. An equally performing identification (i.e. one-to-many) system with the same number of enrolled people operating in access control mode would reject access for 1 in 150 valid (i.e. enrolled) users while granting access for 1 in 1000 invalid (i.e. unenrolled) users. Matchers from two other participants follow closely behind with FNIRs of 0.0081 (1 in 123) and 0.0083 (1 in 120) respectively at the same FPIR. More than half of the matchers yield an FNIR less than 0.02 at FPIR = 10^{-3}.

  The most accurate *one-to-one* matcher yields an FNMR (False Non-Match Rate) of 0.0057 (about 1 in 175) at an FMR of 10^{-5} (1 in $100\,000$). Submissions from four other participants follow closely behind with FNMRs between 0.0066 (1 in 152) and 0.0070 (1 in 143). The differences in accuracy between these matchers are unlikely to be statistically significant.

- **Error Rates:** The error rates in this evaluation are much lower than in previous IREX evaluations. There are many reasons for this. More aggressive steps were taken to mitigate ground truth errors in the current evaluation. Recognition technology has also advanced in the four years since the last IREX evaluation. Finally, the test dataset was collected with more modern two eye cameras under more cooperative conditions, yielding generally better quality samples. The accuracy of iris recognition is dominated by the small fraction of samples that suffer from significant quality-related problems (e.g. motion blur, eyelid occlusion). Such problems appear to be much less common in the current test dataset. NIST authored *IREX V: Guidance for Iris Image Collection* [2] to address the problem of poor sample quality. The document provides guidance on the proper collection, storage, and handling of iris data.

- **Matching Speed:** The fastest matcher was able to search against an enrolled population of 160 thousand with a median search duration of 11 milliseconds using just one processing core. This is faster than any matcher from IREX IV and almost 50 times faster than any other matcher in IREX IX. Search duration varies widely across matchers, with a roughly one thousand factor difference between the fastest and slowest matchers. Search duration can dictate computational hardware requirements for applications that have high search volumes.

  Search time scales approximately linearly with the size of the enrolled population for nearly every matcher. That is, a doubling of the enrolled population size approximately doubles the search time. There is one exception where the relationship is sub-linear. For this matcher, every doubling of the enrolled population size seems to increase the median search time by about 2.2 milliseconds.

  Comparison time for one-to-one comparisons also varied widely across matchers. The fastest properly functioning matcher compares two-eye templates with a median time of 0.006 milliseconds, or about $167\,000$ comparisons per second on one processing core. The slowest matcher is almost 10 thousand times slower with a median comparison time of 54.11 milliseconds. Most of the matchers compare templates with a median time under a millisecond.

  Timing statistics were collected on a Dual Intel Xeon E5-2695 v3 3.3 GHz CPU equipped with AVX instructions using just one processing core. Operationally, multiple cores (and multiple machines) could be employed to reduce these times.

- **Speed-accuracy Tradeoff:** IREX III found that a speed-accuracy tradeoff exists for some iris recognition algorithms,

where improved accuracy can be achieved through slower, but more involved, comparison strategies. For one-to-one comparisons, the fastest matchers tend not to be the most accurate. The correlation coefficient for the log of the median search time and the log of FNMR is -0.30, indicating a weak but apparent speed-accuracy tradeoff.

For one-to-many matching, a speed-accuracy tradeoff is less apparent. Nevertheless, one participant submitted a very fast matcher along with more accurate (but slower) matchers. For this participant, a roughly one-hundred-fold improvement in search time is realized but at the expense of a 37% increase in FNIR (at FPIR=10^{-3}) against an enrolled population of 160 thousand.

- **Template Creation Time:** The time it takes to create a comparable template from an iris sample (or pair of samples) can affect throughput rates (i.e. the number of transactions a system can handle per unit time). Very long template creation times can lead to backlogs at, for example, access control gates. Template creation time is particularly important for one-to-one comparisons since the time it takes to compare two templates is nearly instantaneous by comparison. The median time to create a template from a pair of iris samples (one of each eye) is 1007.8 milliseconds for the most accurate matcher. The shortest median creation time for any submission is 37.9 milliseconds. Most submissions have a median creation time under 200 milliseconds, but a few have times between 1 and 3 seconds. All times were computed using a single processing core. Dedicating more computational resources to template creation would reduce these times.
- **Demographic Effects:** Ideally, a biometric system will not perform substantially better or worse for members of any particular demographic group. Sex has a significant impact on accuracy for some matchers, but the effect is not consistent. Some matchers perform better on males while others perform better on females. For one participant's matchers, FNMR is double for females compared to males, but typically the magnitude of the difference is less. With respect to race, the matchers tend to perform best on Whites and poorest on Asians. This is not true in all cases and sometimes the differences are negligible. When a race effect is noticeable, comparisons between Whites tend to be less likely to false non-match but more likely to false match compared to Asians. Eye colour is discretized into the binary categories light (blue, green, and grey) and dark (brown and black). The most accurate matchers tend to perform slightly better on lighter eyes, but eye colour covaries with many other factors and demographic traits which could be responsible for the true effect.

  Because the test dataset consists of samples collected in various environments over a period of years, we cannot discount the possibility that any apparent demographic effects are due to confounding factors. Further investigation in necessary before drawing any solid conclusions.

Caution is advised when attempting to extrapolate numerical results from this evaluation to other scenarios. This evaluation assesses performance over a particular set of images collected under certain environmental conditions using specific hardware. It is difficult to predict how changing any of these parameters might affect performance.

Future Work

The next IREX IX report will document recognition of irides illuminated at different wavelengths. *ISO/IEC 19794-6: Iris Image Data* recommends illuminating the iris using near-infrared wavelengths "between approximately 700 and 900 nanometres (nm)". The next report will investigate how well iris recognition algorithms can segment and compare iris samples captured at illumination wavelengths ranging from the visible (405 nm - 700 nm) to the infrared.

Acknowledgements

The authors would like to thank the sponsor of this activity, the Federal Bureau of Investigation.

Disclaimer

Specific hardware and software products identified in this report were used in order to perform the evaluations described in this document. In no case does identification of any commercial product, trade name, or vendor, imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

Institutional Review Board

The National Institute of Standards and Technology (NIST) Human Subjects Protection Office (HSPO) reviewed the protocol for this project (ITL-16-0026) and determined it is not human subjects research as defined in Department of Commerce Regulations, 15 CFR 27, also known as the Common Rule for the Protection of Human Subjects (45 CFR 46, Subpart A).

Contents

List of Figures

List of Tables

1 Introduction

1.1. Purpose

The aim of this study is to evaluate the performance of iris recognition over operational test data. As a technology evaluation, it is very similar to *IREX IV Part 1: Evaluation of Iris Identification Algorithms* [1]. However, unlike IREX IV it assesses both verification (one-to-one) and identification (one-to-many) performance. Thirteen research institutions submitted recognition algorithms for evaluation, more than any other IREX evaluation.

The main goals of this evaluation are to:

- *Assess the current state of the art:* Biometric evaluations promote industrial competitiveness by providing a fair platform for comparison. This evaluation aims to impartially assess the current state of the art of automated iris recognition. Rather than concentrating on any specific application, performance is assessed for the common tasks of person identification and verification to ensure relevance to a wide range of applications.
- *Facilitate research and development:* The current evaluation seeks to identify areas for future research and development with an eye on the needs of our sponsors. IREX IX also offers algorithm developers, including participants from previous IREX evaluations, an opportunity to further improve and test their recognition algorithms.
- *Assess the impact of demographics:* IREX IX aims to identify possible disparities in performance for certain demographic groups. If comparison accuracy is markedly poorer for any particular group, it can disproportionately impact members of that group. Three demographic factors are considered: sex, race, and eye colour.

As a technology evaluation IREX IX focuses predominantly on algorithm performance rather than other factors relevant to the operation of a biometric system. It does not address the costs associated with operating a biometric system, or the system's usability, or possible security issues such as algorithm vulnerabilities. As an off-line evaluation, it does not include a live image acquisition component or any interaction with real users.

1.2. The IREX Program

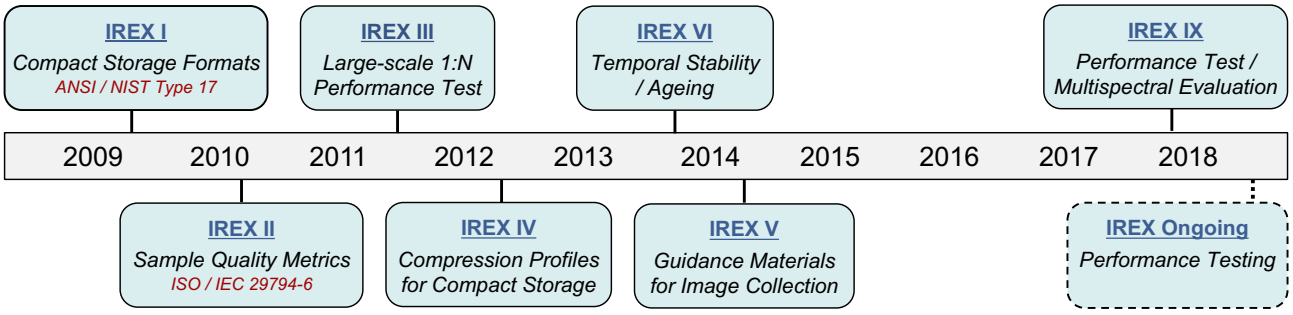

Figure 1.1: Timeline of the IREX program, including a planned future installment.

The IREX Program was initiated by National Institute of Standards and Technology (NIST) to support an expanded marketplace of iris-based applications. IREX provides quantitative support for iris recognition standardization, development, and deployment. To date, 6 activities have been completed and two more are tentatively planned. Each is summarized below.

- **IREX I** [3] was a large-scale, independently administered, evaluation of one-to-many iris recognition. It was conducted in cooperation with the iris recognition industry to develop and test standard formats for storing iris images. Standard formats are important for maintaining interoperability and preventing vendor lock-in. The evaluation was conducted in support of the ISO/IEC 19794-6 and ANSI/NIST-ITL 1-2011 standards.
- **IREX II** [4] supported industry by establishing a standard set of quality metrics for iris samples. Although iris recognition has the potential to be extremely accurate, it is highly dependent on the quality of the samples. The evaluation tested the efficacy of 14 automated quality assessment algorithms in support of the ISO/IEC 29794-6 standard [5].
- **IREX III** was a performance test of the latest iris recognition algorithms over operational data. Despite growing interest in iris-based technology, at the time there was a paucity of experimental data to support published theoretical

considerations and accuracy claims. IREX III constituted the first public presentation of large-scale performance results using operational data.

- **IREX IV** built upon IREX III as a performance test of one-to-many iris recognition. In addition to providing participants from previous evaluations an opportunity to further develop and test their recognition algorithms, this evaluation explores the potential for using a cost equation model for optimizing algorithms for specific applications.
- **IREX V** is an ongoing effort that provides best practice recommendations and guidelines for the proper collection and handling of iris images.
- **IREX VI** explored a possible aging effect for iris recognition. The intrinsic features of the iris may naturally change over time in a way that affects recognition accuracy. Factors such as subject habituation and aging of the camera may also introduce a time dependency.
- **IREX VII** is planned to develop a framework for communication and interaction between components in an iris recognition system. By introducing layers of abstraction that isolate underlying vendor-specific implementation details, a system can become more flexible, extensible, and modifiable. That framework is currently in use internally at NIST.
- **IREX VIII** was never conducted.
- **IREX IX** is the topic of the current report - a performance test of iris recognition over operational test data. The second report will be a multispectral evaluation of iris recognition.
- **IREX Ongoing** is tentatively planned a successor to IREX IX. It will be an ongoing, largely automated, evaluation similar to *Ongoing MINEX* and *FRVT Ongoing*.

The latest information on the IREX Program can be found on the IREX website [6].

1.3. Industry Growth

Iris recognition has experienced rapid growth in the last 20 years. Government-sponsored evaluations such as the IREX program have facilitated this growth through 1) the development of standards and 2) by affirming the potential for iris recognition to meet the demands of large-scale deployments. IREX IV found that some matching algorithms are capable of searching a single iris image against an enrolled population of millions in under a second (using just one processing core). The evaluation also found that for the most accurate matchers, identification failures were almost always the result of poor sample quality, where the eye is closed, off-axis, highly rotated, etc. Many of these errors can be avoided through the use of more advanced cameras or improved image collection and data handling practices.

In recent years, several government agencies have deployed (or are in the process of deploying) iris recognition systems that operate on a national scale. The largest is India's Unique Identity Authority of India (UIDAI) program [7] which contains iris images of hundreds of millions of Indian residents. The program was initiated to better manage the allocation of government resources and to provide improved services to citizens. The United Arab Emirates (UAE) also employs iris recognition as part of its border-crossing control system [8, 9]. At ports of entry, visitors are searched against a watch list of several hundred thousand people previously expelled from the country for various violations. The Federal Bureau of Investigation (FBI) includes iris recognition technology on its technical roadmap. Since 2006, the Department of Defense (DOD) has been using handheld devices to collect iris images of people in various theaters of operation. The images are consolidated into a central repository known as the Automated Biometric Identification System (ABIS) to support a variety of missions from tactical operations to detention management. More locally, iris recognition is being used at correctional facilities for employee access authentication [10].

2 Methodology

The current technology evaluation focuses on matcher performance as opposed to other factors that might be relevant to the deployment and operation of a biometric system (e.g. societal, economic, legal factors). Performance metrics are selected to objectively compare different matcher implementations, primarily in terms of accuracy and speed. While recognition accuracy is always an important performance factor, the importance of speed depends much more on the application. For example, biometric systems that perform identification in real-time require rapid response times. Off-line tasks (e.g. database de-duplication), on the other hand, tend to have more relaxed time constraints.

2.1. Test Environment

The evaluation was conducted offline at a NIST facility. Offline evaluations are attractive because they allow uniform, fair, repeatable, and large-scale statistically robust testing. However, they do not capture all aspects of an operational system. While this evaluation is designed to mimic operational reality as much as possible, it does not include a live image acquisition component or any interaction with real users.

Testing was performed on high-end PC-class blades running the Linux operating system (CentOS 7.2), which is typical of central server applications. Most of the blades had Dual Intel Xeon E5-2695 v3 3.3 GHz CPUs (56 total cores) with 192 GB of main memory. The test harness used concurrent processing to distribute workload across multiple blades.

2.2. Matching Algorithms

Thirteen commercial organizations and academic institutions submitted 46 iris recognition software libraries for evaluation. The participation window opened on October 7th, 2016 and closed on September 7th, 2017. Participation was open world-wide to anyone with the ability to implement a large-scale one-to-many iris identification algorithm. There was no charge to participate.

Participants provided their submissions to NIST as static or dynamic libraries compiled on a recent Linux kernel. The libraries were then linked against NIST's test driver code to produce executables. A further validation step was performed to ensure that the algorithms produce identical output on both the participants' and NIST's test machines. The full process is described in the IREX IX Application Programming Interface (API) and Concept of Operations (CONOPS) document [11].

Participants submitted their implementations in three rounds referred to as "phases". After the first two phases, participants were provided with rudimentary feedback on the performance of their submissions in the hope that it would assist with algorithm development for the next phase. Although only two phases were planned, a third phase was introduced and the first was designated a test phase (the results of which will not be made public). Table 2.1 lists the IREX IX participants along with the phases in which they participated. The deadline to submit to the second phase was January 21st, 2017 and the deadline for the third phase was September 1st, 2017. Each participant was required to submit at least one one-to-one implementation and one one-to-many implementation for each phase, although participants were allowed to submit up to two of each per phase. Some of the participants are new to the IREX program and some (Iris ID, Neurotechnology, Delta ID, NEC, FotoNation) have participated in previous IREX evaluations.

Participant	Phase 2	Phase 3
Aware Inc.	✓	✓
Decatur	✓	✓
DeltaID	✓	✓
Dermalog	✓	✓
FotoNation	✓	✓
IrisID	✓	✓
NEC	✓	✓
NeuroTechnology	✓	✓
Qualcomm	✓	✓
SOAR Advanced Technologies		✓
Tafirt	✓	✓
Tiger IT	✓	✓
Unique Biometrics	✓	

Table 2.1: Participants of IREX IX along with the submission phases in which they participated.

2.3. Image Dataset

All testing was performed over a single dataset of 673 662 iris samples from 260 809 subjects (both left and right eyes), henceforth referred to as the OPS-III dataset. The samples were field collected from various locations between September 2007 and October 2015. A visual inspection of the samples seems to indicate that they are better quality, at least on average, than the OPS-II samples used for performance testing in previous IREX evaluations. That said, field collected samples tend to suffer more from quality related problems (e.g. motion and focus blur) than samples collected in more controlled laboratory settings. In comparison to OPS-II the samples also appear to have been collected by more contemporary iris cameras.

Considerable effort was taken to filter out errors in the ground truth. Ground truth errors are cases where iris samples are assigned incorrect person identifiers. Failing to remove these mistakes can inflate estimates of error rates. The process of identifying and excluding ground truth errors is detailed in Appendix C.

The dataset is sequestered (i.e. not publicly available). The participants were not allowed to view any of the iris samples and were not provided with a representative set of iris samples, although their basic characteristics were described in the IREX IX CONOPS document [11].

2.4. Performance Metrics

Performance is evaluated for both *one-to-one* and *one-to-many* comparison modes (sometimes referred to, respectively, as *verification* and *identification* modes). In one-to-one mode, a specific claim to identity is made and two biometric templates are compared to determine whether the claim is true. In one-to-many mode, an authentication template is searched against a database of enrolled templates for a match. Although no specific identity claim is made, an implicit claim of enrollment (or lack of enrollment) is made. For example, anyone presenting their biometric features to an access control system is implicitly claiming they are in the enrolled population.

The following sections provide a high-level description of the performance metrics used in this report. Full mathematical definitions of all error metrics are presented in Appendix A.

2.4.1 One-to-one Matching

The degree of dissimilarity between two biometric templates is quantified by a dissimilarity score. In the case of John Daugman's IrisCode algorithm [12], the dissimilarity score is also known as a Hamming Distance. A dissimilarity score is referred to as *mated* if it is the result of comparing two templates representing the same iris (in the case of single-eye comparisons) or pair of irides (in the case of two-eye comparisons). It is known as a *nonmated* score if it is the result of comparing templates representing different irides. An identity claim is accepted if the dissimilarity score is below (or equal to) a preset decision threshold. Otherwise, the identity claim is rejected. As with any binary classification problem, two types of decision errors are possible. The first occurs when a nonmated comparison is misclassified as mated. This is known as a *false match*. The second type of decision error occurs when a mated comparison is misclassified as nonmated. This is known as a *false nonmatch*.

Adjusting the decision threshold reduces the rate of one type of error but at the expense of the other. This relationship is characterized by a DET curve [13], which plots the tradeoff between the two error rates. DET curves have become a standard in biometric testing, superseding the analogous ROC curve. Compared to ROC curves, the logarithmic axes of DET curves provide a superior view of the differences between matchers in the critical high performance region.

Timing statistics are presented as the actual physical time that elapsed for the operations of template creation and template comparison. Timing statistics are collected for single-threaded operations on otherwise unloaded machines. For ease of testing and fair comparison, submissions were required to operate in single-threaded mode. Operationally, software can be designed to exploit multiple cores when available to expedite template creation and comparison.

2.4.2 One-to-many Matching

Open-set biometric systems are tasked with searching a biometric sample against an enrollment database and returning zero or more candidates. A candidate is returned if the matcher determines that its dissimilarity to the searched image is at or below a preset decision threshold. A false positive occurs when a search returns a candidate for an individual that *is not* enrolled in the database. A false negative occurs when a search *does not* return the correct candidate for an individual that *is* enrolled in the database. Brief definitions of the two opposing error rates are provided in Table 2.2. Raising the decision threshold increases the false negative identification rate (FNIR) but decreases the false positive identification rate (FPIR). Although the metrics do not strictly represent error rates in a binary classification system, core accuracy is still presented in the form of Detection Error Tradeoff (DET) plots, this time showing the tradeoff between FPIR and FNIR.

Metric	Definition
False Negative Identification Rate (FNIR)	Fraction of mated searches that do not return the correct mate.
False Positive Identification Rate (FPIR)	Fraction of nonmated searches that return at least one (incorrect) candidate.

Table 2.2: *Informal definitions of the error metrics for a one-to-many iris recognition system.*

For the purposes of testing, the IREX IX API required submissions to return a fixed number of candidates for each search, but only candidates with dissimilarity scores at or below threshold are considered. Candidates with corresponding dissimilarity scores above threshold are effectively discarded.

False positives are computed exclusively from non-mated searches (i.e. searches for which the searched individual is not enrolled in the database). This is more reflective of operation than if false positives had been computed from mated searches with the correct candidates removed from the list. Similarly, false negatives are computed exclusively from mated searches.

Timing statistics are presented as the physical time that elapsed for the operations of template creation and searches. Search time is expected to be proportional to the size of the enrolled population.

2.4.3 Feature Extraction Failures

Participants were instructed to provide submissions that always create comparable templates, even when no useful feature information could be extracted. These "blank templates" are expected to produce high measures of dissimilarity (effectively infinity) when compared. This was done for ease of testing but does not reflect operational reality since, for example, a blank template would never be saved onto a smartcard and used for access control. If the template is being acquired in real-time from a cooperative user, the user could be prompted to provide a new sample or different accommodations could be made (e.g. using fingerprints instead). This inability to handle template creation errors in real time highlights a weakness of off-line testing.

2.4.4 Confidence Intervals

OPS-III was sampled from a larger dataset of field-collected iris samples used by a government agency. We refer to this parent dataset as the *population*. The confidence intervals presented in this report show how well the accuracy statistics calculated over our test data estimate the true population values. All of our confidence intervals are computed at the 90% confidence level. This does not mean that there is a 90% probability that the true population value falls within the interval. Rather, it means that if the population is repeatedly sampled and an interval estimate is computed on each occasion, the interval estimates would contain the true population value 90% of the time.

The iris images in OPS-III are paired in various ways to form comparison sets. These pairings introduce a correlation structure. For example, samples of a person's left and right eye captured during the same session are expected to be highly correlated in terms of sample quality. Wayman [14] found that failing to account for these dependencies can lead to overly optimistic estimates of confidence intervals. Thus, we took steps to factor the correlation structure into our estimates of uncertainty. The full procedure is detailed in Appendix B.

Unfortunately, we do not know how OPS-III was sampled from the larger parent population. In particular, we do not know if there was any sampling bias that might introduce systematic over- or under-estimation of the accuracy metrics as well as their confidence interval estimates. In the absence of any information on how OPS-III was sampled, we are forced to assume simple random sampling even though this is far from ideal.

3 Results

3.1. Accuracy

Accuracy is presented for two-eye comparisons since most iris scanning devices acquire images of both eyes, whether they're captured concurrently with a two-eye camera or successively with a single-eye camera. The costs associated with decision errors are highly dependent on the application and often difficult to quantify. A false positive could result in free access to a theme park [15] or unauthorized access to classified information. Hence the reason accuracy is presented in the form of DET curves, which show classification accuracy over a range of operating thresholds without making assumptions about the costs of errors.

3.1.1 One to One Matching

Although iris recognition is being increasingly deployed for large-scale one-to-many applications, many systems still operate in a verification mode. The Pentagon Force Protection Agency uses iris and fingerprint scanners to control access to the Pentagon Building and Mark Center [16, 17]. Several state and local law enforcement agencies have also expressed interest in using iris recognition for identity management at their prison and jail facilities [18].

Figure 3.1 shows two-eye DET accuracy. For clarity of presentation only the most accurate matcher from each participant is shown (specifically, the submission that yields the lowest FNMR at FMR $= 10^{-5}$). Figure 3.2 shows 90% confidence intervals for all matchers at FMR $= 10^{-5}$. Comprehensive DET plots and tables for both single-eye and two-eye comparisons can be found in Appendix D.

Notable Observations

- **Accuracy:** The most accurate one-to-one matcher (NEC 5) yields an FNMR of 0.0057 at FMR = 10^{-5}. Four submissions follow (NeuroTechnology 5, DeltaID 6, Tiger IT 5, and Decatur 6) with FNMRs between 0.0066 and 0.0070 (the differences are unlikely to be statistically significant). Thirty of the 46 submissions yield an FNMR less than 0.02 at FMR = 10^{-5}. In general, if two confidence intervals do not overlap, then the difference is statistically significant. However, the opposite is not necessarily true: if two confidence intervals overlap, the difference may or may not be statistically significant.
- **Flatness:** The DET curves have lightly sloping DET curves such that FNMR increases only slightly as FMR decreases. For NEC 5, FNMR increases from 0.0043 to 0.0067, an increase of 55% as FMR decreases from 10^{-1} to 10^{-7}. Some matchers have steeper slopes than others. Tafirt 4 and Dermalog 6 and FotoNation 5 perform comparatively better at higher FPIRs.
- **Improvements with Later Submissions:** Nearly all of the most accurate matchers were submitted during the final submission phase of IREX IX. Curves translated downward in relation to previous submissions may indicate improvements in the feature extraction process. Changes in slope or shape may indicate alterations to the comparison strategy.

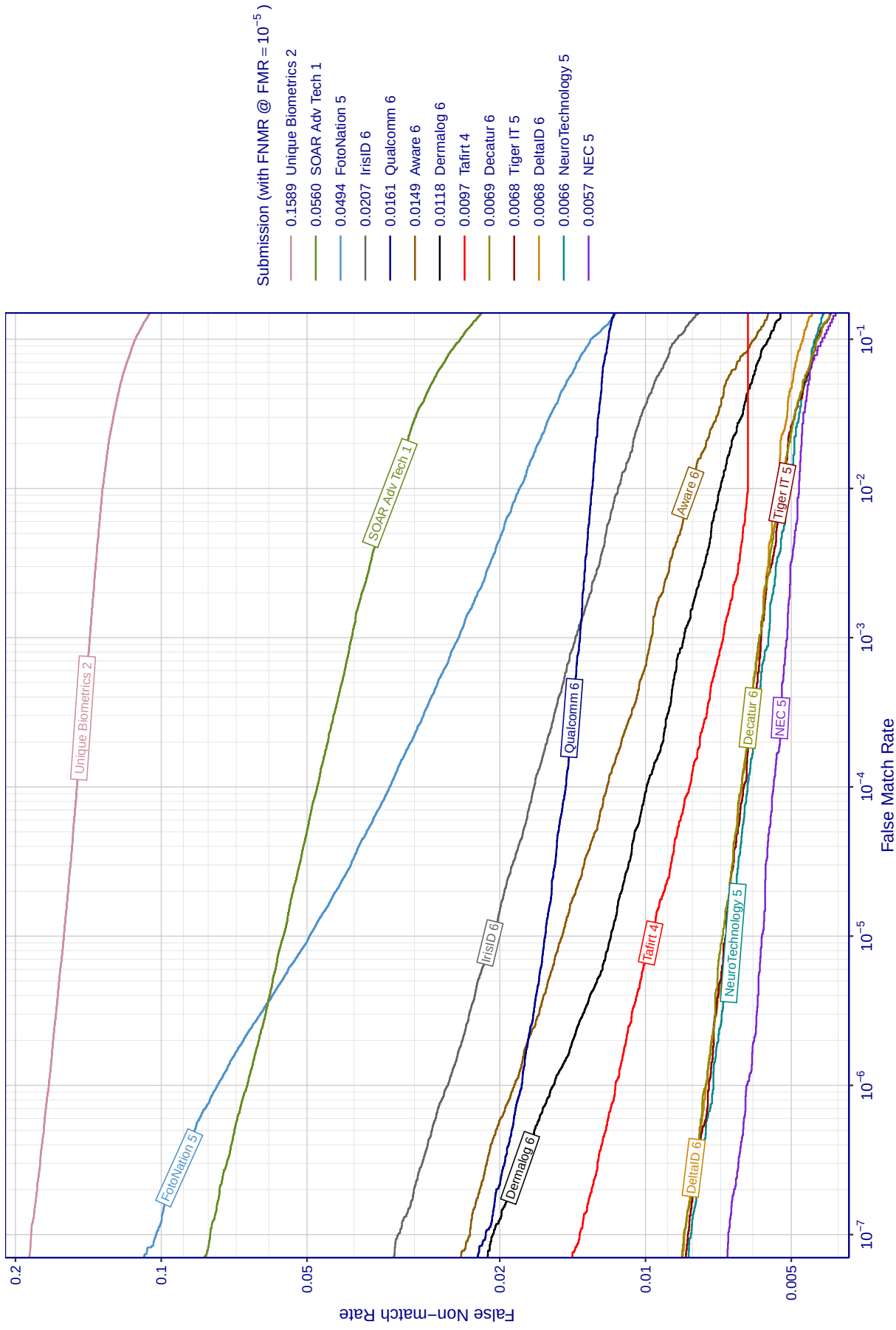

Figure 3.1: DET curves for two-eye comparisons. Only the most accurate matcher (at FMR=10^{-5}) is shown for each participant. Plots were generated from ≈ 83K mated and ≈ 500 million nonmated comparisons.

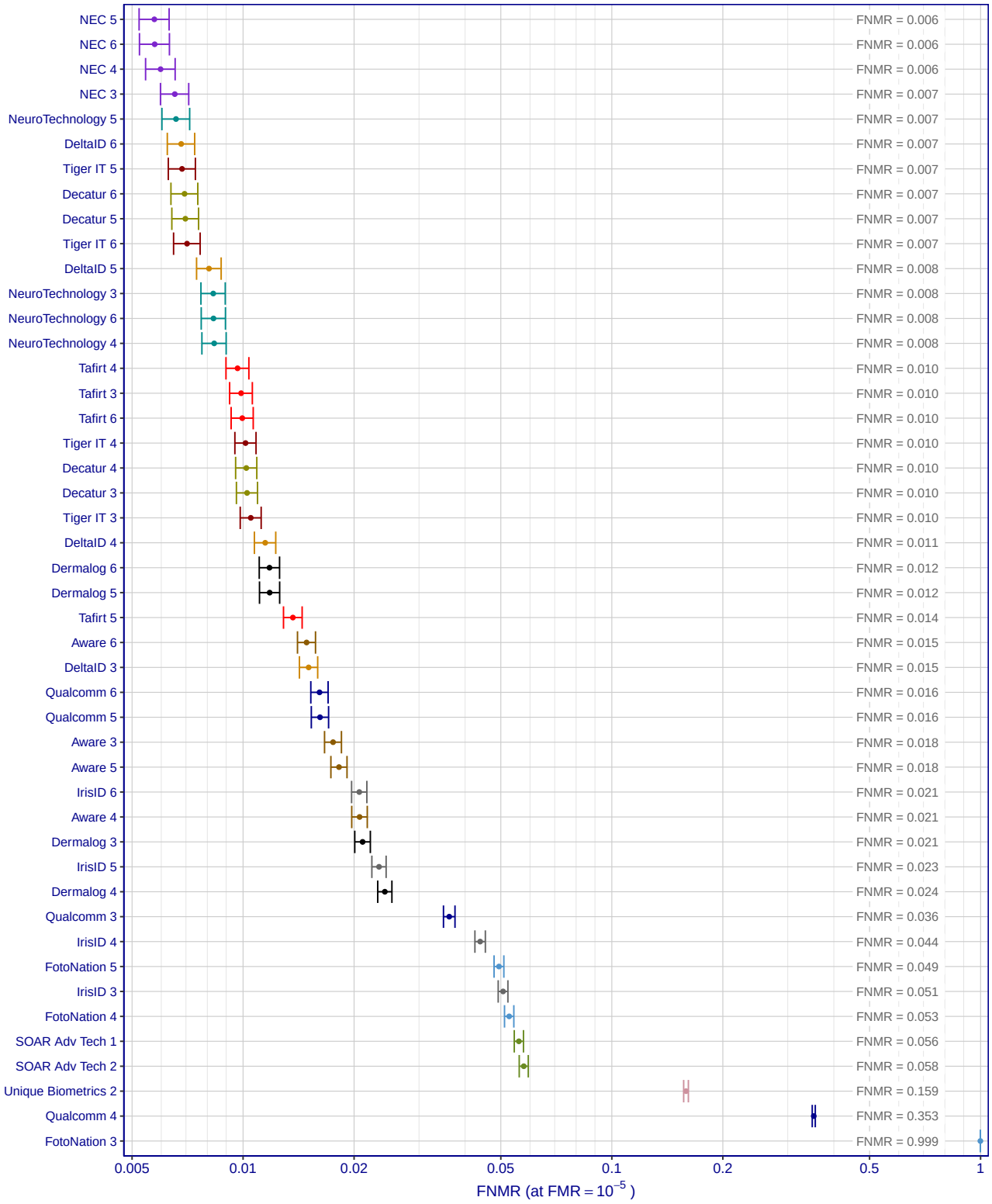

Figure 3.2: Ninety percent confidence intervals for FNMR (at FMR=10^{-5}) for two-eye coparisons. Plots were generated from $\approx$ 83K mated and $\approx$ 500 million nonmated comparisons.

3.1.2 One to Many Matching

The identification task differs from verification in that it does not require the user to provide a claim of identity. Thus, the user is not required to enter a user-specific pin number or present a smart card to use the system. The self-service iris kiosks used by UK IRIS [19] operated in this way, as does the current NEXUS program which offers expedited processing for travelers between the United States and Canada. These are both access control systems, a specific type of *positive identification system* that grants special privileges to enrolled users. Positive identification systems verify the implicit claim that the user is enrolled in the system. They contrast with *negative identification systems* which verify the implicit claim that the user *is not* enrolled in the system. The most common example of a negative identification system is a watchlist system, which typically denies special privileges to enrolled users. For example, the United Arab Emirates (UAE) maintains a border-crossing system that prevents those previously expelled from the country from reentering.

Iris identification is particularly robust to increases in the enrolled population size because of 1) its ability to rapidly perform searches against the entire enrolled population, and 2) the fact that false negatives are often the result of poor quality captures that would occur at any enrolled population size. This is evidenced by the nearly-flat (i.e. low-slope) appearance of iris DET curves that has been noted in previous reports and evaluations [20, 21]. IREX IV [1] also showed that large increases in the enrolled population size translate into only minor decreases in accuracy.

This report presents two-eye accuracy for enrollment populations ranging from 10 to 160 thousand people. Figure 3.3 shows DET accuracy when the enrolled population is 10 thousand. For clarity of presentation only the most accurate submission from each participant is shown (specifically, the submission that yields the lowest FNIR at FPIR $= 10^{-3}$). Figure 3.4 shows 90% confidence intervals for all submissions at FPIR $= 10^{-3}$. Figures 3.5 and 3.6 present the same information when the enrolled population is 160 thousand. Comprehensive DET plots for both single-eye and two-eye comparisons can be found in Appendix D.

Threshold calibration tends to be easier for iris recognition compared to other biometric modalities due to the relative stability and predictability of the nonmated distribution. Daugman [22] asserts that when the comparison scores are Hamming Distances, the nonmated distribution can be derived by applying extreme value theory to the binomial distribution. Figures 3.8 and 3.7 plot FPIR and FNIR as a function of the enrolled population size when the decision threshold is fixed.

Notable Observations

- **Accuracy:** The most accurate one-to-many matcher (NEC 6) yields an FNIR of 0.0067 at FPIR = 10^{-3} with an enrolled population of 160 thousand. NeuroTechnology 5 and DeltaID 5 follow with FNIRs of 0.0081 and 0.0083 respectively. Twenty four of the 46 submissions yield an FNIR less than 0.02 at FPIR = 10^{-3} with the 160 thousand enrolled population size.
- **Flatness:** The DET curves have lightly sloping DET curves such that FNIR increases only slightly as FPIR decreases. For NEC 6, FNIR increases from 0.00572 to 0.0074, an increase of 29% as FPIR decreases from 10^{-1} to 10^{-4}. Some matchers have steeper slopes than others. Tafirt 4 and Dermalog 6 and FotoNation 5 perform comparatively better at higher FPIRs.
- **Improvements with Later Submissions:** Nearly all of the most accuracy matchers were submitted during the final submission phase of IREX IX. Curves translated downward in relation to previous submissions may indicate improvements in the feature extraction process. Changes in slope or shape may indicate alterations to the comparison strategy.
- **Threshold Calibration** Figure 3.7 reveals that at a fixed operating threshold, FNIR remains relatively fixed despite large variations in the enrolled population size. The only exceptions are Unique Biometrics 2 and the submissions from Aware, where FNIR increases with the enrolled population size. Figure 3.8 shows that at a fixed operating threshold, FPIR increases as the size of the enrolled population grows for nearly every matcher. Most matchers experience a 5 to 20 fold increase in FPIR as the enrollment population size goes from 10 thousand to 160 thousand (a factor of 16 increase in population size). The exceptions are the matchers submitted by Aware, where FPIR actually trends downward. It is possible that Aware is adjusting its comparison scores to accommodate changes in the size of the enrolled population.

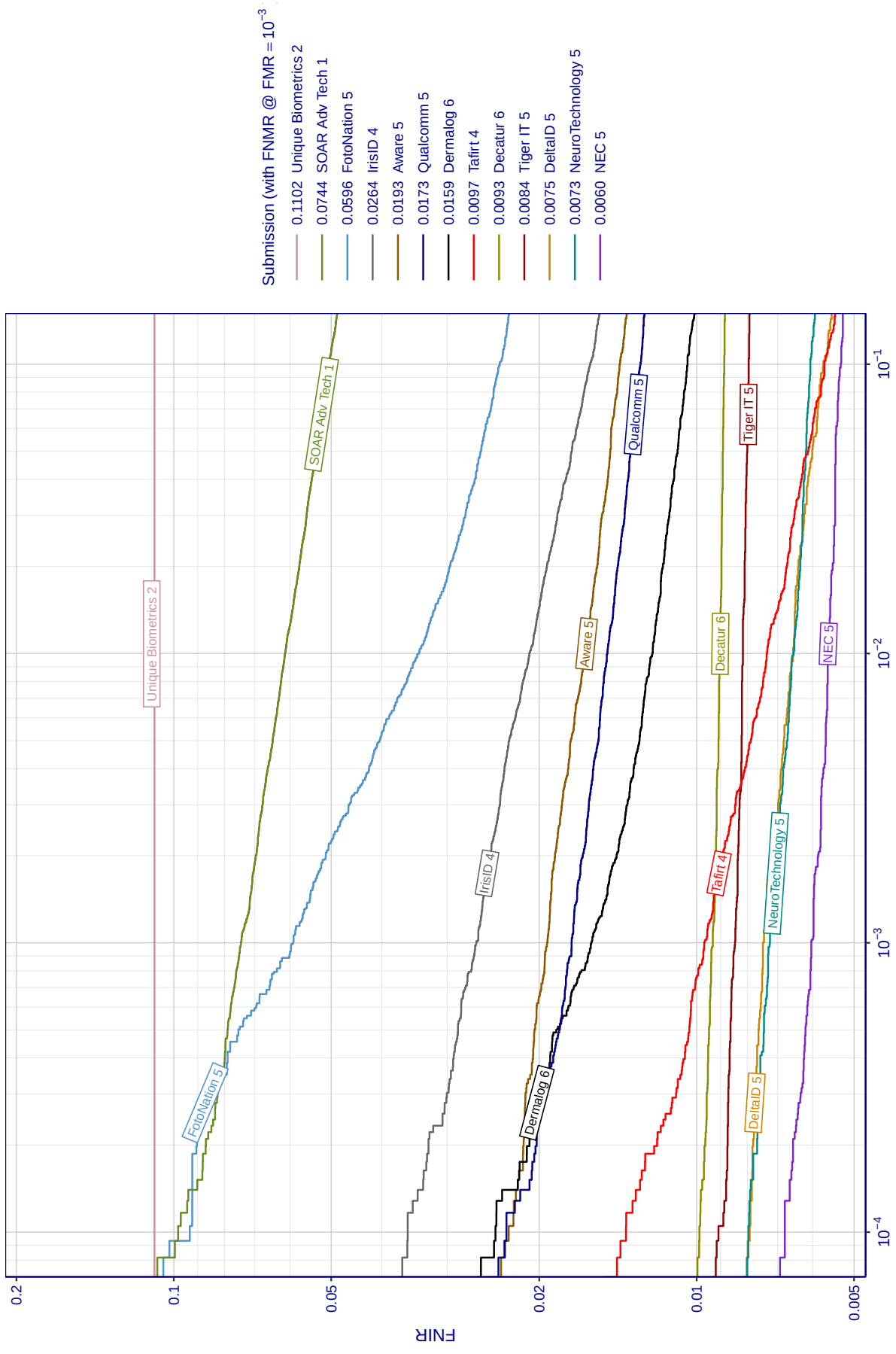

Figure 3.3: DET curves for two-eye comparisons against an enrollment database of 10 thousand. Only the most accurate matcher (at FPIR=10^{-3}) is shown for each participant. Plots were generated from ≈ 83K mated and ≈ 86K nonmated searches.

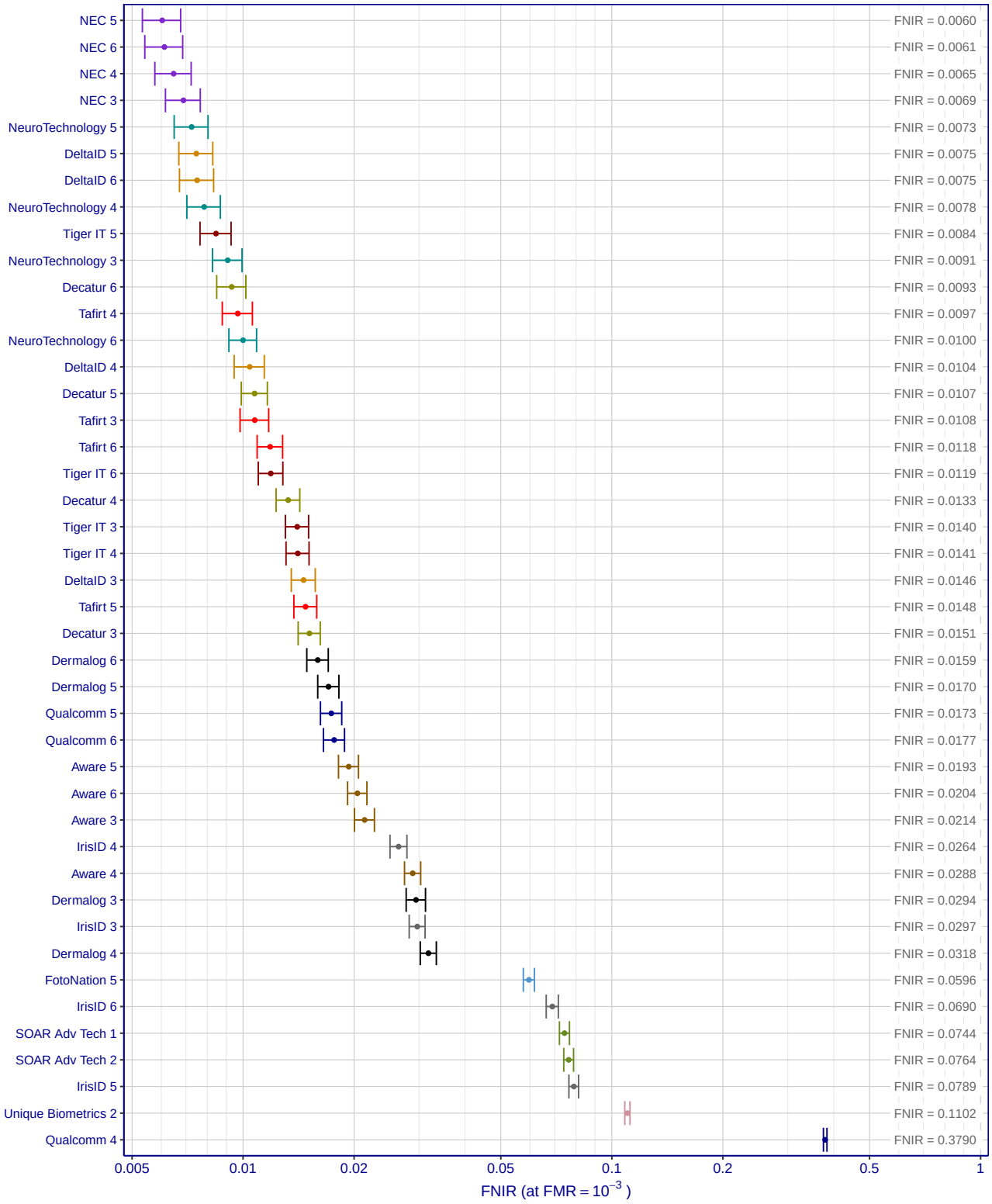

Figure 3.4: Ninety percent confidence intervals for FNIR (at FPIR = 10^{-3}) for two-eye comparisons against an enrolled population of 10 thousand. Plots were generated from ≈ 83K mated and ≈ 86K nonmated searches.

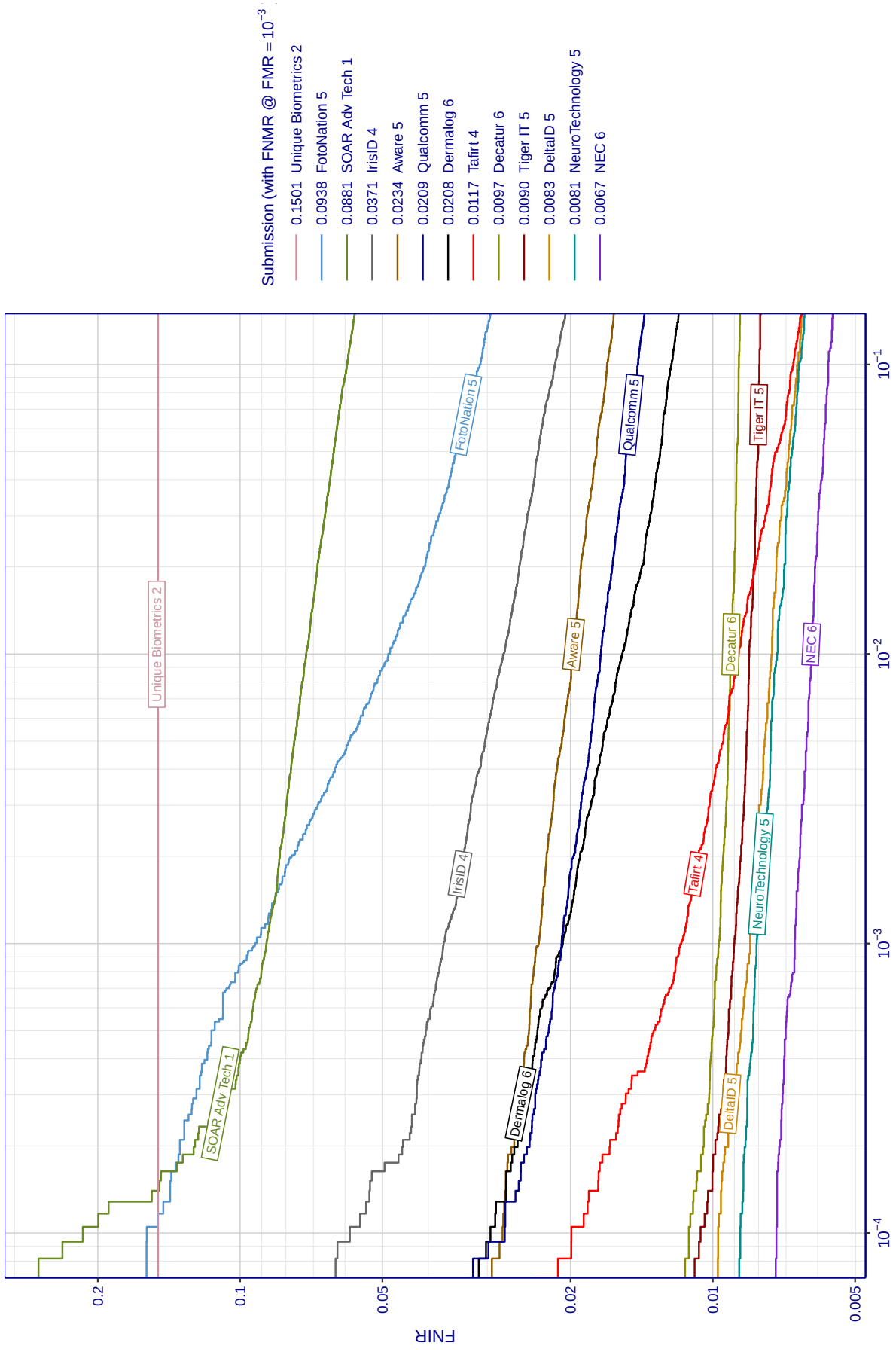

Figure 3.5: DET curves for two-eye comparisons against an enrollment database of 160 thousand. Only the most accurate matcher (at FPIR=10^{-3}) is shown for each participant. Plots were generated from ≈ 83K mated and ≈ 86K nonmated searches.

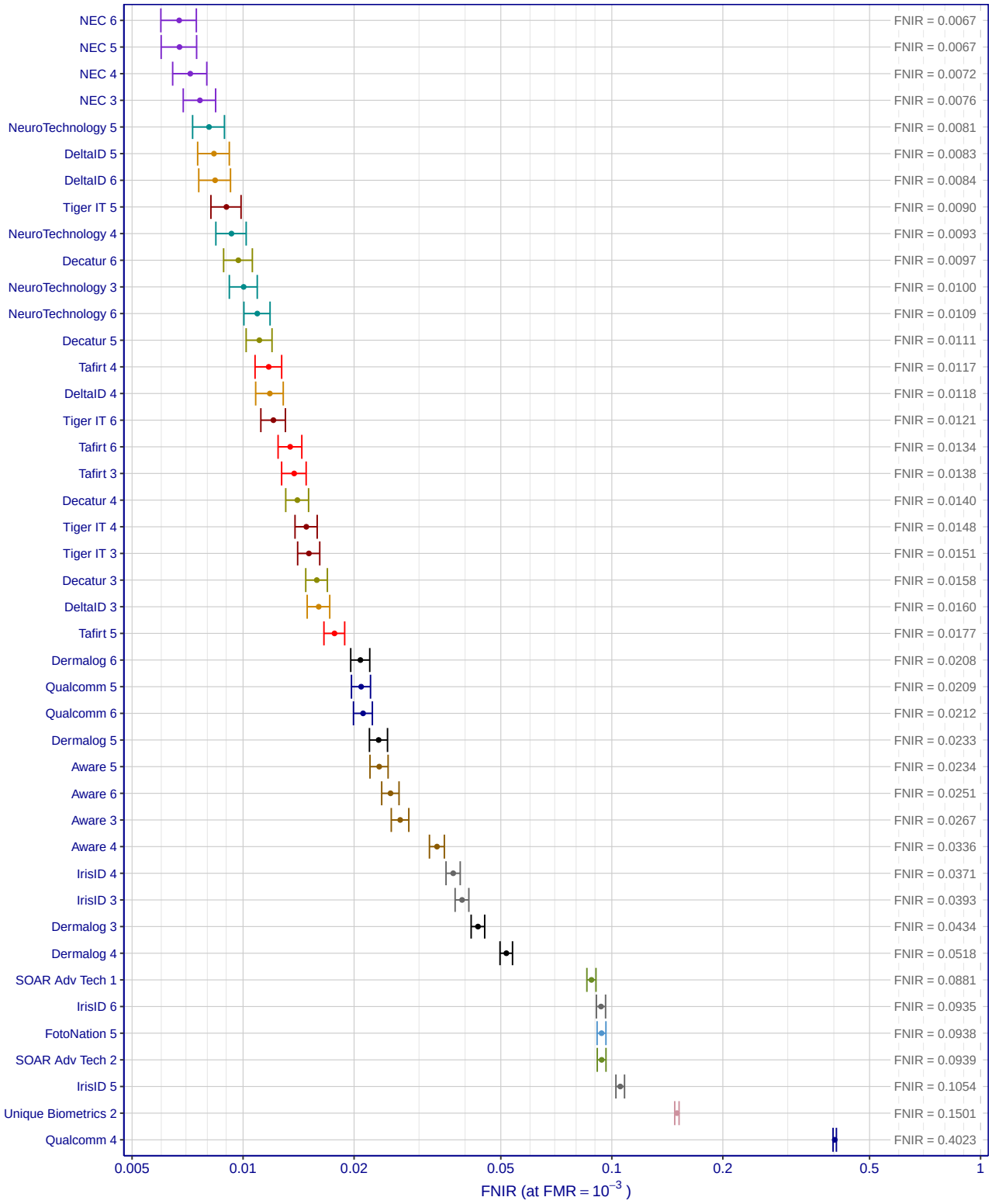

Figure 3.6: Ninety percent confidence intervals for FNIR (at FPIR = 10^{-3}) for two-eye comparisons against an enrolled population of 160 thousand. Plots were generated from ≈ 83K mated and ≈ 86K nonmated searches.

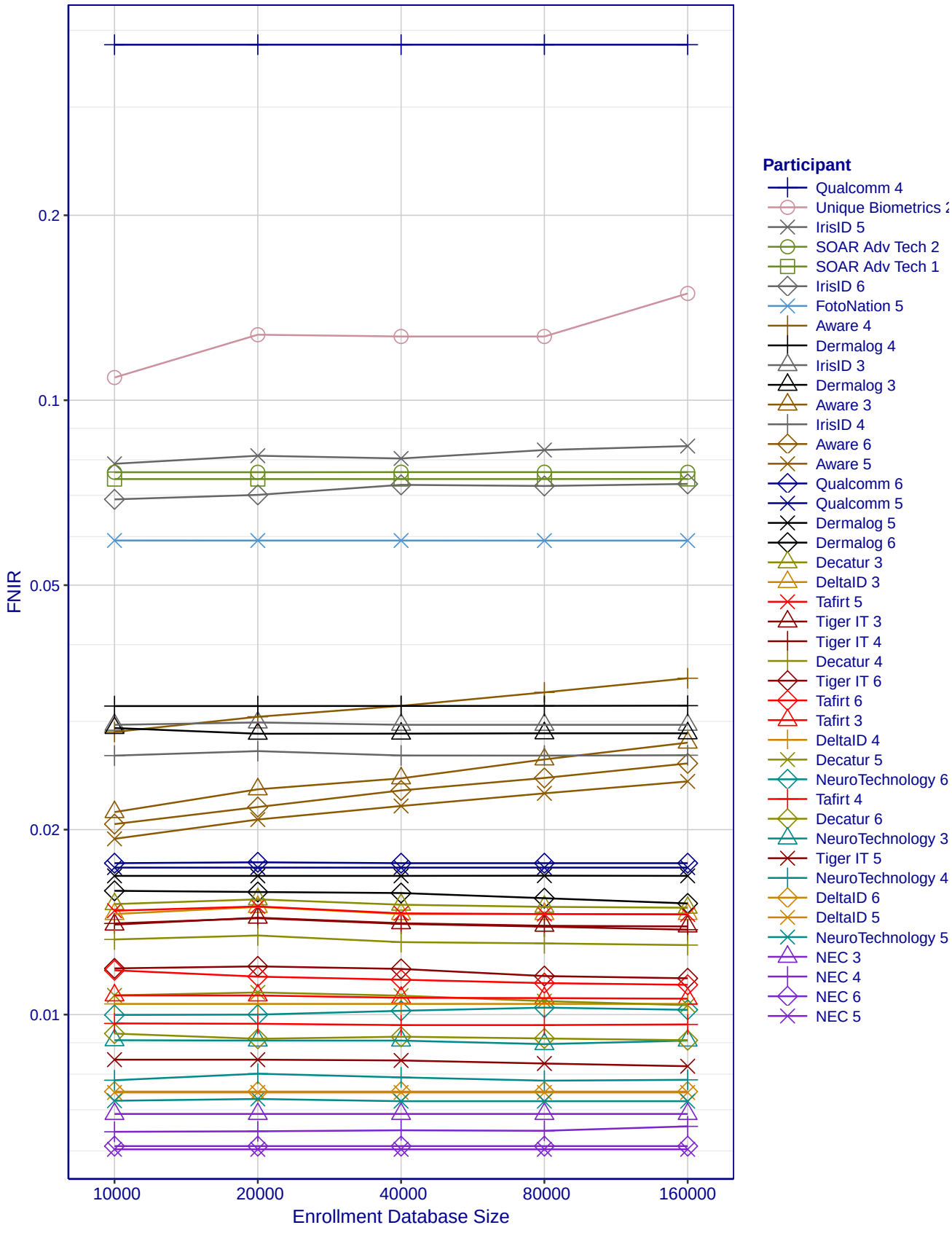

Figure 3.7: FNIR as a function of enrollment database size when the decision threshold is fixed. The threshold is fixed to elicit an FPIR of 10^{-3} at an enrollment database size of 10k.

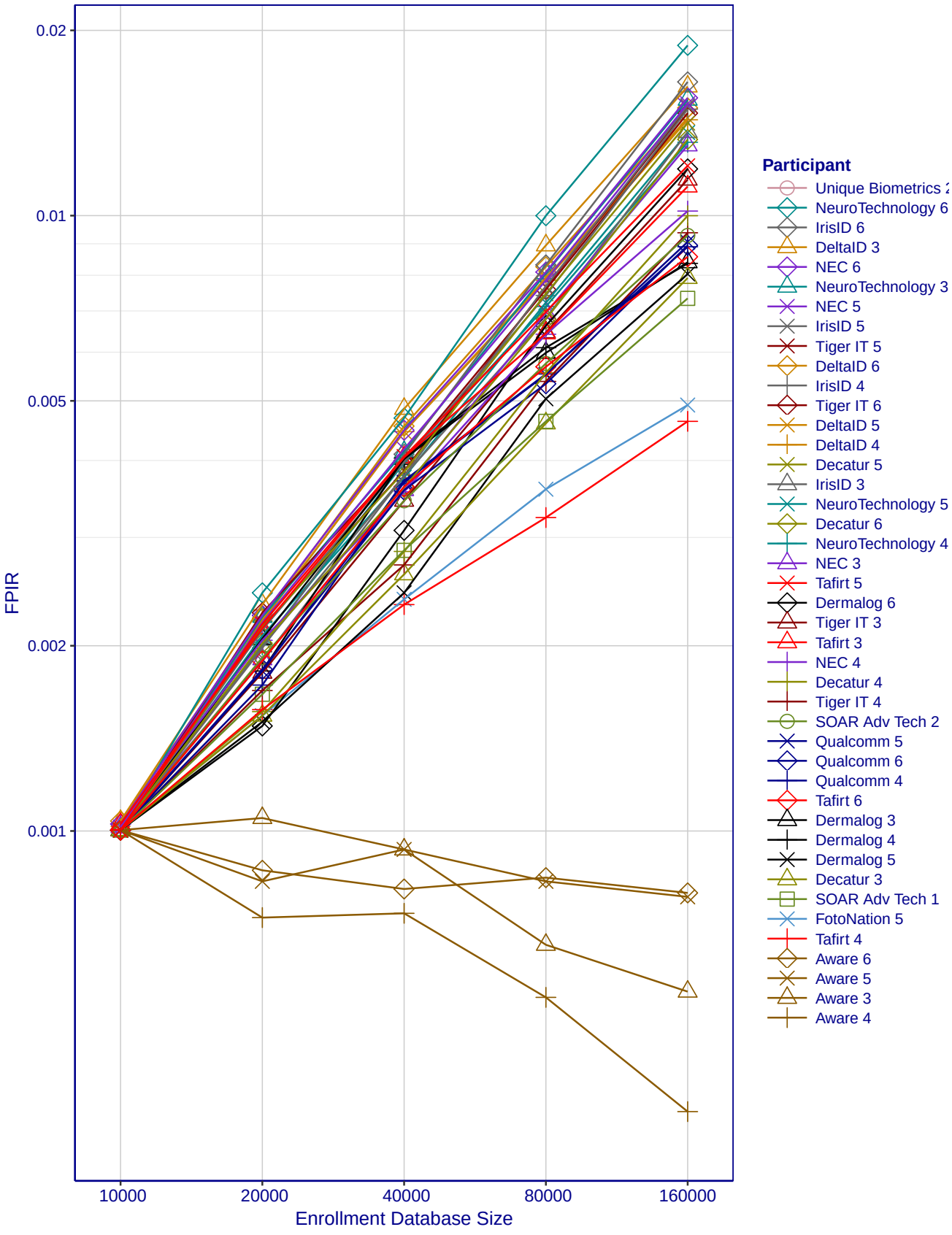

Figure 3.8: FPIR as a function of enrollment database size when the decision threshold is fixed. The threshold is fixed to elicit an FPIR of 10^{-3} at an enrollment database size of 10k.

3.2. Speed

3.2.1 One to One Matching

3.2.1.1 Template Creation Time

Template creation time refers to the amount of time that elapses while a comparable template is created from a raw iris image (or images). The relevant factor is turnaround time (i.e. the speed at which a response can be returned after an iris sample is acquired). Short turnaround times are critical for maintaining high throughput at, for example, access control gates. Longer template creation times mean more time waiting to verify an identity claim. Since comparisons are relatively fast, template creation time is the more important speed metric for most verification systems. A transaction in a centralized system involves several steps (presentation of the iris, image acquisition, network transfer, etc.) which together are likely to take longer than the comparison step alone.

Figure 3.9 shows the distribution of template creation times for each submission when provided with both left and right iris samples (i.e. two iris samples per template). The times do not include any pre-processing steps performed by the testing harness such as loading the iris samples from disk. The timing machine was a Dual Intel Xeon E5-2695 running at 3.3 GHz. Further details on the testing environment can be found in Section 2.1.

Notable Observations

Unique Biometrics 2 creates their templates in the least amount of time with a median creation time of 37.9 milliseconds. Thirteen submissions have median creation times under 100 milliseconds and 23 have median creation times under 200 milliseconds. Eight submissions have median creation times over one second (i.e. 1000 milliseconds). The fastest submission (Unique Biometrics 2) creates templates about 33 times faster than the slowest submission (Tiger IT 5). Multiple cores could be used to reduce these times.

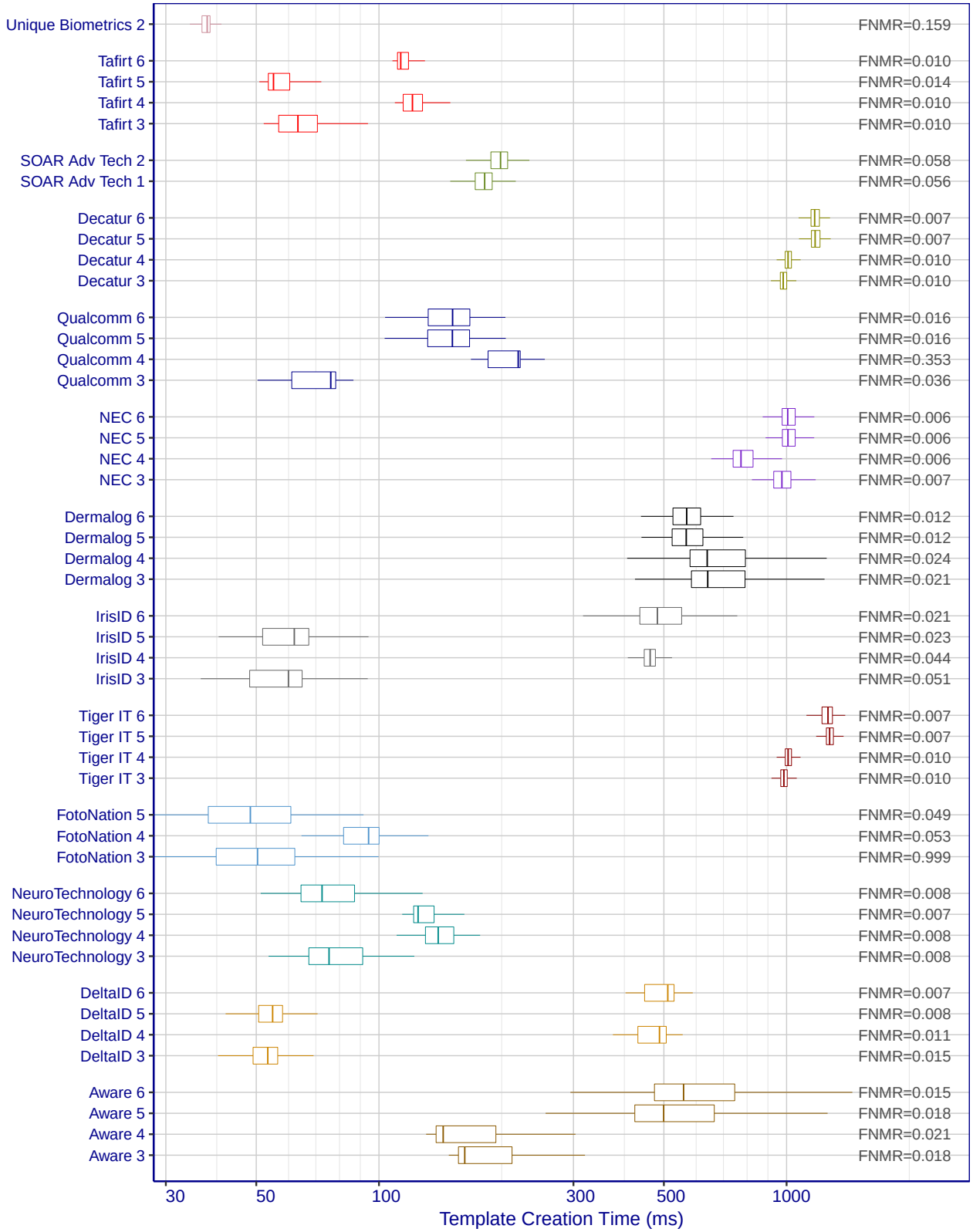

Figure 3.9: Boxplots summarizing the distribution of template creation times for each one-to-one submission. Each template is created from a pair of iris samples (one of the left eye and one of the right). For reference, FNMR at an FMR of 10^{-5} is reported on the right. Boxplots were generated from 1000 created templates.

3.2.1.2 Comparison Time

Comparison time refers to the amount of real-world time it takes to compare two templates and return a dissimilarity score. Timing statistics were collected for $10\,000$ mated comparisons using a single processing core. Figure 3.10 shows the distribution of comparison times for each matcher. For reference, FNMR at an FMR of 10^{-5} for each matcher is reported along the right-hand side of the figure.

Notable Observations

The fastest functioning matcher compares two-eye templates with a median time of 0.06 milliseconds. The fastest functioning matcher is almost $10\,000$ times faster than the slowest matcher. Thirty-six of the 46 matchers compare templates with a median time of under a millisecond. Comparison times tend to vary noticeably even for different matchers from the same participant, suggesting a variety of comparison strategies are being attempted.

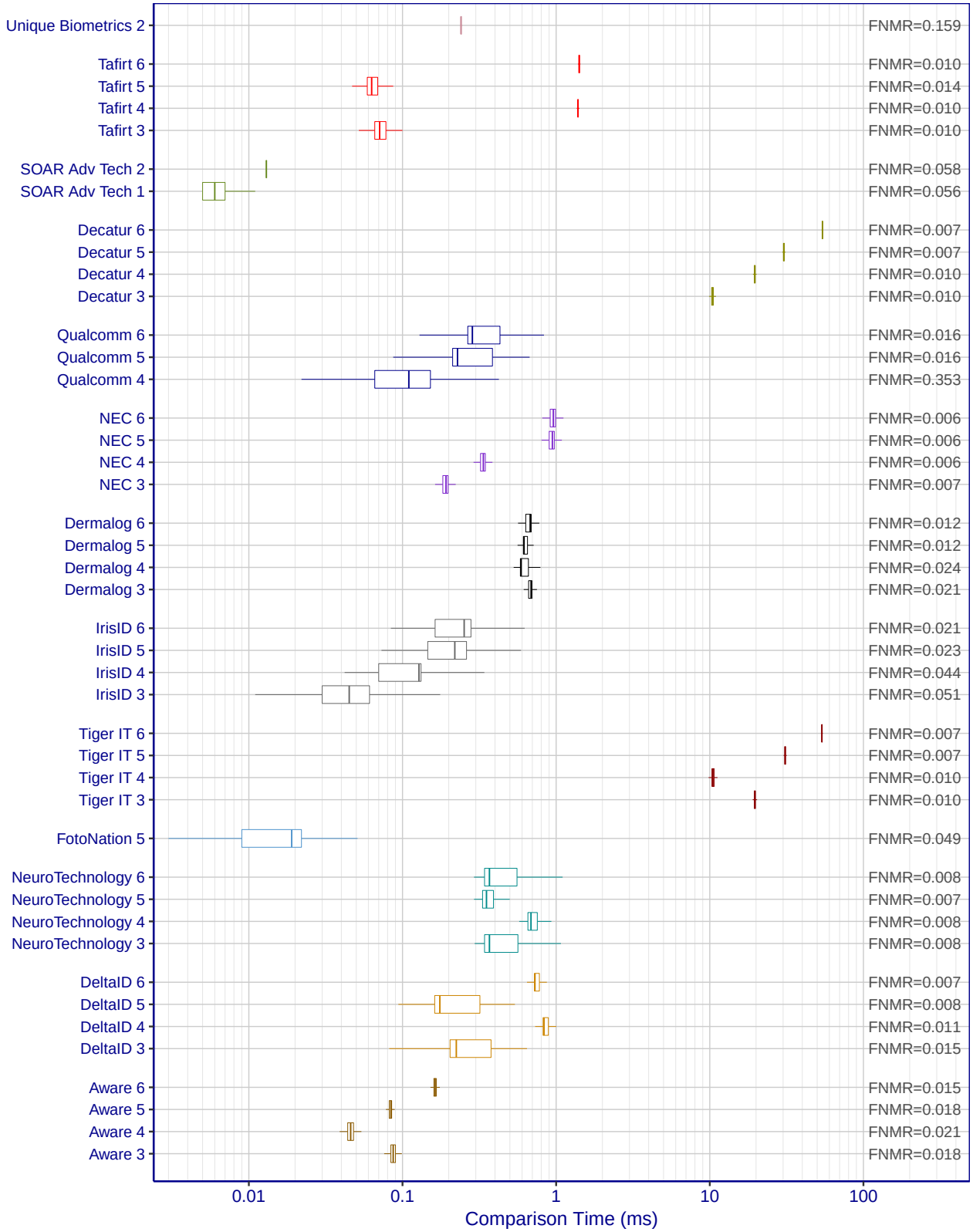

Figure 3.10: Boxplots summarizing the distribution of comparison times for each submission. Each template is created from a pair of iris samples (one of the left eye and one of the right). For reference, FNMR at an FMR of 10^{-5} is reported on the right. Boxplots were generated from 10 000 comparisons.

3.2.1.3 Speed-Accuracy Tradeoff

IREX III found that sometimes a speed-accuracy tradeoff exists for iris recognition, where improved accuracy can be achieved through slower, but more involved, comparison strategies. Figure 3.11 plots accuracy (FNMR at an FMR of 10^{-5}) vs. comparison time for mated comparisons. SOAR Advanced Technologies and FotoNation submitted the fastest matchers, but their accuracy lags compared to other submissions. The correlation coefficient for the log of the median search time and the log of FNMR is -0.30, indicating a weak but apparent speed-accuracy trade off.

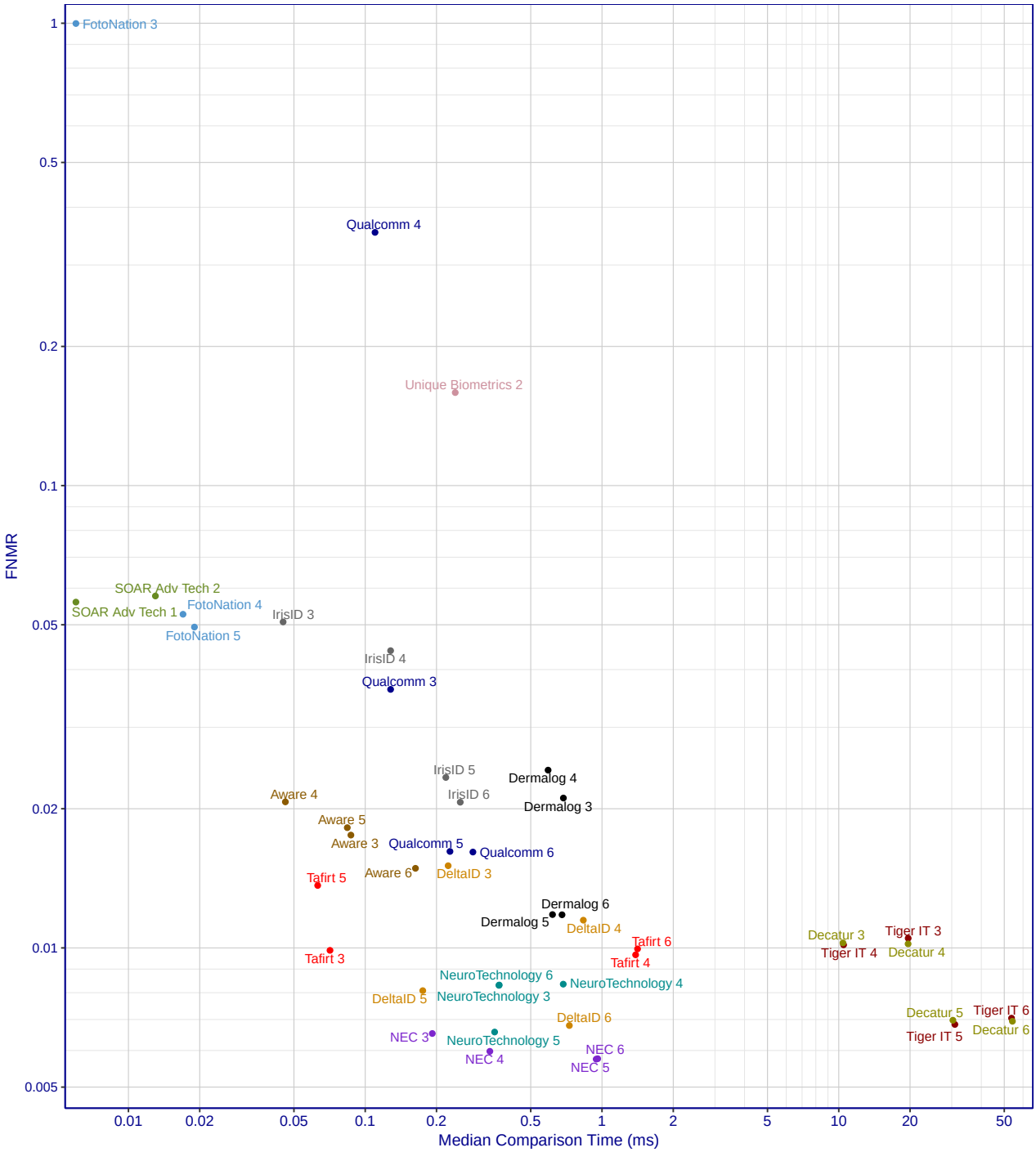

Figure 3.11: Comparison time vs accuracy for each submission. FNMR at an FMR of 10^{-5} is reported. All comparisons are two-eye. Timing statistics are for mated comparisons only.

3.2.2 One to Many Matching

The time it takes to create a template and search it against the enrolled population can affect throughput rates at physical access control points, biometric scanning stations, service kiosks, etc. Iris recognition matchers are capable of rapidly identifying users against large databases, which has led several countries to adopt iris-based methods of authenticating travelers at airports and border crossings [23]. Even when comparisons are performed offline, search speed can dictate computational hardware requirements.

In a typical centralized live-capture system, a biometric sample is acquired and then transferred over a network to a central facility. Once at the facility, a template is created from the sample and searched against the enrolled population. The results of the search are then used to send a response back over the network. Turnaround time is affected by the time it takes to both create a template and search it against the enrolled population. However, throughput rates may not be affected if the steps can be performed concurrently with other tasks. For example, a CBP officer could manually inspect a visitor's credentials while waiting for a response.

3.2.2.1 Search Time

Search time refers to the amount of time that elapses when a template is searched against an enrollment database. Timing statistics were collected for 1000 nonmated searches using a single processing core. Machines that have multiple cores can perform concurrent processing to speedup searches. IREX III found that using 16 cores simultaneously results in an 8 to 16 fold improvement in search time (for one matching algorithm that could operate in both single-threaded and multi-threaded mode). IREX IV found that the fastest matchers can search against an enrolled database of 1.6 million in under a second (although for single-eye matching). Figure 3.12 shows search times for each matcher when the enrolled population is $160\,000$. Figure 3.13 plots search time as a function of the size of the enrolled population.

Notable Observations

The fastest matcher (NeuroTechnology 6) was able to search against an enrolled population of $160\,000$ with a median search time under 11 milliseconds, faster than any matcher from IREX 4 and almost 50 times faster than any other matcher submitted to IREX IX. The other matchers submitted by NeuroTechnology achieve lower error rates, suggesting the participant may have deliberately traded some accuracy for speed. In terms of FNMR, the difference is 0.008 for NeuroTechnology 5 compared to 0.011 for NeuroTechnology 6 with roughly a hundred-fold difference in speed.

Figure 3.13 reveals that search time scales linearly with the size of the enrolled population for nearly all matchers (i.e. a doubling of the enrolled population size results in a doubling of search time). An exception might be FotoNation 5, which seems to begin with a sublinear relationship but quickly becomes linear by the time the enrolled population size reaches about $40\,000$. Another exception is NeuroTechnology 6. The slopes of its line segments suggest a sublinear relationship. When the enrolled population increases from $10\,000$ to $160\,000$ (a factor increase of 16), median search time increases from 3.3 milliseconds to 12.5 milliseconds (a factor increase of 3.8). Every doubling of the enrolled population size seems to increase median search time by a factor of 1.24.

Figure 3.12: Boxplots summarizing the distribution of search times for two-eye comparisons against an enrollment database of $160\,000$. For reference, FNIR at an FPIR of 10^{-3} is reported on the right. Each boxplot shows the distribution for 1000 searches.

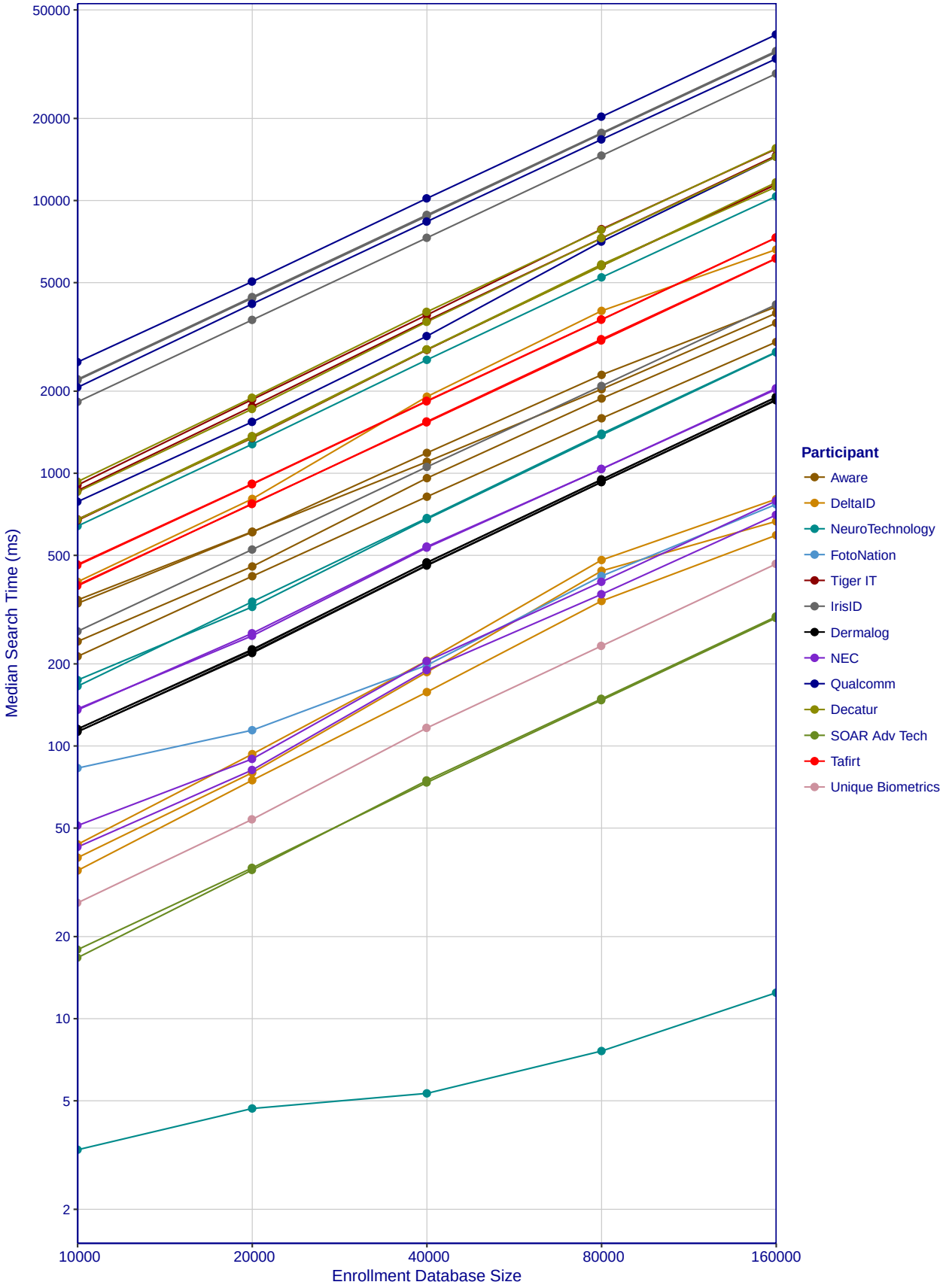

Figure 3.13: Median search time as a function of the enrolled population size for each submission. Both axes are on log scales. Most curves have a slope of one, indicating that a doubling of the enrolled population size doubles the search time. An exception is the bottom cyan curve representing Neurotechnology 5.

3.2.2.2 Speed-Accuracy Tradeoff

Although there is no pronounced speed-accuracy tradeoff, the most accurate matchers tend not to be the fastest. As was noted earlier, NeuroTechnology 6 has, by far, the shortest median search time but it is not as accurate as the other submissions from NeuroTechnology. It is roughly a hundred-fold faster than NeuroTechnology 5 but has an FNIR that is 35% greater (at FPIR=10^{-3}). Other participants that appear to exhibit a speed-accuracy trade off are NEC, Decatur, and Tiger IT.

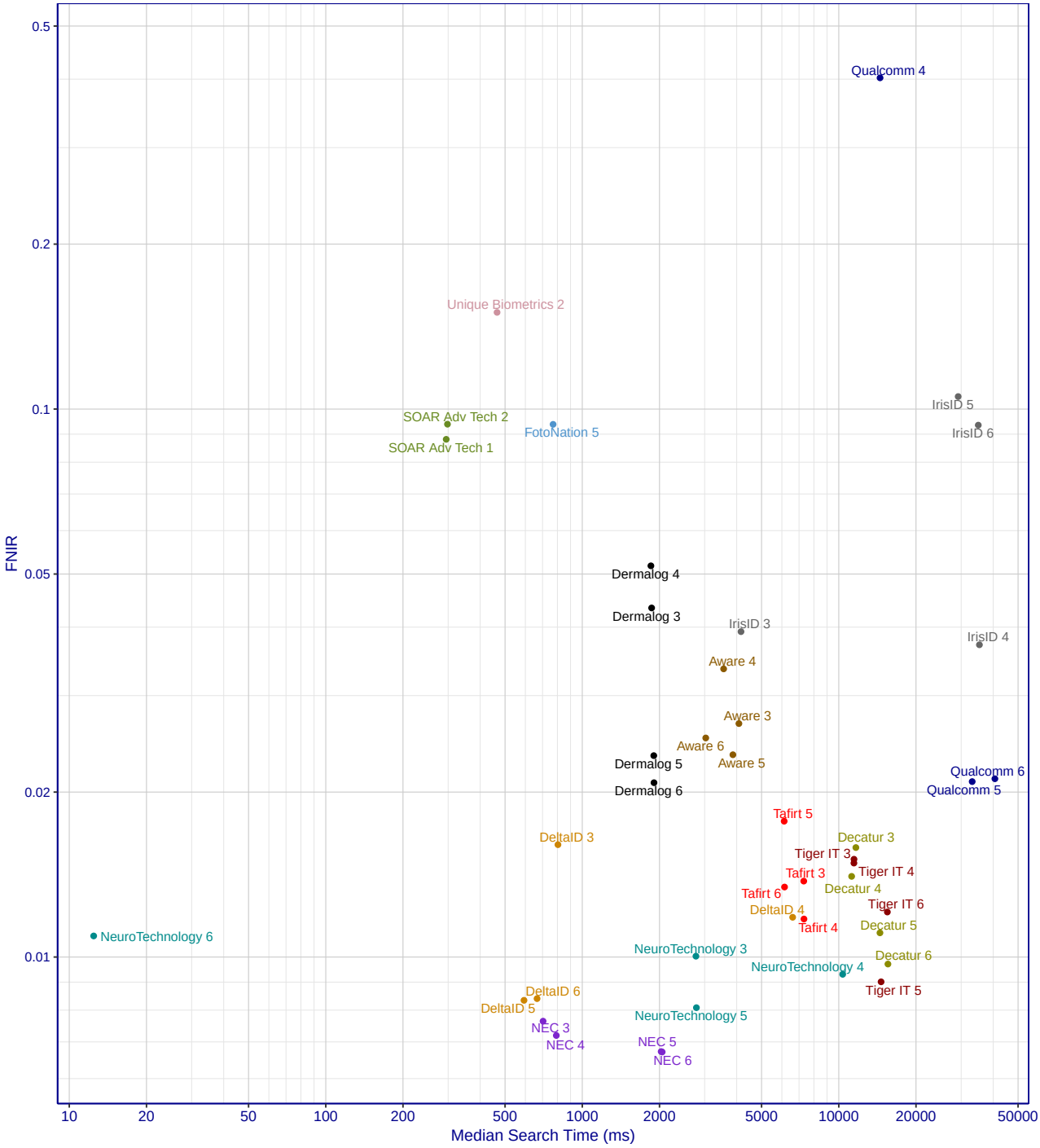

Figure 3.14: Median search time vs accuracy for each submission. The vertical axis is FNIR at an FPIR of 10^{-3} for two-eye comparisons against an enrolled population of $160\,000$ people. Timing statistics are for mated searches only.

3.3. Template Size

An iris template is a proprietary representation of the features from an iris sample (or samples). Although these templates are often smaller than the original iris samples, storage and exchange of iris data should be performed using the standard image formats defined in ANSI/NIST-ITL 1-2011 Update 15 and ISO/IEC 19794-6 to maintain interoperability and prevent vendor lock-in. Nevertheless, the size of proprietary templates can still dictate machine requirements. For example, in centralized systems the enrollment templates are often permanently loaded in memory to facilitate rapid searches against the database. The size of search templates is less important since they are typically only loaded into memory for the duration of the search. Table 3.1 shows summarary statistics on template sizes for one-to-many submissions. Zero-sized "blank" templates were excluded prior to computing mean template sizes.

Notable Observations

- **Range of Template Sizes:** The mean single-eye enrollment template size varies from 579 bytes (FotoNation 4) to 18 229 bytes (Decature 5, Decature 6, TigerIT 5, TigerIT 6). For search templates the mean size varies from 752 bytes (Unique Biometrics 2) to 17 914 bytes (Decatur 4 and TigerIT 4).
- **Asymmetrical Template Sizes:** Search and enrollment templates often have different mean sizes. FotoNation's enrollment templates are less than a tenth the size of their search templates. IrisID's enrollment templates are about half the size of their search templates. Tafirt's enrollment templates are also smaller than their search templates. Occasionally the search templates are slightly larger (e.g. all submissions from Qualcomm).
- **Variable Template Sizes:** Many submissions do not produce fixed-size templates. Templates created by Tiger IT and Decatur vary in size from one template to the other. As one would expect, two-eye templates tend to be about twice the size of single-eye templates. The exceptions are SOAR Adv Tech 1 and SOAR Adv Tech 2 where single-eye and two-eye templates are the same size. These submissions may not be performing effective two-eye comparisons (see Section 3.5). Some submissions create two-eye templates that take one of two sizes (e.g. DeltaID 3, Tafirt 4). Most likely, the two sizes correspond to whether features could be extracted from one or both iris samples.

	One-eye		Two-eye	
	Search Templates	Enrollment Templates	Search Templates	Enrollment Templates
Aware 3	1806 ± 0	1806 ± 0	3612 ± 0	3612 ± 0
Aware 4	1806 ± 0	1806 ± 0	3612 ± 0	3612 ± 0
Aware 5	1062 ± 0	1062 ± 0	2124 ± 0	2124 ± 0
Aware 6	1062 ± 0	1062 ± 0	2124 ± 0	2124 ± 0
DeltaID 3	2048 ± 0	2048 ± 0	4084 ± 157	4082 ± 171
DeltaID 4	6144 ± 0	6144 ± 0	12267 ± 359	12263 ± 393
DeltaID 5	2048 ± 0	2048 ± 0	4096 ± 0	4096 ± 0
DeltaID 6	6144 ± 0	6144 ± 0	12288 ± 0	12288 ± 0
NeuroTechnology 3	2348 ± 0	2348 ± 0	4676 ± 0	4676 ± 0
NeuroTechnology 4	4676 ± 0	4676 ± 0	9332 ± 0	9332 ± 0
NeuroTechnology 5	2348 ± 0	2348 ± 0	4676 ± 0	4676 ± 0
NeuroTechnology 6	2348 ± 0	2348 ± 0	4676 ± 0	4676 ± 0
FotoNation 3	7481 ± 0	581 ± 0		1153 ± 73
FotoNation 4	7479 ± 0	579 ± 0		1155 ± 42
FotoNation 5	7481 ± 0	581 ± 0	14886 ± 752	1157 ± 53
Tiger IT 3	17403 ± 2889	17709 ± 3000	34830 ± 5427	35370 ± 5585
Tiger IT 4	17914 ± 2895	18219 ± 3010	35852 ± 5438	36391 ± 5602
Tiger IT 5	17817 ± 3156	18229 ± 3222	35673 ± 5994	36412 ± 6132
Tiger IT 6	17817 ± 3156	18229 ± 3222	35673 ± 5994	36412 ± 6132
IrisID 3	3080 ± 0	1544 ± 0	6160 ± 0	3088 ± 0
IrisID 4	4104 ± 0	2056 ± 0	8208 ± 0	4112 ± 0
IrisID 5	3080 ± 0	1544 ± 0	6160 ± 0	3088 ± 0
IrisID 6	4104 ± 0	2056 ± 0	8208 ± 0	4112 ± 0
Dermalog 3	1899 ± 1	1887 ± 149	3765 ± 247	3773 ± 227
Dermalog 4	1899 ± 1	1887 ± 149	3765 ± 247	3773 ± 227
Dermalog 5	1899 ± 1	1897 ± 62	3792 ± 105	3793 ± 91
Dermalog 6	1899 ± 1	1897 ± 62	3792 ± 105	3793 ± 91
NEC 3	4632 ± 0	4632 ± 0	9248 ± 0	9248 ± 0
NEC 4	4632 ± 0	4632 ± 0	9248 ± 0	9248 ± 0
NEC 5	6178 ± 0	6178 ± 0	12330 ± 0	12330 ± 0
NEC 6	6178 ± 0	6178 ± 0	12330 ± 0	12330 ± 0
Qualcomm 3	2104 ± 0	2128 ± 0	4208 ± 0	4256 ± 0
Qualcomm 4	2104 ± 0	2128 ± 0	4208 ± 0	4256 ± 0
Qualcomm 5	3256 ± 0	3280 ± 0	6512 ± 0	6560 ± 0
Qualcomm 6	3256 ± 0	3280 ± 0	6512 ± 0	6560 ± 0
Decatur 3	17403 ± 2889	17709 ± 3000	34830 ± 5427	35370 ± 5585
Decatur 4	17914 ± 2895	18219 ± 3010	35852 ± 5438	36391 ± 5602
Decatur 5	17817 ± 3156	18229 ± 3222	35673 ± 5994	36412 ± 6132
Decatur 6	17817 ± 3156	18229 ± 3222	35673 ± 5994	36412 ± 6132
SOAR Adv Tech 1	965 ± 0	965 ± 0	965 ± 0	965 ± 0
SOAR Adv Tech 2	1925 ± 0	1925 ± 0	1925 ± 0	1925 ± 0
Tafirt 3	3850 ± 0	2464 ± 0	7692 ± 177	4923 ± 107
Tafirt 4	6946 ± 0	5560 ± 0	13884 ± 235	11111 ± 222
Tafirt 5	3850 ± 0	2464 ± 0	7650 ± 437	4902 ± 251
Tafirt 6	6946 ± 0	5560 ± 0	13878 ± 307	11108 ± 261
Unique Biometrics 2	752 ± 0	752 ± 0	1471 ± 155	1473 ± 150

Table 3.1: Mean template size in bytes along with standard deviations.

3.4. Impact of Demographics

Biometric systems may perform better or worse for certain demographic groups. When such a system is deployed for common activities (e.g. access control, border crossing) poor matching accuracy for a particular group can disproportionately impact members of that group. This section breaks out accuracy for three demographic factors: age, sex, and eye colour. Because the dataset consists of samples collected under various circumstances from several locations over a period of years, we cannot discount the possibility that any apparent demographic effects are actually due to some form of selection bias.

Figure 3.15 shows single-eye DET accuracy for each matcher broken out by sex. Figure 3.17 shows DET accuracy for each matcher broken out by race. It should be noted that 'Asian' can refer to anyone from the Asian continent, including East Asia and most of the Middle East. Race is problematic as a method of categorization because it is subjective and based at least partially on social and cultural traits. The Face Recognition Vendor Test (FRVT) Ongoing [24] uses the less ambiguous term 'geographic region'. However, we are limited by the format in which the OPS-III meta data was collected and stored. Figure 3.16 shows DET accuracy for each matcher broken out by eye colour.

Notable Observations

- **Sex:** Sex appears to have a significant impact on accuracy for some matchers, but the effect is not consistent. Some matchers perform better on males (e.g. SOAR Adv. Tech 1, Iris ID 3, Dermalog 4) while others perform better on females (e.g. NEC 6, Tafirt 6, DeltaID 4). This inconsistent behavior sometimes holds even for different matchers from the same participant. For example, Aware's earlier submissions (Aware 3 and Aware 4) perform better on males while their later submissions (Aware 5 and Aware 6) perform better on females. The gray lines of equal threshold reveal that, in most cases, the accuracy differences have more to do with changes in FNMR rather than FMR. That is, the matcher has an easier time recognizing that two samples represent the same iris when they come from one sex over the other. The cause of this behavior is unknown and may have to do with ease of localizing the iris boundaries across sexes. Sometimes females wear mascara which can make localization of the iris boundaries more difficult.

- **Eye Colour:** Eye colour was consolidated into the binary categories *light* (blue, green, and grey) and *dark* (brown and black). Eye colour was not recorded for all subjects. The mated comparison sets contain 32 thousand dark-eye comparisons and 8 thousand light-eye comparisons. The nonmated comparison sets contain 71 million dark-eye comparisons and 2 million light-eye comparisons.

  Thirteen matchers perform noticeably better on dark eyes while 27 appear to perform better on light eyes, although the difference is often small. Tiger IT's earlier submissions (TigerIT 3 and Tiger IT 4) yield much lower error rates on dark eyes while their later submissions (Tiger IT 5 and Tiger IT 6) yield lower error rates on light eyes. The same holds true for Decatur's matchers. The lines of equal threshold reveal that lighter eyes are more likely to false match for most matchers. The most accurate matchers (e.g. NEC 5, NeuroTechnology 5, Decatur 5) appear to perform better on lighter eyes. There are several possible explanations for these behaviors:

  – Different behaviors from those with lighter coloured eyes. Eye colour covaries with race and other demographic factors that might be responsible for the true effect.
  – Differences in the ease of localization of the iris boundaries for different eye colours.
  – Difference in statistical richness of the iris "texture" between lighter and darker coloured irides.

- **Race:** The three races considered are White, Black, and Asian. The mated comparison sets contain 30 thousand comparisons between eyes from Whites, 6 216 comparisons between eyes from Blacks, and 2 858 comparisons between eyes from Asians. The nonmated comparison sets contain 65 million comparisons between eyes from Whites, 3.5 million comparisons between eyes from Blacks, and 2.2 million comparisons between eyes from Asians.

  The matchers tend to perform best on Whites and poorest on Asians. This is not true in all cases and sometimes the differences are negligible. Race effects could be due to biases in training data. Companies based in Europe and the United States may have easier access to iris samples from Whites. That said, NEC is based in Japan but performs better on Whites. The effect could be because East Asians have smaller palperbal fissures on average [25]. It is also possible that apparent race effects are simply the result of random variation that could be resolved with more test data. The lines of equal threshold show that for most matchers, comparisons between Whites are less likely to false non-match but more likely to false match in relation to Asians. *Further investigation in necessary before drawing any solid conclusions.*

3.4.1 Sex

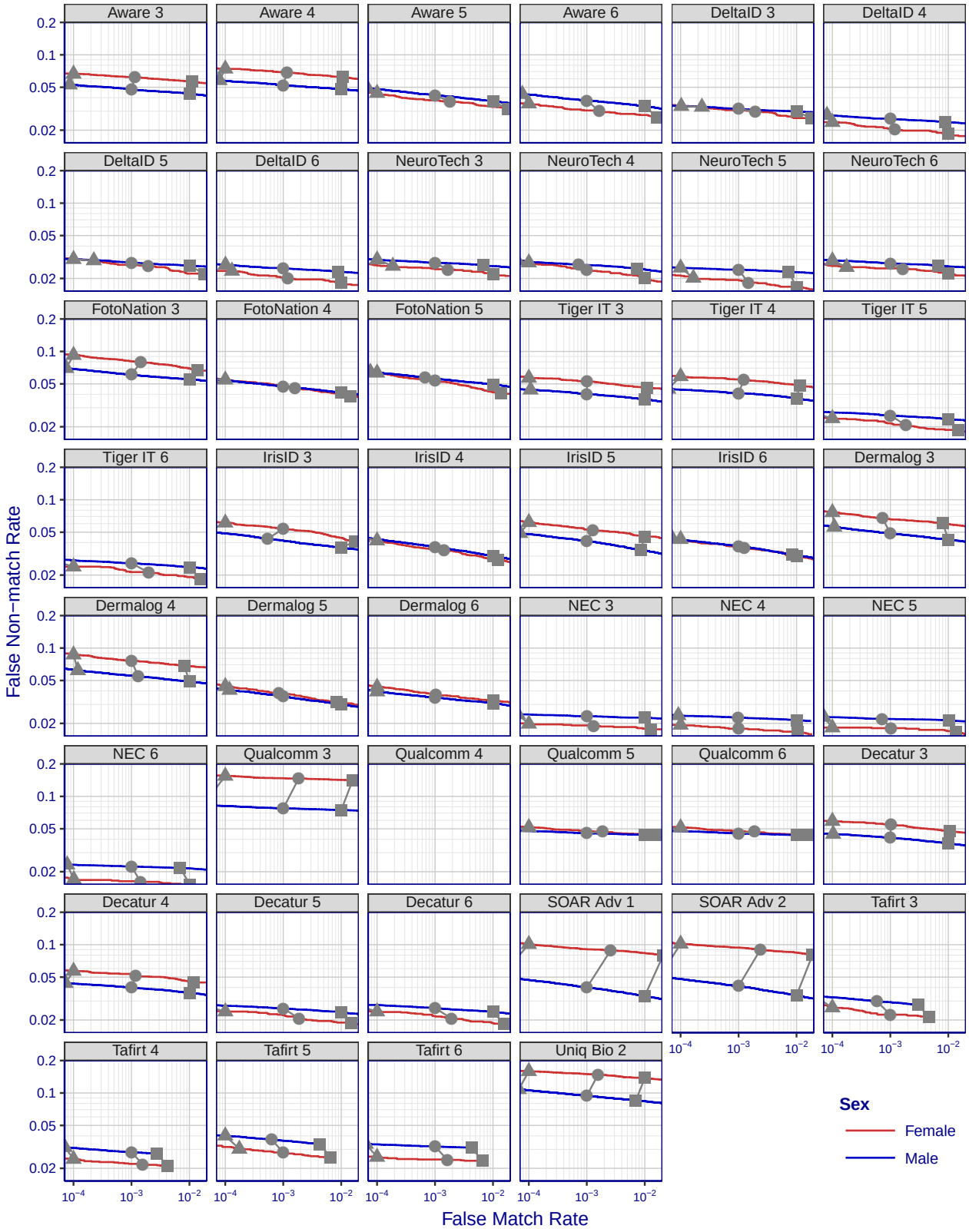

Figure 3.15: Impact of sex on accuracy for each submission. Grey lines connect points of equal threshold.

3.4.2 Eye Colour

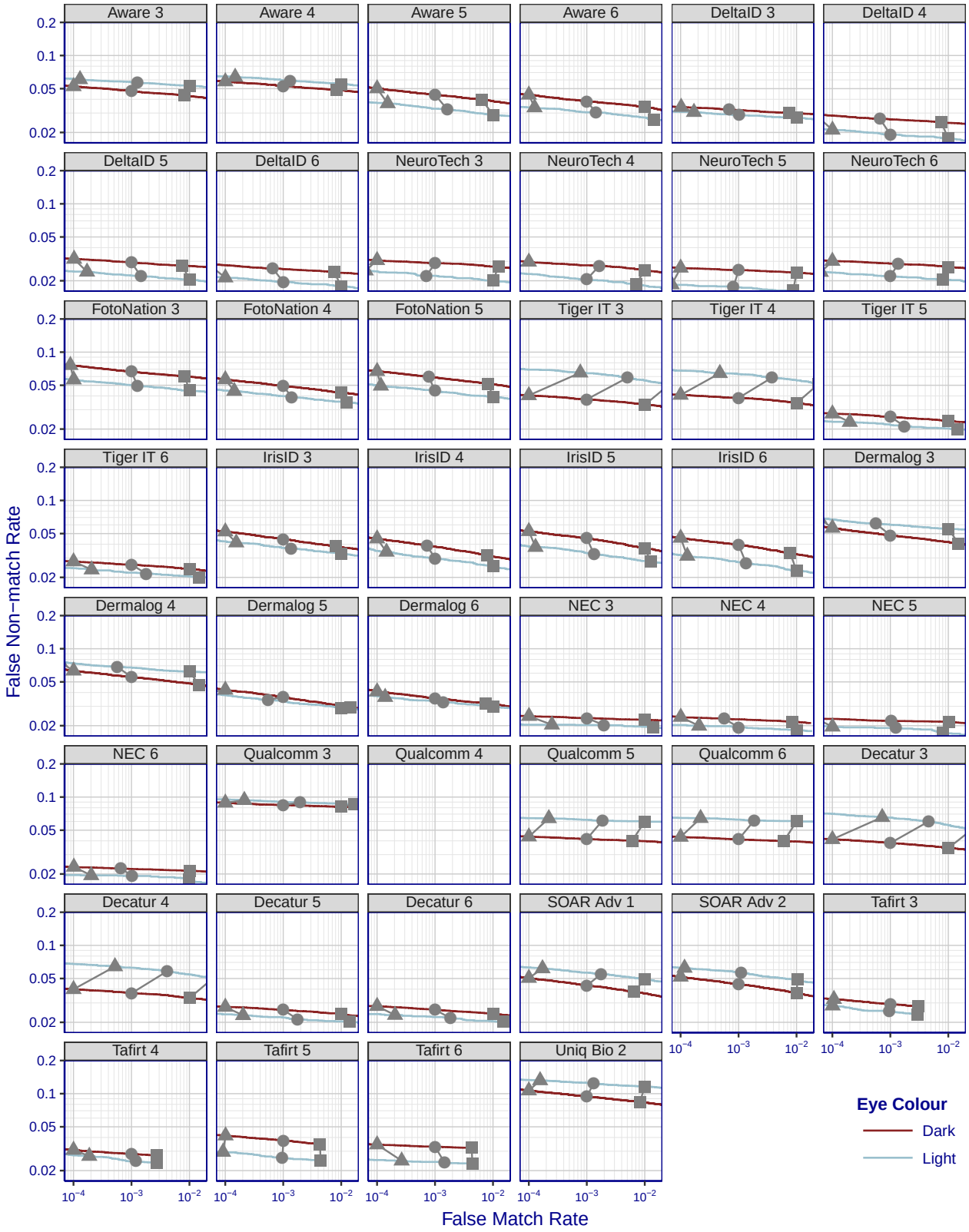

Figure 3.16: Impact of eye colour on accuracy for each submission. Grey lines connect points of equal threshold.

3.4.3 Race

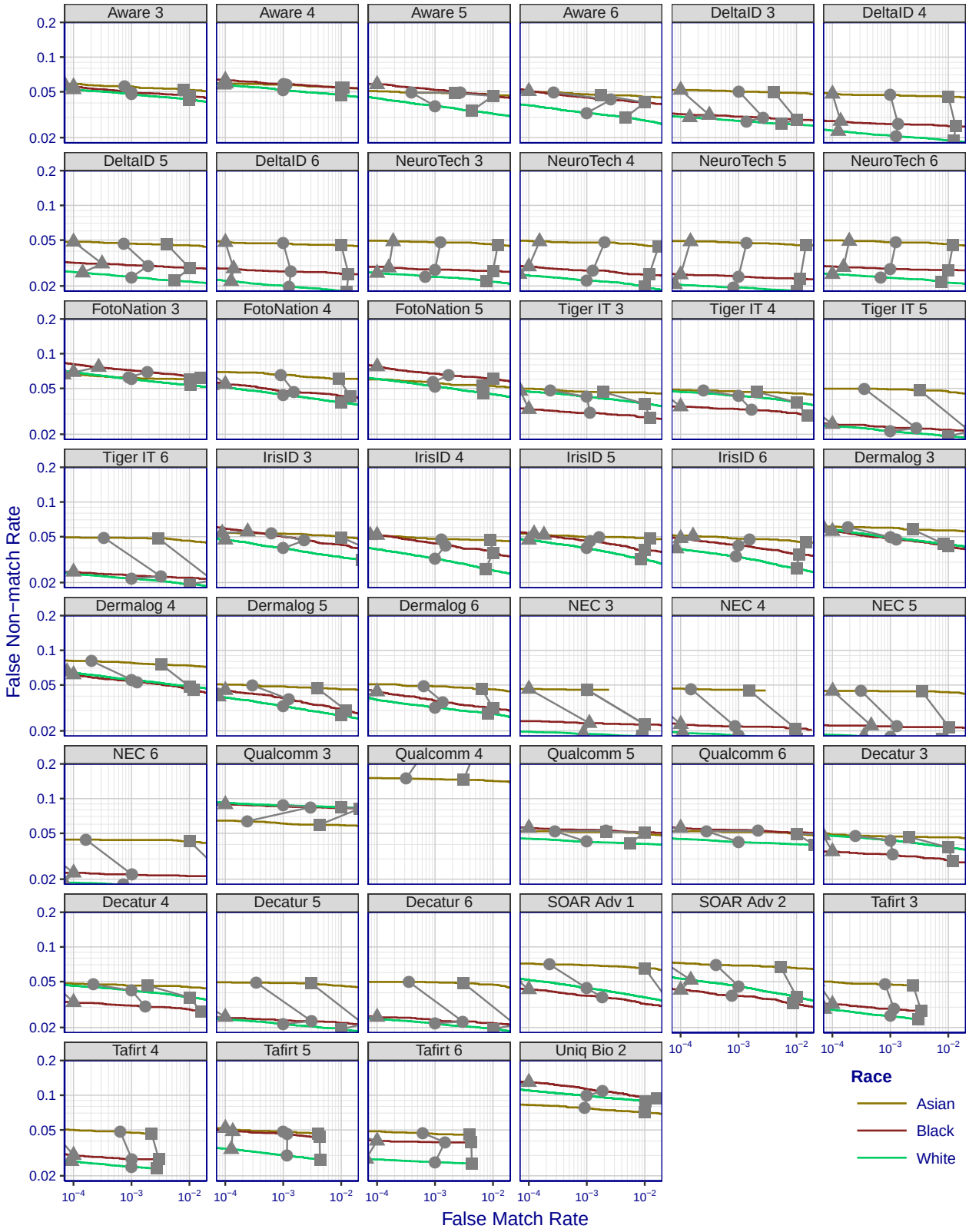

Figure 3.17: Impact of race on accuracy for each submission. Grey lines connect points of equal threshold.

3.5. Single vs. Dual Eye Accuracy

Most contemporary iris recognition systems use both eyes for verification where samples of both eyes are captured concurrently. As a result, the image quality characteristics for the left and right eyes are highly correlated. Examples of highly correlated effects include: (1) blinks, (2) squints, (3) excessive pupil dilation caused by the consumption of certain drugs and (4) bilateral congenital defects such as non-circular pupils (as shown in the IREX III Supplement [26]), although there are certainly individual cases where these dependencies do not occur. Even when a single-eye camera is used to capture the images in succession, many of these correlations will remain because of subject and/or operator habits or environmental factors. We note the advantages of dual-eye recognition over single-eye recognition in most scenarios, but point out that it does not provide the overall level of performance improvement that would be expected if the left and right eye captures were statistically independent events [1]. In one important respect, the correlation in a dual-eye capture is beneficial – the "roll" of the subject's head is the same for both eyes and that enables the roll to be compensated for in software, which in turn leads to a reduction in the range of roll that needs to be searched. To first order, cutting the range of roll by 2X reduces the computational complexity of the match calculation by 2X – an important effect in system optimization.

Though specific scenarios might benefit from single-eye capture, in most deployments dual-eye capture is likely the better choice. This issue will be discussed in depth in a separate paper on iris camera properties that is under development at NIST.

Notable Observations

Figure 3.18 compares one-eye and two-eye DET accuracy. The IREX III [28] and IREX IV [1] reports state that switching from one-eye to two-eye comparisons appears to result in a downward translation (on a log scale) of the DET curve. The number reported in the upper right-hand corner of each pane is the mean factor increase in FNMR at fixed FMR when switching from two-eye to single-eye comparisons. So, for example, $r = 2$ would correspond to a factor of two increase in FNMR. Mean factor reductions vary from a low of 0.75 to a high of 3.93. Values less than one indicate worse performance for two-eye comparisons, obviously indicating a problem.

An iris recognition system could be designed to support both one-eye and two-eye comparisons. This would certainly be reasonable if images of both eyes were only available some of the time. At a fixed decision threshold, most matchers are more likely to return false non-matches for single-eye comparisons. Behavior for false matches is less consistent, with some matchers more likely to return false matches for single eye comparisons (e.g. Dermalog 5) and others less likely to return false matches for single-eye comparisons (e.g. DeltaID 3). Some participants may be normalizing their comparison scores to keep the false match rate consistent (e.g. Aware 3 and Aware 4).

[1] similar effects have been observed for fingerprints [27]

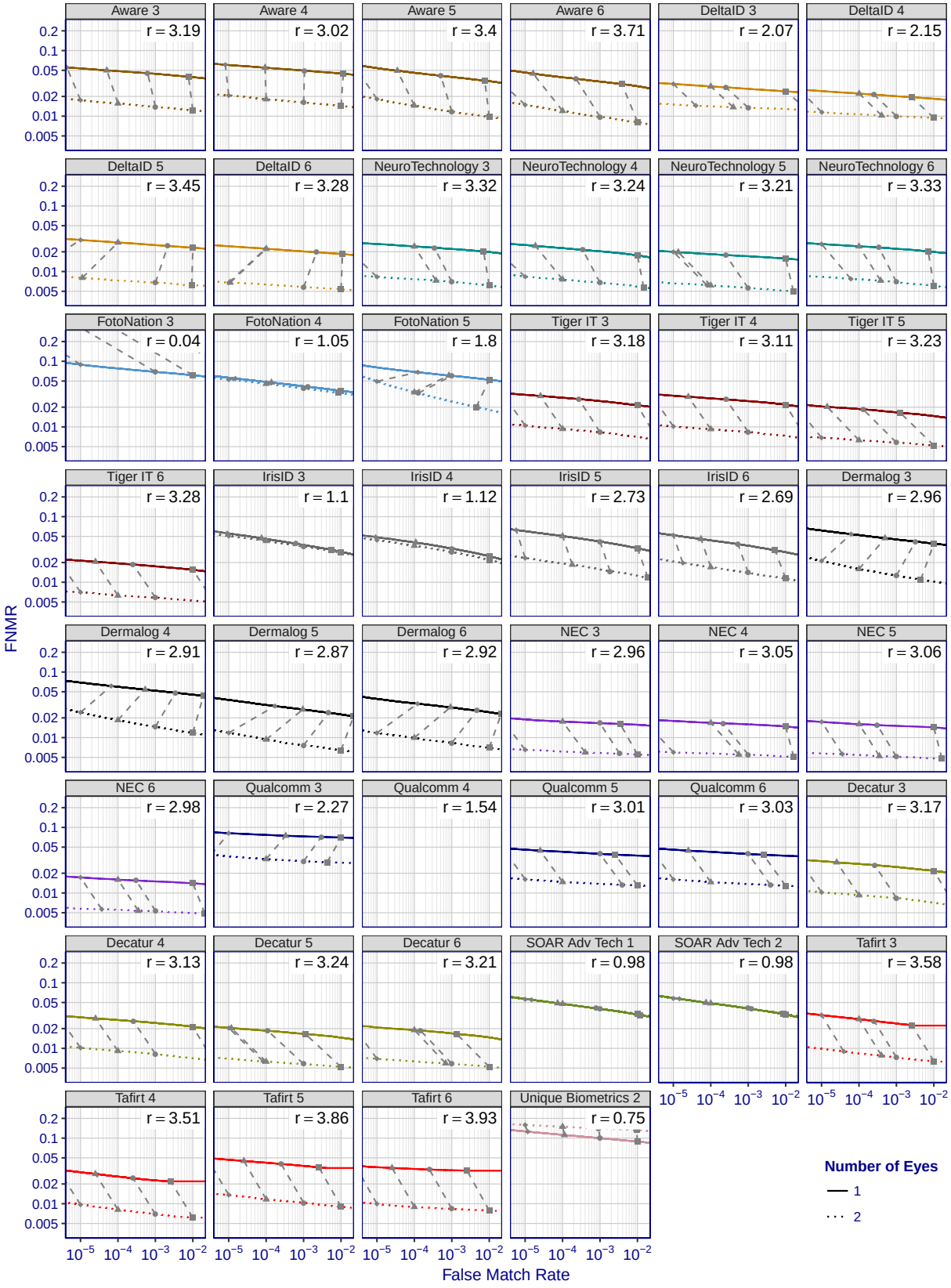

Figure 3.18: Comparison of one-eye and two-eye accuracy for one-to-one comparisons. Gray line segments connect points of equal threshold, showing how error rates for single-eye comparisons relate to two-eye comparisons at the same decision threshold. The number reported in the upper right corner is the mean factor increase in FNMR at fixed FMR when switching from two-eye to single-eye comparisons.

4 References

[1] G. W. Quinn, P. Grother, and M. Ngan, "IREX IV Part 1: Evaluation of Iris Identification Algorithms." https://www.nist.gov/publications/irex-iv-part-1-evaluation-iris-identification-algorithms, 2014. 1, 7, 15, 37, 46

[2] G. W. Quinn, J. Matey, E. Tabassi, and P. Grother, "IREX V: Guidance for Iris Image Collection." https://www.nist.gov/itl/iad/image-group/irex-v-homepage, 2014. 1

[3] P. Grother, E. Tabassi, G. W. Quinn, and W. Salamon, "Performance of Iris Recognition Algorithms on Standard Images." https://www.nist.gov/itl/iad/image-group/irex-i, 2009. 7

[4] E. Tabassi, P. Grother, and W. Salamon, "IREX - IQCE Performance of Iris Image Quality Assessment Algorithms." https://www.nist.gov/itl/iad/image-group/irex-ii-iqce, 2011. 7

[5] *ISO/IEC 29794-6 - Biometric Sample Quality Standard- Part 6: Iris Image*. 2012. 7

[6] "The IREX Program." https://www.nist.gov/programs-projects/iris-exchange-irex-overview. 8

[7] "Prime Minister launches Aadhaar Enabled Service Delivery." Press Release, October 2012. http://uidai.gov.in/images/2nd_anniversary/uidai_press_release_for_oct_20.pdf. 8

[8] A. N. Al-Raisi and A. M. Al-Khouri, "Iris recognition and the challenge of homeland and border control security in UAE," *Telematics and Informatics*, vol. 25, no. 2, pp. 117–132, 2008. 8

[9] "United Arab Emirates Iris System." https://www.ica.gov.ae/. 8

[10] 8

[11] G. Quinn, P. Grother, and J. Matey, "Multi-Spectral Iris Evaluation Concept, Evaluation Plan, and API Specification Version 1.3." https://www.nist.gov/sites/default/files/documents/2017/02/23/irex9_conops.pdf, Sept. 2017. 9, 10

[12] J. Daugman, "How iris recognition works," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, pp. 21–30, Jan 2004. 10

[13] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech*, pp. 1895–1898, 1997. 10, 42

[14] J. L. Wayman, "Confidence Interval and Test Size Estimation for Biometric Data," in *IEEE Proc. AutoID*, pp. 177–184, 1999. 11, 46

[15] G. Levin, "Real World. Most Demanding Biometric Applications," in *Biometric Consortium Conference*, 2007. 12

[16] "Interagency Advisory Board Meeting Agenda." http://fips201.com/resources/audio/iab_1111/iab_110111_nagel.pdf. 12

[17] "DoD to use iris scans, fingerprints for building security." http://www.federalnewsradio.com/396/2817314/DoD-to-use-iris-scans-fingerprints-for-building-security. 12

[18] "Southwestern Border Sheriffs' Coalition (SBSC) to immediately begin improving the biometric identification capabilities of the 31 Sheriffs' Offices along the U.S. and Mexico Border to increase border security and combat criminal activity." https://www.businesswire.com/news/home/20170406006132/en/Southwestern-Border-Sheriffs-Coalition-SBSC-immediately-improving. 12

[19] "Iris recognition immigration system (IRIS)." https://web.archive.org/web/20140109063540/http://www.ukba.homeoffice.gov.uk/customs-travel/Enteringtheuk/usingiris/. 15

[20] J. Daugman, "Evolving Methods in Iris Recognition," in *Biometrics: Theory, Applications and Systems*, 2007. 15

[21] J. Daugman, "Iris Recognition: Algorithms, Performance, and Challenges," in *Biometrics Consortium Conference*, 2007. 15

[22] J. Daugman, "The importance of being random: statistical principles of iris recognition.," *Pattern Recognition*, vol. 36, no. 2, pp. 279–291, 2003. 15

[23] J. Daugman and I. Malhas, "Iris Recognition Border-Crossing System in the UAE," *International Airport Review*, vol. 8, no. 2, pp. 49–53, 2004. 27

[24] P. Grother, M. Ngan, and K. Hanaoka, "Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification." https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing, 2018. 33

[25] L. Farkas, *Anthropometry of the head and face*. Raven Press, 1994. 33

[26] G. Quinn and P. Grother, "IREX III Supplement I: Failure Analysis." https://www.nist.gov/itl/iad/image-group/irex-iii-homepage, 2011. 37

[27] E. Tabassi, "Quality measure workshop lessons learned," in *Latent Testing Workshop 2006*, NIST, 2006. 37

[28] P. Grother, G. Quinn, J. Matey, M. Ngan, W. Salamon, G. Fiumara, and C. Watson, "IREX III: Performance of Iris Identification Algorithms." https://www.nist.gov/itl/iad/image-group/irex-iii-homepage, 2011. 37

[29] Lehmann-Scheffe, "Completeness, similar regions, and unbiased estimation," *Sankhya*, vol. 10, pp. 305–340, 1950. 42

[30] "Information Technology - Vocabulary, Part 37: Biometrics," standard, International Organization for Standardization, Geneva, CH, Feb. 2017. 42, 44

[31] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior, *Guide to Biometrics*. Springer, 2004. 43

[32] L. D. Brown, T. T. Cai, and A. Dasgupta, "Interval Estimation for a Binomial Proportion," *Statistical Science*, vol. 16, pp. 101–133, 2001. 45

[33] G. W. Quinn, P. Grother, M. Ngan, and N. Rymer, "IREX IV: Part 2 Compression Profiles for Iris Image Compression." https://www.nist.gov/publications/irex-iv-part-2-compression-profiles-iris-image-compression, 2014. 46

[34] "Biometric Ideal Test: CASIA-IrisV4." http://biometrics.idealtest.org/dbDetailForUser.do?id=4. 46

[35] K. W. Bowyer and P. J. Flynn, "The ND-IRIS-0405 iris image dataset," *CoRR*, vol. abs/1606.04853, 2016. 46

[36] A. J. Mansfield and J. L. Wayman, "Best Practices in Testing and Reporting Performance of Biometric Devices," tech. rep., National Physical Laboratory. 46

[37] M. E. S. , *Computational Methods in Biometric Authentication*. Information Science and Statistics„ London :: Springer London :, 2010. 46

[38] P. Bickel, "Response to SAG Problem 97-23." University of California, Berkeley, Department of Statistics. 46

Appendices

A Computation of Performance Statistics

This appendix formally defines the statistics that characterize the performance of iris recognition algorithms. As an offline evaluation, IREX IX cannot test all aspects of an operation system. It does not include a live image acquisition component or any interaction with real users. The core accuracy statistics are defined in A.1 and A.2 while timing statistics are described in A.3.

A.1. One to One Matching

One-to-one iris comparisons produce measures of dissimilarity between biometric templates. If the dissimilarity score is at or below a preset decision threshold, the comparison is classified as "mated", meaning the templates represent the same biometric characteristic. If the dissimilarity score is above the decision threshold, the comparison is classified as "nonmated", meaning the two samples represent different biometric characteristics. Two types of decision error are possible. The first is a *false match*, where a nonmated comparison is erroneously classified as mated. The second is a *false nonmatch*, where a mated comparison is erroneously classified as nonmated. The False Match Rate (FMR) is the rate at which false matches occur for nonmated comparisons. Formally, if u_i are nonmated dissimilarity scores (with $i = 1,...,N$), then FMR is estimated as

$$\text{FMR}(t) = \frac{1}{N}\sum_{i=0}^{N} H(u_i - t) \tag{A.1}$$

where t is the decision threshold and $H(.)$ is the Heaviside step function,

$$H(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1, & \text{otherwise.} \end{cases} \tag{A.2}$$

The FNMR is similarly estimated for mated comparisons υ_i (with $i = 1,...,M$):

$$\text{FNMR}(t) = \frac{1}{M}\sum_{i=0}^{M} H(t - \upsilon_i). \tag{A.3}$$

Adjusting the decision threshold, t, reduces the rate of one type of error, but at the expense of the other. This relationship is characterized by a DET curve, which plots the tradeoff between the two error rates. DET curves has become a standard in biometric testing, superseding the analogous ROC curve. Compared to ROC curves, the logarithmic axes of DET curves provide a superior view of the differences between matchers in the critical high performance region [13].

As estimates of the true population parameters, Equations A.1 and A.3 are both *complete* and *sufficient*, making them Uniform Minimum Variance Unbiased Estimates (UMVUEs) [29]. Given no other information about the comparisons other than the dissimilarity scores, they provide the best (*i.e.* lowest variance) estimates of FMR and FNMR among all unbiased estimates.

A Failure to Enroll (FTE) occurs when a reference template could not be created, usually because no useful feature information could be extracted from the image(s). The analogous case for biometric probes is a Failure to Acquire (FTA). Participants were instructed to submit matchers that always create comparable templates, even when no feature information could be extracted. These "blank templates" are expected to produce high measures of dissimilarity (effectively infinity) when compared. This was done for ease of testing but does not accurately reflect operational reality since, for example, a blank template would never be saved onto a smartcard and used for access control. This inability to handle template creation errors in realtime highlights a weakness of offline testing.

Biometric comparisons test whether biometric characteristics match, which is slightly different than testing whether the individuals themselves match. It is possible to falsely match a sample of a person's left eye to a sample of his right eye. Operationally, this might lead to the correct decision being made for the wrong reason (sometimes referred to as a "Type III error"). Left and right irides from the same person are never directly compared in this evaluation. But even if they were, credit would not be given for erroneously classifying such comparisons as mated.

This report uses the terms FMR and FNMR rather than the analogous but less general terms "false accept rate" (FAR) and "false reject rate" (FRR). The latter terms only apply when an authoritative claim is made about the origin of the biometric sample [30]. Many biometric applications involve making a negative identity claim - i.e. that a biometric sample is not represented by the reference sample. For example, if fingerprints are lifted from a crime scene, they may be used to exclude certain suspects. In this scenario, a "false accept" would describe the event where the negative identity claim is falsely rejected.

A.2. One to Many Matching

Open-set biometric systems are tasked with searching a biometric characteristic against an enrollment database and returning zero or more candidates. A candidate is returned if the matcher determines that its dissimilarity to the searched image is below a pre-determined decision threshold. A false positive occurs when a search returns a candidate for an individual that *is not* enrolled in the database. A false negative occurs when a search *does not* return the correct candidate for an individual that *is* enrolled in the database. Raising the decision threshold increases the rate of false positives but decreases the rate of false negatives.

False positives are computed exclusively from *non-mated* searches (i.e. searches for which the searched biometric characteristic is not enrolled in the database). This is more reflective of operation than if false positives had been computed from mated searches with the correct candidates removed from the list.

Formally, let s_i be the dissimilarity score between the ith searched sample and its enrolled mate (with $i = 1, ..., M$). Additionally, let r_i be the rank of the enrolled mate in the candidate list. Then the estimate of FNIR is

$$\text{FNIR}(t, \ell) = \frac{1}{M} \sum_{i=1}^{M} H(r_i - \ell) \wedge H(s_i - t). \tag{A.4}$$

where t is the decision threshold and ℓ is the rank requirement. The first call to the Heaviside step function ensures the mate fulfills the rank requirement. The second ensures it fulfills the score requirement. A biometric system may only return this information if the enrolled mate does, in fact, fulfill both of these requirements. The FPIR depends only on the dissimilarity score for the top ranked candidates for each search. Let q_i be the dissimilarity score for the top ranked candidate for the ith nonmated search (with $i = 1, ..., Q$). Then the FPIR is estimated by

$$\text{FPIR}(t) = \frac{1}{Q} \sum_{i=1}^{Q} H(t - q_i). \tag{A.5}$$

Although the above metrics do not represent error rates in a binary classification system, core accuracy is still presented in the form of Detection Error Tradeoff (DET) plots, this time showing the tradeoff between the FPIR and the FNIR as the decision threshold is adjusted. When evaluating iris matchers, we did not look for the correct mate past the 10th most similar candidate (thus, $\ell = 10$). For iris recognition the rank requirement has little effect on the estimate of FNIR since, at any reasonably selective threshold, the correct mate rarely fulfills the threshold requirement without also fulfilling the rank requirement.

Equation A.5 defines FPIR as the fraction of non-mated searches for which *at least one* candidate has a dissimilarity score at or below threshold. Selectivity-Reliability curves [31] compute *selectivity* as the average number of false positives returned for a non-mated search. This differs from our metric in that it takes into account the actual number of false positives returned for a particular search beyond just the first. *Selectivity* is a better metric for investigational mode applications where each candidate must be inspected by a human examiner (and thus workload scales with the number of returned candidates). That said, our definition of FPIR is grounded on the assumption that most operational uses of iris recognition result in similar outcomes regardless of whether the search returns one or several false positives. For example, one-to-many access control systems grant access to users as long as they match at least one enrolled individual.

Some DETs in this report include line segments between curves that connect points of equal threshold. The two curves might differ by enrolled population size or the number of iris samples used, and a connecting line segment shows how error rates for one curve compare to the other at the same decision threshold.

A.3. Computation Timing

Timing statistics are presented for primary operations (e.g. searching, template generation) as the actual physical time that elapsed. The C++ chrono library is used, which has nanosecond resolution on the test machines. Timing statistics are collected for single-threaded operations on otherwise unloaded machines. For ease of testing and fair comparison, algorithms were required to operate in single-threaded mode. Operationally, an algorithm can be designed to exploit multiple cores when available to expedite searching and enrollment.

B Uncertainty Estimation

This appendix describes how estimates of variability are computed in this report. Estimates of variability do not directly describe the population. Rather, they convey information about the primary statistics that are used to make inferences about the population. The primary statistics in this report are the core accuracy metrics defined in Appendix A. Variability refers to how tightly these statistics represent the true population parameters.

The core accuracy metrics are computed over a sample of data subjects [1] selected from a larger population. In our case, the population is a set of field-collected iris samples used by a government agency. We do not know the sampling methodology used to procure our test data so we are forced to assume simple-random-sampling of the data subjects. The iris images collected from the data subjects are paired in various ways to form comparison sets. These pairings introduce a correlation structure that must be incorporated into the estimates of variability.

B.1. One-to-one Matching

The correlation structure for one-to-one comparisons is characterized by the positive correlations between comparisons. For example, two comparisons are expected to be positively correlated if they share an enrollment template in common. Table B.1 defines three distinct types of dependency for single-eye mated comparison. The final column shows the strength of each type of dependency as the mean Pearson Correlation Coefficient (PCC) across all submissions over OPS-III. The correlations are measured with respect to the decisions made at an FMR of 10^{-4}. Although the correlation values are threshold dependent, they tend to change little for varying FMR due to the relative "flatness" of iris DET curves.

Correlation Type	Same Person	Same Eye	Same Verification Session	Same Enrollment Session	Correlation at FMR $= 10^{-4}$
1	Yes	Yes	No	Yes	0.50 ± 0.04
2	Yes	No	Yes	Yes	0.36 ± 0.06
3	Yes	No	No	Yes	0.18 ± 0.02

Table B.1: *A basic correlation structure for single-eye mated comparisons. The rows describe different types of dependency that can exist between comparisons. The values in the final column are mean correlation coefficients across all submissions (along with the standard deviations across submissions).*

The first type of dependency is the strongest and refers to comparisons that share a reference sample in common. The second refers to comparisons that share both capture sessions in common but compare different eyes. Finally, the third and weakest type of dependency refers to comparisons that share only the enrollment session in common but also compare different eyes. Despite sharing no actual iris samples in common, the comparisons are correlated because the sample quality of left and right iris images captured during the same session tend to be highly correlated.

The mated comparison sets for OPS-III were constructed by assigning the first chronological capture instance for each person as the reference sample, and all subsequent capture instances as probe samples. Thus, all mated comparisons for a particular person share the same enrollment session in common. This greatly simplifies the correlation structure to the point that Table B.1 fully captures the significant sources of dependency for mated comparisons. Different approaches to comparison set construction can lead to much more complicated correlation structures. For two-eye mated comparisons, the only source of dependency is when both comparisons involve the same person.

General equations are presented for estimating the variability of the core accuracy metrics given arbitrary correlation structures. Formally, let $\hat{p}(t)$ be the estimate of either FMR or FNMR as computed in Appendix A. Let $d_i(t)$ be the decision at threshold t for the ith comparison ($i = 1,...,N$). If the comparison is mated, then $d_i(t) = H(\upsilon_i - t)$. Furthermore, let S_k be the set of comparison pairs fulfilling the criteria for dependency type k ($k = 1,...,K$). The elements of S_k are pairs of indices where a given element (i, j) refers to comparisons i and j respectively. An unbiased estimate of the covariance for the kth dependency type is

$$\hat{\sigma}_k^2(t) = \left(\frac{N}{N-1}\right)\left(\frac{1}{|S_k|}\right) \sum_{(i,j)\,\in\, S_k} \left(d_i(t) - \hat{p}(t)\right)\left(d_j(t) - \hat{p}(t)\right). \tag{B.1}$$

The first term is Bessel's Correction. The rest of the equation is just the standard computation for covariance. The estimate

[1] Terminology such as "data subject" is now used by NIST to conform with ISO/IEC 2382-3: Vocabulary, Part 37: Biometrics [30]

of variance is

$$\hat{\sigma}^2(t) = \frac{\hat{p}(t)(1-\hat{p}(t))}{N} + \hat{c}(t) \tag{B.2}$$

where

$$\hat{c}(t) = \frac{1}{N^2}\sum_{i=1}^{K} |S_k| \hat{\sigma}_k^2(t). \tag{B.3}$$

Equation B.3 consolidates the contribution of all of the covariances to the overall estimate of the variance.

Confidence intervals can be constructed using estimates of both p and the σ^2. The simplest and most straightforward approach is to invoke the Central Limit Theorem (CLT) and define the interval as

$$\hat{p}(t) \pm z_{\alpha/2}\sqrt{\hat{\sigma}^2(t)} \tag{B.4}$$

where $z_{\alpha/2}$ is the $\alpha/2$th quantile of the standard normal distribution. However, Brown et. al [32] identify several problems with this approach. First, they note that always defining $\hat{p}(t)$ as the center of the interval can introduce a systematic negative bias to the coverage probability. Second, the actual distribution of $\hat{p}(t)$ is significantly nonnormal when p is close to 0 or 1, even for large N. Finally, due to the fact that $\hat{p}(t)$ is a discretized estimator, Equation B.4 severely underestimates the true coverage probability for certain "unlucky pairs" of p and N. For these reasons, we adopt their recommendation to use the Wilson Score method. However, the method must be modified to account for the correlation structure.

The Wilson Score interval is formed by inverting the normal approximation to the equal-tailed hypothesis test of $H_0 : p = p_0$. At significance level α, the hypothesis is accepted if $\hat{p}(t)$ falls within the interval

$$\frac{|\hat{p}(t) - p_0|}{\sqrt{\frac{1}{N}p_0(1-p_0) + \hat{c}(t)}} \leq \pm z_{\alpha/2}. \tag{B.5}$$

The denominator is the standard deviation of the test statistic. Unlike Equation B.4, it does not require a full estimate of the variance. The additional $\hat{c}(t)$ term incorporates the contribution of the correlation structure to the estimate. Since knowing p_0 does not reveal the true value of $c(t)$, the latter must be approximated using Equation B.3. The Wilson Interval is derived by regarding p_0 as the unknown parameter. Using the quadratic equation to solve Equation B.5 for p_0 yields the interval:

$$CI_W = \frac{\hat{p}(t) + \frac{1}{2N}z_{\alpha/2}^2 \pm z_{\alpha/2}\sqrt{\frac{1}{N}\hat{p}(t)(1-\hat{p}(t)) + \frac{1}{4N^2}z_{\alpha/2}^2 + (1+\frac{1}{N}z_{\alpha/2}^2)\hat{c}(t)}}{1+\frac{1}{N}z_{\alpha/2}^2}. \tag{B.6}$$

The Wilson Score interval still loses accuracy (though not as severely as Equation B.4) when np or $n(p-1)$ is small. For this reason, we conservatively opt not to apply the Wilson Score Interval to cases where $np < 10^3$.

B.2. One-to-Many Matching

The correlation structure for one-to-many comparisons is characterized by the positive correlations between searches. Each identification template is searched against a database of enrolled templates. The correlation structure has the potential to be much more complex compared to one-to-one comparisons. Dependences can exist across different enrollment databases as well as between templates enrolled within any one database. Worse, it is unclear how to measure and incorporate these dependencies into any estimates of variability. The simplest solution is to construct the enrollment databases such that these dependencies are never introduced. This is primarily accomplished by ensuring each database consists of an entirely disparate set of individuals. Additionally, no person should be represented by more than one entry in any database. So, for example, left and right eyes from the same person should not be enrolled as separate entries.

Correlation Type	Same Person	Same Eye	Same Verification Session	Same Enrollment Database	Correlation at FMR $= 10^{-3}$
1	Yes	Yes	No	Yes	0.458 ± 0.008
2	Yes	No	Yes	Yes	0.35 ± 0.03
3	Yes	No	No	Yes	0.23 ± 0.02
4	No	Yes	No	Yes	0.0 ± 0.1
5	No	No	No	Yes	0.0 ± 0.1

Table B.2: *A basic correlation structure for single-eye mated searches against an enrolled population of* 10000*. The rows describe different types of dependency that can exist between comparisons. The values in the final column are mean correlation coefficients across all submissions.*

Table B.2 describes five remaining sources of dependency for single-eye mated searches. All involve searches against the same enrollment database. The first refers to identification templates created from separate captures of the same eye. The second refers to identification templates created from opposite eyes captured during the same session. The third refers to identification templates created from opposite eyes from the same person acquired during different capture sessions. The fourth and fifth forms of dependency involve identification templates that are not expected to be correlated with each other but are both searched against the same database. Generally speaking, any two identification templates searched against the same enrollment database are expected to be correlated. Equations B.3 and B.6 are used along with the aforementioned dependency types to construct confidence intervals for FNIR (as well as FNMR).

For the case of two-eye comparisons, we consider two types of dependency when estimating confidence intervals. The first involves identification templates representing the same person searched against the same enrollment database. The second involves identification templates representing different people that are both searched against the same database.

B.3. Discussion and Further Considerations

Previous NIST evaluations [1, 33] used the Wilson Score Method under the assumption that all comparisons are independent. Failing to account for the dependencies probably led to overly optimistic estimates of variability. In the current evaluation, we found that the independence assumption leads to significant underestimates of variance, sometimes by a factor of 2 or 3. Many academic iris datasets (*e.g.* CASIA [34], Notre Dame 0405 [35]) consist of iris samples collected from comparatively small numbers of subjects, typically a few hundred at most. Thus, the dependencies are expected to contribute even more toward the variability of accuracy statistics computed over these datasets.

Mansfield *et. al* [36] provide estimates of variability for the false match rate given a particular sampling strategy. Schuckers [37] expands upon their work by defining a general correlation structure for fingerprint recognition. Although his proposed method of estimating confidence intervals is common and asymptotically valid, it still suffers from the same weaknesses identified by Brown *et. al.* in relation to Equation B.4. Bootstrapping also fails to offer a viable alternative because it cannot be generalized to work for arbitrary correlation structures. Altering the resampling strategy can perhaps compensate for one or two types of dependency [36]. Beyond that, the problem becomes too complex. Wayman [14] tests the accuracy of uncertainty bounds calculated using equations defined by Bickel [38] and finds them to be accurate when full cross comparisons are available.

Sometimes we report estimates of variability for FNMR at fixed FMR when in fact the decision threshold is fixed. Uncertainty with respect to what decision threshold corresponds to the targeted FMR results in increased uncertainty about the true value of FNMR. That said, our estimates of FMR are expected to be very tight given the large number of nonmated comparisons performed (often in excess of a billion). Additionally, even at very low FMRs, the lightly sloping nature of iris DET curves means that small discrepancies in FMR are not expected to significantly impact FNMR. Similar logic holds for estimating FNIR at fixed FPIR.

C Removing Ground Truth Errors from OPS-III

In an ideal world, every iris sample would be assigned the correct person identifier. In reality, we find this is rarely ever the case (possibly due to clerical / human error or some mistake in data handling). If not addressed, these ground truth errors will inflate estimates of FNMR and FMR. If two samples of the same iris are falsely labeled as different irides, then correctly matching them will be counted as a false match. To address this, probe samples were horizontally flipped prior to template creation when the comparison was nonmated. The flipping converts the probe sample into a mirror image itself, so even comparisons against the same iris are less likely to produce low measures of dissimilarity. This strategy was used in both IREX III and IREX IV (see IREX III Section 6.4 for a detailed explanation and analysis). Some have noted that Purkinje images (i.e. reflections of objects on the eye) might introduce false similarities between iris samples if the reflections are similar, and that horizontally flipping would cancel out this effect. Although horizontal flipping may not be a perfect solution, we believe it is preferable to ignoring this type of ground truth error.

Another type of ground truth error is when two samples of different irides are assigned the same person identifier. This can inflate estimates of FNMR (or FNIR in the one-to-many case). We attempted to identify these labeling errors and filter them out of the comparison sets. We began by identifying all two-eye mated comparisons that were 'missed' by all 46 matchers at an FMR of 10^{-4}. Figure C.1 plots how frequently mated comparisons were missed by a specific number of matchers. So, for example, the figure shows that 6100 comparisons were missed by exactly 3 matchers. Two-hundred fifty seven comparisons were missed by all 46 matchers. We then manually inspected the face images that were captured concurrently with the iris samples to verify whether the comparison is truly mated. Two-hundred twenty one of the comparisons were determined to be nonmated and were removed from the comparison sets. A comparison was only removed if it was obvious that the comparison is not mated. The remaining 36 comparisons that were not removed generally involve extremely poor quality iris samples (closed eyes, patterned contact lenses, etc.). Our reason for considering only those comparisons missed by all matchers was to avoid matcher-specific bias.

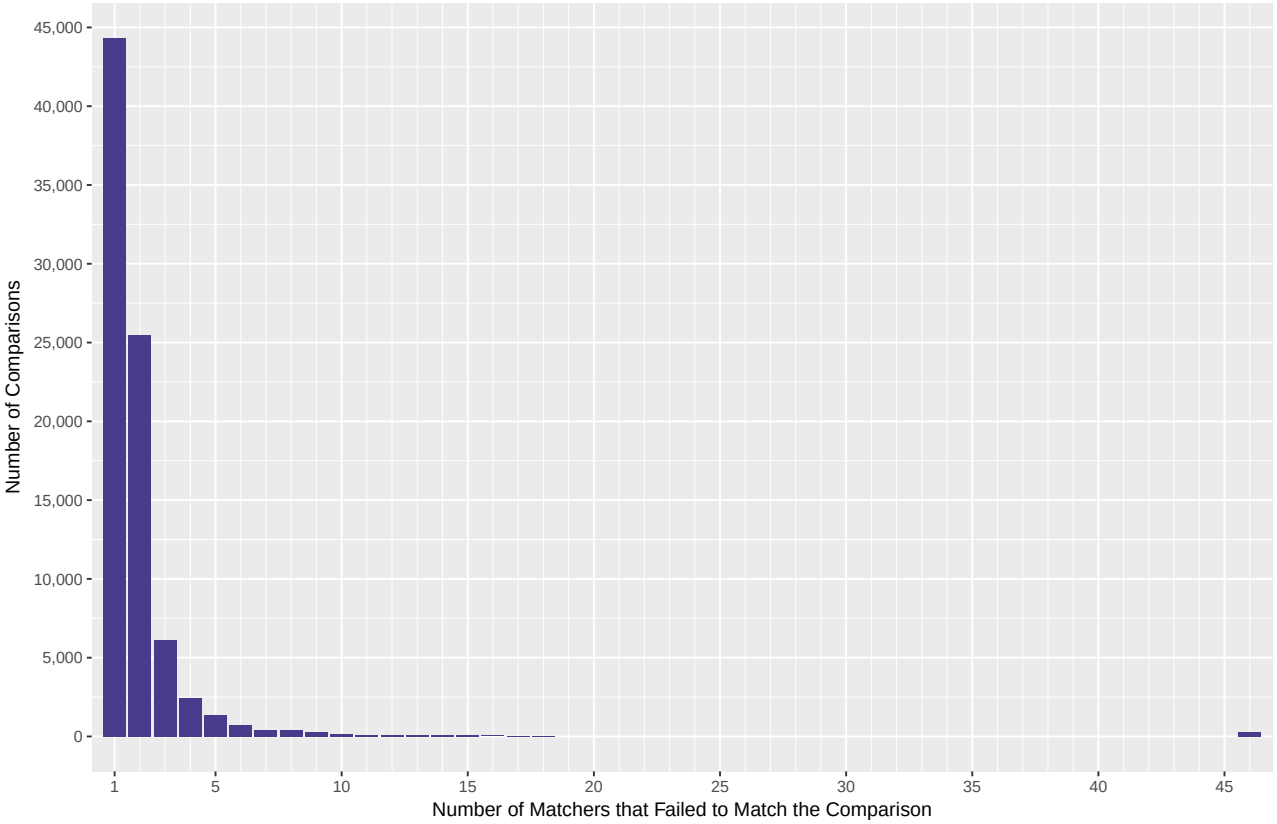

Figure C.1: Histogram showing how frequently comparisons were missed by a specific number of matchers. Note the short bar on the far right representing the 257 comparisons missed by all 46 matchers.

D Additional Figures and Tables

Appendix A contains supplementary DET plots and additional summary statistics for all recognition algorithms.

D.1. One-to-One

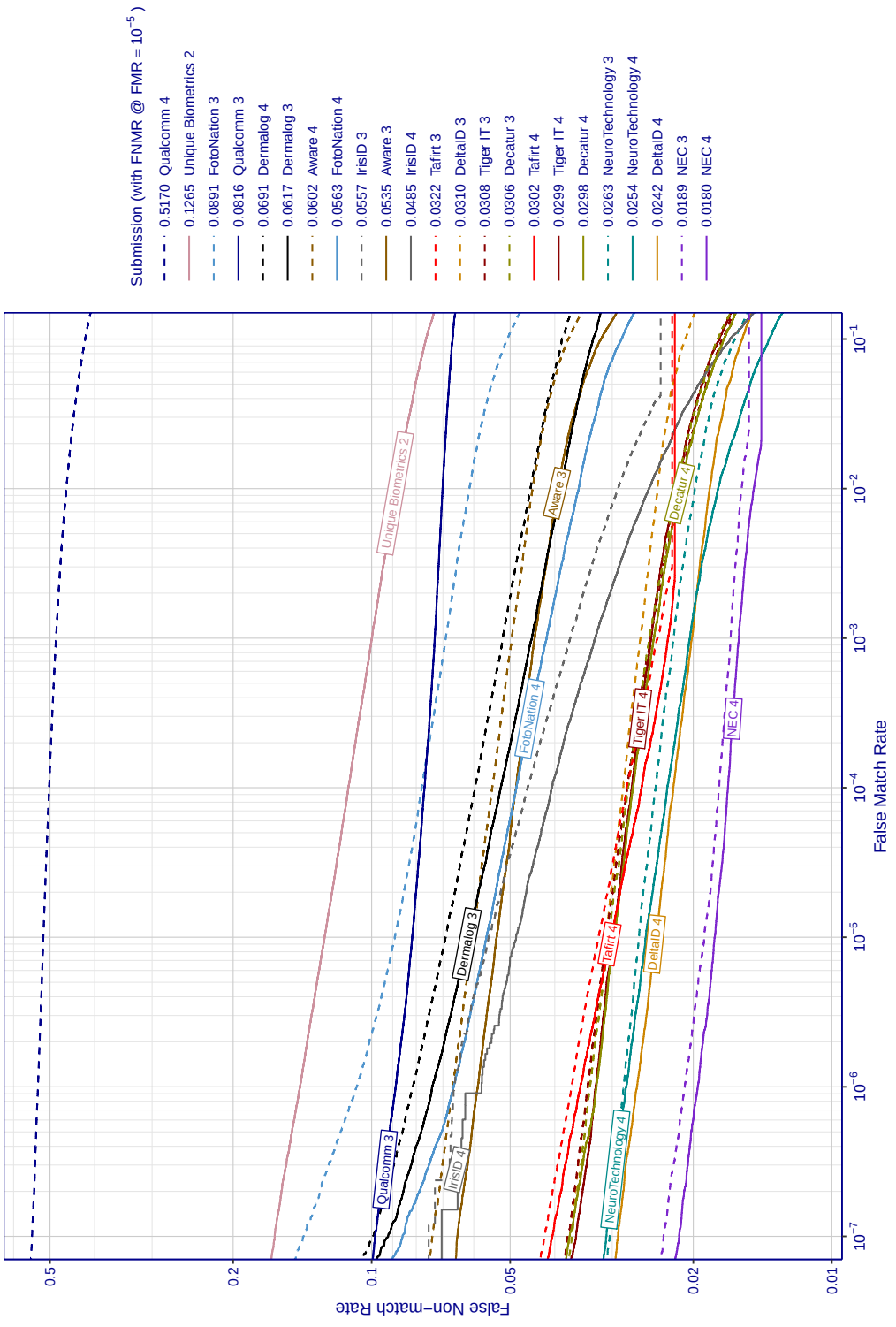

Figure D.1: DET curves for (one-to-one) ***single-eye comparisons***. Only ***Phase 2 submissions*** are shown. Plots were generated from 166 thousand mated and 1 billion nonmated comparisons.

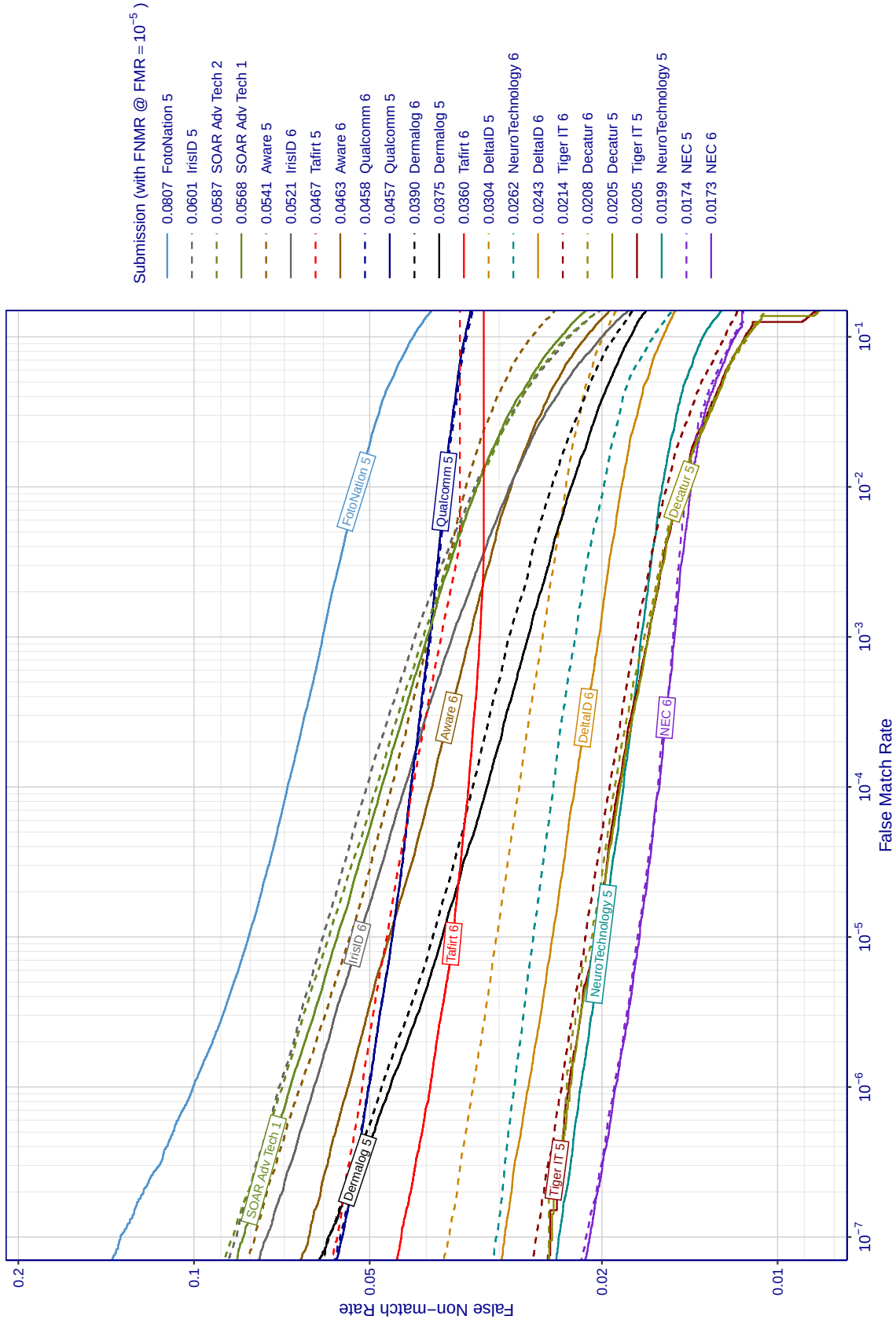

Figure D.2: DET curves for (one-to-one) ***single-eye comparisons***. Only ***Phase 3 submissions*** are shown. Plots were generated from 166 thousand mated and 1 billion nonmated comparisons.

Submission	FMR=10^{-2}	FMR=10^{-4}	FMR=10^{-6}
NEC 6	0.0141 ± 0.0009	0.016 ± 0.001	0.019 ± 0.001
NEC 5	0.0144 ± 0.0009	0.016 ± 0.001	0.019 ± 0.001
NEC 4	0.0148 ± 0.0007	0.0168 ± 0.0007	0.0196 ± 0.0008
NEC 3	0.0157 ± 0.0006	0.0176 ± 0.0007	0.0209 ± 0.0007
NeuroTechnology 5	0.0158 ± 0.0006	0.0184 ± 0.0007	0.0218 ± 0.0007
Decatur 5	0.0146 ± 0.0006	0.0186 ± 0.0007	0.0227 ± 0.0007
Tiger IT 5	0.015 ± 0.001	0.019 ± 0.001	0.023 ± 0.001
Decatur 6	0.0147 ± 0.0008	0.0190 ± 0.0009	0.023 ± 0.001
Tiger IT 6	0.0155 ± 0.0006	0.0194 ± 0.0007	0.0234 ± 0.0008
DeltaID 4	0.0184 ± 0.0006	0.0221 ± 0.0007	0.0265 ± 0.0008
DeltaID 6	0.0185 ± 0.0007	0.0223 ± 0.0009	0.027 ± 0.001
NeuroTechnology 4	0.0177 ± 0.0007	0.0228 ± 0.0009	0.028 ± 0.001
NeuroTechnology 6	0.0199 ± 0.0009	0.024 ± 0.001	0.028 ± 0.001
NeuroTechnology 3	0.0198 ± 0.0009	0.024 ± 0.001	0.028 ± 0.001
Decatur 4	0.0213 ± 0.0006	0.0271 ± 0.0006	0.0325 ± 0.0006
Tiger IT 4	0.0218 ± 0.0006	0.0272 ± 0.0006	0.0326 ± 0.0006
DeltaID 5	0.023 ± 0.001	0.028 ± 0.001	0.033 ± 0.001
Decatur 3	0.022 ± 0.002	0.028 ± 0.002	0.034 ± 0.002
Tiger IT 3	0.0214 ± 0.0006	0.0278 ± 0.0007	0.0339 ± 0.0008
DeltaID 3	0.0239 ± 0.0006	0.0284 ± 0.0007	0.0339 ± 0.0008
Tafirt 4	0.0219 ± 0.0004	0.0259 ± 0.0007	0.0352 ± 0.0008
Tafirt 3	0.0222 ± 0.0004	0.0279 ± 0.0007	0.0368 ± 0.0008
Tafirt 6	0.032 ± 0.001	0.034 ± 0.001	0.039 ± 0.002
Dermalog 5	0.0226 ± 0.0008	0.0315 ± 0.0009	0.046 ± 0.001
Dermalog 6	0.0244 ± 0.0007	0.0333 ± 0.0008	0.048 ± 0.001
Qualcomm 5	0.0372 ± 0.0007	0.0425 ± 0.0007	0.0502 ± 0.0008
Qualcomm 6	0.0370 ± 0.0006	0.0424 ± 0.0007	0.0503 ± 0.0007
Tafirt 5	0.0350 ± 0.0006	0.0422 ± 0.0006	0.0515 ± 0.0007
Aware 6	0.0286 ± 0.0006	0.0391 ± 0.0007	0.0544 ± 0.0007
Aware 3	0.039 ± 0.001	0.049 ± 0.001	0.059 ± 0.001
IrisID 4	0.0250 ± 0.0006	0.0406 ± 0.0006	0.0596 ± 0.0007
IrisID 6	0.0288 ± 0.0006	0.0435 ± 0.0006	0.0623 ± 0.0007
Aware 5	0.0342 ± 0.0008	0.046 ± 0.001	0.066 ± 0.001
FotoNation 4	0.0355 ± 0.0008	0.0485 ± 0.0009	0.066 ± 0.001
IrisID 3	0.0291 ± 0.0007	0.0458 ± 0.0008	0.0664 ± 0.0009
Aware 4	0.0447 ± 0.0007	0.0547 ± 0.0008	0.0667 ± 0.0009
SOAR Adv Tech 1	0.0327 ± 0.0006	0.0477 ± 0.0006	0.0684 ± 0.0006
SOAR Adv Tech 2	0.0327 ± 0.0006	0.0486 ± 0.0006	0.0706 ± 0.0006
IrisID 5	0.0330 ± 0.0009	0.0506 ± 0.0009	0.071 ± 0.001
Dermalog 3	0.0388 ± 0.0009	0.0524 ± 0.0009	0.074 ± 0.001
Dermalog 4	0.0452 ± 0.0006	0.0596 ± 0.0006	0.0814 ± 0.0007
Qualcomm 3	0.0701 ± 0.0006	0.0768 ± 0.0006	0.0887 ± 0.0007
FotoNation 5	0.0524 ± 0.0008	0.0692 ± 0.0009	0.100 ± 0.001
FotoNation 3	0.0614 ± 0.0008	0.0778 ± 0.0009	0.108 ± 0.001
Unique Biometrics 2	0.0886 ± 0.0006	0.1124 ± 0.0009	0.143 ± 0.001
Qualcomm 4	0.4607 ± 0.0005	0.5022 ± 0.0008	0.5311 ± 0.0009

Table D.1: Accuracy table for (one-to-one) single-eye matching. Standard deviations are presented after the plus/minus.

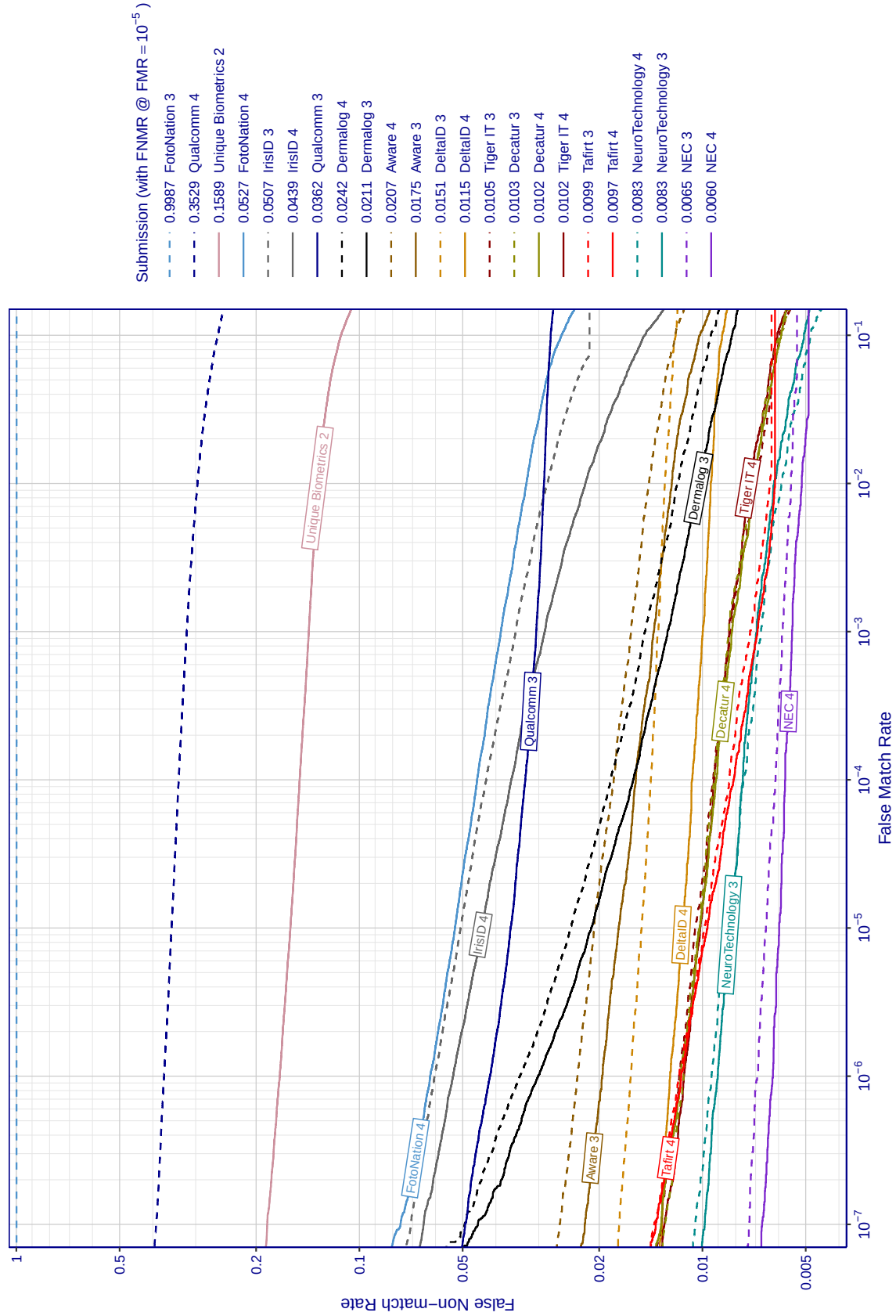

Figure D.3: DET curves for one-to-one ***two-eye comparisons***. Only ***Phase 2 submissions*** are shown. Plots were generated from 83 thousand mated and 500 million nonmated comparisons.

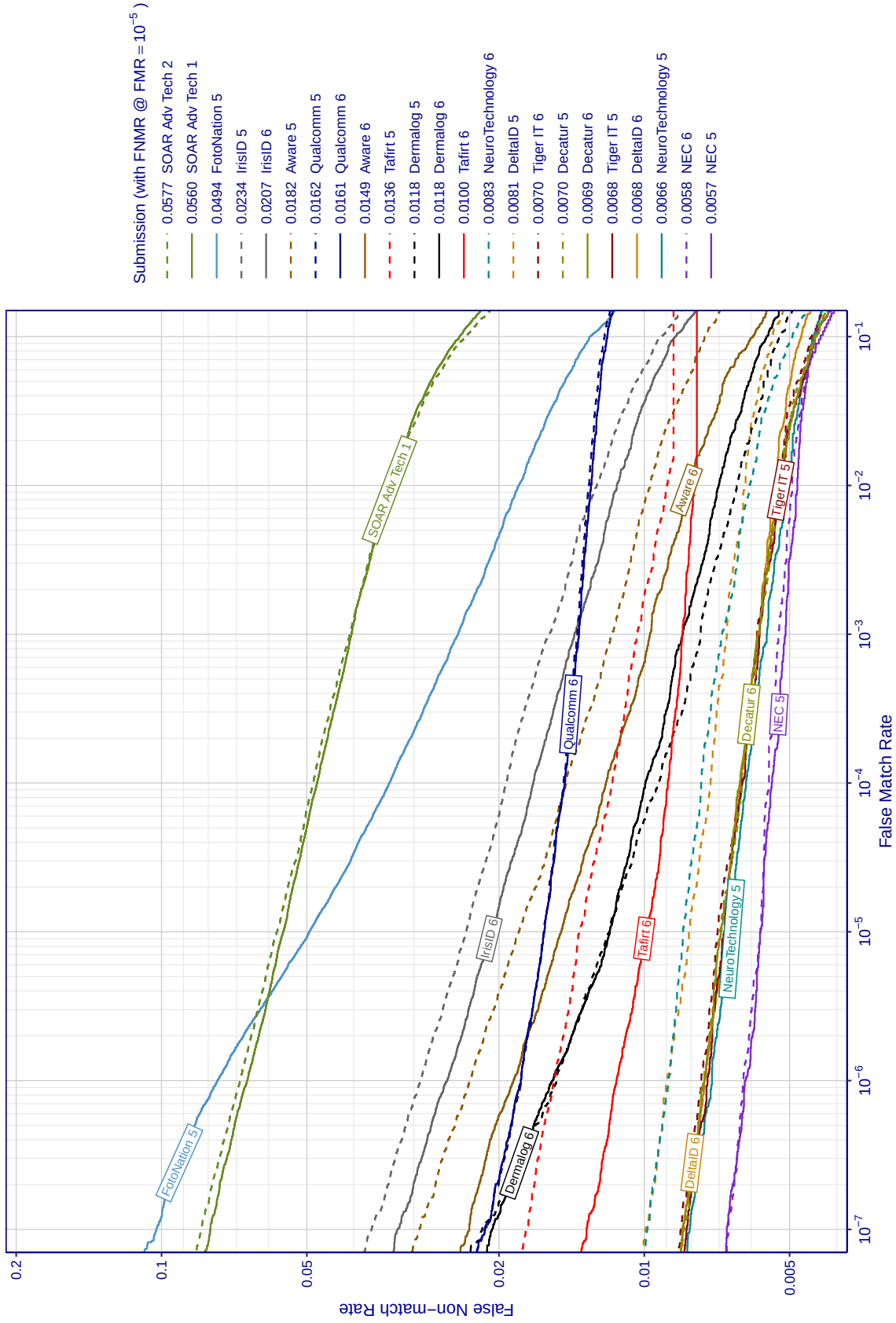

Figure D.4: DET curves for one-to-one ***two-eye comparisons***. Only ***Phase 3 submissions*** are shown. Plots were generated from 83 thousand mated and 500 million nonmated comparisons.

Submission	FMR=10^{-2}	FMR=10^{-4}	FMR=10^{-6}
NEC 5	0.0048 ± 0.0008	0.0054 ± 0.0009	0.006 ± 0.001
NEC 4	0.0051 ± 0.0009	0.006 ± 0.001	0.006 ± 0.001
NEC 6	0.0049 ± 0.0007	0.0055 ± 0.0008	0.0062 ± 0.0008
NEC 3	0.0055 ± 0.0006	0.0061 ± 0.0007	0.0069 ± 0.0007
NeuroTechnology 5	0.0051 ± 0.0006	0.0062 ± 0.0006	0.0072 ± 0.0007
Tiger IT 5	0.0052 ± 0.0006	0.0063 ± 0.0006	0.0074 ± 0.0007
DeltaID 6	0.0054 ± 0.0008	0.0063 ± 0.0009	0.008 ± 0.001
Decatur 5	0.0052 ± 0.0009	0.006 ± 0.001	0.008 ± 0.001
Decatur 6	0.0052 ± 0.0006	0.0063 ± 0.0007	0.0075 ± 0.0007
Tiger IT 6	0.0053 ± 0.0006	0.0063 ± 0.0007	0.0076 ± 0.0007
NeuroTechnology 6	0.0061 ± 0.0008	0.008 ± 0.001	0.009 ± 0.001
DeltaID 5	0.0062 ± 0.0007	0.007 ± 0.001	0.009 ± 0.001
NeuroTechnology 3	0.0062 ± 0.0008	0.008 ± 0.001	0.009 ± 0.001
NeuroTechnology 4	0.0059 ± 0.0008	0.008 ± 0.001	0.009 ± 0.001
Tiger IT 4	0.0073 ± 0.0005	0.0093 ± 0.0006	0.0113 ± 0.0006
Decatur 3	0.0071 ± 0.0005	0.0092 ± 0.0006	0.0113 ± 0.0006
Tafirt 6	0.008 ± 0.001	0.009 ± 0.001	0.011 ± 0.001
Tafirt 4	0.006 ± 0.003	0.008 ± 0.003	0.012 ± 0.003
Decatur 4	0.0071 ± 0.0006	0.0091 ± 0.0007	0.0116 ± 0.0007
Tiger IT 3	0.0071 ± 0.0006	0.0093 ± 0.0007	0.0117 ± 0.0007
Tafirt 3	0.0064 ± 0.0005	0.0083 ± 0.0007	0.0117 ± 0.0008
DeltaID 4	0.0095 ± 0.0005	0.0107 ± 0.0007	0.0124 ± 0.0008
Dermalog 5	0.006 ± 0.001	0.009 ± 0.001	0.015 ± 0.002
Tafirt 5	0.0090 ± 0.0007	0.0117 ± 0.0008	0.015 ± 0.001
Dermalog 6	0.0070 ± 0.0007	0.0100 ± 0.0008	0.0155 ± 0.0009
DeltaID 3	0.0129 ± 0.0006	0.0142 ± 0.0007	0.0162 ± 0.0007
Qualcomm 5	0.0130 ± 0.0006	0.0147 ± 0.0006	0.0180 ± 0.0007
Qualcomm 6	0.0129 ± 0.0005	0.0146 ± 0.0006	0.0180 ± 0.0006
Aware 6	0.0081 ± 0.0006	0.0120 ± 0.0006	0.0187 ± 0.0007
Aware 3	0.0123 ± 0.0009	0.016 ± 0.001	0.020 ± 0.002
Aware 5	0.0097 ± 0.0005	0.0145 ± 0.0006	0.0230 ± 0.0006
Aware 4	0.0143 ± 0.0005	0.0181 ± 0.0006	0.0231 ± 0.0006
IrisID 6	0.0114 ± 0.0007	0.0170 ± 0.0009	0.026 ± 0.001
IrisID 5	0.0125 ± 0.0007	0.0193 ± 0.0009	0.029 ± 0.001
Dermalog 3	0.0102 ± 0.0006	0.0159 ± 0.0007	0.0302 ± 0.0009
Dermalog 4	0.0118 ± 0.0006	0.0185 ± 0.0008	0.0341 ± 0.0009
Qualcomm 3	0.0290 ± 0.0005	0.0328 ± 0.0005	0.0416 ± 0.0006
IrisID 4	0.0219 ± 0.0005	0.0361 ± 0.0005	0.0533 ± 0.0006
IrisID 3	0.0283 ± 0.0008	0.0427 ± 0.0008	0.0604 ± 0.0009
FotoNation 4	0.0326 ± 0.0008	0.0454 ± 0.0008	0.0622 ± 0.0009
SOAR Adv Tech 1	0.0335 ± 0.0005	0.0476 ± 0.0006	0.0665 ± 0.0006
SOAR Adv Tech 2	0.0335 ± 0.0005	0.0486 ± 0.0006	0.0688 ± 0.0006
FotoNation 5	0.018 ± 0.001	0.034 ± 0.001	0.077 ± 0.001
Unique Biometrics 2	0.132 ± 0.001	0.149 ± 0.001	0.171 ± 0.001
Qualcomm 4	0.2965 ± 0.0007	0.3363 ± 0.0008	0.3709 ± 0.0009
FotoNation 3	0.9983 ± 0.0006	0.9985 ± 0.0007	0.9988 ± 0.0008

Table D.2: Accuracy table for (one-to-one) two-eye matching. Standard deviations are presented after the plus/minus.

D.2. One-to-Many

D.3. 10K Enrollment Size

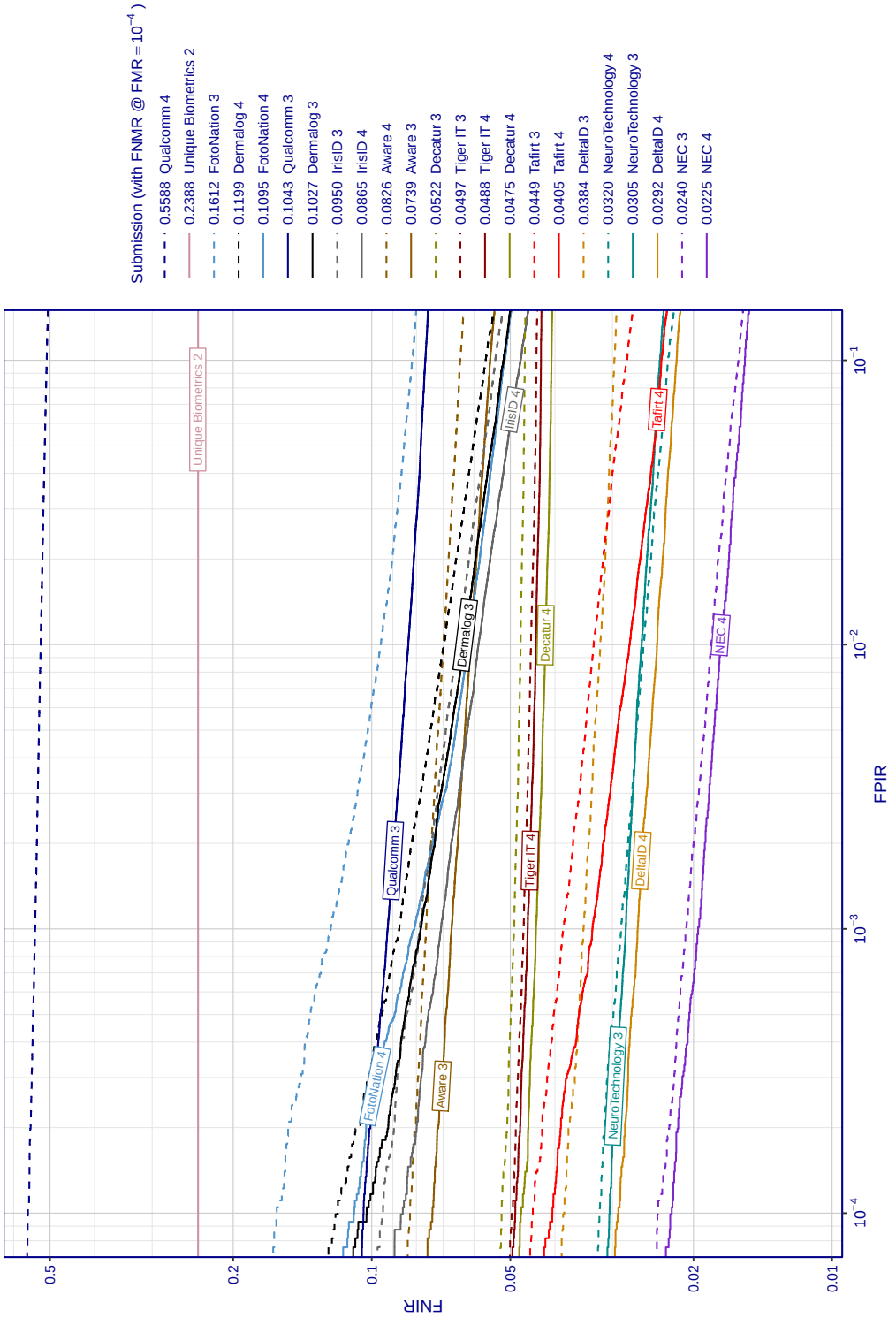

Figure D.5: DET curves for ***single-eye comparisons*** against a database of ***10 thousand enrolled individuals***. Only ***Phase 2 submissions*** are shown. Plots were generated from 166 thousand mated and 171 thousand nonmated searches.

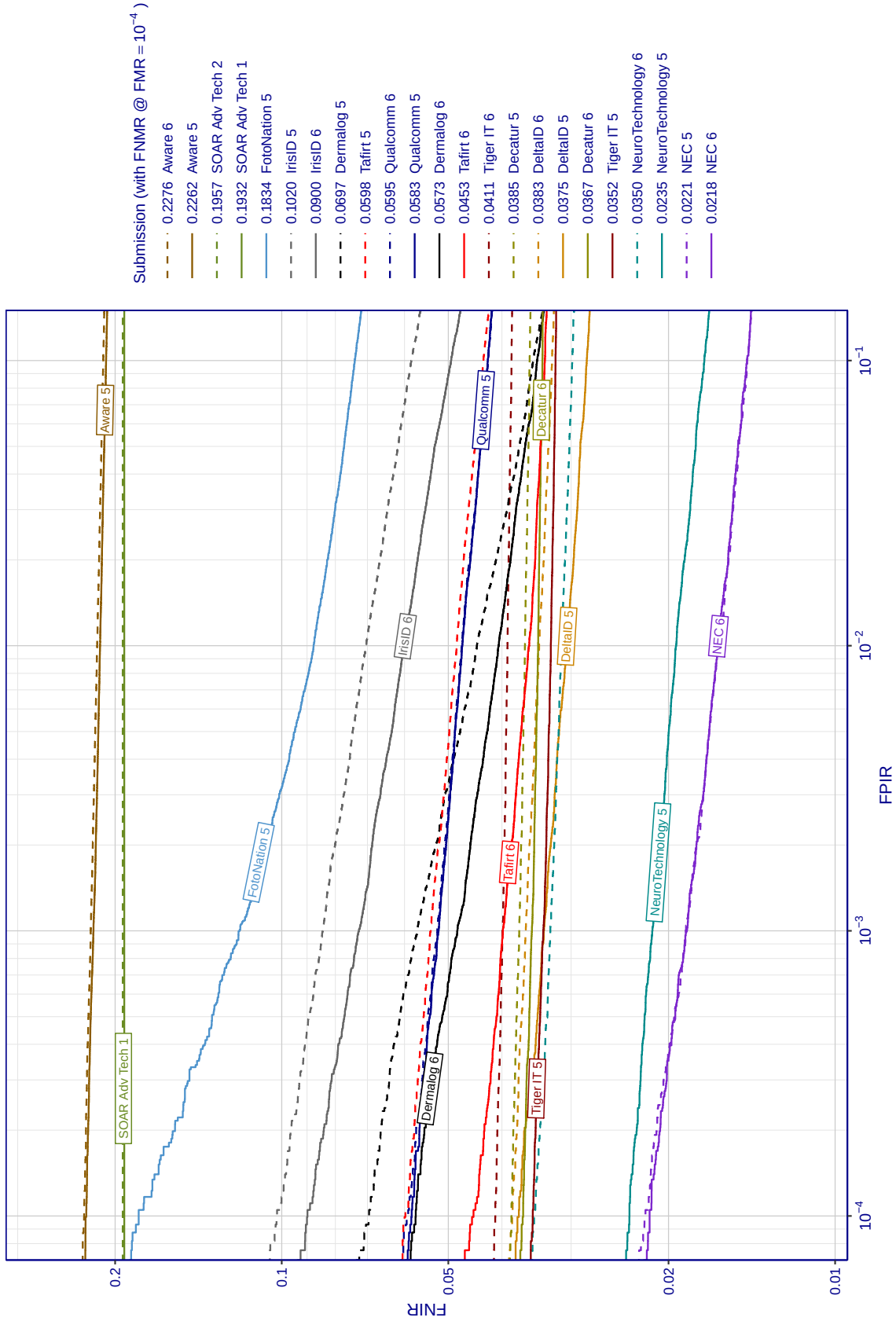

Figure D.6: DET curves for ***single-eye comparisons*** against a database of ***10 thousand enrolled individuals.*** Only ***Phase 3 submissions*** are shown. Plots were generated from 166 thousand mated and 171 thousand nonmated searches.

Submission	FPIR=10^{-2}	FPIR=10^{-3}	FPIR=10^{-4}
NEC 6	0.016 ± 0.001	0.019 ± 0.001	0.022 ± 0.002
NEC 5	0.016 ± 0.001	0.019 ± 0.001	0.022 ± 0.002
NEC 4	0.018 ± 0.001	0.020 ± 0.001	0.022 ± 0.001
NeuroTechnology 5	0.0194 ± 0.0009	0.0214 ± 0.0009	0.0235 ± 0.0009
NEC 3	0.0183 ± 0.0009	0.0207 ± 0.0009	0.024 ± 0.001
DeltaID 4	0.0240 ± 0.0009	0.026 ± 0.001	0.029 ± 0.001
NeuroTechnology 3	0.026 ± 0.002	0.028 ± 0.002	0.030 ± 0.002
NeuroTechnology 4	0.026 ± 0.001	0.029 ± 0.002	0.032 ± 0.002
NeuroTechnology 6	0.031 ± 0.001	0.033 ± 0.001	0.035 ± 0.001
Tiger IT 5	0.033 ± 0.001	0.034 ± 0.001	0.035 ± 0.001
Decatur 6	0.034 ± 0.001	0.035 ± 0.002	0.037 ± 0.002
DeltaID 5	0.031 ± 0.001	0.034 ± 0.001	0.038 ± 0.002
DeltaID 6	0.034 ± 0.001	0.036 ± 0.002	0.038 ± 0.002
DeltaID 3	0.032 ± 0.001	0.035 ± 0.002	0.038 ± 0.002
Decatur 5	0.0364 ± 0.0008	0.0372 ± 0.0008	0.0385 ± 0.0009
Tafirt 4	0.0275 ± 0.0008	0.0332 ± 0.0008	0.0405 ± 0.0009
Tiger IT 6	0.039 ± 0.002	0.040 ± 0.002	0.041 ± 0.002
Tafirt 3	0.033 ± 0.003	0.038 ± 0.003	0.045 ± 0.003
Tafirt 6	0.036 ± 0.001	0.040 ± 0.001	0.045 ± 0.001
Decatur 4	0.042 ± 0.001	0.044 ± 0.001	0.047 ± 0.001
Tiger IT 4	0.044 ± 0.001	0.046 ± 0.001	0.049 ± 0.001
Tiger IT 3	0.0451 ± 0.0009	0.047 ± 0.001	0.050 ± 0.001
Decatur 3	0.047 ± 0.001	0.049 ± 0.001	0.052 ± 0.001
Dermalog 6	0.041 ± 0.003	0.048 ± 0.003	0.057 ± 0.003
Qualcomm 5	0.047 ± 0.003	0.052 ± 0.003	0.058 ± 0.003
Qualcomm 6	0.0473 ± 0.0009	0.052 ± 0.001	0.059 ± 0.001
Tafirt 5	0.048 ± 0.001	0.054 ± 0.001	0.060 ± 0.001
Dermalog 5	0.0444 ± 0.0008	0.0569 ± 0.0009	0.0697 ± 0.0009
Aware 3	0.061 ± 0.001	0.067 ± 0.001	0.074 ± 0.001
Aware 4	0.070 ± 0.002	0.076 ± 0.002	0.083 ± 0.002
IrisID 4	0.0587 ± 0.0009	0.0708 ± 0.0009	0.086 ± 0.001
IrisID 6	0.060 ± 0.001	0.073 ± 0.001	0.090 ± 0.001
IrisID 3	0.065 ± 0.001	0.079 ± 0.002	0.095 ± 0.002
IrisID 5	0.070 ± 0.001	0.084 ± 0.002	0.102 ± 0.002
Dermalog 3	0.063 ± 0.001	0.078 ± 0.001	0.103 ± 0.001
Qualcomm 3	0.083 ± 0.001	0.092 ± 0.001	0.104 ± 0.001
FotoNation 4	0.0619 ± 0.0007	0.0809 ± 0.0008	0.1095 ± 0.0008
Dermalog 4	0.0696 ± 0.0007	0.0872 ± 0.0008	0.1199 ± 0.0008
FotoNation 3	0.096 ± 0.001	0.123 ± 0.001	0.161 ± 0.001
FotoNation 5	0.088 ± 0.001	0.120 ± 0.001	0.183 ± 0.001
SOAR Adv Tech 1	0.193 ± 0.001	0.193 ± 0.001	0.193 ± 0.001
SOAR Adv Tech 2	0.1957 ± 0.0009	0.196 ± 0.001	0.196 ± 0.001
Aware 5	0.213 ± 0.001	0.219 ± 0.001	0.226 ± 0.001
Aware 6	0.216 ± 0.001	0.221 ± 0.001	0.228 ± 0.001
Unique Biometrics 2	0.239 ± 0.001	0.239 ± 0.001	0.239 ± 0.001
Qualcomm 4	0.524 ± 0.001	0.539 ± 0.001	0.559 ± 0.001

Table D.3: Accuracy table for single-eye matching against an enrolled population of 10 thousand. Standard deviations are presented after the plus/minus.

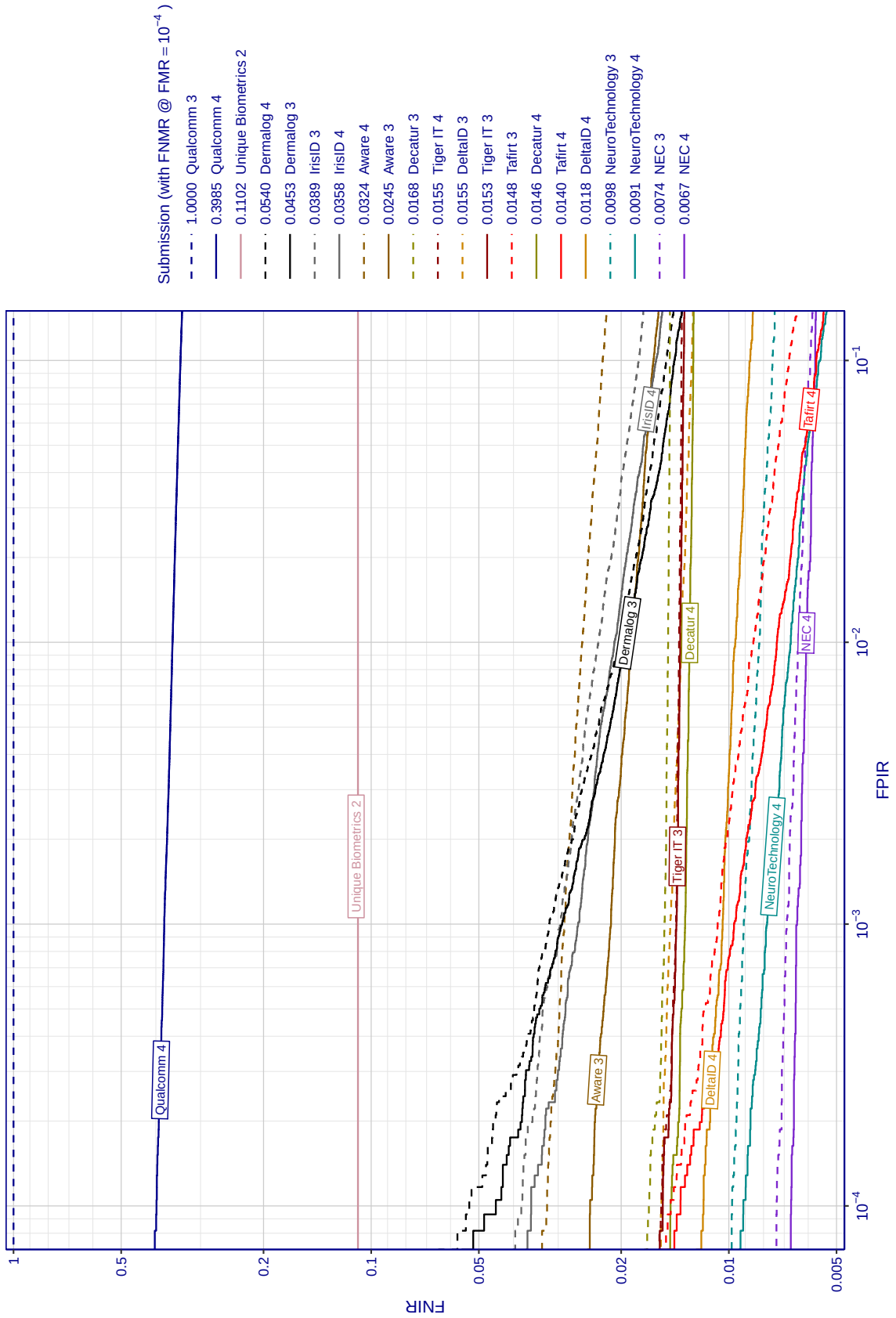

Figure D.7: DET curves for ***two-eye comparisons*** against a database of ***10 thousand enrolled individuals***. Only ***Phase 2 submissions*** are shown. Plots were generated from 166 thousand mated and 171 thousand nonmated searches.

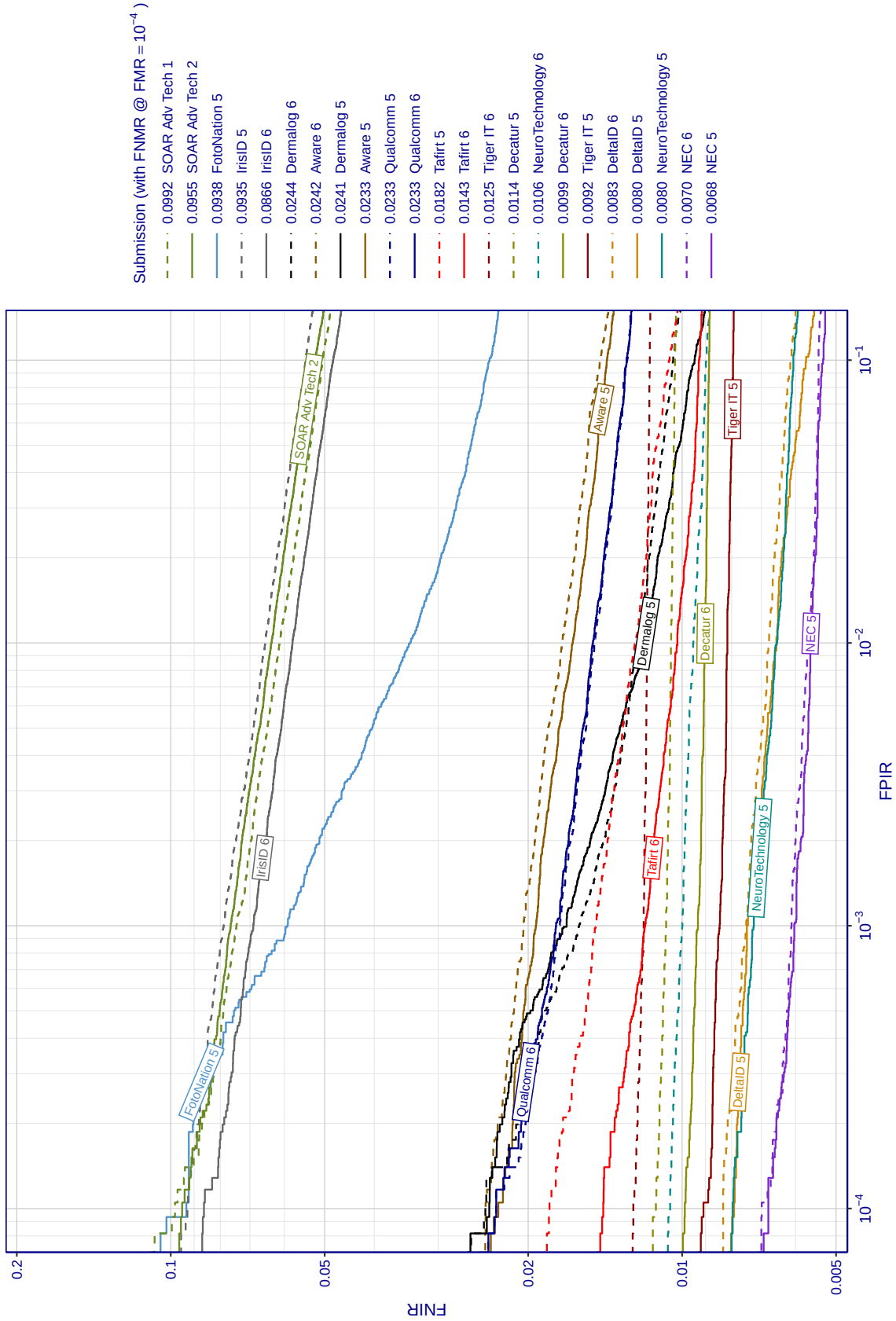

Figure D.8: DET curves for ***two-eye*** comparisons against a database of ***10 thousand enrolled individuals.*** Only ***Phase 3 submissions*** are shown. Plots were generated from 83 thousand mated and 86 thousand nonmated searches.

Submission	FPIR=10^{-2}	FPIR=10^{-3}	FPIR=10^{-4}
NEC 4	0.006 ± 0.001	0.006 ± 0.001	0.007 ± 0.001
NEC 5	0.006 ± 0.001	0.006 ± 0.001	0.007 ± 0.001
NEC 6	0.0057 ± 0.0008	0.0061 ± 0.0008	0.0070 ± 0.0008
NEC 3	0.0064 ± 0.0007	0.0069 ± 0.0007	0.0074 ± 0.0008
NeuroTechnology 5	0.0065 ± 0.0007	0.0073 ± 0.0007	0.0080 ± 0.0007
DeltaID 5	0.0065 ± 0.0007	0.0075 ± 0.0007	0.0080 ± 0.0007
DeltaID 6	0.0068 ± 0.0008	0.0075 ± 0.0009	0.0083 ± 0.0009
NeuroTechnology 4	0.0067 ± 0.0008	0.0078 ± 0.0009	0.0091 ± 0.0009
Tiger IT 5	0.008 ± 0.001	0.008 ± 0.001	0.009 ± 0.001
NeuroTechnology 3	0.008 ± 0.001	0.009 ± 0.001	0.010 ± 0.001
Decatur 6	0.009 ± 0.001	0.009 ± 0.001	0.010 ± 0.002
NeuroTechnology 6	0.010 ± 0.001	0.010 ± 0.001	0.011 ± 0.002
Decatur 5	0.0105 ± 0.0006	0.0107 ± 0.0006	0.0114 ± 0.0006
DeltaID 4	0.0096 ± 0.0006	0.0104 ± 0.0006	0.0118 ± 0.0006
Tiger IT 6	0.012 ± 0.003	0.012 ± 0.003	0.012 ± 0.003
Tafirt 4	0.0073 ± 0.0009	0.0097 ± 0.0009	0.0140 ± 0.0009
Tafirt 6	0.0102 ± 0.0008	0.0118 ± 0.0008	0.0143 ± 0.0009
Decatur 4	0.0129 ± 0.0007	0.0133 ± 0.0008	0.0146 ± 0.0009
Tafirt 3	0.0085 ± 0.0007	0.0108 ± 0.0007	0.0148 ± 0.0008
Tiger IT 3	0.014 ± 0.001	0.014 ± 0.001	0.015 ± 0.001
DeltaID 3	0.014 ± 0.002	0.015 ± 0.002	0.015 ± 0.002
Tiger IT 4	0.014 ± 0.002	0.014 ± 0.002	0.016 ± 0.002
Decatur 3	0.0148 ± 0.0007	0.0151 ± 0.0007	0.0168 ± 0.0008
Tafirt 5	0.0124 ± 0.0007	0.0148 ± 0.0008	0.0182 ± 0.0008
Qualcomm 6	0.0149 ± 0.0006	0.0177 ± 0.0006	0.0233 ± 0.0006
Qualcomm 5	0.0148 ± 0.0007	0.0173 ± 0.0007	0.0233 ± 0.0008
Aware 5	0.016 ± 0.001	0.019 ± 0.001	0.023 ± 0.002
Dermalog 5	0.0119 ± 0.0007	0.0170 ± 0.0007	0.0241 ± 0.0007
Aware 6	0.0172 ± 0.0008	0.0204 ± 0.0008	0.0242 ± 0.0008
Dermalog 6	0.012 ± 0.001	0.016 ± 0.001	0.024 ± 0.002
Aware 3	0.019 ± 0.001	0.021 ± 0.002	0.024 ± 0.002
Aware 4	0.0257 ± 0.0009	0.029 ± 0.001	0.032 ± 0.001
IrisID 4	0.0209 ± 0.0008	0.0264 ± 0.0009	0.036 ± 0.001
IrisID 3	0.0232 ± 0.0006	0.0297 ± 0.0006	0.0389 ± 0.0006
Dermalog 3	0.0193 ± 0.0006	0.0294 ± 0.0006	0.0453 ± 0.0006
Dermalog 4	0.0202 ± 0.0009	0.0318 ± 0.0009	0.054 ± 0.001
IrisID 6	0.0578 ± 0.0009	0.0690 ± 0.0009	0.087 ± 0.001
IrisID 5	0.0653 ± 0.0008	0.0789 ± 0.0008	0.0935 ± 0.0008
FotoNation 5	0.0342 ± 0.0007	0.0596 ± 0.0008	0.0938 ± 0.0008
SOAR Adv Tech 2	0.064 ± 0.001	0.076 ± 0.001	0.096 ± 0.001
SOAR Adv Tech 1	0.061 ± 0.001	0.074 ± 0.001	0.099 ± 0.001
Unique Biometrics 2	0.1102 ± 0.0009	0.1102 ± 0.0009	0.110 ± 0.001
Qualcomm 4	0.3593 ± 0.0008	0.3790 ± 0.0008	0.3985 ± 0.0008

Table D.4: Accuracy table for two-eye matching against an enrolled population of 10 thousand. Standard deviations are presented after the plus/minus.

D.4. 160K Enrollment Size

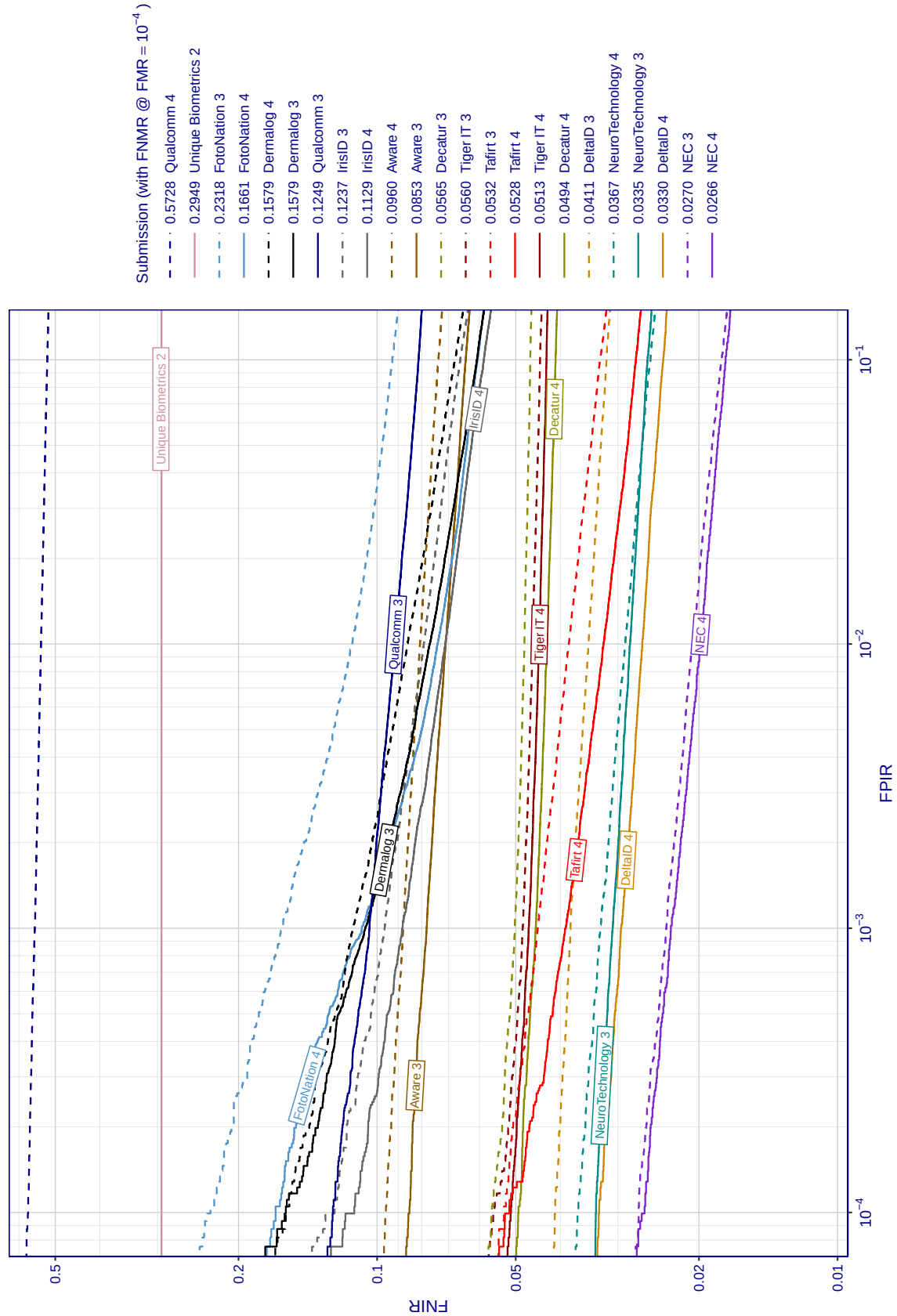

Figure D.9: DET curves for ***single-eye*** comparisons against a database of ***160 thousand enrolled individuals***. Only ***Phase 2*** submissions are shown. Plots were generated from 166 thousand mated and 171 thousand nonmated searches.

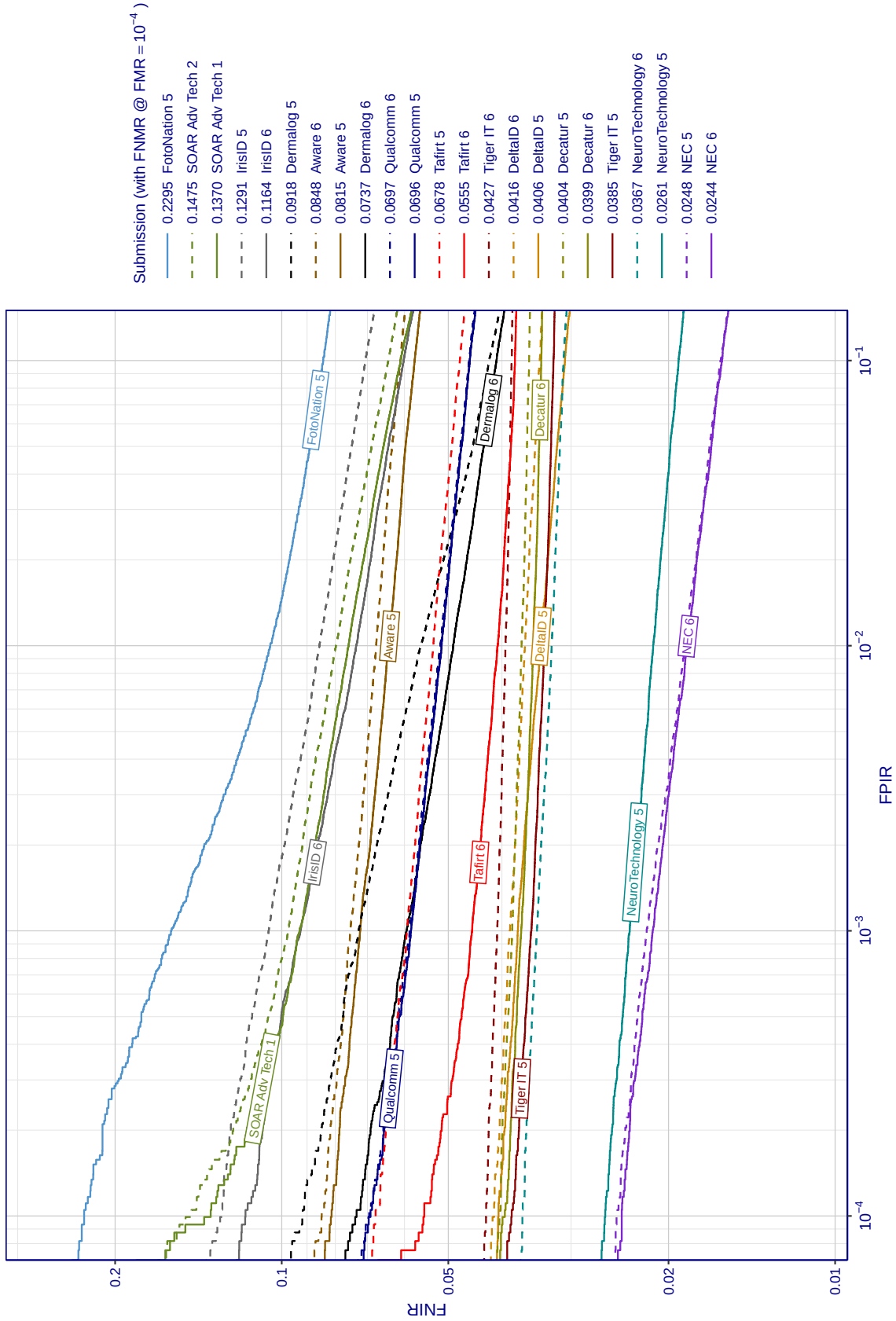

Figure D.10: DET curves for ***single-eye*** comparisons against a database of ***160 thousand enrolled individuals***. Only ***Phase 3*** submissions are shown. Plots were generated from 166 thousand mated and 171 thousand nonmated searches.

Submission	FPIR=10^{-2}	FPIR=10^{-3}	FPIR=10^{-4}
NEC 6	0.019 ± 0.001	0.021 ± 0.001	0.024 ± 0.001
NEC 5	0.019 ± 0.001	0.022 ± 0.001	0.025 ± 0.001
NeuroTechnology 5	0.0213 ± 0.0008	0.0235 ± 0.0009	0.0261 ± 0.0009
NEC 4	0.0199 ± 0.0007	0.0230 ± 0.0008	0.0266 ± 0.0008
NEC 3	0.0207 ± 0.0008	0.0236 ± 0.0008	0.0270 ± 0.0008
DeltaID 4	0.0266 ± 0.0008	0.0296 ± 0.0008	0.0330 ± 0.0009
NeuroTechnology 3	0.028 ± 0.001	0.031 ± 0.002	0.034 ± 0.002
NeuroTechnology 4	0.029 ± 0.001	0.033 ± 0.001	0.037 ± 0.002
NeuroTechnology 6	0.0325 ± 0.0009	0.0346 ± 0.0009	0.037 ± 0.001
Tiger IT 5	0.0336 ± 0.0009	0.0354 ± 0.0009	0.038 ± 0.001
Decatur 6	0.035 ± 0.001	0.037 ± 0.001	0.040 ± 0.002
Decatur 5	0.037 ± 0.001	0.038 ± 0.001	0.040 ± 0.002
DeltaID 5	0.034 ± 0.001	0.037 ± 0.001	0.041 ± 0.002
DeltaID 3	0.035 ± 0.001	0.038 ± 0.001	0.041 ± 0.002
DeltaID 6	0.0364 ± 0.0007	0.0388 ± 0.0007	0.0416 ± 0.0007
Tiger IT 6	0.0394 ± 0.0007	0.0408 ± 0.0007	0.0427 ± 0.0007
Decatur 4	0.043 ± 0.001	0.045 ± 0.001	0.049 ± 0.002
Tiger IT 4	0.044 ± 0.002	0.047 ± 0.002	0.051 ± 0.002
Tafirt 4	0.0322 ± 0.0009	0.0391 ± 0.0009	0.053 ± 0.001
Tafirt 3	0.0384 ± 0.0009	0.0449 ± 0.0009	0.0532 ± 0.0009
Tafirt 6	0.0407 ± 0.0009	0.0453 ± 0.0009	0.055 ± 0.001
Tiger IT 3	0.0459 ± 0.0008	0.0483 ± 0.0009	0.056 ± 0.001
Decatur 3	0.048 ± 0.001	0.050 ± 0.001	0.057 ± 0.001
Tafirt 5	0.053 ± 0.001	0.060 ± 0.001	0.068 ± 0.001
Qualcomm 5	0.051 ± 0.001	0.058 ± 0.001	0.070 ± 0.001
Qualcomm 6	0.0517 ± 0.0008	0.0588 ± 0.0008	0.0697 ± 0.0009
Dermalog 6	0.0490 ± 0.0008	0.0595 ± 0.0009	0.0737 ± 0.0009
Aware 5	0.0642 ± 0.0007	0.0722 ± 0.0007	0.0815 ± 0.0008
Aware 6	0.0675 ± 0.0008	0.0752 ± 0.0008	0.0848 ± 0.0009
Aware 3	0.071 ± 0.001	0.078 ± 0.002	0.085 ± 0.002
Dermalog 5	0.0561 ± 0.0008	0.0726 ± 0.0008	0.0918 ± 0.0008
Aware 4	0.0807 ± 0.0008	0.0878 ± 0.0008	0.0960 ± 0.0009
IrisID 4	0.071 ± 0.001	0.088 ± 0.001	0.113 ± 0.002
IrisID 6	0.073 ± 0.001	0.093 ± 0.001	0.116 ± 0.002
IrisID 3	0.079 ± 0.001	0.096 ± 0.001	0.124 ± 0.001
Qualcomm 3	0.091 ± 0.001	0.104 ± 0.001	0.125 ± 0.001
IrisID 5	0.0855 ± 0.0006	0.1057 ± 0.0007	0.1291 ± 0.0007
SOAR Adv Tech 1	0.0757 ± 0.0006	0.0929 ± 0.0007	0.1370 ± 0.0007
SOAR Adv Tech 2	0.080 ± 0.001	0.097 ± 0.001	0.148 ± 0.001
Dermalog 3	0.077 ± 0.001	0.106 ± 0.001	0.158 ± 0.001
Dermalog 4	0.0860 ± 0.0008	0.1131 ± 0.0008	0.1579 ± 0.0008
FotoNation 4	0.0739 ± 0.0008	0.1080 ± 0.0008	0.1661 ± 0.0008
FotoNation 5	0.105 ± 0.001	0.159 ± 0.001	0.229 ± 0.002
FotoNation 3	0.114 ± 0.001	0.162 ± 0.001	0.232 ± 0.002
Unique Biometrics 2	0.295 ± 0.001	0.295 ± 0.001	0.295 ± 0.001
Qualcomm 4	0.5370 ± 0.0009	0.5539 ± 0.0009	0.573 ± 0.001

Table D.5: Accuracy table for single-eye matching against an enrolled population of 160 thousand. Standard deviations are presented after the plus/minus.

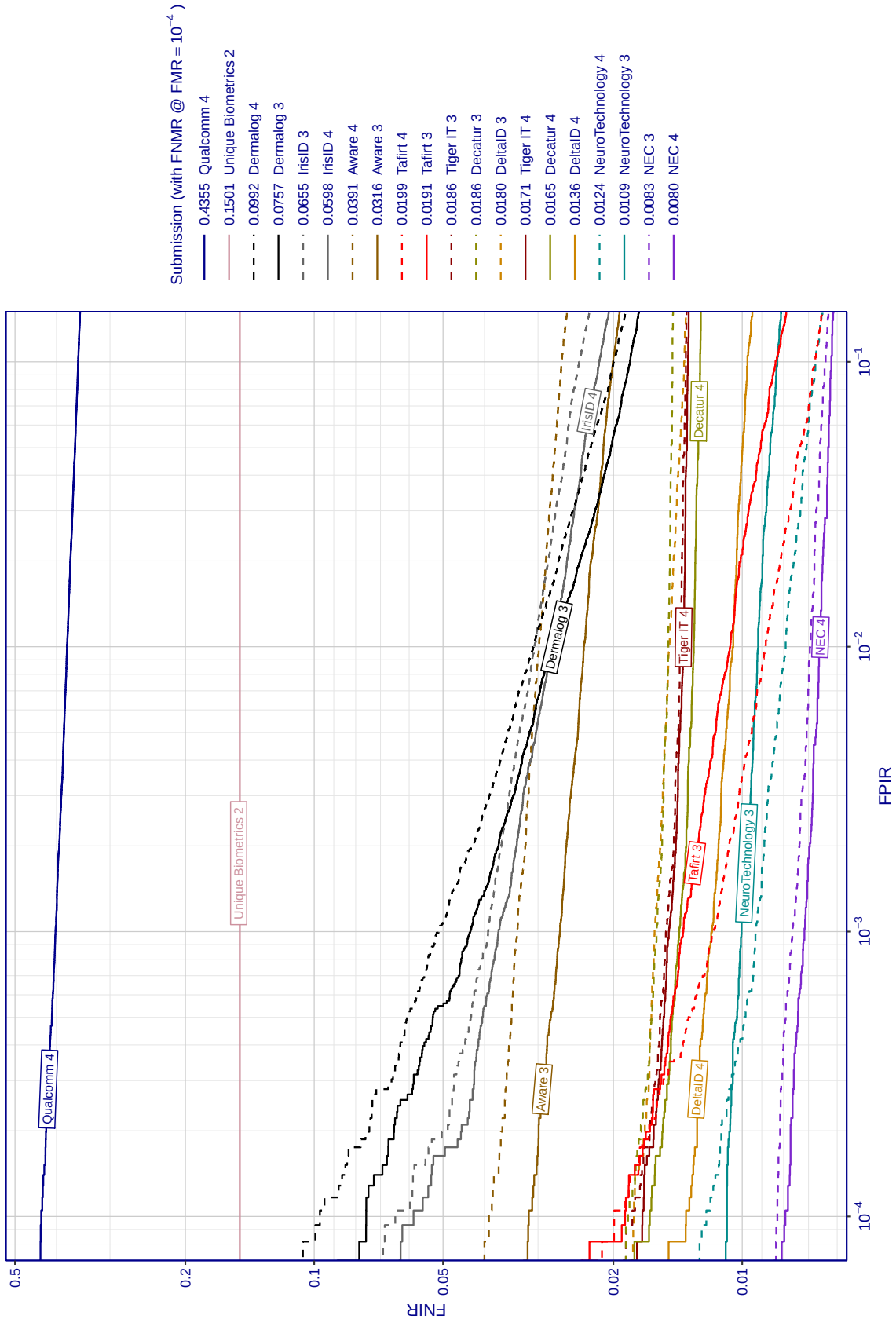

Figure D.11: DET curves for ***two-eye*** comparisons against a database of ***160 thousand enrolled individuals***. Only ***Phase 2*** submissions are shown. Plots were generated from 83 thousand mated and 86 thousand nonmated searches.

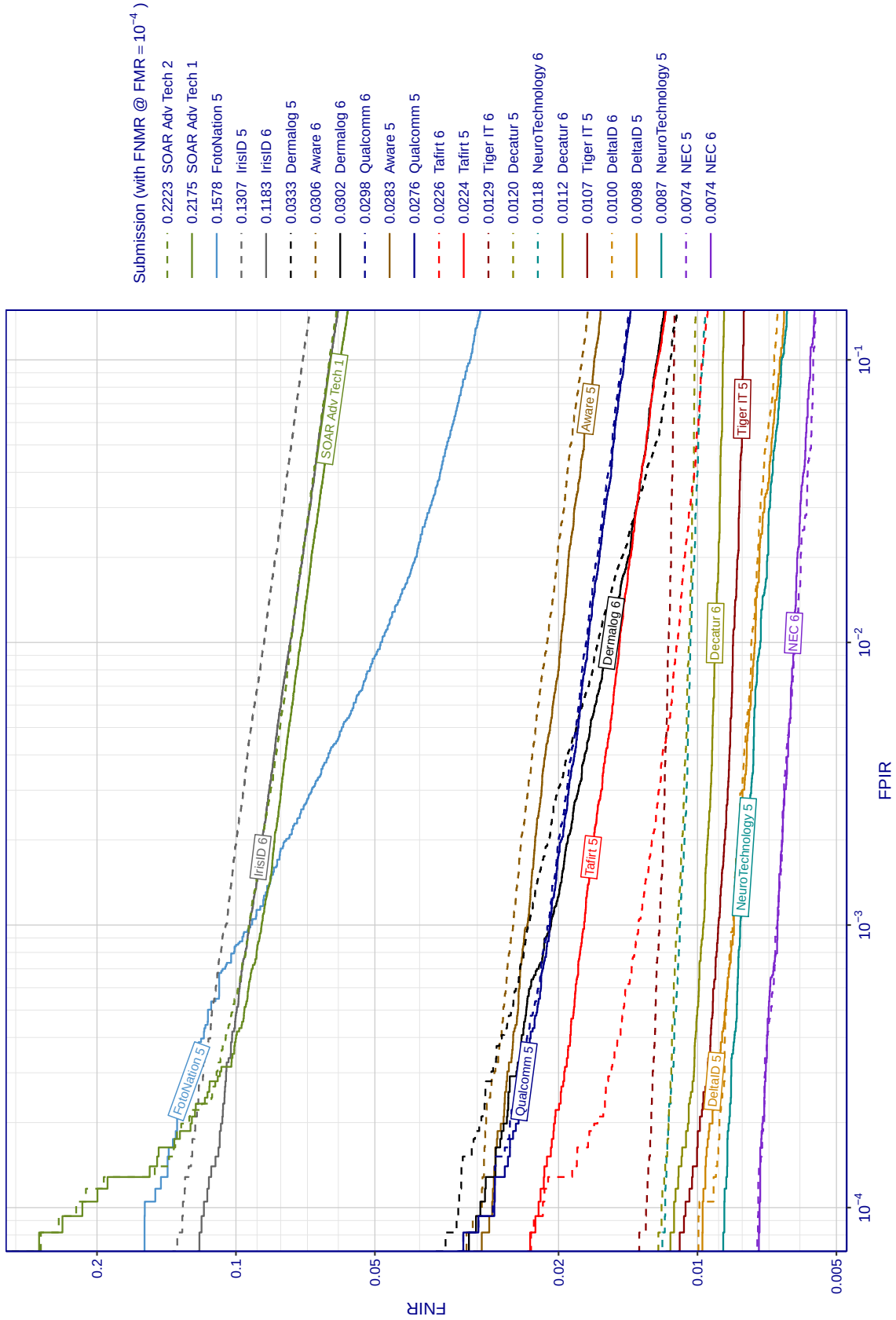

Figure D.12: DET curves for ***two-eye*** comparisons against a database of ***160 thousand enrolled individuals***. Only ***Phase 3*** submissions are shown. Plots were generated from 83 thousand mated and 86 thousand nonmated searches.

Submission	FMR=10^{-2}	FMR=10^{-3}	FMR=10^{-4}
NEC 6	0.0062 ± 0.0009	0.007 ± 0.001	0.007 ± 0.001
NEC 5	0.006 ± 0.001	0.007 ± 0.001	0.007 ± 0.001
NEC 4	0.0066 ± 0.0006	0.0072 ± 0.0007	0.0080 ± 0.0007
NEC 3	0.0070 ± 0.0006	0.0076 ± 0.0006	0.0083 ± 0.0006
NeuroTechnology 5	0.0073 ± 0.0006	0.0081 ± 0.0006	0.0087 ± 0.0006
DeltaID 5	0.0075 ± 0.0006	0.0083 ± 0.0006	0.0098 ± 0.0007
DeltaID 6	0.0077 ± 0.0007	0.0084 ± 0.0007	0.0100 ± 0.0008
Tiger IT 5	0.0084 ± 0.0007	0.0090 ± 0.0007	0.0107 ± 0.0008
NeuroTechnology 3	0.009 ± 0.001	0.010 ± 0.001	0.011 ± 0.001
Decatur 6	0.0092 ± 0.0009	0.010 ± 0.001	0.011 ± 0.001
NeuroTechnology 6	0.010 ± 0.001	0.011 ± 0.001	0.012 ± 0.002
Decatur 5	0.010 ± 0.001	0.011 ± 0.001	0.012 ± 0.002
NeuroTechnology 4	0.0079 ± 0.0005	0.0093 ± 0.0005	0.0124 ± 0.0006
Tiger IT 6	0.0115 ± 0.0005	0.0121 ± 0.0005	0.0129 ± 0.0006
DeltaID 4	0.011 ± 0.002	0.012 ± 0.002	0.014 ± 0.002
Decatur 4	0.0130 ± 0.0007	0.0140 ± 0.0008	0.0165 ± 0.0008
Tiger IT 4	0.0137 ± 0.0007	0.0148 ± 0.0007	0.0171 ± 0.0008
DeltaID 3	0.0148 ± 0.0007	0.0160 ± 0.0007	0.0180 ± 0.0008
Tiger IT 3	0.0140 ± 0.0006	0.0151 ± 0.0007	0.0186 ± 0.0008
Decatur 3	0.015 ± 0.002	0.016 ± 0.002	0.019 ± 0.002
Tafirt 3	0.0107 ± 0.0009	0.0138 ± 0.0009	0.019 ± 0.001
Tafirt 4	0.0088 ± 0.0009	0.0117 ± 0.0009	0.020 ± 0.001
Tafirt 5	0.0148 ± 0.0007	0.0177 ± 0.0007	0.0224 ± 0.0007
Tafirt 6	0.0110 ± 0.0007	0.0134 ± 0.0007	0.0226 ± 0.0007
Qualcomm 5	0.0173 ± 0.0005	0.0209 ± 0.0005	0.0276 ± 0.0006
Aware 5	0.0198 ± 0.0006	0.0234 ± 0.0006	0.0283 ± 0.0007
Qualcomm 6	0.017 ± 0.001	0.021 ± 0.001	0.030 ± 0.002
Dermalog 6	0.0155 ± 0.0006	0.0208 ± 0.0006	0.0302 ± 0.0007
Aware 6	0.0212 ± 0.0007	0.0251 ± 0.0007	0.0306 ± 0.0007
Aware 3	0.023 ± 0.001	0.027 ± 0.001	0.032 ± 0.001
Dermalog 5	0.016 ± 0.001	0.023 ± 0.001	0.033 ± 0.001
Aware 4	0.0298 ± 0.0009	0.034 ± 0.001	0.039 ± 0.001
IrisID 4	0.0277 ± 0.0008	0.0371 ± 0.0009	0.060 ± 0.001
IrisID 3	0.0306 ± 0.0005	0.0393 ± 0.0005	0.0655 ± 0.0006
Dermalog 3	0.0275 ± 0.0005	0.0434 ± 0.0005	0.0757 ± 0.0006
Dermalog 4	0.0307 ± 0.0008	0.0518 ± 0.0008	0.0992 ± 0.0009
IrisID 6	0.0763 ± 0.0008	0.0935 ± 0.0008	0.1183 ± 0.0009
IrisID 5	0.0869 ± 0.0007	0.1054 ± 0.0007	0.1307 ± 0.0007
Unique Biometrics 2	0.1501 ± 0.0006	0.1501 ± 0.0007	0.1501 ± 0.0007
FotoNation 5	0.049 ± 0.001	0.094 ± 0.001	0.158 ± 0.002
SOAR Adv Tech 1	0.072 ± 0.001	0.088 ± 0.001	0.218 ± 0.002
SOAR Adv Tech 2	0.0760 ± 0.0008	0.0939 ± 0.0008	0.2223 ± 0.0009
Qualcomm 4	0.3780 ± 0.0007	0.4023 ± 0.0007	0.4355 ± 0.0008

Table D.6: Accuracy table for two-eye matching against an enrolled population of 160 thousand. Standard deviations are presented after the plus/minus.