

NISTIR 8271 DRAFT SUPPLEMENT

Face Recognition Vendor Test (FRVT) Part 2: Identification

Patrick Grother
Mei Ngan
Kayee Hanaoka
*Information Access Division
Information Technology Laboratory*

This document is a draft supplement of [NIST Interagency Report 8271](#)

2021/08/05



NISTIR 8271 DRAFT SUPPLEMENT

Face Recognition Vendor Test (FRVT) Part 2: Identification

Patrick Grother
Mei Ngan
Kayee Hanaoka
*Information Access Division
Information Technology Laboratory*

This document is a draft supplement of [NIST Interagency Report 8271](#)

June 2021



U.S. Department of Commerce
Wynn Coggins, Acting Secretary

National Institute of Standards and Technology
James Altoff, Under Secretary of Commerce for Standards and Technology and Director, acting

RELEASE NOTES

2021-08-02: The 1:N track of the FRVT remains open. Three news items:

- ▷ This document is the ninth draft update to [NIST Interagency Report 8271](#). It includes results for algorithms recently submitted by eight participants: Cyberlink Corp, NEC Corp, N-Tech Lab, Realnetworks Inc., Sensetime Group, Veridas Digital, Viettel Group, and Vigilant Solutions.
- ▷ Algorithms submitted since July 24 will be included in the next update scheduled for September 9, 2021.
- ▷ A new report, NIST Interagency Report 8381 - FRVT Part 7: Identification for Paperless Travel and Immigration, has been released [[PDF](#), [webpage](#)]. It documents the use of FRVT 1:N algorithms in positive access control and immigration status update travel applications where the enrolled population size is as low as 420 people for aircraft boarding, and 42 000 for an airport security line. These population sizes are much smaller than those used in the main [1:N evaluation](#). Going forward, we will update the report and webpage with results for new algorithms.

2021-07-07: The 1:N track of the FRVT remains open. One update:

- ▷ This document is the eighth draft update to [NIST Interagency Report 8271](#). It include results for an algorithm from one participant: Kakao Enterprises.

2021-06-22: The 1:N track of the FRVT remains open. Three updates:

- ▷ This is the seventh draft of the update to [NIST Interagency Report 8271](#). It includes results for algorithms from three new participants: Line Corporation, Rendip, and Samsung S1 Corp.
- ▷ We have also added results for algorithms from five returning developers: Imagus Technology, Kneron, Tevian, Visidon, and Xforward AI Technology.
- ▷ The algorithm-specific report cards (examples: [1](#), [2](#), and [3](#)) now include figures showing how low threshold values can be used to reduce candidate list lengths for human review, while (usually) elevating miss rates (FNIR) only modestly. The reports also feature some minor additions and clarifications.

2021-03-26: The 1:N track of the FRVT remains open. Three updates:

- ▷ This is the sixth draft of the update to [NIST Interagency Report 8271](#). It includes results for algorithms from three returning developers: Neurotechnology, Guangzhou Pixel Solutions, and Tech5 SA.
- ▷ We have added results on the webpage and in the report for a new ageing dataset in which border crossing photos are searched against a gallery of border crossing photos collected between 10 and 15 years prior to the mated search photos. See section [2](#) for a description of the images. Table [1](#) has a new entry describing the experiment.
- ▷ We will mostly discontinue running the mugshot ageing test, reserving it for algorithms that show high accuracy on the new border-crossing set.

2021-03-26: Regarding the fifth draft of the update to [NIST Interagency Report 8271](#):

- ▷ In addition have added results for first algorithms from two new participants: Viettel Group and Veridas Digital Authentication Solutions.
- ▷ We have added results for algorithms from two returning developers: Idemia and Cognitec Systems.
- ▷ In addition to the report, the [results page](#) and its hyperlinked [report cards](#) have been updated.

2021-02-08: Regarding the fourth draft of the update to [NIST Interagency Report 8271](#):

- ▷ We have added results for eight algorithms submitted by eight developers: Cyberlink, Dermalog, Imagus, Paravision, Sensetime, Trueface, Vigilant Solutions, and X-Forward AI. With the exception of Trueface, all of these developers have participated previously.

- ▷ We anticipate updating this report again in the first week of March 2021.
- ▷ The main [results page](#) has been revised with tabs for the investigative and lights-out identification tables, and a new tab dedicated to speed and resource consumption.
- ▷ The report cards (example [here](#)) hyperlinked from the [results page](#) have been revised to improve content and format.

2020-12-14: Regarding third draft of the update to [NIST Interagency Report 8271](#):

- ▷ We have added results for fifteen algorithms submitted by thirteen developers. The four first-time participants are: Acer, Akurat Satu Indonesia, Canon, and Xforward AI Technology. The ten returning developers are: AllGoVision, Cyberlink Corp, Dahua Technology, Deepglint, Guangzhou Pixel Solutions, IIT Vision, Innovatrics, Rank One Computing, Scanovate, Sensetime Group, Synesis, and VisionLabs.
- ▷ We have added two new datasets to the evaluation: First a set of “visa-border” photos, representing search of an airport immigration lane photo against a database of closely ISO standard portraits; second a “visa-kiosk” set representing search of a photo collected in a registered traveller kiosk against the same ISO portrait gallery. The images are described in section [2.1](#).
- ▷ As in previous reports, we include results for searching mugshots against a mugshot gallery containing a single image of each of 12 million people. However we have suspending running searches against a gallery in which multiple lifetime photos per person are present, because this is computationally expensive. We retain a $N = 3$ million search test dedicated to ageing in which mugshots taken up to 18 years after the first photograph are searched - see Table [6](#).
- ▷ Tables containing computational resource information, Table [2](#) . . . , now include duration of the finalization step, in which search algorithms can, at their option, build fast-search data structures.
- ▷ We have linked revised per-algorithm PDF report cards from the main [results page](#).
- ▷ We have regenerated all figures and tables to drop algorithms submitted before June 2018. Results for prior algorithms appear in [archived editions](#) of this report.
- ▷ Going forward, we anticipate producing more frequent updates to this report. Developers may submit one algorithm to this evaluation every four calendar months.

2020-03-24: Regarding the second draft of the update to [NIST Interagency Report 8271](#):

- ▷ Adds results for three algorithms from three developers, Dermalog, Innovatrics, and Synesis.
- ▷ Adds Table [6](#) on ageing showing the increase in false negative rates with time elapsed between two photos. Some of the results were contained in graphs in prior editions of this report, but the table adds results for some newly submitted algorithms.
- ▷ Adjusts frontal mugshot results (for recent and lifetime consolidated galleries) to include the effect of removing some images that should not have been included in image test sets. These images were mostly profile views, images of tattoos containing faces, images of faces on tee shirts, and images of photographs on walls behind the intended subject. This affects many tables and reduces false negative identification rates for all algorithms. The reduction is larger for “recent” enrollments than for “lifetime consolidated” ones with the consequence that accuracy on recent images is now superior.

2020-02-26: Regarding the first draft of the update to [NIST Interagency Report 8271](#):

- ▷ Adds results for 38 algorithms from 31 different developers, eleven of whom are entirely new to the 1:N track of FRVT. These are Allgovision, Cyberlink, Deepsea Tencent, Farbar F8, Imperial College London, Intsys MSU, Kedacom, Kneron, Pixelall, and Scanovate.

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's Research Protections Office reviewed the protocol for this project 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).

ACKNOWLEDGMENTS

The authors are grateful for the support and collaboration of the the Department of Homeland Security's Science & Technology Directorate (S&T), Office of Biometric Identity Management (OBIM), and Customs and Border Protection (CBP).

Additionally, the authors are grateful to staff in the NIST Biometrics Research Laboratory for infrastructure supporting rapid evaluation of algorithms.

Executive Summary

This document is a draft revision of the September 2019 report [NIST Interagency Report 8271](#). That report gave extensive documentation of face recognition applied to mugshots. This report extends that by adding more two more challenging datasets containing images with serious departures from canonical frontal image standards. The report also adds results for algorithms submitted to NIST since in 2019 and 2020. The algorithms, which implement one-to-many identification of faces appearing in two-dimensional images, are prototypes from the research and development laboratories of mostly commercial suppliers, and are submitted to NIST as compiled black-box libraries implementing a NIST-specified C++ test interface. The report therefore does not describe how algorithms operate. The report lists accuracy results alongside developer names and will therefore be useful for comparison of face recognition algorithms and assessment of absolute capability. The report is accompanied by a [webpage](#) with sortable results.

The evaluation uses six datasets: frontal mugshots, profile view mugshots, desktop webcam photos, visa-like immigration application photos, immigration lane photos, and registered traveler kiosk photos. These datasets are sequestered at NIST, meaning that developers do not have access to them for training or testing. This aspect is important because face recognition algorithms are very often deployed without the developer having access to the customers image data. A possible exception to this would be in a cloud-based application where the operational image data is uploaded to a cloud operated by a face recognition developer.

The major result in NIST IR 8271 was that massive gains in accuracy have been achieved in the years 2013 to 2018 and these far exceed improvements made in the prior period, 2010 to 2013. While the industry gains were broad - at least 30 developers' algorithms outperformed the most accurate algorithm from late 2013, there remains a wide range of capability. While this report shows accuracy gains only over the period 2018-2020, the most accurate algorithm reported here is substantially more accurate than anything reported in NIST IR 8271. This is evidence that face recognition development continues apace, and that FRVT reports are but a snapshot of contemporary capability.

From discussion with developers, the accuracy gains stem from the adoption of deep convolutional neural networks. As such, face recognition has undergone an industrial revolution, with algorithms increasingly tolerant of poorly illuminated and other low quality images, and poorly posed subjects. One related result is that a few algorithms correctly match side-view photographs to galleries of frontal photos, with search accuracy approaching that of the best c. 2010 algorithms operating on purely frontal images. The capability to recognize under a 90-degree change in viewpoint - pose invariance - has been a long-sought milestone in face recognition research.

With good quality portrait photos, the most accurate algorithms will find matching entries, when present, in galleries containing 12 million individuals, with rank one miss rates of approaching 0.1%. The remaining errors are in large part attributable to long-run ageing, facial injury and poor image quality. Given this impressive achievement - close to perfect recognition - an advocate might claim that cooperative face recognition is a solved problem, a statement that can be refuted with the following context and caveats:

- ▷ **Mugshots vs. less constrained captures:** The low error rates reported here are attained using mostly excellent cooperative live-capture mugshot images collected with an attendant present. Recognition in other circumstances, particularly those without a dedicated photographic environment and human or automated quality control checks, will lead to declines in accuracy. This is documented here for side-view images, poorer quality webcam images, and, particularly, for newly introduced ATM-style kiosk photos that were not originally intended for automated face recognition. In this case, recognition error rates are much higher, often in excess of 20% even with the more accurate algorithms which variously remain intolerant of face cropping (at image edge) and of large downward head pitch.
- ▷ **Algorithm accuracy spectrum:** Recognition accuracy is very strongly dependent on the algorithm and, more

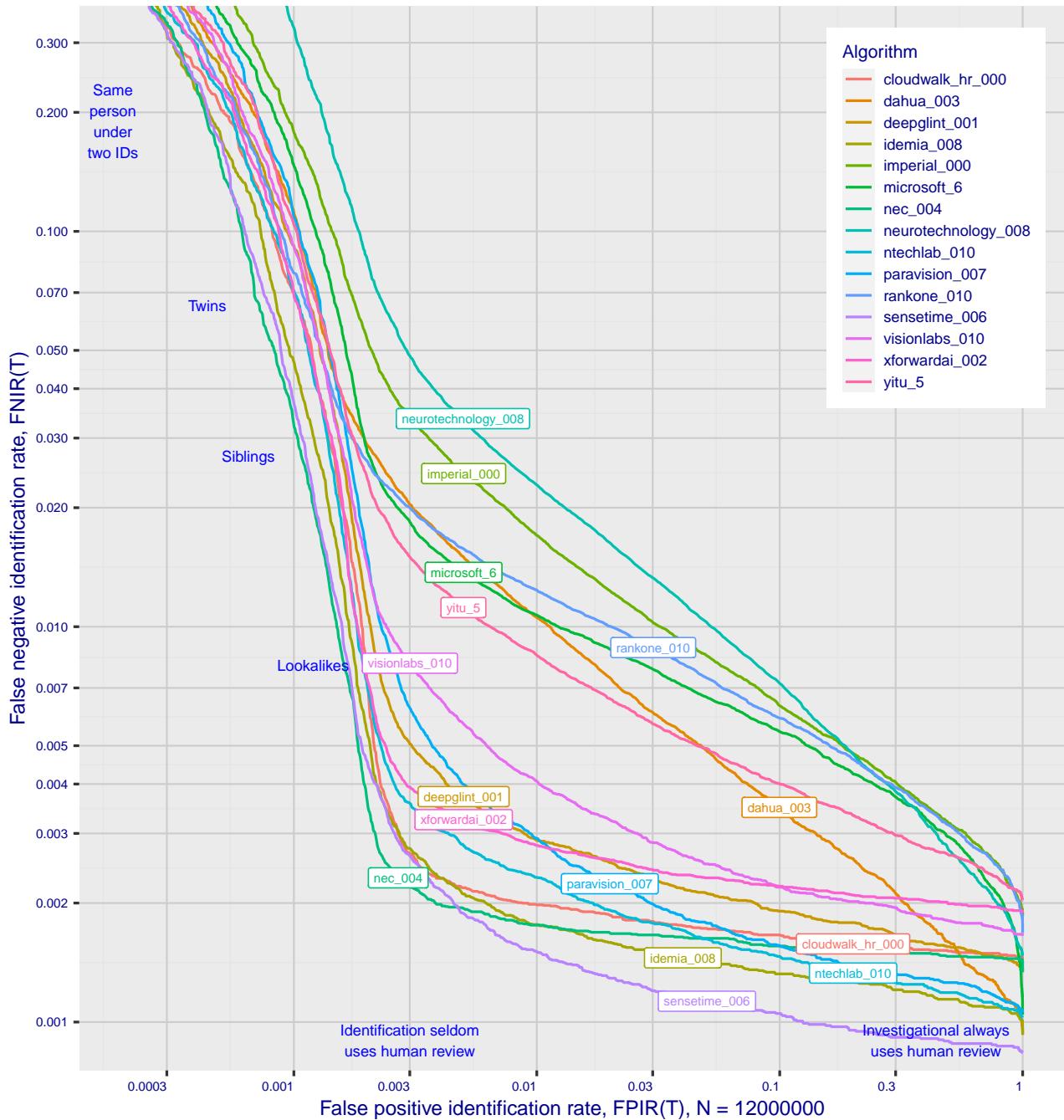


Figure 1: Identification miss rates across the false positive range. N = 12 million individuals are enrolled with one recent image.

generally, on the developer of the algorithm. False negative error rates in a particular scenario range from a few tenths of one percent to beyond fifty percent. This is tabulated exhaustively later: For example Table 9 shows accuracy across datasets. Figure 1 here compares algorithms on mugshot searches in a consolidated gallery of 12 million subjects and 12 million photos. Many algorithms do not achieve the low error rates noted above, and while many of those may still be useful and valuable to end-users, only the most accurate excel on poor quality images and those collected long after the initial enrollment sample.

▷ **Versioning:** While results for up to ten algorithms from each developer are reported here, the intra-provider

accuracy variations are usually smaller than the inter-provider variations. That said different versions give an order of magnitude fewer misses. Some developers demonstrate speed-accuracy tradeoffs¹. See Figs. 18, 19.

- ▷ **Low similarity scores:** In thousands of mugshot cases the correct gallery image is returned at rank 1 but its similarity score is nevertheless low, below some operationally required score threshold. This is not so important when face recognition is used for “lead generation” in investigational applications because human reviewers are specifically required to review potentially long candidate lists and the threshold is effectively 0. In applications where search volumes are higher and labor is not available to review the results from searches, a higher threshold must be applied. This reduces the length of candidate lists and false positive identification rates at the expense of increased false negative miss rates. The tradeoff between the two error rates is reported extensively later.
- ▷ **Population size:** As the number of enrolled subjects grows, some mates are displaced from rank one, decreasing accuracy. As tabulated later for N up to 12 million, false negative rates generally rise slowly with population size. This enables use of face recognition in very large populations. However in most positive and negative identification applications², a score threshold is set to limit the rate at which non-mate searches produce false positives. This has the consequence that some mated searches will report the mate below threshold, i.e. a miss, even if it is at rank 1. The utility of this is that many non-mated searches will return no candidate identities at all. As the error-tradeoff characteristic shows, investigational miss rates on the right side are very low but then rise steadily (in the center region) as threshold is increased to support “lights-out” applications, and ultimately rise quickly (left side) as discussed below. Thus, if we demand that just one in one thousand non-mate searches produce any false positives, the most accurate algorithms there (Sensetime-004 and NEC-3) would fail on between 3 and 5% of mated searches. Even though the graph shows results for the most accurate algorithms, all but two would fail to find the mate in more than 8% of mated searches. While the two most accurate algorithms produce a relatively flat error tradeoff until the threshold is raised to limit false positives to about 1 in 400 non-mated searches³.

Thereafter, as the threshold is raised to further reduce false positives, miss rates rise rapidly. This means that low false positive identification rates are inaccessible with these algorithms, a result that does not apply for ten-finger identification algorithms. The rapid rise occurs because the lower mate scores are mixed with very high non-mate scores, the low scores from poor image quality and ageing, the high non-mates from the presence of lookalikes persons (doppelgangers), twins (discussed next) and, ultimately, the presence of a few unconsolidated subjects i.e. persons present under multiple IDs.

- ▷ **False negatives from ageing:** A large source of error in long-run applications where subjects are not re-enrolled on a set schedule is ageing. Changes in facial appearance increase with the time elapsed between photographs. These will depress similarity scores and eventually cause false negatives. All faces age and while this usually proceeds in a graceful and progressive manner, drug use can accelerate this [28]. Elective surgery may be effective in delaying it although this has not been formally quantified with face recognition. As ageing is essentially unavoidable, it can only be mitigated by scheduled re-capture, as in passport re-issuance. To quantify ageing effects, we used the more accurate algorithms to enroll the earliest image of 3.1 million adults and then search

¹For example, NEC-0 prepares templates much faster than NEC-2 but gives twenty times more misses. Dermalog-5 executes a template search much more quickly than Dermalog-6 but is also much less accurate.

²In a positive identification application such as a registered traveler system, a user is making an implicit claim to be enrolled in the system - most users will be. In a negative application, such as with deportees, the implicit claim is that the subject is not enrolled - most will not be.

³The gallery size here is 12 million people, one image per person. Given 331 201 non-mated searches, an exhaustive implementation of one-too-many search would execute almost 4 trillion comparisons. At a false positive identification rate of 0.0025 the number of false positives is, to first order, 828 corresponding to single-comparison false match rate of $828 / 4 \text{ trillion} = 2.1 \times 10^{-10}$ i.e. about 1 in 5 billion. Strictly this FMR computation is meaningful only for algorithms that implement 1:N search using N 1:1 comparisons, which is not always the case.

with 10.3 million newer photos taken up to 18 years after the initial enrollment photo. Figure 2 puts ageing into context by contrasting it with the increase in false negatives that occurs when the number of individuals in an enrollment database becomes larger and the chance of a false positive increases such that higher thresholds may become necessary⁴.

The Figure shows, from top to bottom, increases in false negative identification rates (FNIR) with the algorithm being tested. This applies to increases due to N on the left side, and increases due to ageing on the right side. The relative spacing of the dots shows that for all algorithms the dependency of FNIR on N (up to 12 million) is considerably less than on ΔT (up to 18 years).

In the inset table, accuracy is seen to degrade progressively with time, as mate scores decline and non-mates displace mates from rank 1 position. More accurate algorithms tend to be less sensitive to ageing. The more accurate algorithms give fewer errors after 18 years of ageing than middle tier algorithms give after four. Note also we do not quantify an ageing rate - more formal methods [2] borrowed from the longitudinal analysis literature have been published for doing so (given suitable repeated measures data). See Figures 60, 80 and 90.

⁴Some algorithms implement strategies to automatically adjust scores to account for increased population size. This relieves the system owner of having to increase thresholds as N increases.

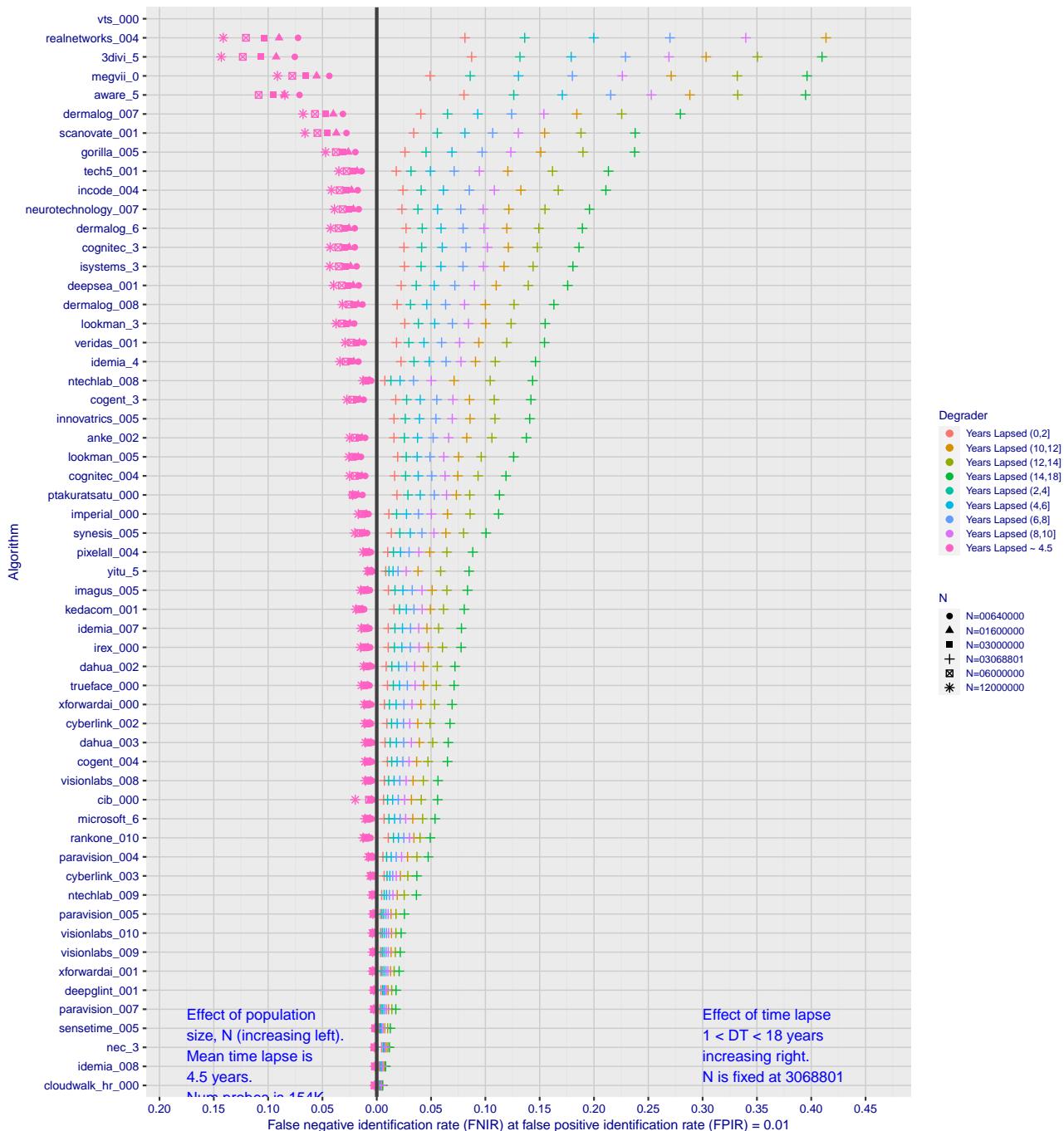


Figure 2: Identification miss rates as a function of enrolled population size, N , and time-lapse, ΔT .

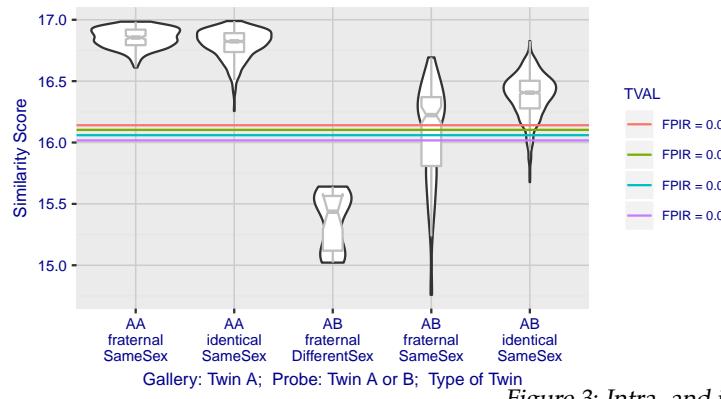


Figure 3: Intra- and inter-twin scores

▷ **False positives from twins:** By enrolling 640 000 mugshots, adding photos of one twin, and then searching photos of those subjects and their twin the inset figure shows, for one typical algorithm, the similarity is generally greater when searching twins against themselves (A) than when searching twins against their sibling (B) but very often still above even stringent thresholds i.e. those corresponding to one in one thousand searches producing a false positive. Thus twins will very often produce a high-scoring non-match on a candidate list and a false alarm in an online identification system. The plot of Fig. 3 shows that fraternal twins are sometimes correctly rejected at those thresholds - including most different sex twins (at center). Figure ?? shows substantially similar behavior for all algorithms tested. In an investigative search, a twin would typically appear at rank 1, or rank 2 if their sibling happened to also be the gallery. Twins (and triplets etc.) constituted 3.3% of all live births [17] in recent years⁵, and because that number is higher today than when the individuals in current adult databases were born, the false positives that arise from twins are now, and will increasingly be, an operational problem. Relative to the United States, twins are born with considerable regional variation. For example they are much less common in East Asia, and much more common in Sub-Saharan Africa [21].

The presence of twins in the mugshot database is inevitable given its size, around 12.3 million people. As this is not an insignificant sample of the domestic United States population, people with other familial ties will be present also. The data was collected over an extended period and because location information is not available, we are unable to estimate the proportion of the domestic population that is present in the dataset. However, if we assume twins are neither more or less disposed to arrest than the general population, we can estimate that hundreds of thousands of individuals in the dataset are twins. This will affect false positive rates because we randomly set aside 331 201 individuals for nonmate searches, and some proportion of those will be twins with siblings in the gallery.

▷ **Database integrity:** An operational error rate should be added to all false negative rates in this report reflecting the proportion of images in a real database that are un-matchable. Such anomalies arise from images that: do not contain a face; include multiple persons; cannot be decoded; are rotated by 90° or 180°; depict a face on clothing; and others introduced by a long tail of various clerical errors. While the mugshot trials in this report have been constructed to minimize such effects, they are a real problem in actual operations.

This report is being updated continuously as new algorithms are submitted to FRVT, and run on new datasets. Participation in the [one-to-many identification track](#) is independent of participation in the [one-to-one verification track](#) of FRVT.

⁵See the CDC's National Vital Statistics Report for 2017: https://www.cdc.gov/nchs/data/nvsr/nvsr67/nvsr67_08-508.pdf

Scope and Context

Audience: This report is intended for developers, integrators, end users, policy makers and others who have some familiarity with biometrics applications. The methods and metrics documented here will be of interest to organizations engaged in tests of face recognition algorithms. Some of these have been incorporated in the ISO/IEC 19795 Part 1 Biometric Testing and Reporting Framework standard, now nearing publication.

Prior benchmarks: Automated face recognition accuracy has improved massively in the two decades since initial commercialization of the various technologies. NIST has tracked that improvement through its conduct of regular independent, free, open, and public evaluations. These have fostered improvements in the state of the art. This report serves as an update to the [NIST Interagency Report 8271](#) on performance of face identification algorithms, published in September 2019.

Demographics: In December 2019, NIST published a first report on demographic dependencies in face recognition, [NIST Interagency Report 8280](#) that documented age, sex and race differentials in one-to-one and one-to-many false positive and false negative rates.

Scope: NIST IR 8271 documented recognition results for four databases containing in excess of 30.2 million still photographs of 14.4 million individuals. That constituted the largest public and independent evaluation of face recognition ever conducted. It includes results for accuracy, speed, investigative vs. identification applications, scalability to large populations, use of multiple images per person, images of cooperative and non-cooperative subjects.

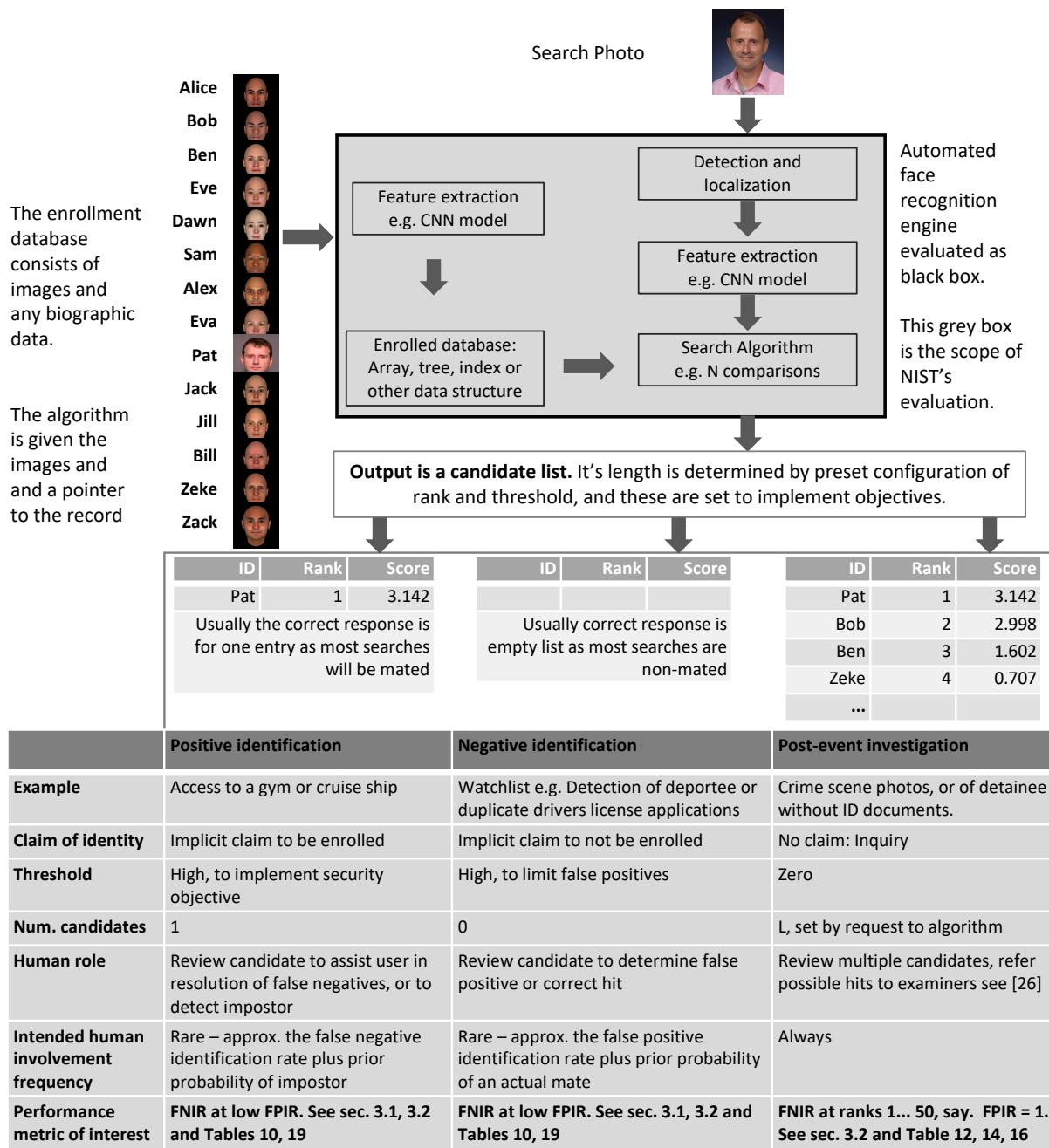
The report also includes results for ageing, recognition of twins, and recognition of profile-view images against frontal galleries. It otherwise does not address causes of recognition failure, neither image-specific problems nor subject-specific factors including demographics. Separate reports on demographic dependencies in face recognition will be published in the future. Additionally out of scope are: performance of live [human-in-the-loop transactional systems](#) like automated border control gates; human recognition accuracy as used in forensic applications; and recognition of persons in video sequences (which NIST evaluated separately [9]). Some of those applications share core matching technologies that *are* tested in this report.

Images: Five kinds of images are employed; these are either compared with images of the same kind, or against others from different capture environments as follows. The primary dataset is a set of law enforcement mugshot images (Fig. 5) which are enrolled and then searched with three kinds of images: other mugshots (i.e. within-domain); profile-view photographs (90 degree cross-view); and lower quality webcam images (Fig. 6) collected in similar detention operations (cross-domain). Additionally we compare high quality visa-like photos collected in immigration offices, with: medium quality border crossing images collected in primary immigration lanes; poor quality images collected in ATM-like registered traveller kiosks.

Participation and industry coverage: The report includes performance figures for prototype algorithms from the research laboratories of commercial developers and a few universities. This represents a substantial majority of the face recognition industry, but only a tiny minority of the academic community. Participation was open worldwide. While there is no charge for participation, developers incur some software engineering expense in implementing their algorithms behind the NIST application programming interface (API). The test is a black-box test where the function of the algorithm, and the intellectual property associated with it, is hidden inside pre-compiled libraries.

Recent technology development: Most face recognition research with deep convolutional neural networks (CNNs) has been aimed at achieving invariance to pose, illumination and expression variations that characterize photojournalism and social media images. The initial research [18, 22] employed large numbers of images of relatively few ($\sim 10^4$) individuals to learn invariance. Inevitably much larger populations ($\sim 10^7$) were employed for training [11, 20] but the benchmark, Labeled Faces in the Wild with (essentially) an equal error rate metric [12], represents an easy task,

one-to-one verification at very high false match rates. While a larger scale identification benchmark duly followed, Megaface [15], its primary metric, rank one hit rate, contrasts with the high threshold discrimination task required in most large-population applications of face recognition, namely credential de-duplication, and background checks. There, identification in galleries containing up to 10^8 individuals must be performed using a) very few images per individual and b) stringent thresholds to afford very low false positive identification rates. This track of FRVT was launched to measure the capability of the new technologies, including in these two cases. FRVT has included open-set identification tests since 2002, reporting both false negative and positive identification rates [7].



Performance metrics for applications: This report documents the performance of one-to-many face recognition algorithms. The word "performance" here refers to recognition accuracy and computational resource usage, as measured

by executing those algorithms on massive sequestered datasets.

This report includes extensive tabulation of recognition error rates germane to the main use-cases for face search technology. The Figure below, inspired by the Figure 1 in [23] differentiates different applications of the technolgy. The last row directs readers to the main tables relevant to those applications, respectively threshold-based and rank-based metrics that are special cases of the metrics given in section 3. The terms negative identification and positive identification are taken from the ISO/IEC 2382-37:2017 standardized biometrics vocabulary.

The algorithms are specifically configured for these applications by setting thresholds and candidate list lengths. Both rank-based metrics and threshold-based metrics include tradeoffs. In investigation, overall accuracy will be reduced if labor is only available to review a few candidates from the automated system. Note that when a fixed number of candidates are returned, the false positive identification rate of the automated face recognition engine will be 100%, because a probe image of anyone not enrolled will still return candidates. In identification applications where false positives must be limited to satisfy reviewer labor availability or a security objective, higher false negative rates are implied. This report includes extensive quantification of this threshold-based tradeoff.

See Sec. 3

Template diversity: The FRVT is designed to evaluate black-box technologies with the consequence that the templates that hold features extracted from face images are entirely proprietary opaque binary data that embed considerable intellectual property of the developer. Despite migration to CNN-based technologies there is no consensus on the optimal feature vector dimension. This is evidenced by template sizes ranging from below 100 bytes to more than four kilobytes. This diversity of approaches, suggests there is no prospect of a standard template something that would require a common feature set to be extracted from faces. Interoperability in automated face recognition remains solidly based on images and documentary standards for those, in particular the ICAO portrait [27] specification deriving from the ISO/IEC 19794-5 Token frontal [24] standard, which are similar to certain ANSI/NIST Type 10 [26] formats.

Training: The algorithms submitted to NIST have been developed using image datasets that developers do not disclose. The development will often include application of machine learning techniques and will additionally involve iterative training and testing cycles. NIST itself does not perform any training and does not refine or alter the algorithm in any way. Thus the model, data files, and libraries that define an algorithm are fixed for the duration of the tests. This reflects typical operational reality where recognition software, once installed, is fixed and constant until upgraded. This situation persists because on-site training of algorithms on customer data is atypical essentially because training is not a turnkey process.

Automated search and human review: Virtually all applications using automated face search require human review of the outputs at some frequency: Always for investigational applications; rarely in positive identification applications, after rejection (false or otherwise); and rarely in negative identification applications, after an alarm (false or otherwise). The human role is usually to compare a reference image with the query image or the live-subject if present, to render either a definitive decision on “exclusion” (different subjects), or “identification” (same subject), or a declaration that one or both images have “no value” and that no decision can be made. Note that automated face recognition algorithms are not built to do exclusion - low scores from a face comparison arise from different faces *and* poor quality images of the same face.

Human reviewers make recognition errors [5, 19, 25] and are sensitive to image acquisition and quality. Accurate human review is supported by high resolution - as specified in the Type 50, 51 acquisition profiles of the ANSI/NIST Type 10 record [26], and by multiple non-frontal views as specified in the same standard. These often afford views of the ear. Organizations involved in image collection should consider supporting human adjudication by collecting high-resolution frontal and non-frontal views, preparing low resolution versions for automated face recognition [24], and retaining both for any subsequent resolution of candidate matches. Along these lines, the ISO/IEC Joint Technical

Committee 1 subcommittee 37 on biometrics has just initiated projects on image quality assessment and face-aware capture.

Release Notes

FRVT Activities: Since February 2017, NIST has been evaluating one-to-one verification algorithms on an ongoing basis. NIST then restarted FRVT's one-to-many track in February 2018, inviting participants to send up to prototype algorithms. Both tracks allows developers to submit updated algorithms to NIST at any time but no more frequently than four calendar months. This more closely aligns development and evaluation schedules. Results are posted to the web within a few weeks of submission. Details and full report are linked from the [Ongoing FRVT site](#).

FRVT Reports: The results of the FRVT appear in the series NIST Interagency Reports tabulated below. The reports were developed separately and released on different schedules. In prior years NIST has mostly reported FRVT results as a single report; this had the disadvantage that results from completed sub-studies were not published until all other studies were complete.

Date	Link	Title	No.
2014-03-20	PDF	FRVT Performance of Automated Age Estimation Algorithms	7995
2015-04-20	PDF	Face Recognition Vendor Test (FRVT) Performance of Automated Gender Classification Algorithms	8052
2014-05-21	PDF	FRVT Performance of face identification algorithms	8009
2017-03-07	PDF	Face In Video Evaluation (FIVE) Face Recognition of Non-Cooperative Subjects	8173
2017-11-23	PDF	The 2017 IARPA Face Recognition Prize Challenge (FRPC)	8197
2018-11-27	PDF	Face Recognition Vendor Test - Part 2: Identification	8271
2019-09-11	PDF	Face Recognition Vendor Test - Part 2: Identification	8271
2019-12-11	PDF	Face Recognition Vendor Test - Part 3: Demographic Effects	8280
2020-01-03	WWW	Face Recognition Vendor Test (FRVT) - Part 1 Verification	Draft

Details appear on pages linked from <https://www.nist.gov/programs-projects/face-projects>.

Appendices: This report is accompanied by appendices which present exhaustive results on a per-algorithm basis. These are machine-generated and are included because the authors believe that visualization of such data is broadly informative and vital to understanding the context of the report.

Typesetting: Virtually all of the tabulated content in this report was produced automatically. This involved the use of scripting tools to generate directly type-settable L^AT_EX content. This improves timeliness, flexibility, maintainability, and reduces transcription errors.

Graphics: Many of the Figures in this report were produced using the **ggplot2** package running under **R**, the capabilities of which extend beyond those evident in this document.

Contents

Release Notes	1
Disclaimer	3
Institutional Review Board	3
Acknowledgments	3
Executive Summary	4
Scope and Context	10
Release Notes	14
1 Introduction	16
2 Evaluation datasets	17
3 Performance metrics	23
4 Results	39
Appendices	71
A Accuracy on large-population FRVT 2018 mugshots	71
B Effect of time-lapse: Accuracy after face ageing	116
C Effect of enrolling multiple images	209
D Accuracy with poor quality webcam images	216
E Accuracy for profile-view to frontal recognition	226
F Search duration	230
G Gallery Insertion Timing	237

1 Introduction

One-to-many identification represents the largest market for face recognition technology. Algorithms are used across the world in a diverse range of biometric applications: detection of duplicates in databases, detection of fraudulent applications for credentials such as passports and driving licenses, token-less access control, surveillance, social media tagging, lookalike discovery, criminal investigation, and forensic clustering.

This report contains a breadth of performance measurements relevant to many applications. Performance here refers to accuracy and resource consumption. In most applications, the core accuracy of a facial recognition algorithm is the most important performance variable. Resource consumption will be important also as it drives the amount of hardware, power, and cooling necessary to accommodate high volume workflows. Algorithms consume processing time, they require computer memory, and their static template data requires storage space. This report documents these variables.

1.1 Open-set searches

FRVT tested open-set identification algorithms. Real-world applications are almost always “open-set”, meaning that some searches have an enrolled mate, but some do not. For example, some subjects have truly not been issued a visa or drivers license before; some law enforcement searches are from first-time arrestees⁶. In an “open-set” application, algorithms make no prior assumption about whether or not to return a high-scoring result, and for a mated search, the ideal behaviour is that the search produces the correct mate at high score and first rank. For a non-mate search, the ideal behavior is that the search produces zero high-scoring candidates.

Many academic benchmarks execute only closed-set searches. The proportion of mates found in the rank one position is the default accuracy metric. This hit rate metric ignores the score with which a mate is found; weak hits count as much as strong hits. This ignores the real-world imperative that in many applications it is necessary to elevate a threshold to reduce the number of false positives.

⁶Operationally closed-set applications are rare because it is usually not the case that all searches have an enrolled mate. One counter-example, however, is a cruise ship in which all passengers are enrolled and all searches should produce exactly one identity. Another example is forensic identification of dental records from an aircraft crash.

2 Evaluation datasets

This report documents accuracy for four kinds of images - mugshots, webcam, profiles and wild - as described in the following sections.

2.1 Immigration-related images

This report includes benchmark tests sharing a common enrollment of high quality frontal portrait images collected while subject make applications for various immigration benefits. We then search that with two kinds of images, webcam images collected during in-bound immigration and also images collected from registered travelers using a ATM-style kiosk. These are described below and depicted in Figure 4.



Figure 4: Example photos.

- ▷ **Application reference photos:** The images are collected in an attended interview setting using dedicated capture equipment and lighting. The images, at size 300x300 pixels, are smaller than normally indicated by ISO. The images are all high-quality frontal portraits collected in immigration offices and with a white background. As such, potential quality related drivers of high false match rates (such as blur) can be expected to be absent. The images are encoded as ISO/IEC 10918-1 i.e. JPEG. Older images had a compression ration of about 16:1, while newer images, since 2010, are more lightly compressed at 4:1. When these images are provided as input into the algorithm, they are labeled with the type "iso". This report enrols 1 600 000 application images, one per person.
- ▷ **Border crossing photos:** Most images are have width 320 and height 240 pixels. They are JPEG compressed at 16:1 i.e. filesize just below 15KB. The images present challenges for face recognition in that subjects often exhibit non-zero yaw and pitch (associated with the rotational degrees of freedom of the camera mount), low contrast (due to varying and intense background lights), and poor spatial resolution (due to inexpensive cameras). There are often subjects standing in the background, usually at very low resolution (see Figure 4b). In such cases, algorithms should detect all faces and determine which is the largest and most centered. When these images are provided as input into the algorithm, they are labeled with the type "wild".
- ▷ **Kiosk photos:** These photos were collected from subjects whose attention was focused on interaction with an immigration kiosk. They images were not intended for use with automated face recognition. The camera is situated above a display which the user touches, and is triggered either without directing the subject to look at it, or without waiting for the subject to comply. The images are therefore characterized by pitch-down pose, sometimes exceeding 45 degrees, as in Figure 4c. Yaw-angle variation is mild, with most images close to frontal. The images

have width 320 pixels and height 240 pixels and therefore tall individuals are sometimes cropped. This is often just above the eyes and can occur at the nose or mouth. Conversely, short individuals are sometimes cropped such that only the top part of the face is visible. In a quite small number of cases, there other subjects standing just behind the primary subject such that algorithms should detect all faces and determine which is the largest and most centered. Background ceiling lighting is often visible and this sometimes leads to under-exposure of the face. When these images are provided as input into the algorithm, they are labeled with the type "wild".

2.2 Law enforcement images

The main mugshot dataset used is referred to as the FRVT 2018 set. This set was collected over the period 2002 to 2017 in routine United States law enforcement operations. This set yields three subsets

- ▷ **Mugshots:** Mugshots comprise about 86% of the database. They have reasonable compliance with the ANSI/NIST ITL1-2011 Type 10 standard's subject acquisition profiles levels 10-20 for frontal images [26]. The most common departure from the standard's requirements is the presence of mild pose variations around frontal - the images of Figure 5 are typical. The images vary in size, with many being 480x600 pixels with JPEG compression applied to produce filesizes of between 18 and 36KB with many images outside this range, implying that about 0.5 bits are being encoded per pixel. When these images are provided as input into the algorithm, they are labeled with the type "mugshot".

Example images appear in Fig. 5

[NIST Interagency Report 8238](#) includes a comparison of this set of mugshots with the smaller and easier sets of mugshots used in tests run in 2010 and 2014.

- ▷ **Profile images:** Profile-view images have been collected in law enforcement for more than 100 years, as human capability is improved with orthogonal information. The profile images used in this report were collected during the same session as the frontal mugshot photograph, in the same standardized photographic setup. These would not therefore be used with automated face recognition. A small subset, 200 000 images, were set aside for testing. When these images are provided as input into the algorithm, they are labeled with the type "wild".

Example images appear in Fig. 7

- ▷ **Webcam images:** The remaining 14% of the images were collected using an inexpensive webcam attached to a flexible operator-directed mount. These images are all of size 240x240 pixels, that are in considerable violation of most quality-related clauses of all face recognition standards. As evident in the figure, the most common defects are non-frontal pose (associated with the rotational degrees of freedom of the camera mount), low contrast (due to varying and intense background lights), and poor spatial resolution (due to inexpensive camera optics) - see examples in Fig 6. The images are overly JPEG compressed, to between 4 and 7KB, implying that only 0.5 to 1 bits are being encoded per color pixel. When these images are provided as input into the algorithm, they are labeled with the type "wild".

Example images appear in Fig. 6

These are drawn from NIST Special Database 32 which may be downloaded [here](#).

These images were partitioned in galleries and probesets for the various experiment listed in Table 1.

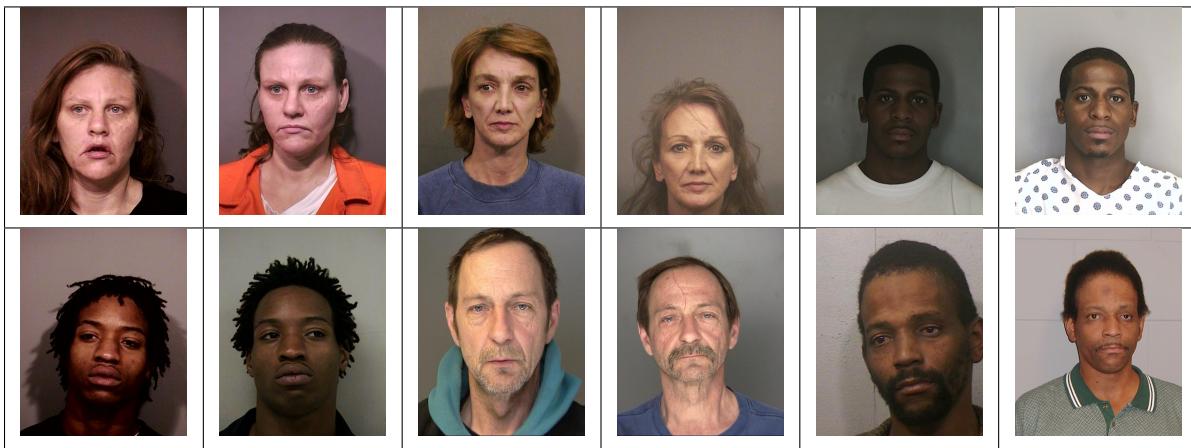


Figure 5: Six mated mugshot pairs representative of the FRVT-2014 (LEO) and FRVT-2018 datasets. The images are collected live, i.e. not scanned from paper. Image source: NIST Special Database 32 the Multiple Encounter Deceased Subjects dataset.



Figure 6: Twelve webcam images representative of probes against the FRVT-2018 mugshot gallery. The first eight images are four mated pairs. Such images present challenges to recognition including pose, non-uniform illumination, low contrast, compression, cropping, and low spatial sampling rate. Image source: NIST Special Database 32 the Multiple Encounter Deceased Subjects dataset.



Figure 7: **[Profile views]** The three images are a frontal enrollment, subsequent frontal probe, and same-session ninety degree profile view. While collection of both frontal and profile views has been typical in law enforcement for more than a century, the recognition of profile to frontal views has essentially been impossible. However, reasonably high accuracy results is now possible - see section E.

Image				
Encounter	1	...	$K_i - 1$	K_i
Capture Time	T_1	...	T_{K_i-1}	T_{K_i}
Role RECENT	Not used	Not used	Enrolled	Search
Role LIFETIME	Enrolled	Enrolled	Enrolled	Search

Figure 8: Depiction of the “recent” and “lifetime” enrollment types. Image source: NIST Special Database 32

2.3 Enrollment strategies

Many operational applications include collection and enrollment of biometric data from subjects on more than one occasion. This might be done on a regular basis, as might occur in credential (re-)issuance, or irregularly, as might happen in a criminal recidivist situation [4]. The number of images per person will depend on the application area. In civil identity credentialing (e.g. passports, driver’s licenses), the images will be acquired approximately uniformly over time (e.g. ten years for a passport). While the distribution of dates for such images of a person might be assumed uniform, a number of factors might undermine this assumption⁷. In criminal applications, the number of images would depend on the number of arrests. The distribution of dates for arrest records for a person (i.e. the recidivism distribution) has been modeled using the exponential distribution but is recognized to be more complicated⁸.

In any case, the 2010 NIST evaluation of face recognition showed that considerable accuracy benefits accrue with retention and use of *all* historical images [6].

To this end, the FRVT API document provides $K \geq 1$ images of an individual to the enrollment software. The software is tasked with producing a single proprietary undocumented “black-box” template⁹ from the K images. This affords the algorithm an ability to generate a *model* of the individual, rather than to simply extract features from each image on a sequential basis.

As depicted in Figure 8, the i -th individual in the FRVT 2018 dataset has K_i images. These are labelled as x_k for $k = 1 \dots K_i$ in chronological order of capture date. To measure the utility of having multiple enrollment images, this report evaluates three kinds of enrollment:

- ▷ **Recent:** Only the second most recent image, x_{K_i-1} is enrolled. This strategy of enrollment mimics the operational policy of retaining the imagery from the most recent encounter. This might be done operationally to ameliorate the effects of face ageing. Obviously retaining only the most recent image should only be done if the identity of the person is trusted to be correct. For example, in an access control situation retention of the most recent successful *authentication* image would be hazardous if it could be a false positive.
- ▷ **Lifetime-consolidated:** All but the most recent image are enrolled, $x_1 \dots x_{K_i-1}$. This subject-centric strategy might be adopted if quality variations exist where an older image might be more suitable for matching, despite the ageing effect.

⁷For example, a person might skip applying for a passport for one cycle, letting it expire. In addition, a person might submit identical images (from the same photography session) to consecutive passport applications at five year intervals.

⁸A number of distributions have been considered to model recidivism, see for example [3].

⁹There are no formal face template standards. Template standards only exist for fingerprint minutiae - see ISO/IEC 19794-2:2011.

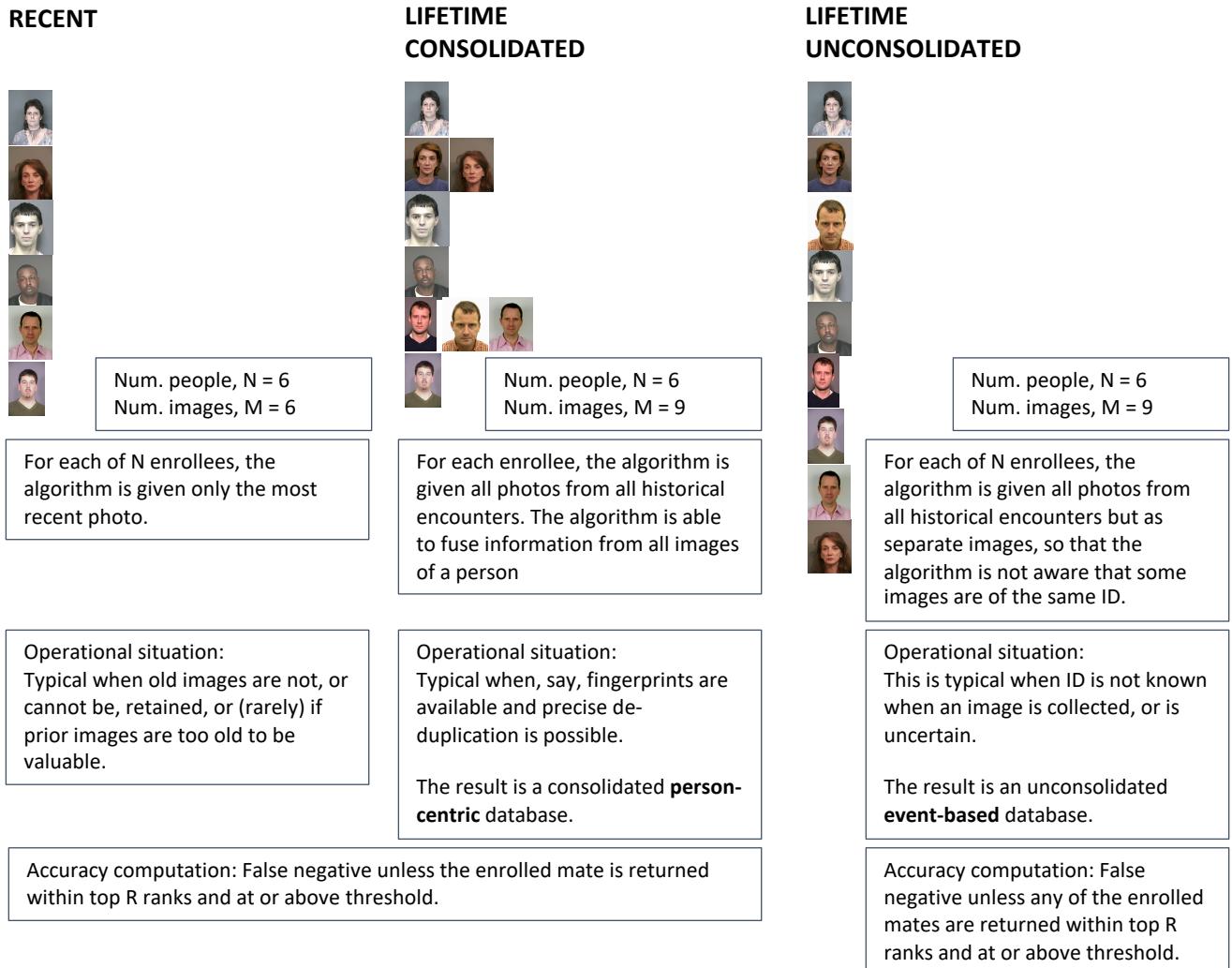


Figure 9: Enrollment strategies. The figure shows the three kinds of enrollment databases examined in this report. Image source: NIST Special Database 32

	ENROLLMENT				SEARCH			
	TYPE SEE SECTION 2.3	POPULATION FILTER	N-SUBJECTS	N-IMAGES	MATE N-SUBJECTS	NON-MATE N-IMAGES	N-SUBJECTS	N-IMAGES
Mugshot trials from enrollment of single images								
1	RECENT	NATURAL	640 000	640 000	154 549	154 549	331 254	331 254
2	RECENT	NATURAL	1 600 000	1 600 000				
3	RECENT	NATURAL	3 000 000	3 000 000				
4	RECENT	NATURAL	6 000 000	6 000 000				
5	RECENT	NATURAL	12 000 000	12 000 000				
Cross-domain								
13	MUGSHOTS AS ON ROW 2				82 106 WEBCAM	82 106 WEBCAM	331 254 WEBCAM	331 254 WEBCAM
Cross-view								
14	MUGSHOTS AS ON ROW 2				100 000 PROFILE	100 000 PROFILE	100 000 PROFILE	100 000 PROFILE
Mugshot ageing								
17	OLDEST	NATURAL	3 068 801	3 068 801	2 853 221	10 951 064	0	0
Border crossing ageing								
17	OLDEST	NATURAL	1 600 000	1 600 000	1 922 437	1 922 437	1 920 000	1 920 000
Visa-border								
19	PRIOR	NATURAL	1 600 000 VISA	1 600 000 VISA	80 000 BORDER	80 000 BORDER	80 000 BORDER	80 000 BORDER
20	VISA AS ON ROW 18				21 016 BORDER	21 016 BORDER	21 016 BORDER	21 016 BORDER

Table 1: Enrollment and search sets. Each row summarizes one identification trial. Unless stated otherwise, all entries refer to mugshot images. The term “natural” means that subjects were selected without heed to demographics, i.e. in the distribution native to this dataset. The probe images were collected in a different calendar year to the enrollment image. Missing values in rows 2-12 are the same as in row 1.

▷ **Lifetime-unconsolidated:** Again all but the most recent image are enrolled $x_1 \dots x_{K_i-1}$ but now separately, with different identifiers, such that the algorithm is not aware that the images are from the same face. This kind of event- or encounter-centric enrollment is very common when operational constraints preclude reliable consolidation of the historical encounters into a single identity. This aspect also prevents the recognition algorithm from a) building a holistic model of identity (as is common in speaker recognition systems) and b) implementing fusion, for example template-level fusion of feature vectors, or post-search score-level fusion. The result is that searches will typically yield more than one image of a person in the top ranks. This has consequences for appropriate metrics, as detailed in section 3.2.1

NIST first evaluated this kind of enrollment in mid 2018, and the results tables include some comparison of accuracy available from all three enrollment styles.

In all cases, the most recent image, x_{K_i} , is reserved as the search image. For the 1.6 million subject enrollment partition of the FRVT 2018 data, $1 \leq K_i \leq 33$ with $K_i = 1$ in 80.1% of the individuals, $K_i = 2$ in 13.4%, $K_i = 3$ in 3.7%, $K_i = 4$ in 1.4%, $K_i = 5$ in 0.6%, $K_i = 6$ in 0.3%, and $K_i > 6$ is 0.2% for everyone else. This distribution is substantially dependent on United States recidivism rates.

We did not evaluate the case of retaining only the highest quality image, since automated quality assessment is out of scope for this report. We do not anticipate that such strategies will prove beneficial when the quality assessment apparatus is imperfect and unvalidated.

3 Performance metrics

This section gives specific definitions for accuracy and timing metrics. Tests of open-set biometric algorithms must quantify frequency of two error conditions:

- ▷ **False positives:** Type I errors occur when search data from a person who has never been seen before is incorrectly associated with one or more enrollees' data.
- ▷ **Misses:** Type II errors arise when a search of an enrolled person's biometric does not return the correct identity.

Many practitioners prefer to talk about "hit rates" instead of "miss rates" - the first is simply one minus the other as detailed below. Sections 3.1 and 3.2 define metrics for the Type I and Type II performance variables.

Additionally, because recognition algorithms sometimes fail to produce a template from an image, or fail to execute a one-to-many search, the occurrence of such events must be recorded. Further because algorithms might elect to not produce a template from, for example, a poor quality image, these failure rates must be combined with the recognition error rates to support algorithm comparison. This is addressed in section 3.5.

Finally, section 3.7 discusses measurement of computation duration, and section 3.8 addresses the uncertainty associated with various measurements. Template size measurement is included with the results.

3.1 Quantifying false positives

It is typical for a search to be conducted into an enrolled population of N identities, and for the algorithm to be configured to return the closest L candidate identities. These candidates are ranked by their score, in descending order, with all scores required to be greater than or equal to zero. A human analyst might examine either all L candidates, or just the top $R \leq L$ identities, or only those with score greater than threshold, T . The workload associated with such examination is discussed later, in 3.6.

False alarm performance is quantified in two related ways. These express how many searches produces false positives, and then, how many false positives are produced in a search.

False positive identification rate: The first quantity, FPIR, is the proportion of non-mate searches that produce an adverse outcome:

$$\text{FPIR}(N, T) = \frac{\text{Num. non-mate searches where one or more enrolled candidates are returned with score at or above threshold}}{\text{Num. non-mate searches attempted.}} \quad (1)$$

Under this definition, FPIR can be computed from the highest non-mate candidate produced in a search - it is not necessary to consider candidates at rank 2 and above. FPIR is the primary measure of Type I errors in this report.

Selectivity: However, note that in any given search, several non-mate may be returned above threshold. In order to quantify such events, a second quantity, selectivity (SEL), is defined as the *number* of non-mates returned on a candidate list, averaged over all searches.

$$\text{SEL}(N, T) = \frac{\text{Num. non-mate enrolled candidates returned with score at or above threshold}}{\text{Num. non-mate searches attempted.}} \quad (2)$$

where $0 \leq \text{SEL}(N, T) \leq L$. Both of these metrics are useful operationally. FPIR is useful for targeting how often an

adverse false positive outcome can occur, while SEL as a number is related to workload associated with adjudicating candidate lists. The relationship between the two quantities is complicated - it depends on whether an algorithm concentrates the false alarms in the results of a few searches or whether it disburses them across many. This was detailed in FRVT 2014, NISTIR 8009. It has not yet been detailed in FRVT 2018.

3.2 Quantifying hits and misses

If L candidates are returned in a search, a shorter candidate list can be prepared by taking the top $R \leq L$ candidates for which the score is above some threshold, $T \geq 0$. This reduction of the candidate list is done because thresholds may be applied, and only short lists might be reviewed (according to policy or labor availability, for example). It is useful then to state accuracy in terms of R and T , so we define a “miss rate” with the general name **false negative identification rate** (FNIR), as follows:

$$\text{FNIR}(N, R, T) = \frac{\text{Num. mate searches with enrolled mate found outside top } R \text{ ranks or score below threshold}}{\text{Num. mate searches attempted.}} \quad (3)$$

This formulation is simple for evaluation in that it does not distinguish between causes of misses. Thus a mate that is not reported on a candidate list is treated the same as a miss arising from face finding failure, algorithm intolerance of poor quality, or software crashes. Thus if the algorithm fails to produce a candidate list, either because the search failed, or because a search template was not made, the result is regarded as a miss, adding to FNIR.

Hit rates, and true positive identification rates: While FNIR states the “miss rate” as how often the correct candidate is either not above threshold or not at good rank, many communities prefer to talk of “hit rates”. This is simply the **true positive identification rate**(TPIR) which is the complement of FNIR giving a positive statement of how often mated searches are successful:

$$\text{TPIR}(N, R, T) = 1 - \text{FNIR}(N, R, T) \quad (4)$$

This report does not report true positive “hit” rates, preferring false negative miss rates for two reasons. First, costs rise linearly with error rates. For example, if we double FNIR in an access control system, then we double user inconvenience and delay. If we express that as decrease of TPIR from, say 98.5% to 97%, then we mentally have to invert the scale to see a doubling in costs. More subtly, readers don’t perceive differences in numbers near 100% well, becoming inured to the “high nineties” effect where numbers close to 100 are perceived indifferently.

Reliability is a corresponding term, typically being identical to TPIR, and often cited in automated (fingerprint) identification system (AFIS) evaluations.

An important special case is the **cumulative match characteristic**(CMC) which summarizes accuracy of mated-searches only. It ignores similarity scores by relaxing the threshold requirement, and just reports the fraction of mated searches returning the mate at rank R or better.

$$\text{CMC}(N, R) = 1 - \text{FNIR}(N, R, 0) \quad (5)$$

We primarily cite the complement of this quantity, $\text{FNIR}(N, R, 0)$, the fraction of mates *not* in the top R ranks.

The **rank one hit rate** is the fraction of mated searches yielding the correct candidate at best rank, i.e. $\text{CMC}(N, 1)$. While this quantity is the most common summary indicator of an algorithm’s efficacy, it is not dependent on similarity scores, so it does not distinguish between strong (high scoring) and weak hits. It also ignores that an adjudicating reviewer is often willing to look at many candidates.

3.2.1 False negative rates for unconsolidated galleries

As detailed in section 2.3 a common type of gallery, here referred to as the lifetime unconsolidate type, is populated with all images of an individual without any association between them. That is, the gallery construction algorithm is not provided with any ID labels that would support processing of a person's images jointly. This contrasts with the lifetime consolidate type where an algorithm may explicitly fuse features from multiple images of a person, or select a best image. In such cases, where the number of enrolled images is a random variable, we define two false negative rates as follows.

The first demands that the algorithm place any of the K_i mates in the top $R \geq 1$ ranks. The proportion of searches for which this does not occur forms a false negative identification rate:

$$\text{FNIR}_{\text{any}}(N, R, T) = 1 - \frac{\text{Num. mate searches where any enrolled mate is found in the top } R \text{ ranks and at-or-above threshold}}{\text{Num. mate searches attempted.}} \quad (6)$$

The second demands that the algorithm place all K_i mates in the top $R \geq K_i$ ranks. The proportion of searches for which this does not occur forms a false negative identification rate:

$$\text{FNIR}_{\text{all}}(N, R, T) = 1 - \frac{\text{Num. mate searches where all enrolled mates are found in the top } R \text{ ranks and at-or-above threshold}}{\text{Num. mate searches attempted.}} \quad (7)$$

Placing all mates in the top ranks is a more difficult task than correctly retrieving any image, so it holds that: $\text{FNIR}_{\text{all}} \geq \text{FNIR}_{\text{any}}$. This is evident in the results presented for November 2018 algorithms in Tables starting at ??.

The information retrieval community might prefer to compute and plot *precision* and *recall*; this is a valid approach, but we advance the two metrics above because they relate to our normal definition of consolidated FNIR, and they cover the two extreme use-cases of wanting any hit vs. all hits.

3.3 DET interpretation

In biometrics, a false negative occurs when an algorithm fails to match two samples of one person – a Type II error. Correspondingly, a false positive occurs when samples from two persons are improperly associated – a Type I error.

Matches are declared by a biometric system when the native comparison score from the recognition algorithm meets some threshold. Comparison scores can be either similarity scores, in which case higher values indicate that the samples are more likely to come from the same person, or dissimilarity scores, in which case higher values indicate different people. Similarity scores are traditionally computed by fingerprint and face recognition algorithms, while dissimilarities are used in iris recognition. In some cases, the dissimilarity score is a distance possessing metric properties. In any case, scores can be either mate scores, coming from a comparison of one person's samples, or nonmate scores, coming from comparison of different persons' samples.

The words "genuine" or "authentic" are synonyms for mate, and the word "impostor" is used as a synonym for non-mate. The words "mate" and "nonmate" are traditionally used in identification applications (such as law enforcement search, or background checks) while genuine and impostor are used in verification applications (such as access control).

An error tradeoff characteristic represents the tradeoff between Type II and Type I classification errors. For identification this plots false negative vs. false positive identification rates i.e. FNIR vs. FPIR parametrically with T. Such plots

are often called detection error tradeoff (DET) characteristics or receiver operating characteristic (ROC). These serve the same function – to show error tradeoff – but differ, for example, in plotting the complement of an error rate (e.g. $TPIR = 1 - FNIR$) and in transforming the axes, most commonly using logarithms, to show multiple decades of FPIR. More rarely, the function might be the inverse of the Gaussian cumulative distribution function.

The slides of Figures 10 through 15 discuss presentation and interpretation of DETs used in this document for reporting face identification accuracy. Further detail is provided in formal biometrics testing standards, see the various parts of ISO/IEC 19795 Biometrics Testing and Reporting. More terms, including and beyond those to do with accuracy, appear in ISO/IEC 2382-37 Information technology – Vocabulary – Part 37: Harmonized biometric vocabulary.

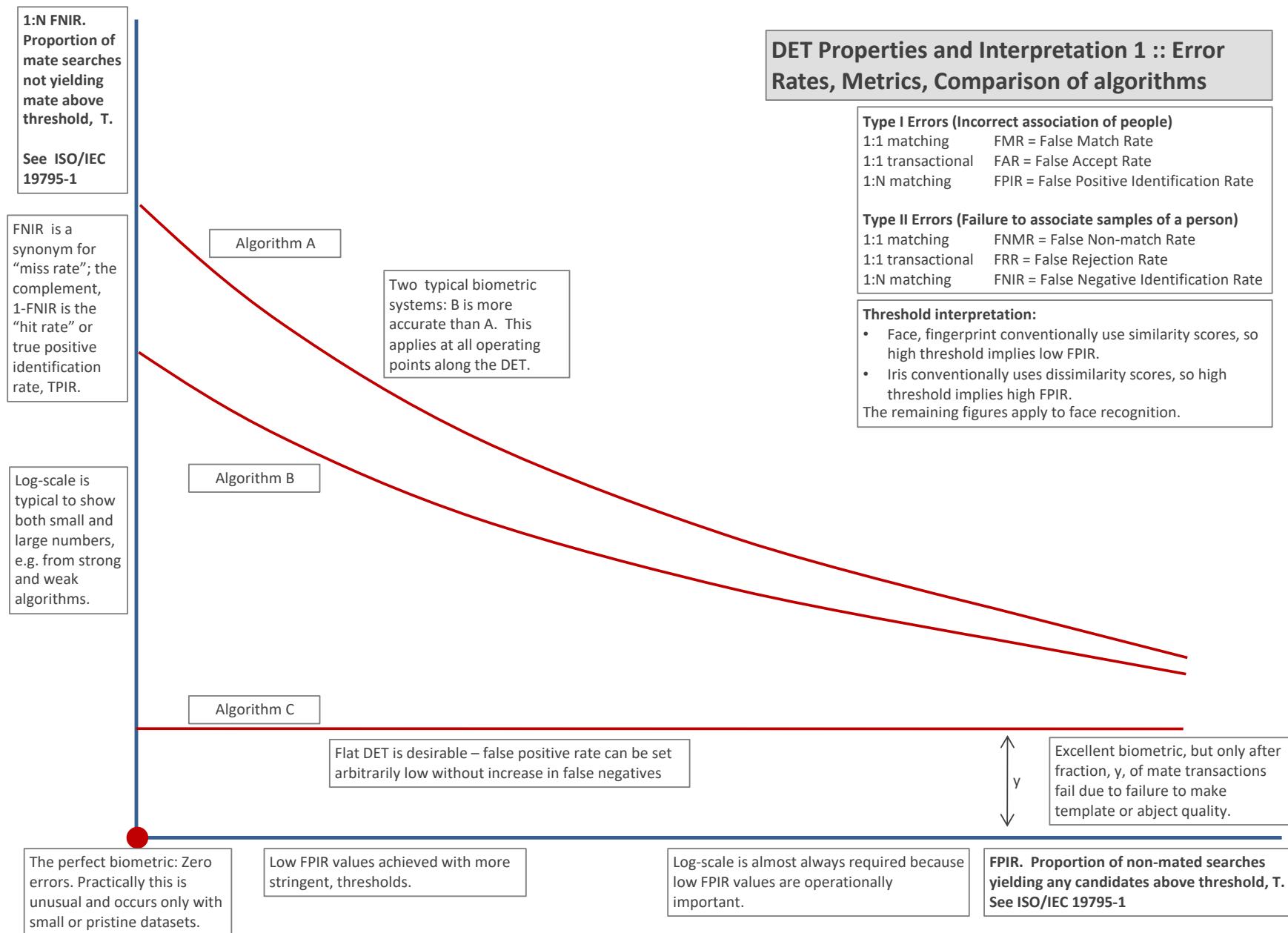


Figure 10: DET as the primary performance reporting mechanism.

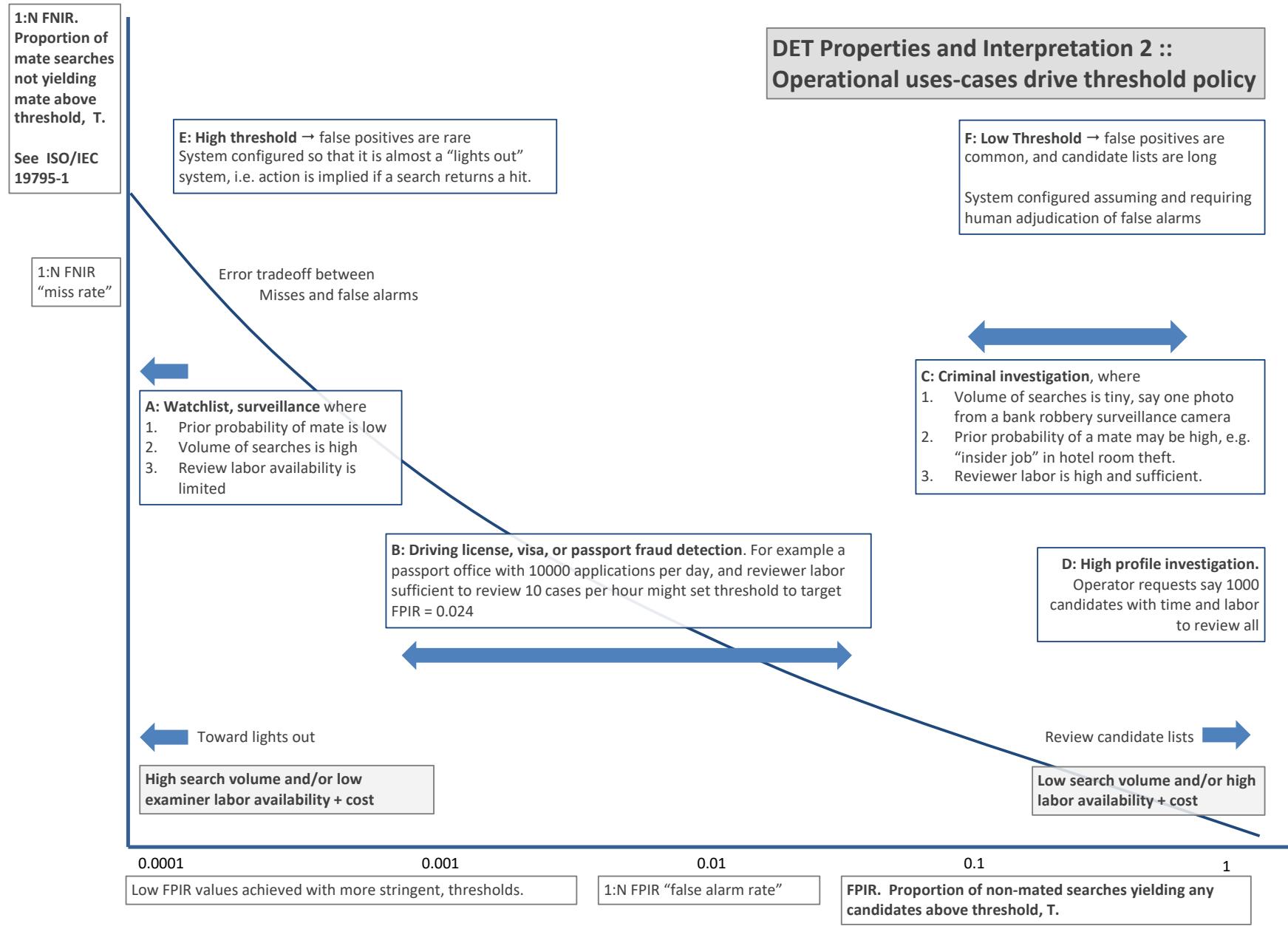
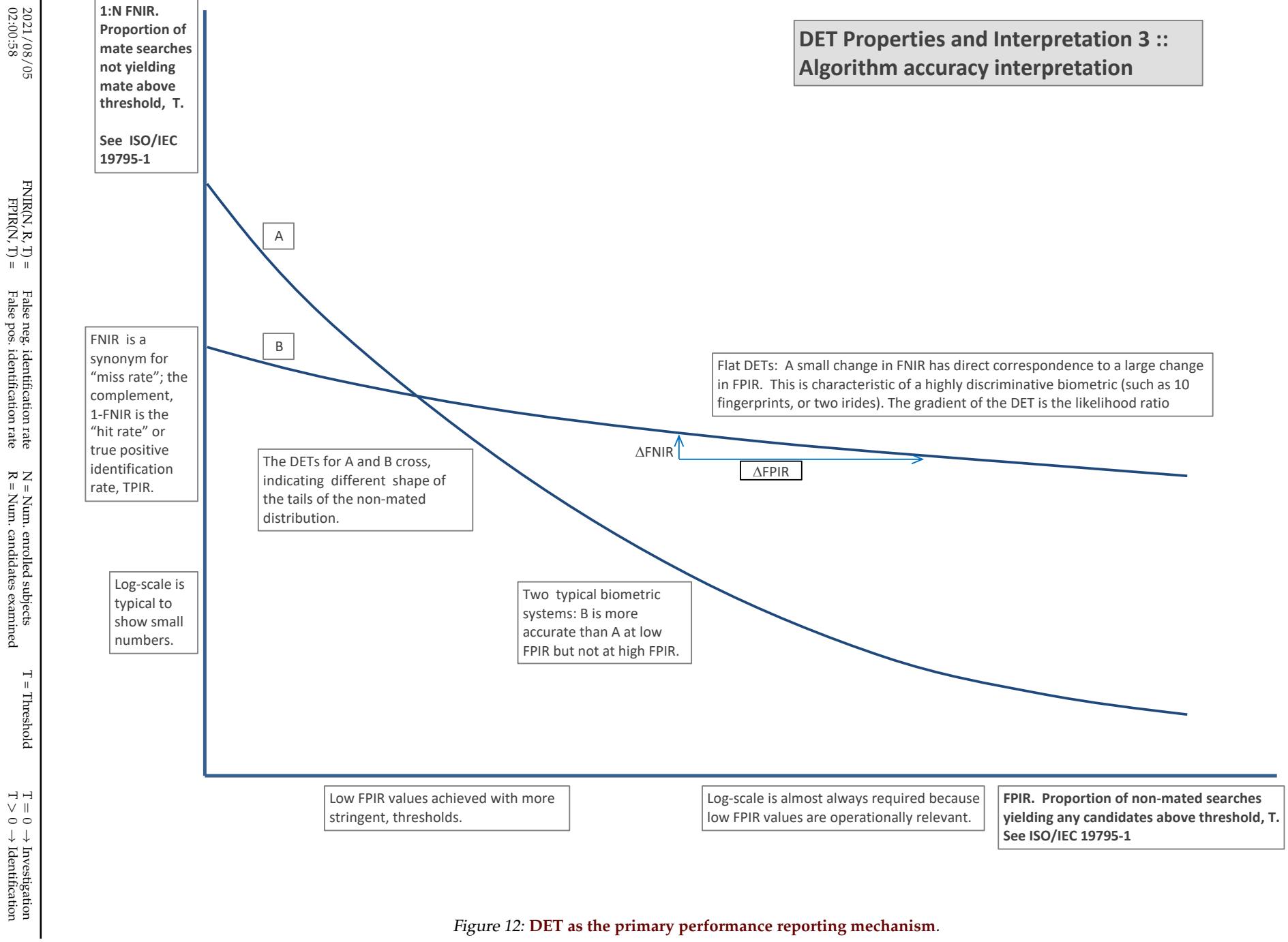
2021/08/05
02:00:58FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = ThresholdT = 0 → Investigation
T > 0 → Identification

Figure 11: DET as the primary performance reporting mechanism.



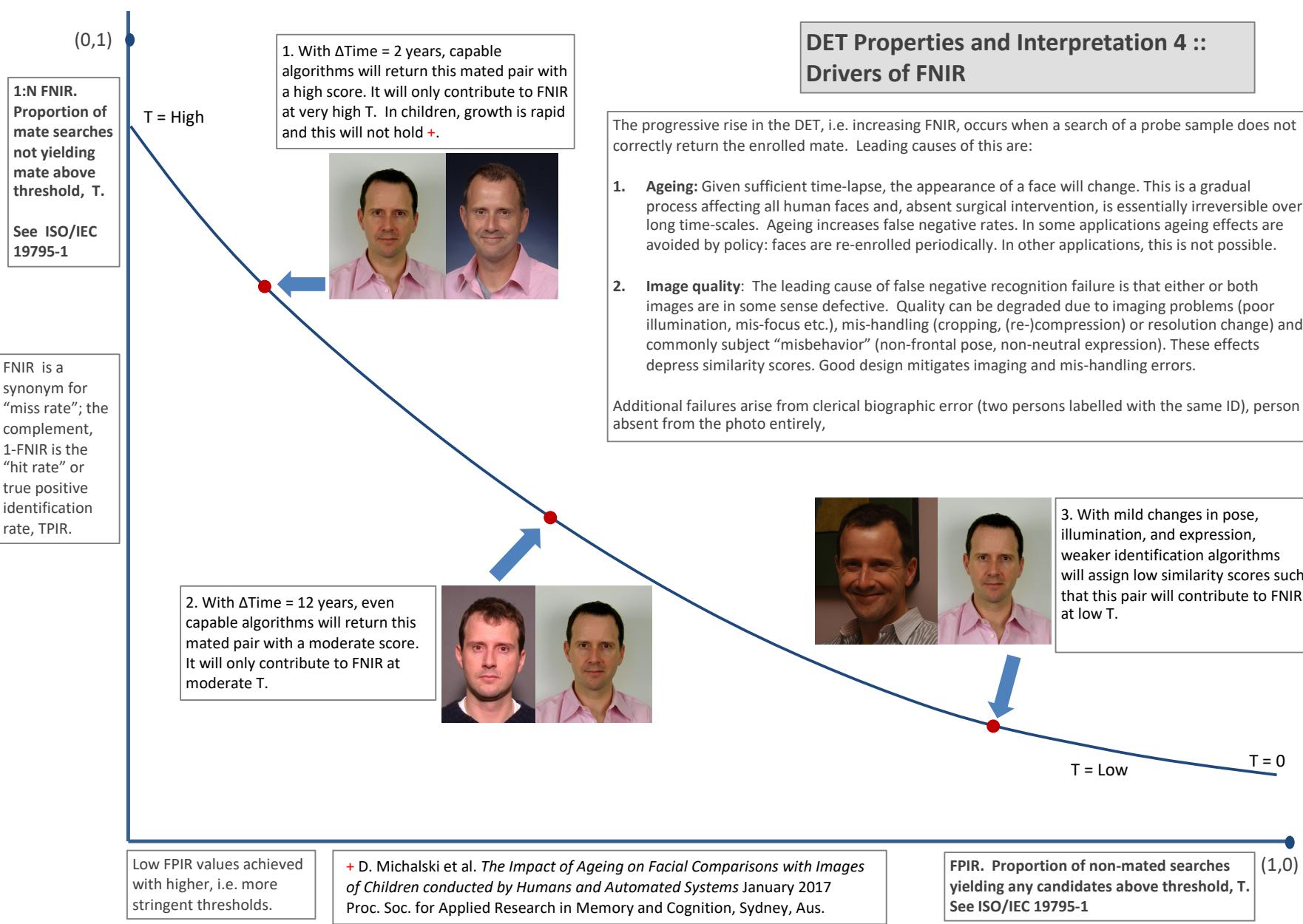


Figure 13: DET as the primary performance reporting mechanism.

2021/08/05
02:00:58

$\text{FNIR}(N, R, T) =$ False neg. identification rate
 $\text{FPIR}(N, T) =$ False pos. identification rate

$N =$ Num. enrolled subjects
 $R =$ Num. candidates examined

$T =$ Threshold
 $T = 0 \rightarrow$ Investigation
 $T > 0 \rightarrow$ Identification

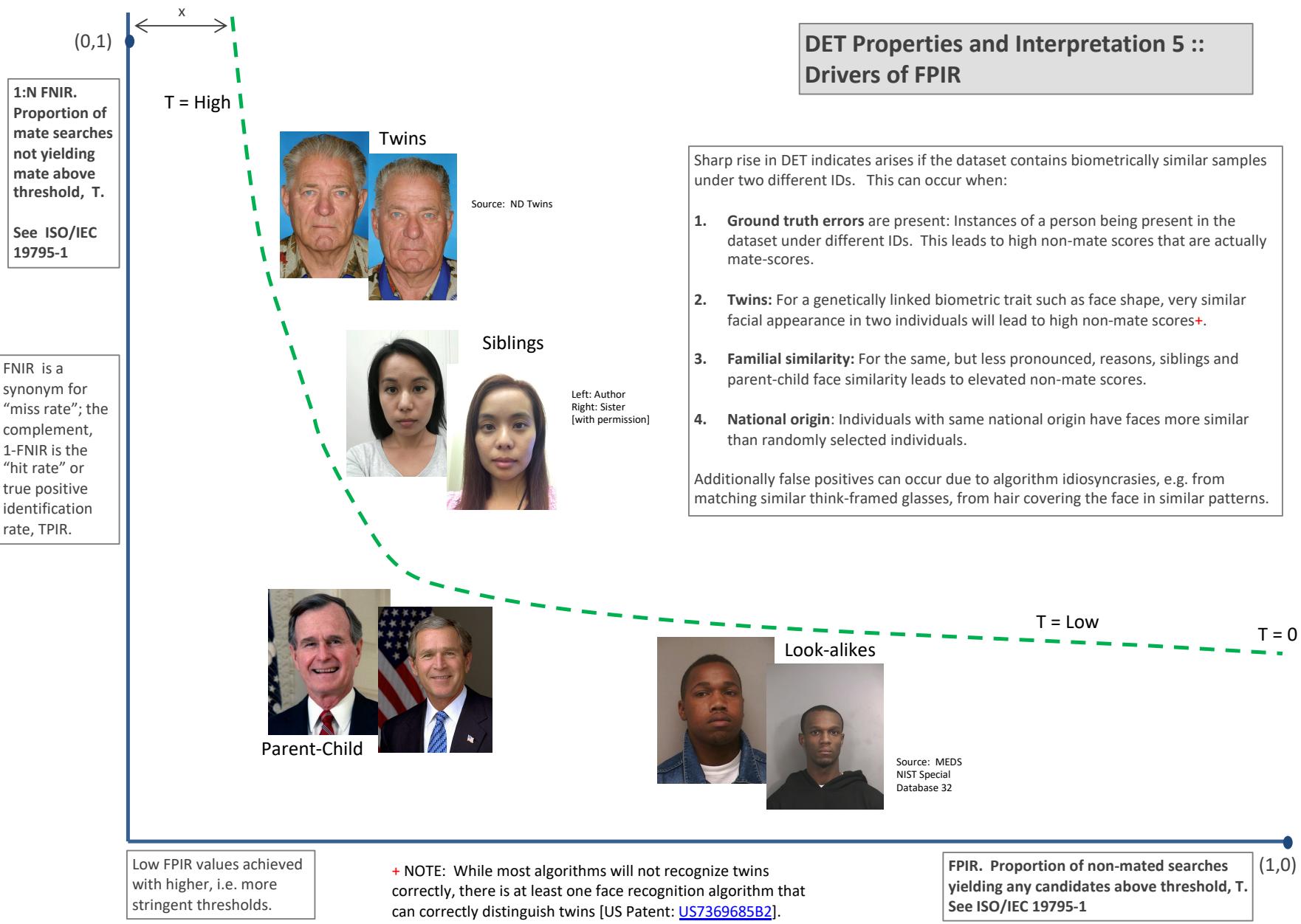


Figure 14: DET as the primary performance reporting mechanism.

2021/08/05
02:00:58

$\text{FNIR}(N, R, T) =$
False neg. identification rate
 $\text{FPIR}(N, T) =$
False pos. identification rate

$N = \text{Num. enrolled subjects}$
 $R = \text{Num. candidates examined}$

$T = \text{Threshold}$

$T = 0 \rightarrow \text{Investigation}$
 $T > 0 \rightarrow \text{Identification}$

DET Properties and Interpretation 6 :: Fixed thresholds, change in image properties or demographics

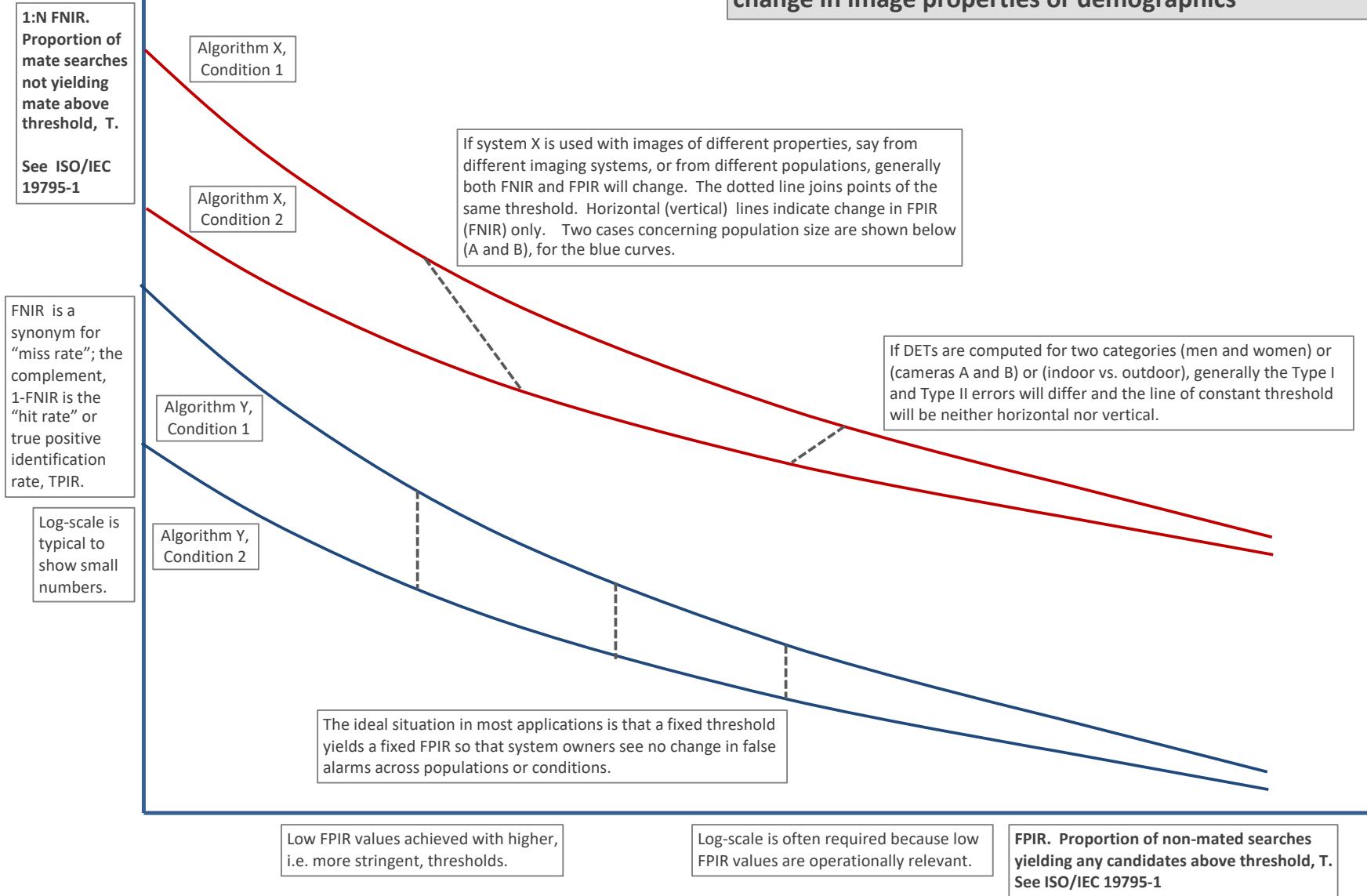


Figure 15: DET as the primary performance reporting mechanism.

1:N FNIR.
Proportion of mate searches not yielding mate above threshold, T.
See ISO/IEC 19795-1

FNIR is a synonym for "miss rate"; the complement, 1-FNIR is the "hit rate" or true positive identification rate, TPIR.

Log-scale is typical to show small numbers.

A: Typical case: In theory, and often in practice, a 1:N search is implemented by executing N 1:1 comparisons independently and then sorting by similarity score:

Mate scores: A mate comparison score is independent of the rest of enrollment data, and so independent of N. This implies the horizontal line above $\text{FNIR}(T, N) = \text{FNMR}(T, 1)$.

Non-mate scores: FPIR increases linearly with N from binomial theory: $\text{FPIR}(N, T) = 1 - (1 - \text{FMR}(T))^N \rightarrow N \text{ FMR}(T)$ for small FPIR.

Pop. N1

Pop. N2 > N1

B: Special case: An enrollment database is not just a linear data structure, it could be an index, or tree, then search is not simply N 1:1 comparisons and a sort. In that case:

Mate scores become dependent on the enrollment data, either its size or actual content, then generally $\text{FNIR}(T, N) \neq \text{FNIR}(T, 1)$.

Non-mate scores are normally no longer just the highest 1:1 comparison score. Instead, for example, scores may be normalized as the implementation attempts to make FPIR independent of N will yield the vertical line linking points of equal threshold.

Low FPIR values achieved with higher, i.e. more stringent, thresholds.

Log-scale is often required because low FPIR values are operationally important.

DET Properties and Interpretation 7 :: Effect of enrolled population size.

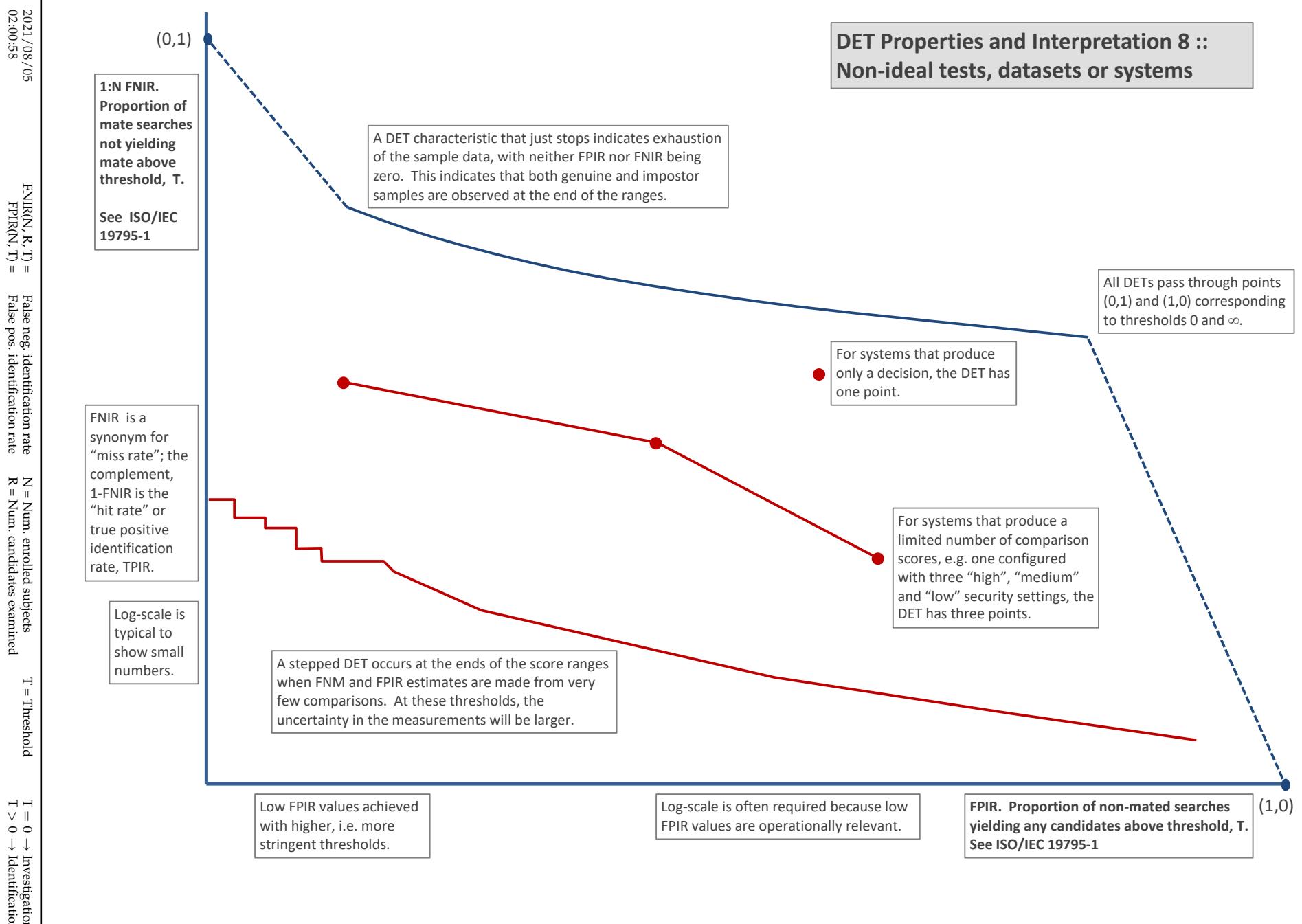


Figure 17: DET as the primary performance reporting mechanism.

3.4 Best practice testing requires execution of searches with and without mates

FRVT embeds 1:N searches of two kinds: Those for which there is an enrolled mate, and those for which there is not. The respective numbers for these types of searches appear in Table 1. However, it is common to conduct only mated searches¹⁰. The cumulative match characteristic is computed from candidate lists produced in mated searches. Even if the CMC is the only metric of interest, the actual trials executed in a test should nevertheless include searches for which no mate exists. As detailed in Table 1 the FRVT reserved disjoint populations of subjects for executing true non-mate searches.

3.5 Failure to extract features

During enrollment some algorithms fail to convert a face image to a template. The proportion of failures is the failure-to-enroll rate, denoted by FTE. Similarly, some search images are not converted to templates. The corresponding proportion is termed failure-to-extract, denoted by FTX.

We do not report FTX because we assume that the same underlying algorithm is used for template generation for enrollment and search.

Failure to extract rates are incorporated into FNIR and FPIR measurements as follows.

- ▷ **Enrollment templates:** Any failed enrollment is regarded as producing a zero length template. Algorithms are required by the API [10] to transparently process zero length templates. The effect of template generation failure on search accuracy depends on whether subsequent searches are mated, or non-mated: Mated searches will fail giving elevated FNIR; non-mated searches will not produce false positives so, to first order, FPIR will be reduced by a factor of $1 - \text{FTE}$.
- ▷ **Search templates and 1:N search:** In cases where the algorithm fails to produce a search template from input imagery, the result is taken to be a candidate list whose entries have no hypothesized identities and zero score. The effect of template generation failure on search accuracy depends on whether searches are mated, or non-mated: Mated searches will fail giving elevated FNIR; Non-mated searches will not produce false positives, so FPIR will be reduced. Thus given a measurement of false negative and positive rates made over only those where failures-to-extract did not occur, those rates - call them FNIR^\dagger and FPIR^\dagger - could be adjusted by an explicit measurement of FTX as follows

$$\text{FNIR} = \text{FTX} + (1 - \text{FTX})\text{FNIR}^\dagger \quad (8)$$

$$\text{FPIR} = (1 - \text{FTX})\text{FPIR}^\dagger \quad (9)$$

This approach is the correct treatment for positive-identification applications such as access control where cooperative users are enrolled and make attempts at recognition. This approach is not appropriate to negative identification applications, such as visa fraud detection, in which hostile individuals may attempt to evade detection by submitting poor quality samples. In those cases, template generation failures should be investigated as though a false alarm had occurred.

¹⁰For example, the [Megaface benchmark](#). This is bad practice for several reasons: First, if a developer knows, or can reasonably assume, that a mate always exists, then unrealistic gaming of the test is possible. A second reason is that it does not put FPIR on equal footing with FNIR and that matters because in most applications, not all searches have mates - not everyone has been previously enrolled in a driving license issuance or a criminal justice system - so addressing between-class separation becomes necessary.

3.6 Fixed length candidate lists, threshold independent workload

Suppose an automated face identification algorithm returns L candidates, and a human reviewer is retained to examine up to R candidates, where $R \leq L$ might be set by policy, preference or labor availability. For now, assume also that the reviewer is not provided with, or ignores, similarity scores, and thresholds are not applied. Given the algorithm typically places mates at low (good) ranks, the number of candidates a reviewer can be expected to review can be derived as follows. Note that the reviewer will:

- ▷ Always inspect the first ranked image Frac. reviewed = 1
- ▷ Then inspect those candidates where mate not confirmed at rank 1 Frac. reviewed = 1-CMC(1)
- ▷ Then inspect those candidates where mate not confirmed at rank 1 or 2 Frac. reviewed = 1-CMC(2)

etc. Thus if the reviewer will stop after a maximum of R candidates, the expected number of candidate reviews is

$$M(R) = 1 + (1 - CMC(1)) + (1 - CMC(2)) + \dots + (1 - CMC(R - 1)) \quad (10)$$

$$= R - \sum_{r=1}^{R-1} CMC(r) \quad (11)$$

A recognition algorithm that front-loads the cumulative match characteristic will offer reduced workload for the reviewer. This workload is defined only over the searches for which a mate exists. In the cases where there truly is no mate, the reviewer would review all R candidates. Thus, if the proportion of searches for which a mate does exist is β , which in the law enforcement context would be the recidivism rate [3], the full expression for workload becomes:

$$M(R) = \beta \left(R - \sum_{r=1}^{R-1} CMC(r) \right) + (1 - \beta)R \quad (12)$$

$$= R - \beta \sum_{r=1}^{R-1} CMC(r) \quad (13)$$

3.7 Timing measurement

Algorithms were submitted to NIST as implementations of the application programming interface(API) specified by NIST in the Evaluation Plan [10]. The API includes functions for initialization, template generation, finalization, search, gallery insert, and gallery delete. Two template generation functions are required, one for the preparation of an enrollment template, and one for a search template.

In NIST's test harness, all functions were wrapped by calls to the C++ std::chrono::high_resolution_clock which on the dedicated timing machine counts 1ns clock ticks. Precision is somewhat worse than that however.

3.8 Uncertainty estimation

3.8.1 Random error

This study leverages operational datasets for measurement of recognition error rates. This affords several advantages. First, large numbers of searches are conducted (see Table 1) giving precision to the measurements. Moreover, for the two mugshot datasets, these do not involve reuse of individuals so binomial statistics can be expected to apply to recognition error counts. In that case, an observed count of a particular recognition outcome (i.e. a false negative or false positive) in M trials will sustain 95% confidence that the actual error rate is no larger than some value.

As an example, the minimum number of mugshot searches conducted in this report is $M = 154\,549$, and for an observed FNIR around 0.002, the measurement supports a conclusion that the actual FNIR is no higher than 0.00228 at 99% confidence level. On the false positive side, we tabulate FNIR at FPIR values as low as 0.001. Given estimates based on 331 254 non-mate trials, the actual FPIR values will be below 0.00115 at 99% confidence. In conclusion, large scale evaluation, without reuse of subjects, supports tight uncertainty bounds on the measured error rates.

3.8.2 Systematic error

The FRVT 2018 dataset includes anomalies discovered as a result of inspecting images involved in recognition failures from the most accurate algorithms. Two kinds of failure occur: False negatives (which, for the purpose here, include failures to make templates) and false positives.

False negative errors: We reviewed 600 false negative pairs for which either or both of the leading two algorithms did not put the correct mate in the top 50 candidates. Given 154 549 searches, this number represents 0.39% of the total, resulting in $\text{FNIR} \sim 0.0039$. Of the 600 pairs:

- ▷ **A: Poor quality:** About 20% of the pairs included images of very low quality, often greyscale, low resolution, blurred, low contrast, partially cropped, interlaced, or noisy scans of paper images. Additionally, in a few cases, the face is injured or occluded by bandages or heavy cosmetics.
- ▷ **B: Ground truth identity label bugs:** About 15% of the pairs are not actually mated. We only assigned this outcome when a pair is clearly not mated.
- ▷ **C: Profile views:** About 35% included an image of a profile (side) view of the face, or, more rarely, an image that was rotated 90 degrees in-plane (roll).
- ▷ **D: Tattoos:** About 30% included an image of a tattoo that contained a face image. These arise from mis-labelling in the parent dataset metadata.
- ▷ **E: Ageing:** There is considerable time-lapse between the two captures.

All these estimates are approximate. Of these, the tattoo and mislabelled images can never be matched. These constitute an accuracy floor in the sample implying that FNIR cannot be below 0.0018¹¹. The profile-views, low-quality images, and images with considerable ageing can, in principle, be successfully matched - indeed some algorithms do so - so are not part of the accuracy floor.

¹¹This value is the sum of two partial false negative rates: $\text{FNIR}_B = 0.15 * 0.0039$ plus $\text{FNIR}_D = 0.3 * 0.0039$

For the microsoft-4 algorithm the lowest miss rate from (recent entry in Table 20) is $\text{FNIR}(640\,000, 50, 0) = 0.0018$. This is close to the value estimated from the inspection of misses. It is below the 0.0039 figure because the algorithm does match some profile and poor quality images, that the yitu-2 algorithm does not.

For many tables (e.g. Table 20), the FNIR values obtained for the FRVT-2018 mugshots could be corrected by reducing them by 0.0018. The best values would then be indistinct from zero. The results in this report *were not* adjusted to account for this systematic error.

False positive errors: As shown in Figure 1 and discussed in Figure 14 many of the DET characteristics in this report exhibit a pronounced turn upward at low false positive rates. The shape can be caused by identity labelling errors in the ground truth of a dataset, specifically persons present in the database under two IDs such that some proportion of non-mate pairs are actually mated. To look for such possibilities, we merged the highest 1000 non-mate pairs produced by three different algorithms which resulted in 1839 unique pairs. This constitutes 0.56% of all non-mate searches. We assert that it is *very* difficult for human reviewers to assign the pairs into the following three categories: twins; doppelgangers; or ground-truth errors (instances of the same person under two IDs). Given this difficulty we made no attempt to correct any possible ground truth errors except by removing 57 pairs in the following categories:

- ▷ **A: Profile views:** Thirteen pairs included one or two profile-view images. As described in Figure 165, these can cause false positives.
- ▷ **B: Same-session photographs:** For twelve pairs, the images were identical or trivially altered (e.g. cropped) versions of the same photo. These were present under a different ID likely due to some clerical or procedural mistake.
- ▷ **C: Tattoos of faces:** There were fourteen instances of tattoo photographs that contained faces causing false matches.
- ▷ **D: T-shirt faces:** There were six instances of T-shirt photographs (of Bob Marley and Che Guevara) being detected instead of the face and causing false positives.
- ▷ **E: Background faces:** There were twelve instances of one subject appearing in the background of two otherwise correct portrait photos.

Note we did not remove any images where there was a chance that the pair was actually a different person.

In any case, the results in this report have not been adjusted for this systematic error.

4 Results

This section gives extensive results for algorithms submitted to FRVT 2018. Three page “report cards” for each algorithm are contained in a [separate supplement](#). Performance metrics were described in section 3. The main results are summarized in tabular form with more exhaustive data included as DET, CMC and related graphs in appendices as follows:

- ▷ The three tables 2-4 list algorithms alongside full developer names, acceptance date, size of the provided configuration data, template size and generation time, and search duration data.
 - The **template generation duration** is most important to applications that require fast response. For example, an eGate taking more than two seconds to produce a template might be unacceptable. Note that GPUs may be of utility in expediting this operation for some algorithms, though at additional expense. Two additional factors should be considered¹²¹³.
 - The **search duration** is the time taken for a search of a search template into a gallery of N enrollment templates. This performance variable, together with the volume of searches, is influential on the amount of hardware needed to sustain an operational deployment. This is measured here with the algorithm running on a single core of a contemporary CPU. Search is most simply implemented as N computations of a distance metric followed by a sort operation to find the closest enrollments. However, considerable optimization of this process is possible, up to and including fast-search algorithms that, by various means, avoid computation of all N distances.
 - The **template size** is the size of the extracted feature vector (or vectors) and any needed header information. Large template sizes may be influential on bus or network bandwidth, storage requirements, and on search duration. While the template itself is an opaque data blob, the feature dimensionality might be estimated by assuming a four-bytes-per-float encoding. There is a wide range of encodings. For the more accurate algorithm, sizes range from 256 bytes to about 2KB bytes, indicating essentially no consensus on face modeling and template design.
 - The **template size multiplier** column shows how, given k input images, the size of the template grows. Most implementations internally extract features from each image and concatenate them, and implement some score-level fusion logic during search. Other implementations, including many of the most accurate algorithms, produce templates whose size does not grow with k . This could be achieved via selection of the best quality image - but this is not optimal in handling ageing where the oldest image could be the best quality. Another mechanism would be feature-level fusion where information is fused from all k inputs. In any case, as a black-box test, the fusion scheme is proprietary and unknown.
 - The size of the **configuration data** is the total size of all files resident in a vendor-provided directory that contains arbitrary read-only files such as parameters, recognition models (e.g caffe). Generally a large value for this quantity may prohibit the use of the algorithm on a resource-constrained device.

¹²The FRVT 2018 API prohibited threading, so some gains from parallelism may be available on multiple-cores or multiple processors, if the feature extraction code could be distributed across them.

¹³Note also that factors of two or more may be realizable by exploiting modern vector processing instructions on CPUs. It is not clear in our measurements whether all developers exploited Intel’s AVX2 instructions, for example. Our machine was so equipped, but we insisted that the same compiled library should also run on older machines lacking that instruction. The more sophisticated implementations may have detected AVX2 presence and branched accordingly. The less sophisticated may be defaulted to the reduced instruction set. Readers should see the FRVT 2018 API document for the specific chip details.

▷ Tables 20-21 report core rank-based accuracy for mugshot images. The population size is limited to $N = 1.6$ million identities because this is the largest gallery size on which all algorithms were executed. Notable observations from these tables are as follows:

- **Accuracy gains since 2018:** NIST Interagency Report 8238 documented massive gains over those reported in the FRVT 2014 report, NIST Interagency Report 8009. Further gains are documented in this report. Comparing the most accurate algorithm in November 2018, NEC-3, the value of $\text{FNIR}(N, L, T)$ reduced from 0.0031 to 0.0024 for the Sensetime-004 algorithm with $N = 12$ million recent images. The tables show broader gains: many developers have made advances since 2018 with between two and five-fold reduction in errors.
 - **Wide range in accuracy:** The rank-1 miss rates vary from $\text{FNIR}(N, 1, 0) = 0.0012$ for sensetime-004 up to about 0.5 for the very fast but inaccurate microfocus-x algorithms. Among the developers who are superior to NEC in 2013, the range is from 0.002 to 0.035 for camvi-3. This large accuracy range is consistent with the buyer-beware maxim, and indicates that face recognition software is far from being commoditized.
- ▷ Tables 24-25 report threshold-based error rates, $\text{FNIR}(N, L, T)$, for $N = 1.6$ million for mugshot-mugshot accuracy on FRVT 2014, FRVT 2018, and also (in pink) mugshot-webcam accuracy using FRVT 2018 enrollments. Notable observations from these tables are as follows:
- **Order of magnitude accuracy gains since 2014:** As with rank-based results, the gains in accuracy are substantial, though somewhat reduced. At $\text{FPIR} = 0.01$, the best improvement over NEC in 2014 is a 27 fold reduction in FNIR using the NEC_2 algorithm. At $\text{FPIR} = 0.001$, the largest gain is a six-fold reduction in FNIR via the NEC_3 algorithm.
 - **Broad gains across the industry:** About 19 companies realize accuracy better than the NEC benchmark from 2014. This is somewhat lower than the 28 developers who succeeded on the rank-1 metric. This may be due to the ubiquity of, and emphasis on, the rank-1 metric in many published algorithm development papers.
 - **Webcam images:** Searches of webcam images give $\text{FNIR}(N, T)$ values around 2 to 3 times higher than mugshot searches. Notably the leading developers with mugshots are approximately the same with poorer quality webcams. But some developers e.g. Camvi, Megvii, TongYi, and Neurotechnology do improve their relative rankings on webcams, perhaps indicating their algorithms were tailored to less constrained images.
- ▷ Tables 14, 17, 18 and show, respectively, high-threshold, rank 1, and rank 50 FNIR values for all algorithms performing searches into five different gallery sizes, $N = 640\,000$, $N = 1\,600\,000$, $N = 3\,000\,000$, $N = 6\,000\,000$ and $12\,000\,000$. The $\text{FPIR} = 0.001$ table is included to inform high-volume duplicate detection applications. The Rank-1 table is included as a primary accuracy indicator. The Rank-50 table is included to inform agencies who routinely produce 50 candidates for human-review. The notable results are:

- **Slow growth in rank-based miss rates:** $\text{FNIR}(N, R)$ generally grows as a power law, aN^b . From the straight lines of many graphs of Figure 20 this is clearly a reasonable model for most, but not all, algorithms. The coefficient a can be interpreted as FNIR in a gallery of size 1. The more important coefficient b indicates scalability, and often, $b \ll 1$, implies very benign growth in FNIR . The coefficients of the models appear in the Tables 17 and 18.
- **Slow growth in threshold-based miss rates:** $\text{FNIR}(N, T)$ also generally grows as a power law, aN^b except at the high threshold values corresponding to low FPIR values. This is visible in the plots of Figure 36 which

show straight lines except for $FPIR = 0.001$, which increase more rapidly with N above 3 000 000. Each trace in those figures shows $FNIR(N, T)$ at fixed $FPIR$ with both N and T varying. Thus at large N , it is usually necessary to elevate T to maintain fixed $FPIR$. This causes increased $FNIR$. Why that would no-longer obey a power-law is not known. However, if we expect large galleries to contain individuals with familial relations to the non-mate search images - in the most extreme case, twins - then suppression of false positives becomes more difficult. This is discussed in the Figures starting at Fig. 10

▷ Figure ?? shows false positives from twins against their enrolled siblings, broken out by type of twin: fraternal or identical. The Figure is based on the enrollment of 104 single images on one of a pair of twins, and then the search of 2354 second images. Note that the dataset is heavily skewed towards identical twins which is not representative of the true population. There is also a skew towards same sex fraternal twin pairs compared to different sex fraternal twin pairs again not representative of the true population.

The notable results are:

- For all algorithms tested, the 1087 mated searches (Twin A vs. Twin A) produce scores almost always above typical operational thresholds, with (not shown) matches at rank 1. The images are of good quality, so this is the result expected from the rest of this report.
- For the 1066 identical twin searches (AB), almost all produce the twin at rank 1, with a few producing the mate at further down the candidate lists rank and low score.
- For the 169 fraternal searches (AB) from same sex pairs, most algorithms give a large number of very high scores, implying false positives at all thresholds. However, there are long tails containing lower scores that are correctly below threshold. In general, scores that are higher in this distribution are all rank 1 whereas the lower scores have much higher ranks.
- (Not shown) Of the 169, there are 24 fraternal searches (AB) involving different sex twins. Here most algorithms correctly report scores well below the lowest threshold, and usually not on the candidate list at all.

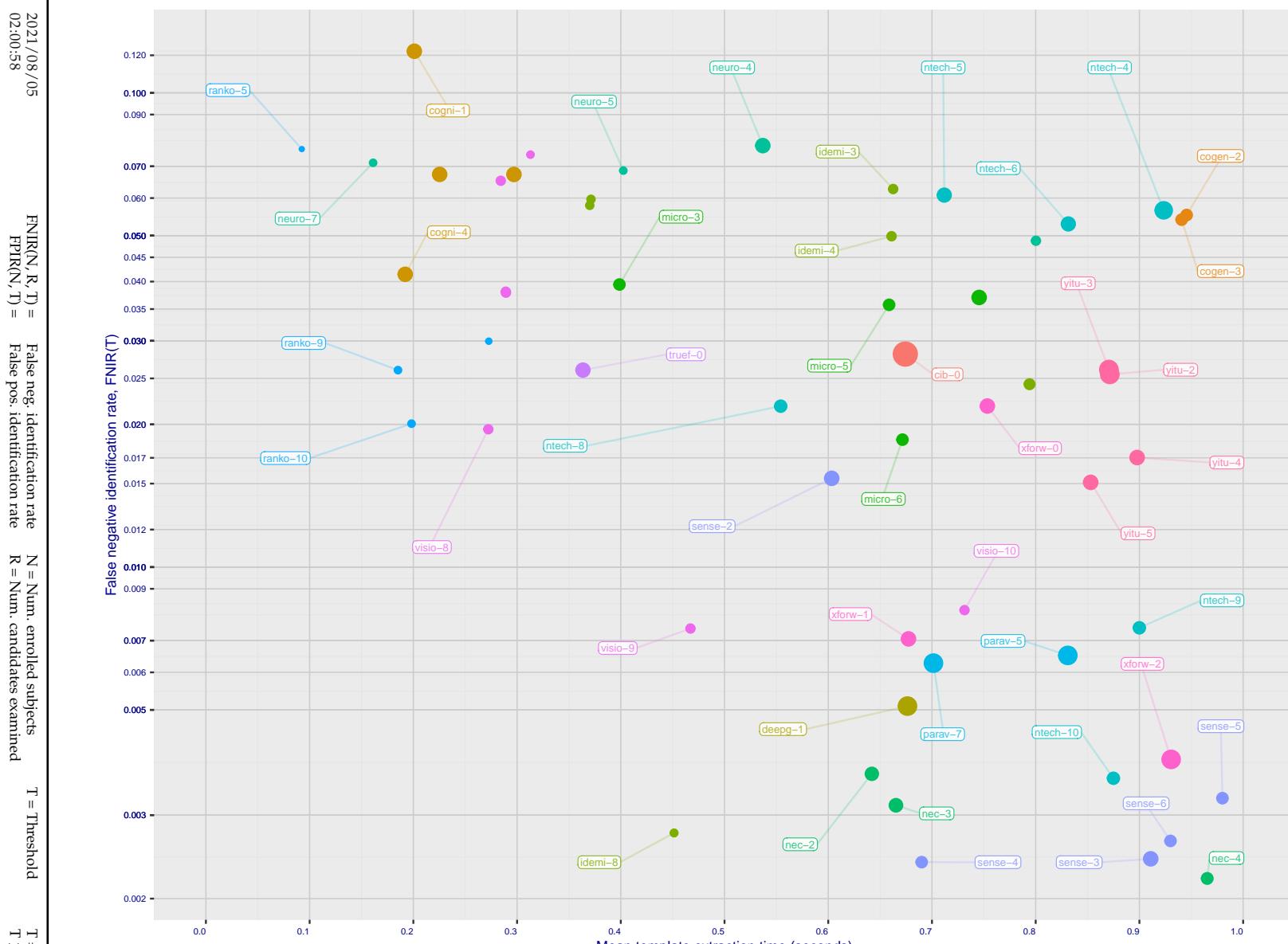


Figure 18: [Mugshot Dataset] Speed-accuracy tradeoff. For developers of the more accurate algorithms the plot shows the tradeoff of high-threshold recognition miss-rates, $\text{FNIR}(N, N, T)$ for $\text{FPIR}(N, T) = 0.003$, and template generation time. Developers are coded by color. Template size is encoded by the size of the circle. Some labels are quite distant from the respective point, to avoid superposing text. Without any other influences, the assumption would be that taking time to localize the face, and extract features, would lead to better accuracy. The most notable result, for NEC, is that their slower algorithms are much more accurate than the version that extract features in fewer than 90 milliseconds.

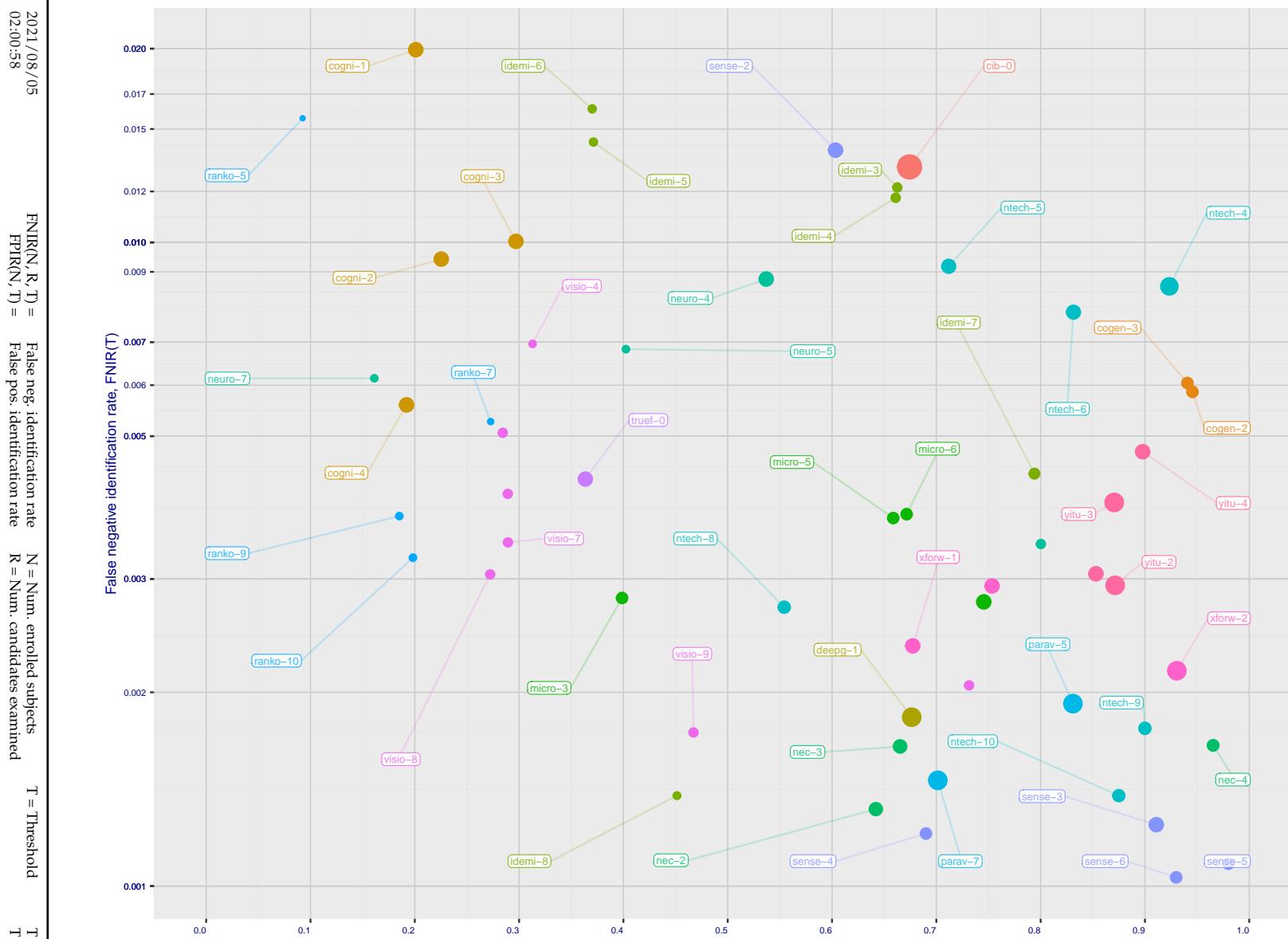


Figure 19: [Mugshot Dataset] Speed-accuracy tradeoff. For developers of the more accurate algorithms the plot shows the tradeoff of rank-one recognition miss-rates, $\text{FNIR}(N, 1, 0)$, and template generation time. Developers are coded by color. Template size is encoded by the size of the circle. Some labels are quite distant from the respective point, to avoid superposing text. Without any other influences, the assumption would be that taking time to localize the face, and extract features, would lead to better accuracy. This occurs for NEC with their slower algorithm being much accurate than the version that extract features in fewer than 90 milliseconds.

	DEVELOPER	SHORT	SEQ.	VALIDATION	CONFIG ¹		LIB ¹	TEMPLATE GENERATION			FINALIZE ²		SEARCH DURATION ⁵						POWER LAW (μ s)
					NAME	NUM.		DATA (MB)	DATA (MB)	SIZE (B)	MULT ³	TIME (MS) ⁴	TIME (S)	L=1	L=50	L=50	L=50	L=50	N=1.6M
1	3Divi	3divi	5	2018-10-26	186	51	148	4096	k	84,638	120 ²⁸	(73) 538	(73) 537	(66) 1377	(64) 2614	(60) 5530	101	0.07N ^{1.1}	
2	3Divi	3divi	6	2018-10-26	187	51	31	528	k	85,640	21 ⁵	(11) 33	(12) 33	-	-	-	-	-	N=1.6M
3	Acer Incorporated	acer	000	2020-08-12	35	67	22	512	-	15,198	13 ⁴	(50) 295	(51) 295	(41) 623	(38) 2302	(53) 4915	128	0.00N ^{1.3}	
4	Akurat Satu Indonesia	ptakuratsatu	000	2020-10-23	0	572	32	538	-	154,905	161 ²⁸⁶³³	(6) 15	(6) 16	(6) 17	(5) 17	(4) 17	3	6827.74N ^{0.1}	
5	Alchera Inc	alchera	2	2018-10-30	7	14	81	2048	k	6,114	14 ³	(35) 2923	(38) 2929	-	-	-	-	-	N=1.6M
6	Alchera Inc	alchera	3	2018-10-30	251	14	80	2048	k	69,531	146 ⁶³	(36) 2955	(39) 2956	(119) 6546	(120) 15013	(120) 35262	123	0.10N ^{1.2}	
7	Alvia / Innovation Sys	isystems	3	2018-10-30	350	784	94	2048	1	133,825	95 ¹⁶	(60) 385	(61) 389	(53) 679	(53) 1822	(76) 9348	129	0.00N ^{1.3}	
8	AllGoVision	allgovision	000	2019-07-30	168	150	72	2048	k	45,404	52 ¹²	(138) 3226	(141) 3193	(117) 6129	(117) 12449	(117) 25835	55	1.40N ^{1.0}	
9	AllGoVision	allgovision	001	2020-07-14	283	126	77	2048	-	122,777	58 ¹³	(137) 3174	(140) 3183	(116) 6073	(115) 12284	(116) 25701	53	1.42N ^{1.0}	
10	Anke Investments	anke	0	2018-10-30	779	27	134	2072	k	50,429	93 ¹⁶	(83) 675	(87) 748	(71) 1483	(70) 2968	(64) 6148	74	0.21N ^{1.1}	
11	Anke Investments	anke	1	2018-10-30	779	27	133	2072	k	51,430	85 ¹⁵	(87) 707	(90) 769	-	-	-	-	-	N=1.6M
12	Anke Investments	anke	002	2019-06-27	341	401	126	2056	k	80,623	67 ¹³	(82) 624	(82) 682	(64) 1306	(60) 2403	(57) 5082	48	0.30N ^{1.0}	
13	Aware	aware	5	2018-10-30	368	27	142	3100	k	127,792	120 ³⁴	(15) 95	(18) 98	(16) 203	(15) 371	(12) 252	13	4.13N ^{0.7}	
14	Aware	aware	6	2018-10-30	368	27	124	2048	k	126,789	2 ²	(28) 158	(28) 162	-	-	-	-	-	N=1.6M
15	Ayonix	ayonix	1	2018-10-29	74	2	46	1036	k	2,12	47 ¹¹	(46) 279	(47) 279	-	-	-	-	-	N=1.6M
16	Ayonix	ayonix	2	2018-10-30	74	2	47	1036	1	1'11	72 ¹⁴	(45) 279	(46) 276	(33) 535	(31) 1087	(31) 2284	62	0.11N ^{1.0}	
17	Camvi Technologies	camvitech	4	2018-10-30	233	220	40	1024	1	99,686	127 ³¹	(12) 33	(11) 32	(10) 38	(9) 40	(7) 48	4	8492.66N ^{0.1}	
18	Camvi Technologies	camvitech	5	2018-10-30	257	220	41	1024	1	118,751	125 ³¹	(10) 31	(9) 30	-	-	-	-	-	N=1.6M
19	Canon Inc	cib	000	2020-10-19	426	127	163	8196	-	94,674	149 ¹¹³	(139) 3589	(143) 3604	(120) 6738	(118) 13495	(118) 27114	23	2.33N ^{1.0}	
20	Cloudwalk - Hengrui AI Technology	hr	000	2021-02-10	501	392	71	2048	-	155,905	80 ¹⁵	(47) 282	(45) 276	(35) 539	(38) 1268	(44) 3177	105	0.03N ^{1.1}	
21	Cognitec Systems GmbH	cognitec	2	2018-10-30	463	26	123	2052	k	18,225	115 ²⁷	(122) 1733	(124) 1763	(106) 3660	(104) 7279	(103) 13895	52	0.83N ^{1.0}	
22	Cognitec Systems GmbH	cognitec	3	2018-10-30	465	26	115	2052	k	28,297	91 ¹⁶	(121) 1719	(125) 1791	(105) 3638	(103) 7277	(106) 14904	68	0.66N ^{1.0}	
23	Cognitec Systems GmbH	cognitec	004	2021-03-08	384	60	116	2052	-	13,192	67 ¹³	(120) 1673	(123) 1727	(98) 2904	(96) 5801	(94) 11707	29	1.15N ^{1.0}	
24	Cyberlink Corp	cyberlink	000	2019-06-12	217	93	122	2052	1	88,654	123 ³⁰	(85) 696	(84) 701	(67) 1379	(65) 2639	(66) 6214	64	0.28N ^{1.0}	
25	Cyberlink Corp	cyberlink	001	2019-10-07	459	102	118	2052	1	48,423	121 ²⁸	(86) 698	(80) 700	(65) 1350	(94) 5524	(97) 12031	12	0.00N ^{1.3}	
26	Cyberlink Corp	cyberlink	002	2020-07-31	333	109	158	4140	-	113,724	153 ⁸⁷⁵	(113) 1353	(142) 3198	(118) 6138	(114) 12205	(101) 13106	15	16.71N ^{0.8}	
27	Cyberlink Corp	cyberlink	003	2021-01-05	333	100	160	6212	-	102,691	135 ³⁵	(68) 488	(85) 723	(69) 1415	(68) 2886	(61) 5643	87	0.12N ^{1.1}	
28	Cyberlink Corp	cyberlink	004	2021-07-16	371	100	161	6212	-	115,728	112 ²³	(69) 492	(70) 504	(53) 923	(43) 1448	(46) 3350	18	0.73N ^{0.9}	
29	Dahua Technology Co Ltd	dahua	0	2018-10-29	276	167	105	2048	k	39,374	110 ²²	-	(43) 258	-	-	-	-	N=1.6M	
30	Dahua Technology Co Ltd	dahua	1	2018-10-29	276	167	86	2048	k	35,369	117 ²⁸	-	(42) 257	(39) 602	(36) 1202	(42) 3007	112	0.02N ^{1.2}	
31	Dahua Technology Co Ltd	dahua	002	2019-12-02	607	137	82	2048	k	98,685	105 ¹⁹	(38) 243	(44) 269	(60) 1189	(69) 2950	(69) 6732	133	0.00N ^{1.5}	
32	Dahua Technology Co Ltd	dahua	003	2020-11-18	889	154	84	2048	-	112,723	98 ¹⁸	(48) 283	(41) 249	(30) 468	(29) 935	(28) 1871	24	0.16N ^{1.0}	
33	Deepglint	deepglint	001	2019-11-15	448	265	151	4096	-	96,676	135 ³⁵	(84) 677	(116) 1495	(75) 1724	(66) 2747	(67) 6246	14	25.27N ^{0.8}	
34	Dermalog	dermalog	5	2018-10-26	0	440	128	128	1	68,528	152 ³¹⁵⁵	(1) 0	(1) 0	(1) 0	(1) 0	(1) 0	5	66.21N ^{0.2}	
35	Dermalog	dermalog	6	2018-10-26	0	453	13	256	1	64,507	3 ²	(25) 142	(25) 144	(21) 269	(20) 531	(19) 1294	69	0.05N ^{1.0}	
36	Dermalog	dermalog	007	2020-02-12	0	424	128	128	1	47,410	1 ¹	(18) 98	(16) 96	(18) 218	(16) 429	(16) 1013	94	0.01N ^{1.1}	
37	Dermalog	dermalog	008	2021-01-25	0	531	21	512	-	37,370	15 ⁴	(54) 335	(38) 246	(29) 462	(28) 924	(27) 1849	27	0.15N ^{1.0}	
38	FarBar Inc	f8	001	2019-10-03	266	19	102	2048	k	131,810	70 ¹⁴	-	-	-	-	-	-	N=1.6M	
39	Gorilla Technology	gorilla	2	2018-10-29	91	1252	52	1132	k	32,338	114 ²⁴	(26) 145	(26) 146	(22) 293	(21) 612	(23) 1509	92	0.02N ^{1.1}	
40	Gorilla Technology	gorilla	3	2018-10-26	94	1252	135	2156	k	71,559	157 ²²	(108) 1273	(110) 1307	(88) 2474	(87) 5198	(91) 11141	70	0.46N ^{1.0}	
41	Gorilla Technology	gorilla	004	2020-01-06	182	1244	136	2192	k	41,388	134 ⁴¹	(49) 286	(50) 285	(61) 1191	(61) 2416	(56) 5036	126	0.00N ^{1.3}	
42	Gorilla Technology	gorilla	005	2021-02-22	306	1420	162	6288	-	61,483	148 ⁷⁸	(91) 802	(91) 799	(72) 1514	(77) 4454	(72) 8820	114	0.05N ^{1.2}	
43	Guangzhou Pixel Solutions Co Ltd	pixelall	002	2019-07-01	0	165	139	2560	k	12,190	84 ¹⁵	(111) 1296	(113) 1334	(91) 2526	(86) 5136	(90) 11045	63	0.52N ^{1.0}	
44	Guangzhou Pixel Solutions Co Ltd	pixelall	003	2019-11-05	0	690	137	2560	k	106,703	111 ²²	(108) 1273	(110) 1307	(88) 2474	(87) 5198	(91) 11141	70	0.46N ^{1.0}	
45	Guangzhou Pixel Solutions Co Ltd	pixelall	004	2020-07-02	0	538	140	2560	k	52,449	97 ¹⁷	(107) 1259	(109) 1300	(87) 2465	(92) 5492	(92) 11443	70	0.34N ^{1.1}	
46	Guangzhou Pixel Solutions Co Ltd	pixelall	005	2021-03-23	0	717	138	2560	-	138,840	46 ¹¹	(117) 1606	(115) 1528	(83) 2609	(83) 4926	(95) 11770	41	0.73N ^{1.0}	
47	Hikvision Research Institute	hikvision	5	2018-10-29	593	9	57	1408	1	77,607	96 ¹⁶	(96) 883	(95) 895	(77) 1908	(73) 3792	(77) 9387	102	0.10N ^{1.1}	
48	Hikvision Research Institute	hikvision	6	2018-10-29	593	9	56	1408	1	75,598	92 ¹⁶	(94) 871	(96) 877	-	-	-	-	-	N=1.6M
49	Idemia	idemia	5	2018-10-29	417	48	13	352	1	38,371	20 ⁵	(22) 137	(23) 138	(27) 437	(25) 724	(24) 1630	120	0.01N ^{1.2}	
50	Idemia	idemia	6	2018-10-29	417	48	20	352	1	36,370	19 ⁴	(23) 137	(22) 138	(28) 442	(27) 827	(25) 1646	123	0.01N ^{1.2}	
51	Idemia	idemia	007	2020-01-17	738	113	37	860	1	128,794	71 ¹⁴	(27) 151	(27) 152	(44) 683	(45) 1481	(43) 3022	131	0.00N ^{1.4}	
52	Idemia	idemia	008	2021-03-15	378	65	18	300	-	54,451	11 ³	(21) 132	(21) 131</td						

2021/08/05
02:00:58FNIR(N, R, T) = False neg. identification rate
FPTR(N, T) = False pos. identification rateN = Num. enrolled subjects
R = Num. candidates examined

T = Threshold

T = 0 → Investigation
T > 0 → Identification

	DEVELOPER	SHORT NAME	SEQ. NUM.	VALIDATION DATE	CONFIG ¹ DATA (MB)	LIB ¹ DATA (MB)	TEMPLATE GENERATION			FINALIZE ² TIME (S)	SEARCH DURATION ⁵ MILLISEC						POWER LAW (μs)
							SIZE (B)	MULT ³	TIME (MS) ⁴		L=1 N=1.6M	L=50 N=1.6M	L=50 N=3M	L=50 N=6M	L=50 N=12M		
157	VisionLabs	visionlabs	008	2019-06-18	348	17	²³ 512	1	²⁴ 272	¹⁵⁸ 12747	⁽⁸⁾ 23	⁽⁷⁾ 24	⁽⁷⁾ 26	⁽⁶⁾ 29	⁽⁵⁾ 33	⁶ 2539.61 $N^{0.2}$	
158	VisionLabs	visionlabs	009	2020-08-04	689	20	²⁶ 512	-	⁵⁷ 467	¹⁵⁹ 13245	⁽⁹⁾ 23	⁽⁸⁾ 29	⁽⁸⁾ 34	⁽¹¹⁾ 61	⁽¹¹⁾ 145	¹¹ 8.88 $N^{0.6}$	
159	VisionLabs	visionlabs	010	2021-02-05	1042	20	²⁴ 512	-	¹¹⁶ 731	¹⁵⁵ 11837	⁽⁷⁾ 21	⁽¹⁰⁾ 32	⁽⁹⁾ 36	⁽⁷⁾ 39	⁽⁶⁾ 43	⁷ 3183.79 $N^{0.2}$	
160	Vocord	vocord	5	2018-10-30	1035	185	³⁶ 768	k	¹²³ 780	²⁷ 7	⁽²⁹⁾ 158	⁽³⁴⁾ 204	⁽²⁶⁾ 383	⁽²⁶⁾ 767	⁽²²⁾ 1466	³² 0.12 $N^{1.0}$	
161	Vocord	vocord	6	2018-10-30	1035	185	¹⁶⁴ 10240	k	¹²⁴ 785	¹⁵⁰ 243	⁽³⁰⁾ 170	⁽³⁶⁾ 216	-	-	-	-	
162	Xforward AI Technology	xforwardai	000	2020-07-24	236	171	¹¹³ 2048	-	¹¹⁹ 753	⁶⁶ 13	⁽¹⁴¹⁾ 4603	⁽¹⁵⁸⁾ 7647	⁽¹³⁴⁾ 15723	⁽¹²⁴⁾ 23900	⁽¹²⁷⁾ 53729	¹⁰⁴ 0.56 $N^{1.1}$	
163	Xforward AI Technology	xforwardai	001	2021-01-21	332	50	¹⁰⁰ 2048	-	⁹⁷ 677	⁹⁴ 16	⁽¹⁴⁹⁾ 5887	⁽¹⁴⁵⁾ 4384	⁽¹²³⁾ 8798	⁽¹²²⁾ 18553	⁽¹²³⁾ 48993	¹¹¹ 0.32 $N^{1.1}$	
164	Xforward AI Technology	xforwardai	002	2021-05-24	691	50	¹⁴⁷ 4096	-	¹⁵⁸ 930	¹⁰⁰ 18	⁽¹⁵³⁾ 6957	⁽¹⁵⁵⁾ 6400	⁽¹³¹⁾ 12659	⁽¹³⁰⁾ 31077	⁽¹³⁰⁾ 65158	¹⁰⁹ 0.52 $N^{1.1}$	

Notes

- Configuration size does not capture static data present in libraries. Libraries are included but the size also includes any ancillary libraries for image processing (e.g. openCV) or numerical computation (e.g. blas).
- Finalization is the processing of converting N = 1600000 templates into a searchable data structure an operation which can be a simple copy, or the building of an index or tree, for example. The duration of the operation may be data dependent, and may not be linear in the number of input templates.
- This multiplier expresses the increase in template size when k images are passed to the template generation function.
- All durations are measured on Intel®Xeon®CPU E5-2630 v4 @ 2.20GHz processors. Estimates are made by wrapping the API function call in calls to std::chrono::high_resolution_clock which on the machine in (3) counts 1ns clock ticks. Precision is somewhat worse than that however.
- Search durations are measured as in the prior note. The power-law model in the final column mostly fits the empirical results in Figure 166. However in certain cases the model is not correct and should not be used numerically.

Table 5: Summary of algorithms and properties included in this report. The blue superscripts give ranking for the quantity in that column. Missing search durations, denoted by “-”, are absent because those runs were not executed, usually because we did not run on the larger galleries. Caution: The power-law model is sometimes an incorrect model. It is included here only to show broad sublinear behavior, which is flagged in green. The models should not be used for prediction.

MISS RATES		INVESTIGATION, FNIR(N, R = 1, T = 0)								IDENTIFICATION, FNIR(N, R = L, T ≥ 0) FOR FPIR = 0.001							
#	ALGORITHM	(0, 2]	(2, 4]	(4, 6]	(6, 8]	(8, 10]	(10, 12]	(12, 14]	(14, 18]	(0, 2]	(2, 4]	(4, 6]	(6, 8]	(8, 10]	(10, 12]	(12, 14]	(14, 18]
89	SENSETIME-002	⁹¹ 0.0186	⁸⁷ 0.0191	⁷⁹ 0.0183	⁷⁰ 0.0179	⁶ 0.0173	⁴³ 0.0133	²¹ 0.0089	¹² 0.0059	³⁶ 0.0220	²⁶ 0.0236	¹⁵ 0.0237	¹⁵ 0.0240	⁹ 0.0245	⁷ 0.0219	⁸ 0.0195	⁷ 0.0222
90	SENSETIME-003	⁸ 0.0021	⁹ 0.0028	⁸ 0.0031	⁵ 0.0033	⁴ 0.0035	⁵ 0.0040	⁵ 0.0047	⁵ 0.0033	⁸ 0.0046	⁵ 0.0064	⁴ 0.0076	³ 0.0086	³ 0.0101	² 0.0122	³ 0.0155	³ 0.0196
91	SENSETIME-004	² 0.0016	² 0.0022	⁴ 0.0025	² 0.0028	² 0.0030	² 0.0035	³ 0.0043	² 0.0025	³ 0.0036	³ 0.0066	² 0.0081	² 0.0099	⁴ 0.0126	³ 0.0169	⁷ 0.0230	
92	SENSETIME-005	¹ 0.0015	¹ 0.0020	¹ 0.0024	¹ 0.0026	¹ 0.0029	¹ 0.0035	² 0.0043	³ 0.0028	⁴ 0.0036	⁴ 0.0059	⁵ 0.0089	⁶ 0.0128	⁷ 0.0177	⁸ 0.0240	⁹ 0.0345	⁷ 0.0493
93	SLAT-002	¹¹ 0.8309	¹¹ 0.8310	¹¹ 0.8311	¹¹² 0.8306	¹¹² 0.8296	¹¹² 0.8302	¹¹ 0.8300	¹¹ 0.8301	¹⁰ 0.8340	¹⁶ 0.8368	¹⁰ 0.8404	¹⁰ 0.8445	¹⁰ 0.8480	¹⁰ 0.8532	¹⁰ 0.8595	¹⁰ 0.8691
94	SYNESIS-003	⁸⁴ 0.0125	⁸⁰ 0.0151	⁷⁸ 0.0174	⁷⁵ 0.0199	⁷⁴ 0.0223	⁷⁰ 0.0240	⁷⁰ 0.0279	⁸⁰ 0.0331	⁸⁰ 0.0658	⁷⁸ 0.1052	⁷⁸ 0.1483	⁷⁷ 0.1968	⁷⁷ 0.2399	⁷⁵ 0.2834	⁷⁵ 0.3405	⁷⁴ 0.4046
95	SYNESIS-005	³⁵ 0.0044	³² 0.0058	³² 0.0070	³⁸ 0.0084	³⁸ 0.0091	³¹ 0.0103	³¹ 0.0125	³² 0.0152	⁴¹ 0.0262	⁴⁰ 0.0444	⁴⁰ 0.0666	⁴⁰ 0.0923	³⁹ 0.1156	³⁶ 0.1399	³⁸ 0.1736	³⁵ 0.2185
96	TECH5-001	⁵² 0.0061	⁵⁶ 0.0093	⁶¹ 0.0128	⁶⁶ 0.0171	⁷² 0.0221	⁷⁷ 0.0289	⁸⁰ 0.0412	⁸⁰ 0.0560	⁸¹ 0.0660	⁸² 0.1156	⁸⁵ 0.1733	⁸⁶ 0.2385	⁸⁶ 0.2998	⁸⁶ 0.3629	⁸⁸ 0.4424	⁸⁹ 0.5284
97	TOSHIBA-001	⁶⁸ 0.0086	⁶⁹ 0.0119	⁶⁹ 0.0150	⁶⁹ 0.0178	⁷⁰ 0.0209	⁷¹ 0.0241	⁷¹ 0.0292	⁷⁰ 0.0365								
98	TRUEFACE-000	³¹ 0.0043	³¹ 0.0057	²⁵ 0.0061	²³ 0.0067	²⁹ 0.0073	²⁹ 0.0084	²³ 0.0097	²¹ 0.0099	³⁶ 0.0200	³² 0.0338	³³ 0.0504	³⁰ 0.0705	³⁰ 0.0904	³¹ 0.1112	²⁸ 0.1401	²⁸ 0.1792
99	VERIDAS-001	⁵³ 0.0063	⁵¹ 0.0083	⁵¹ 0.0099	⁵¹ 0.0113	⁵¹ 0.0132	⁴⁹ 0.0148	⁵⁰ 0.0184	⁴⁷ 0.0219	⁵⁶ 0.0403	⁵⁶ 0.0684	⁵⁷ 0.1012	⁵⁷ 0.1386	⁵⁷ 0.1741	⁵⁷ 0.2113	⁵⁷ 0.2611	⁵⁸ 0.3233
100	VISIONLABS-004	³⁸ 0.0048	⁴¹ 0.0069	⁴⁷ 0.0091	⁵⁰ 0.0111	⁵⁰ 0.0130	⁵² 0.0152	⁵¹ 0.0187	⁵³ 0.0242	⁶⁹ 0.0540	⁷² 0.0916	⁷³ 0.1358	⁷³ 0.1855	⁷⁴ 0.2303	⁷³ 0.2745	⁷² 0.3312	⁶⁸ 0.3913
101	VISIONLABS-005	³⁴ 0.0044	³⁴ 0.0063	³⁸ 0.0081	⁴¹ 0.0095	⁴¹ 0.0109	³⁹ 0.0125	⁴⁰ 0.0151	⁴¹ 0.0187	⁶² 0.0479	⁶² 0.0812	⁶³ 0.1212	⁶⁵ 0.1664	⁶⁴ 0.2078	⁶⁴ 0.2473	⁶³ 0.2999	⁶² 0.3577
102	VISIONLABS-006	²⁵ 0.0035	²⁵ 0.0048	²⁷ 0.0061	²⁵ 0.0069	²⁹ 0.0077	²⁹ 0.0087	²⁶ 0.0105	²⁹ 0.0120	⁴¹ 0.0273	⁴² 0.0465	⁴² 0.0702	⁴² 0.0970	⁴² 0.1228	³⁹ 0.1486	³⁹ 0.1847	³⁹ 0.2295
103	VISIONLABS-008	¹⁸ 0.0028	¹⁷ 0.0037	¹⁸ 0.0047	¹⁸ 0.0053	¹⁸ 0.0058	¹⁷ 0.0067	¹⁹ 0.0081	²⁰ 0.0085	²² 0.0143	²² 0.0241	²³ 0.0373	²² 0.0519	²² 0.0677	²⁰ 0.0850	²⁰ 0.1104	²⁰ 0.1444
104	VISIONLABS-009	⁷ 0.0020	⁷ 0.0026	⁷ 0.0030	⁷ 0.0034	⁸ 0.0038	⁸ 0.0044	⁹ 0.0052	⁹ 0.0046	¹¹ 0.0065	¹² 0.0105	¹² 0.0156	¹² 0.0217	¹³ 0.0289	¹³ 0.0368	¹² 0.0499	¹² 0.0681
105	VISIONLABS-010	⁶ 0.0020	⁶ 0.0025	⁶ 0.0030	⁸ 0.0034	⁷ 0.0036	⁷ 0.0043	⁷ 0.0051	¹⁰ 0.0047	¹³ 0.0069	¹³ 0.0113	¹³ 0.0170	¹³ 0.0238	¹⁴ 0.0316	¹⁴ 0.0411	¹⁴ 0.0557	¹⁴ 0.0740
106	VTS-000	¹¹ 0.5878	¹¹¹ 0.6312	¹¹¹ 0.6602	¹¹⁰ 0.6863	¹¹⁰ 0.7073	¹¹⁰ 0.7246	¹¹⁰ 0.7458	¹¹⁰ 0.7747	¹⁰³ 0.5929	¹⁰³ 0.6397	¹⁰³ 0.6729	¹⁰³ 0.7034	¹⁰² 0.7279	¹⁰² 0.7493	¹⁰¹ 0.7739	¹⁰¹ 0.8076
107	XFORWARDAI-000	¹⁷ 0.0027	¹⁵ 0.0034	¹⁷ 0.0044	¹⁷ 0.0052	¹⁷ 0.0058	¹⁹ 0.0067	¹⁷ 0.0079	¹⁸ 0.0076	²³ 0.0157	²⁶ 0.0281	²⁵ 0.0443	²⁶ 0.0635	²⁷ 0.0834	²⁷ 0.1050	²⁶ 0.1330	²⁶ 0.1714
108	XFORWARDAI-001	¹² 0.0023	⁸ 0.0028	⁷ 0.0034	⁷ 0.0037	⁷ 0.0045	⁸ 0.0052	⁸ 0.0043	¹⁰ 0.0060	¹¹ 0.0096	¹⁰ 0.0144	⁹ 0.0200	¹⁰ 0.0260	¹⁰ 0.0334	¹⁰ 0.0435	¹⁰ 0.0586	
109	YITU-002	⁵⁵ 0.0066	⁵² 0.0083	⁴⁸ 0.0094	⁴⁵ 0.0101	⁴⁵ 0.0121	⁵¹ 0.0150	⁵⁰ 0.0223	⁶³ 0.0328	²⁸ 0.0189	²⁹ 0.0317	³⁰ 0.0494	³⁴ 0.0750	³⁶ 0.1066	⁴² 0.1494	⁴⁹ 0.2171	⁵⁵ 0.2958
110	YITU-003	³⁸ 0.0072	³⁵ 0.0089	³² 0.0100	⁴⁰ 0.0107	⁴⁰ 0.0125	⁵⁰ 0.0153	⁵⁸ 0.0226	⁶⁶ 0.0334	²⁸ 0.0194	³⁶ 0.0321	³¹ 0.0500	³⁶ 0.0756	³⁷ 0.1071	⁴³ 0.1500	⁵⁰ 0.2177	⁵⁵ 0.2964
111	YITU-004	⁵⁰ 0.0061	⁴⁶ 0.0075	³⁹ 0.0081	³⁶ 0.0081	³⁶ 0.0092	³⁹ 0.0107	⁴² 0.0154	⁴⁴ 0.0207	¹⁸ 0.0125	¹⁷ 0.0204	¹⁸ 0.0314	²⁰ 0.0469	²¹ 0.0671	²³ 0.0955	³¹ 0.1421	³⁵ 0.2006
112	YITU-005	⁵⁶ 0.0067	⁴⁹ 0.0080	⁴² 0.0087	³⁹ 0.0085	³⁶ 0.0094	³⁹ 0.0108	⁴¹ 0.0151	⁴³ 0.0204	¹⁷ 0.0124	¹⁶ 0.0198	¹⁷ 0.0308	¹⁷ 0.0462	²⁰ 0.0667	²² 0.0953	³⁰ 0.1418	³¹ 0.1930

Table 8: **Accuracy for the FRVT 2018 mugshot sets under ageing.** The second row shows the time lapse between gallery and subsequent probe images, in years. The first two columns identify the algorithm. The next 8 values give rank-based FNIR with $R = 1$, $T = 0$ and $FPIR = 1$. All these are relevant to investigational uses where candidates from all searches would need human review. The second 8 values give threshold-based FNIR with $T \geq 0$, $FPIR = 0.001$ and no rank criterion. The shaded cells indicate the three most accurate algorithms for that elapsed time. The gallery size is 3068801. The total number of searches is 10951064.

2021/08/05

FNIR(N, R, T) = False neg. identification rate

N = Num. enrolled subjects

T = Threshold

T = 0 → Investigation

T > 0 → Identification

02:00:58

#	ALGORITHM	MISSES BELOW THRESHOLD, T	ENROL MOST RECENT			
			DATASET: FRVT 2018 MUGSHOTS			
		N=0.64M	N=1.6M	N=3.0M	N=6.0M	N=12.0M
1	³ DIVI-005	¹⁶ 0.1358	¹⁶ 0.1664	¹⁴ 0.1915	¹³ 0.2370	¹³ 0.3054
2	ACER-000	¹⁶ 0.1185	¹⁶ 0.1455	¹⁴ 0.1714	¹³ 0.2074	¹² 0.2537
3	ALCHERA-003	¹⁵ 0.1176	¹⁶ 0.1553	¹⁴ 0.1853	¹³ 0.2409	¹⁴ 0.3553
4	ALLGOVISION-000	¹³ 0.0688	¹³ 0.0881	¹² 0.1084	¹¹ 0.1389	¹⁰ 0.2129
5	ALLGOVISION-001	¹⁴ 0.0785	¹⁴ 0.1017	¹³ 0.1218	¹² 0.1584	¹¹ 0.2273
6	ANKE-000	¹⁴ 0.0942	¹⁴ 0.1169	¹³ 0.1404	¹² 0.1776	¹² 0.2559
7	ANKE-002	⁷ 0.0229	⁷ 0.0318	⁷ 0.0406	⁶ 0.0605	⁶ 0.1466
8	AWARE-003	¹⁵ 0.1098	¹⁵ 0.1283	¹³ 0.1447	¹² 0.1768	¹¹ 0.2364
9	AWARE-005	¹⁹ 0.3389	¹⁹ 0.3643	¹⁵ 0.3993	¹⁴ 0.4526	¹² 0.5231
10	AYONIX-002	²¹ 0.7862	²¹ 0.8242	¹⁹ 0.8508	¹⁸ 0.8704	¹⁹ 0.8939
11	CAMVI-004	⁹ 0.0367	¹² 0.0716	¹¹ 0.0983	¹¹ 0.2508	¹³ 0.2701
12	CIB-000	²⁶ 0.0086	²⁸ 0.0125	²⁷ 0.0160	³⁴ 0.0303	⁴⁷ 0.1251
13	CLOUDWALK-HR-000	⁷ 0.0019	⁶ 0.0020	⁶ 0.0023	⁷ 0.0072	⁷ 0.0701
14	COGENT-000	¹¹ 0.0430	¹⁰ 0.0527	⁹ 0.0695	¹⁰ 0.1133	⁹ 0.1960
15	COGENT-001	¹¹ 0.0430	⁹ 0.0527	¹⁰ 0.0695	¹⁰ 0.1133	⁹ 0.1960
16	COGENT-002	⁸ 0.0322	⁸ 0.0444	⁸ 0.0610	¹⁰ 0.1116	¹⁰ 0.2180
17	COGENT-003	⁸ 0.0328	⁹ 0.0463	⁹ 0.0683	¹¹ 0.1294	¹² 0.2445
18	COGENT-004	⁶ 0.0210	⁷ 0.0331	⁸ 0.0527	¹⁰ 0.1138	¹⁰ 0.2119
19	COGNITEC-000	¹⁶ 0.1377	¹⁶ 0.1606	¹⁴ 0.1870	¹³ 0.2176	¹³ 0.2831
20	COGNITEC-001	¹⁴ 0.0807	¹⁴ 0.1017	¹³ 0.1214	¹² 0.1513	¹¹ 0.2238
21	COGNITEC-002	¹⁰ 0.0406	¹⁰ 0.0531	⁹ 0.0666	⁸ 0.0935	⁹ 0.1874
22	COGNITEC-003	¹⁰ 0.0400	⁹ 0.0526	⁹ 0.0650	⁸ 0.0895	⁸ 0.1772
23	COGNITEC-004	⁷ 0.0222	⁷ 0.0313	⁶ 0.0388	⁶ 0.0540	³² 0.1103
24	CYBERLINK-000	¹⁰ 0.0414	¹⁰ 0.0565	¹⁰ 0.0707	⁹ 0.1031	¹⁰ 0.2050
25	CYBERLINK-001	¹⁰ 0.0392	¹⁰ 0.0536	¹⁰ 0.0695	⁹ 0.0973	⁸ 0.1794
26	CYBERLINK-002	³² 0.0105	³⁵ 0.0148	³⁶ 0.0202	⁴⁹ 0.0399	⁴⁵ 0.1255
27	CYBERLINK-003	²⁰ 0.0056	²⁰ 0.0077	¹⁹ 0.0100	²² 0.0235	⁴¹ 0.1237
28	CYBERLINK-004	¹⁹ 0.0051	¹⁹ 0.0071	²⁰ 0.0102	¹⁸ 0.0199	⁵⁰ 0.1269
29	DAHUA-001	¹² 0.0569	¹² 0.0727	¹¹ 0.0878	¹⁰ 0.1148	⁹ 0.1867
30	DAHUA-002	³⁶ 0.0108	³⁶ 0.0151	³⁵ 0.0191	³¹ 0.0291	³⁹ 0.1153
31	DAHUA-003	³⁶ 0.0100	³² 0.0139	³⁶ 0.0180	³² 0.0296	³⁶ 0.1130
32	DEEPLINT-001	¹² 0.0027	¹² 0.0033	¹² 0.0043	¹² 0.0121	²¹ 0.0922
33	DEEPSEA-001	⁹ 0.0347	⁹ 0.0462	⁸ 0.0586	⁸ 0.0802	⁸ 0.1708
34	DERMALOG-005	¹⁴ 0.0700	¹³ 0.0880	¹² 0.1144	¹² 0.1578	¹² 0.2451
35	DERMALOG-006	¹⁰ 0.0395	⁹ 0.0517	⁹ 0.0659	⁹ 0.0973	⁸ 0.1745
36	DERMALOG-007	¹³ 0.0691	¹³ 0.0863	¹² 0.1107	¹¹ 0.1504	¹¹ 0.2299
37	DERMALOG-008	⁸ 0.0338	⁸ 0.0455	⁸ 0.0626	⁹ 0.1060	¹¹ 0.2276
38	GORILLA-002	¹⁷ 0.1539	¹⁷ 0.1880	¹⁴ 0.2184	¹⁴ 0.2596	¹⁴ 0.3398
39	GORILLA-004	¹³ 0.0699	¹³ 0.0892	¹² 0.1048	¹¹ 0.1370	⁹ 0.1969
40	GORILLA-005	¹¹ 0.0453	¹¹ 0.0583	¹⁰ 0.0704	⁹ 0.0974	⁶² 0.1474
41	HIK-003	¹⁴ 0.0828	¹⁴ 0.1028	¹² 0.1202	¹² 0.1525	¹² 0.2480
42	HIK-004	¹⁴ 0.0796	¹⁴ 0.0988	¹² 0.1147	¹¹ 0.1474	¹² 0.2483
43	HIK-005	⁸ 0.0312	⁸ 0.0436	⁸ 0.0560	⁸ 0.0911	¹⁰ 0.2129
44	IDEARIA-003	⁹ 0.0346	⁹ 0.0471	¹¹ 0.0892	¹⁴ 0.2789	¹⁴ 0.4311
45	IDEARIA-004	⁸ 0.0300	⁸ 0.0373	⁷ 0.0447	⁶ 0.0617	⁸⁰ 0.1635
46	IDEARIA-005	⁹ 0.0360	⁸ 0.0440	⁸ 0.0537	⁸ 0.0764	⁹ 0.1915
47	IDEARIA-006	⁹ 0.0351	⁸ 0.0433	⁸ 0.0525	⁷ 0.0734	¹⁰ 0.2201
48	IDEARIA-007	⁴ 0.0136	⁴² 0.0181	³⁹ 0.0228	⁴² 0.0357	⁵⁹ 0.1402
49	IDEARIA-008	⁷ 0.0016	⁵ 0.0019	⁸ 0.0024	⁴ 0.0053	⁸ 0.0470
50	IMAGUS-005	⁴⁴ 0.0137	⁴⁵ 0.0185	⁴⁴ 0.0237	⁴³ 0.0368	⁴⁰ 0.1067
51	IMAGUS-006	⁴ 0.0137	⁴⁷ 0.0190	⁴⁷ 0.0244	⁴⁷ 0.0396	⁴⁰ 0.1159
52	IMPERIAL-000	⁵⁸ 0.0187	⁵⁸ 0.0259	⁶⁵ 0.0358	⁷⁷ 0.0733	⁵⁶ 0.1794
53	INCODE-003	¹⁶ 0.1324	¹⁶ 0.1672	¹⁴ 0.1961	¹³ 0.2345	¹³ 0.3123
54	INCODE-004	¹⁰ 0.0403	¹⁰ 0.0538	⁹ 0.0662	⁸⁷ 0.0917	⁷⁷ 0.1619
55	INTSYSMSU-000	²² 0.9982	²² 0.9984	¹⁶ 0.9985	¹⁵ 0.9987	¹⁵ 0.9988
56	IREX-000	⁶² 0.0190	⁶⁶ 0.0280	⁶⁷ 0.0391	⁷³ 0.0677	⁶⁵ 0.1479
57	ISYSTEMS-002	¹² 0.0584	¹² 0.0783	¹¹ 0.0973	¹¹ 0.1373	¹¹ 0.2295
58	ISYSTEMS-003	¹¹ 0.0438	¹¹ 0.0590	¹¹ 0.0807	¹⁰ 0.1259	¹¹ 0.2357
59	KAKAO-000	³⁷ 0.0109	³⁸ 0.0151	³⁷ 0.0196	³⁷ 0.0324	²⁴ 0.1010
60	KEDACOM-001	⁵⁵ 0.0181	⁵⁴ 0.0227	⁴⁸ 0.0265	⁵² 0.0422	⁵⁶ 0.1340
61	LOOKMAN-003	⁹ 0.0346	⁸⁵ 0.0437	⁷⁹ 0.0514	⁷⁶ 0.0724	⁷⁸ 0.1620
62	LOOKMAN-005	⁷² 0.0240	⁶⁸ 0.0301	⁶⁴ 0.0356	⁵⁸ 0.0512	⁵⁵ 0.1334
63	MEGVII-001	¹² 0.0562	¹² 0.0722	¹¹ 0.0872	¹¹ 0.1309	¹³ 0.2713
64	MICROFOCUS-005	²² 0.9732	²¹ 0.8354	¹⁶ 0.8555	¹⁵ 0.8755	¹⁵ 0.8954
65	MICROSOFT-003	⁶⁴ 0.0198	⁶⁴ 0.0278	⁶¹ 0.0356	⁶³ 0.0538	⁷¹ 0.1539
66	MICROSOFT-004	⁵⁷ 0.0185	⁵⁹ 0.0259	⁵⁸ 0.0333	⁵⁹ 0.0517	⁶⁹ 0.1510
67	MICROSOFT-005	⁵⁶ 0.0181	⁵⁷ 0.0256	⁵⁷ 0.0320	⁵⁷ 0.0512	⁶⁷ 0.1491
68	MICROSOFT-006	²⁸ 0.0091	²⁵ 0.0120	²⁸ 0.0162	³³ 0.0301	⁶⁶ 0.1482
69	NEC-000	¹³ 0.0637	¹² 0.0789	¹¹ 0.0933	¹⁰ 0.1163	⁹⁴ 0.1941
70	NEC-001	¹⁴ 0.0863	¹⁴ 0.1055	¹³ 0.1249	¹² 0.1519	¹¹ 0.2253
71	NEC-002	⁹ 0.0020	¹⁰ 0.0026	¹⁰ 0.0033	¹⁴ 0.0135	⁹ 0.0653
72	NEC-003	¹⁰ 0.0021	⁸ 0.0024	⁷ 0.0028	⁶ 0.0059	⁸ 0.0540

Table 14: Identification-mode: Effect of N on FNIR at high threshold. Values are threshold-based miss rates i.e. FNIR at FPIR = 0.001 for five enrollment population sizes, N. The right six columns apply for enrollment of one image. Missing entries usually apply because another algorithm from the same developer was run instead. Some developers are missing because less accurate algorithms were not run on galleries with $N \geq 3\,000\,000$. Throughout blue superscripts indicate the rank of the algorithm for that column.

MISSES BELOW THRESHOLD, T FNIR(N, T > 0, R > L)		ENROL, MOST RECENT DATASET: FRVT 2018 MUGSHOTS					
#	ALGORITHM	N=0.64M	N=1.6M	N=3.0M	N=6.0M	N=12.0M	
73	NEC-004	⁶ 0.0017	² 0.0018	¹ 0.0020	¹ 0.0037	¹ 0.0329	
74	NEUROTECHNOLOGY-003	²⁰⁷ 0.5698	²⁰⁸ 0.6362	¹⁵⁸ 0.7035	¹⁸¹ 0.7602	¹⁴⁸ 0.8224	
75	NEUROTECHNOLOGY-004	¹¹⁸ 0.0466	¹¹⁷ 0.0629	¹⁰⁶ 0.0779	¹⁰⁴ 0.1135	¹⁰⁴ 0.2102	
76	NEUROTECHNOLOGY-005	¹⁰² 0.0396	¹⁰⁶ 0.0538	⁹⁶ 0.0675	⁹¹ 0.0950	⁹⁸ 0.1966	
77	NEUROTECHNOLOGY-007	¹¹⁴ 0.0436	¹¹⁶ 0.0623	¹⁰⁸ 0.0802	¹¹³ 0.1320	¹¹⁹ 0.2393	
78	NEUROTECHNOLOGY-008	²⁰ 0.0339	¹⁰¹ 0.0530	¹¹ 0.0893	¹²⁸ 0.1769	¹³² 0.3288	
79	NTECHLAB-003	¹¹⁰ 0.0421	¹⁰⁴ 0.0537	⁹⁶ 0.0674	⁸⁸ 0.0907	⁷⁵ 0.1582	
80	NTECHLAB-004	⁸² 0.0312	⁸¹ 0.0405	⁸⁰ 0.0519	⁷⁵ 0.0722	⁶⁸ 0.1503	
81	NTECHLAB-005	⁸⁶ 0.0334	⁸² 0.0424	⁸⁴ 0.0537	⁸⁰ 0.0760	⁷³ 0.1543	
82	NTECHLAB-006	⁷⁸ 0.0288	⁷⁶ 0.0367	⁷⁷ 0.0471	⁷² 0.0670	²⁰ 0.1523	
83	NTECHLAB-007	³⁹ 0.0188	⁵⁶ 0.0256	⁵⁰ 0.0317	⁵⁶ 0.0495	³¹ 0.1306	
84	NTECHLAB-008	³⁴ 0.0107	³³ 0.0145	³⁴ 0.0187	²⁹ 0.0286	²² 0.0995	
85	NTECHLAB-009	⁷ 0.0037	¹⁷ 0.0049	¹⁷ 0.0062	¹⁵ 0.0125	¹⁴ 0.0735	
86	NTECHLAB-010	⁸ 0.0020	⁹ 0.0025	⁸ 0.0030	⁸ 0.0077	¹³ 0.0710	
87	PARAVISION-003	⁷⁴ 0.0260	⁷⁴ 0.0351	⁷⁸ 0.0447	⁷¹ 0.0657	⁷⁹ 0.1630	
88	PARAVISION-004	²² 0.0074	²⁴ 0.0101	²⁴ 0.0136	²⁷ 0.0267	⁴⁹ 0.1256	
89	PARAVISION-005	¹⁴ 0.0032	¹⁴ 0.0041	¹⁴ 0.0057	¹⁷ 0.0174	²⁵ 0.1037	
90	PARAVISION-007	¹⁵ 0.0030	¹³ 0.0040	¹⁵ 0.0055	¹⁹ 0.0211	³¹ 0.1097	
91	PIXELALL-002	¹⁰¹ 0.0716	¹⁴⁵ 0.1052	¹³⁹ 0.1475	¹⁴⁰ 0.2489	¹⁴³ 0.3904	
92	PIXELALL-003	³¹ 0.0158	⁵¹ 0.0218	⁵³ 0.0288	⁵³ 0.0474	³⁷ 0.1138	
93	PIXELALL-004	³⁹ 0.0129	⁴⁴ 0.0183	⁴⁷ 0.0245	⁴⁴ 0.0378	⁵⁷ 0.1375	
94	PIXELALL-005	²⁰ 0.0087	²⁷ 0.0121	³⁶ 0.0171	²³ 0.0250	²⁹ 0.1052	
95	PTAKURATSATU-000	⁷⁵ 0.0275	⁷⁵ 0.0366	⁷⁶ 0.0458	⁶¹ 0.0523	⁷ 0.0523	
96	QUANTASOFT-001	²⁰⁹ 0.6387	²⁰⁹ 0.6387	¹⁵⁷ 0.6387	¹⁴⁶ 0.6387		
97	RANKONE-002	¹⁵⁴ 0.0973	¹⁴⁹ 0.1175	¹³³ 0.1359	¹²⁵ 0.1718	¹²⁹ 0.2613	
98	RANKONE-003	¹⁵³ 0.0973	¹⁵⁰ 0.1175	¹³⁴ 0.1359	¹²⁶ 0.1718	¹²⁸ 0.2613	
99	RANKONE-005	¹¹⁹ 0.0473	¹¹⁴ 0.0592	¹⁰² 0.0700	⁸⁹ 0.0944	¹⁰⁰ 0.1998	
100	RANKONE-007	³⁸ 0.0168	⁵³ 0.0222	⁴⁶ 0.0266	⁴⁶ 0.0381	³⁸ 0.1132	
101	RANKONE-009	⁴⁰ 0.0132	⁴¹ 0.0177	⁴¹ 0.0230	³⁹ 0.0344	²⁰ 0.0921	
102	RANKONE-010	³³ 0.0106	³¹ 0.0136	³¹ 0.0174	²⁶ 0.0265	¹⁶ 0.0785	
103	REALNETWORKS-002	¹⁷⁹ 0.1943	¹⁷⁸ 0.2314	¹⁵⁹ 0.2656	¹⁴⁶ 0.3134	¹³ 0.3208	
104	REALNETWORKS-003	¹⁶⁵ 0.1300	¹⁶⁴ 0.1594	¹⁴⁴ 0.1858	¹³⁵ 0.2246	¹³⁴ 0.3076	
105	REALNETWORKS-004	¹⁶⁴ 0.1279	¹⁶³ 0.1581	¹⁴⁷ 0.1857	¹³⁶ 0.2329	¹³⁰ 0.3179	
106	REALNETWORKS-005	⁶⁵ 0.0202	⁶³ 0.0277	⁶² 0.0355	⁶⁷ 0.0560	⁶⁰ 0.1431	
107	REMARKAI-000	¹⁰⁷ 0.0406	¹⁰⁷ 0.0552	⁹⁷ 0.0676	⁹⁵ 0.1028	¹⁰¹ 0.2003	
108	RENDIP-000	²⁵ 0.0085	²⁶ 0.0121	²⁶ 0.0156	²⁸ 0.0277	⁴² 0.1182	
109	S1-000	⁶⁹ 0.0204	⁶⁵ 0.0279	⁶⁷ 0.0382	⁷⁰ 0.0630	⁸¹ 0.1707	
110	SCANOVATE-000	¹²⁰ 0.0498	¹²⁰ 0.0667	¹⁰⁹ 0.0804	⁹⁹ 0.1097	³³ 0.1109	
111	SCANOVATE-001	¹³ 0.0630	¹³⁰ 0.0815	¹²⁸ 0.0993	¹⁰⁹ 0.1292	⁹⁷ 0.1960	
112	SENSETIME-000	⁵⁰ 0.0158	⁴⁹ 0.0208	⁵¹ 0.0270	⁴⁸ 0.0398	⁴³ 0.1232	
113	SENSETIME-001	³² 0.0161	⁵² 0.0219	⁵⁰ 0.0277	⁵¹ 0.0420	³² 0.1304	
114	SENSETIME-002	⁴⁷ 0.0146	³⁴ 0.0148	²⁹ 0.0153	²¹ 0.0234	¹⁰ 0.0657	
115	SENSETIME-003	³ 0.0016	⁴ 0.0018	³ 0.0021	⁵ 0.0054	⁴ 0.0451	
116	SENSETIME-004	² 0.0015	¹ 0.0018	² 0.0021	² 0.0040	¹ 0.0354	
117	SENSETIME-005	⁴ 0.0016	⁷ 0.0022	⁹ 0.0031	¹⁰ 0.0089	⁵ 0.0454	
118	SENSETIME-006	¹ 0.0014	³ 0.0018	³ 0.0023	³ 0.0047	³ 0.0372	
119	SHAMAN-007	¹⁶³ 0.1212	¹⁵⁹ 0.1413	¹⁴⁰ 0.1587	¹³⁰ 0.1879	¹²² 0.2460	
120	SIAT-001	⁴² 0.0136	³⁹ 0.0176	⁴² 0.0230	³⁸ 0.0344	²⁴ 0.1035	
121	SIAT-002	⁴⁹ 0.0154	⁵⁰ 0.0216	⁵² 0.0273	⁵⁰ 0.0404	⁵¹ 0.1283	
122	SYNESIS-003	¹² 0.0499	¹¹⁸ 0.0652	¹¹⁸ 0.0804	⁹⁸ 0.1095	⁹² 0.1916	
123	SYNESIS-003	²⁰⁵ 0.5341	²⁰⁵ 0.5821	¹⁵⁶ 0.6113	¹⁵⁰ 0.6479	¹⁴⁷ 0.6822	
124	SYNESIS-005	⁵⁴ 0.0181	⁵⁵ 0.0248	⁵⁶ 0.0319	⁶⁰ 0.0518	⁷⁴ 0.1580	
125	TECH5-001	¹⁰⁹ 0.0420	¹⁰⁹ 0.0574	¹¹⁸ 0.0911	¹³³ 0.2106	¹⁴² 0.3725	
126	TECH5-002	⁶³ 0.0194	⁶² 0.0269	⁶¹ 0.0346	⁶² 0.0537	⁷⁶ 0.1607	
127	TEVIAN-005	¹⁵⁸ 0.0692	¹³⁵ 0.0873	¹²³ 0.1066	¹¹¹ 0.1301	⁸⁸ 0.1840	
128	TEVIAN-006	²⁴ 0.0078	²² 0.0098	²² 0.0130	²⁵ 0.0261	⁵³ 0.1305	
129	TIGER-002	¹³³ 0.0647	¹³² 0.0861	¹²¹ 0.1036	¹¹⁴ 0.1332	¹¹⁰ 0.2231	
130	TOSHIBA-000	¹¹⁷ 0.0460	¹¹⁵ 0.0620	¹⁰⁷ 0.0780	¹⁰¹ 0.1117	¹⁰³ 0.2082	
131	TRUEFACE-000	⁴¹ 0.0134	⁴³ 0.0182	⁴⁸ 0.0238	⁴⁵ 0.0380	³⁸ 0.1385	
132	VD-001	¹⁷⁵ 0.1642	¹⁷⁸ 0.2015	¹⁵¹ 0.2351	¹⁴³ 0.2736	¹³⁹ 0.3293	
133	VERIDAS-001	⁷ 0.0278	⁷⁹ 0.0373	⁷⁸ 0.0491	⁷⁹ 0.0753	⁷² 0.1541	
134	VERIDAS-002	⁷⁶ 0.0278	⁷⁸ 0.0373	⁶⁶ 0.0373	⁵⁵ 0.0491	¹⁵ 0.0753	
135	VIGILANTSOLUTIONS-008	⁴⁸ 0.0146	⁴⁸ 0.0205	⁵⁰ 0.0269	⁵⁴ 0.0489	⁴¹ 0.1164	
136	VISIONLABS-004	¹¹¹ 0.0427	¹¹⁰ 0.0578	¹⁰³ 0.0703	⁹⁰ 0.0949	⁸⁹ 0.1853	
137	VISIONLABS-005	⁹⁸ 0.0369	⁹⁶ 0.0502	⁸⁸ 0.0626	⁸³ 0.0847	⁸⁷ 0.1815	
138	VISIONLABS-006	⁶⁰ 0.0188	⁶¹ 0.0267	⁶⁹ 0.0336	⁶⁶ 0.0542	⁶³ 0.1478	
139	VISIONLABS-007	⁶¹ 0.0188	⁶⁰ 0.0266	⁵⁹ 0.0335	⁶⁵ 0.0540	⁶¹ 0.1479	
140	VISIONLABS-008	²⁹ 0.0096	²⁹ 0.0131	²⁹ 0.0166	³⁰ 0.0291	⁴⁶ 0.1247	
141	VISIONLABS-009	¹⁵ 0.0034	¹⁵ 0.0046	¹⁵ 0.0060	¹⁵ 0.0140	¹⁸ 0.0881	
142	VISIONLABS-010	¹⁸ 0.0038	¹⁸ 0.0051	¹⁸ 0.0070	¹⁶ 0.0149	¹⁹ 0.0920	
143	VOCORD-005	¹⁶⁰ 0.1179	¹⁶² 0.1577	¹⁴⁸ 0.2183	¹⁴⁵ 0.3122	¹⁴⁵ 0.4490	
144	VTS-001	³¹ 0.0102	³⁰ 0.0133	³⁴ 0.0175	³⁵ 0.0322	⁴⁵ 0.1243	

Table 15: Identification-mode: Effect of N on FNIR at high threshold. Values are threshold-based miss rates i.e. FNIR at FPIR = 0.001 for five enrollment population sizes, N. The right six columns apply for enrollment of one image. Missing entries usually apply because another algorithm from the same developer was run instead. Some developers are missing because less accurate algorithms were not run on galleries with $N \geq 3\,000\,000$. Throughout blue superscripts indicate the rank of the algorithm for that column.

#	ALGORITHM	ENROL MOST RECENT					
		DATASET: FRVT 2018 MUGSHOTS					
		N=0.64M	N=1.6M	N=3.0M	N=6.0M	N=12.0M	
145	XFORWARDAI-000	³⁵ 0.0107	³⁷ 0.0151	³⁶ 0.0195	³⁶ 0.0324	²⁸ 0.1057	
146	XFORWARDAI-001	¹⁶ 0.0037	¹⁶ 0.0049	¹⁶ 0.0060	¹¹ 0.0120	¹⁷ 0.0800	
147	XFORWARDAI-002	¹¹ 0.0026	¹¹ 0.0030	¹¹ 0.0035	¹⁰ 0.0078	¹² 0.0706	
148	YITU-002	³⁸ 0.0129	⁴⁰ 0.0177	⁴⁰ 0.0228	⁴⁰ 0.0345	³⁶ 0.1133	
149	YITU-003	⁴⁶ 0.0138	⁴⁶ 0.0185	⁴⁵ 0.0236	⁴¹ 0.0353	³⁸ 0.1148	
150	YITU-004	²¹ 0.0067	²¹ 0.0096	²¹ 0.0129	²⁰ 0.0232	²⁶ 0.1046	
151	YITU-005	²³ 0.0074	²³ 0.0101	²³ 0.0135	²⁴ 0.0255	²⁸ 0.1057	

Table 16: Identification-mode: Effect of N on FNIR at high threshold. Values are threshold-based miss rates i.e. FNIR at FPIR = 0.001 for five enrollment population sizes, N. The right six columns apply for enrollment of one image. Missing entries usually apply because another algorithm from the same developer was run instead. Some developers are missing because less accurate algorithms were not run on galleries with $N \geq 3\,000\,000$. Throughout blue superscripts indicate the rank of the algorithm for that column.

MISSES AT GIVEN RANK		ENROL MOST RECENT											
FNIR(N, T= 0, r)		RANK 1					RANK 50						
#	ALGORITHM	N=0.64M	N=1.6M	N=3.0M	N=6.0M	N=12.0M	aN ^b	N=0.64M	N=1.6M	N=3.0M	N=6.0M	N=12.0M	aN ^b
145	VTS-001	¹⁷ 0.0014	²³ 0.0015	²² 0.0017	²⁵ 0.0019	²⁴ 0.0023	⁴² 0.0001 N ^{0.179} ⁶⁴	²⁰ 0.0010	¹⁸ 0.0010	¹⁷ 0.0011	¹⁶ 0.0011	⁸⁰ 0.0005 N ^{0.051} ⁴⁷	
146	XFORWARDAI-000	⁵⁶ 0.0021	⁴⁹ 0.0023	⁴⁴ 0.0024	⁴² 0.0027	³⁶ 0.0029	¹⁴ 0.0005 N ^{0.111} ³¹	¹⁰ 0.0019	⁹⁵ 0.0019	⁸⁸ 0.0019	⁷³ 0.0020	⁶⁵ 0.0020	¹³⁰ 0.0015 N ^{0.018} ²⁰
147	XFORWARDAI-001	⁵⁰ 0.0020	⁴⁴ 0.0020	³⁵ 0.0021	²⁹ 0.0022	²⁵ 0.0024	¹²⁹ 0.0009 N ^{0.085} ¹⁰	¹⁰⁹ 0.0019	⁹⁴ 0.0019	⁸⁶ 0.0019	⁷¹ 0.0019	⁵⁸ 0.0019	¹³⁴ 0.0018 N ^{0.004} ⁷
148	XFORWARDAI-002	⁴⁶ 0.0019	³⁸ 0.0020	³² 0.0020	²⁶ 0.0021	²¹ 0.0022	¹³¹ 0.0011 N ^{0.038} ⁵	⁹⁹ 0.0019	⁹³ 0.0019	⁸⁵ 0.0019	⁷⁰ 0.0019	⁵⁶ 0.0019	¹³³ 0.0018 N ^{0.003} ⁶
149	YITU-002	³⁰ 0.0016	³³ 0.0018	³⁶ 0.0021	³⁶ 0.0024	³⁷ 0.0029	³⁰ 0.0001 N ^{0.213} ⁷⁹	¹⁸ 0.0009	²¹ 0.0010	¹⁹ 0.0010	¹⁸ 0.0011	¹⁷ 0.0012	⁶⁹ 0.0004 N ^{0.073} ⁶²
150	YITU-003	⁷² 0.0026	⁶⁹ 0.0029	⁶⁴ 0.0031	⁵⁹ 0.0035	⁵⁶ 0.0039	¹⁰⁸ 0.0004 N ^{0.141} ⁴³	¹⁰ 0.0020	¹⁰² 0.0021	⁹⁵ 0.0022	⁸⁸ 0.0023	⁸⁰ 0.0024	¹⁷⁰ 0.0010 N ^{0.054} ⁴⁸
151	YITU-004	¹¹ 0.0011	¹¹ 0.0013	¹⁴ 0.0015	¹⁷ 0.0017	⁶⁵ 0.0047	³ 0.0000 N ^{0.438} ¹⁴⁸	¹¹ 0.0008	⁹ 0.0009	¹⁰ 0.0009	⁹ 0.0009	⁹⁵ 0.0036	⁵ 0.0000 N ^{0.395} ¹⁴⁵
152	YITU-005	⁶¹ 0.0022	⁵¹ 0.0023	⁴⁷ 0.0025	⁴³ 0.0027	⁴⁰ 0.0031	¹⁶ 0.0005 N ^{0.113} ³²	¹⁰² 0.0020	⁹⁷ 0.0020	⁹¹ 0.0020	⁸⁰ 0.0020	⁶⁷ 0.0020	¹³² 0.0017 N ^{0.012} ¹²

Table 19: Investigation-mode: Effect of N on FNIR on recent images For five enrollment population sizes, N, with T = 0 and FPIR = 1. The left five columns are rank 1 miss rates The right five columns are rank 50 miss rates Missing entries usually apply because another algorithm from the same developer was run instead. Some developers are missing because less accurate algorithms were not run on galleries with N > 1 600 000. Throughout blue superscripts indicate the rank of the algorithm for that column, and yellow highlighting indicates the most accurate value. Caution: The Power-low models are mostly intended to draw attention to the kind of behavior, not as a model to be used for prediction.

MISSES OUTSIDE RANK R		RESOURCE USAGE		ENROL MOST RECENT, N = 1.6M					
#	ALGORITHM	BYTES	MSEC	R=1	R=5	R=10	R=20	R=50	WORK-10
1	3DIVI-003	39	512	120	625	201	0.0833	195	0.0444
2	3DIVI-004	211	4096	121	628	161	0.0175	154	0.0091
3	3DIVI-005	207	4096	122	653	161	0.0176	155	0.0075
4	3DIVI-006	47	528	129	653	172	0.0240	179	0.0171
5	ACER-000	41	512	29	201	14	0.0106	126	0.0051
6	ALCHERA-000	135	2048	42	263	157	0.0161	166	0.0124
7	ALCHERA-001	136	2048	66	22	226	0.9869	226	0.9735
8	ALCHERA-002	146	2048	14	115	201	0.0949	200	0.0555
9	ALCHERA-003	136	2048	109	548	138	0.0104	128	0.0054
10	ALLQOVISION-000	117	2048	81	425	144	0.0114	149	0.0084
11	ALLQOVISION-001	119	2048	79	792	129	0.0090	124	0.0048
12	ANKE-000	188	2072	83	431	159	0.0132	140	0.0073
13	ANKE-001	185	2072	85	433	153	0.0132	141	0.0073
14	ANKE-002	177	2056	123	641	67	0.0028	68	0.0020
15	AWARE-003	187	2076	161	716	179	0.0306	176	0.0162
16	AWARE-004	92	712	19	6679	192	0.0679	187	0.0348
17	AWARE-005	198	3100	188	827	180	0.0311	177	0.0167
18	AWARE-006	3	124	182	818	19	0.0697	194	0.0369
19	AYONIX-000	74	1036	1	10	220	0.4505	221	0.3540
20	AYONIX-001	72	1036	3	12	216	0.3414	215	0.2338
21	AYONIX-002	71	1036	2	11	217	0.3414	216	0.2338
22	CAMVI-003	64	1024	154	707	191	0.0520	199	0.0517
23	CAMVI-004	69	1024	163	718	189	0.0468	197	0.0465
24	CAMVI-005	63	1024	173	769	194	0.0652	201	0.0648
25	CIB-000	227	8196	135	674	24	0.0015	22	0.0013
26	CLOUDWALK-HR-000	144	2048	213	908	18	0.0015	34	0.0014
27	COGENT-000	44	525	110	551	159	0.0105	159	0.0096
28	COGENT-001	45	525	111	552	140	0.0105	160	0.0096
29	COGENT-002	76	1043	226	987	81	0.0036	77	0.0022
30	COGENT-003	75	1043	224	960	83	0.0038	87	0.0024
31	COGENT-004	173	2053	222	952	40	0.0020	42	0.0016
32	COGNITEC-000	169	2052	19	176	174	0.0252	172	0.0136
33	COGNITEC-001	171	2052	30	202	14	0.0117	134	0.0062
34	COGNITEC-002	171	2052	35	227	10	0.0057	107	0.0037
35	COGNITEC-003	162	2052	52	297	110	0.0062	114	0.0040
36	COGNITEC-004	171	2052	26	192	21	0.0032	71	0.0020
37	CYBERLINK-000	164	2052	148	699	85	0.0040	96	0.0028
38	CYBERLINK-001	161	2052	84	433	79	0.0035	83	0.0023
39	CYBERLINK-002	222	4140	169	738	61	0.0026	79	0.0023
40	CYBERLINK-003	62	612	147	696	24	0.0016	24	0.0013
41	CYBERLINK-004	227	6212	170	738	27	0.0017	37	0.0015
42	DAHUA-000	157	2048	69	378	133	0.0093	137	0.0066
43	DAHUA-001	111	2048	65	371	114	0.0067	115	0.0040
44	DAHUA-002	133	2048	149	699	34	0.0018	36	0.0015
45	DAHUA-003	151	2048	166	725	10	0.0012	9	0.0010
46	DEEPLINT-001	210	4096	139	687	16	0.0014	27	0.0014
47	DEEPSA-001	111	2048	177	780	99	0.0043	78	0.0022
48	DERMALOG-003	6	128	33	211	205	0.1259	204	0.0744
49	DERMALOG-004	4	128	31	208	205	0.1251	203	0.0739
50	DERMALOG-005	5	128	103	532	156	0.0149	169	0.0129
51	DERMALOG-006	126	256	100	514	124	0.0081	139	0.0069
52	DERMALOG-007	7	128	79	413	132	0.0092	138	0.0066
53	DERMALOG-008	57	512	64	370	68	0.0029	67	0.0020
54	EYEDEA-003	73	1036	70	385	199	0.0800	196	0.0451
55	FS-001	139	2048	196	851	149	0.0120	161	0.0105
56	GLORY-000	29	418	15	160	20	0.1781	211	0.1391
57	GLORY-001	97	1726	76	405	205	0.1268	206	0.0967
58	GORILLA-001	185	2156	17	169	199	0.0603	187	0.0304
59	GORILLA-002	79	1132	61	341	168	0.0197	156	0.0092
60	GORILLA-003	187	2156	113	563	18	0.0361	174	0.0146
61	GORILLA-004	190	2192	72	395	111	0.0063	104	0.0032
62	GORILLA-005	228	6288	95	483	28	0.0032	58	0.0019
63	HIK-003	84	1408	122	633	146	0.0117	133	0.0060
64	HIK-004	87	1152	91	510	140	0.0113	131	0.0059
65	HIK-005	86	1408	119	619	99	0.0046	88	0.0025
66	HIK-006	85	1408	116	610	94	0.0046	89	0.0025
67	IDEMIA-003	46	528	141	689	117	0.0069	120	0.0045
68	IDEMIA-004	48	528	133	669	113	0.0066	111	0.0038
69	IDEMIA-005	28	352	67	374	124	0.0081	118	0.0044
70	IDEMIA-006	27	352	66	373	136	0.0096	127	0.0052
71	IDEMIA-007	56	860	181	807	69	0.0026	43	0.0016
72	IDEMIA-008	26	300	88	451	6	0.0011	7	0.0009

Table 20: Rank-based accuracy for the FRVT 2018 mugshot sets. In columns 3 and 4 are template size and template generation duration. Thereafter values are rank-based FNIR with $T = 0$ and FPIR = 1. This is appropriate to investigational uses but not those with higher volumes where candidates from all searches would need review. The next column is a workload statistic, a small value shows an algorithm front-loads mates into the first 10 candidates. Throughout, blue superscripts indicate the rank of the algorithm for that column, and the best value is highlighted in yellow.

MISSES OUTSIDE RANK R FNIR(N, T=0, R)		RESOURCE USAGE TEMPLATE		ENROL MOST RECENT, N = 1.6M FRVT 2018 MUGSHOTS					
#	ALGORITHM	BYTES	MSEC	R=1	R=5	R=10	R=20	R=50	WORK-10
217	VOCORD-005	⁵⁵ 768	¹⁸³ 822	¹¹⁸ 0.0070	¹²² 0.0046	¹²⁴ 0.0041	¹²⁹ 0.0038	¹³⁶ 0.0035	¹¹⁹ 1.044
218	VOCORD-006	²²⁸ 10240	¹⁸⁴ 825	²²⁷ 1.0000	²²⁷ 1.0000	²²⁷ 1.0000	²²⁷ 1.0000	²²⁸ 10.000	
219	VTS-000	¹⁴⁹ 2048	⁹⁶ 492	²²³ 0.5937	²²⁴ 0.5936	²²⁴ 0.5936	²²⁴ 0.5936	²²⁴ 6.343	
220	VTS-001	¹³¹ 2048	²⁰⁷ 891	²² 0.0015	¹⁸ 0.0012	¹⁸ 0.0011	¹⁸ 0.0011	¹⁷ 0.0010	¹⁷ 1.011
221	XFORWARDAI-000	¹⁰⁴ 2048	¹⁷² 768	⁴⁹ 0.0023	⁷⁰ 0.0020	⁷⁷ 0.0020	⁸⁹ 0.0019	⁹⁵ 0.0019	⁶⁵ 1.018
222	XFORWARDAI-001	¹¹⁵ 2048	¹³⁸ 681	⁴⁴ 0.0020	⁶⁴ 0.0019	⁷⁴ 0.0019	⁸⁸ 0.0019	⁹⁴ 0.0019	⁵⁸ 1.018
223	XFORWARDAI-002	²⁰⁸ 4096	²¹⁹ 935	³⁸ 0.0020	⁶¹ 0.0019	⁷² 0.0019	⁸⁸ 0.0019	⁹³ 0.0019	⁵⁶ 1.017
224	YISHENG-001	²⁰² 3704	⁷¹ 387	¹⁷⁶ 0.0265	¹²⁰ 0.0130	¹⁶⁷ 0.0102	¹⁶⁵ 0.0080	¹⁵⁶ 0.0059	¹⁷¹ 1.134
225	YITU-002	²²¹ 4138	²⁰² 870	³⁹ 0.0018	²⁰ 0.0012	¹⁹ 0.0011	¹⁹ 0.0011	²¹ 0.0010	²¹ 1.012
226	YITU-003	²²⁰ 4138	²⁰³ 871	⁶⁷ 0.0029	⁸⁰ 0.0023	⁹⁰ 0.0022	⁹⁷ 0.0021	¹⁰² 0.0021	⁷⁷ 1.021
227	YITU-004	¹⁸³ 2070	²¹⁴ 910	¹¹ 0.0013	⁶ 0.0009	⁸ 0.0009	⁹ 0.0009	⁹ 0.0009	⁸ 1.009
228	YITU-005	¹⁸² 2070	¹⁹⁹ 861	⁵¹ 0.0023	⁷² 0.0021	⁷⁸ 0.0020	⁹¹ 0.0020	⁹⁷ 0.0020	⁶⁷ 1.019

Table 23: Rank-based accuracy for the FRVT 2018 mugshot sets. In columns 3 and 4 are template size and template generation duration. Thereafter values are rank-based FNIR with $T = 0$ and FPIR = 1. This is appropriate to investigational uses but not those with higher volumes where candidates from all searches would need review. The next column is a workload statistic, a small value shows an algorithm front-loads mates into the first 10 candidates. Throughout, blue superscripts indicate the rank of the algorithm for that column, and the best value is highlighted in yellow.

Appendices

Appendix A Accuracy on large-population FRVT 2018 mugshots

2021/08/05 02:00:58	$\text{FNIR}(N, R, T) =$ $\text{FPTR}(N, T) =$	False neg. identification rate False pos. identification rate	$N =$ Num. enrolled subjects $R =$ Num. candidates examined	$T =$ Threshold $T > 0 \rightarrow$ Identification	$T = 0 \rightarrow$ Investigation
------------------------	---	--	--	---	-----------------------------------

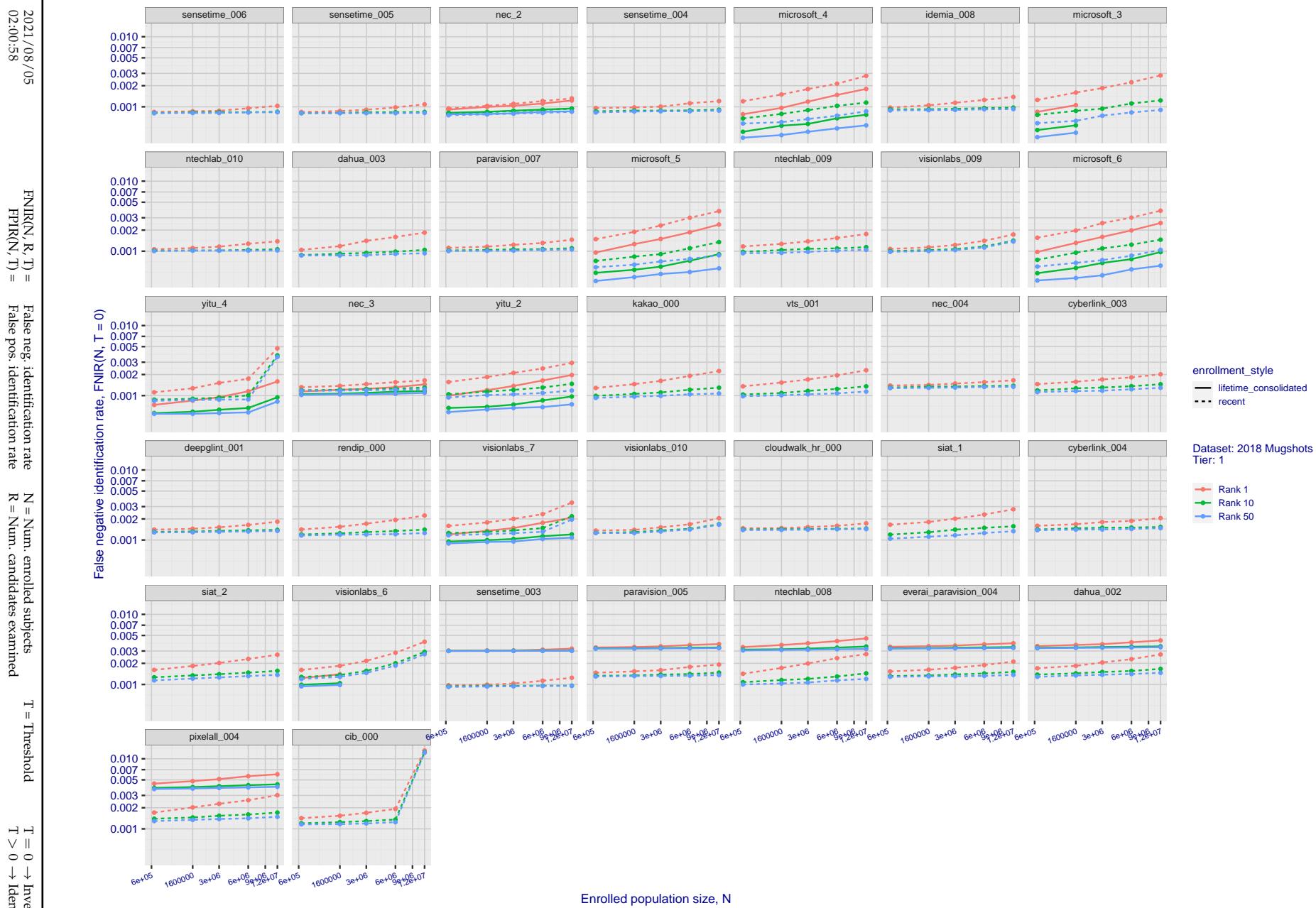


Figure 20: [FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. number of enrolled subjects. The figure shows false negative identification rates, $\text{FNIR}(N, R)$, across various gallery sizes and ranks 1, 10 and 50. The threshold is set to zero, so this metric rewards even weak scoring rank 1 mates. This also means $\text{FPIR} = 1$, so any search without an enrolled mate will return non-mated candidates. For clarity, results are sorted and reported into tiers spanning multiple pages, the tiering criteria being rank 1 hit rate on a gallery size of 640 000.

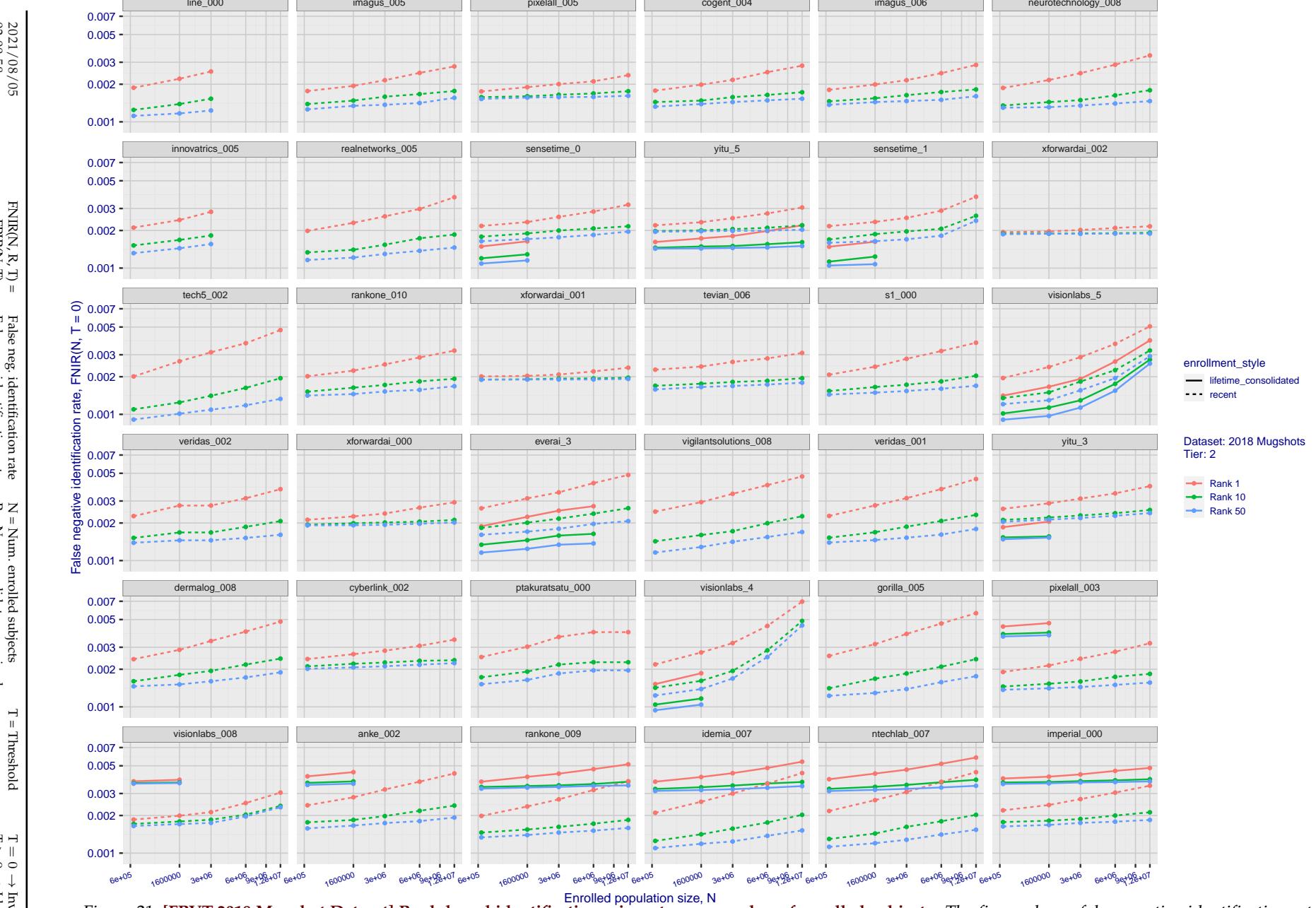


Figure 21: [FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. number of enrolled subjects. The figure shows false negative identification rates, $\text{FNIR}(N, R)$, across various gallery sizes and ranks 1, 10 and 50. The threshold is set to zero, so this metric rewards even weak scoring rank 1 mates. This also means $\text{FPIR} = 1$, so any search without an enrolled mate will return non-mated candidates. For clarity, results are sorted and reported into tiers spanning multiple pages, the tiering criteria being rank 1 hit rate on a gallery size of 640 000.

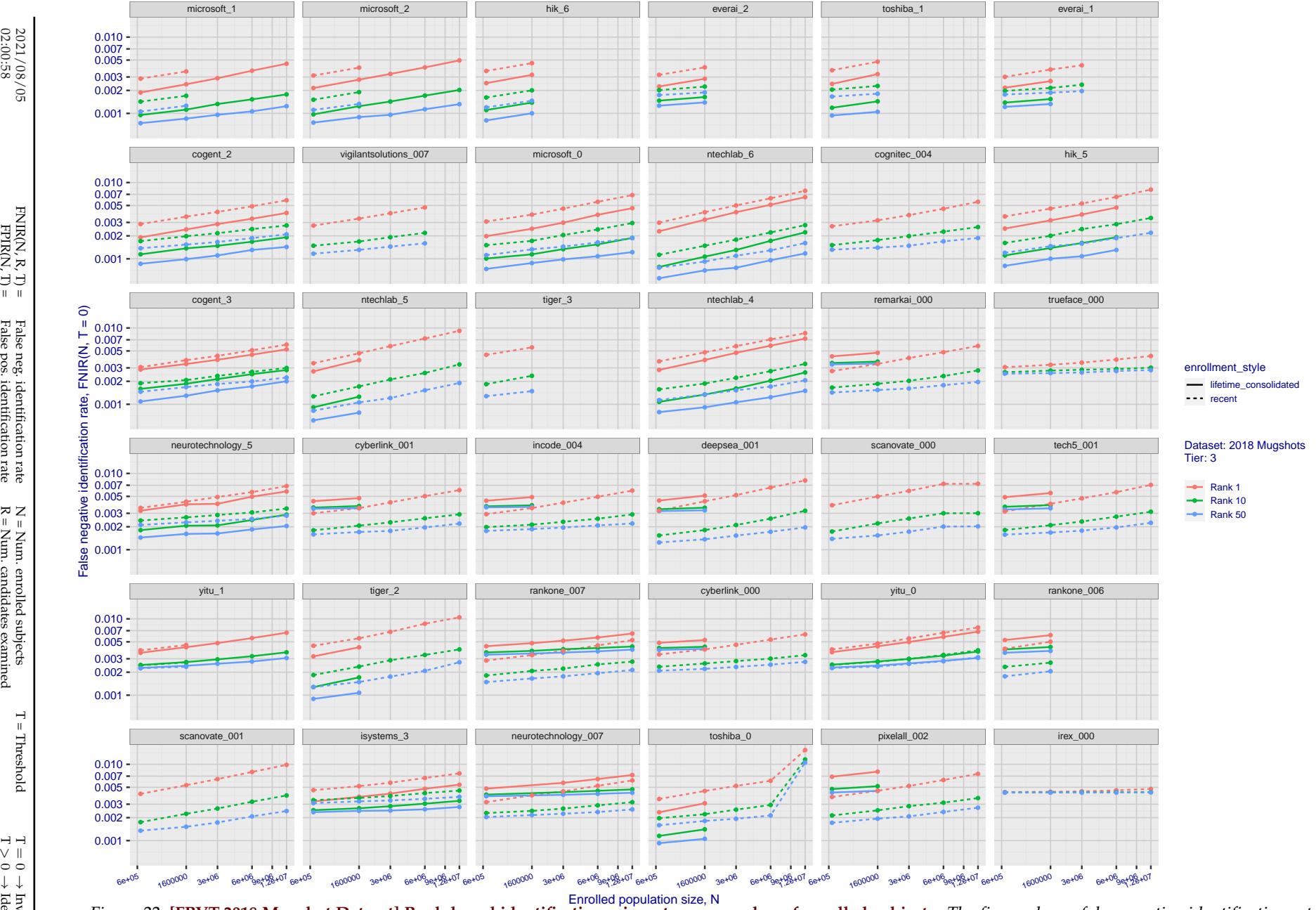


Figure 22: [FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. number of enrolled subjects. The figure shows false negative identification rates, $\text{FNIR}(N, R)$, across various gallery sizes and ranks 1, 10 and 50. The threshold is set to zero, so this metric rewards even weak scoring rank 1 mates. This also means $\text{FPIR} = 1$, so any search without an enrolled mate will return non-mated candidates. For clarity, results are sorted and reported into tiers spanning multiple pages, the tiering criteria being rank 1 hit rate on a gallery size of 640 000.

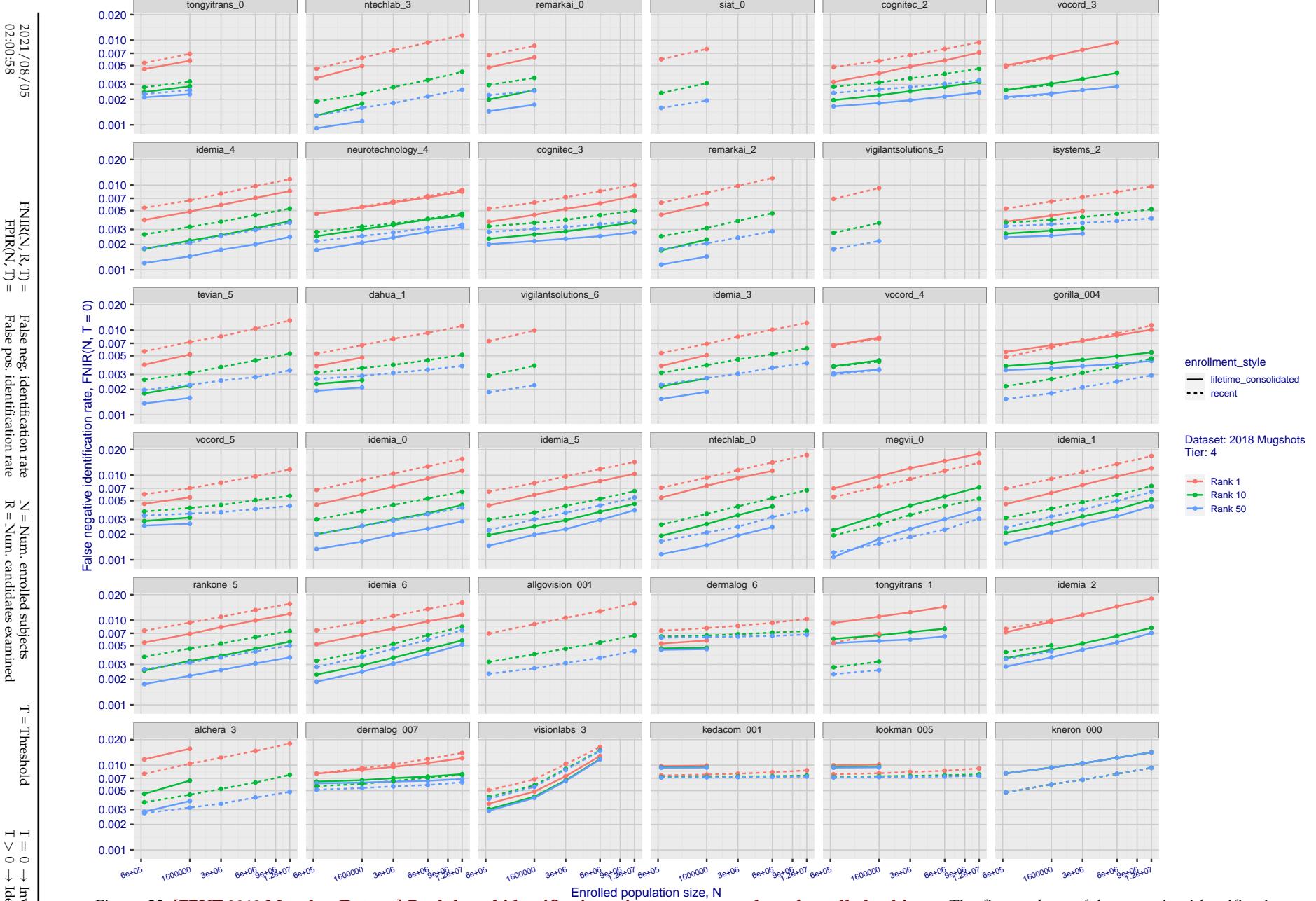


Figure 23: [FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. number of enrolled subjects. The figure shows false negative identification rates, $\text{FNIR}(N, R)$, across various gallery sizes and ranks 1, 10 and 50. The threshold is set to zero, so this metric rewards even weak scoring rank 1 mates. This also means $\text{FPIR} = 1$, so any search without an enrolled mate will return non-mated candidates. For clarity, results are sorted and reported into tiers spanning multiple pages, the tiering criteria being rank 1 hit rate on a gallery size of 640 000.

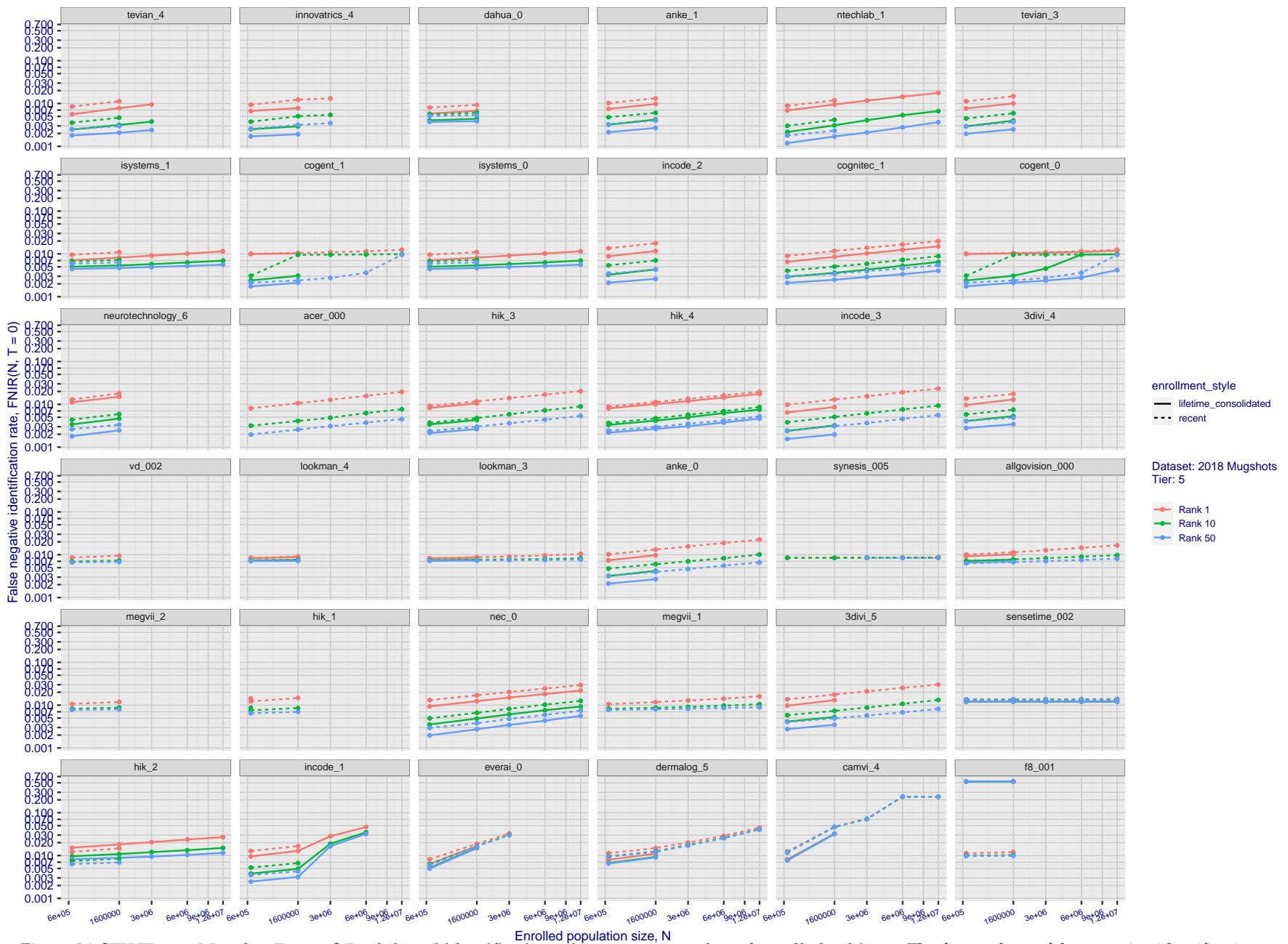


Figure 24: [FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. number of enrolled subjects. The figure shows false negative identification rates, $\text{FNIR}(N, R)$, across various gallery sizes and ranks 1, 10 and 50. The threshold is set to zero, so this metric rewards even weak scoring rank 1 mates. This also means $\text{FPIR} = 1$, so any search without an enrolled mate will return non-mated candidates. For clarity, results are sorted and reported into tiers spanning multiple pages, the tiering criteria being rank 1 hit rate on a gallery size of 640 000.

2021/08/05
02:00:58FNIR(N, R, T) =
False neg. identification rate
FPIR(N, T) =
False pos. identification rateN = Num. enrolled subjects
R = Num. candidates examined

T = Threshold

T = 0 → Investigation
T > 0 → Identification

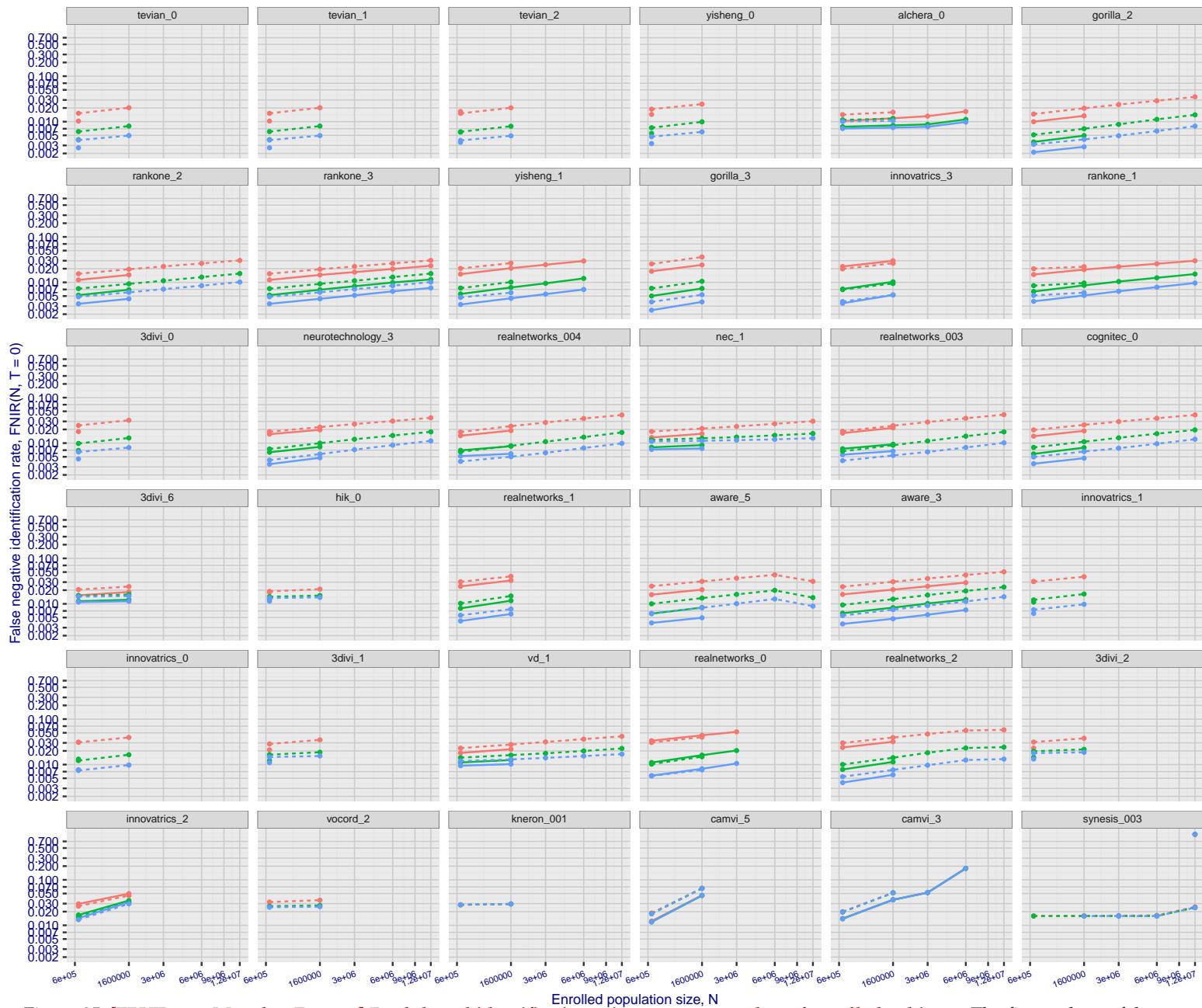


Figure 25: [FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. number of enrolled subjects. The figure shows false negative identification rates, $\text{FNIR}(N, R)$, across various gallery sizes and ranks 1, 10 and 50. The threshold is set to zero, so this metric rewards even weak scoring rank 1 mates. This also means $\text{FPIR} = 1$, so any search without an enrolled mate will return non-mated candidates. For clarity, results are sorted and reported into tiers spanning multiple pages, the tiering criteria being rank 1 hit rate on a gallery size of 640 000.

2021/08/05

02:00:58

FNIR(N, R, T) = False neg. identification rate N = Num. enrolled subjects R = Num. candidates examined T = Threshold $T = 0 \rightarrow$ Investigation

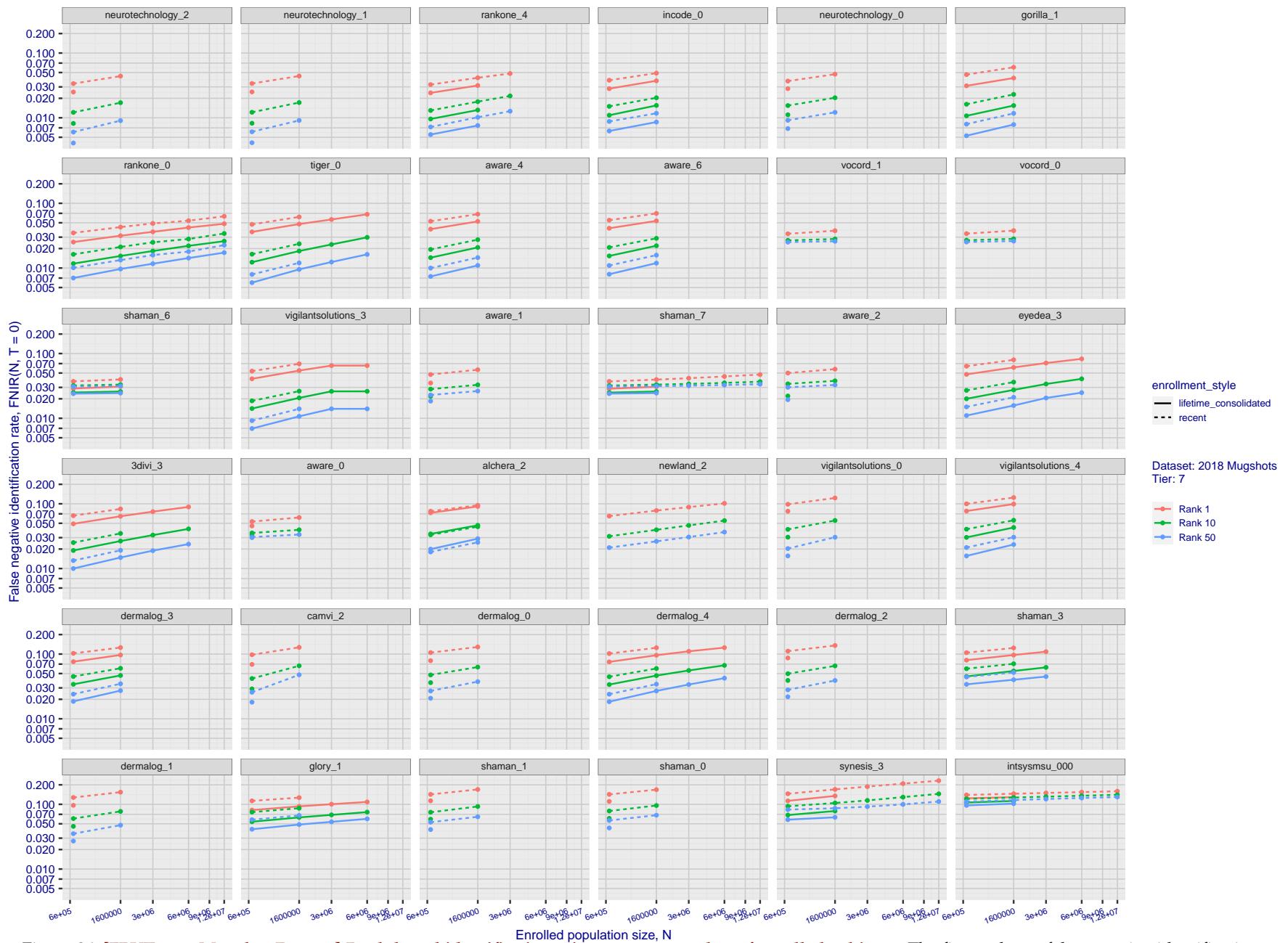


Figure 26: [FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. number of enrolled subjects. The figure shows false negative identification rates, $\text{FNIR}(N, R)$, across various gallery sizes and ranks 1, 10 and 50. The threshold is set to zero, so this metric rewards even weak scoring rank 1 mates. This also means $\text{FPIR} = 1$, so any search without an enrolled mate will return non-mated candidates. For clarity, results are sorted and reported into tiers spanning multiple pages, the tiering criteria being rank 1 hit rate on a gallery size of 640 000.

2021/08/05
02:00:58FNIR(N, R, T) =
False neg. identification rate
FPIR(N, T) =
False pos. identification rateN = Num. enrolled subjects
R = Num. candidates examinedT = Threshold
T = 0 → Investigation
T > 0 → Identification

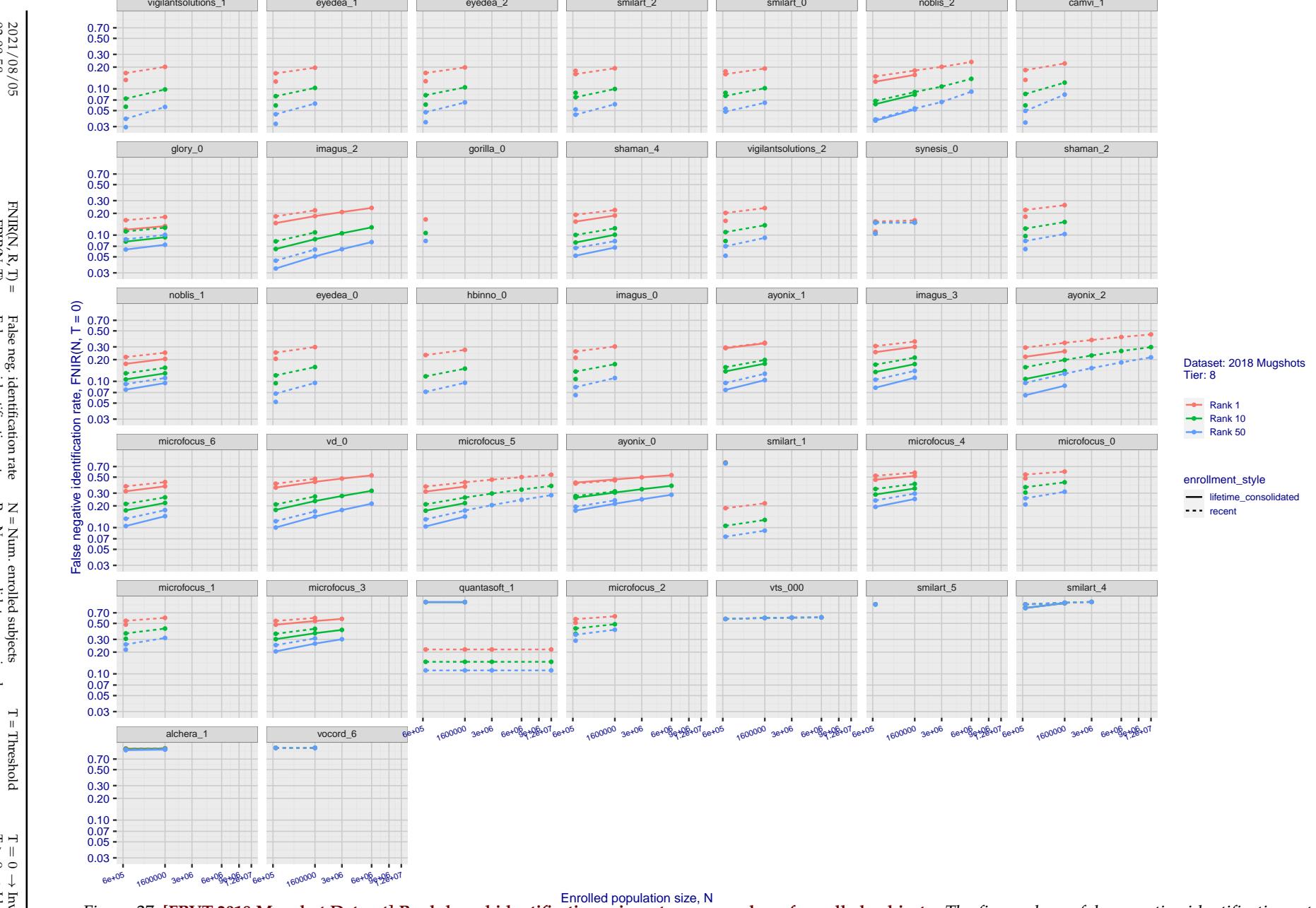


Figure 27: [FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. number of enrolled subjects. The figure shows false negative identification rates, $\text{FNIR}(N, R)$, across various gallery sizes and ranks 1, 10 and 50. The threshold is set to zero, so this metric rewards even weak scoring rank 1 mates. This also means $\text{FPIR} = 1$, so any search without an enrolled mate will return non-mated candidates. For clarity, results are sorted and reported into tiers spanning multiple pages, the tiering criteria being rank 1 hit rate on a gallery size of 640 000.

2021/08/05 02:00:58	$\text{FNIR}(N, R, T) =$ $\text{FPTR}(N, T) =$	False neg. identification rate False pos. identification rate	$N =$ Num. enrolled subjects $R =$ Num. candidates examined	$T =$ Threshold $T > 0 \rightarrow$ Identification	$T = 0 \rightarrow$ Investigation
------------------------	---	--	--	---	-----------------------------------

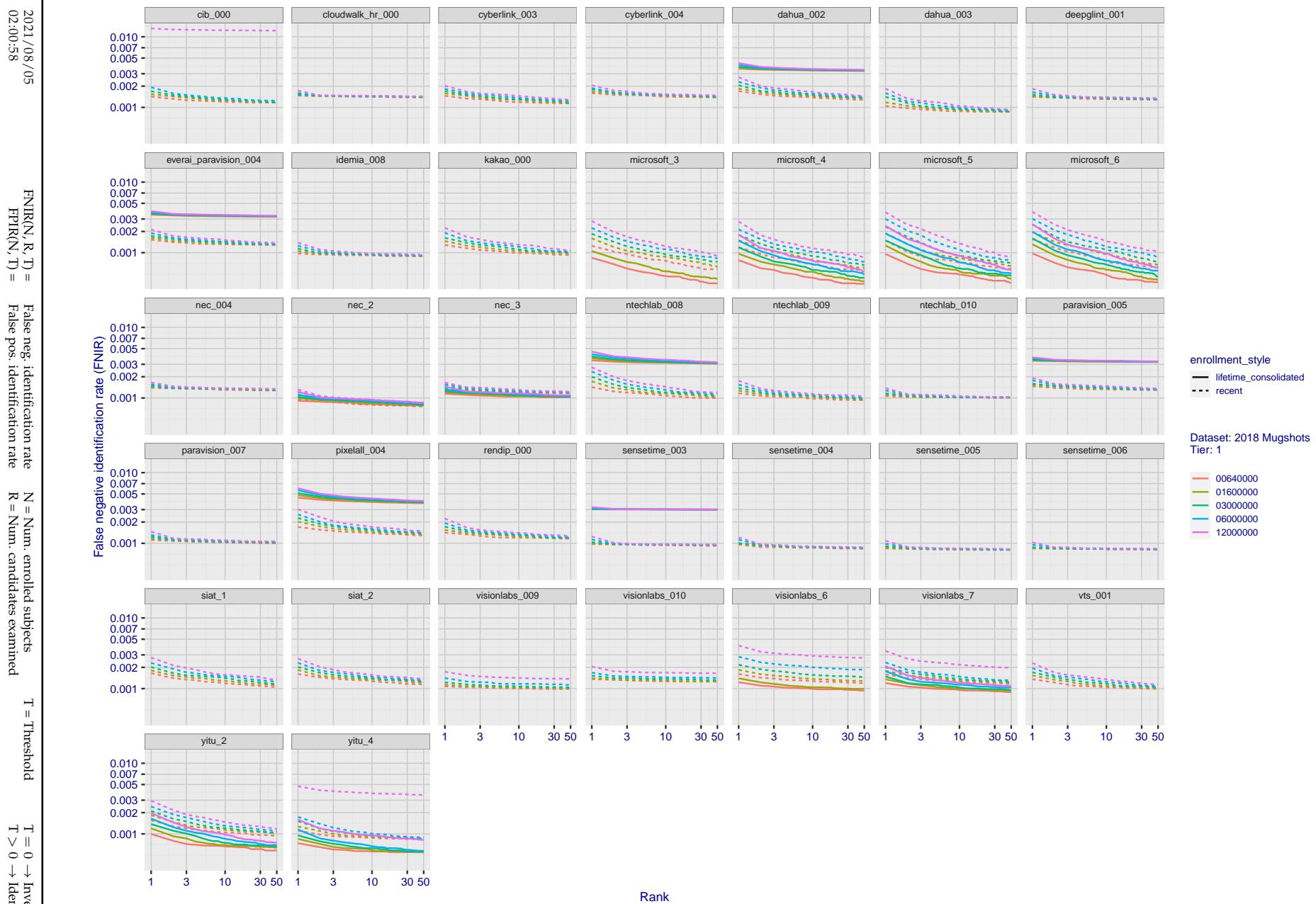


Figure 28: [FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. rank. The figure shows false negative identification rates (FNIR) for ranks up to 50. This metric is appropriate to investigational applications where human reviewers will adjudicate sorted candidate lists. Note that with threshold set to zero, FPIR = 1, i.e. any search without an enrolled mate will return non-mated candidates. Results are sorted and reported into tiers for clarity, with the tiering criteria being rank 1 hit rate on a gallery size of $N = 640\,000$ subjects.

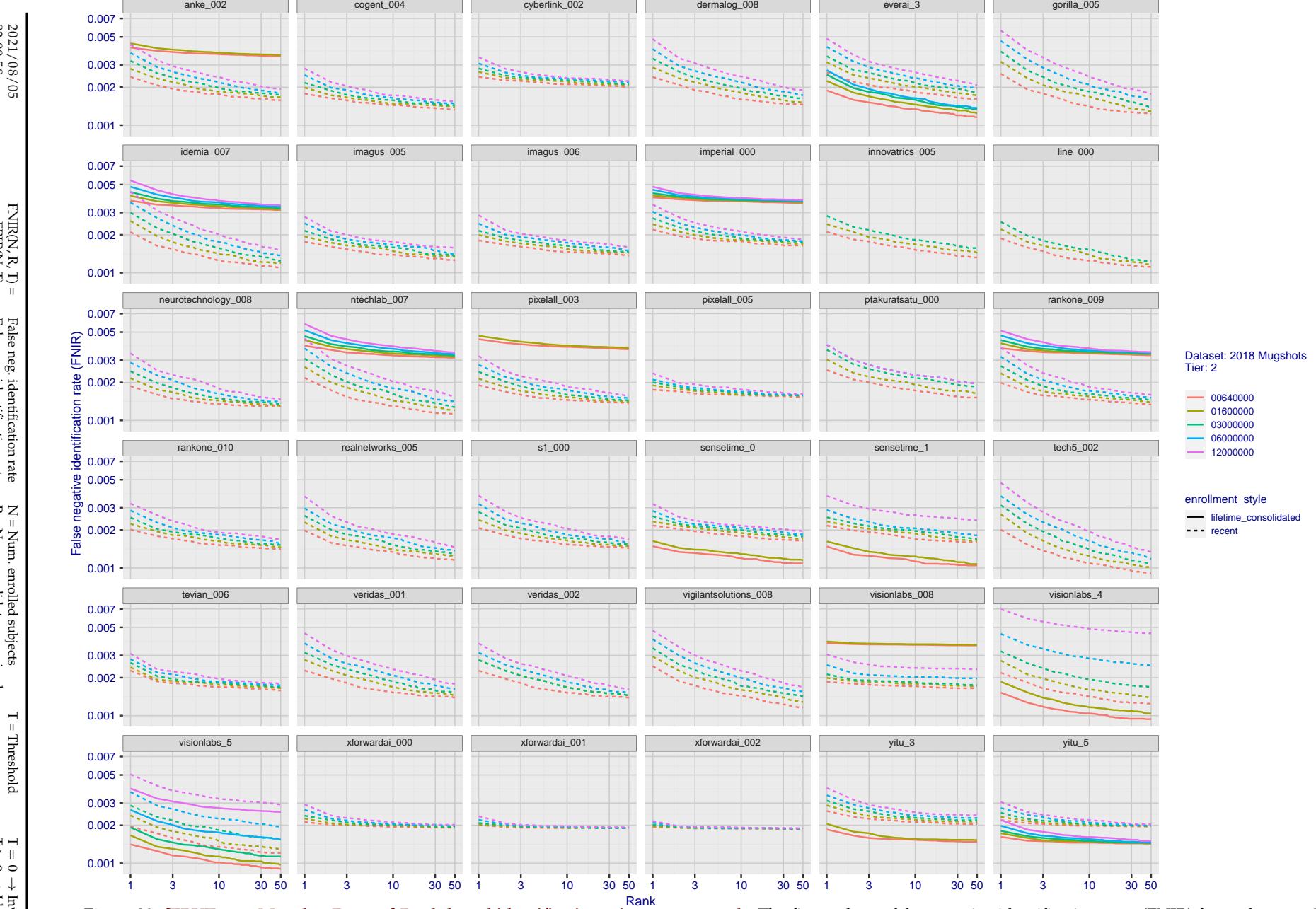


Figure 29: [FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. rank. The figure shows false negative identification rates (FNIR) for ranks up to 50. This metric is appropriate to investigational applications where human reviewers will adjudicate sorted candidate lists. Note that with threshold set to zero, FPIR = 1, i.e. any search without an enrolled mate will return non-mated candidates. Results are sorted and reported into tiers for clarity, with the tiering criteria being rank 1 hit rate on a gallery size of $N = 640\,000$ subjects.

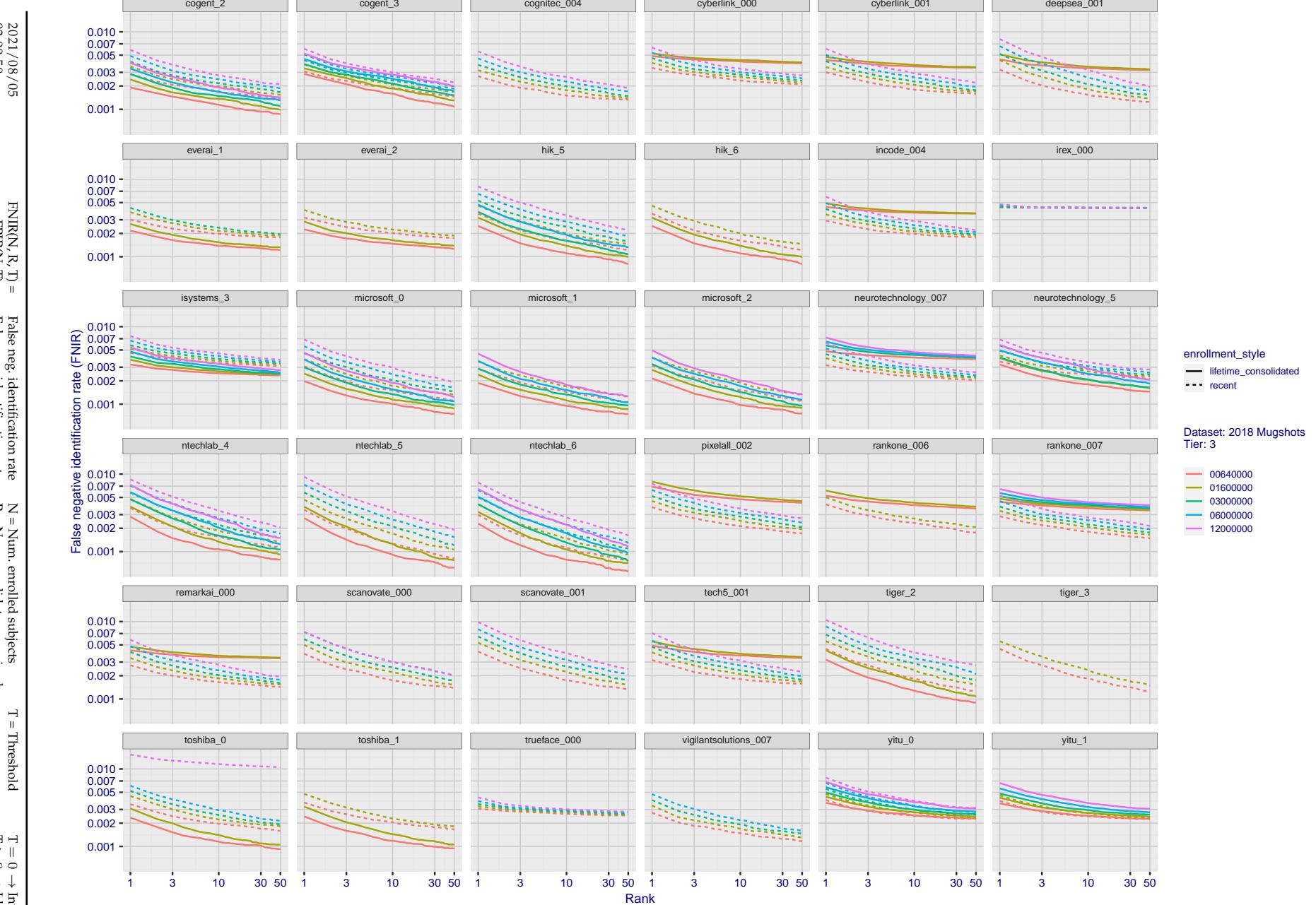


Figure 30: [FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. rank. The figure shows false negative identification rates (FNIR) for ranks up to 50. This metric is appropriate to investigational applications where human reviewers will adjudicate sorted candidate lists. Note that with threshold set to zero, FPTR = 1, i.e. any search without an enrolled mate will return non-mated candidates. Results are sorted and reported into tiers for clarity, with the tiering criteria being rank 1 hit rate on a gallery size of $N = 640\,000$ subjects.

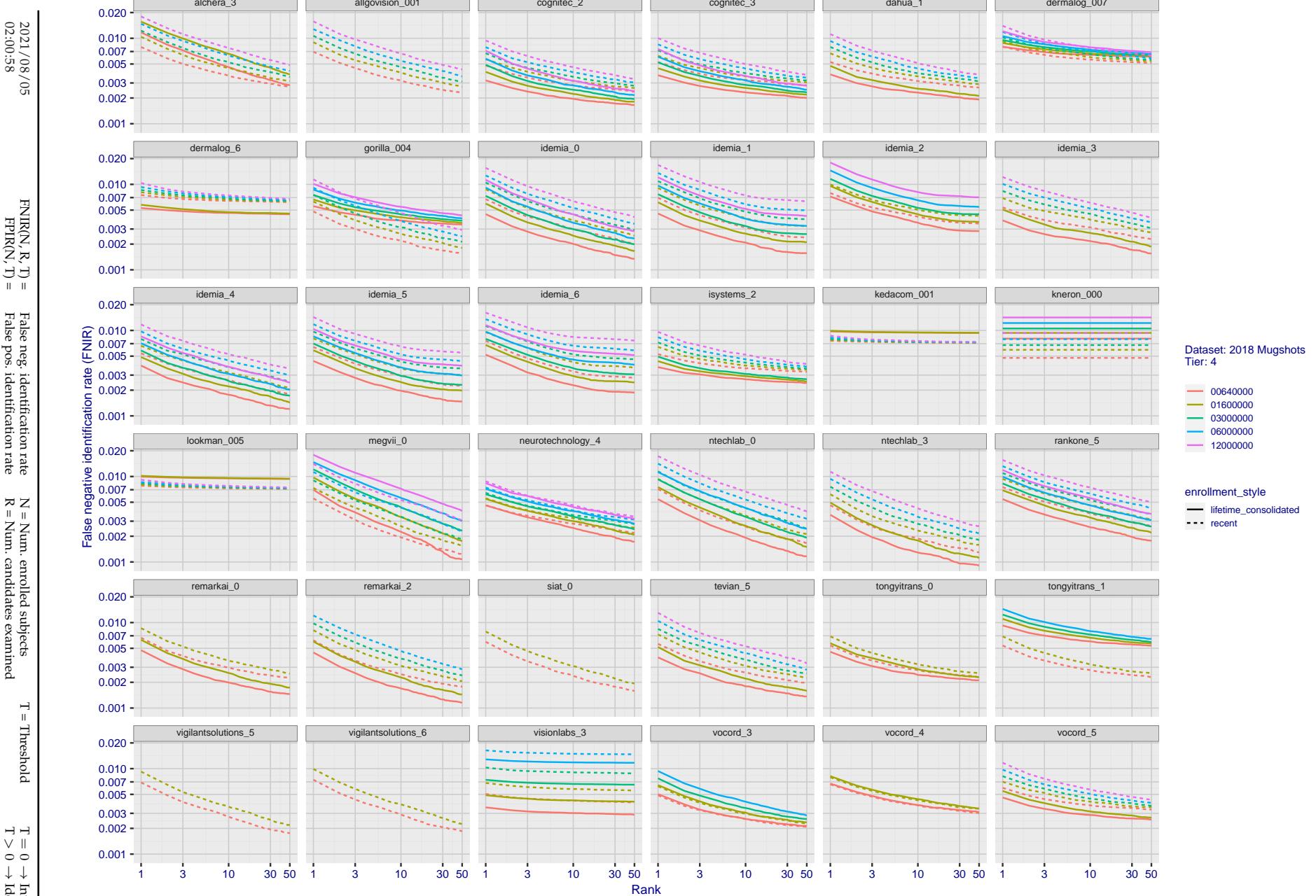


Figure 31: [FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. rank. The figure shows false negative identification rates (FNIR) for ranks up to 50. This metric is appropriate to investigational applications where human reviewers will adjudicate sorted candidate lists. Note that with threshold set to zero, FPIR = 1, i.e. any search without an enrolled mate will return non-mated candidates. Results are sorted and reported into tiers for clarity, with the tiering criteria being rank 1 hit rate on a gallery size of N = 640 000 subjects.

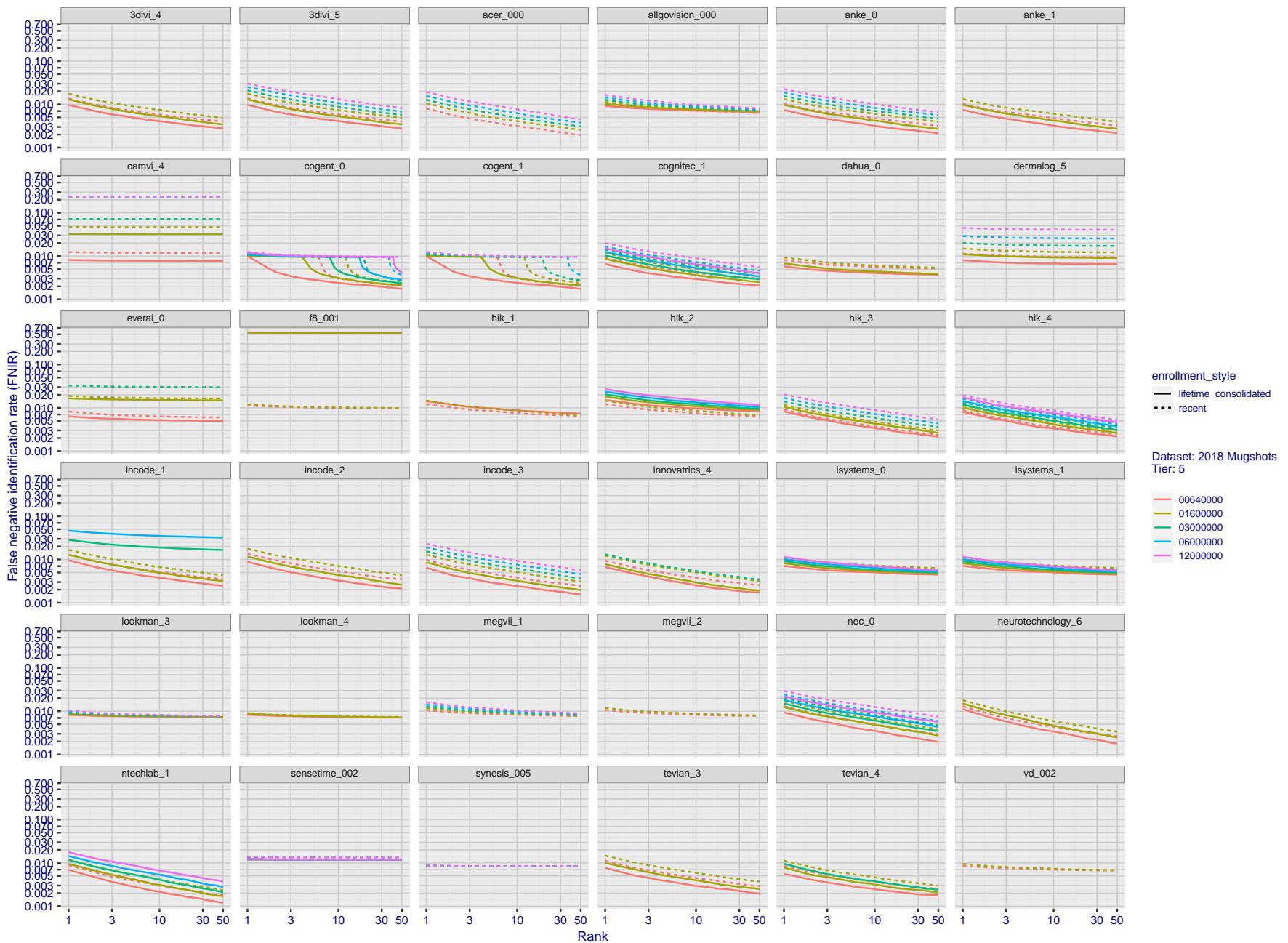


Figure 32: [FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. rank. The figure shows false negative identification rates (FNIR) for ranks up to 50. This metric is appropriate to investigational applications where human reviewers will adjudicate sorted candidate lists. Note that with threshold set to zero, FPIR = 1, i.e. any search without an enrolled mate will return non-mated candidates. Results are sorted and reported into tiers for clarity, with the tiering criteria being rank 1 hit rate on a gallery size of N = 640 000 subjects.

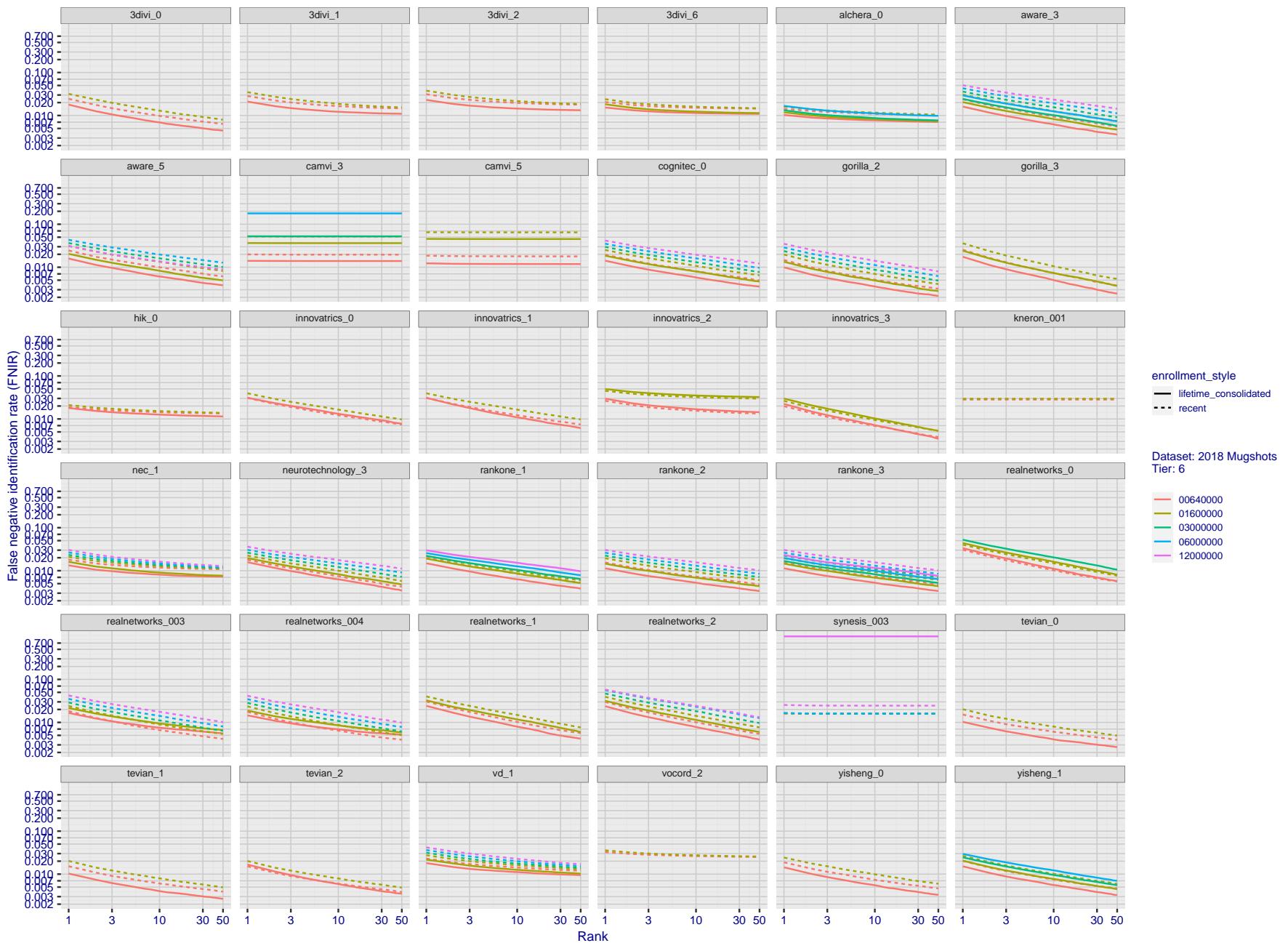


Figure 33: [FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. rank. The figure shows false negative identification rates (FNIR) for ranks up to 50. This metric is appropriate to investigational applications where human reviewers will adjudicate sorted candidate lists. Note that with threshold set to zero, FPIR = 1, i.e. any search without an enrolled mate will return non-mated candidates. Results are sorted and reported into tiers for clarity, with the tiering criteria being rank 1 hit rate on a gallery size of $N = 640\,000$ subjects.

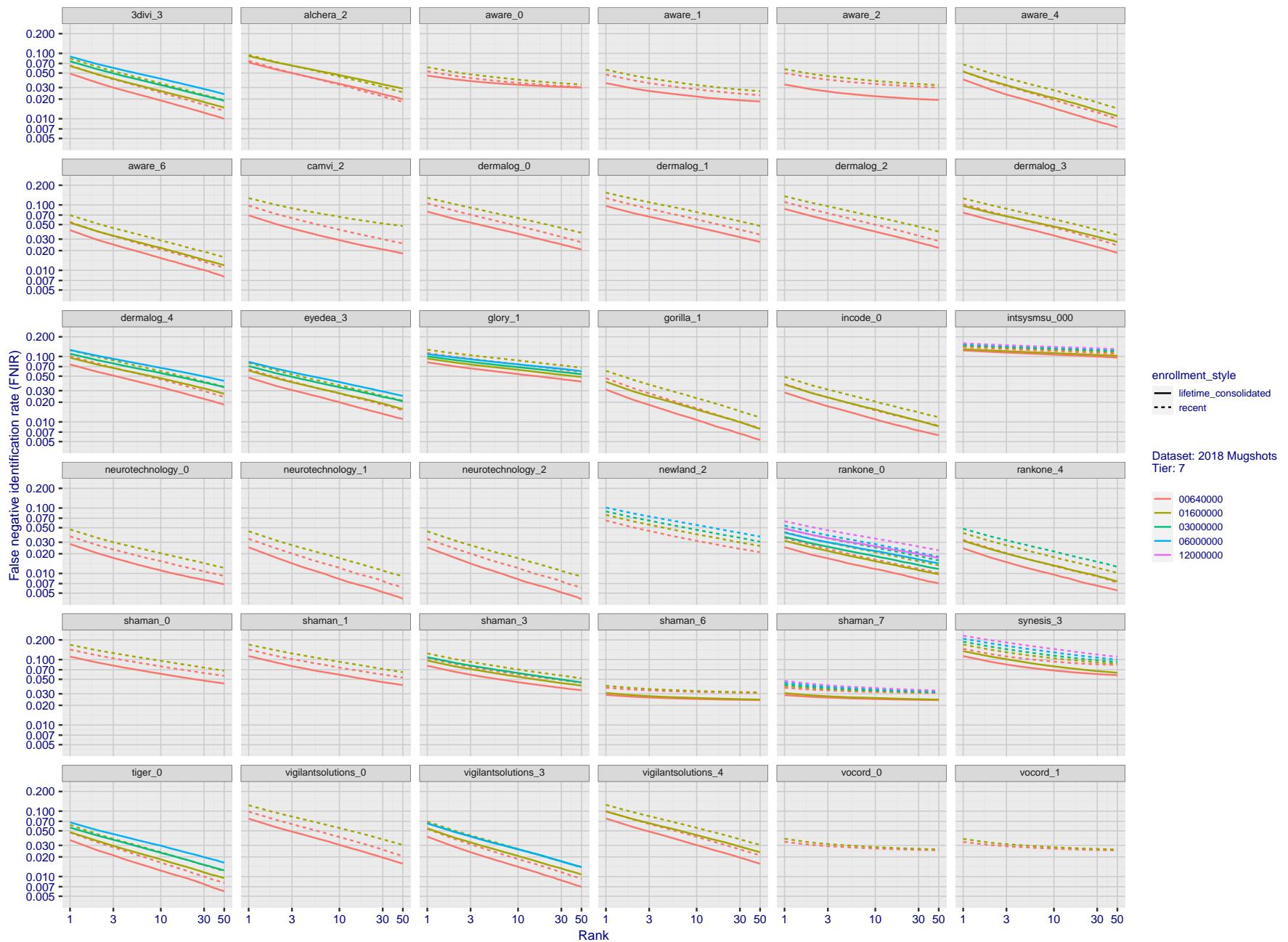


Figure 34: [FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. rank. The figure shows false negative identification rates (FNIR) for ranks up to 50. This metric is appropriate to investigational applications where human reviewers will adjudicate sorted candidate lists. Note that with threshold set to zero, FPTR = 1, i.e. any search without an enrolled mate will return non-mated candidates. Results are sorted and reported into tiers for clarity, with the tiering criteria being rank 1 hit rate on a gallery size of $N = 640\,000$ subjects.

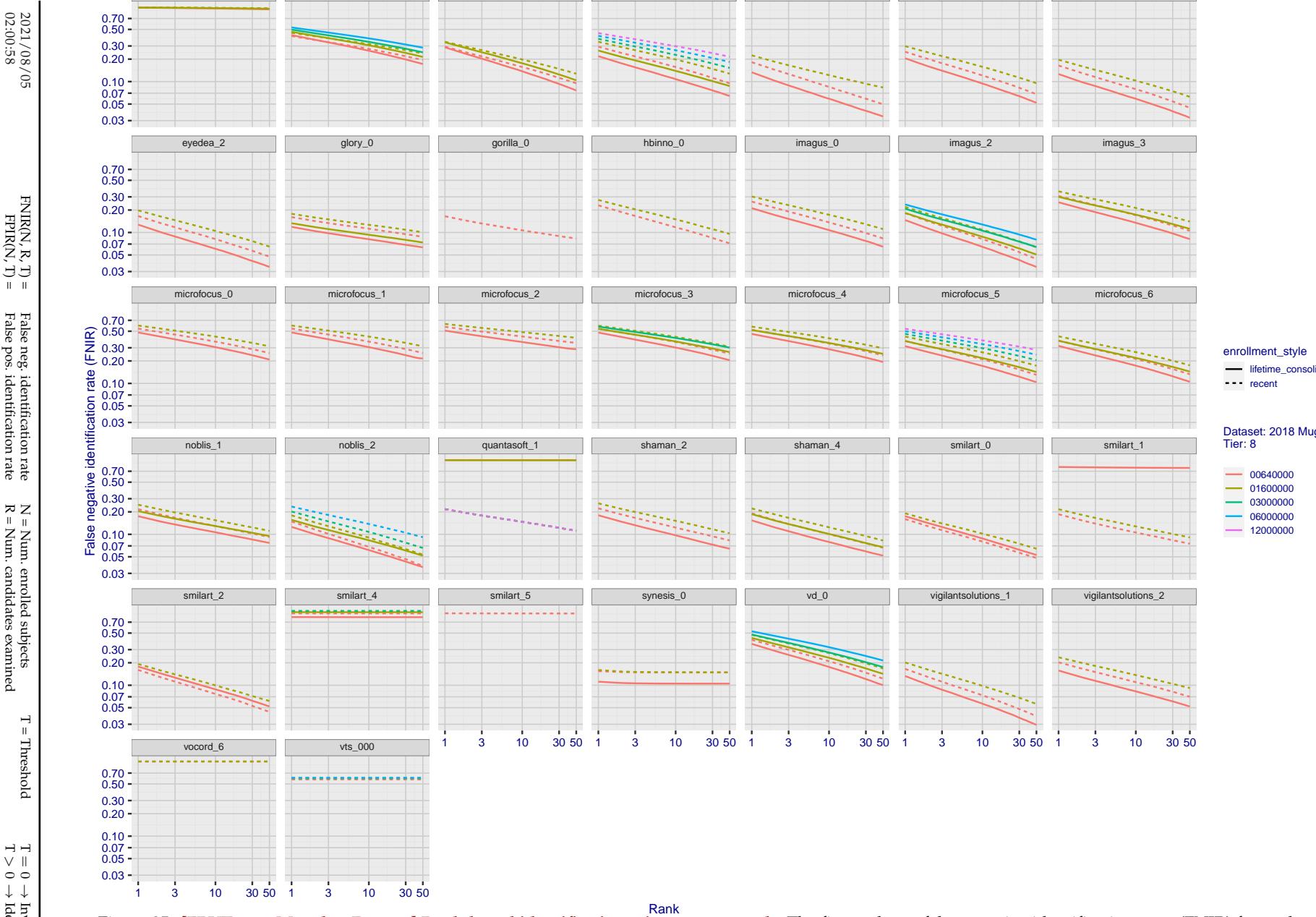


Figure 35: [FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. rank. The figure shows false negative identification rates (FNIR) for ranks up to 50. This metric is appropriate to investigational applications where human reviewers will adjudicate sorted candidate lists. Note that with threshold set to zero, FPIR = 1, i.e. any search without an enrolled mate will return non-mated candidates. Results are sorted and reported into tiers for clarity, with the tiering criteria being rank 1 hit rate on a gallery size of N = 640 000 subjects.

2021/08/05
02:00:58

FNIR(N, R, T) = False neg. identification rate
FPTR(N, T) = False pos. identification rate

N = Num. enrolled subjects
R = Num. candidates examined

T = Threshold
T > 0 → Identification

T = 0 → Investigation

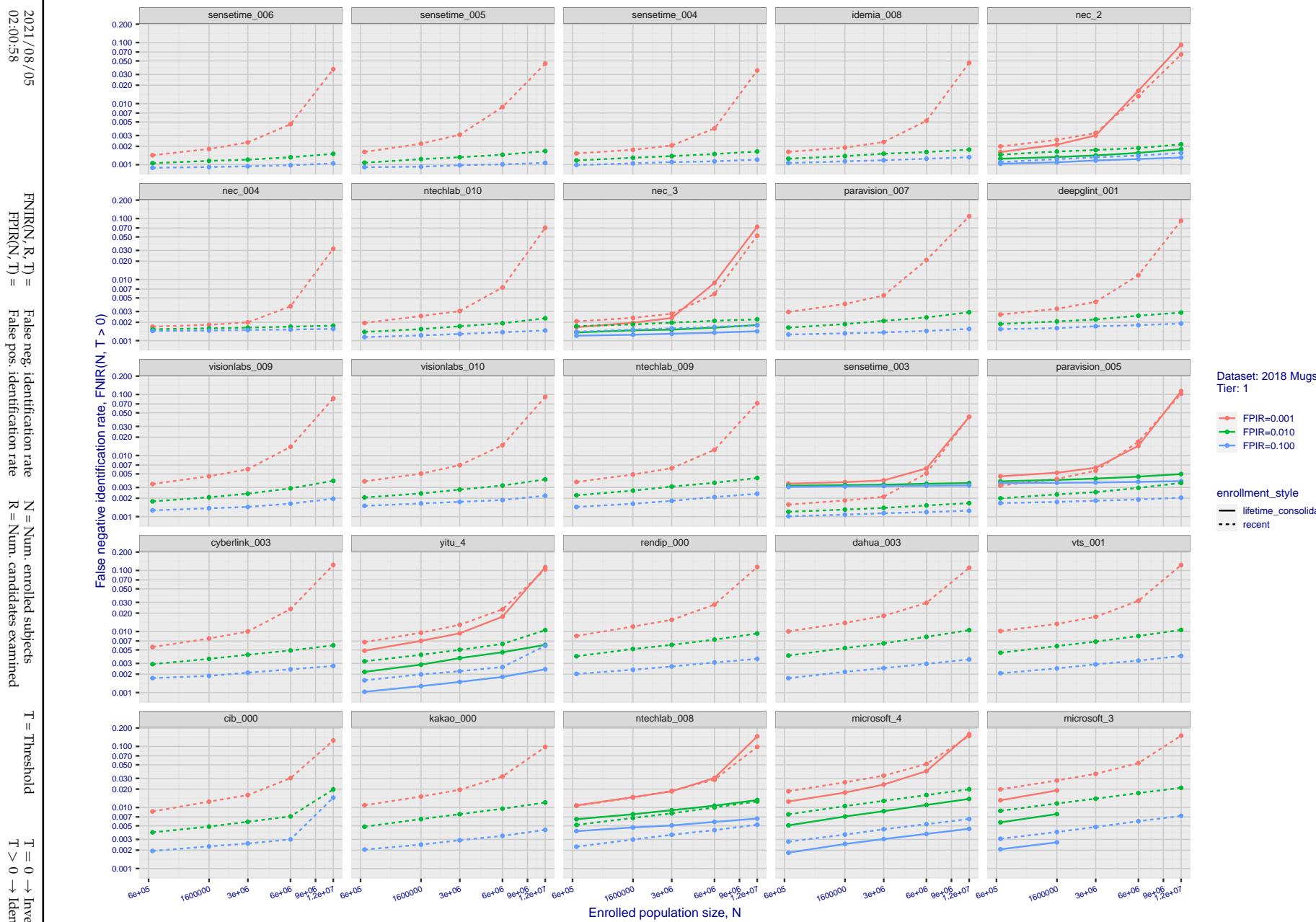


Figure 36: [FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects. The figure shows $\text{FNIR}(N, T)$ across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 1. Less accurate algorithms were not run on large N , so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by $\text{FNIR}(N_b, 1, 0)$, then sorting by median $\text{FNIR}(N_b, T)$, $N_b = 640\,000$.

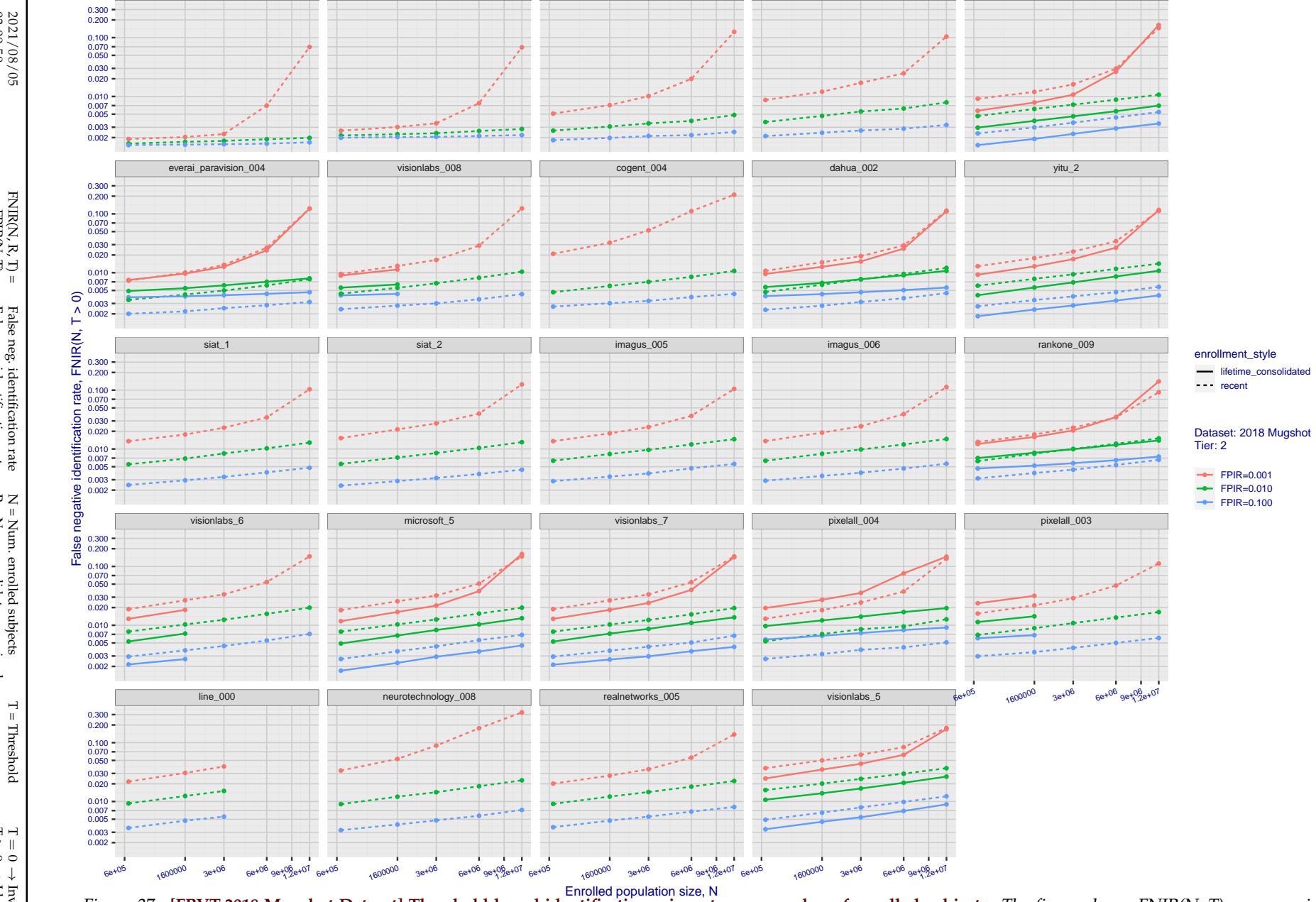


Figure 37: [FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects. The figure shows $\text{FNIR}(N, T)$ across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 1. Less accurate algorithms were not run on large N , so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by $\text{FNIR}(N_b, 1, 0)$, then sorting by median $\text{FNIR}(N_b, T)$, $N_b = 640\,000$.

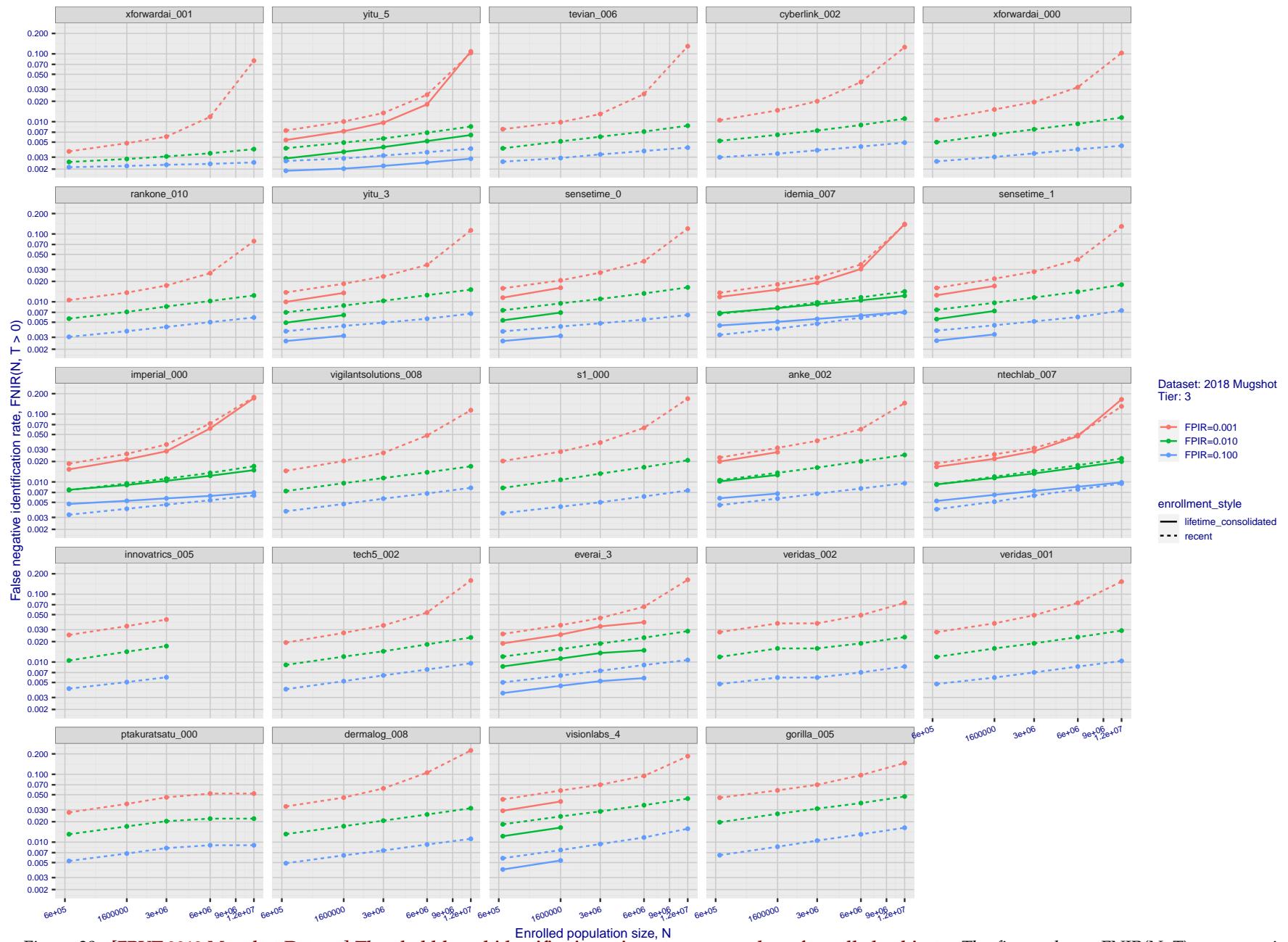


Figure 38: [FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects. The figure shows $\text{FNIR}(N, T)$ across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 1. Less accurate algorithms were not run on large N , so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by $\text{FNIR}(N_b, 1, 0)$, then sorting by median $\text{FNIR}(N_b, T)$, $N_b = 640\,000$.

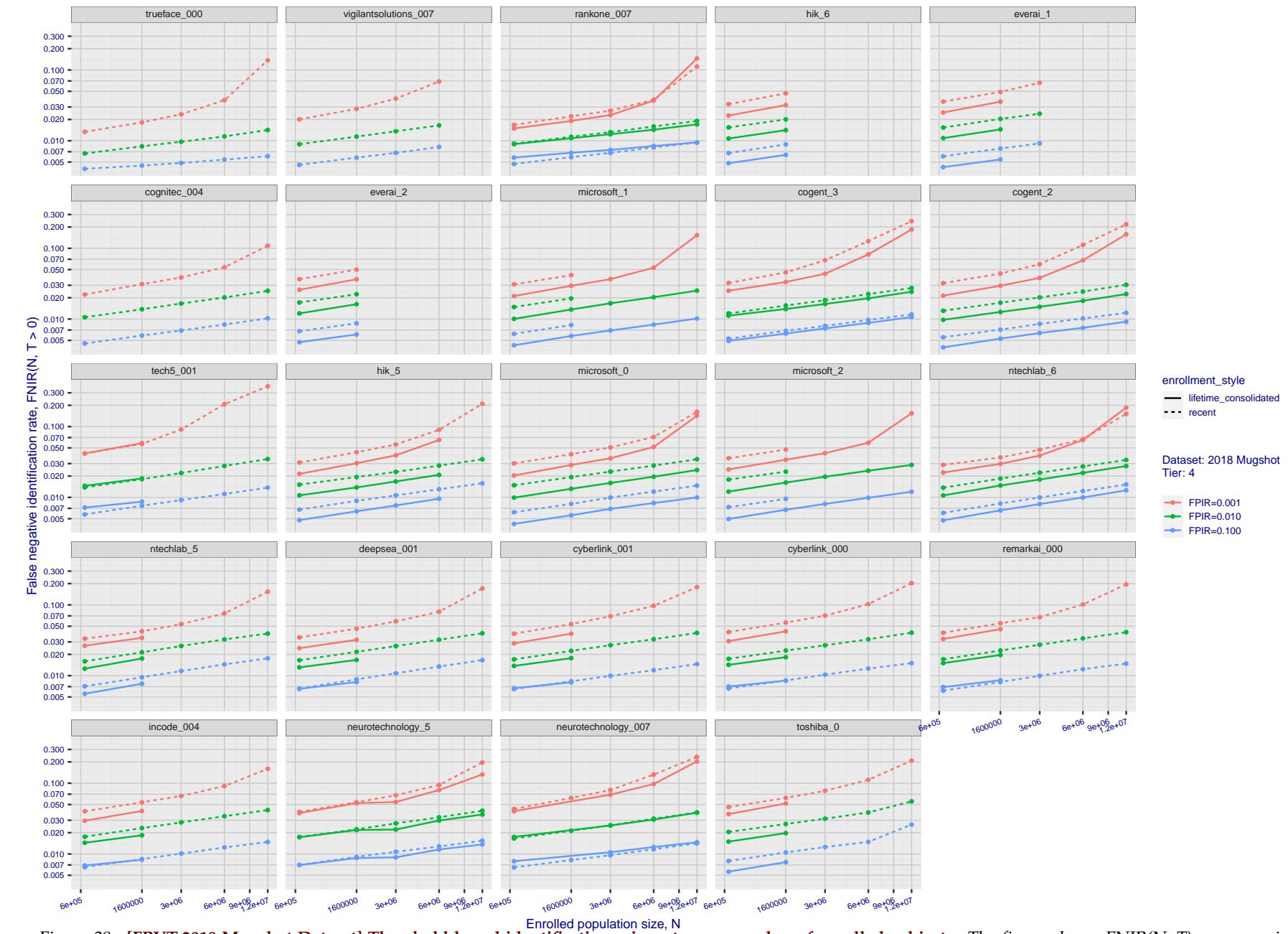


Figure 39: [FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects. The figure shows $\text{FNIR}(N, T)$ across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 1. Less accurate algorithms were not run on large N , so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by $\text{FNIR}(N_b, 1, 0)$, then sorting by median $\text{FNIR}(N_b, T)$, $N_b = 640\,000$.

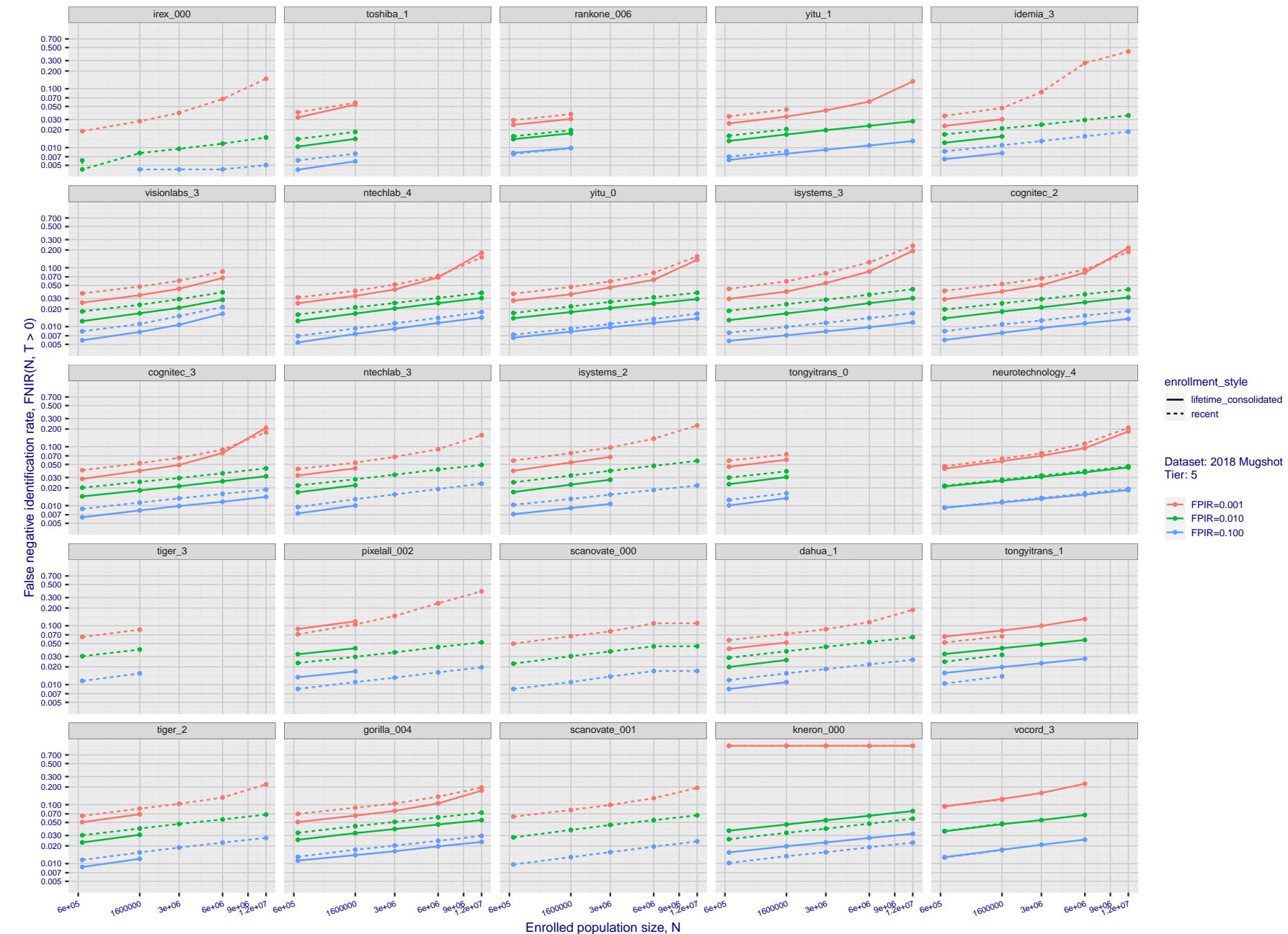


Figure 40: [FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects. The figure shows FNIR(N, T) across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 1. Less accurate algorithms were not run on large N , so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by $\text{FNIR}(N_b, 1, 0)$, then sorting by median $\text{FNIR}(N_b, T)$, $N_b = 640\,000$.

2021/08/05
02:00:58FNIR($N, R, T > 0$)
FPIR(N, T) = False neg. identification rate
FPIR(N) = False pos. identification rate N = Num. enrolled subjects
 R = Num. candidates examined T = Threshold $T = 0 \rightarrow$ Investigation
 $T > 0 \rightarrow$ Identification

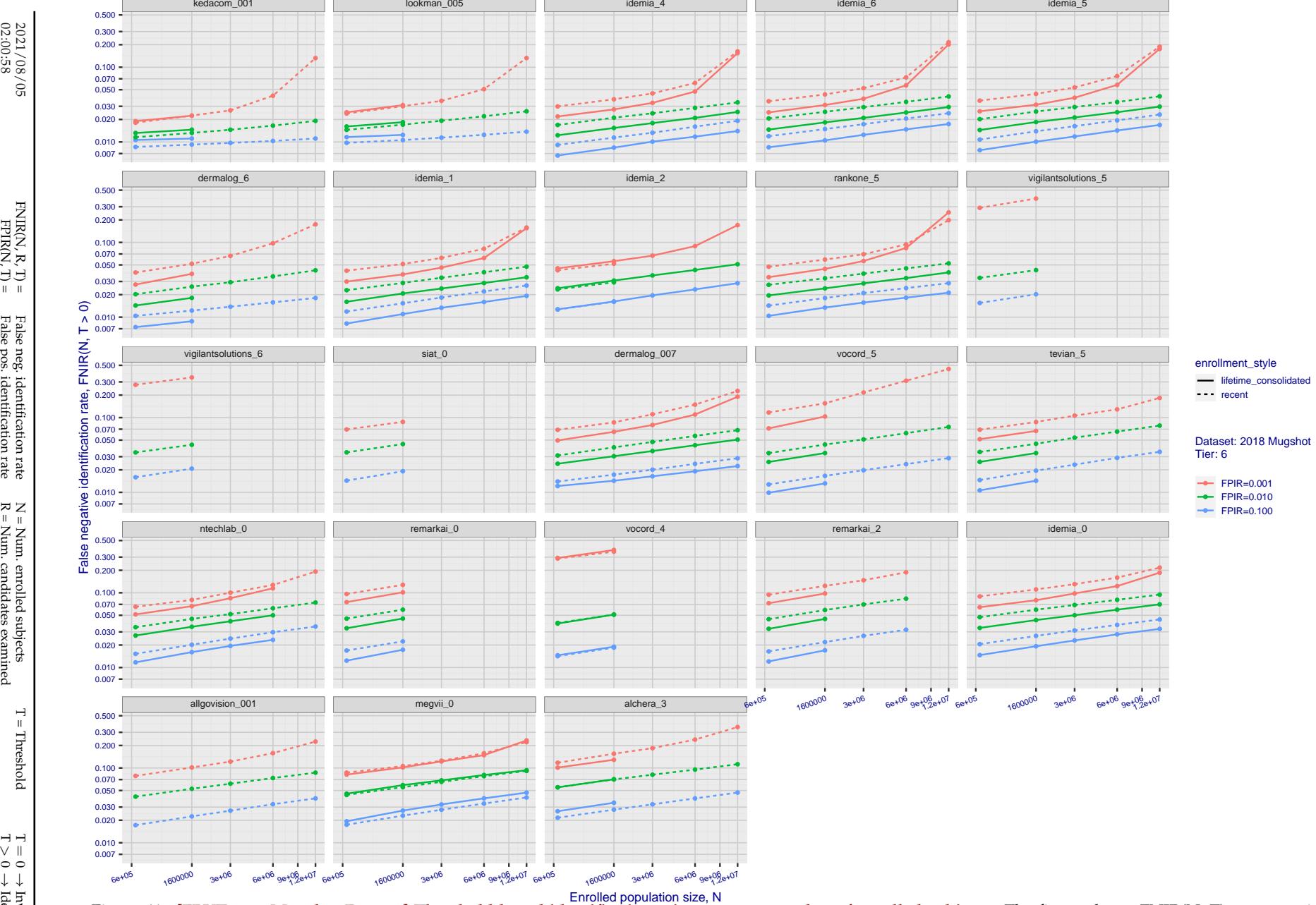


Figure 41: [FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects. The figure shows $\text{FNIR}(N, T)$ across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 1. Less accurate algorithms were not run on large N , so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by $\text{FNIR}(N_b, 1, 0)$, then sorting by median $\text{FNIR}(N_b, T)$, $N_b = 640\,000$.

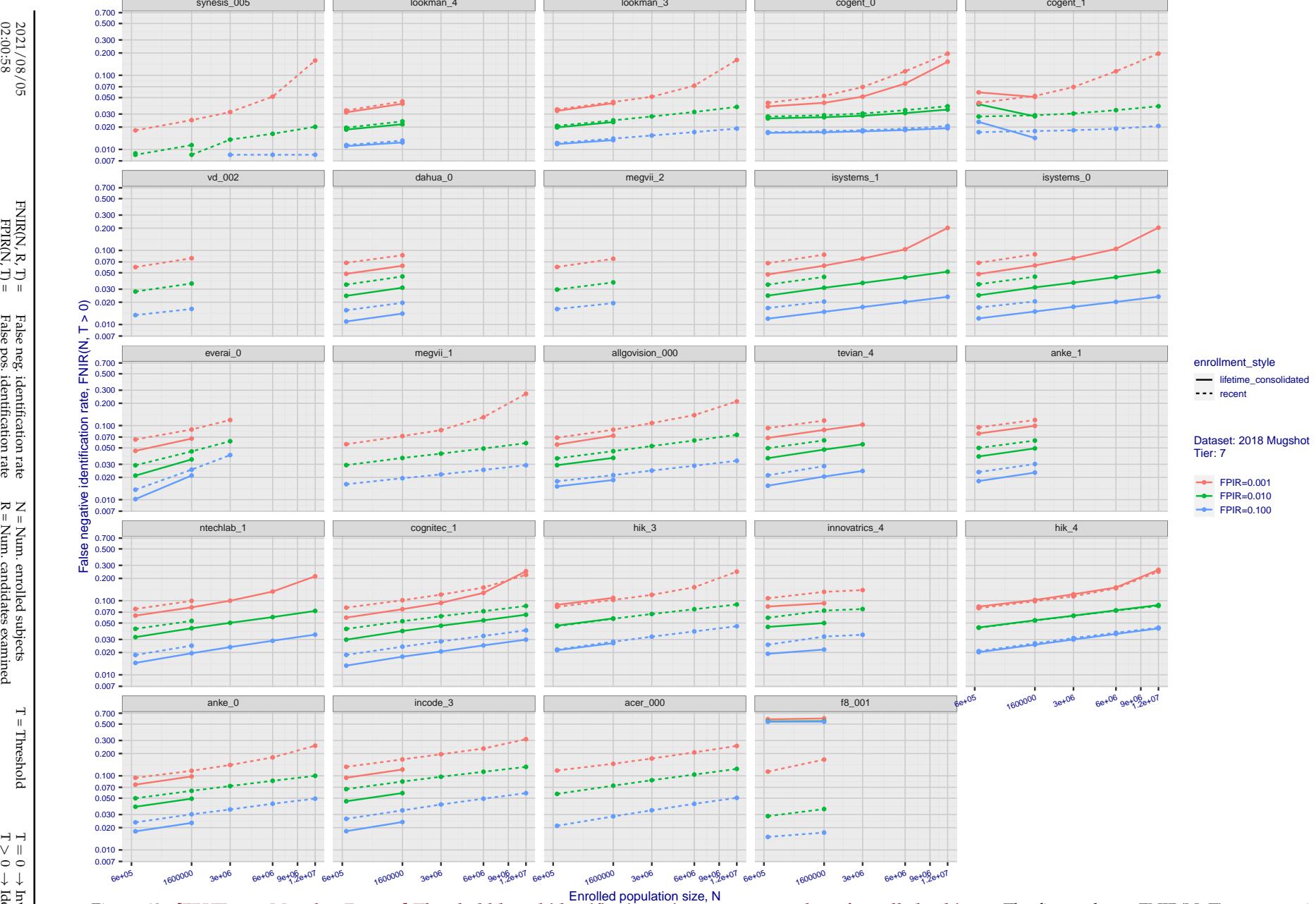


Figure 42: [FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects. The figure shows $\text{FNIR}(N, T)$ across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 1. Less accurate algorithms were not run on large N , so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by $\text{FNIR}(N_b, 1, 0)$, then sorting by median $\text{FNIR}(N_b, T)$, $N_b = 640\,000$.

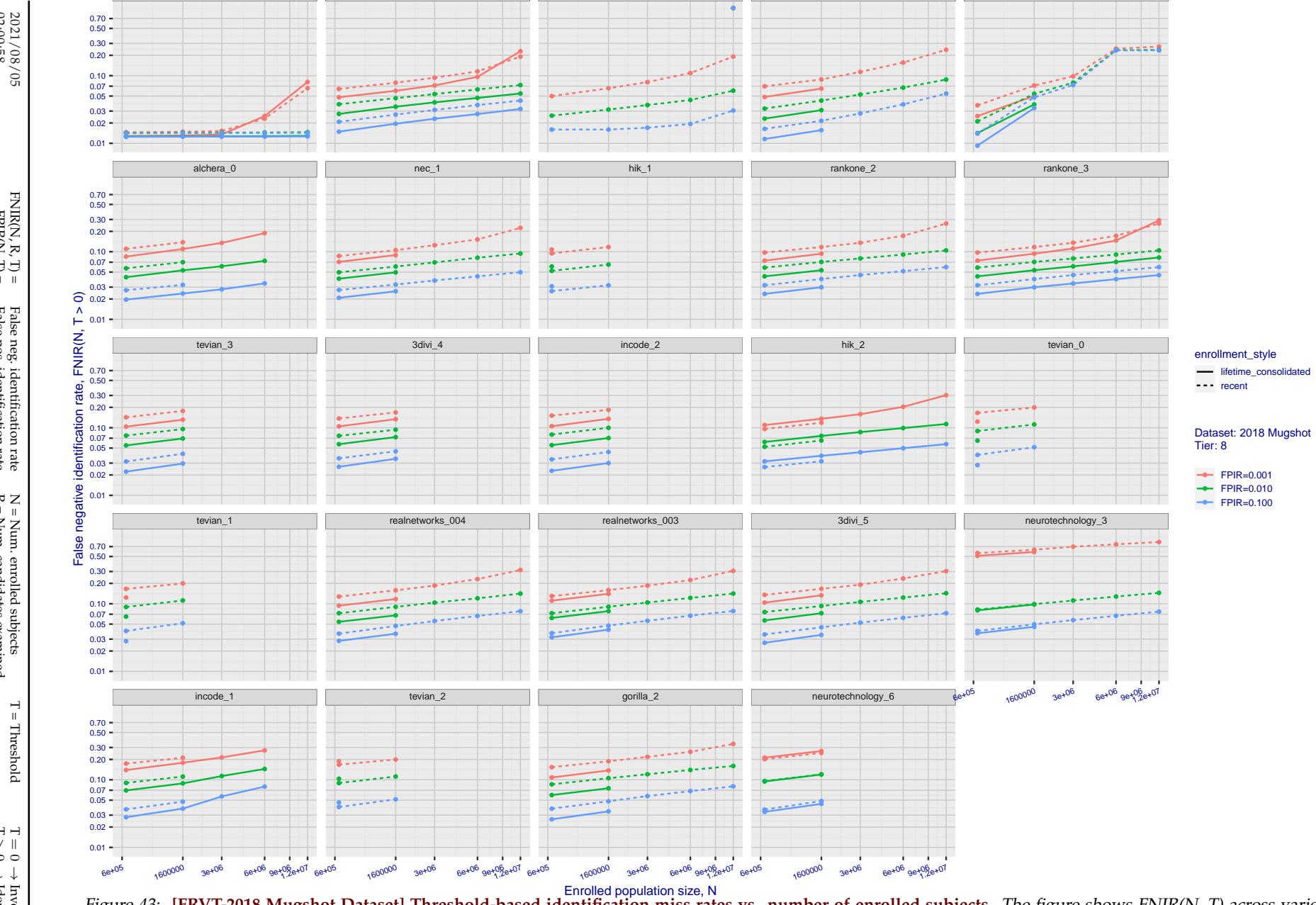


Figure 43: [FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects. The figure shows $\text{FNIR}(N, T)$ across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 1. Less accurate algorithms were not run on large N , so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by $\text{FNIR}(N_b, 1, 0)$, then sorting by median $\text{FNIR}(N_b, T)$, $N_b = 640\,000$.

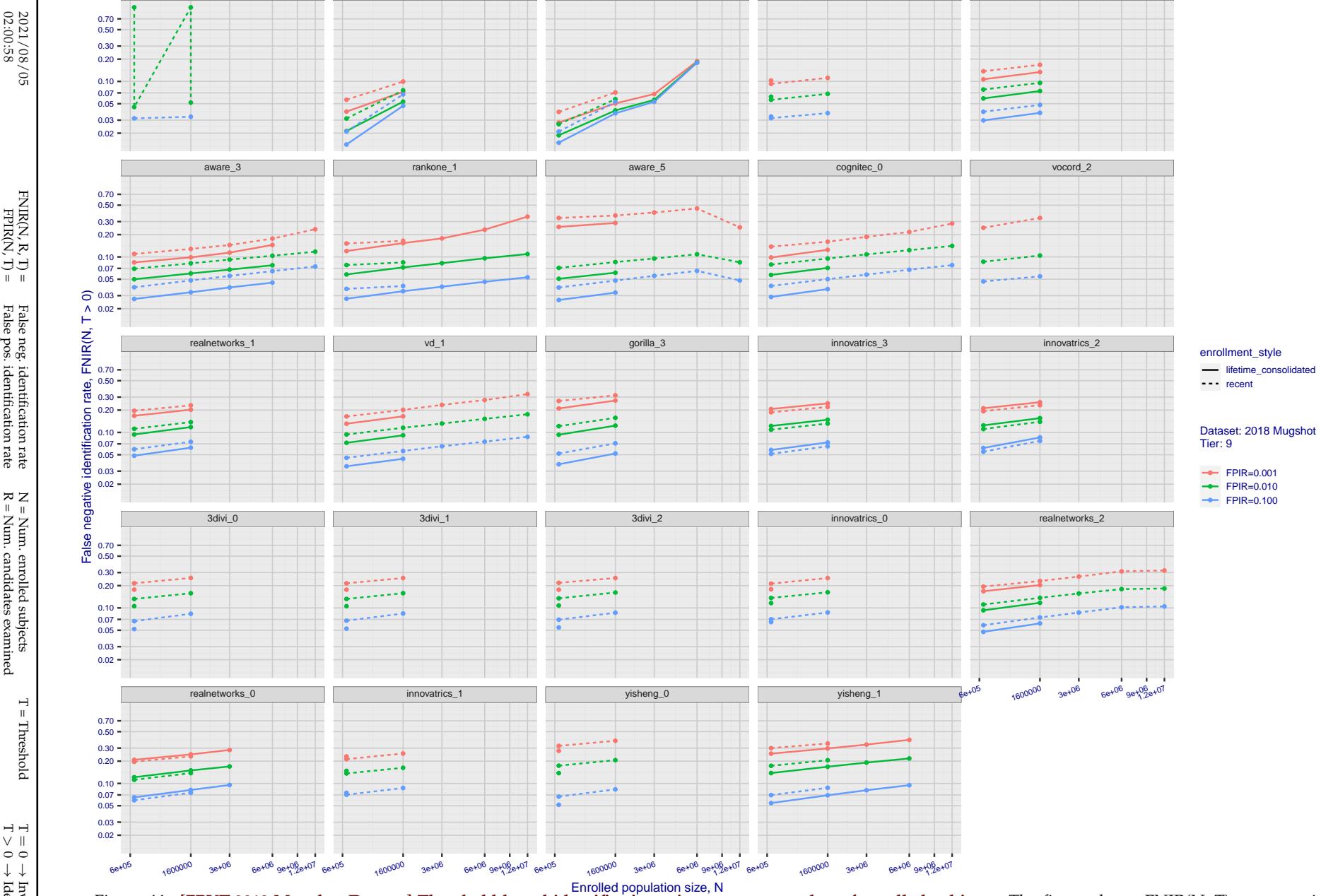


Figure 44: [FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects. The figure shows $\text{FNIR}(N, T)$ across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 1. Less accurate algorithms were not run on large N , so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by $\text{FNIR}(N_b, 1, 0)$, then sorting by median $\text{FNIR}(N_b, T)$, $N_b = 640\,000$.

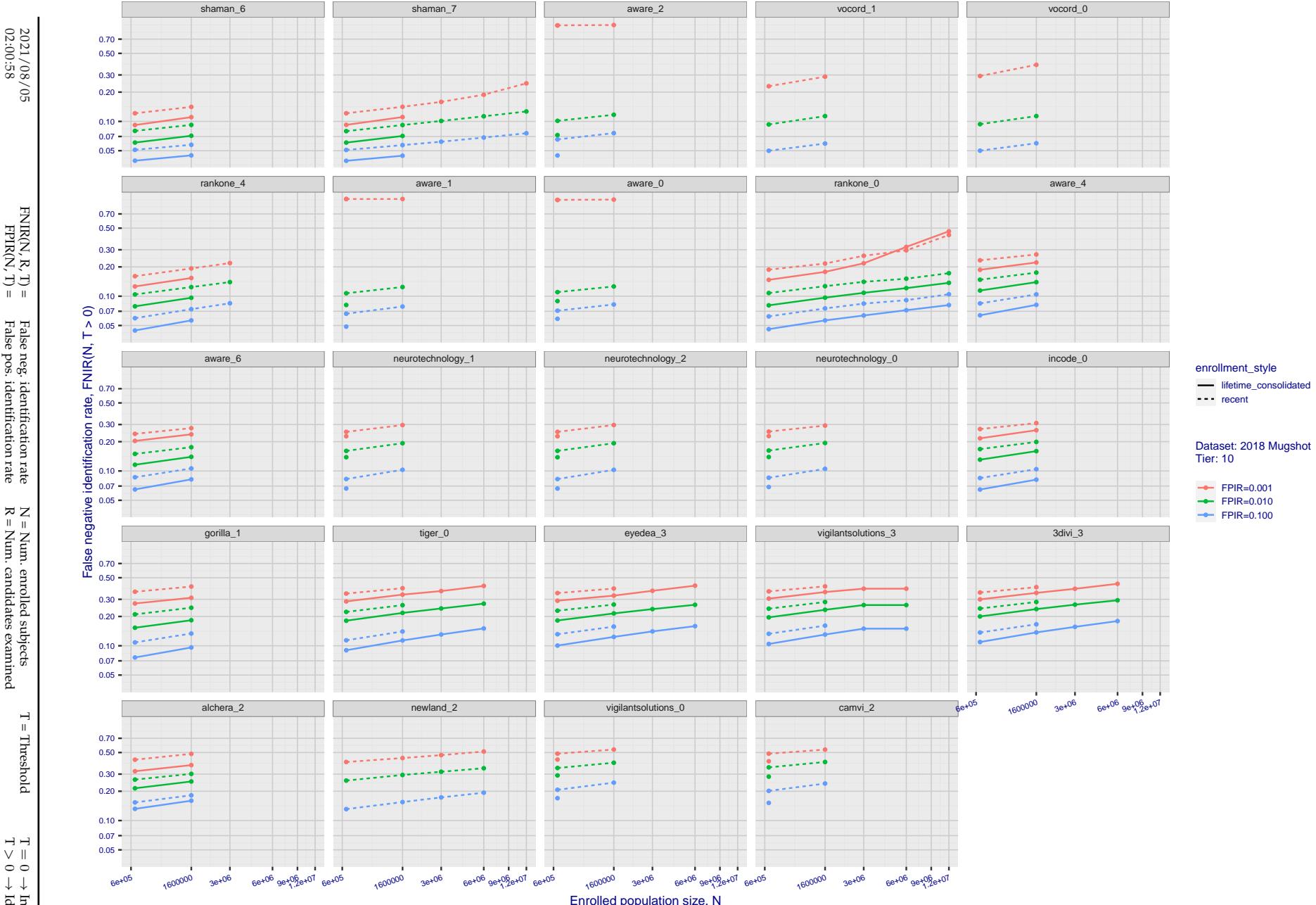


Figure 45: [FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects. The figure shows FNIR(N, T) across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 1. Less accurate algorithms were not run on large N , so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by FNIR($N_b, 1, 0$), then sorting by median FNIR(N_b, T), $N_b = 640\,000$.

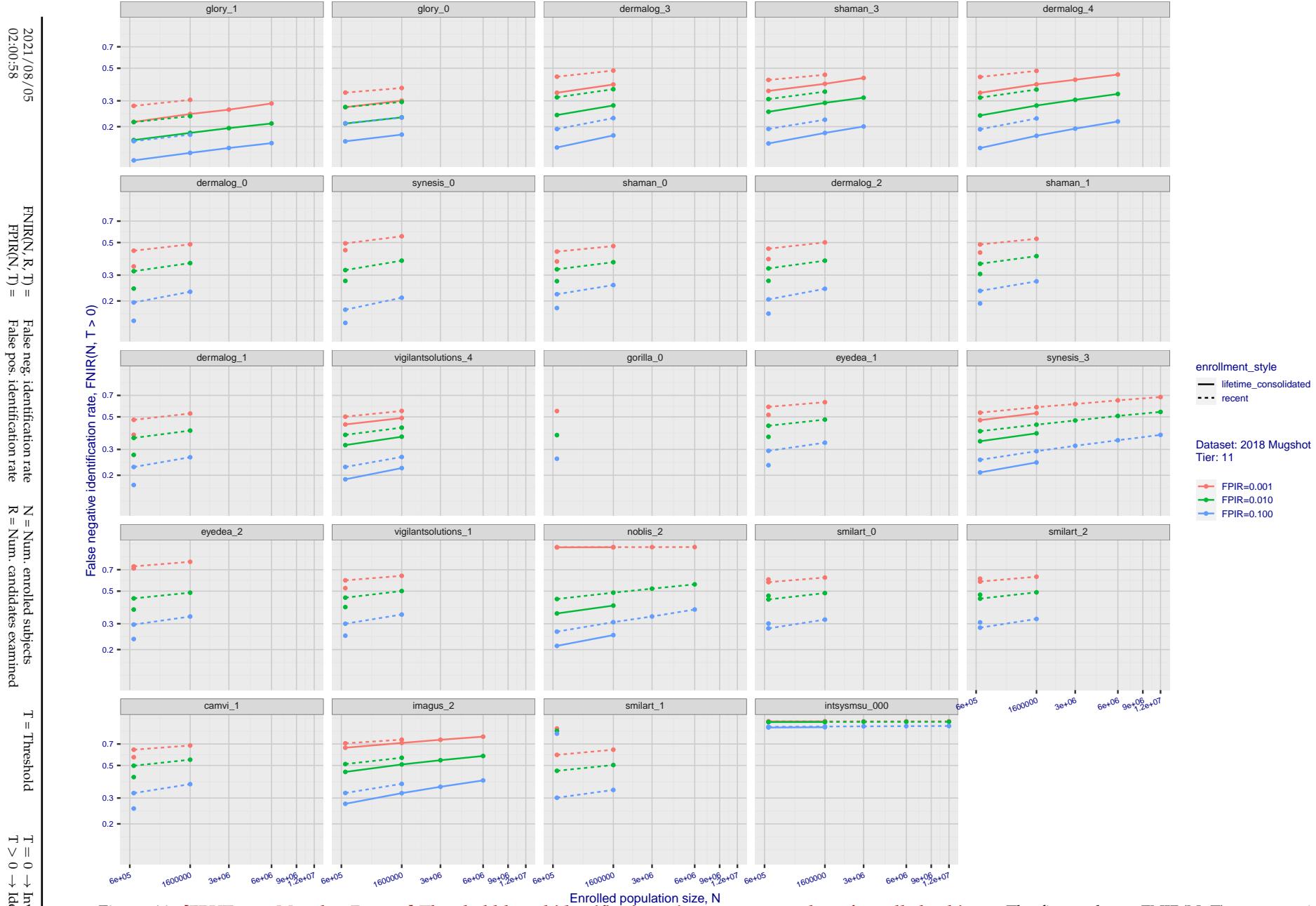


Figure 46: [FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects. The figure shows FNIR(N, T) across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 1. Less accurate algorithms were not run on large N , so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by $\text{FNIR}(N_b, 1, 0)$, then sorting by median $\text{FNIR}(N_b, T)$, $N_b = 640\,000$.

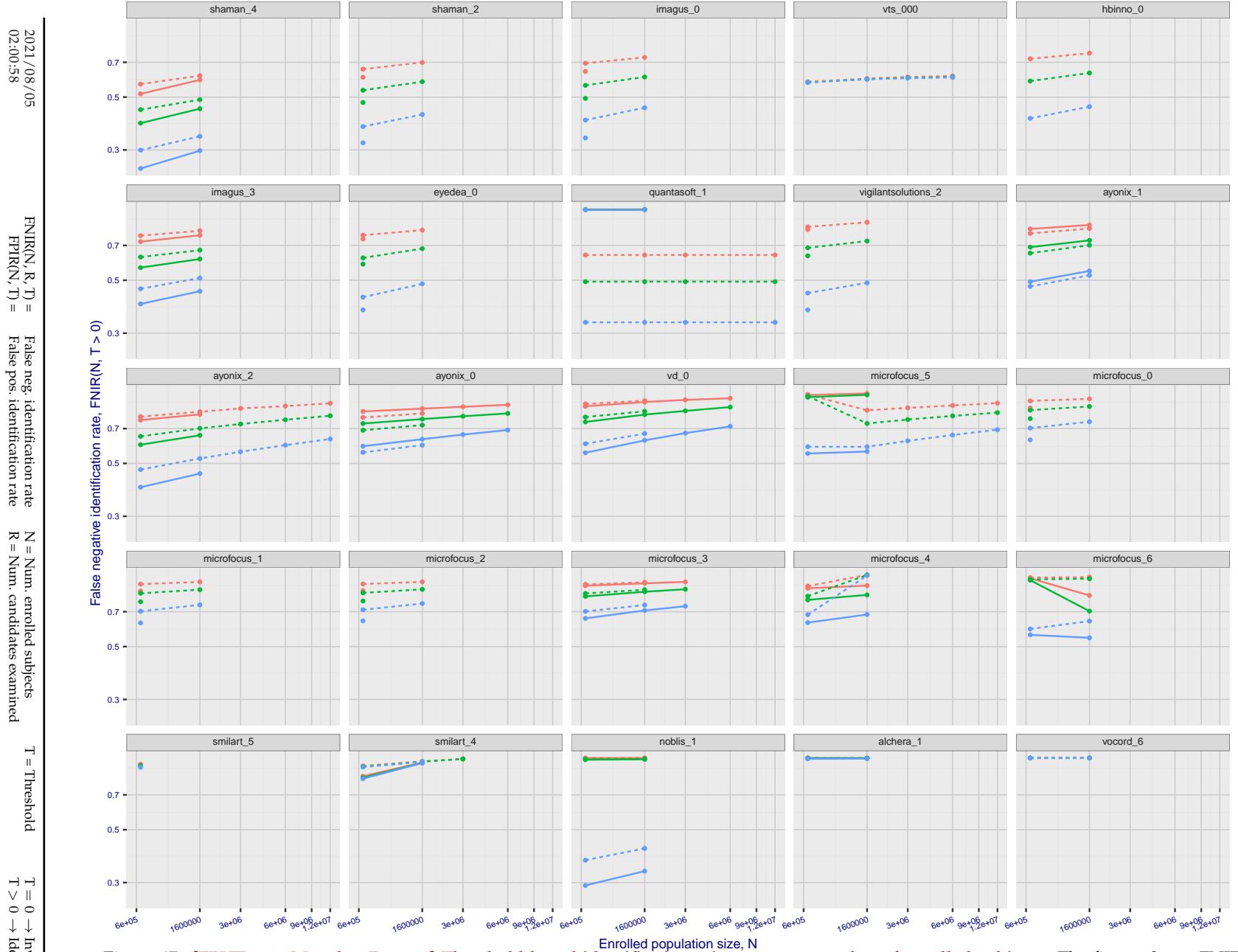


Figure 47: [FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects. The figure shows FNIR(N, T) across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 1. Less accurate algorithms were not run on large N , so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by $\text{FNIR}(N_b, 1, 0)$, then sorting by median $\text{FNIR}(N_b, T)$, $N_b = 640\,000$.

2021/08/05 02:00:58	$\text{FNIR}(N, R, T) =$ $\text{FPTR}(N, T) =$	False neg. identification rate False pos. identification rate	$N =$ Num. enrolled subjects $R =$ Num. candidates examined	$T =$ Threshold $T > 0 \rightarrow$ Identification	$T = 0 \rightarrow$ Investigation
------------------------	---	--	--	---	-----------------------------------

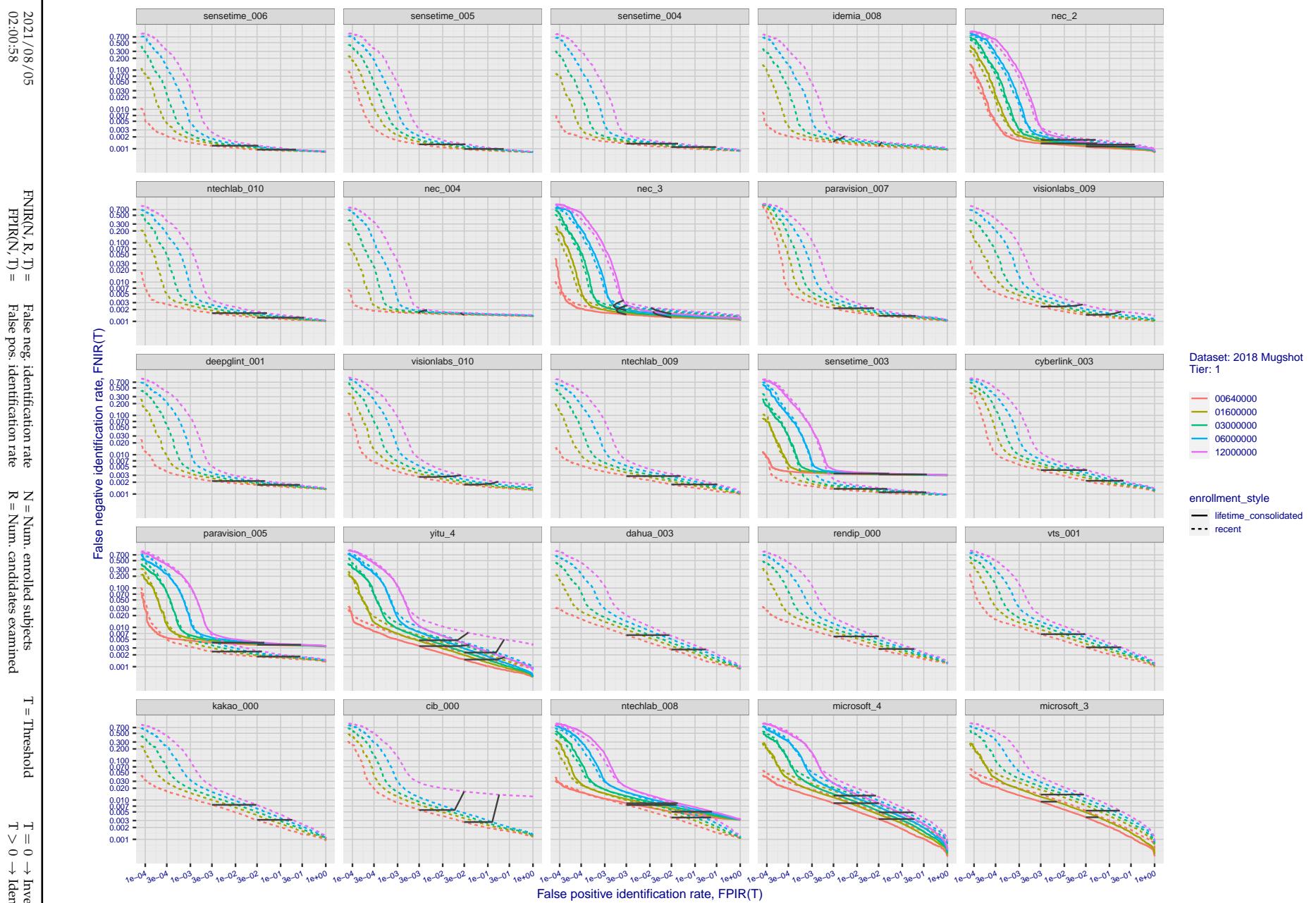


Figure 48: [FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates. The figure shows miss rates $\text{FNIR}(N, L, T)$ as a function of $\text{FPIR}(N, T)$, with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 1. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, $\text{FPIR}(T)$ rises with N , and mate scores are independent of N . Other algorithms adjust scores in an attempt to make FPIR independent of N .

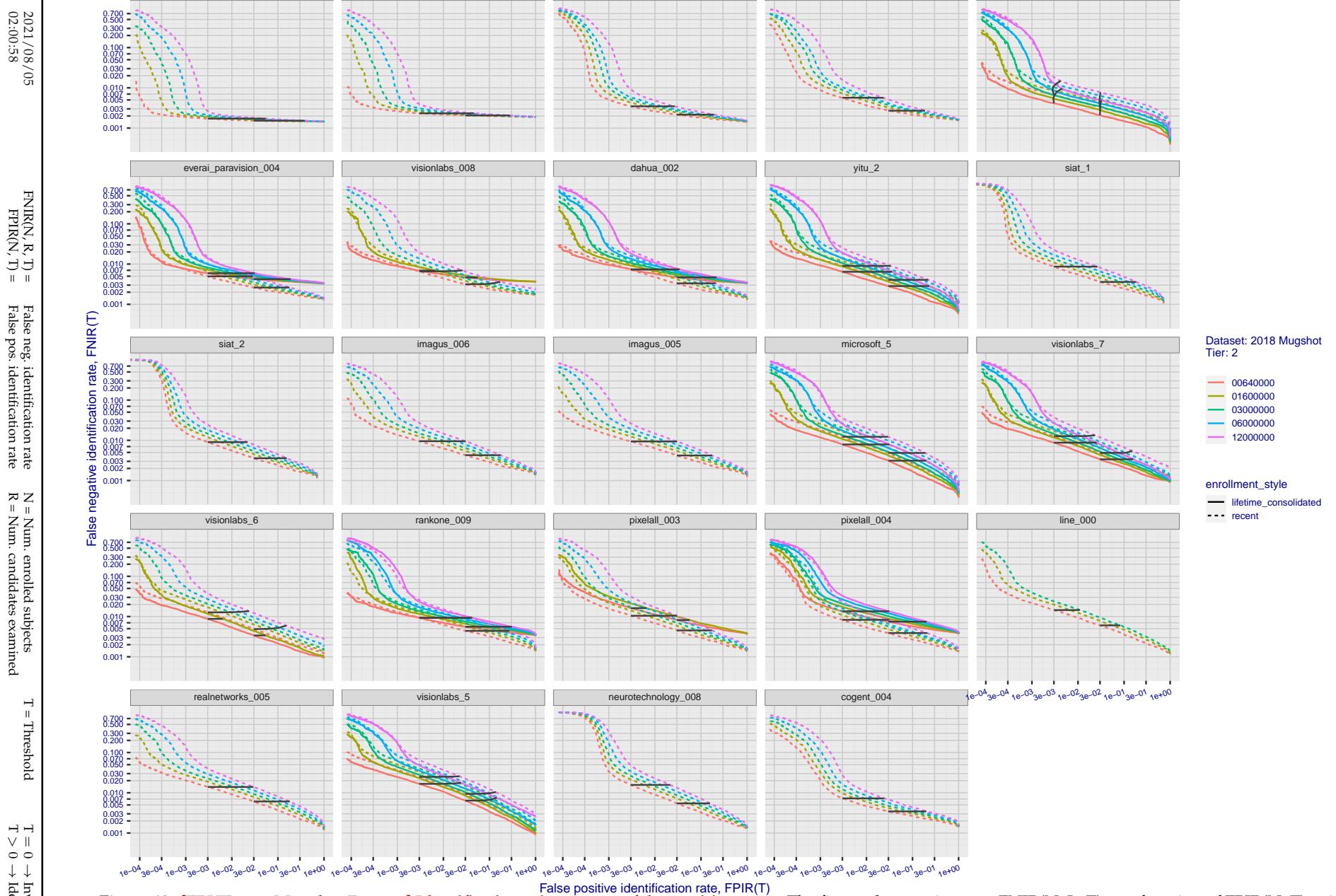


Figure 49: [FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates. The figure shows miss rates $\text{FNIR}(N, L, T)$ as a function of $\text{FPIR}(N, T)$, with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 1. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, $\text{FPIR}(T)$ rises with N , and mate scores are independent of N . Other algorithms adjust scores in an attempt to make FPIR independent of N .

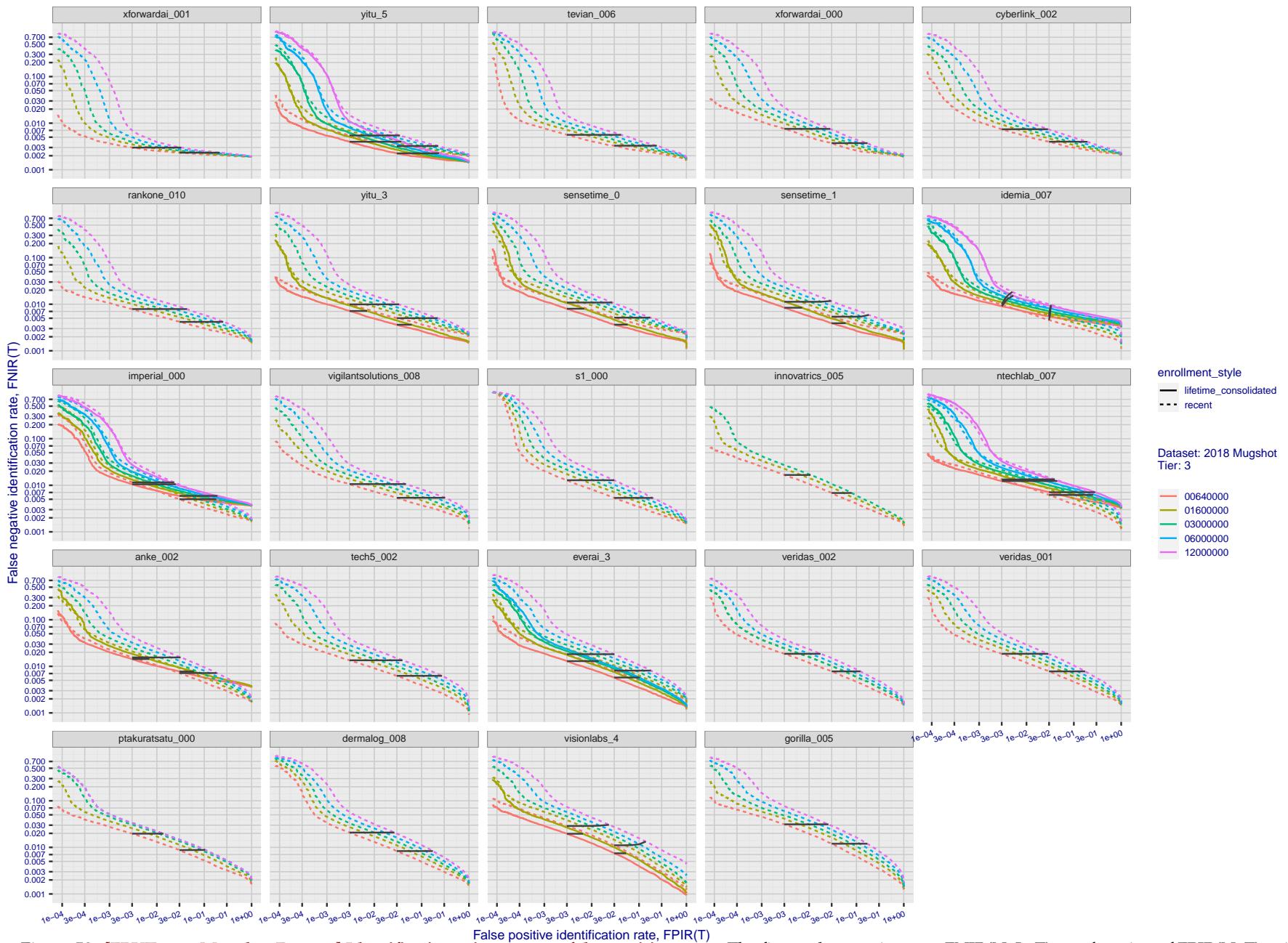


Figure 50: [FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates. The figure shows miss rates $\text{FNIR}(N, L, T)$ as a function of $\text{FPIR}(N, T)$, with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 1. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, $\text{FPIR}(T)$ rises with N , and mate scores are independent of N . Other algorithms adjust scores in an attempt to make FPIR independent of N .

2021 / 08 / 05
02:00:58FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
 N = Num. enrolled subjects
 R = Num. candidates examined T = Threshold $T = 0 \rightarrow$ Investigation
 $T > 0 \rightarrow$ Identification

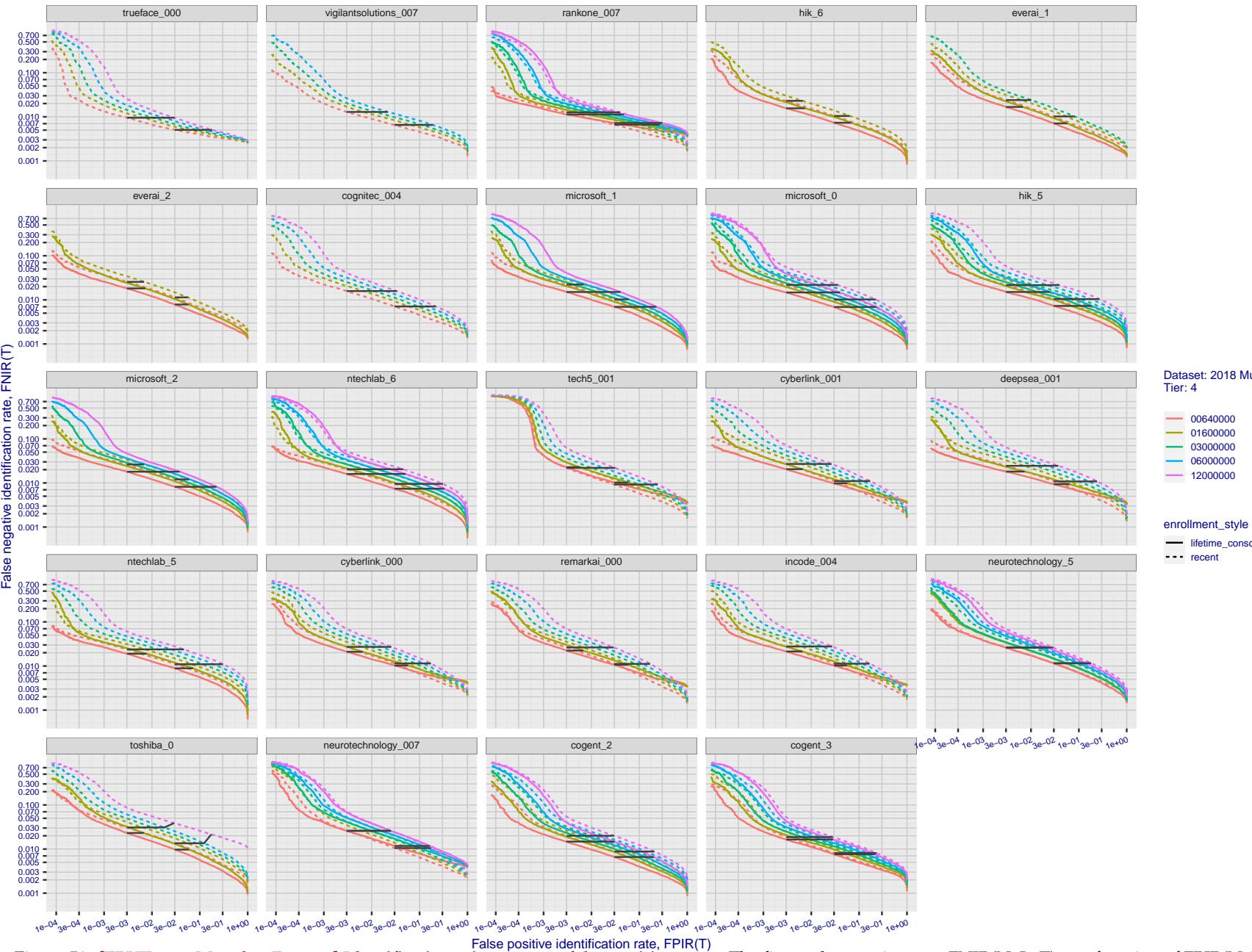


Figure 51: [FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates. The figure shows miss rates $\text{FNIR}(N, L, T)$ as a function of $\text{FPIR}(N, T)$, with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 1. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, $\text{FPIR}(T)$ rises with N , and mate scores are independent of N . Other algorithms adjust scores in an attempt to make FPIR independent of N .

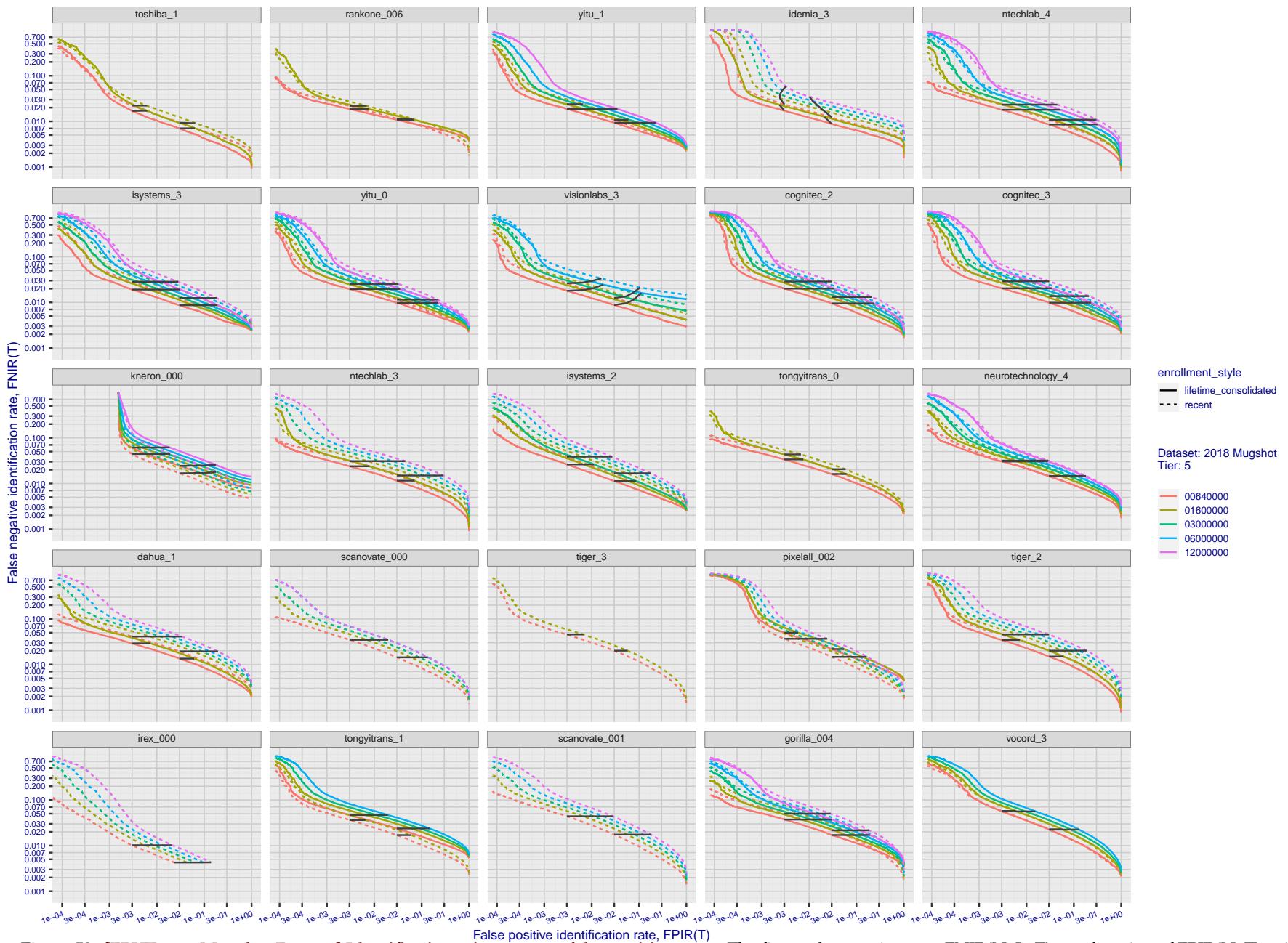


Figure 52: [FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates. The figure shows miss rates $\text{FNIR}(N, L, T)$ as a function of $\text{FPIR}(N, T)$, with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 1. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, $\text{FPIR}(T)$ rises with N , and mate scores are independent of N . Other algorithms adjust scores in an attempt to make FPIR independent of N .

2021/08/05
02:00:58

 $\text{FNIR}(N, R, T) = \text{False neg. identification rate}$
 $\text{FPIR}(N, T) = \text{False pos. identification rate}$
 $N = \text{Num. enrolled subjects}$
 $R = \text{Num. candidates examined}$
 $T = \text{Threshold}$
 $T = 0 \rightarrow \text{Investigation}$
 $T > 0 \rightarrow \text{Identification}$

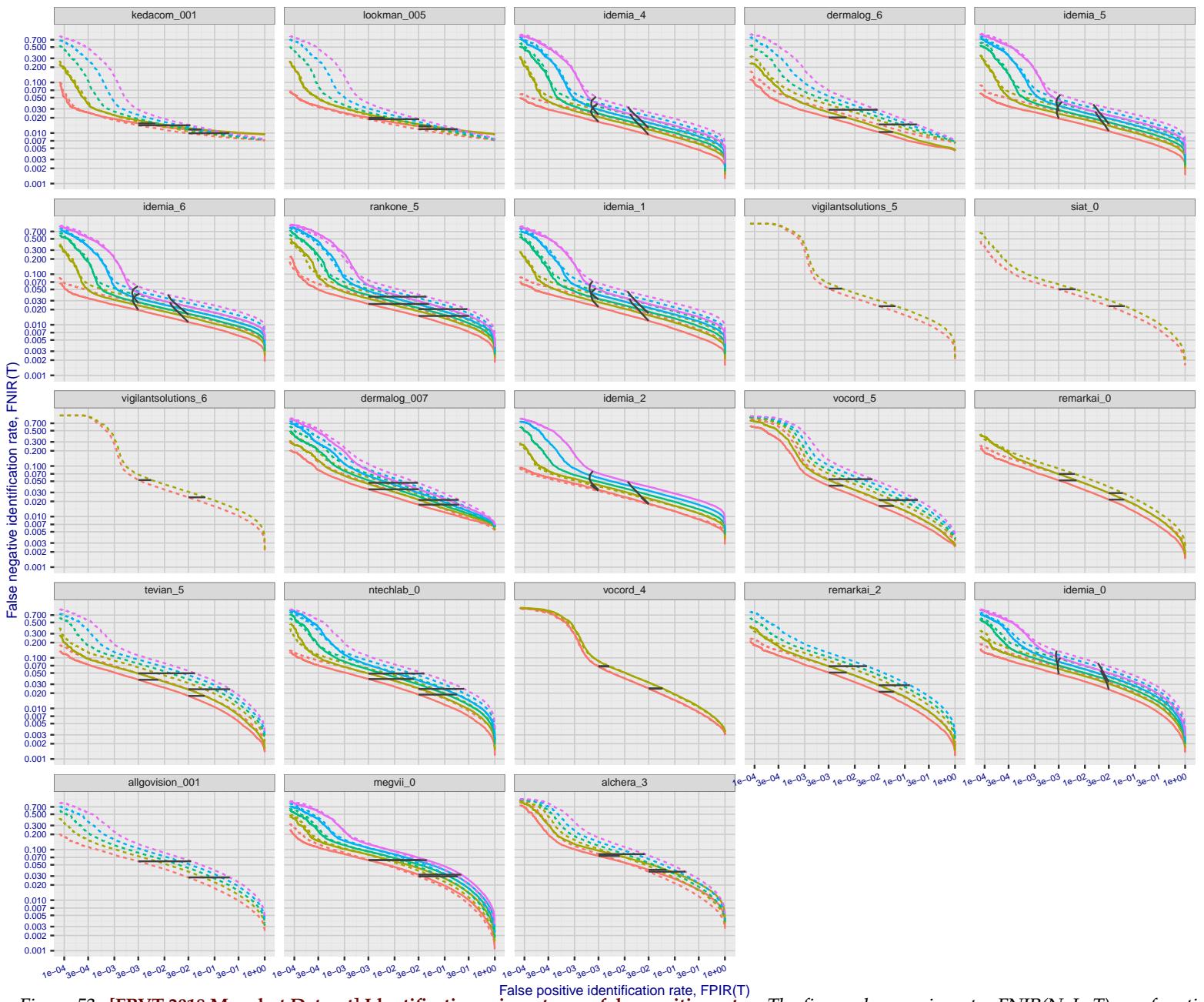


Figure 53: [FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates. The figure shows miss rates $\text{FNIR}(N, L, T)$ as a function of $\text{FPIR}(N, T)$, with N ranging from 64 000 to 12 000 000 as noted in rows 1-10 of Table 1. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, $\text{FPIR}(T)$ rises with N , and mate scores are independent of N . Other algorithms adjust scores in an attempt to make FPIR independent of N .

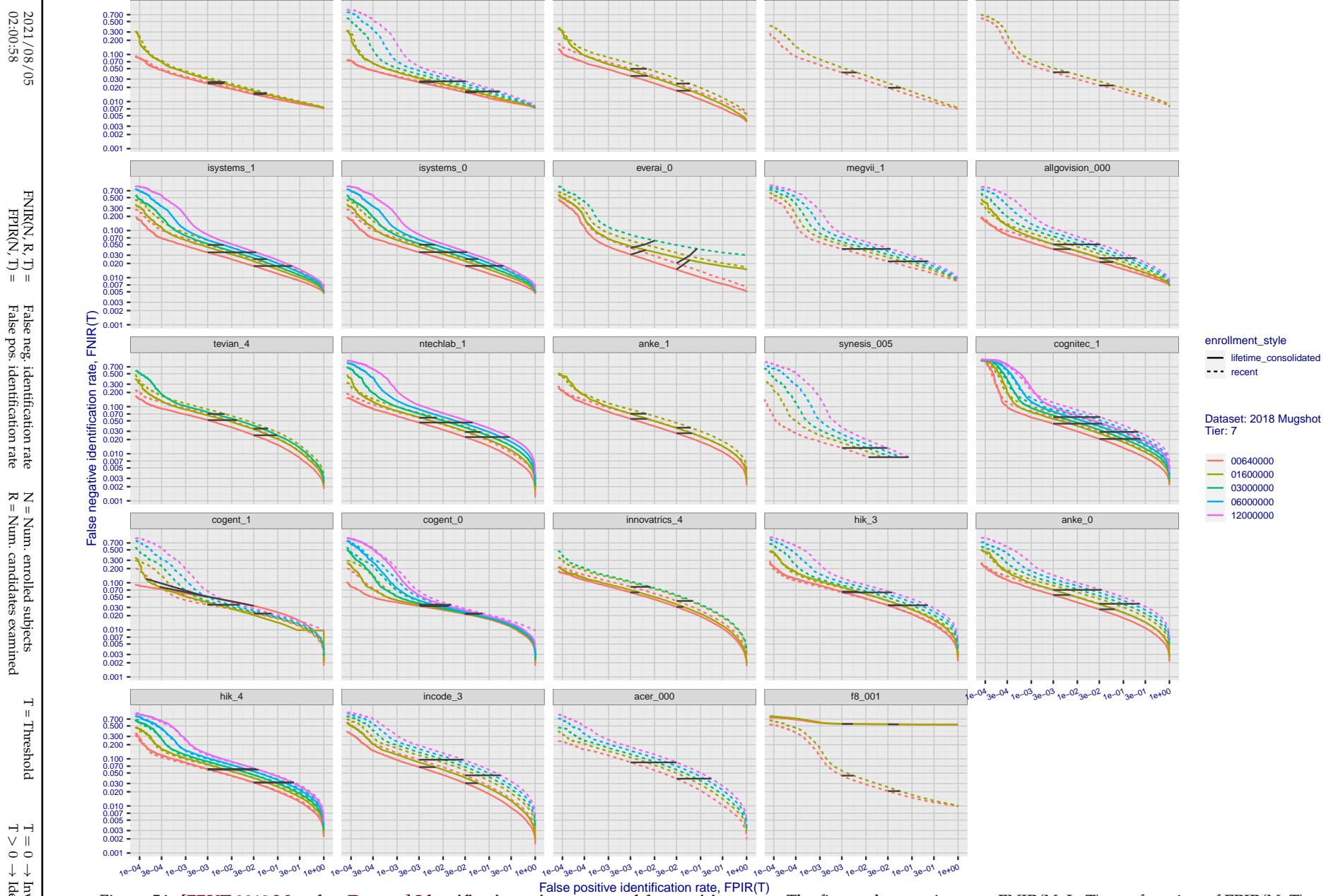


Figure 54: [FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates. The figure shows miss rates $\text{FNIR}(N, L, T)$ as a function of $\text{FPIR}(N, T)$, with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 1. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, $\text{FPIR}(T)$ rises with N , and mate scores are independent of N . Other algorithms adjust scores in an attempt to make FPIR independent of N .

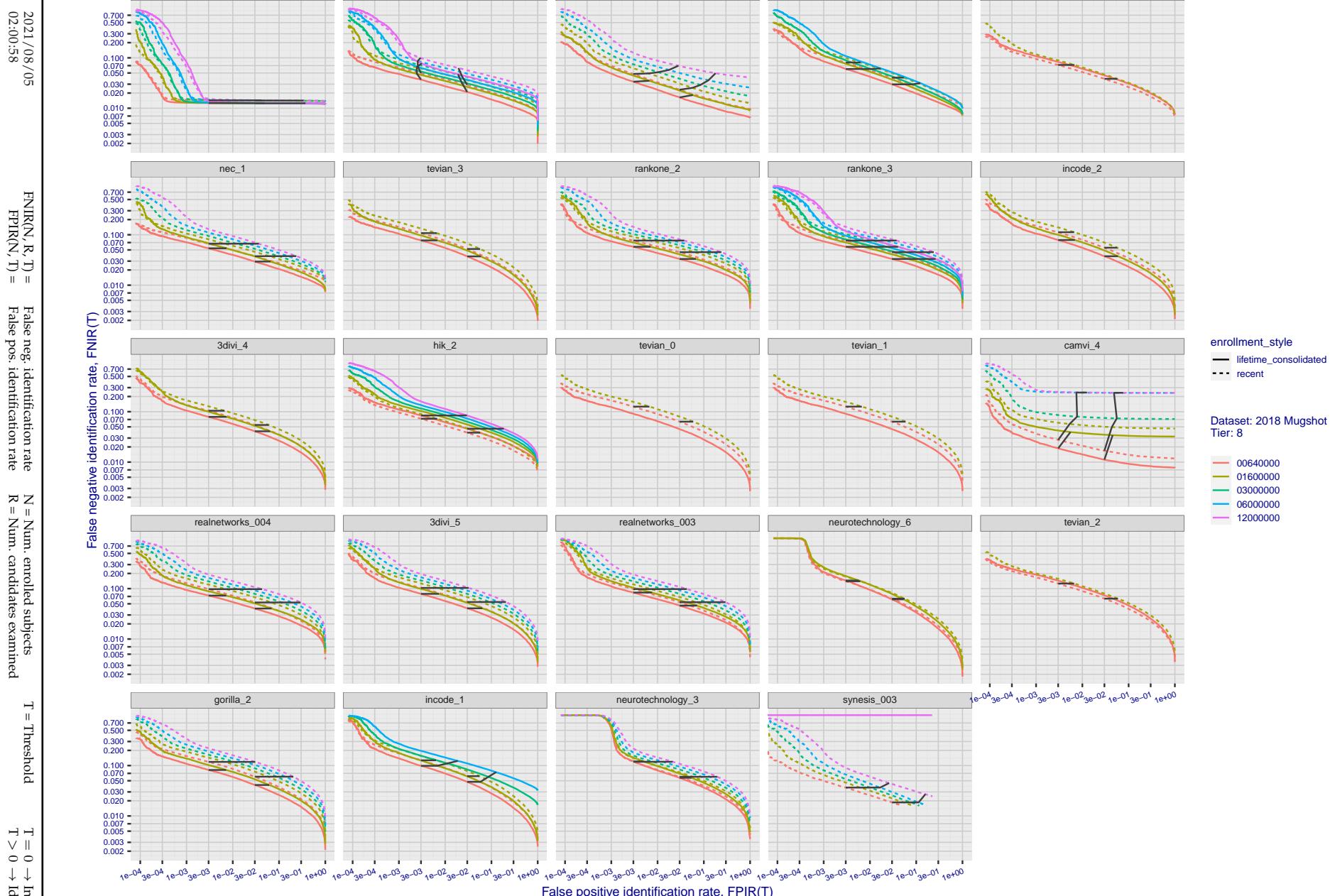


Figure 55: [FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates. The figure shows miss rates $\text{FNIR}(N, L, T)$ as a function of $\text{FPIR}(N, T)$, with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 1. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, $\text{FPIR}(T)$ rises with N , and mate scores are independent of N . Other algorithms adjust scores in an attempt to make FPIR independent of N .

2021/08/05
02:00:58

 $\text{FNIR}(N, R, T) = \text{False neg. identification rate}$
 $\text{FPIR}(N, T) = \text{False pos. identification rate}$
 $N = \text{Num. enrolled subjects}$
 $R = \text{Num. candidates examined}$
 $T = \text{Threshold}$
 $T = 0 \rightarrow \text{Investigation}$
 $T > 0 \rightarrow \text{Identification}$

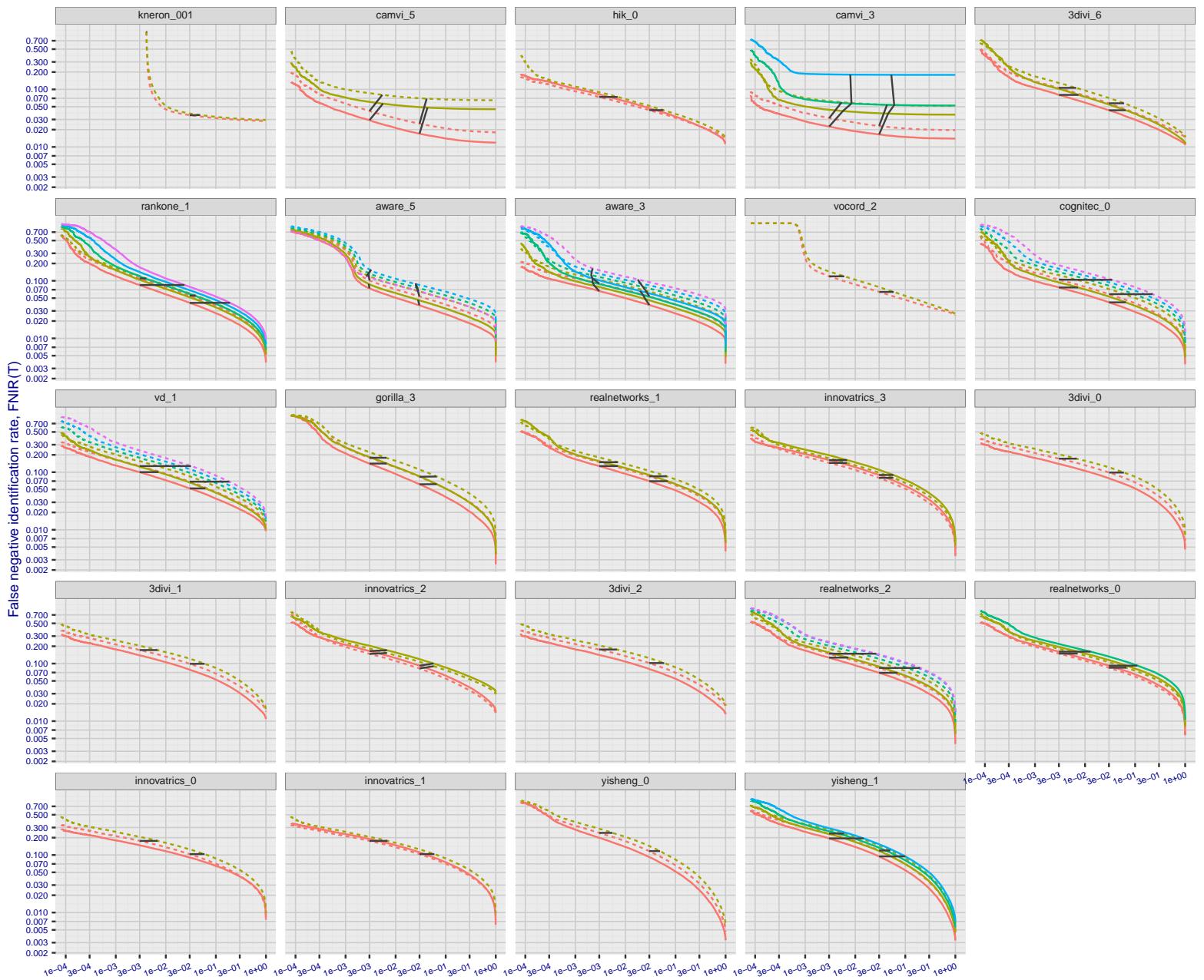


Figure 56: [FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates. The figure shows miss rates $\text{FNIR}(N, L, T)$ as a function of $\text{FPIR}(N, T)$, with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 1. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, $\text{FPIR}(T)$ rises with N , and mate scores are independent of N . Other algorithms adjust scores in an attempt to make FPIR independent of N .

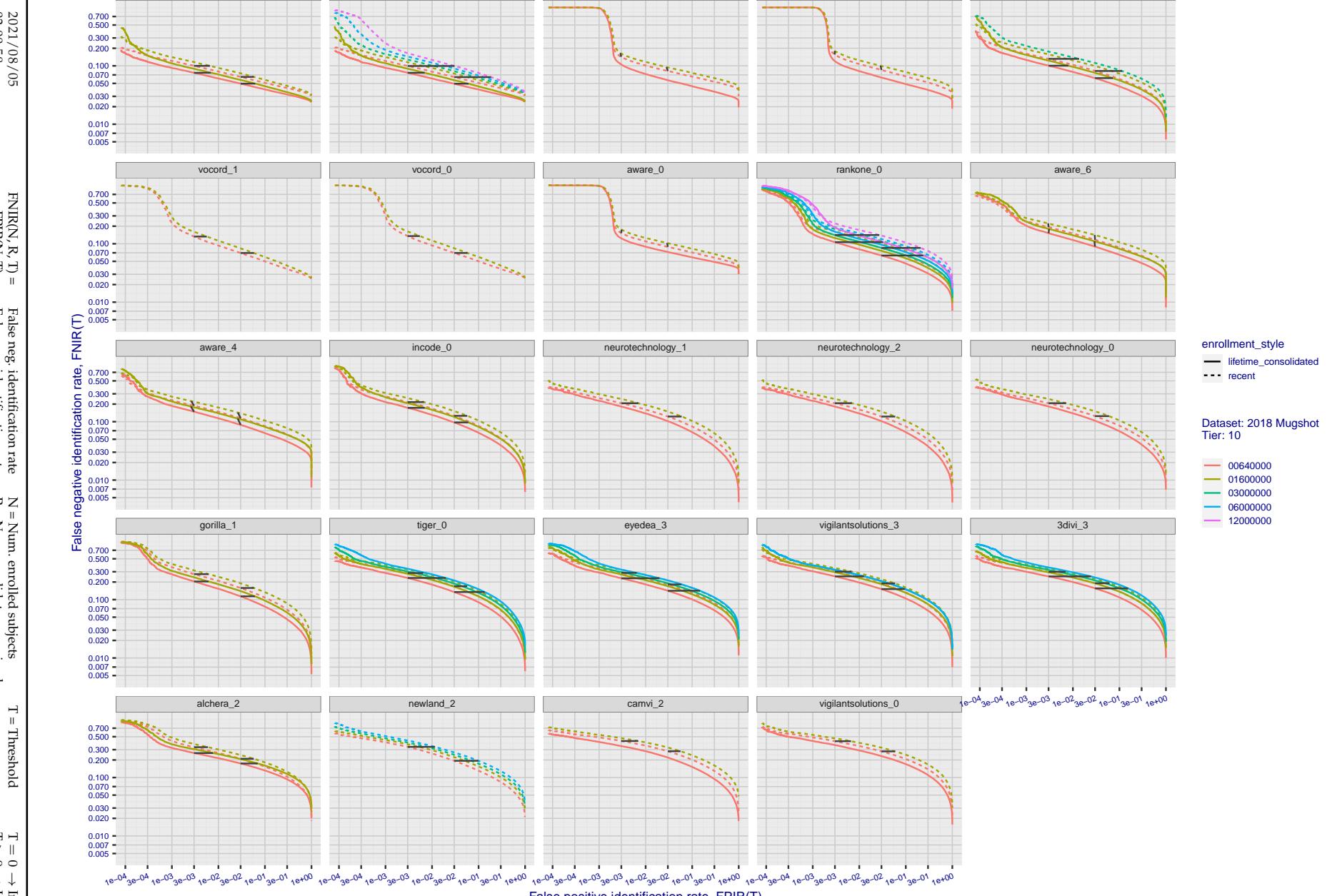


Figure 57: [FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates. The figure shows miss rates $\text{FNIR}(N, L, T)$ as a function of $\text{FPIR}(N, T)$, with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 1. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, $\text{FPIR}(T)$ rises with N , and mate scores are independent of N . Other algorithms adjust scores in an attempt to make FPIR independent of N .

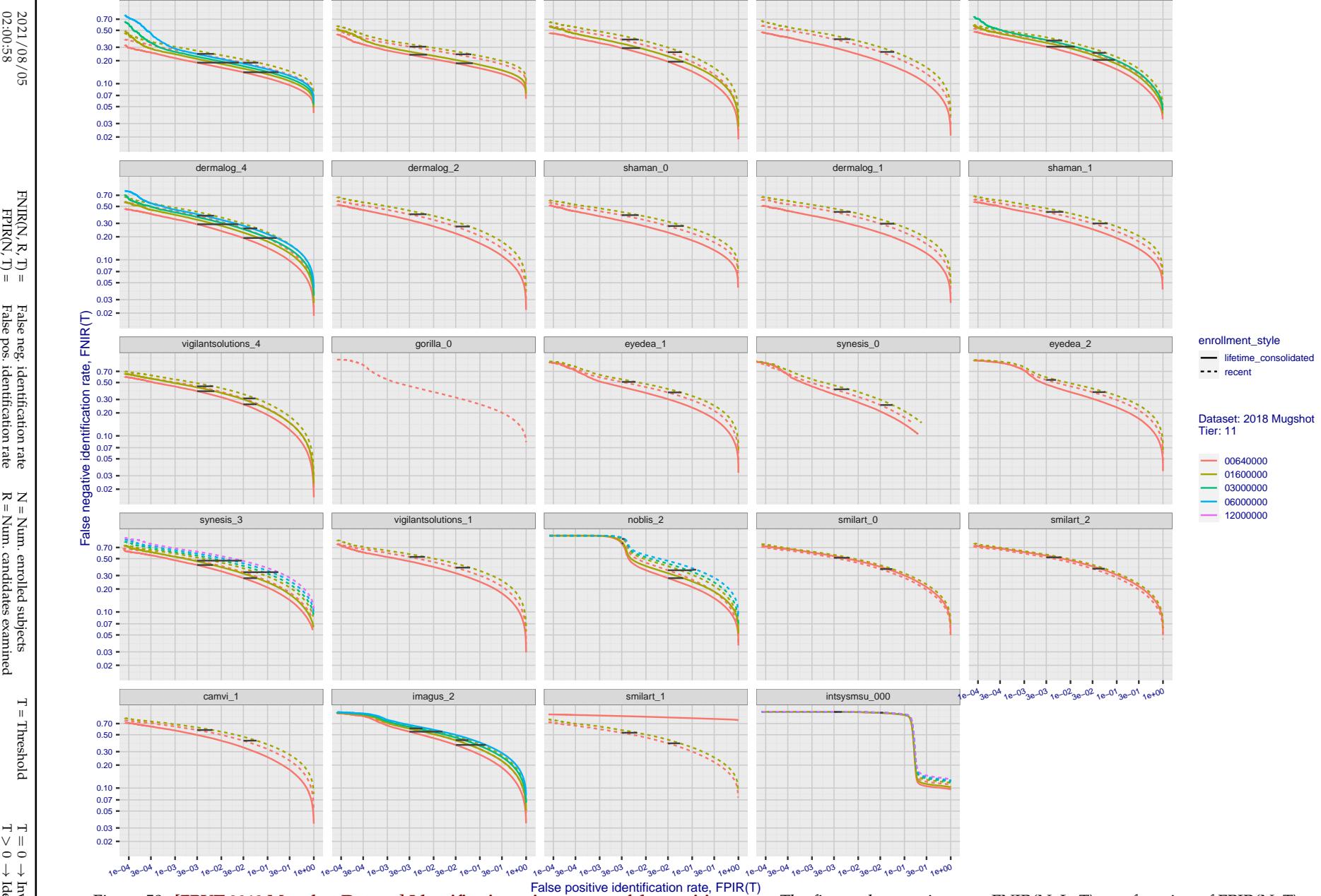


Figure 58: [FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates. The figure shows miss rates $\text{FNIR}(N, L, T)$ as a function of $\text{FPIR}(N, T)$, with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 1. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, $\text{FPIR}(T)$ rises with N , and mate scores are independent of N . Other algorithms adjust scores in an attempt to make FPIR independent of N .

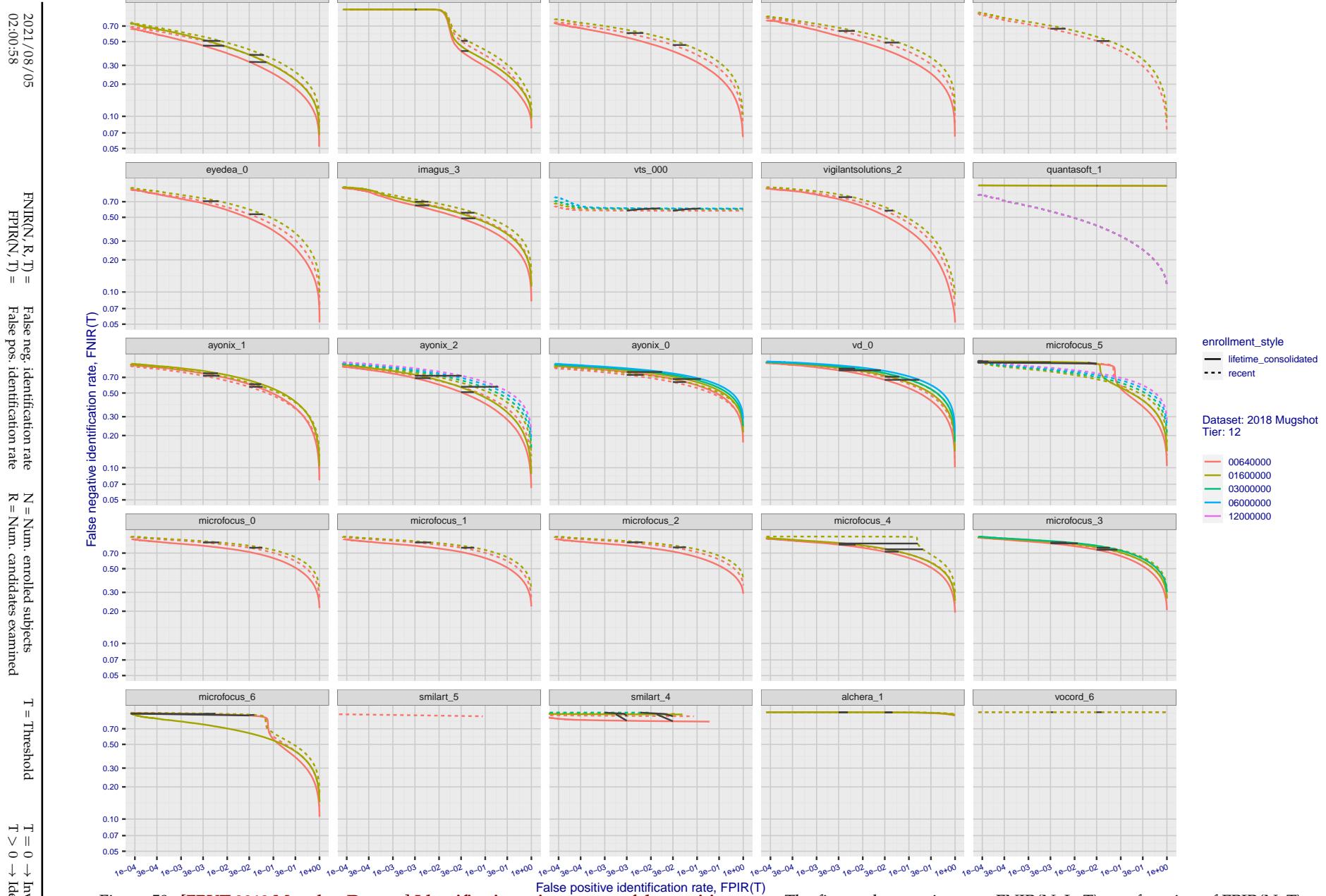


Figure 59: [FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates. The figure shows miss rates $\text{FNIR}(N, L, T)$ as a function of $\text{FPIR}(N, T)$, with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 1. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, $\text{FPIR}(T)$ rises with N , and mate scores are independent of N . Other algorithms adjust scores in an attempt to make FPIR independent of N .

Appendix B Effect of time-lapse: Accuracy after face ageing

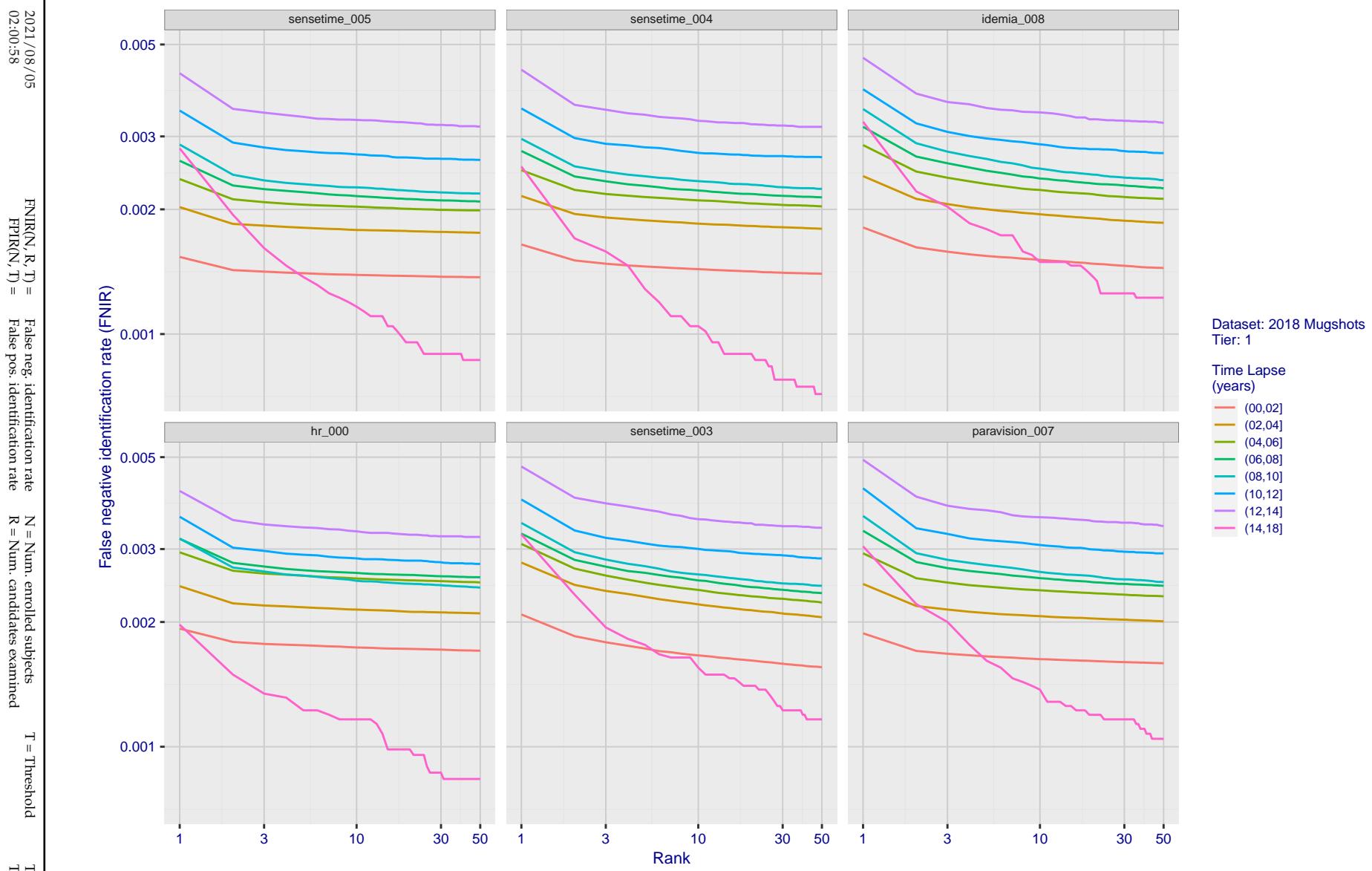


Figure 60: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

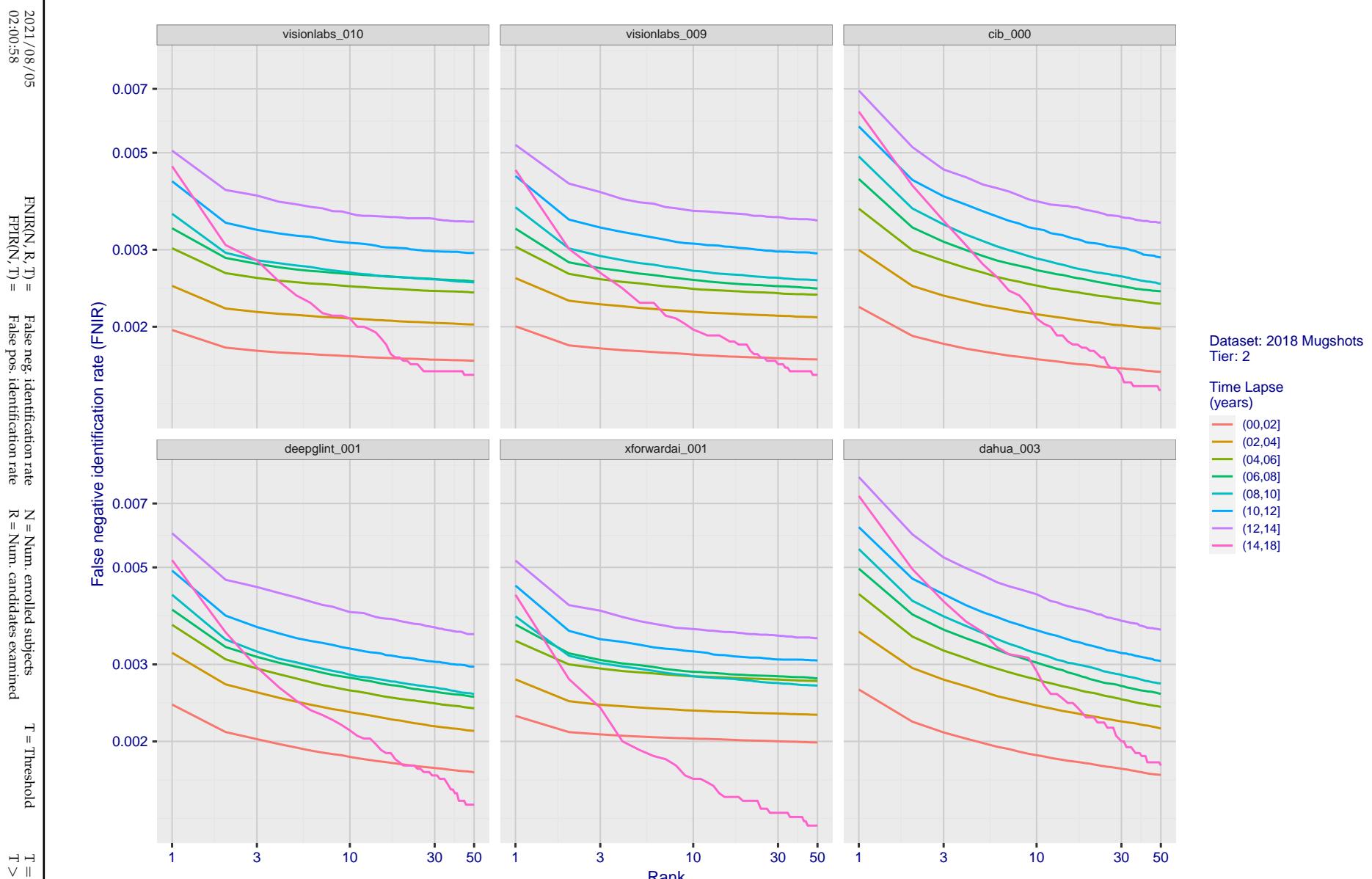


Figure 61: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

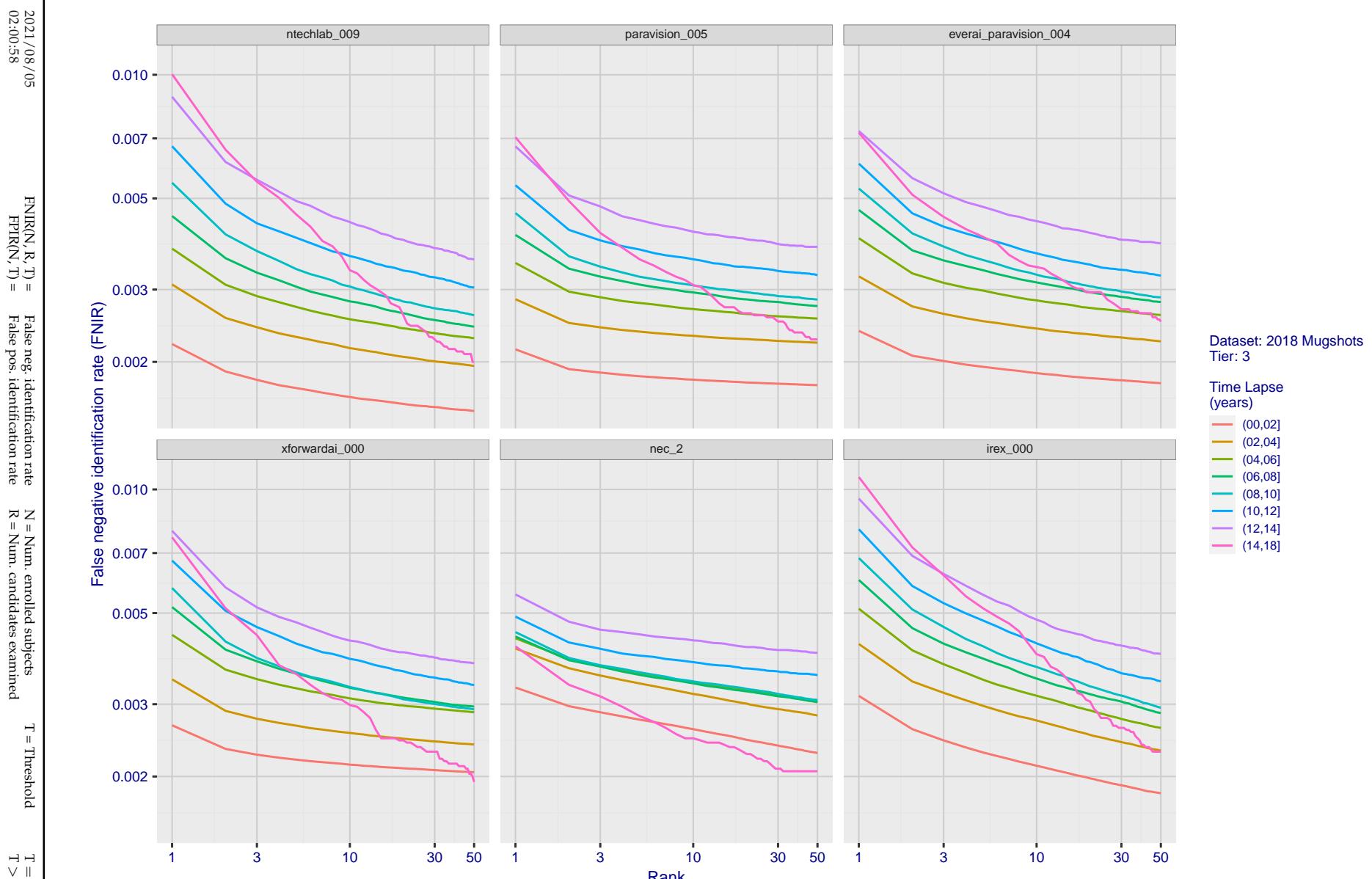


Figure 62: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

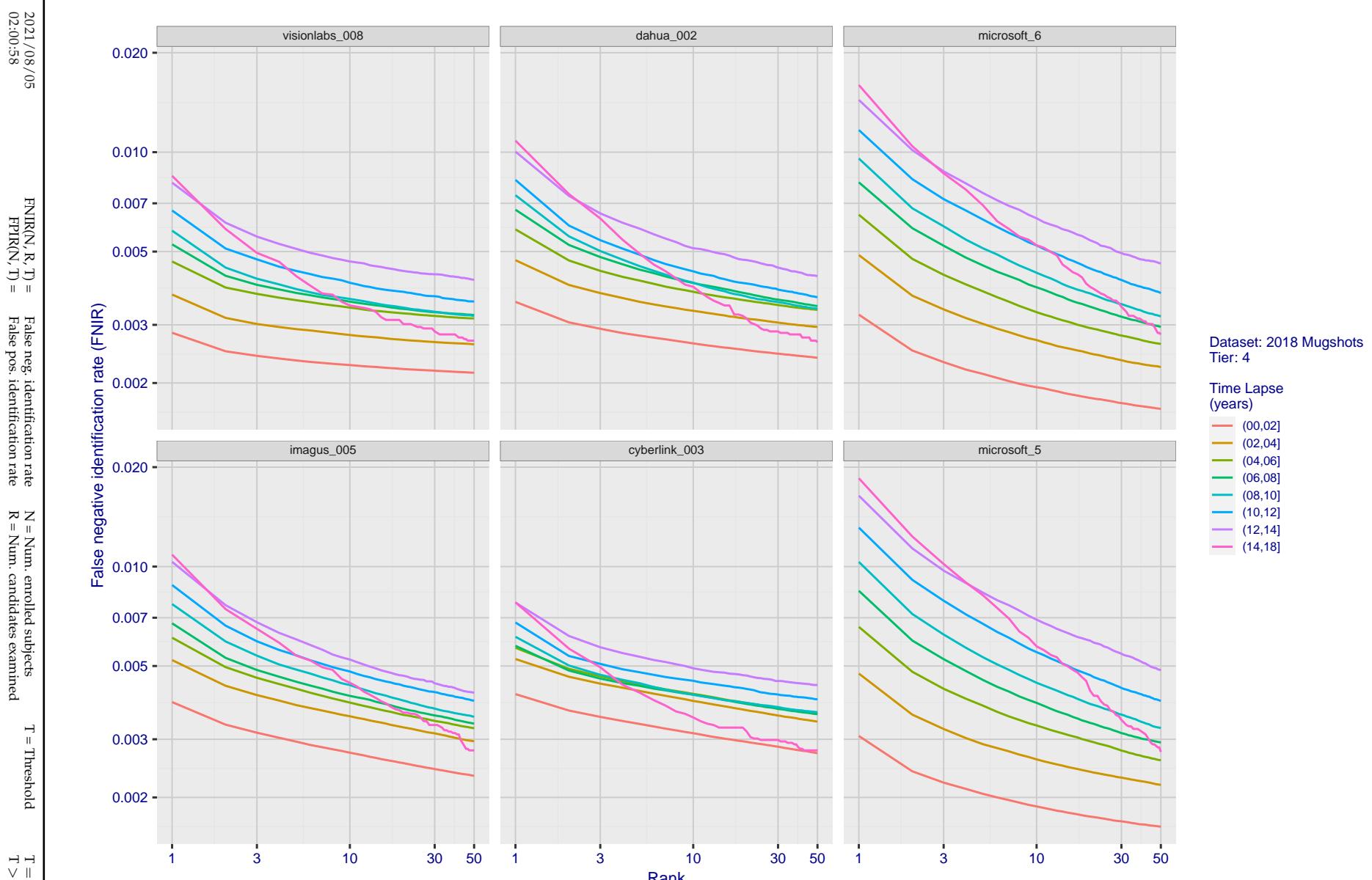


Figure 63: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

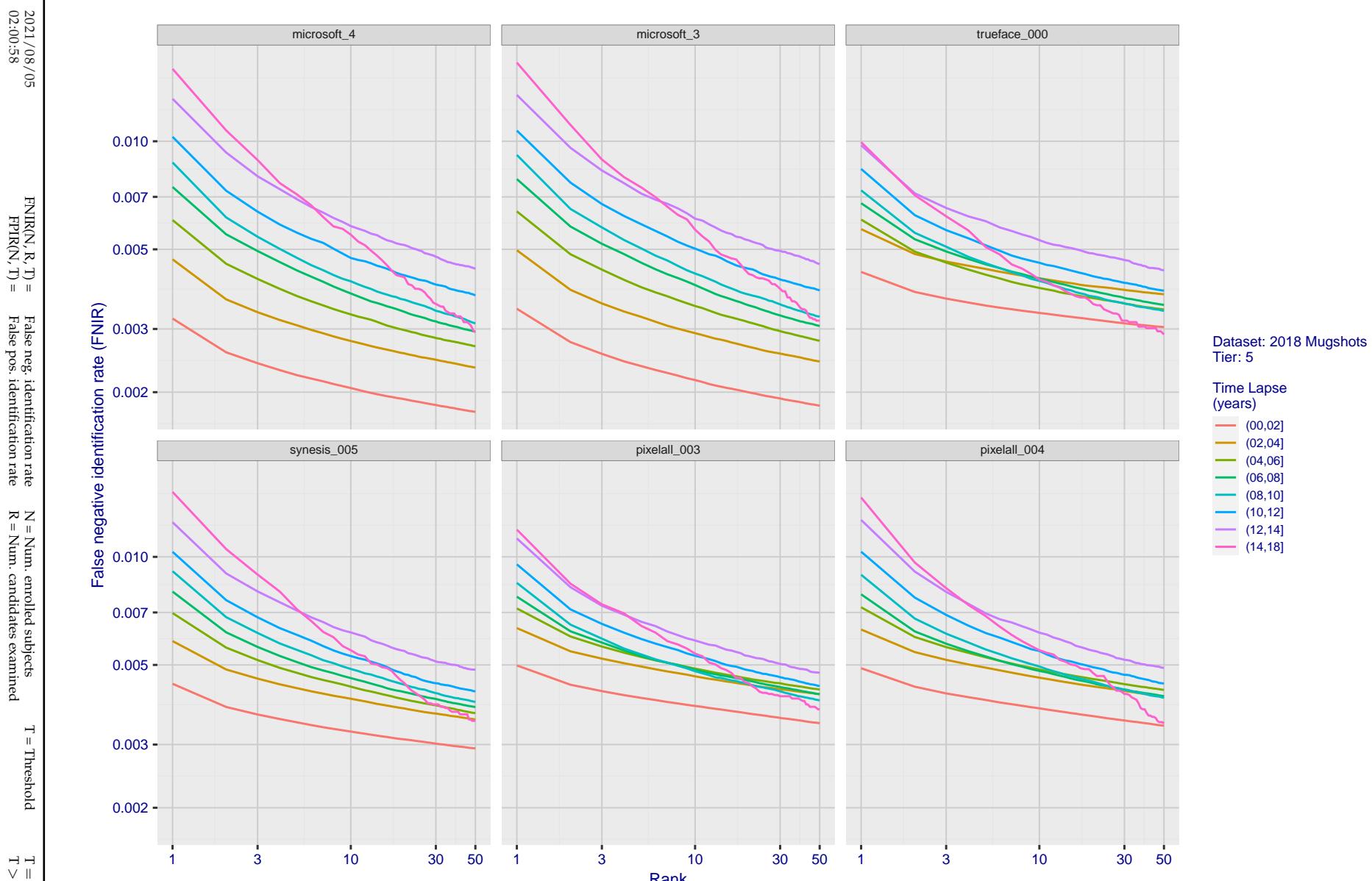


Figure 64: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

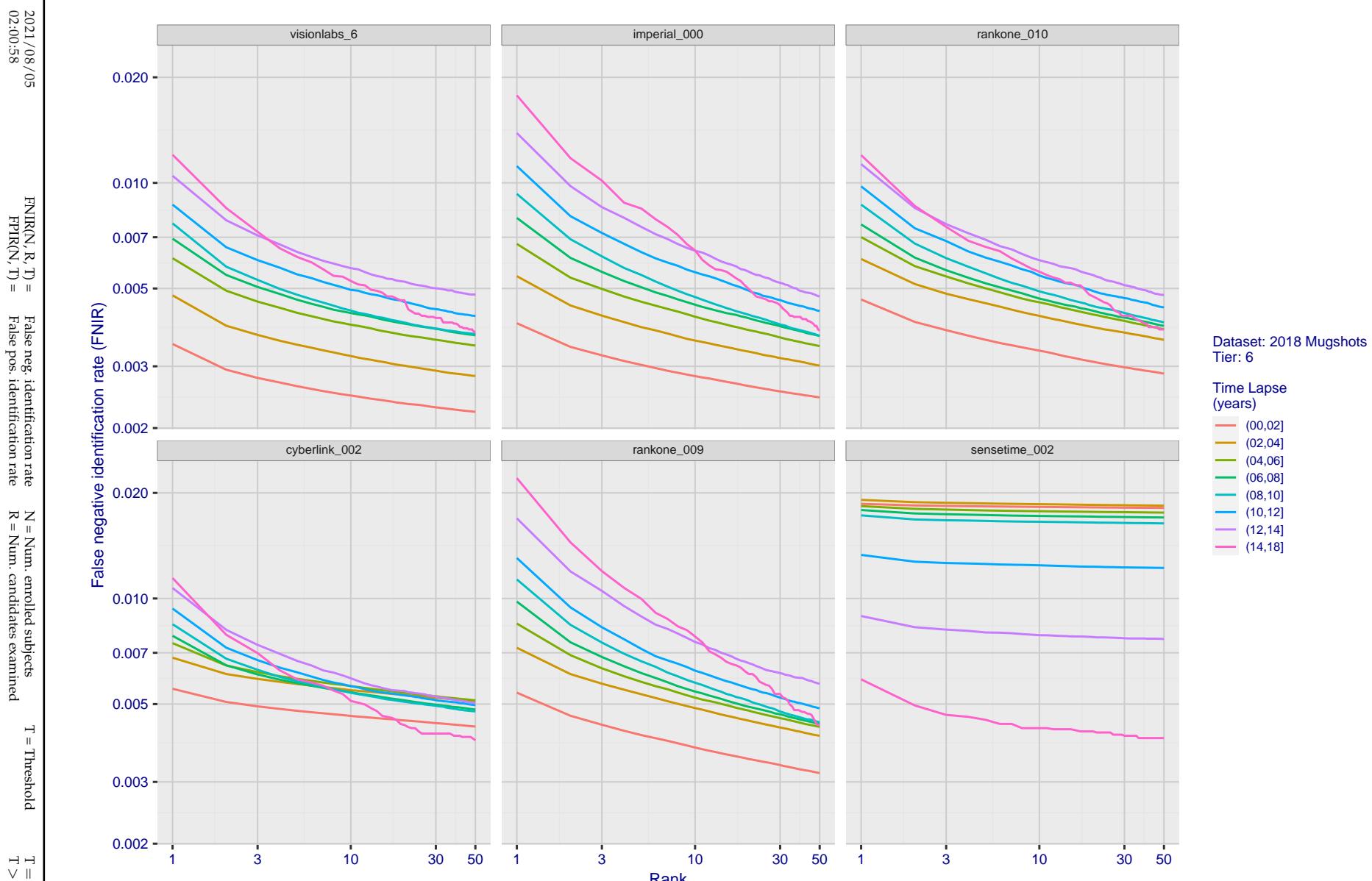


Figure 65: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

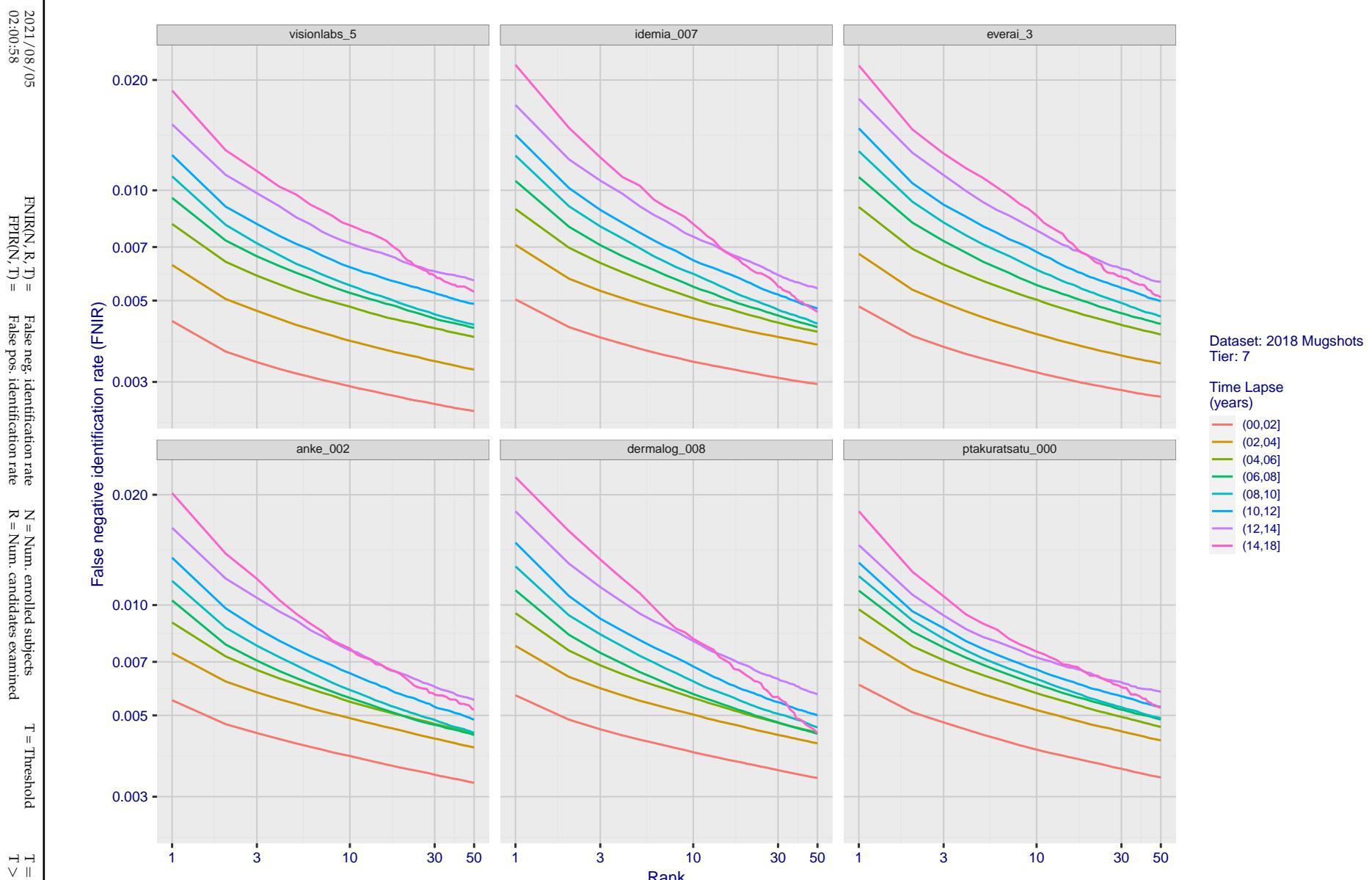


Figure 66: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

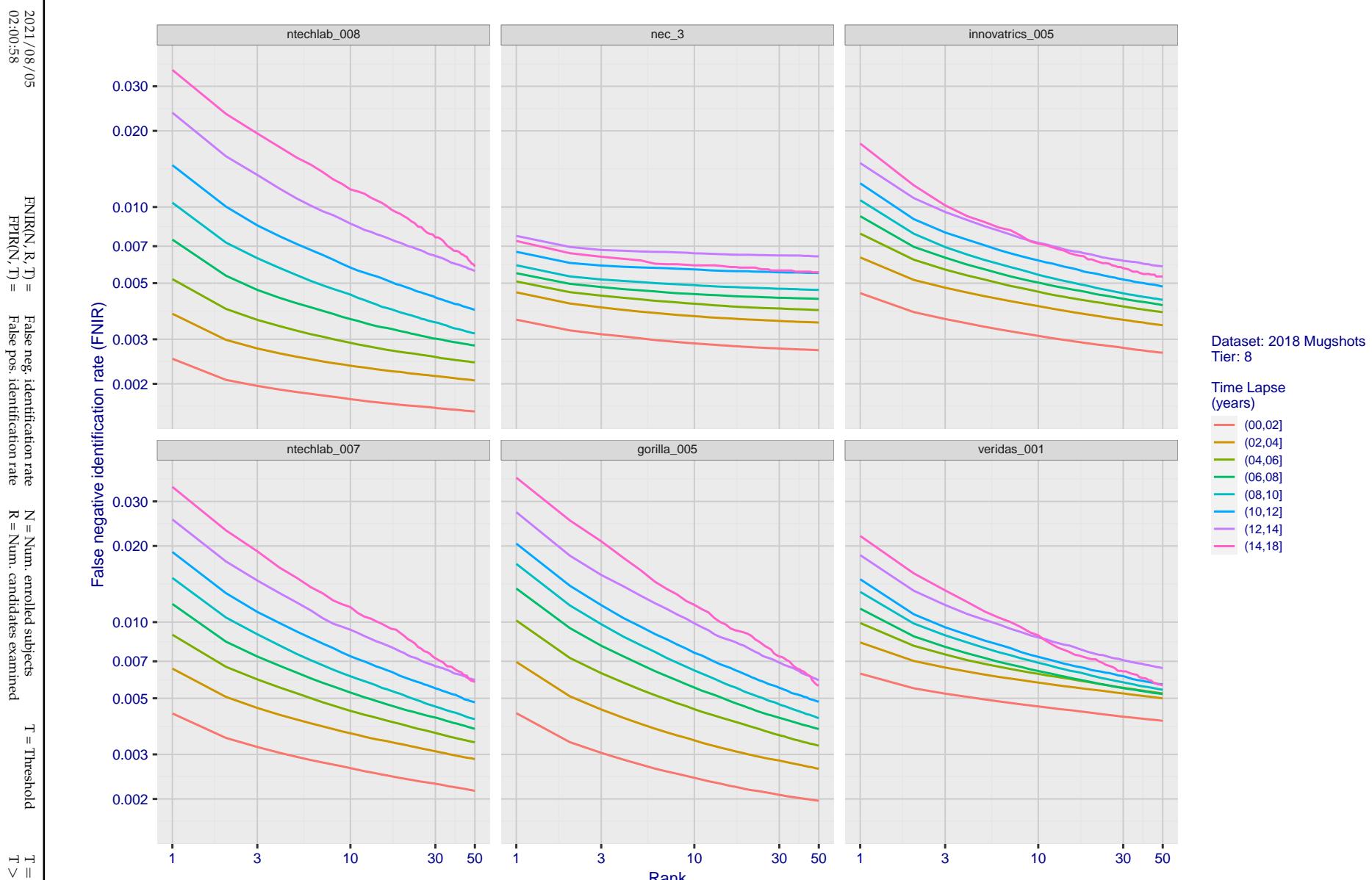


Figure 67: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

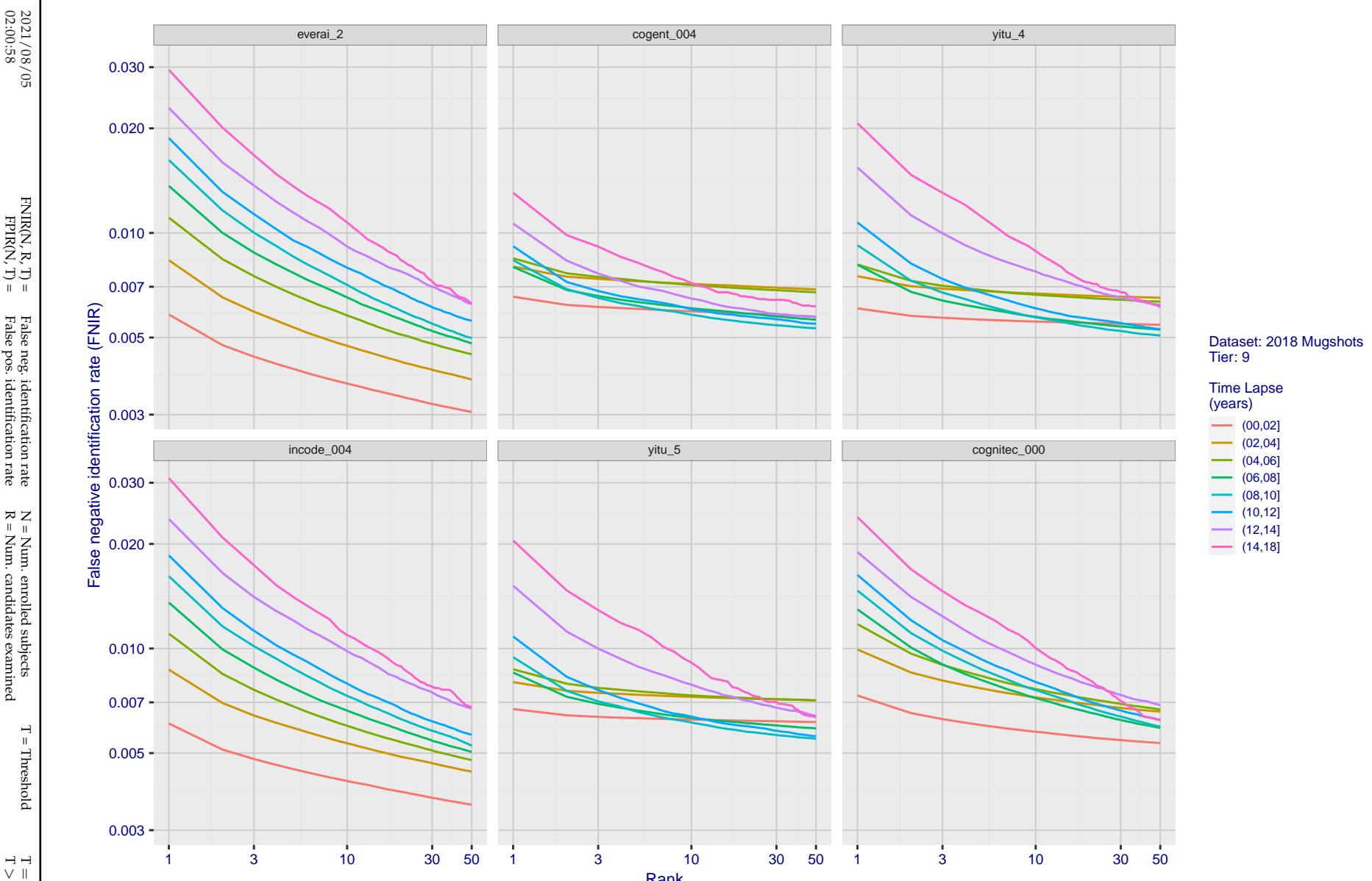


Figure 68: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

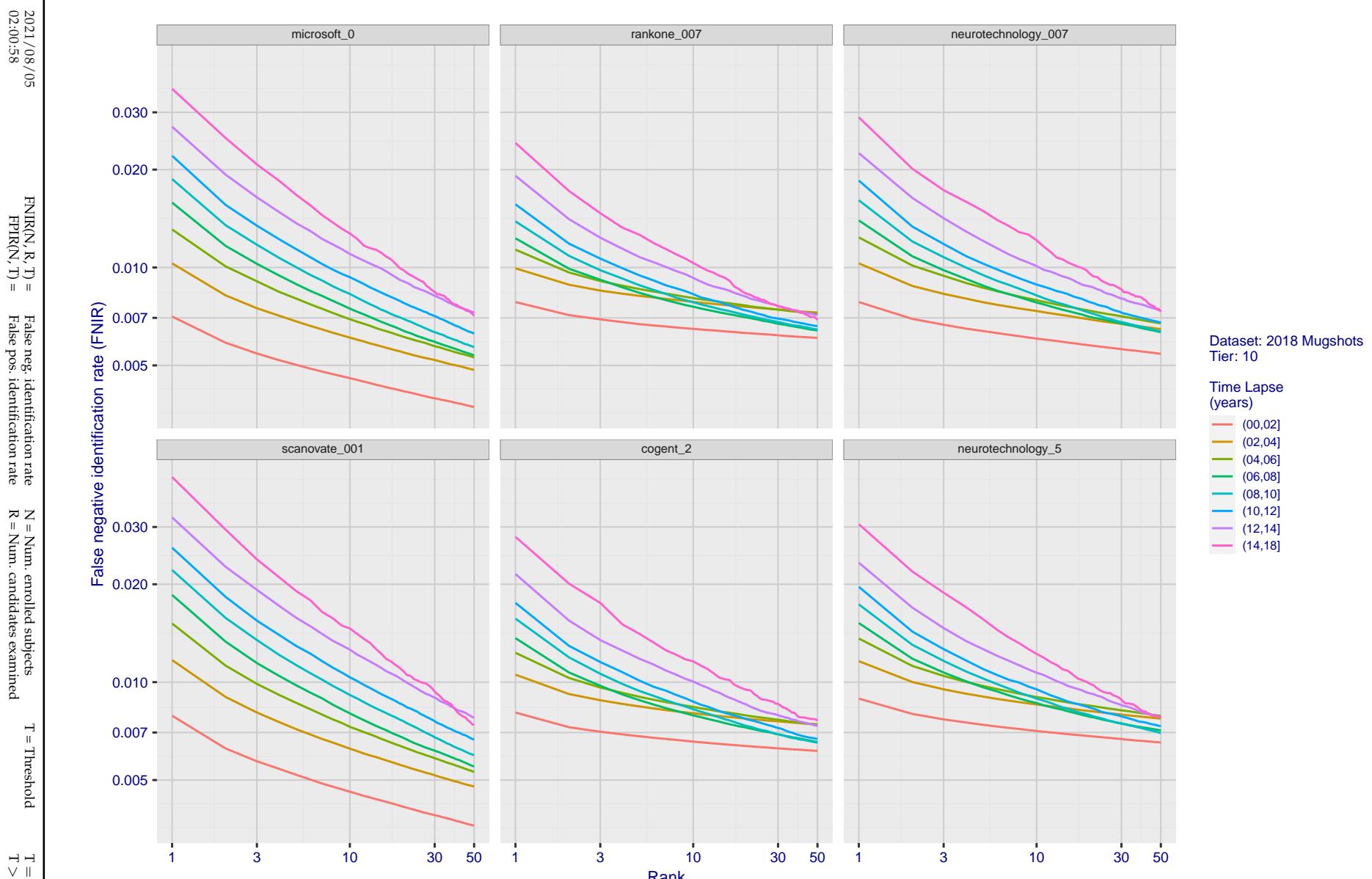


Figure 69: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

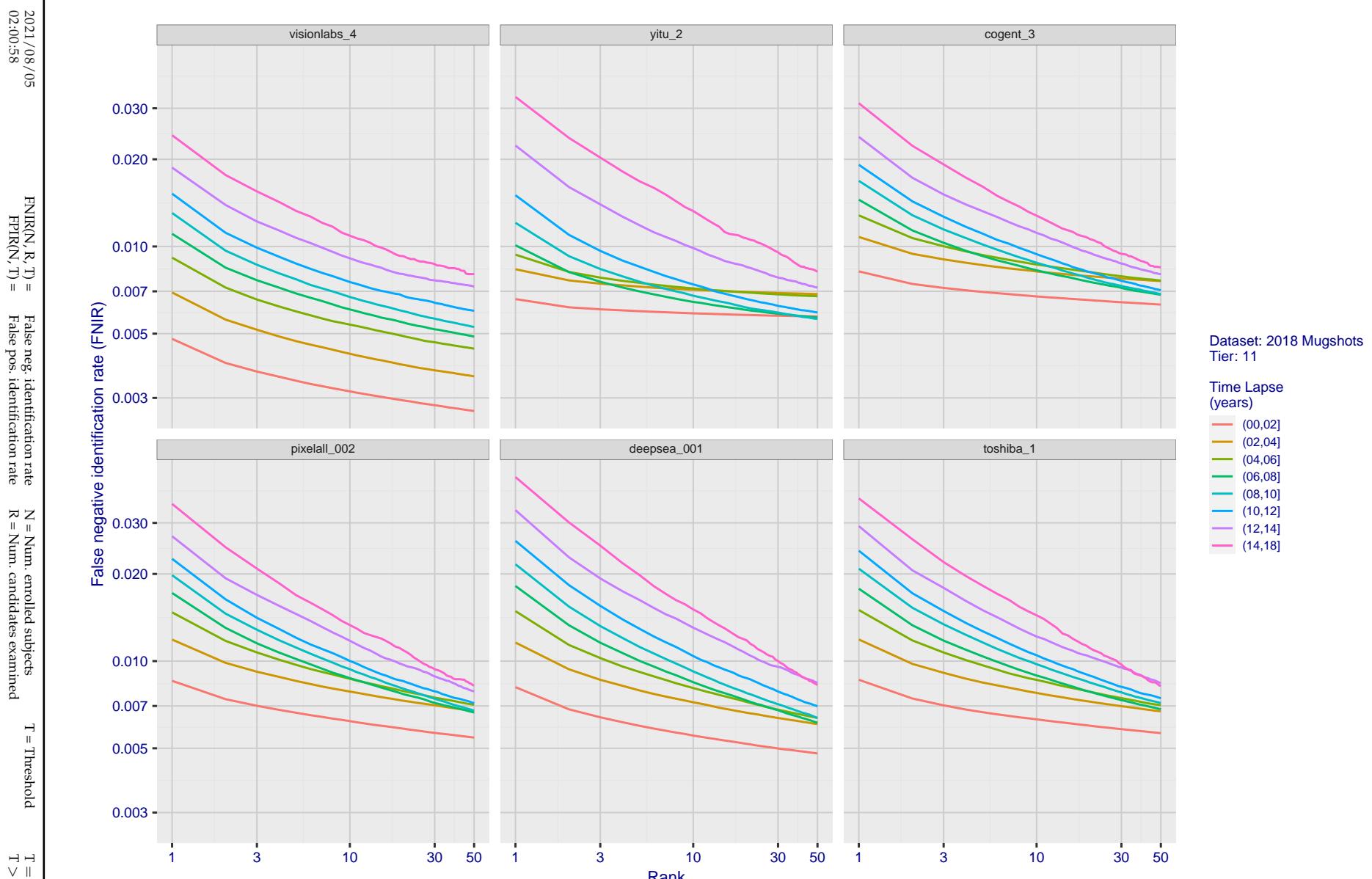


Figure 70: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

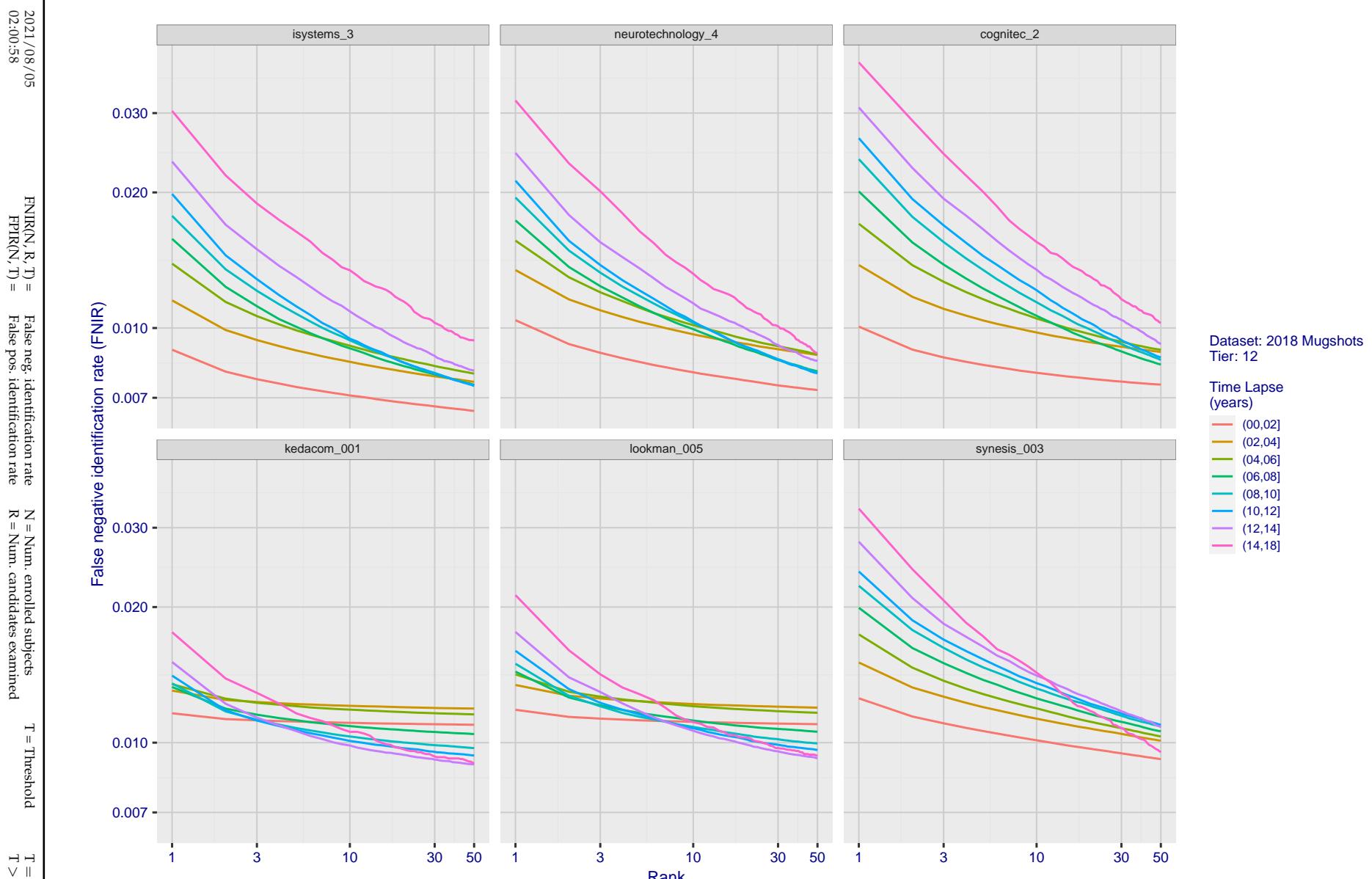


Figure 71: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

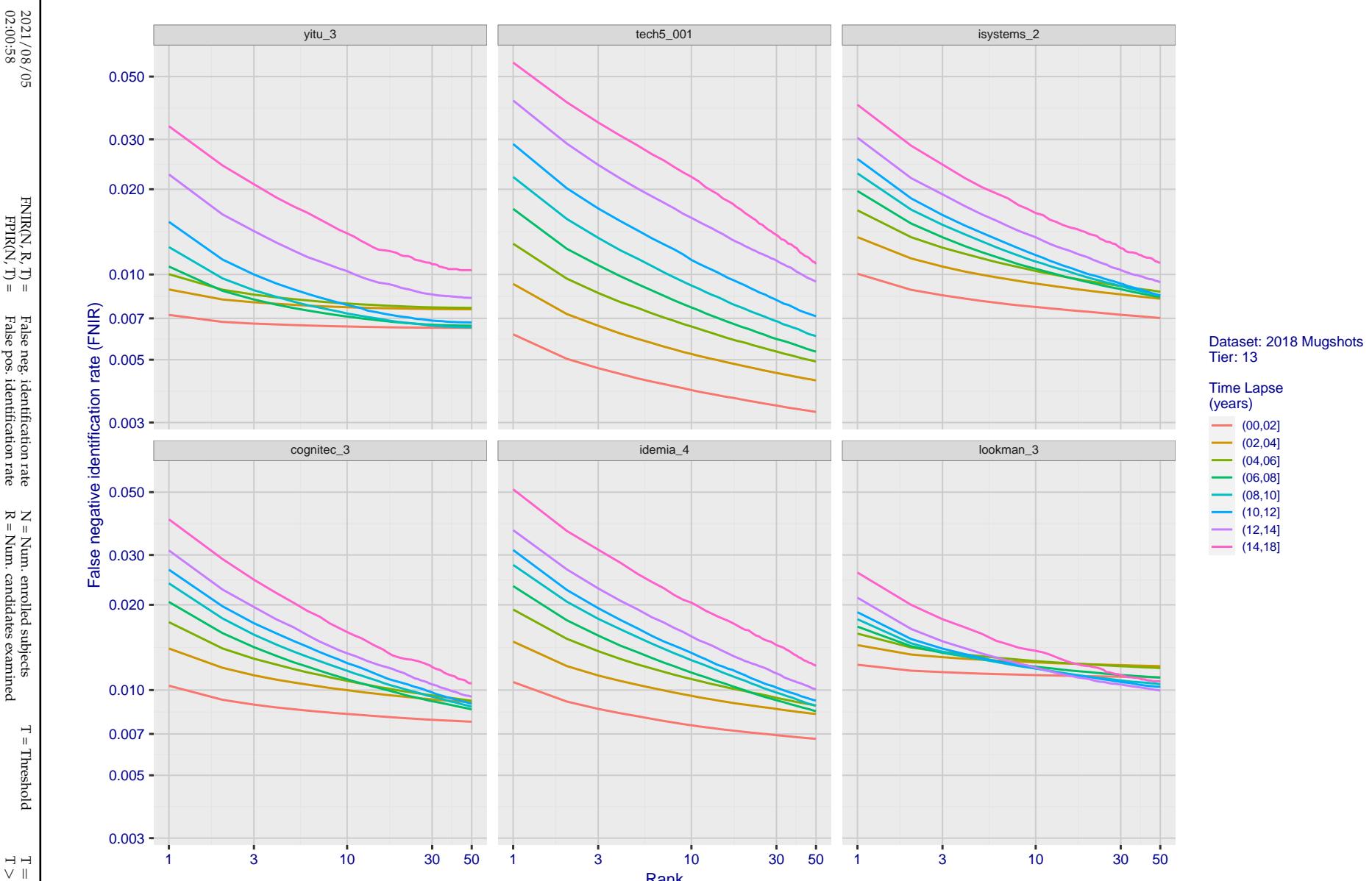


Figure 72: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

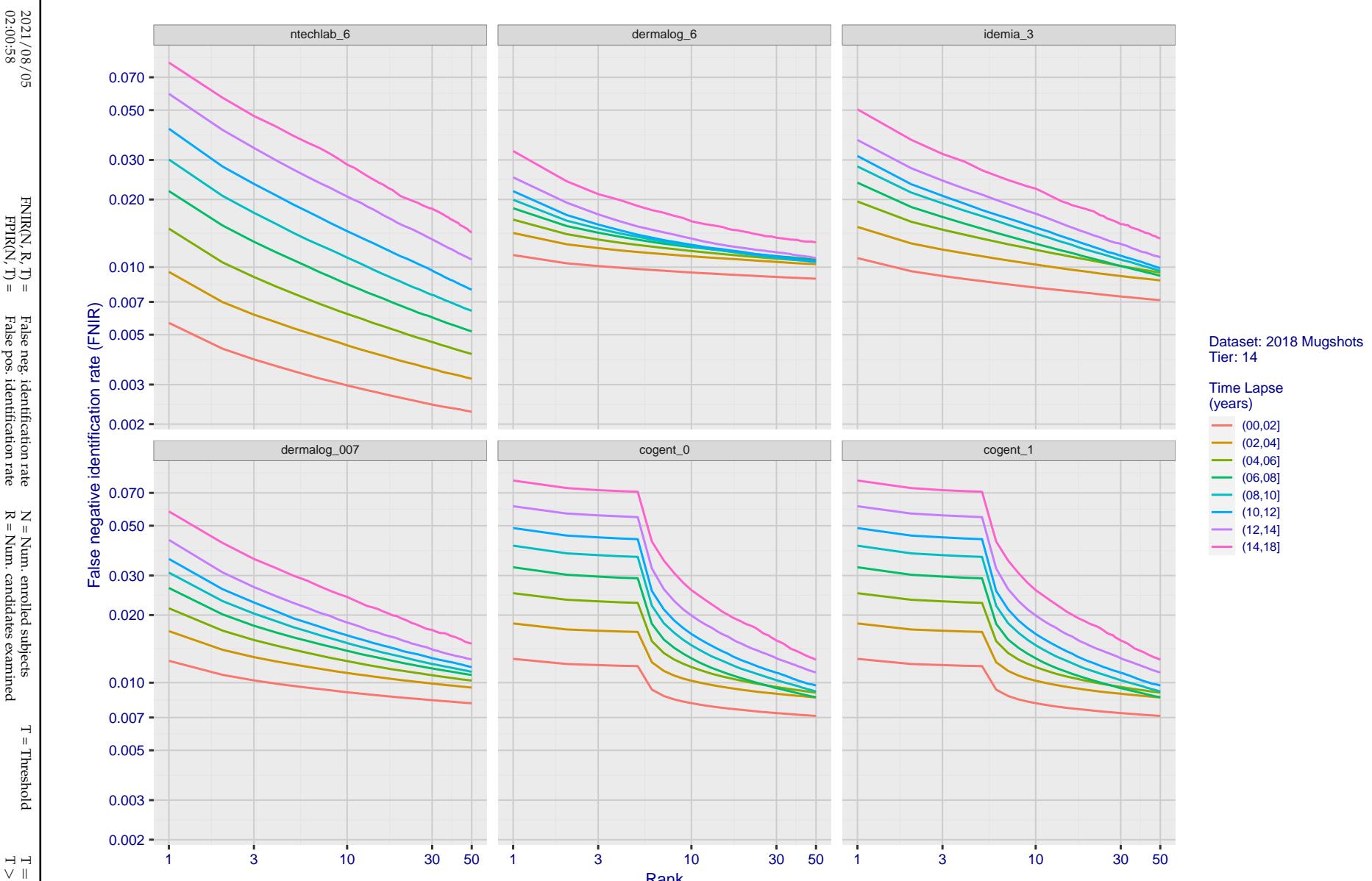


Figure 73: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

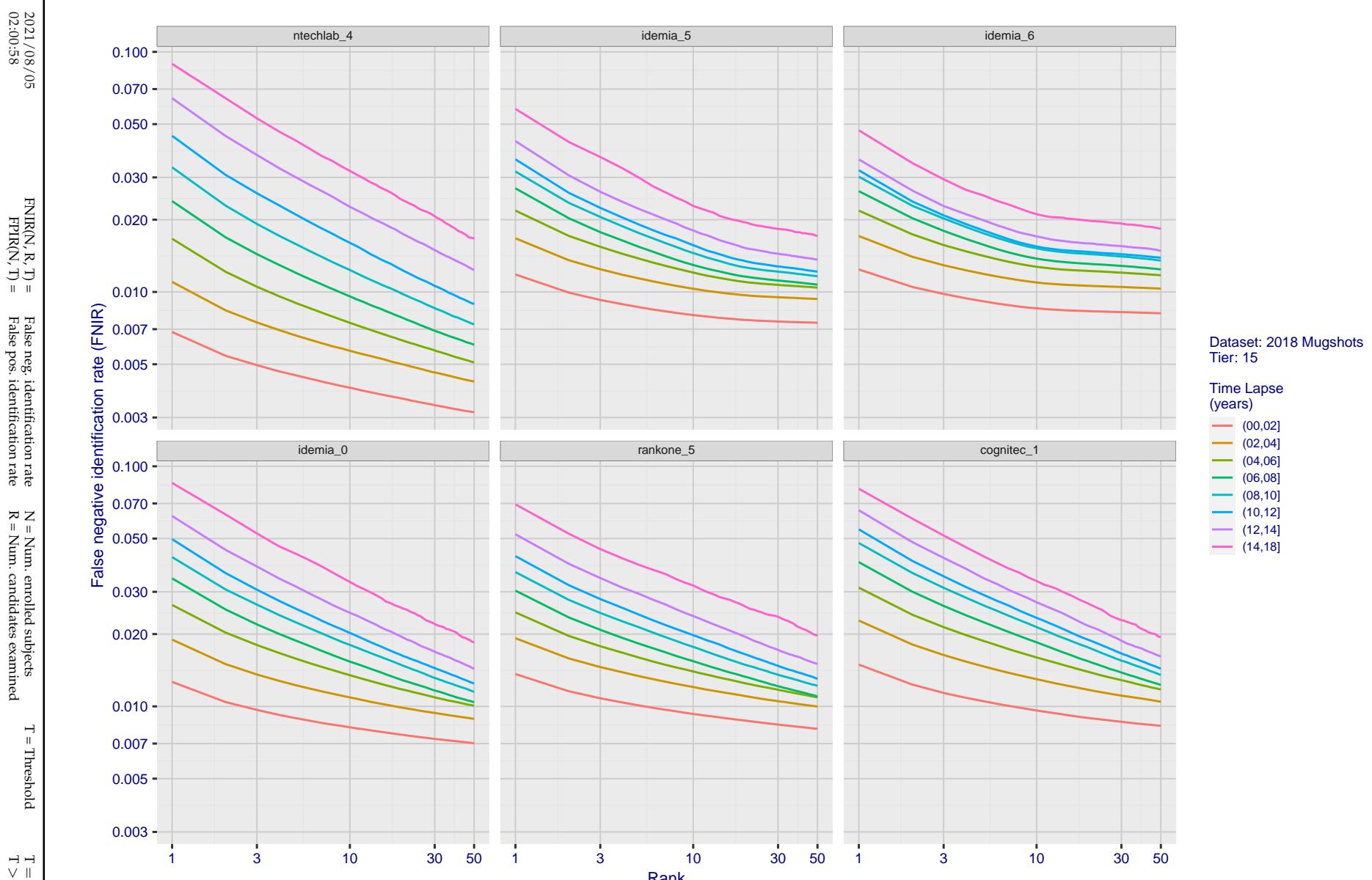


Figure 74: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

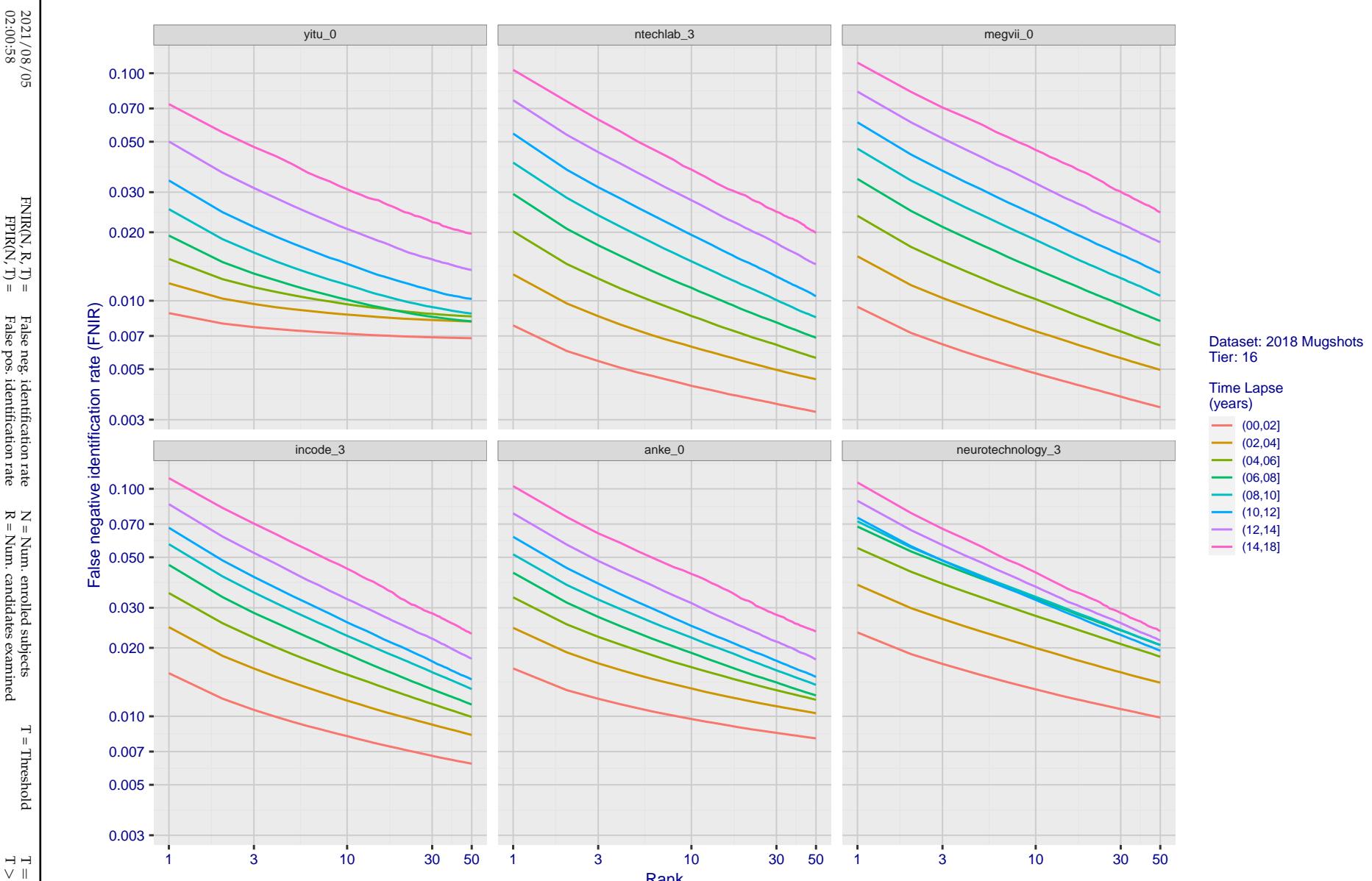


Figure 75: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

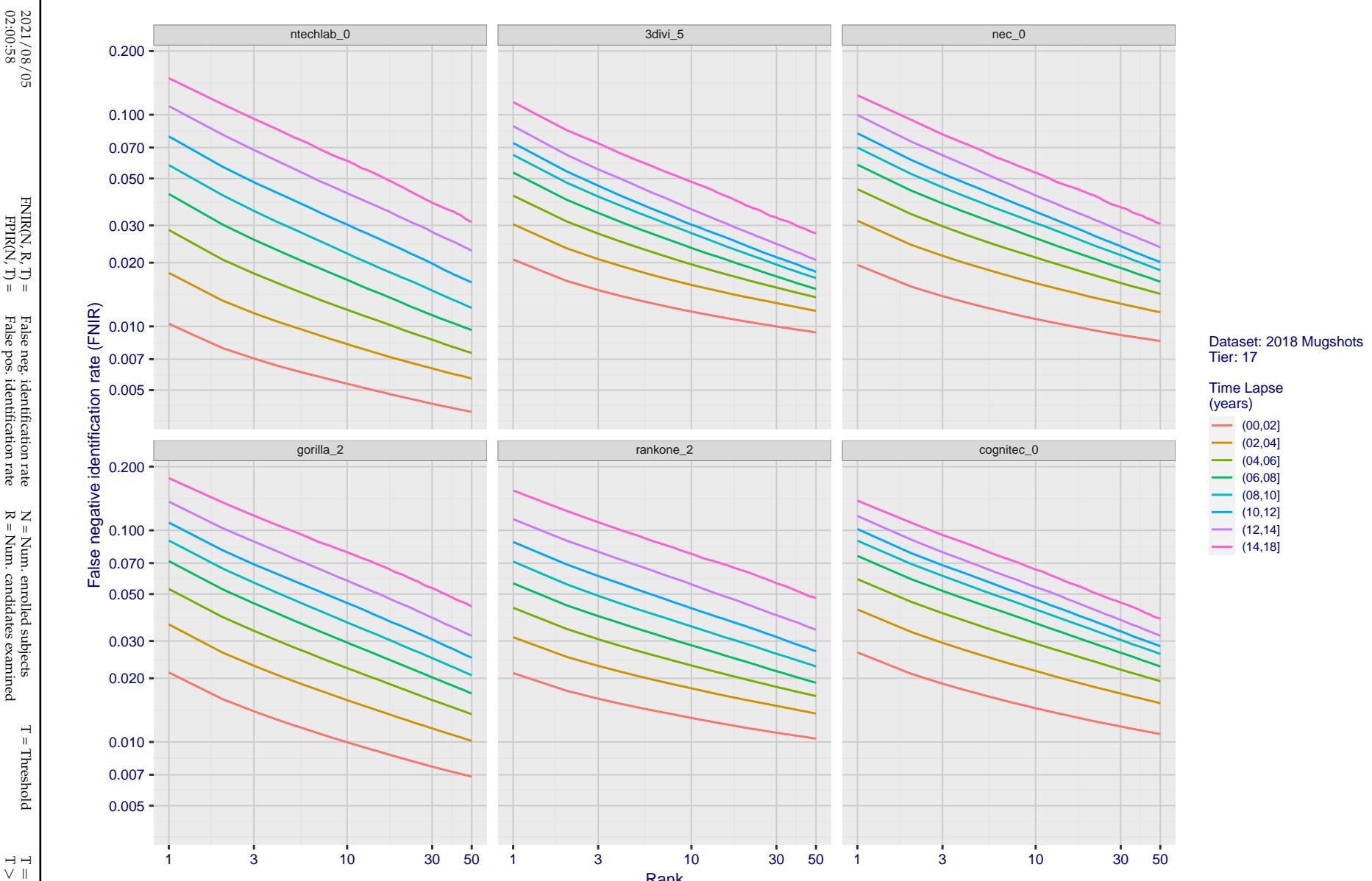


Figure 76: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

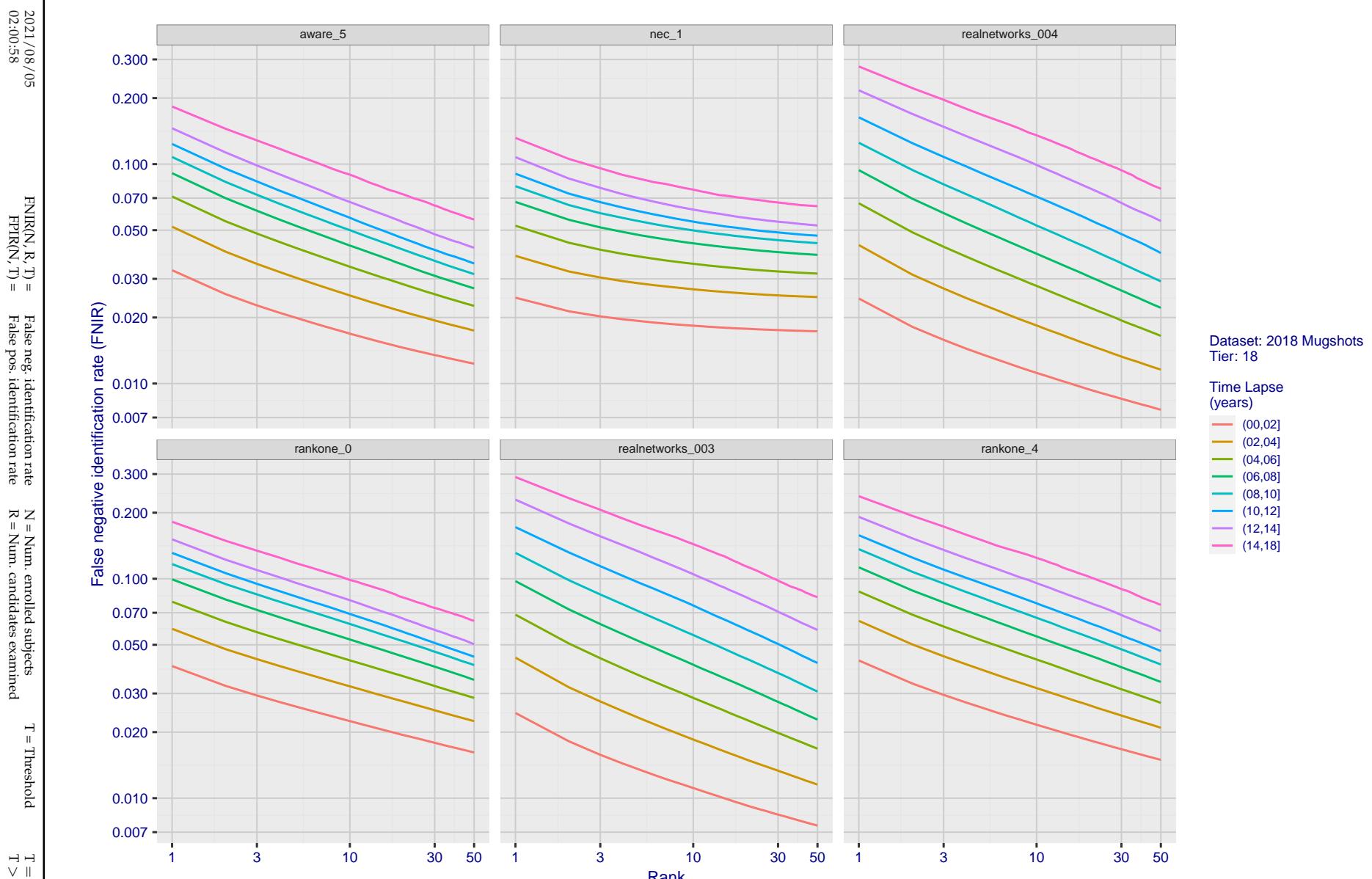


Figure 77: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

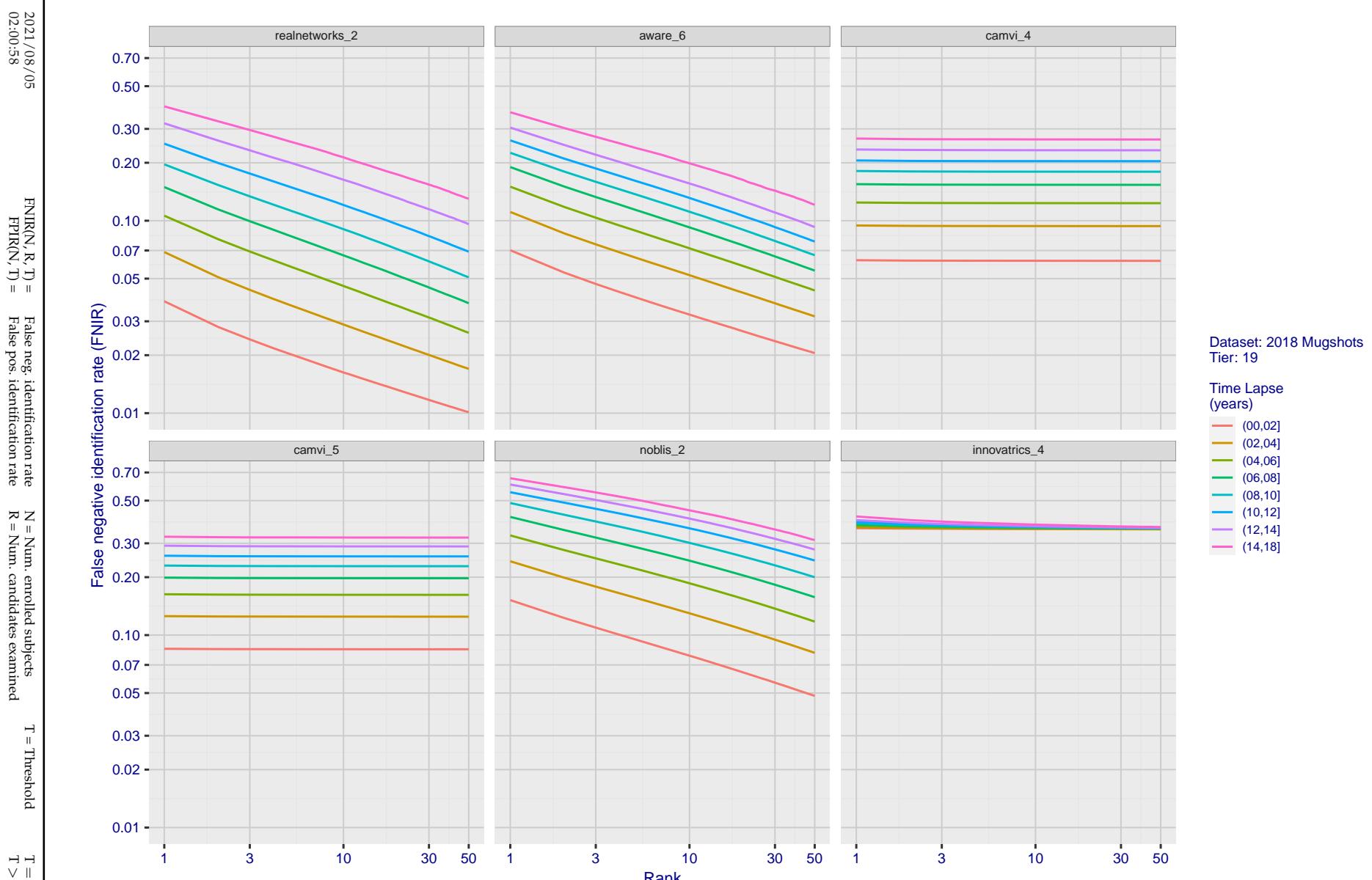


Figure 78: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

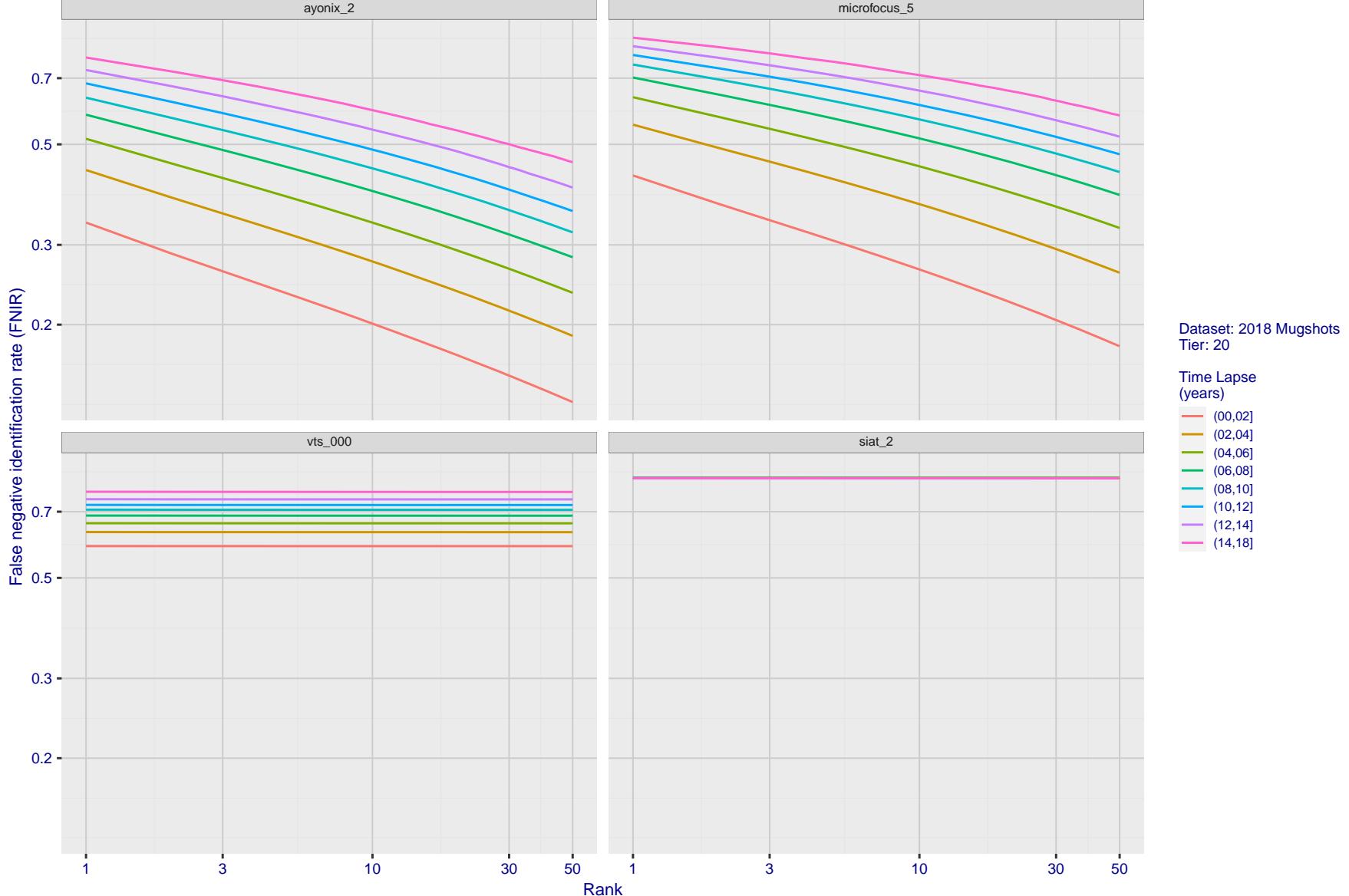


Figure 79: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment.

2021/08/05
02:00:58 FNIR(N, R, T) = False neg. identification rate
 FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

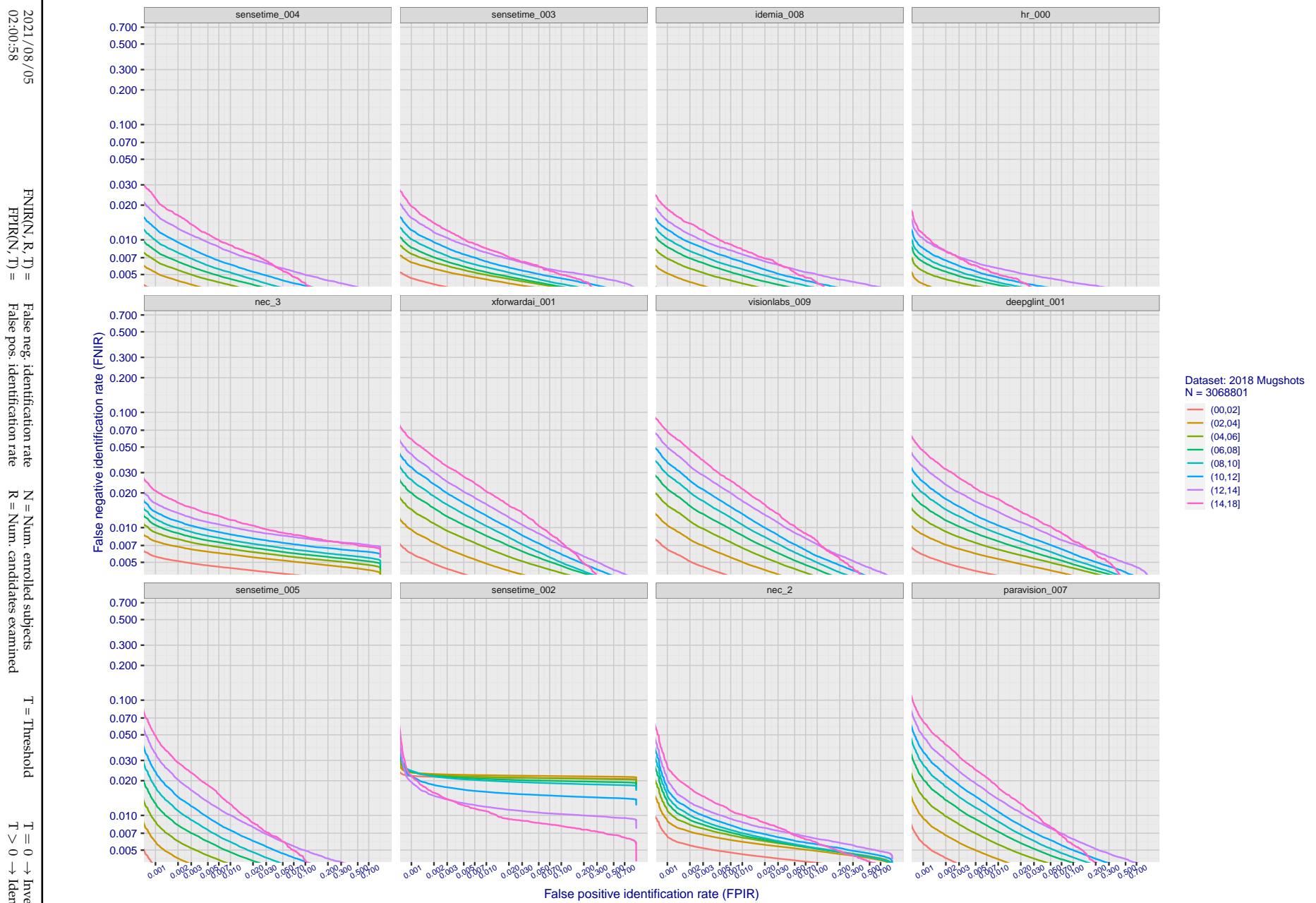


Figure 80: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. FPIR by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment. FPIR is computed from the same FRVT 2018 non-mates noted in row 3 of Table 1 with $N = 3000\,000$.

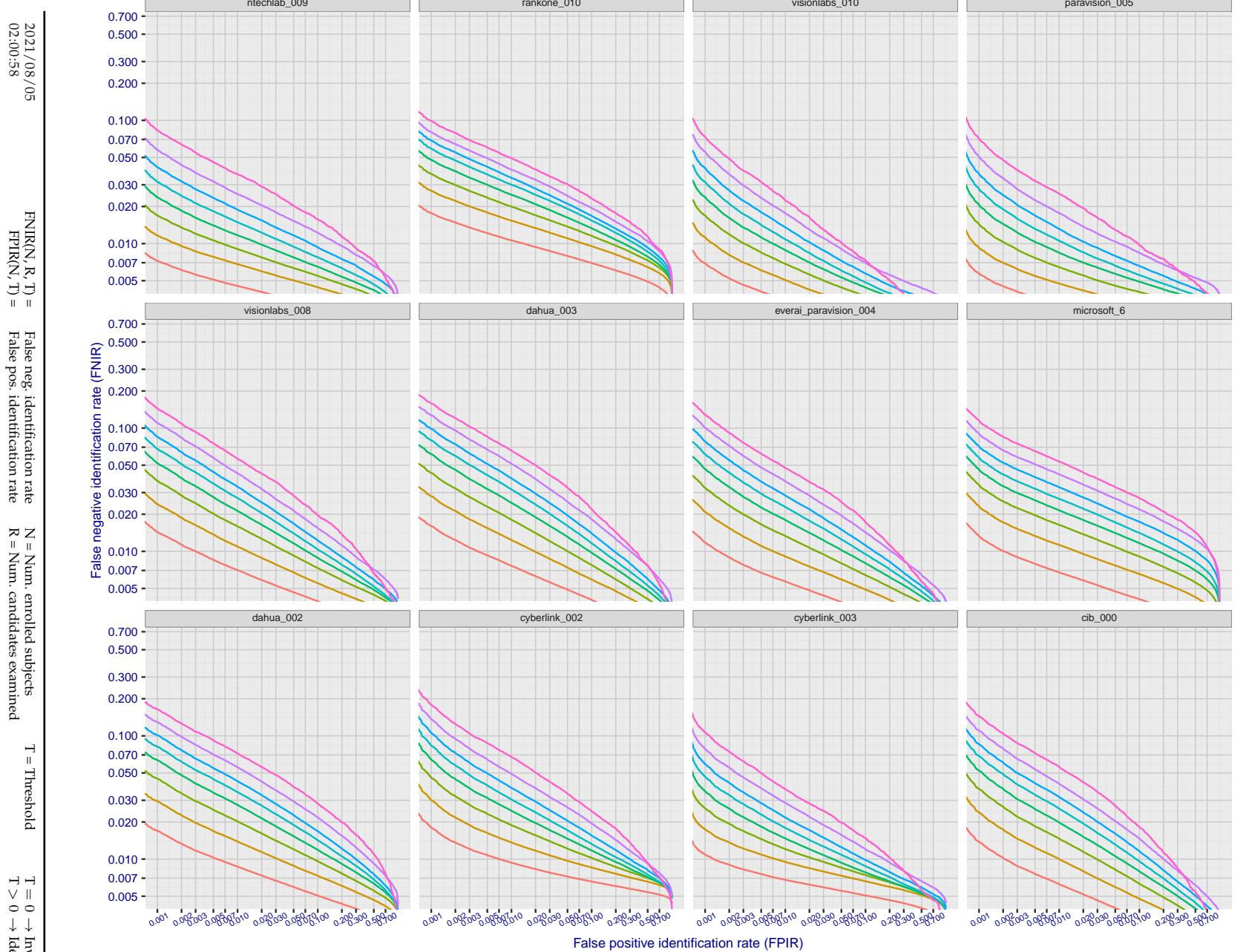


Figure 81: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. FPIR by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment. FPIR is computed from the same FRVT 2018 non-mates noted in row 3 of Table 1 with $N = 3000\,000$.

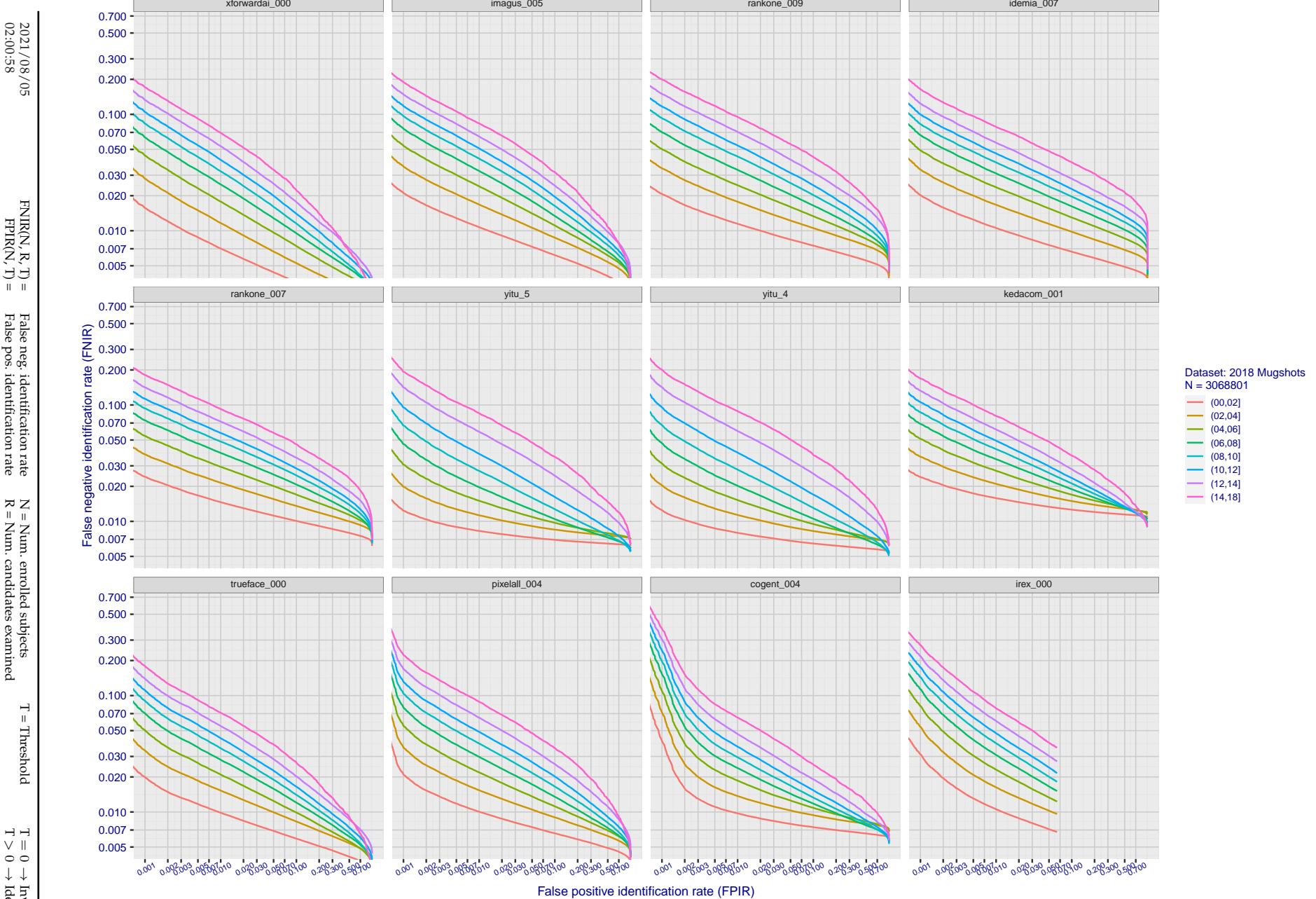


Figure 82: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. FPIR by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment. FPIR is computed from the same FRVT 2018 non-mates noted in row 3 of Table 1 with $N = 3\,000\,000$.

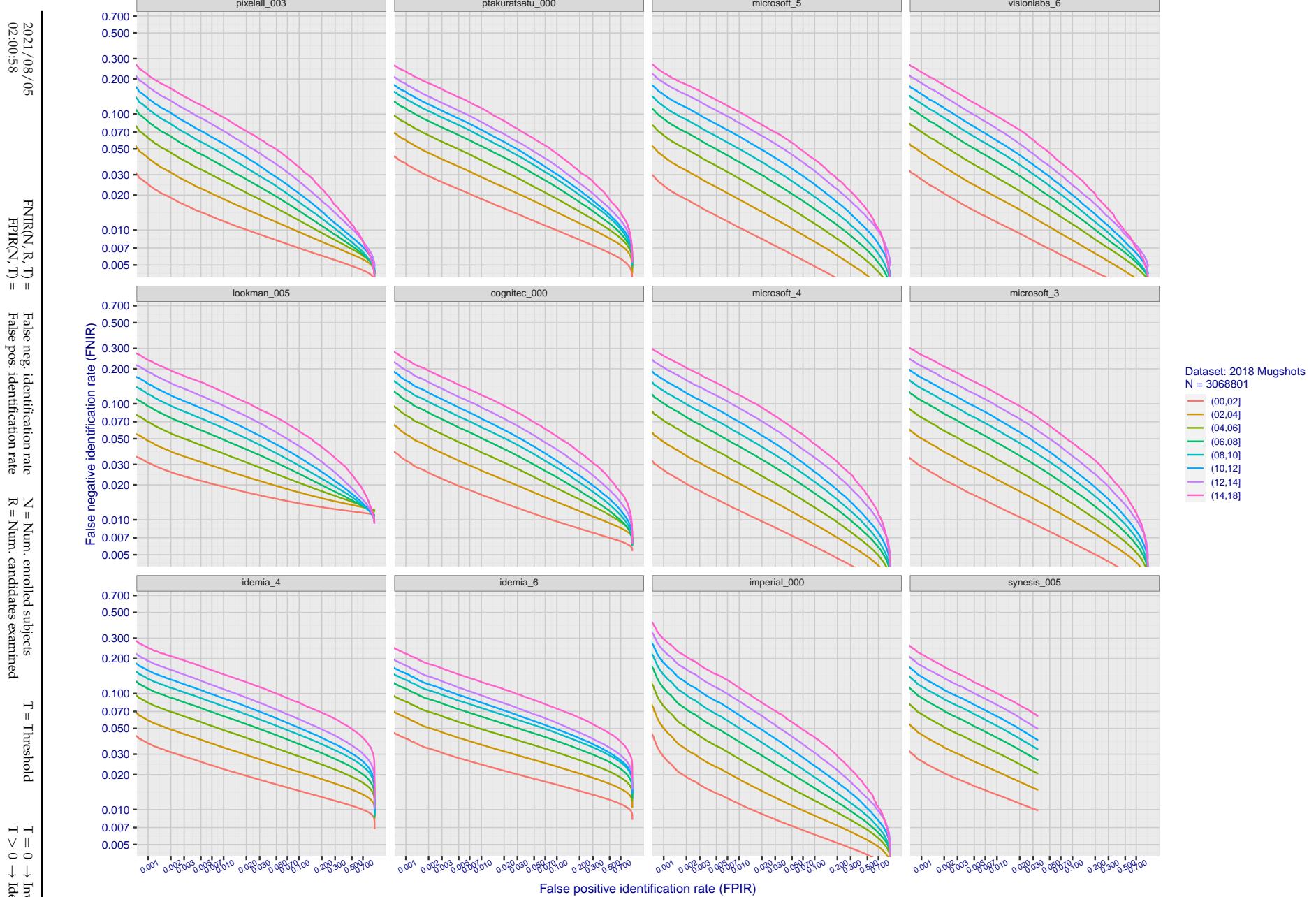


Figure 83: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. FPIR by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment. FPIR is computed from the same FRVT 2018 non-mates noted in row 3 of Table 1 with N = 3 000 000.

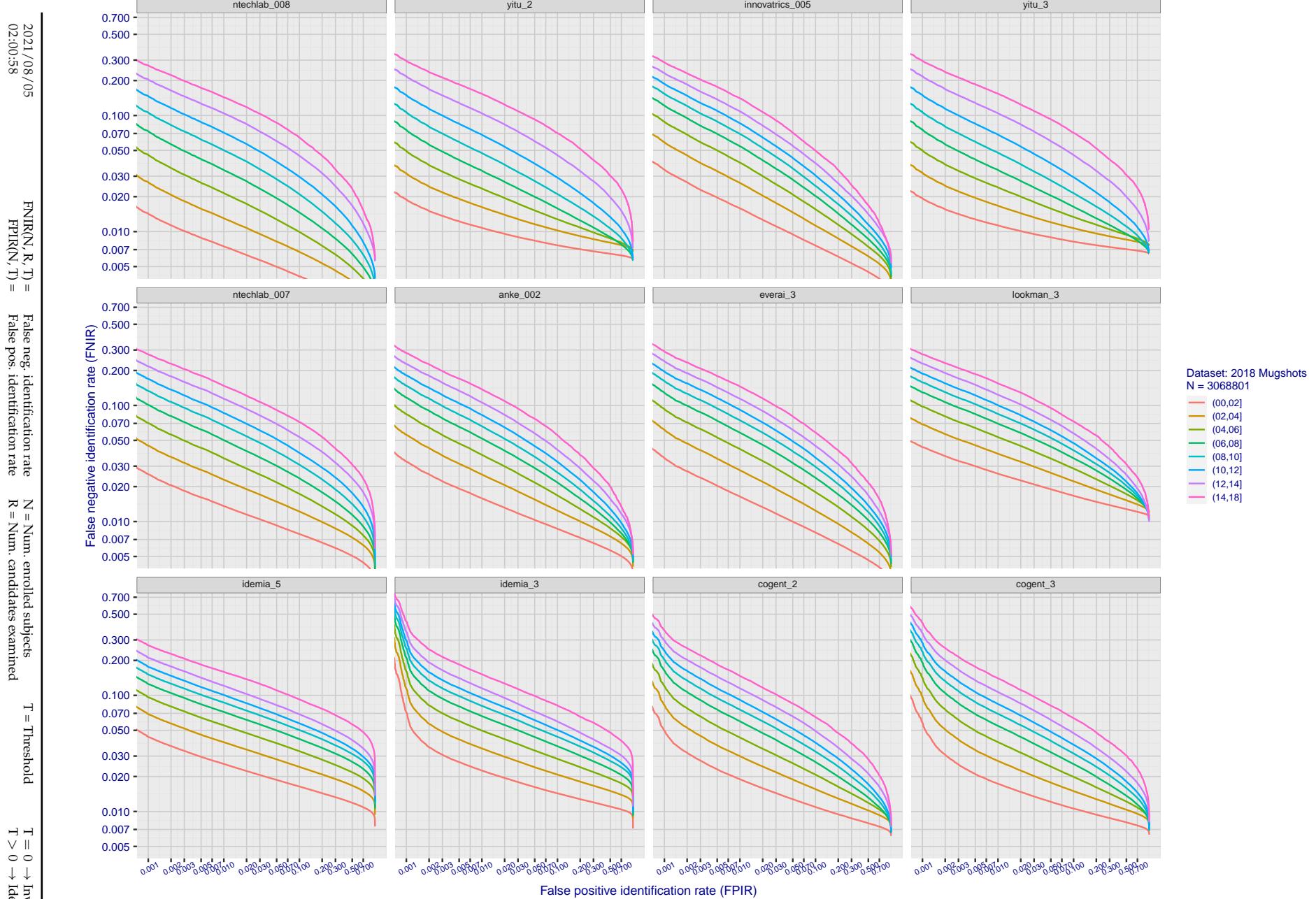


Figure 84: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. FPIR by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment. FPIR is computed from the same FRVT 2018 non-mates noted in row 3 of Table 1 with $N = 3000\,000$.

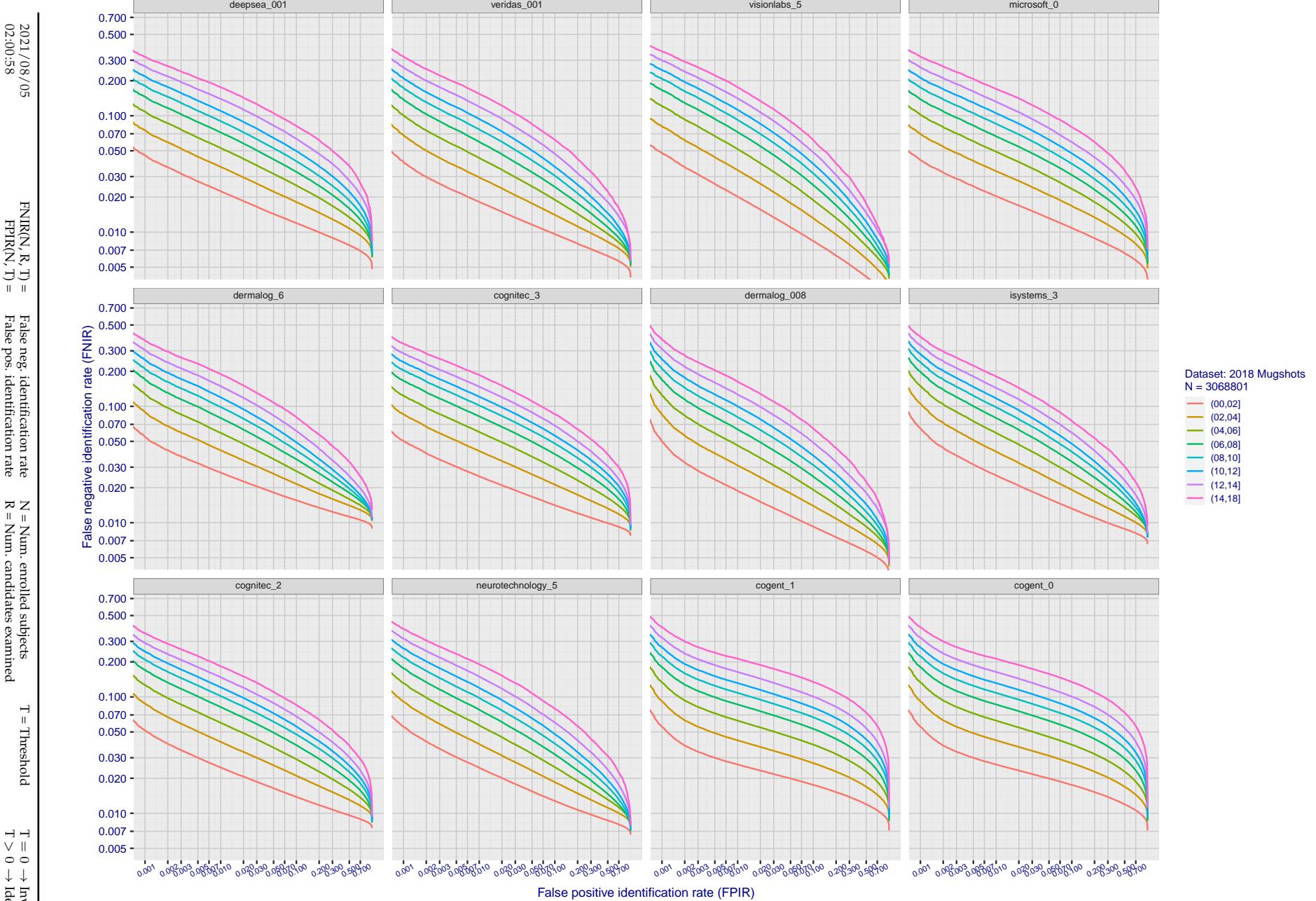


Figure 85: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. FPIR by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment. FPIR is computed from the same FRVT 2018 non-mates noted in row 3 of Table 1 with N = 3 000 000.

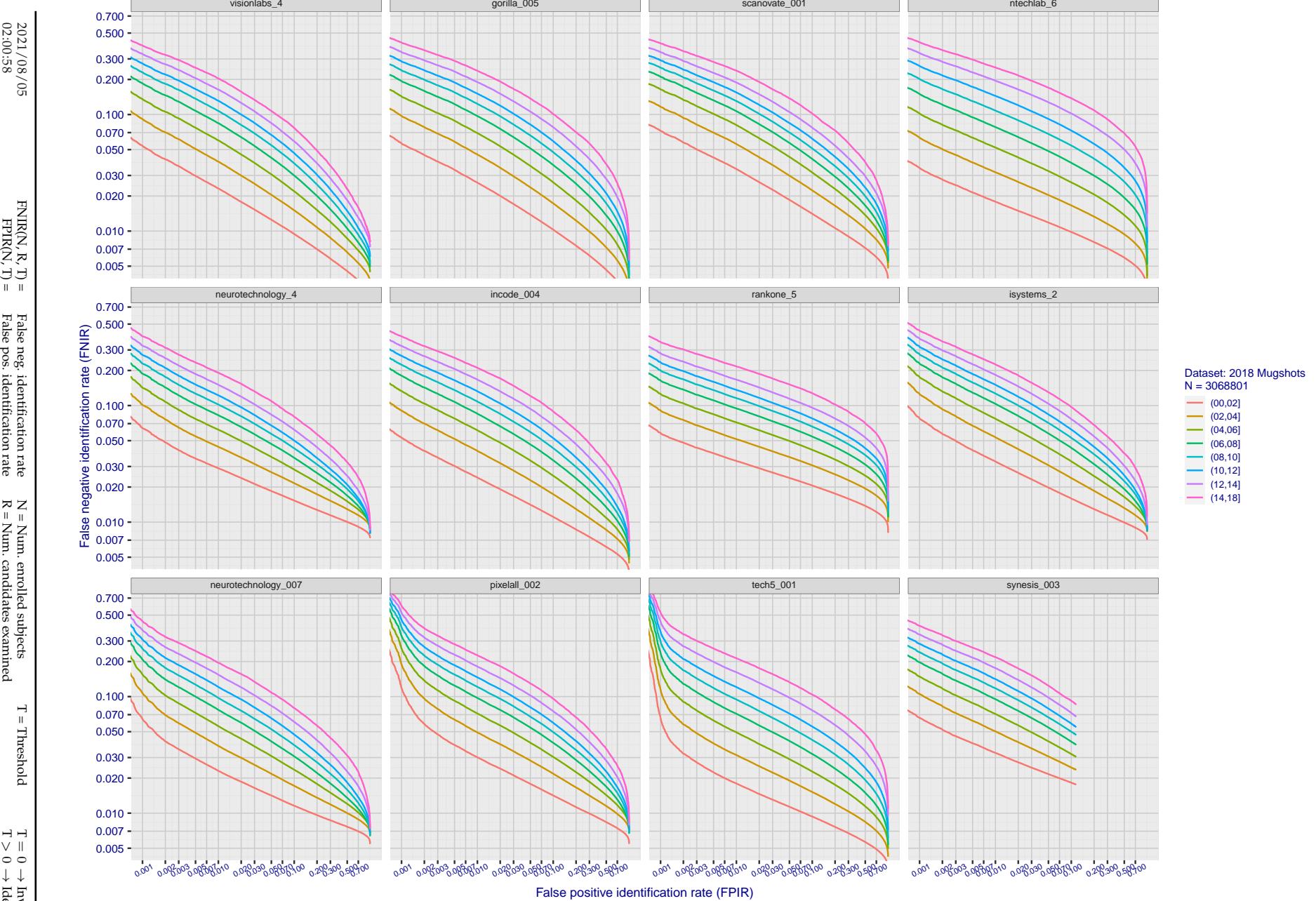


Figure 86: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. FPIR by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment. FPIR is computed from the same FRVT 2018 non-mates noted in row 3 of Table 1 with $N = 3\,000\,000$.

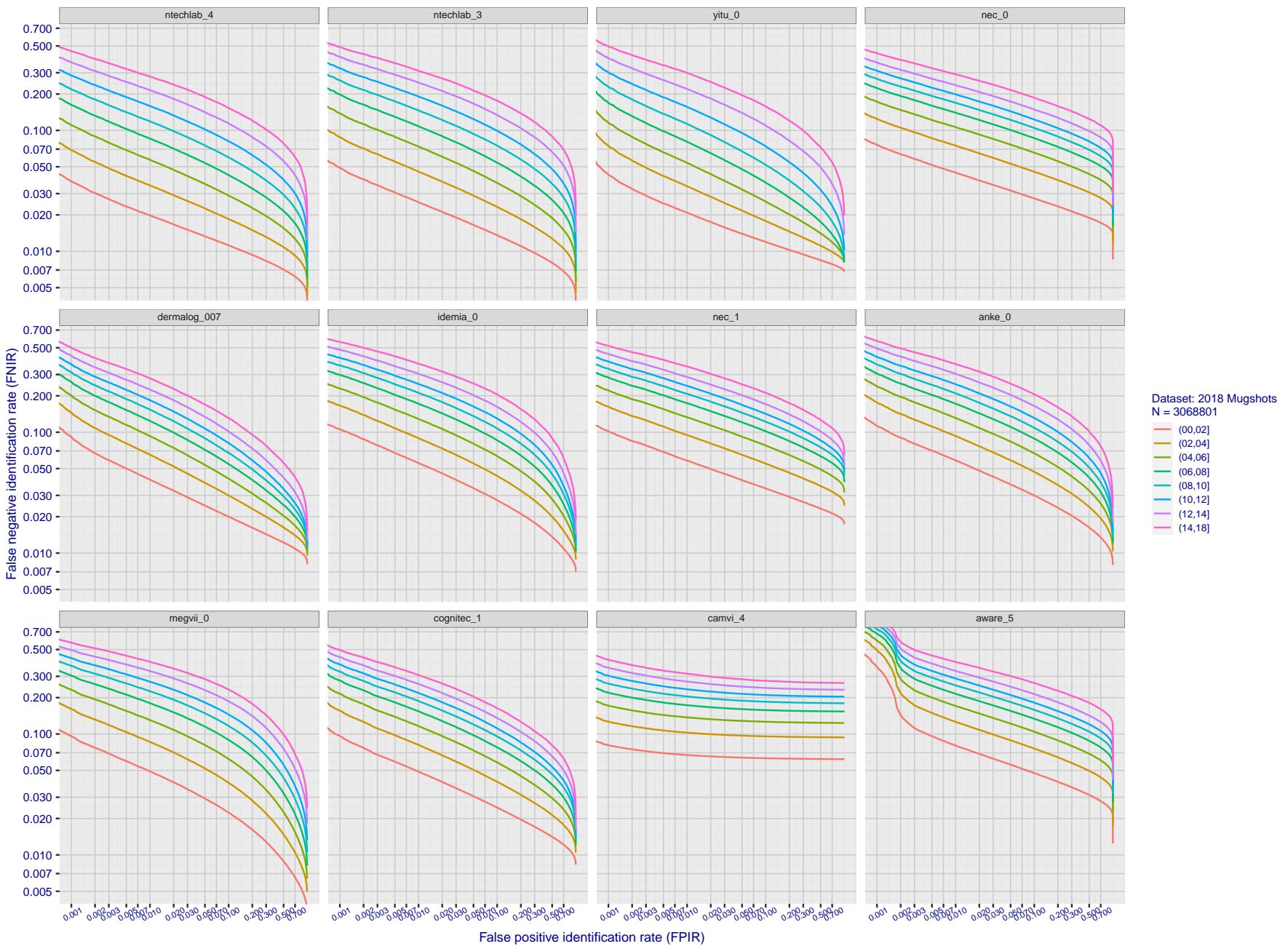


Figure 87: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. FPIR by time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment. FPIR is computed from the same FRVT 2018 non-mates noted in row 3 of Table 1 with $N = 3\,000\,000$.

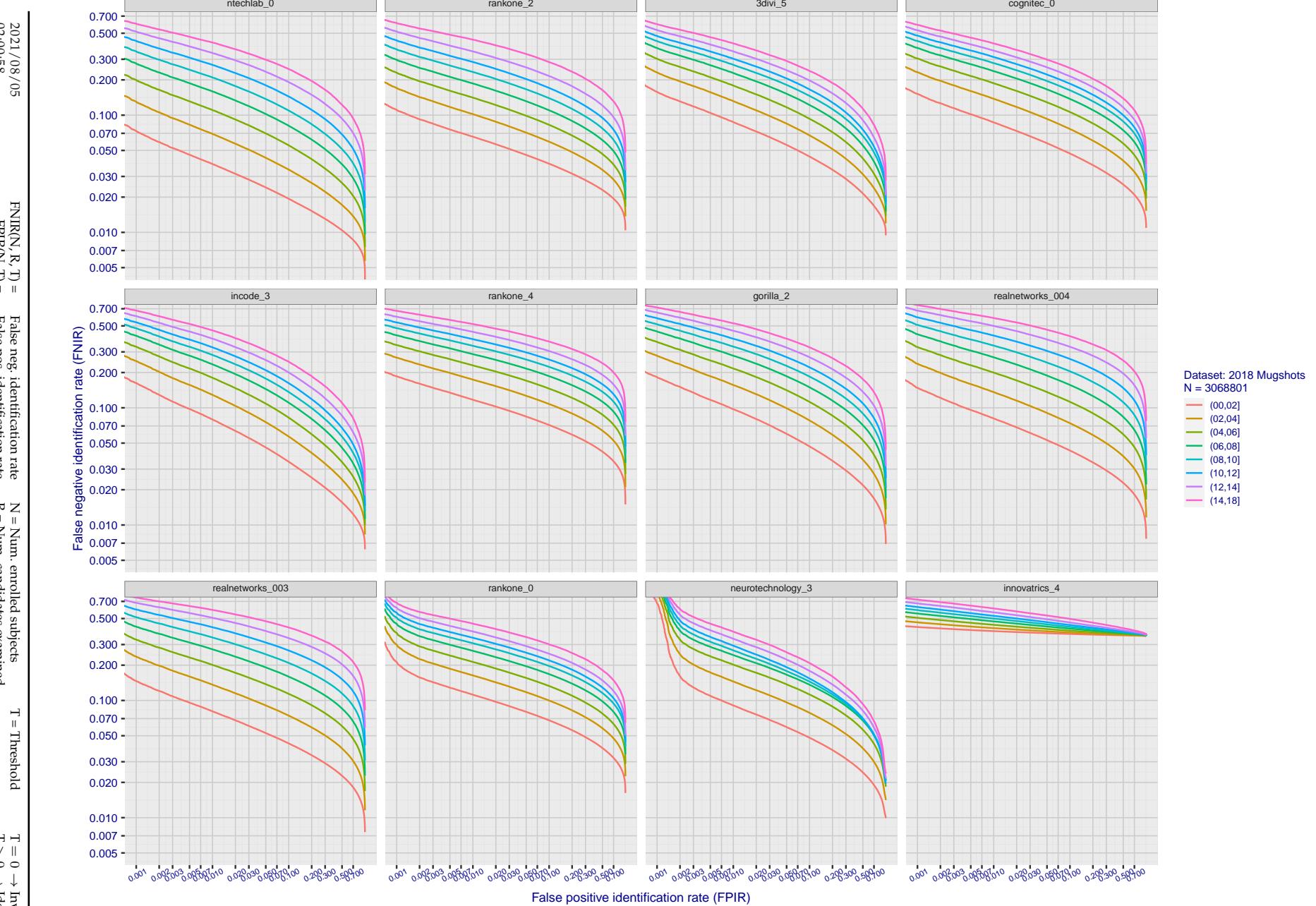


Figure 88: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. FPIR by time elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment. FPIR is computed from the same FRVT 2018 non-mates noted in row 3 of Table 1 with $N = 3\,000\,000$.

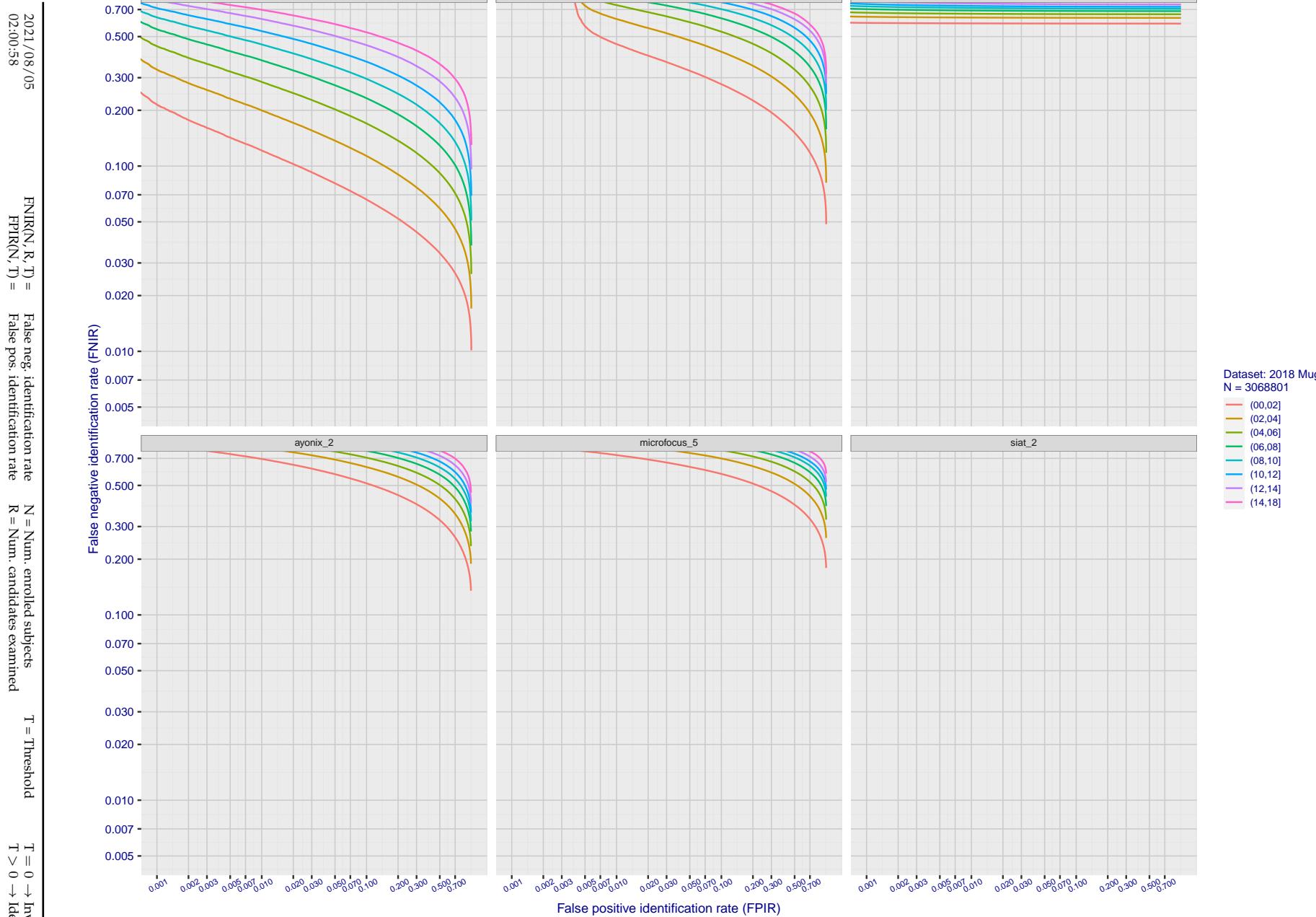


Figure 89: [FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. FPIR by time elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 1 and binned by number of years between search and initial enrollment. FPIR is computed from the same FRVT 2018 non-mates noted in row 3 of Table 1 with $N = 3\,000\,000$.

2021/08/05 02:00:58	$\text{FNIR}(N, R, T) =$ $\text{FPTR}(N, T) =$	False neg. identification rate False pos. identification rate	$N =$ Num. enrolled subjects $R =$ Num. candidates examined	$T =$ Threshold $T > 0 \rightarrow$ Identification	$T = 0 \rightarrow$ Investigation
------------------------	---	--	--	---	-----------------------------------

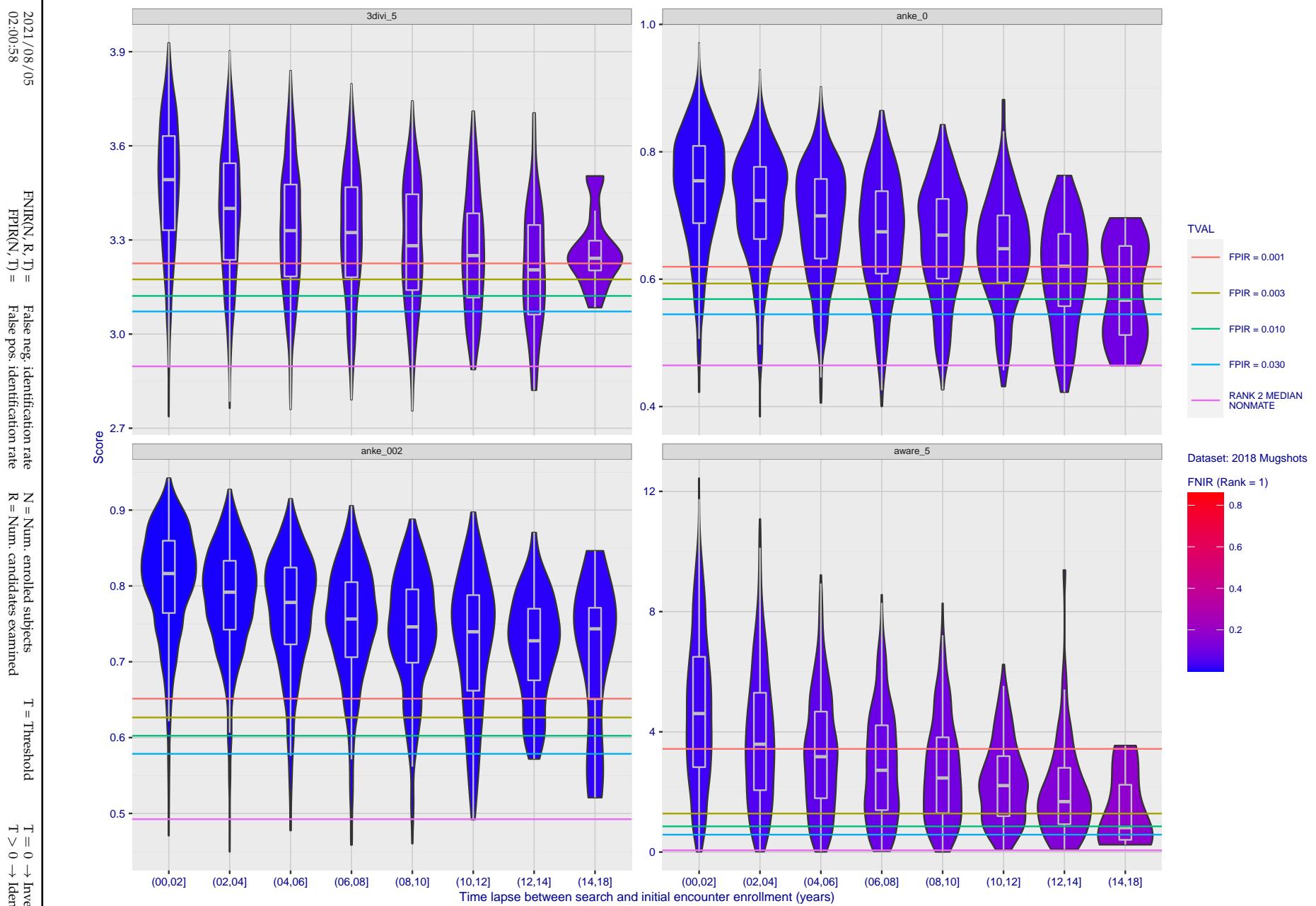


Figure 90: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

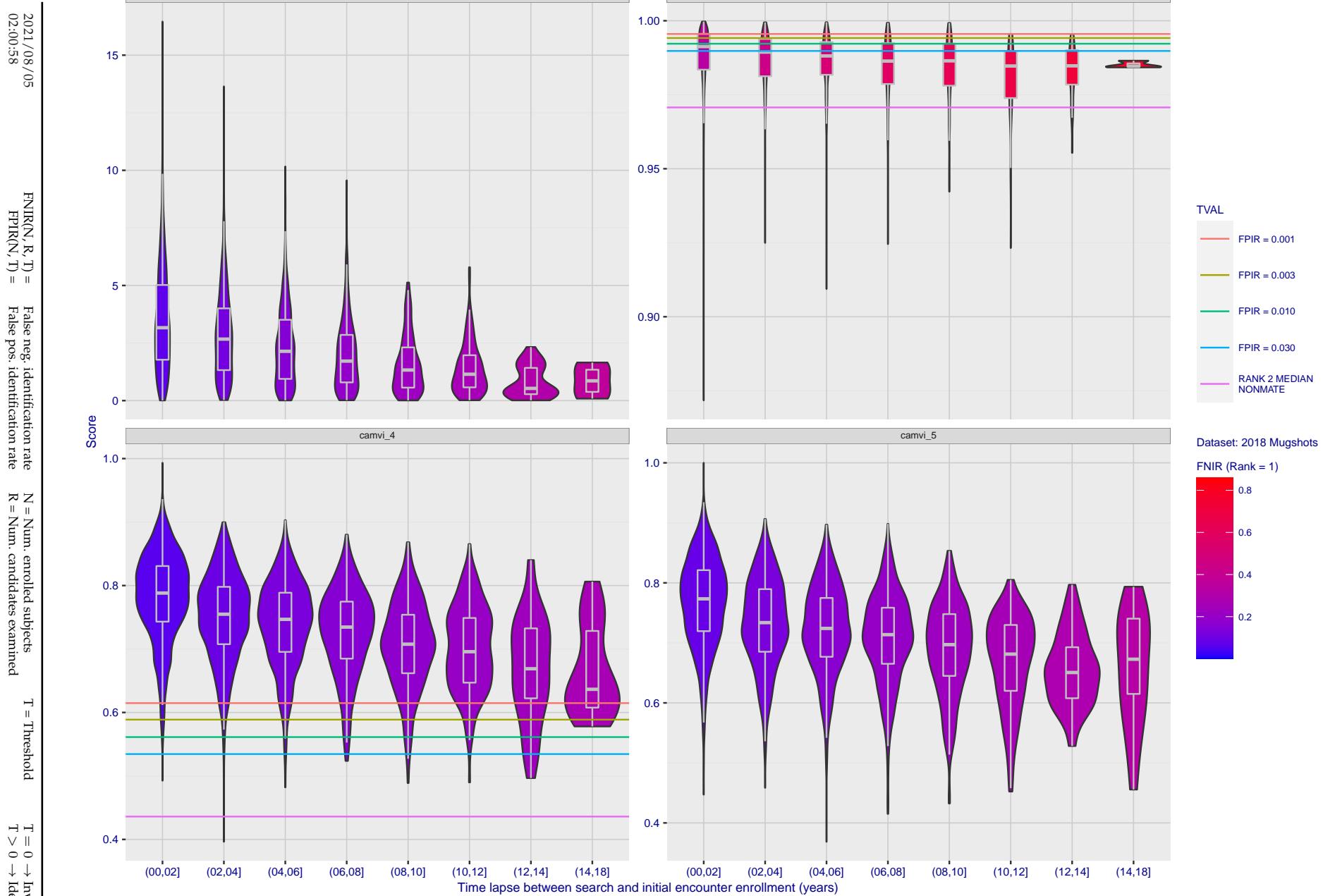


Figure 91: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

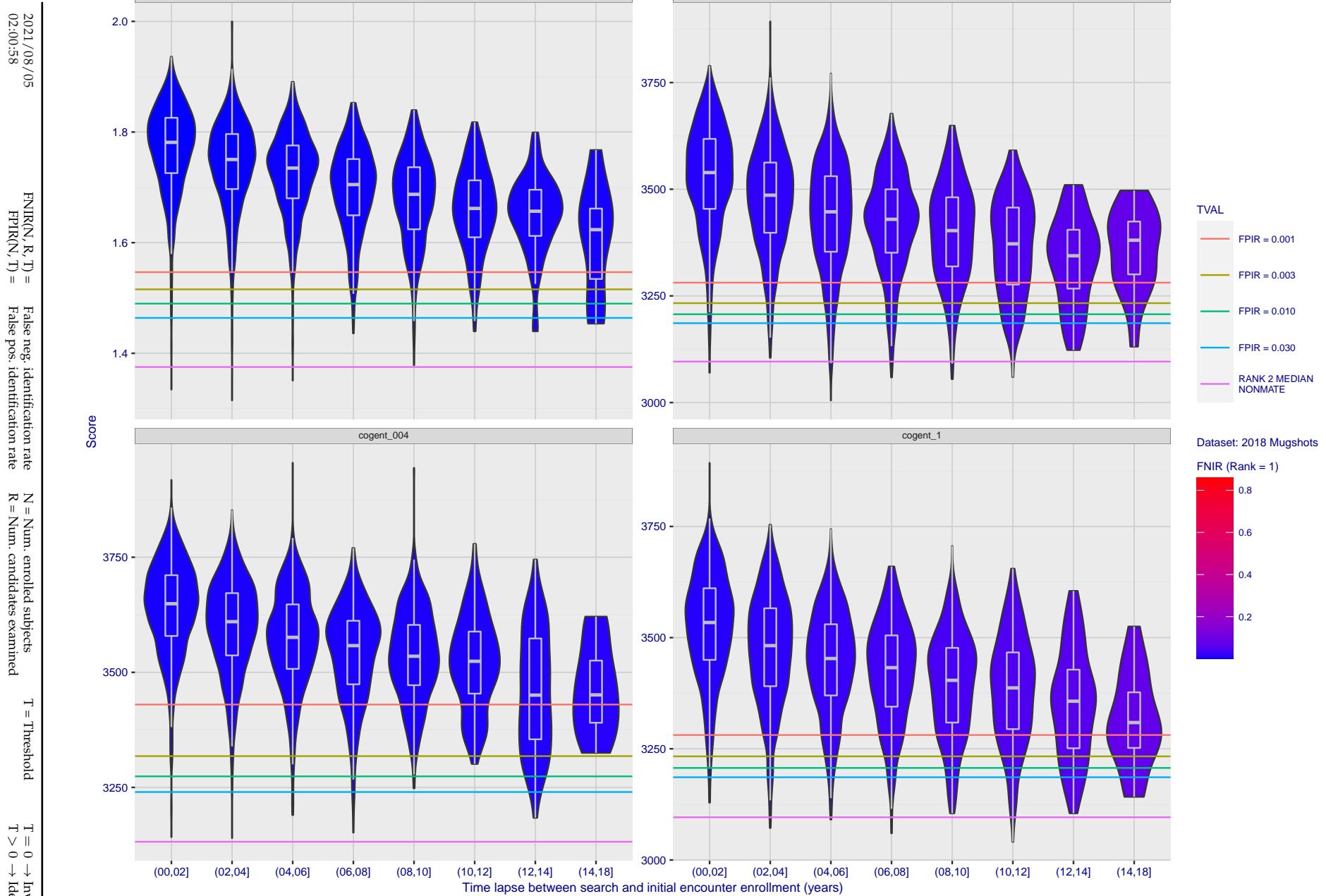


Figure 92: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

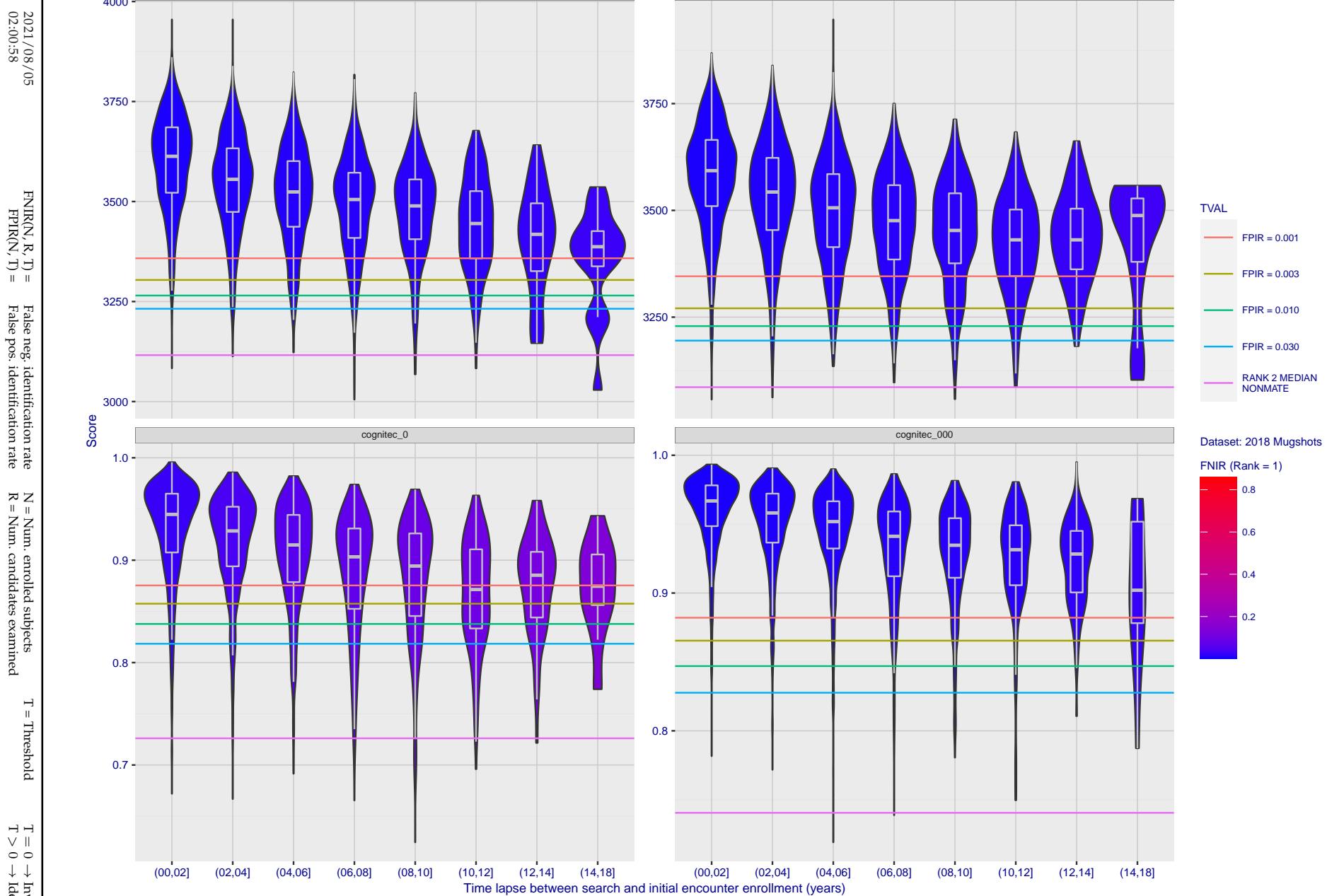


Figure 93: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

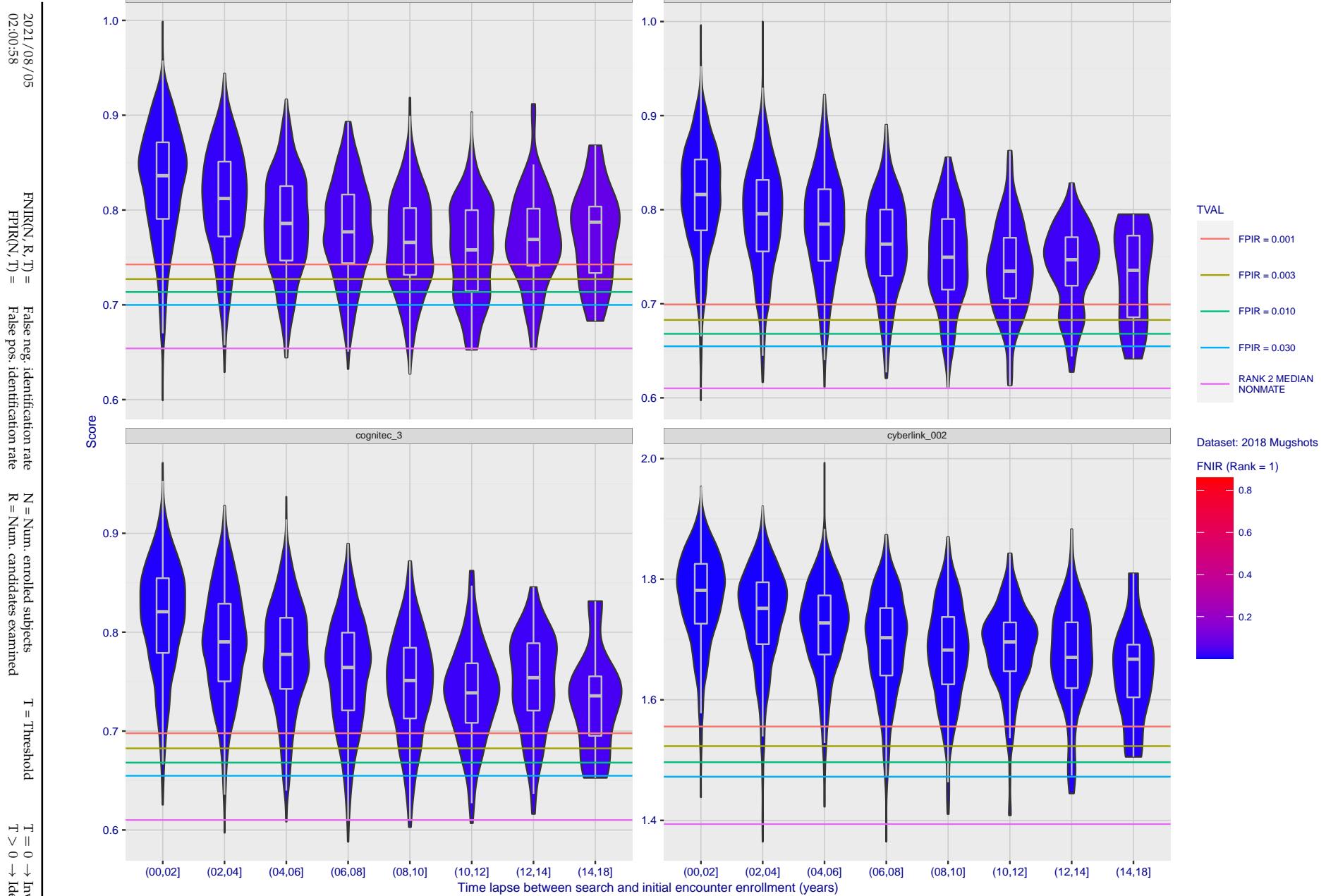


Figure 94: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

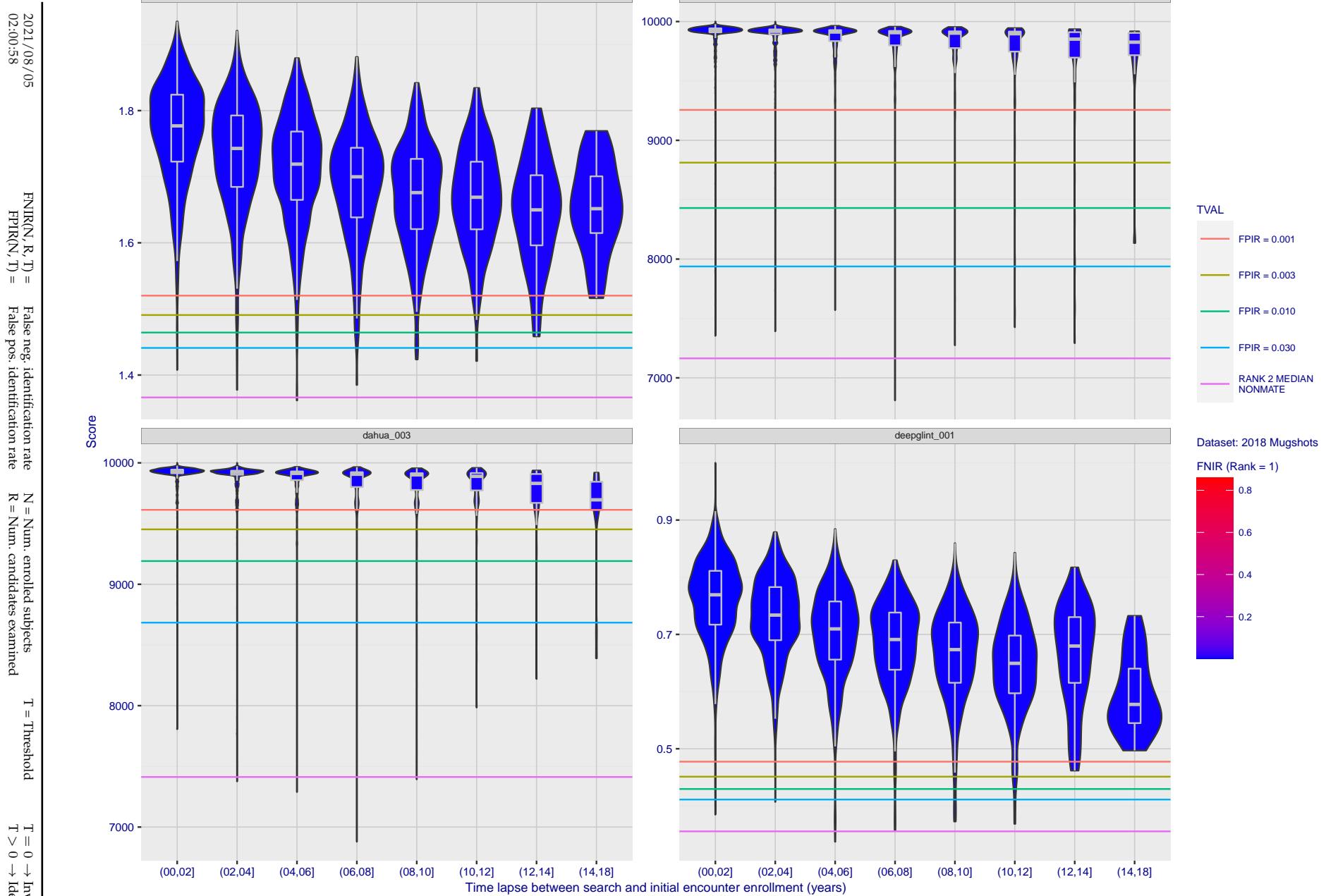


Figure 95: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

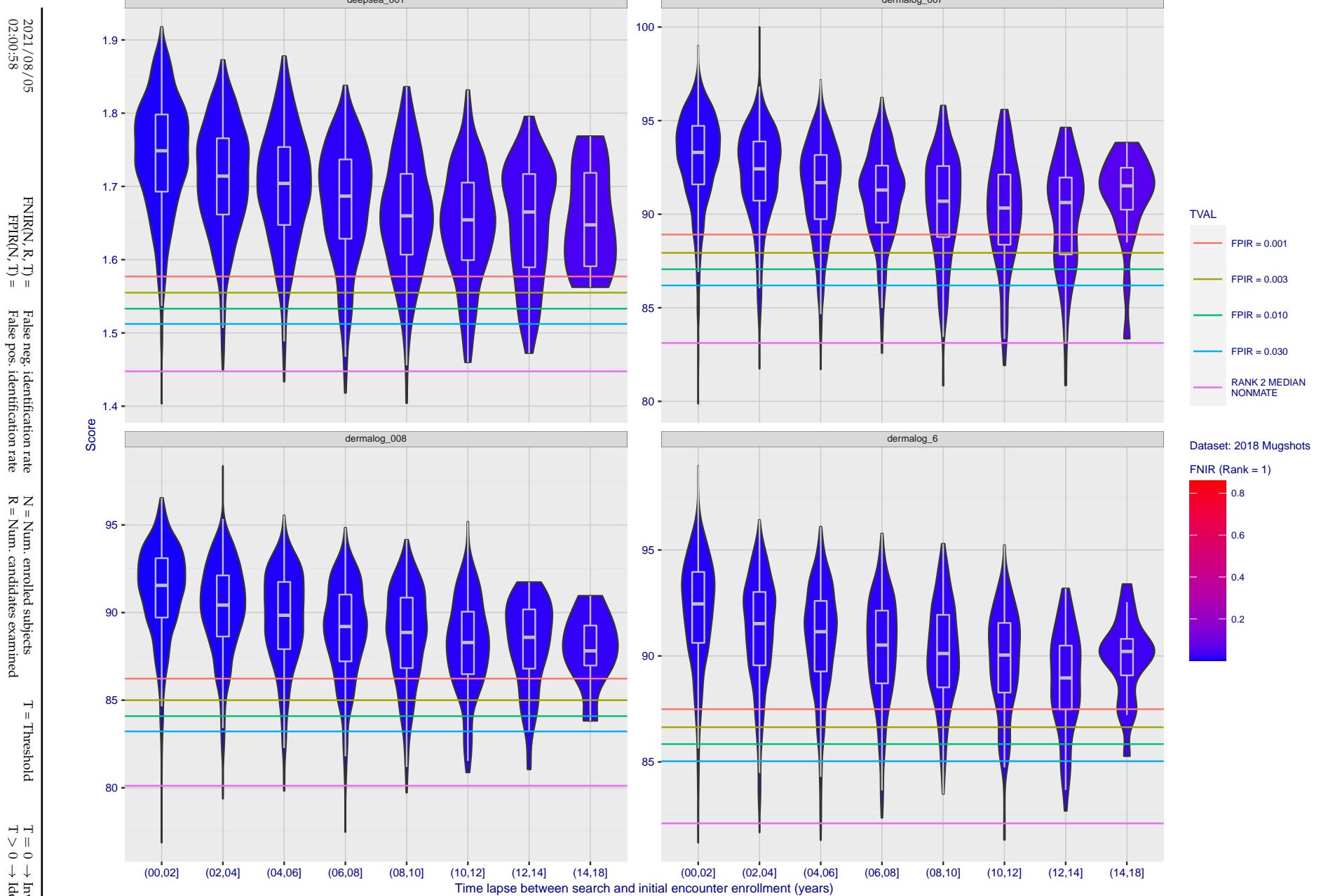


Figure 96: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

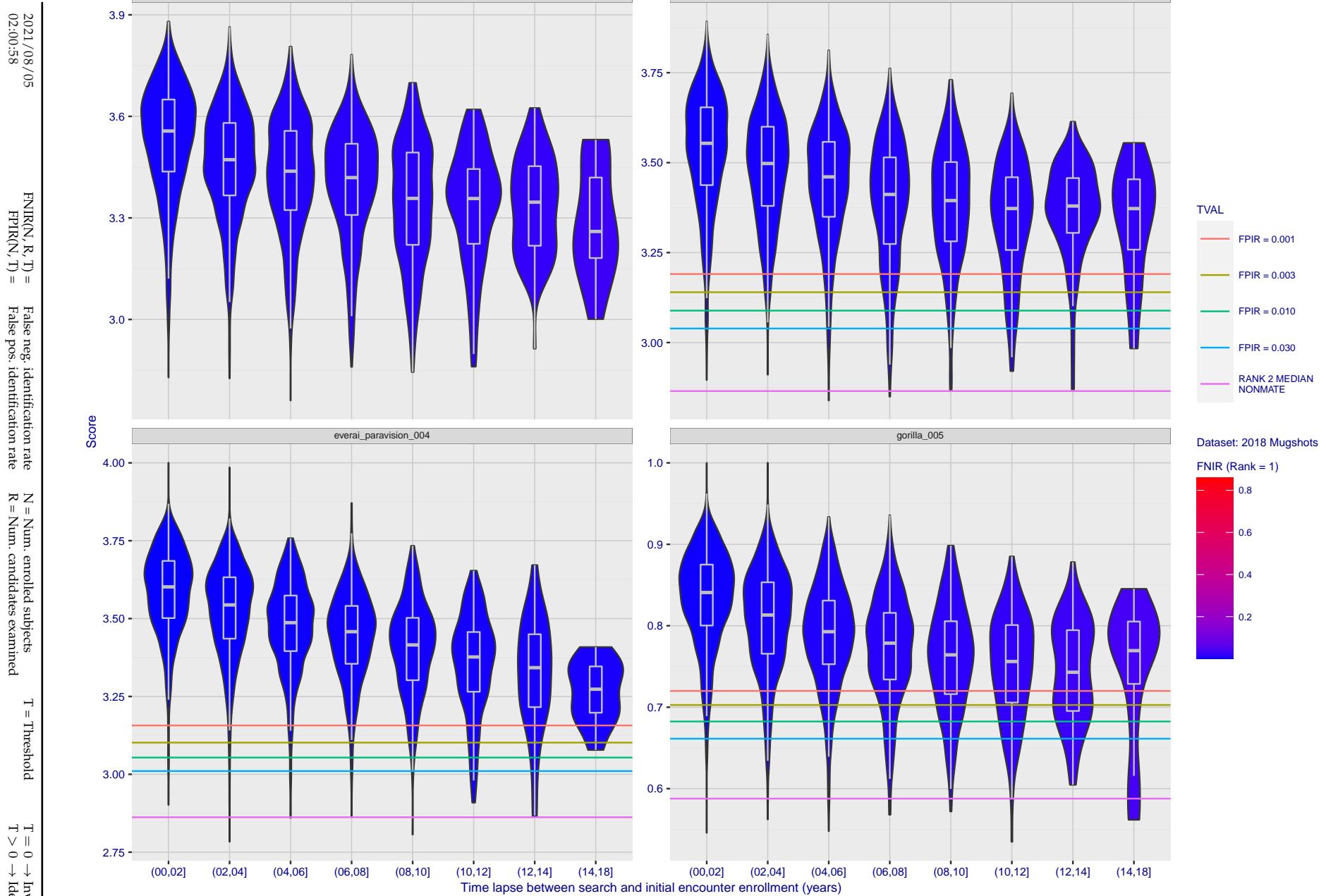


Figure 97: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

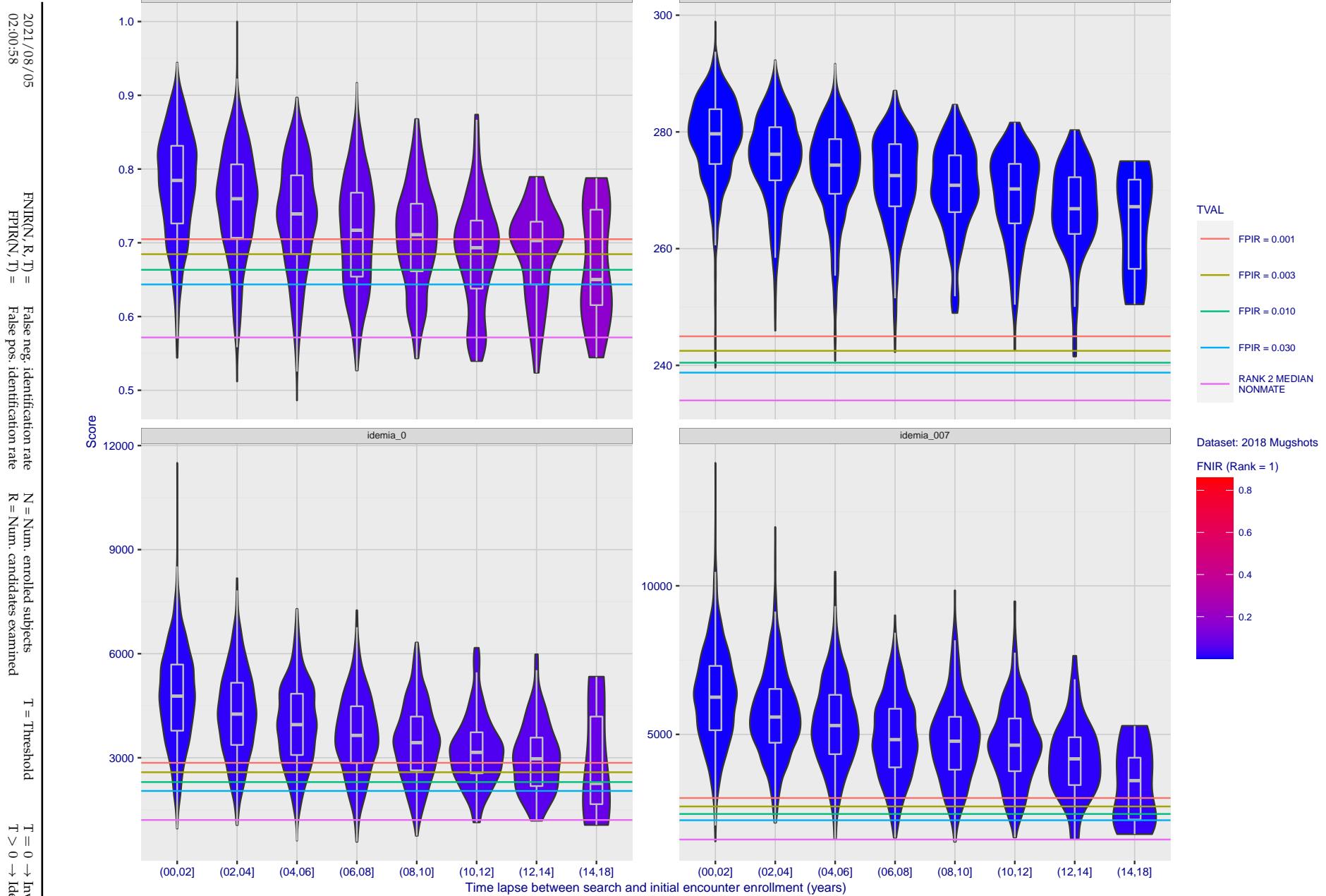


Figure 98: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

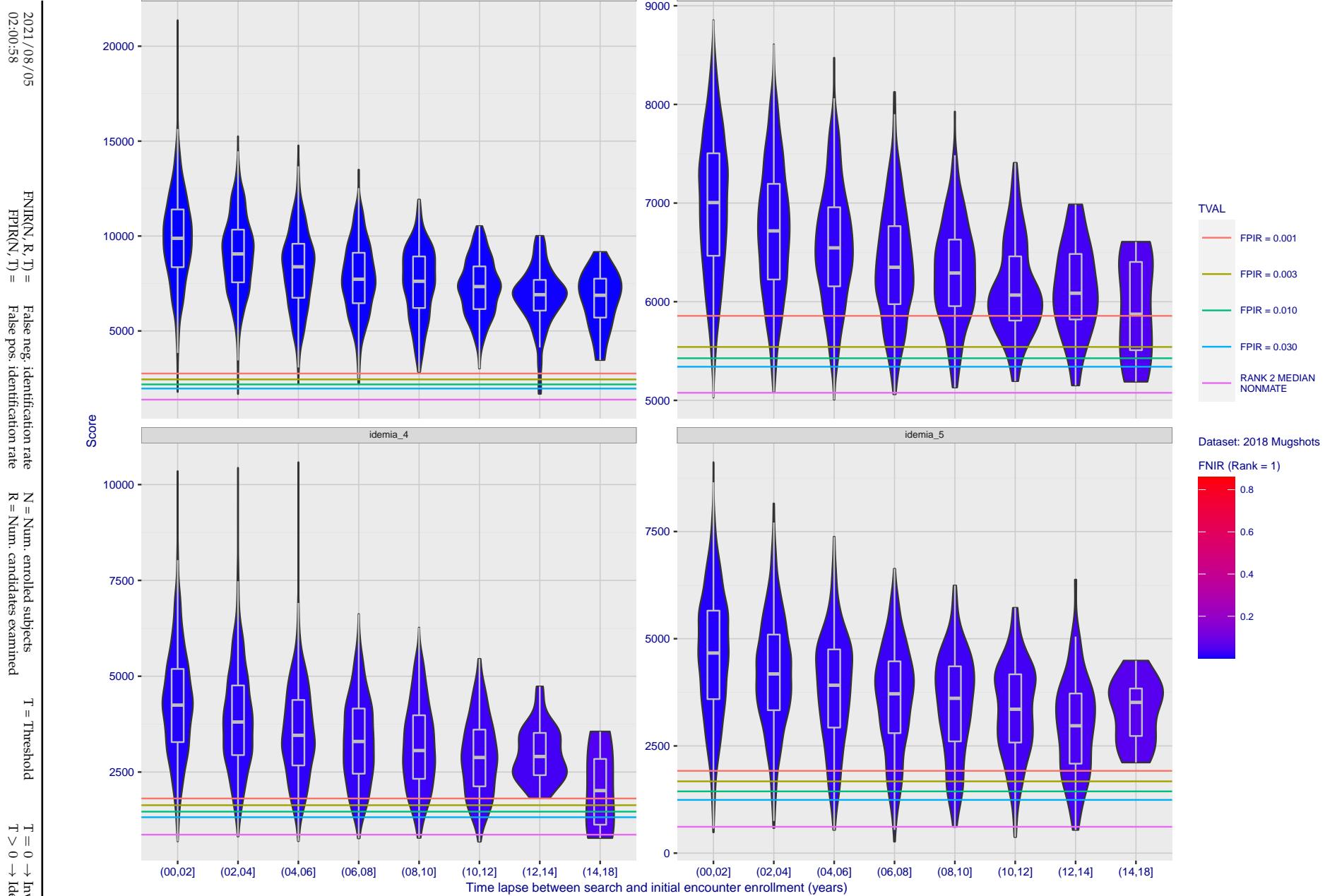


Figure 99: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

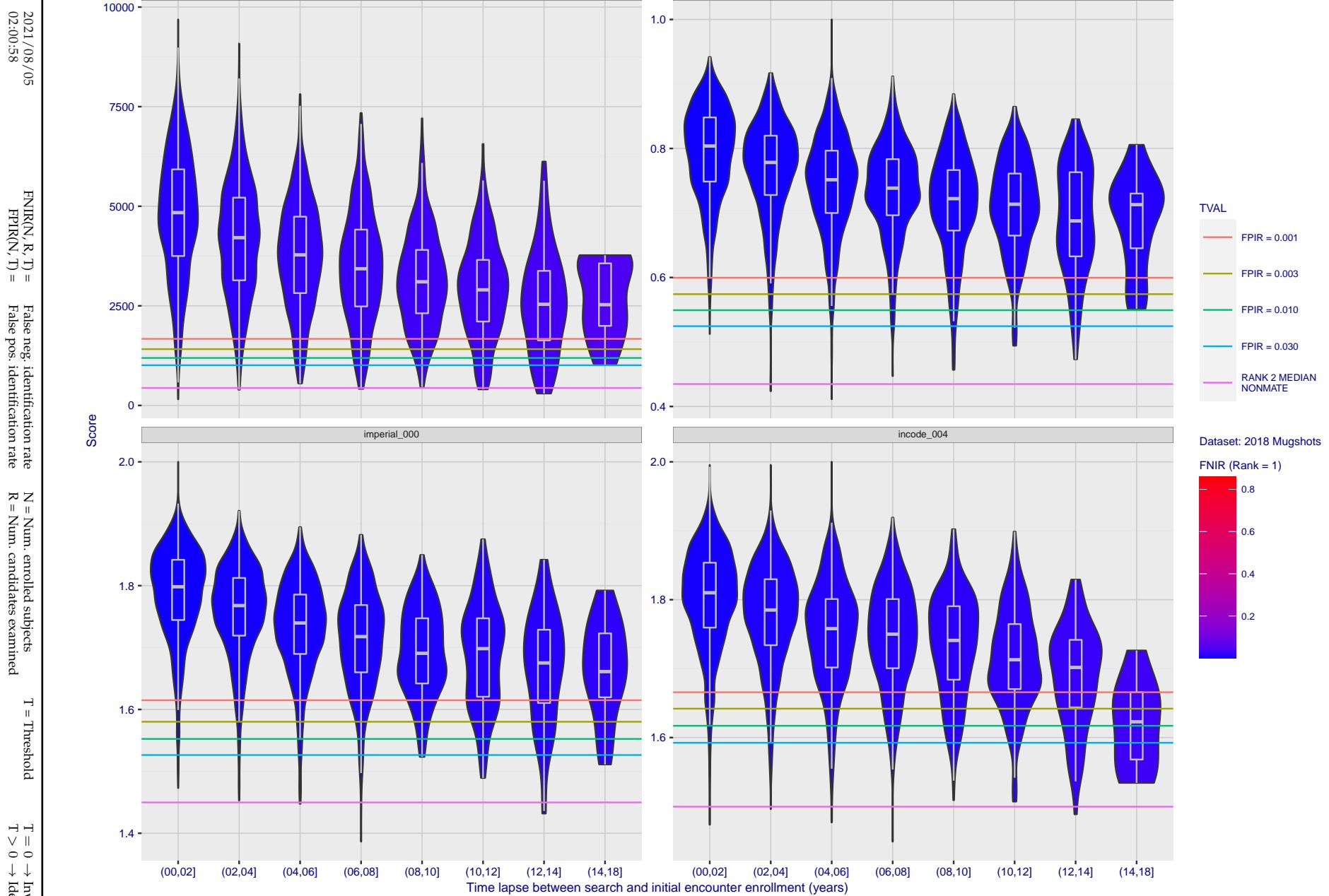


Figure 100: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

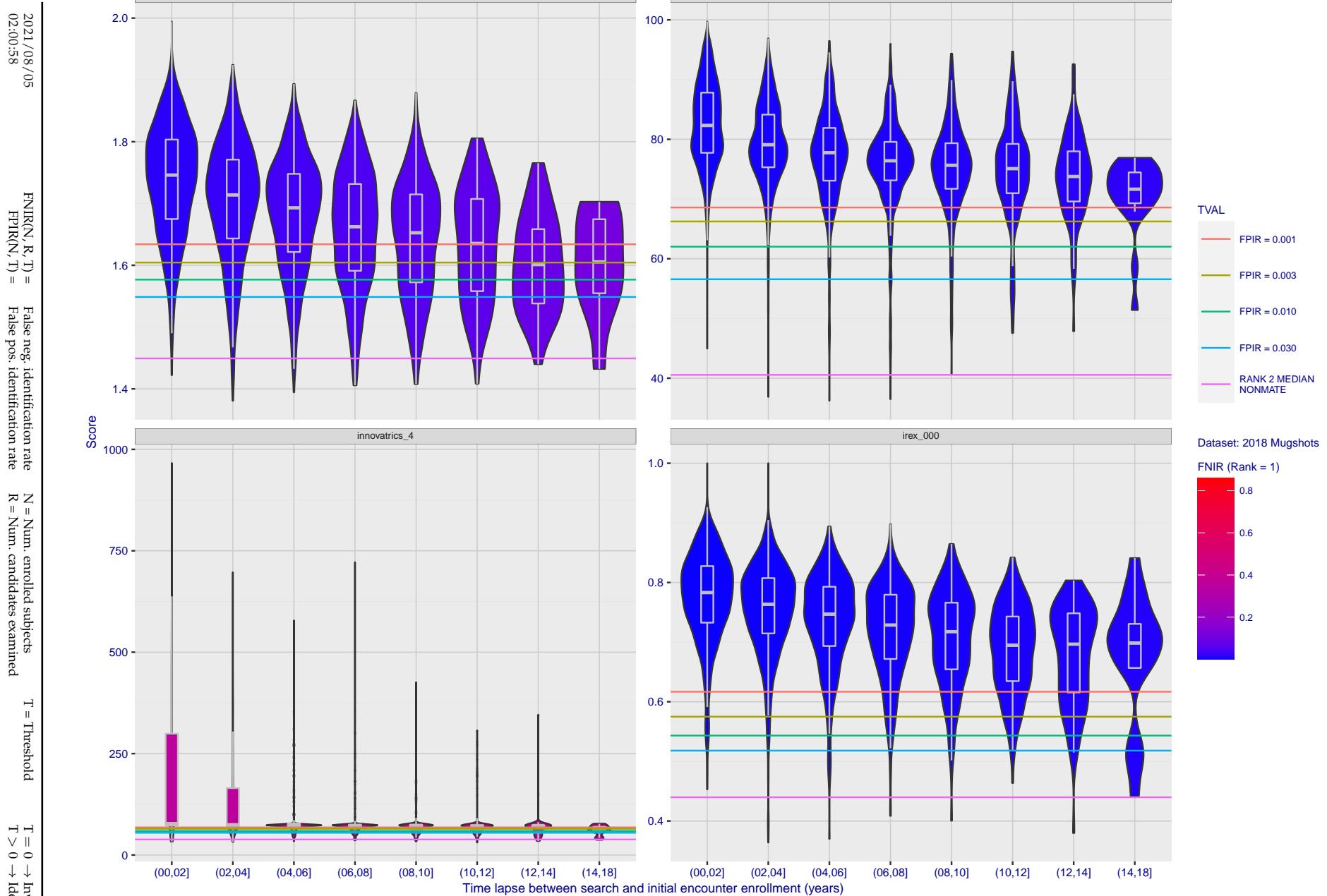


Figure 101: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

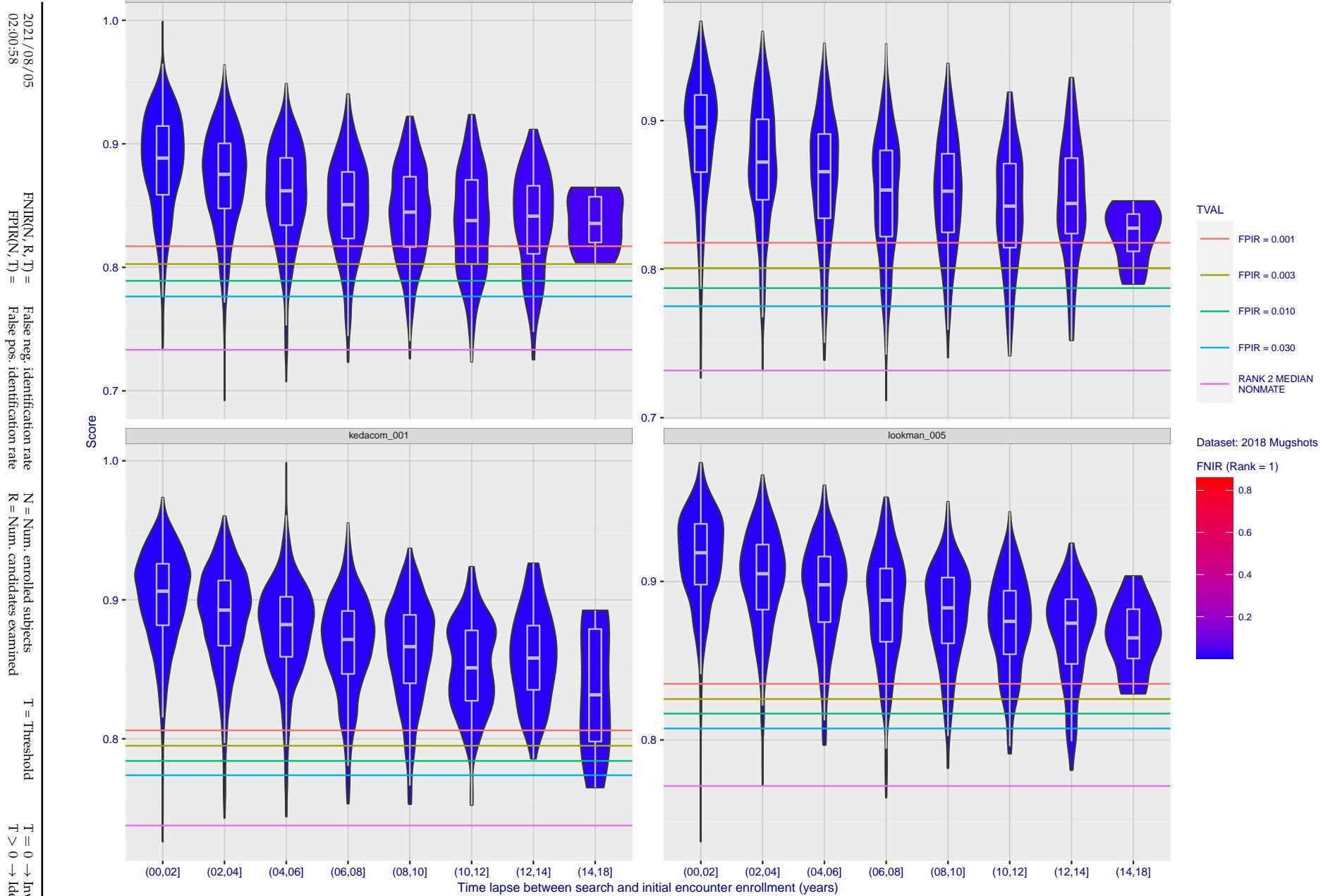


Figure 102: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

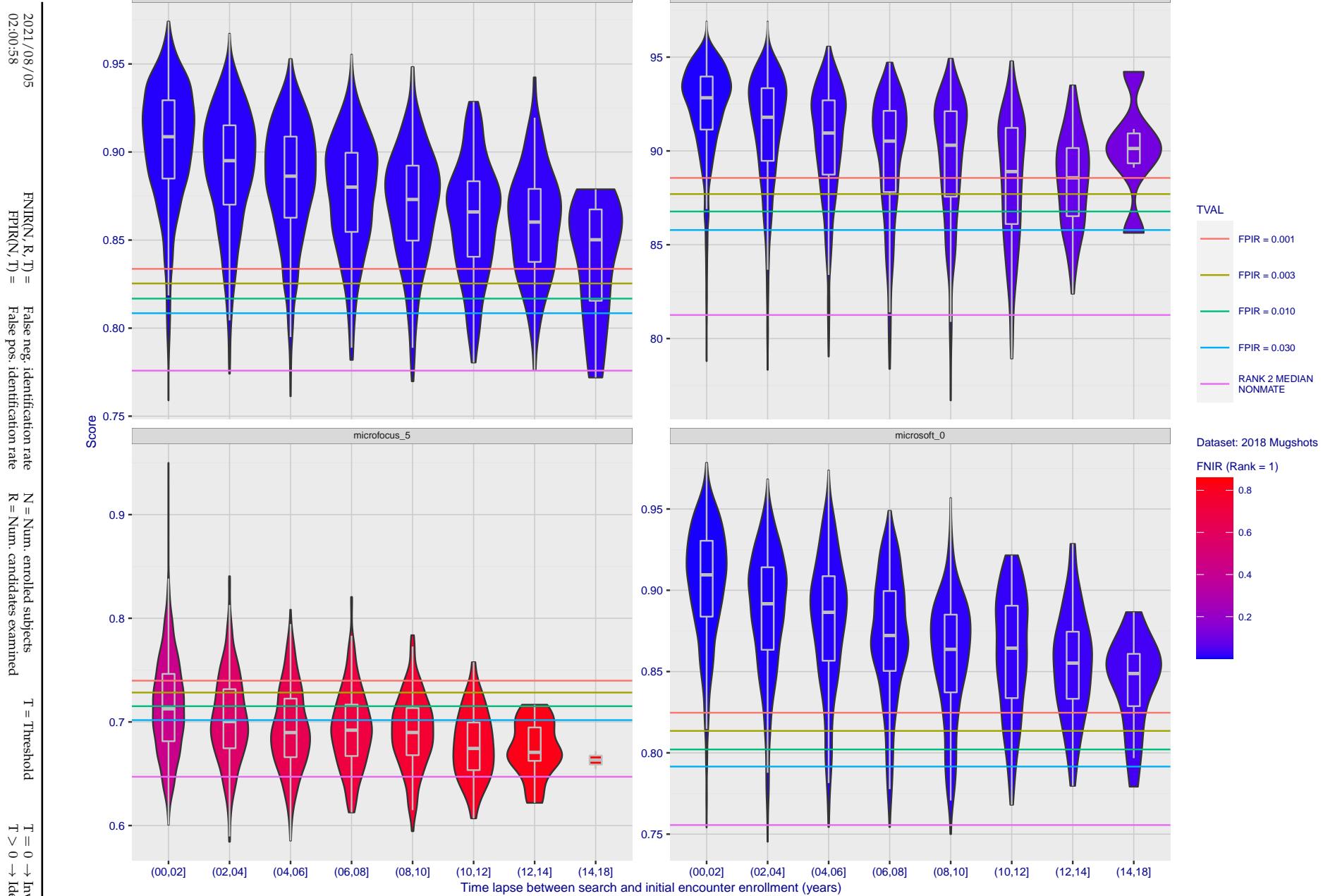


Figure 103: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

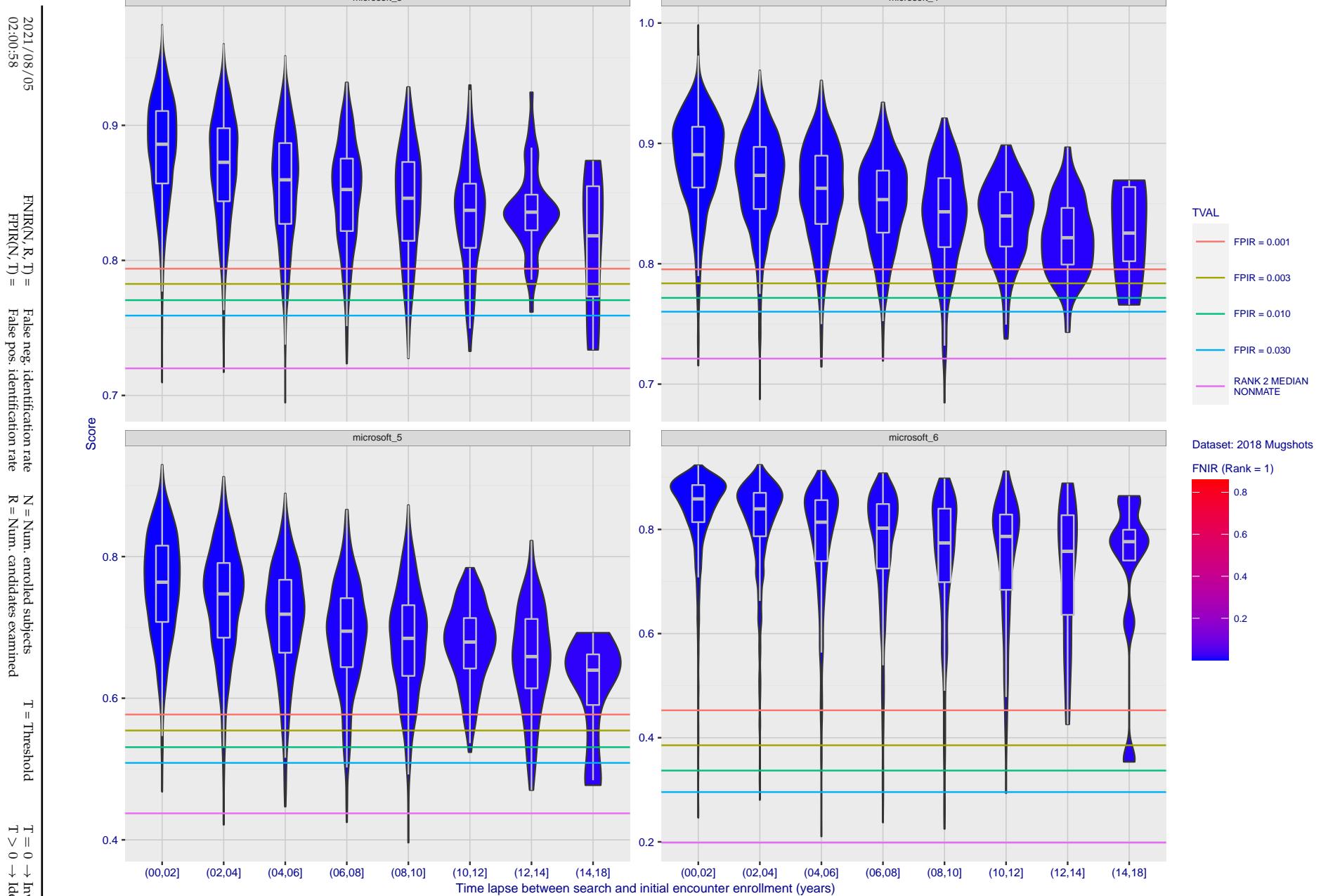


Figure 104: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

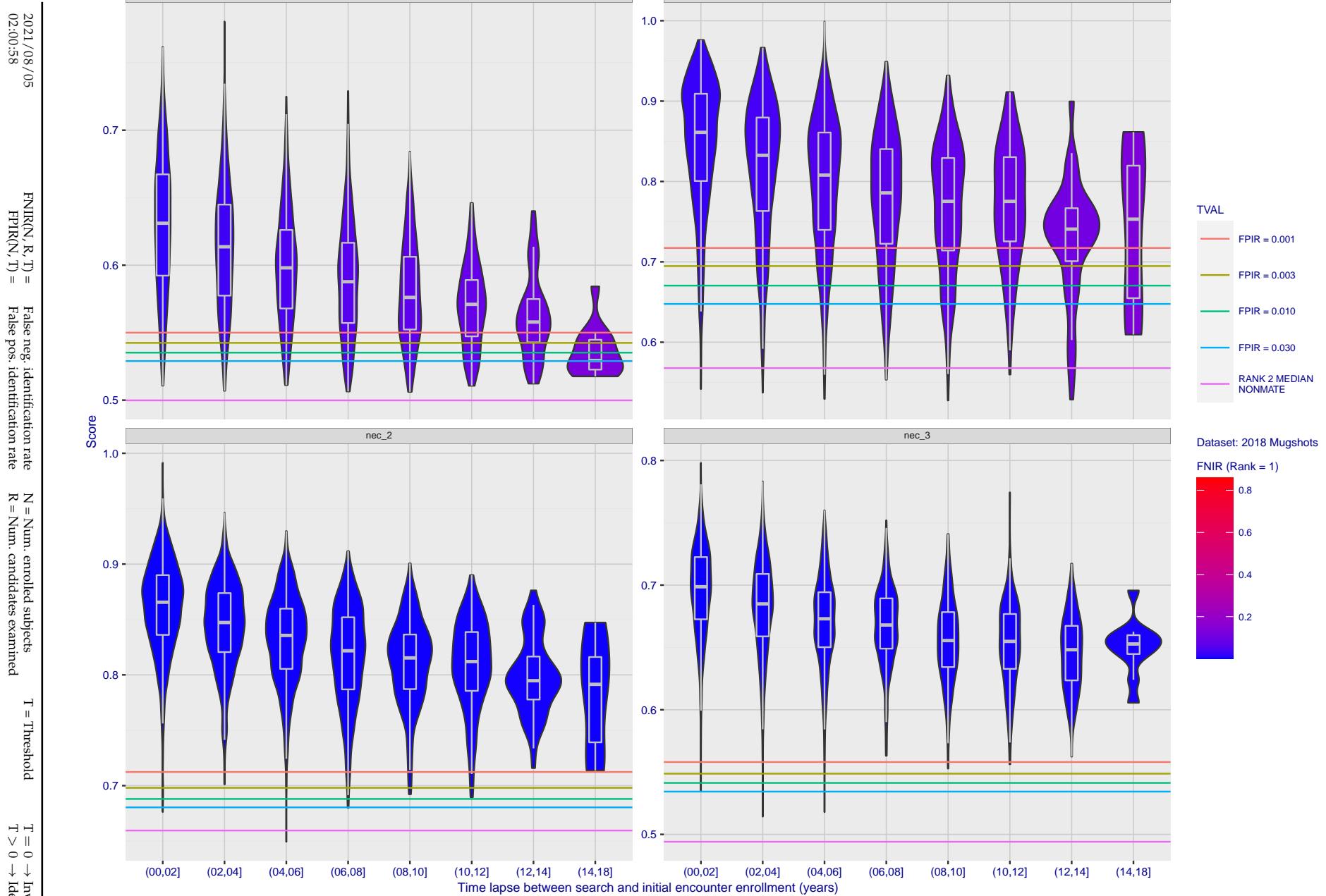


Figure 105: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

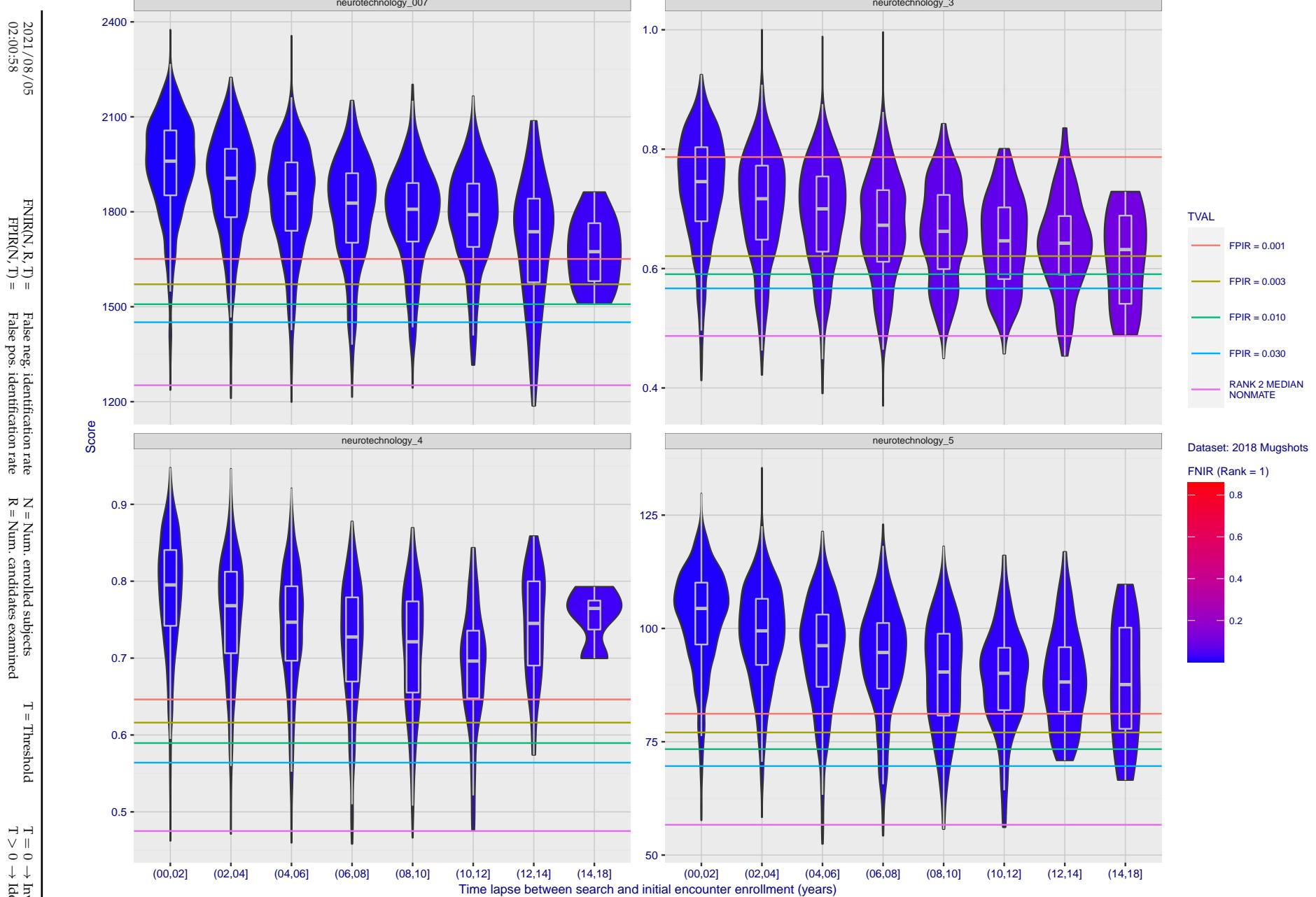


Figure 106: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

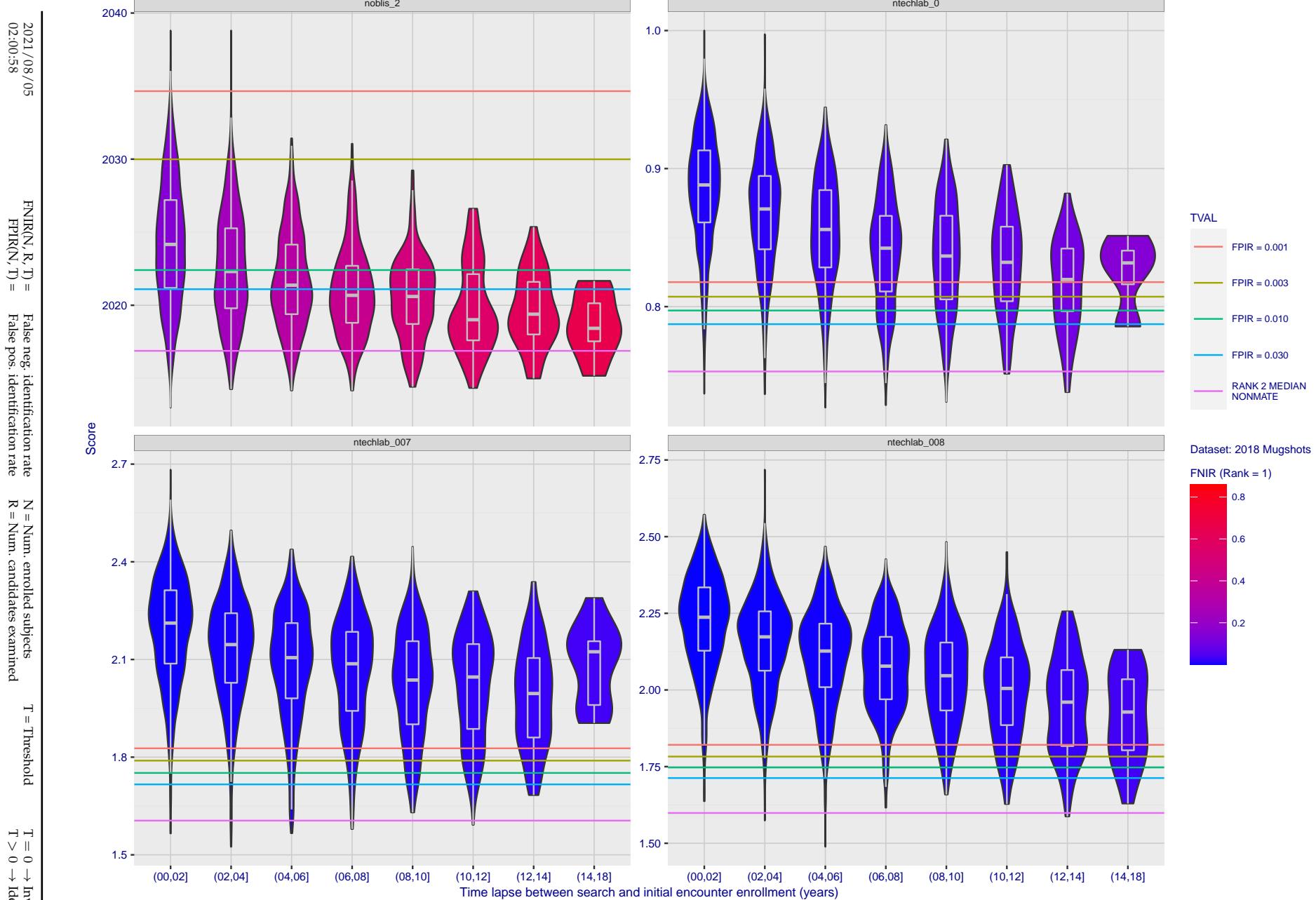


Figure 107: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

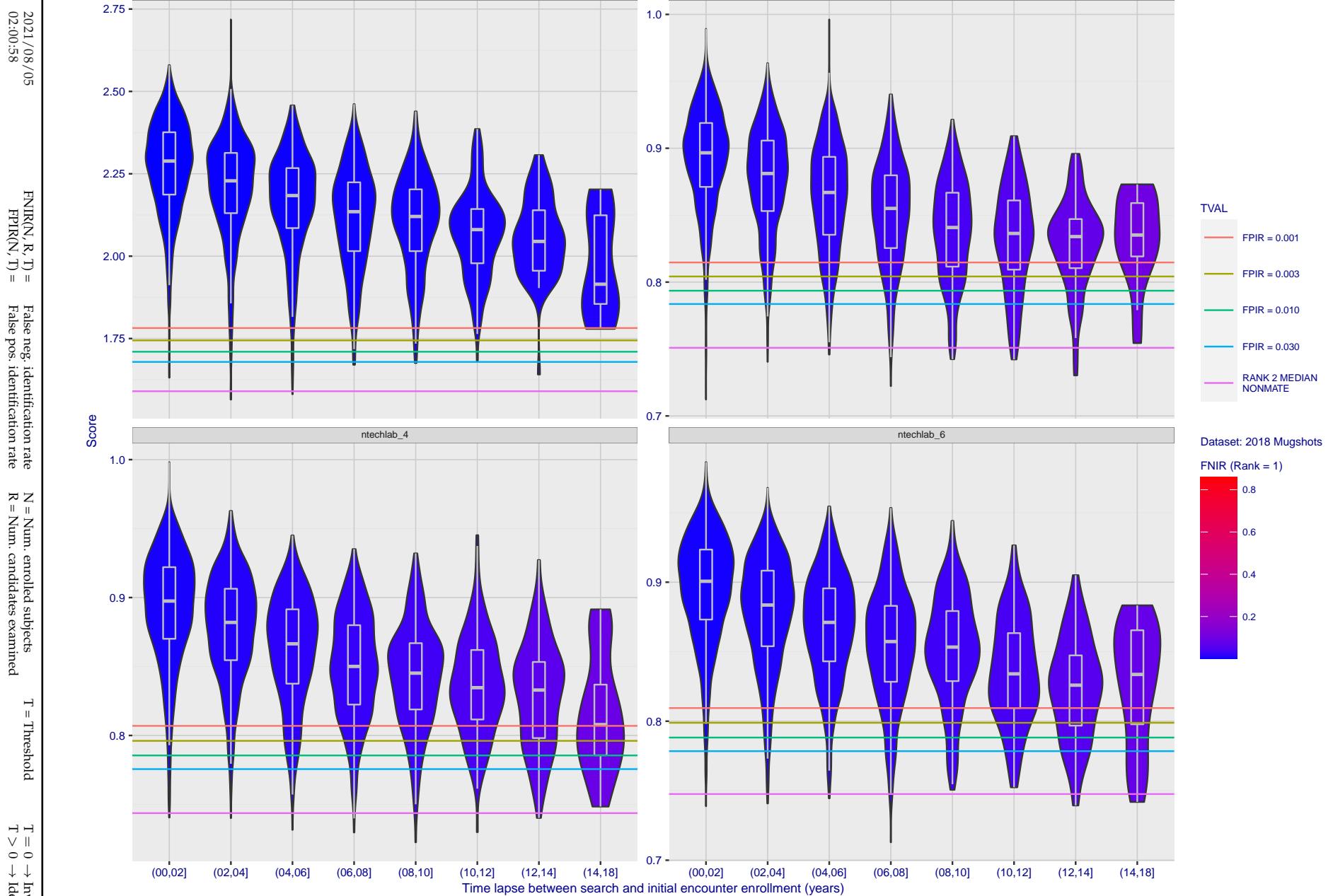


Figure 108: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

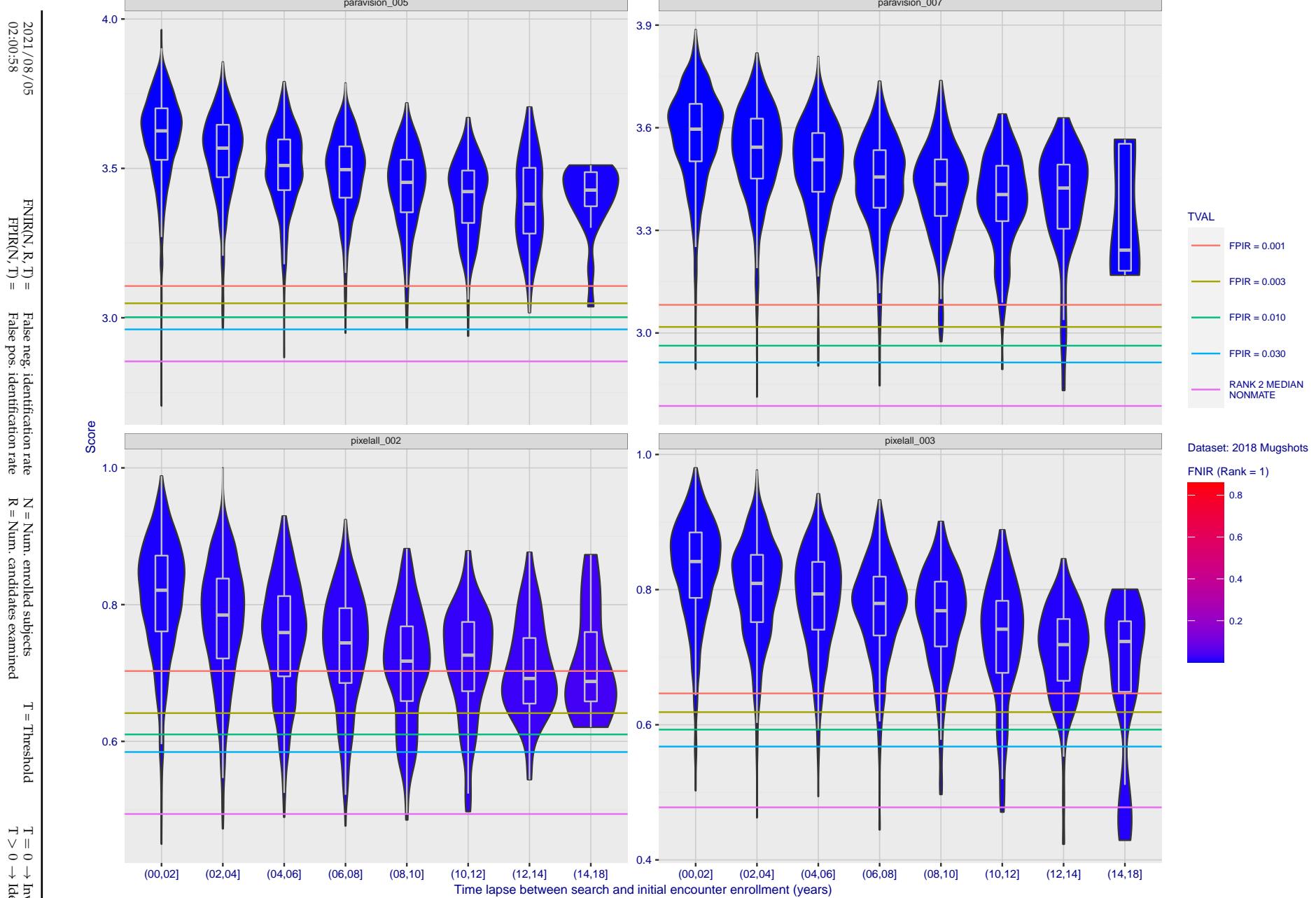


Figure 109: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

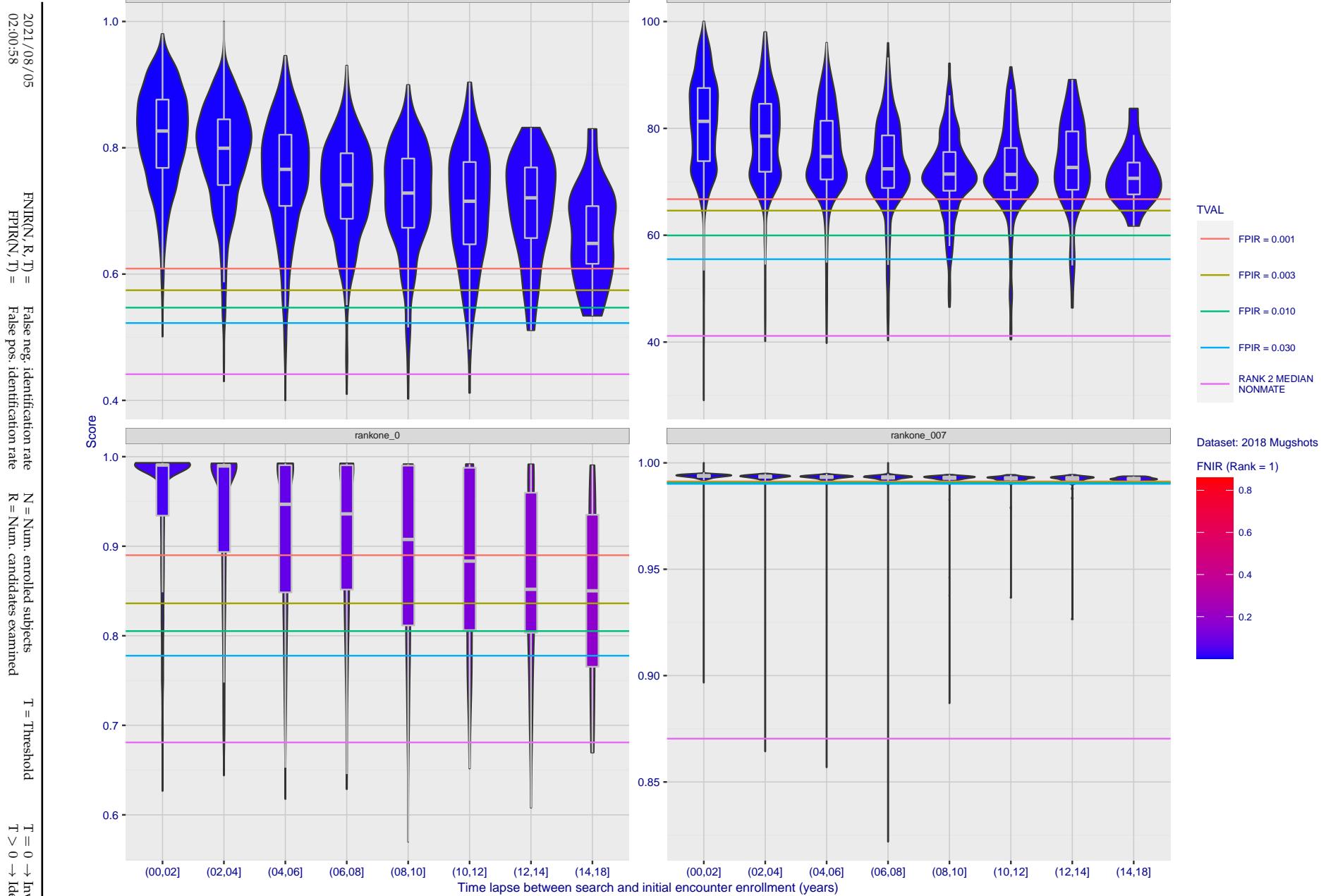


Figure 110: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

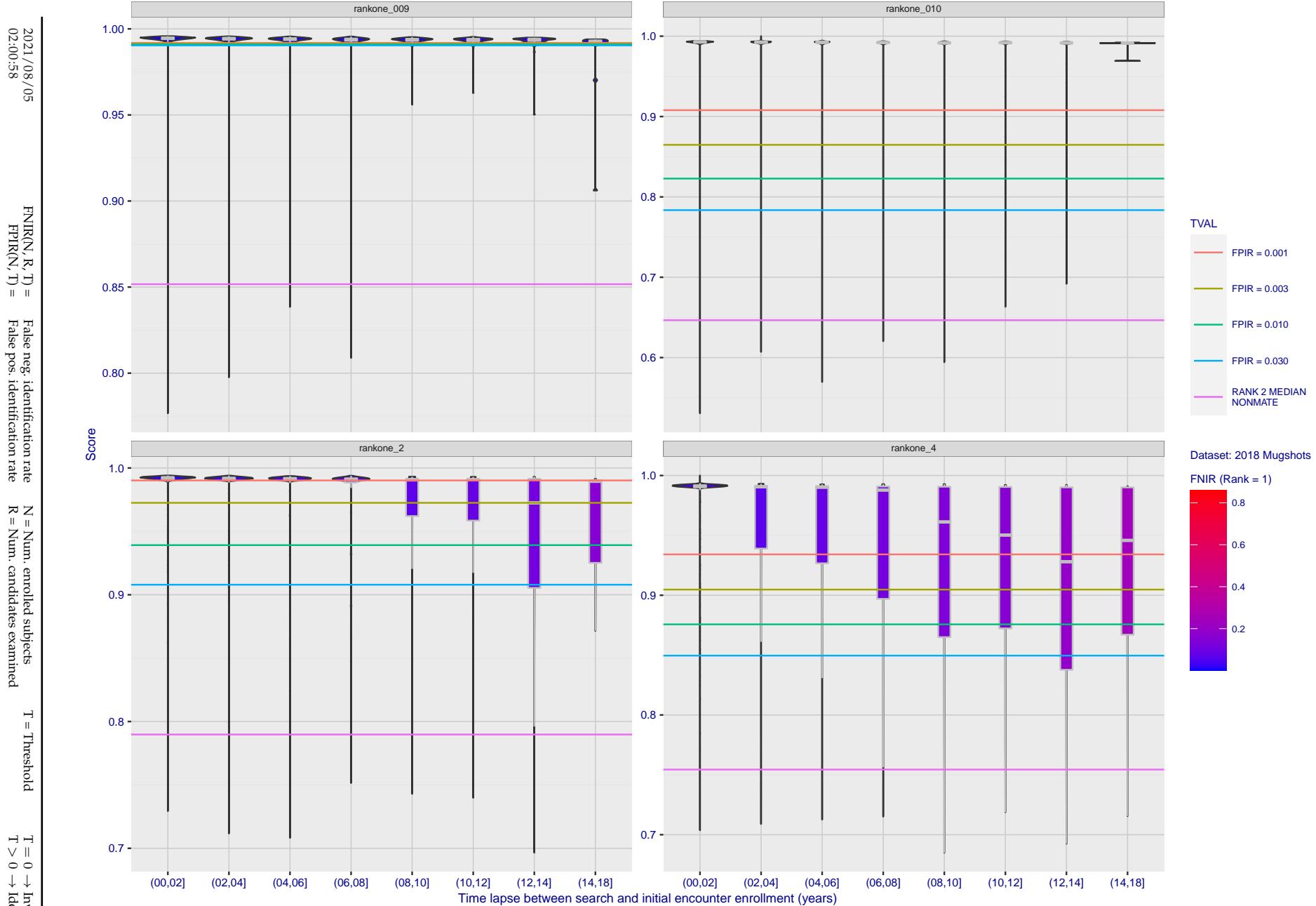


Figure 111: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

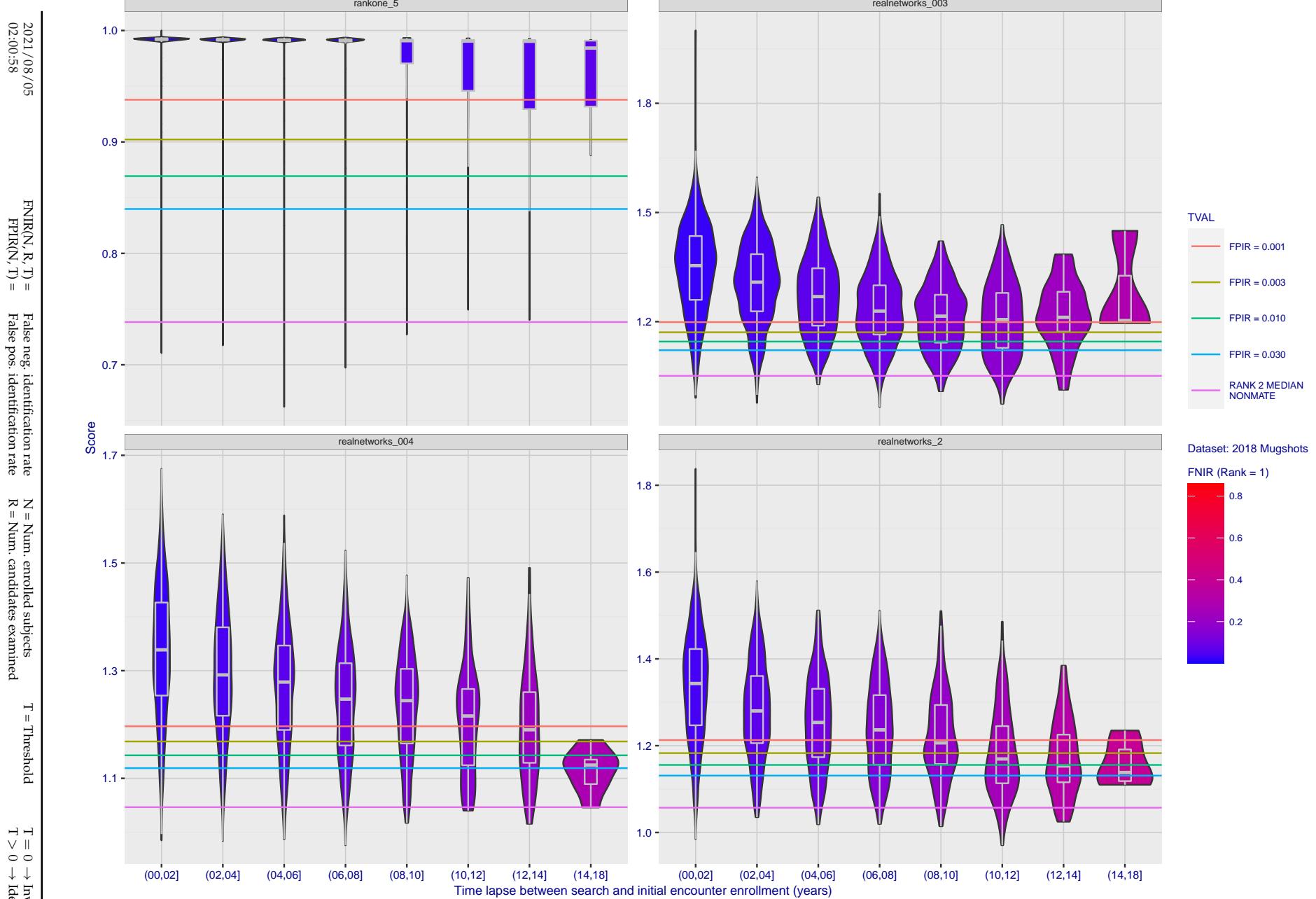


Figure 112: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

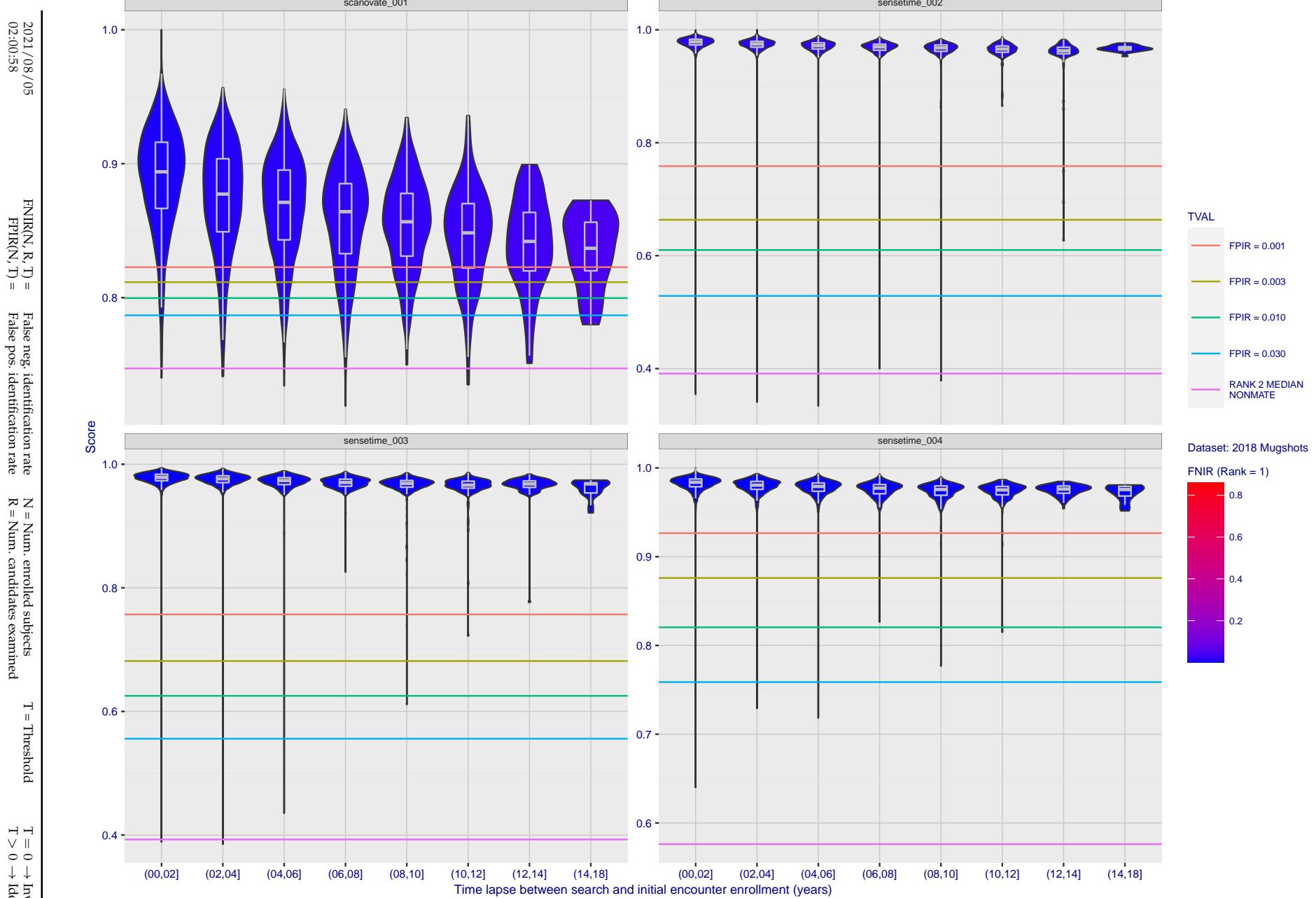


Figure 113: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

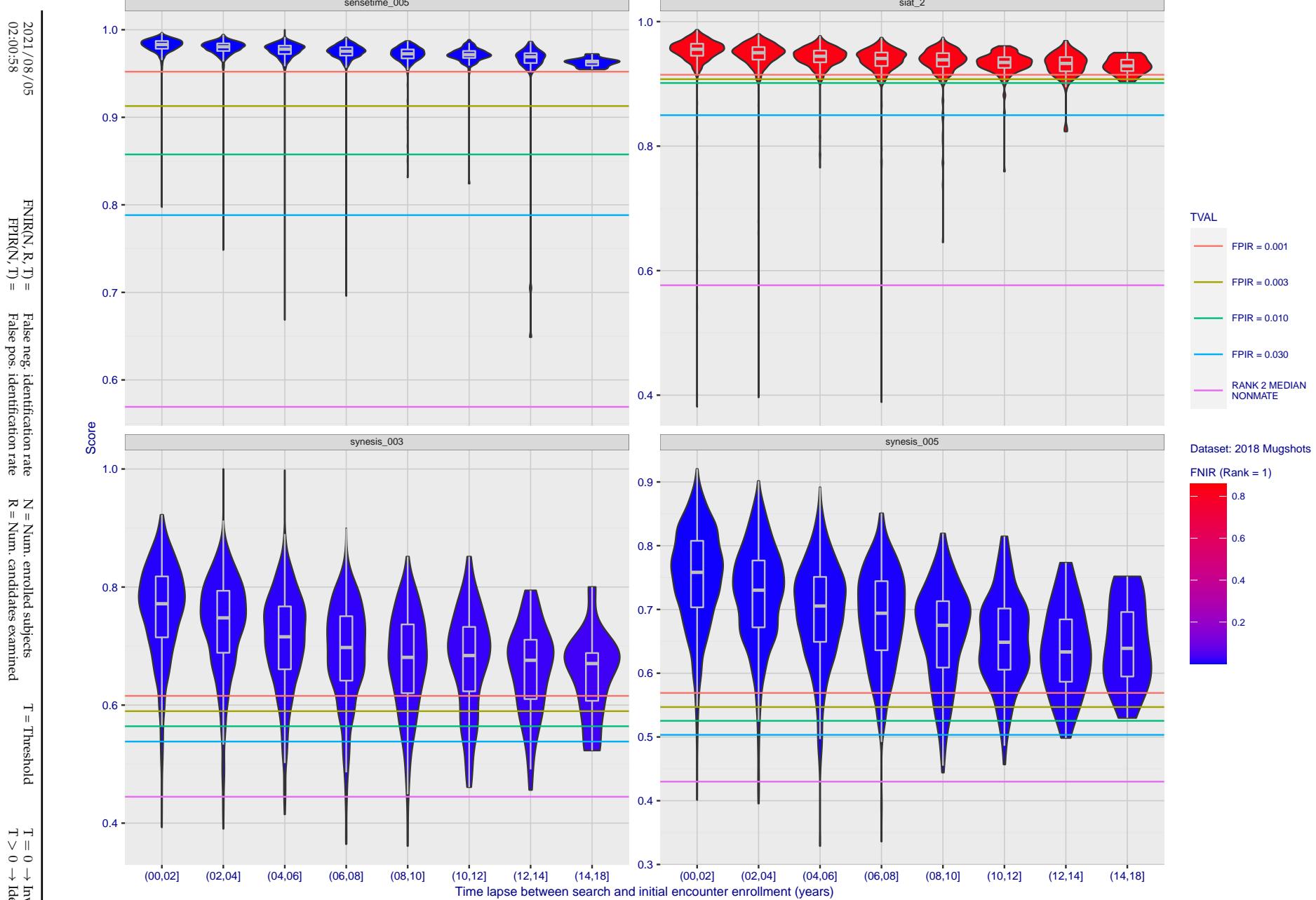


Figure 114: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

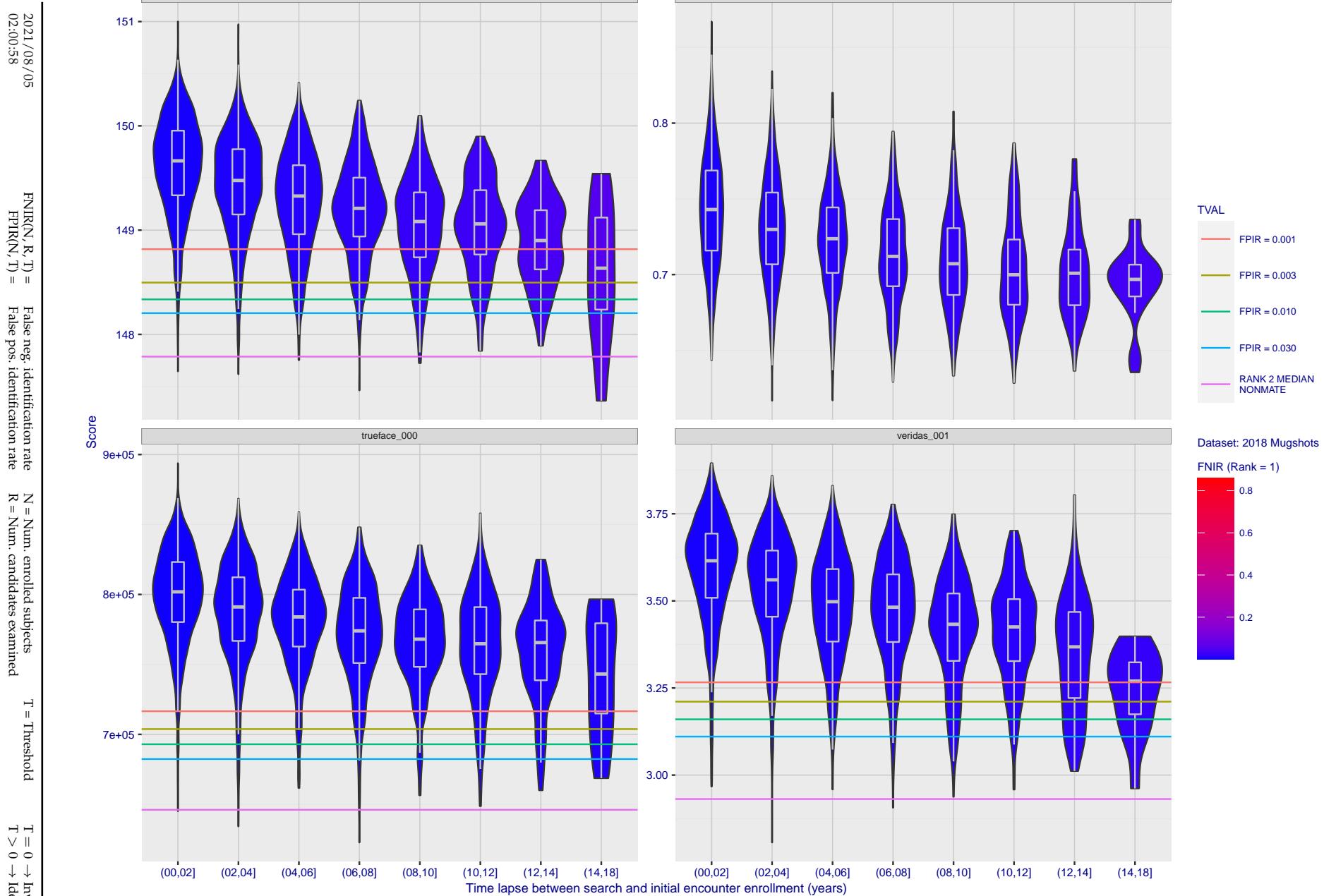


Figure 115: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

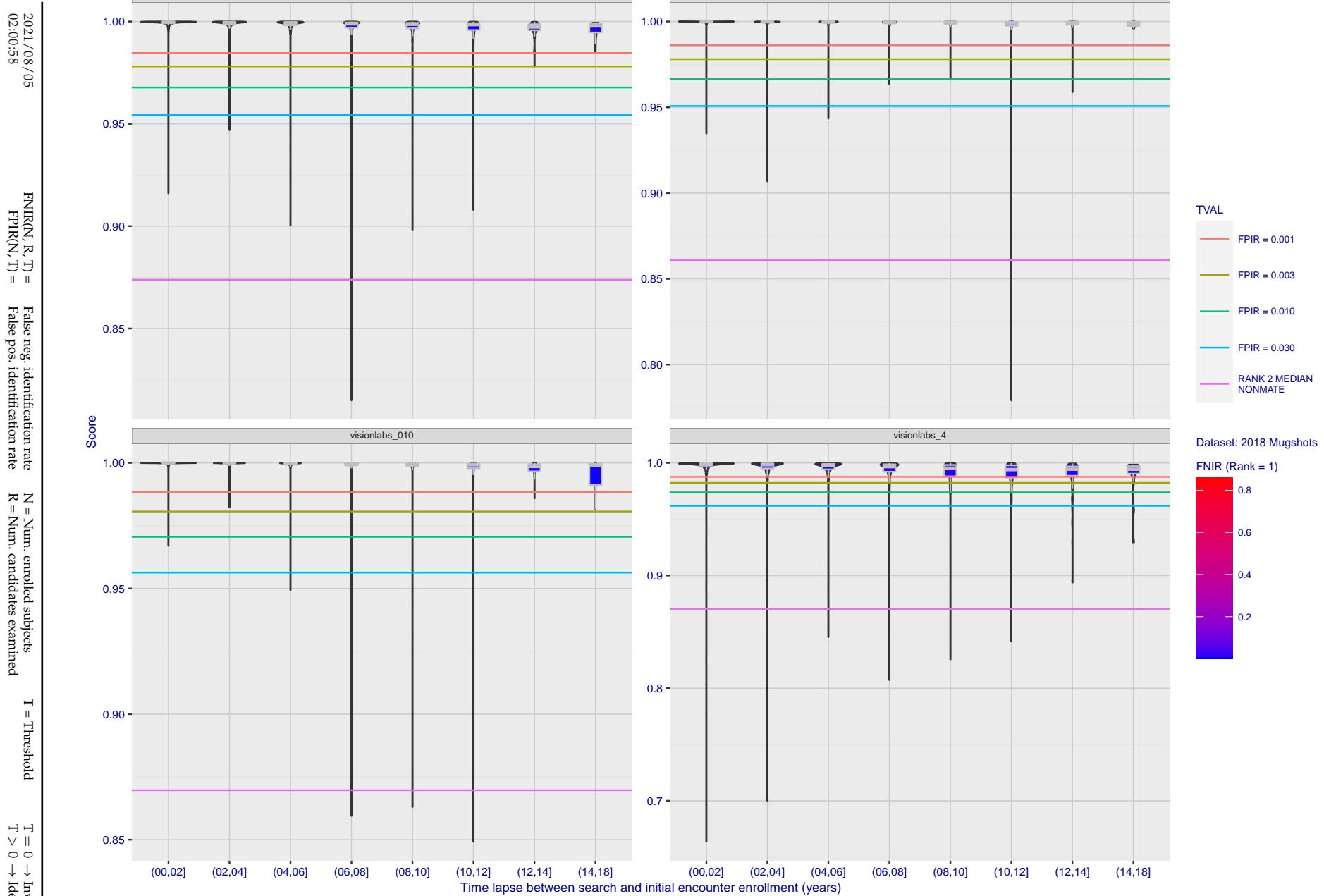


Figure 116: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

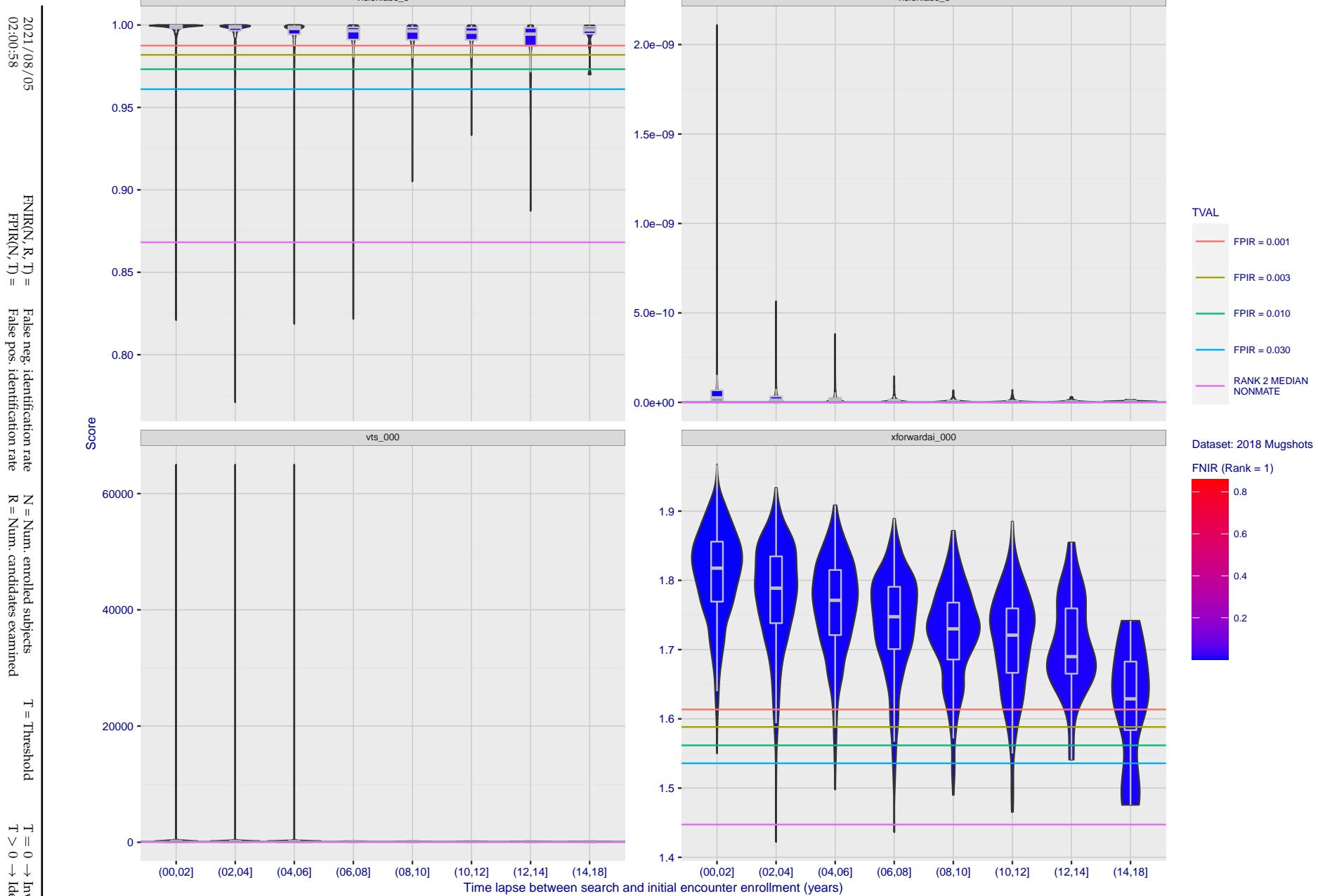


Figure 117: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

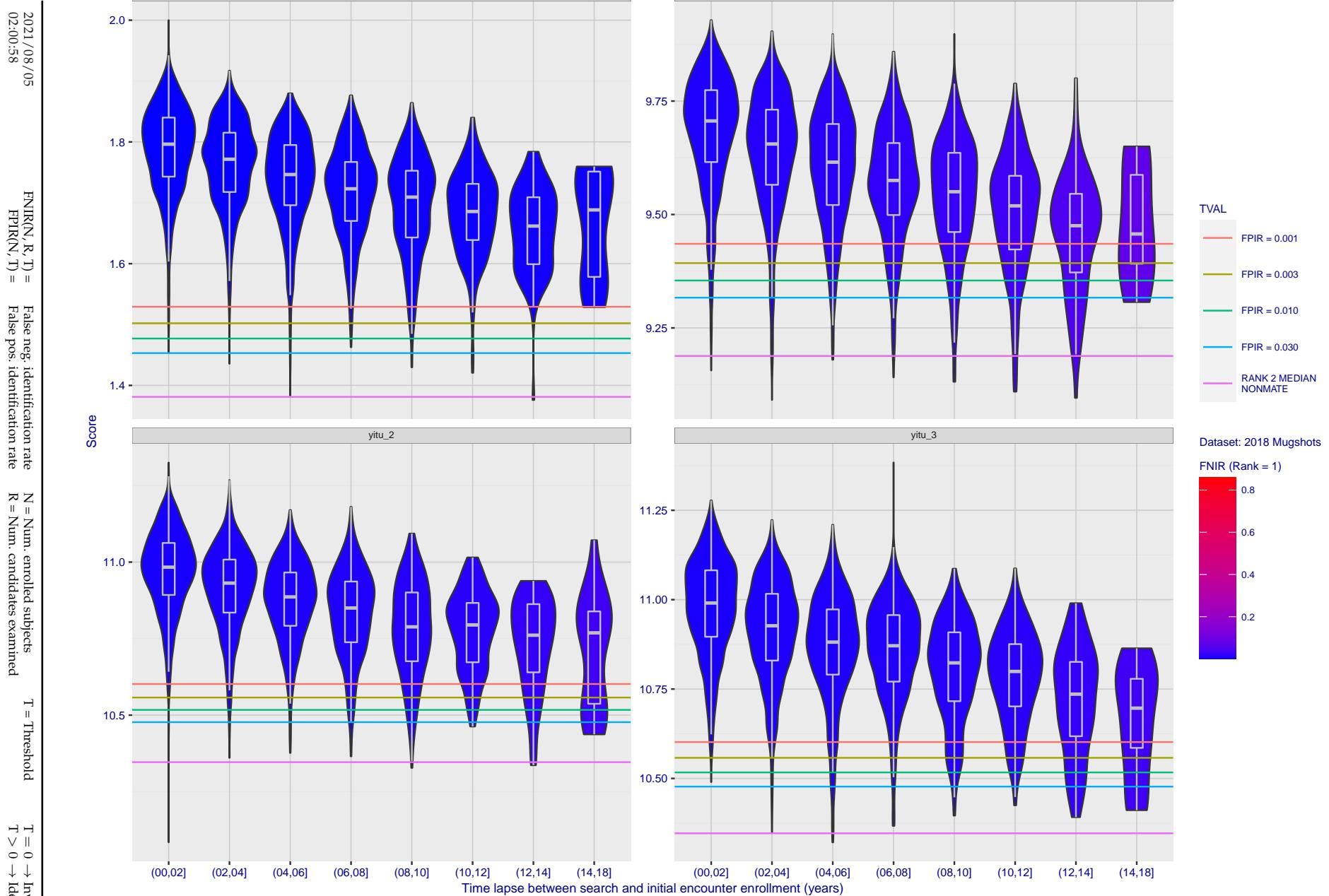


Figure 118: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

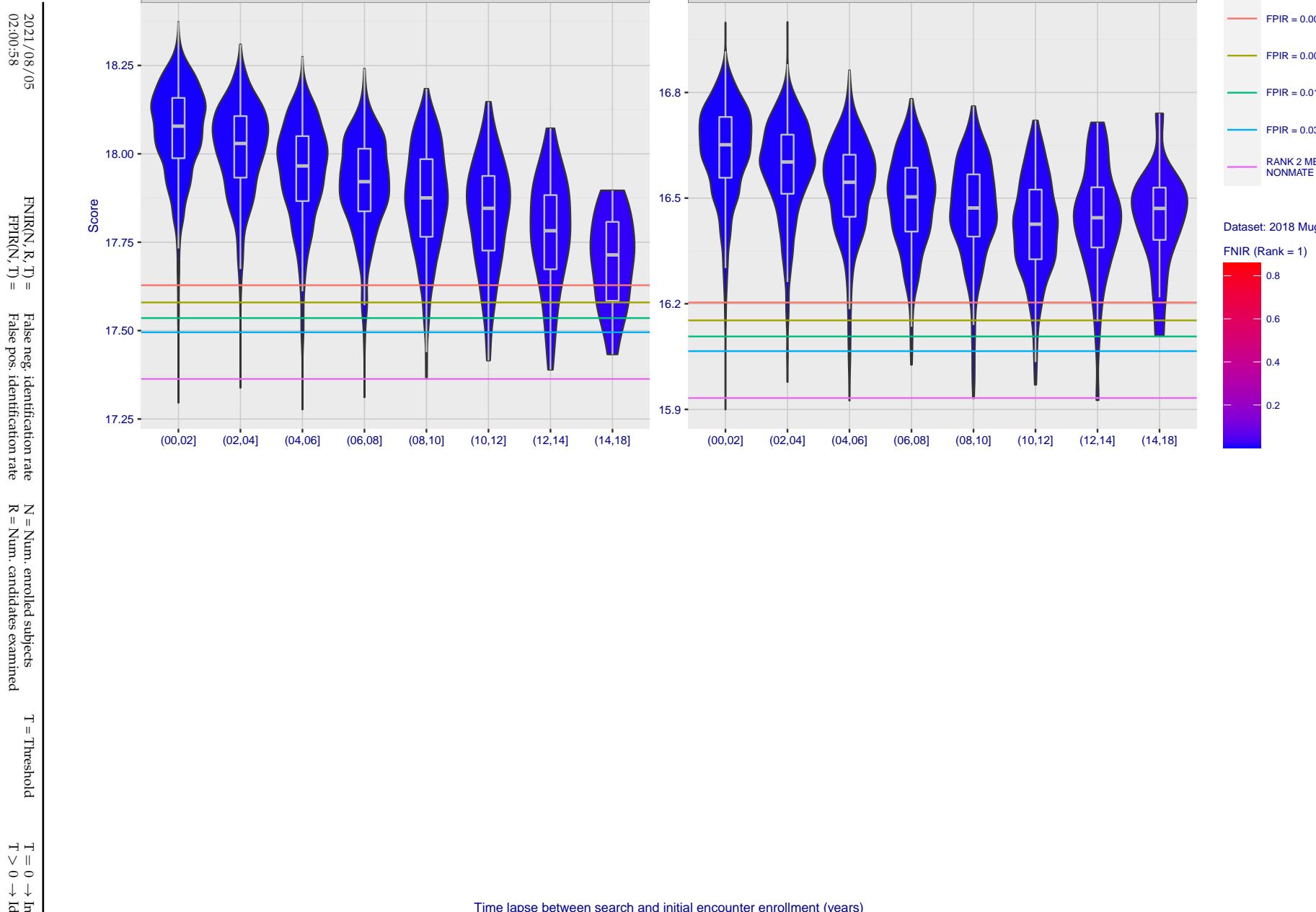


Figure 119: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

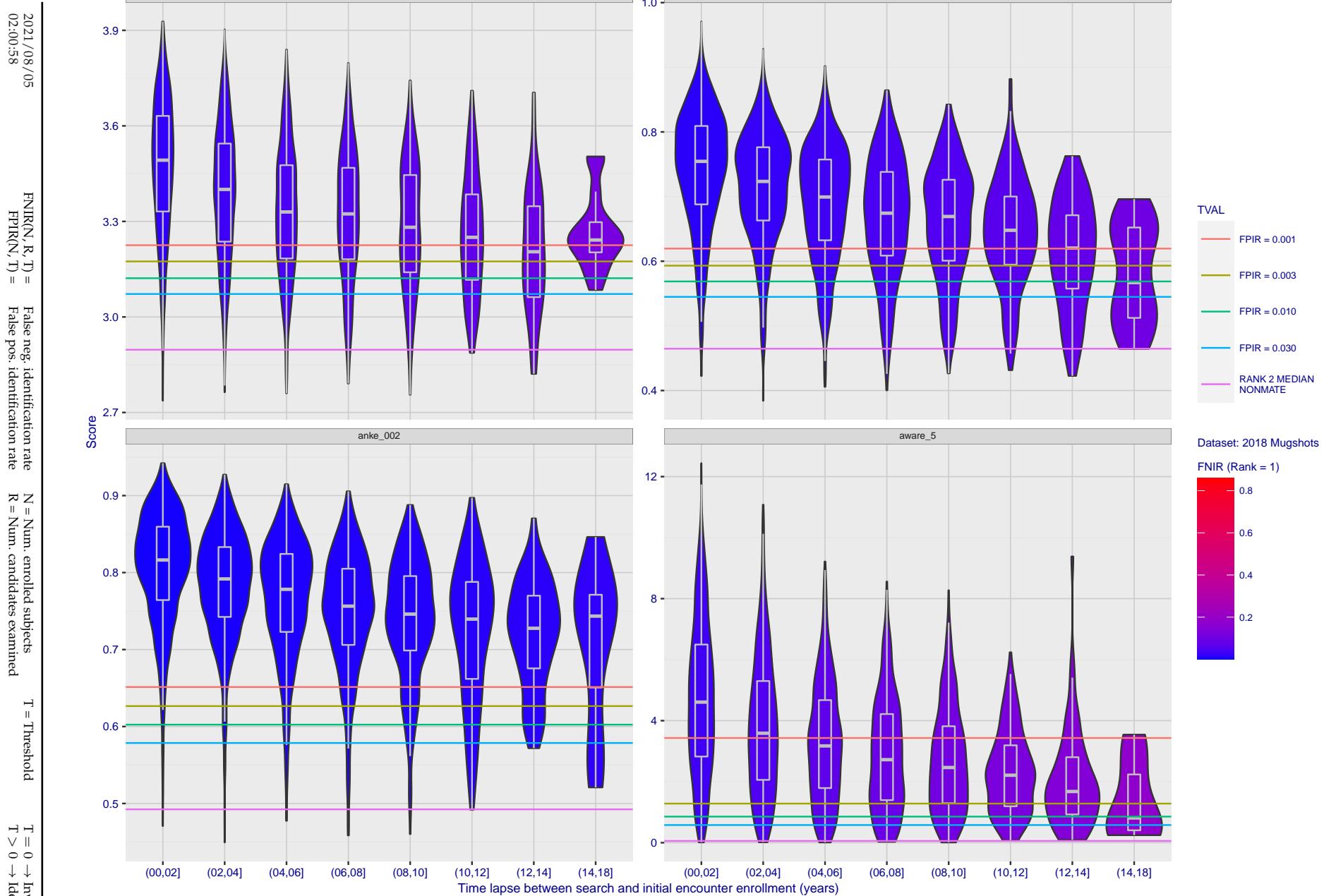


Figure 120: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

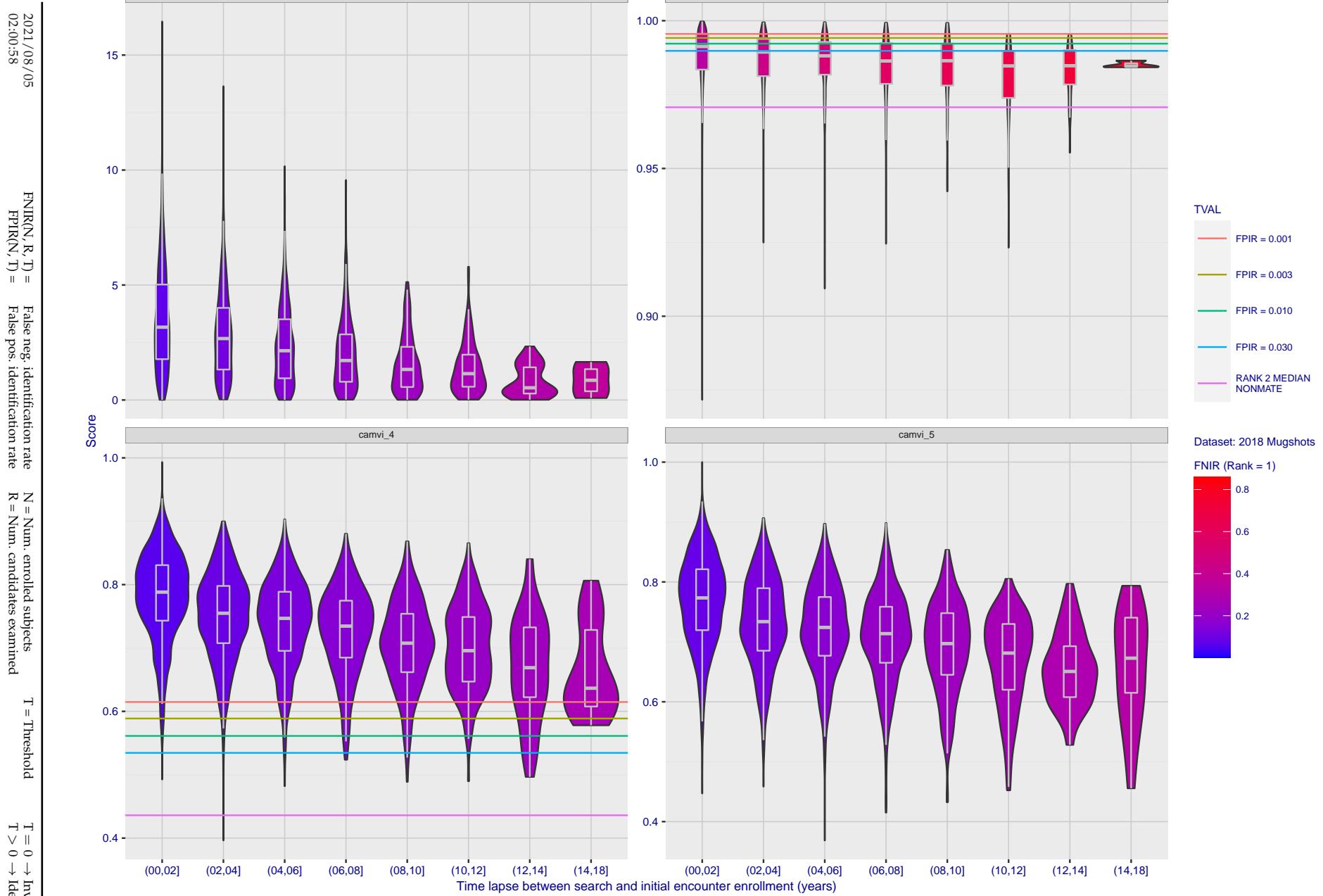


Figure 121: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

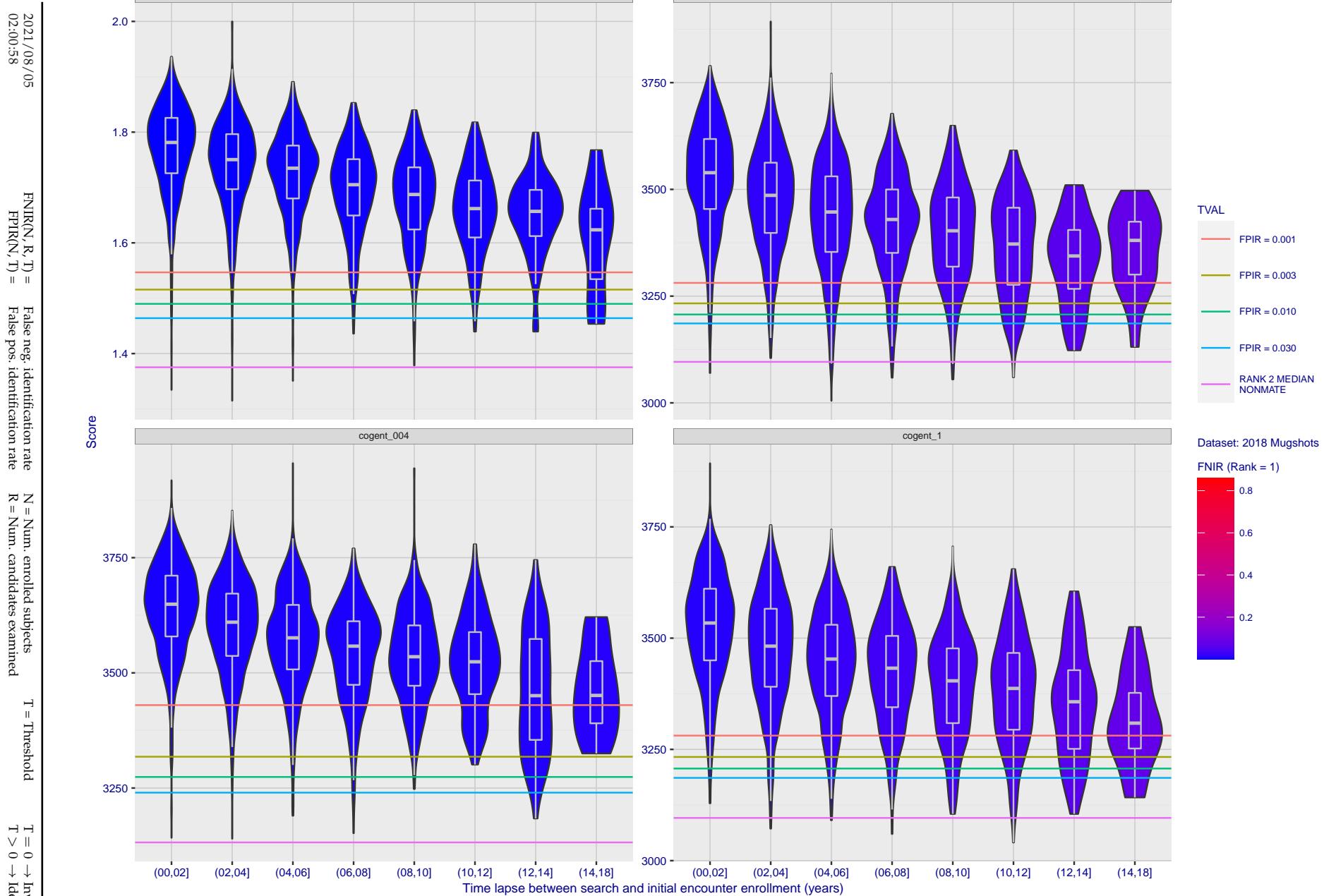


Figure 122: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

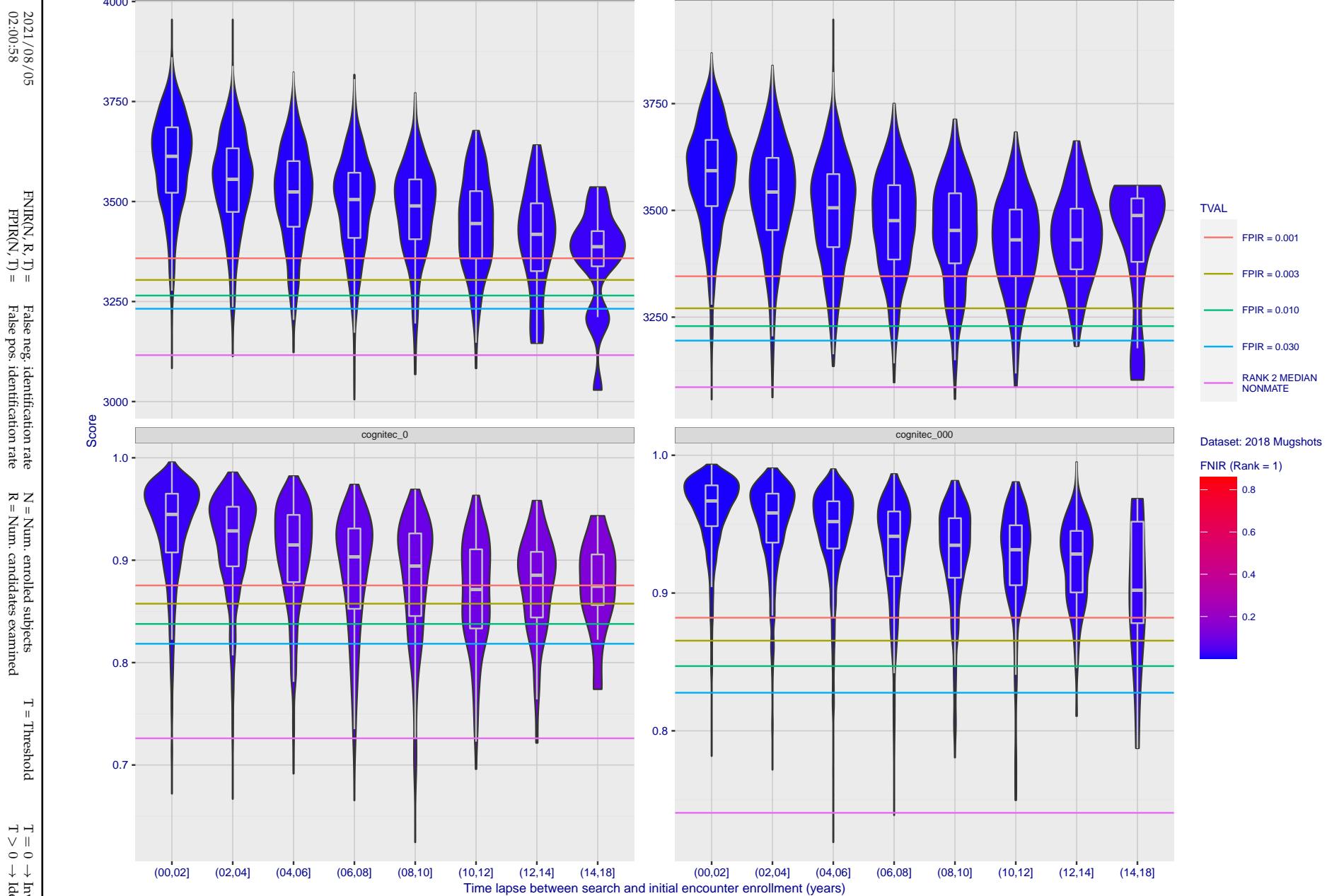


Figure 123: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

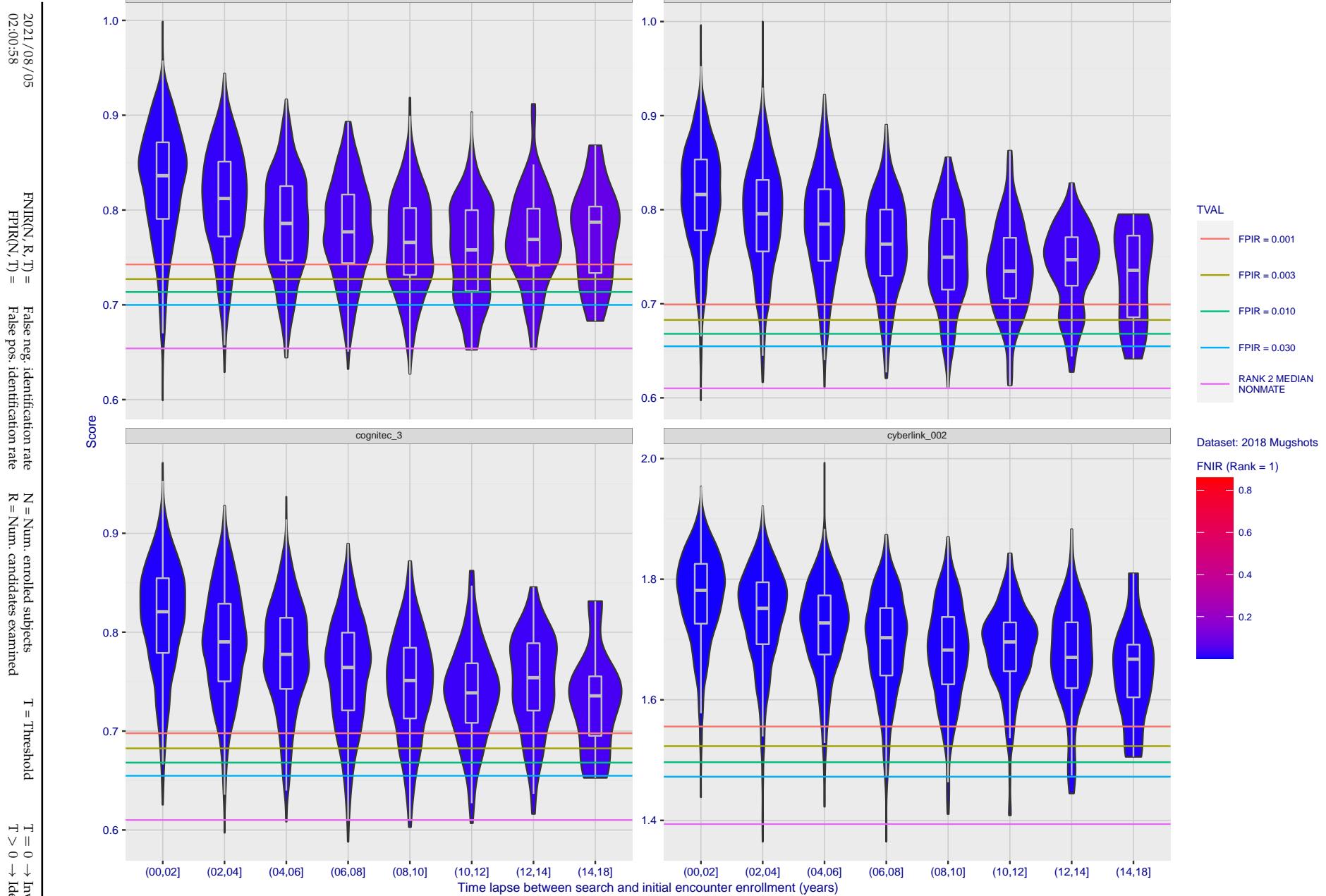


Figure 124: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

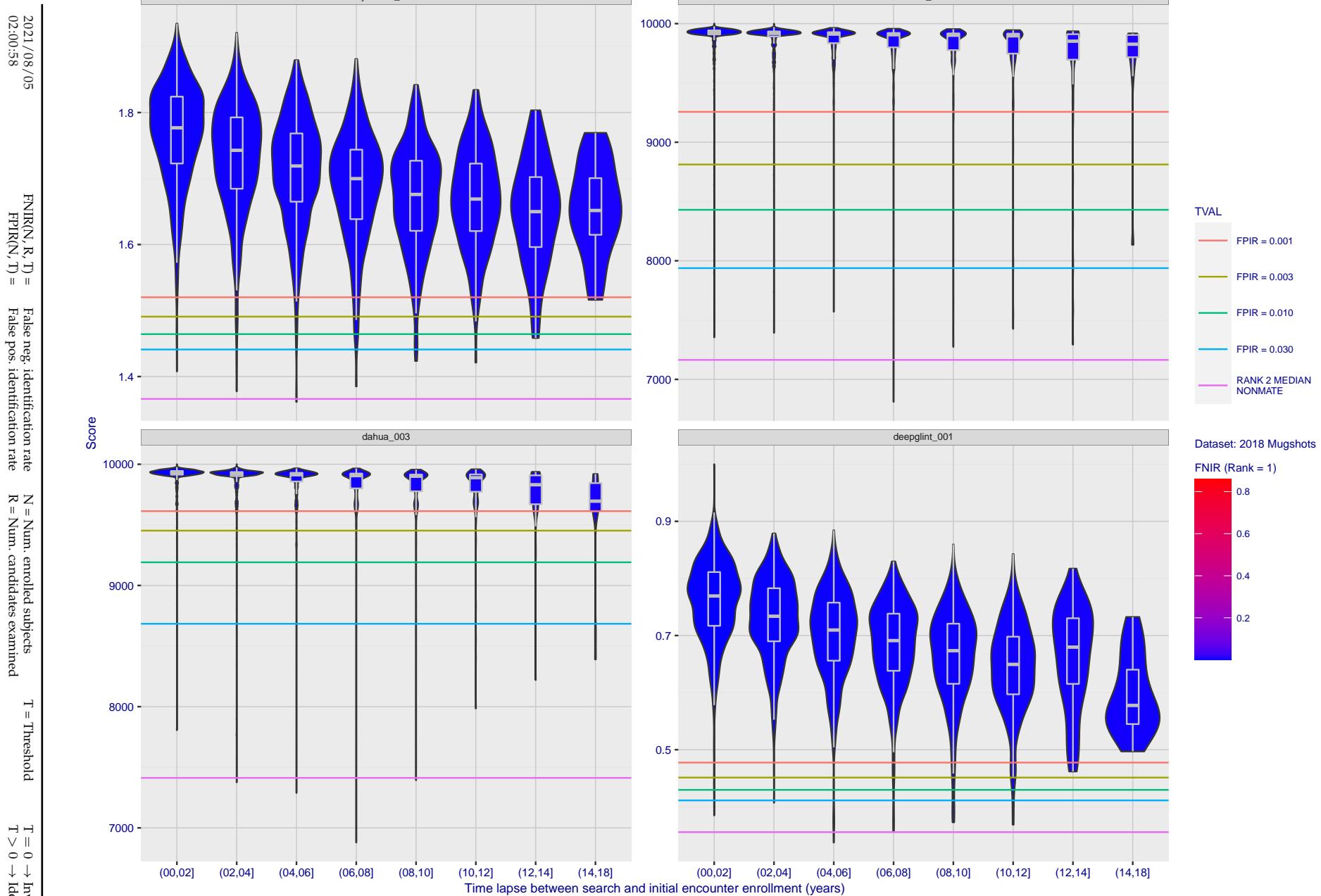


Figure 125: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

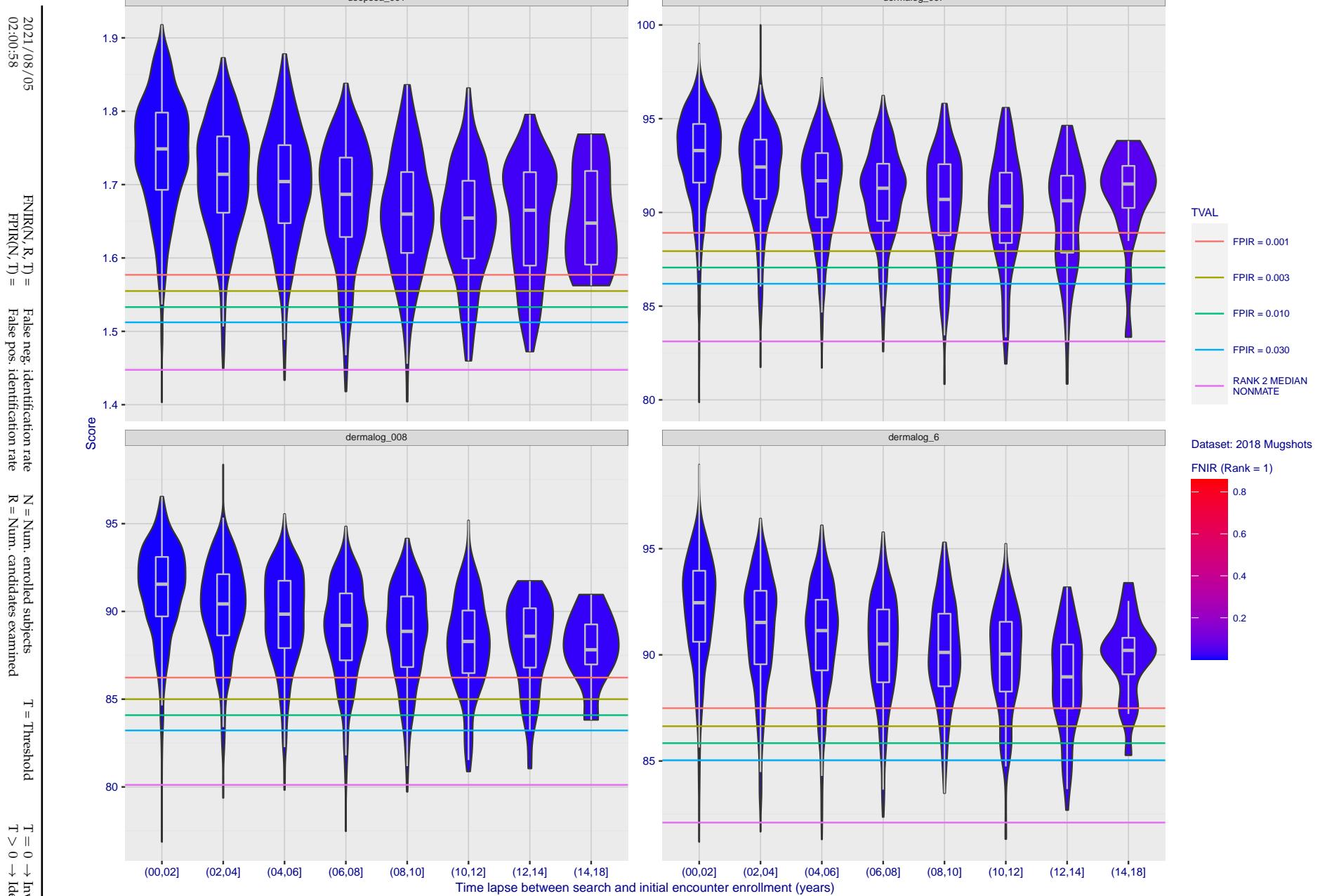


Figure 126: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

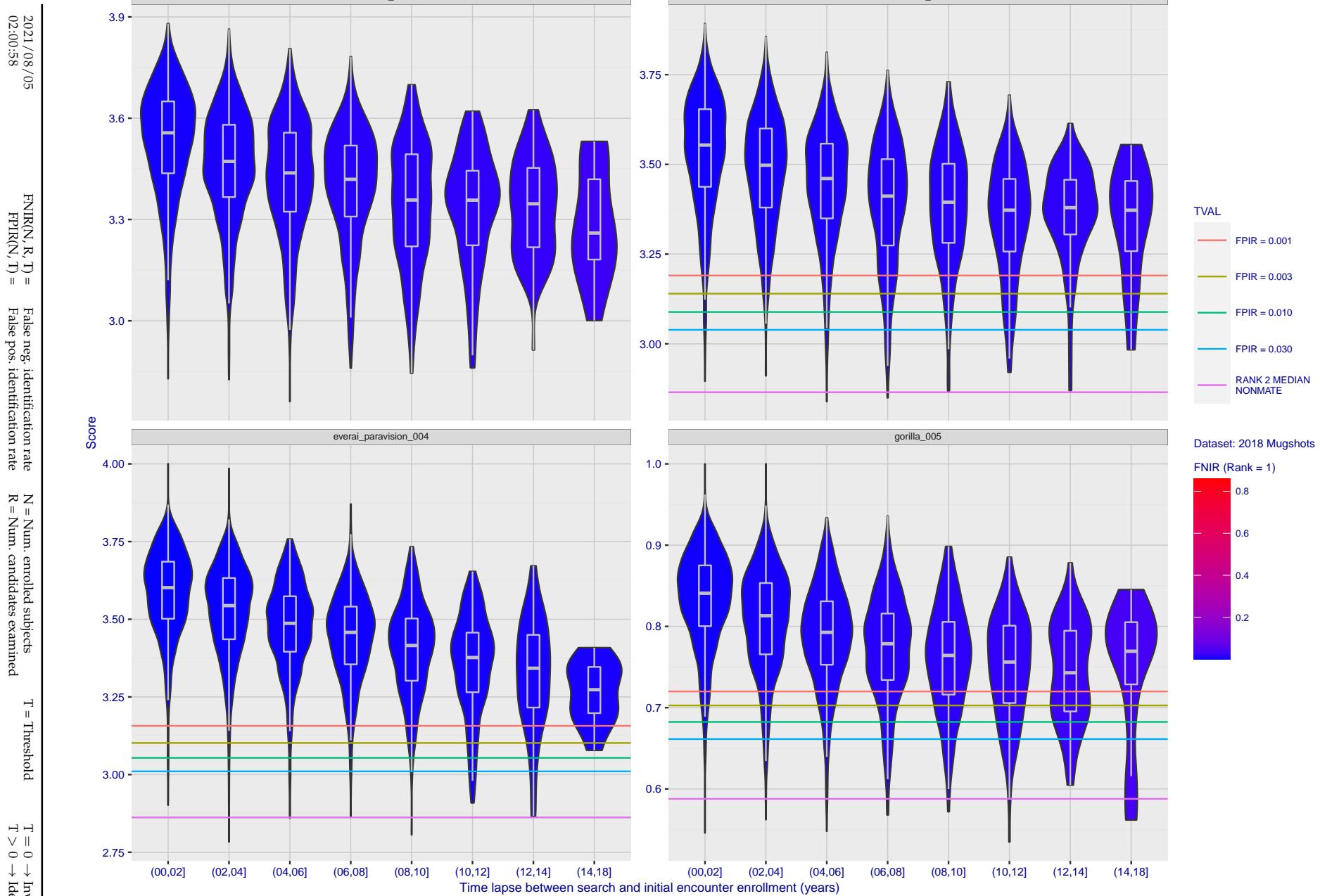


Figure 127: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

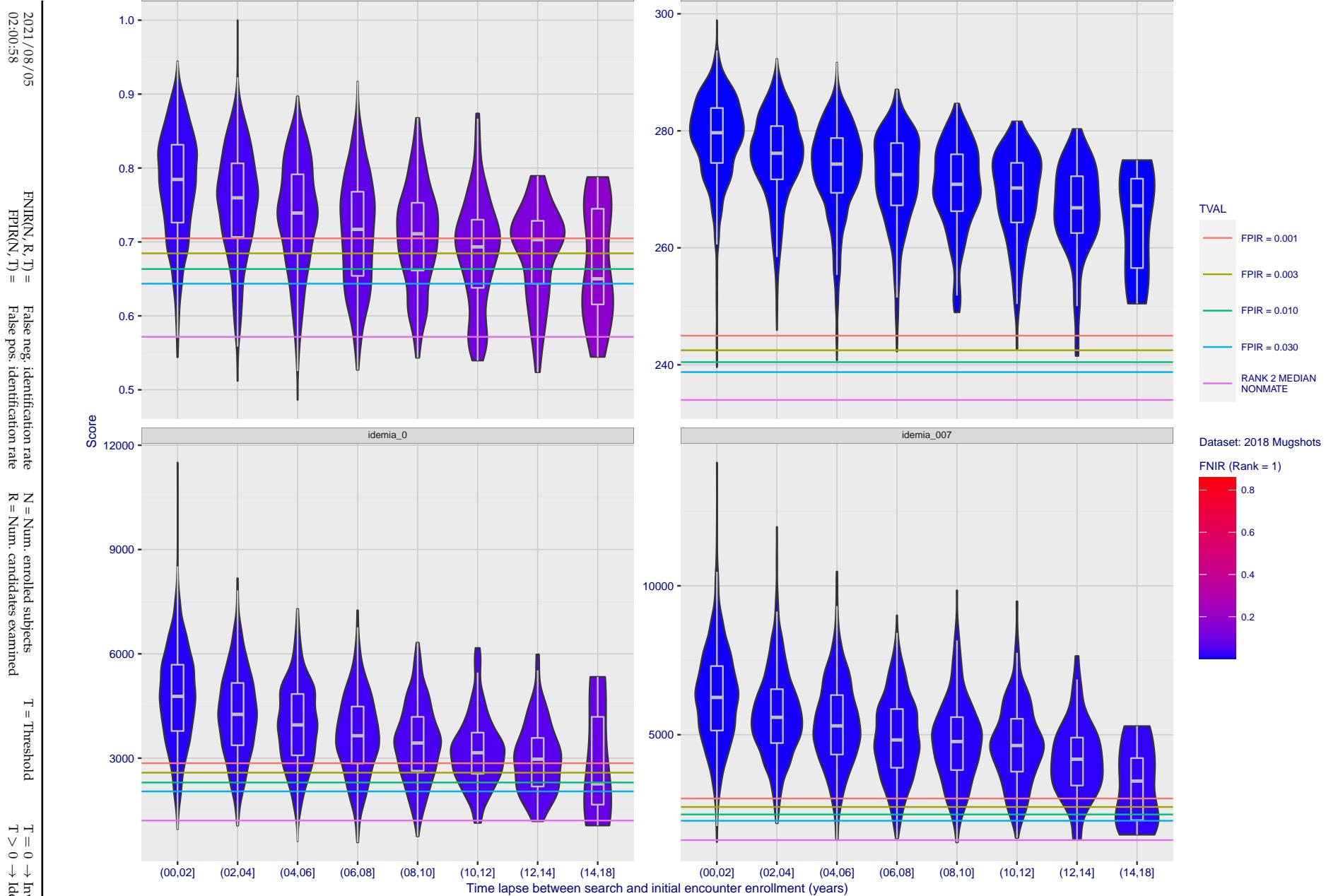


Figure 128: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

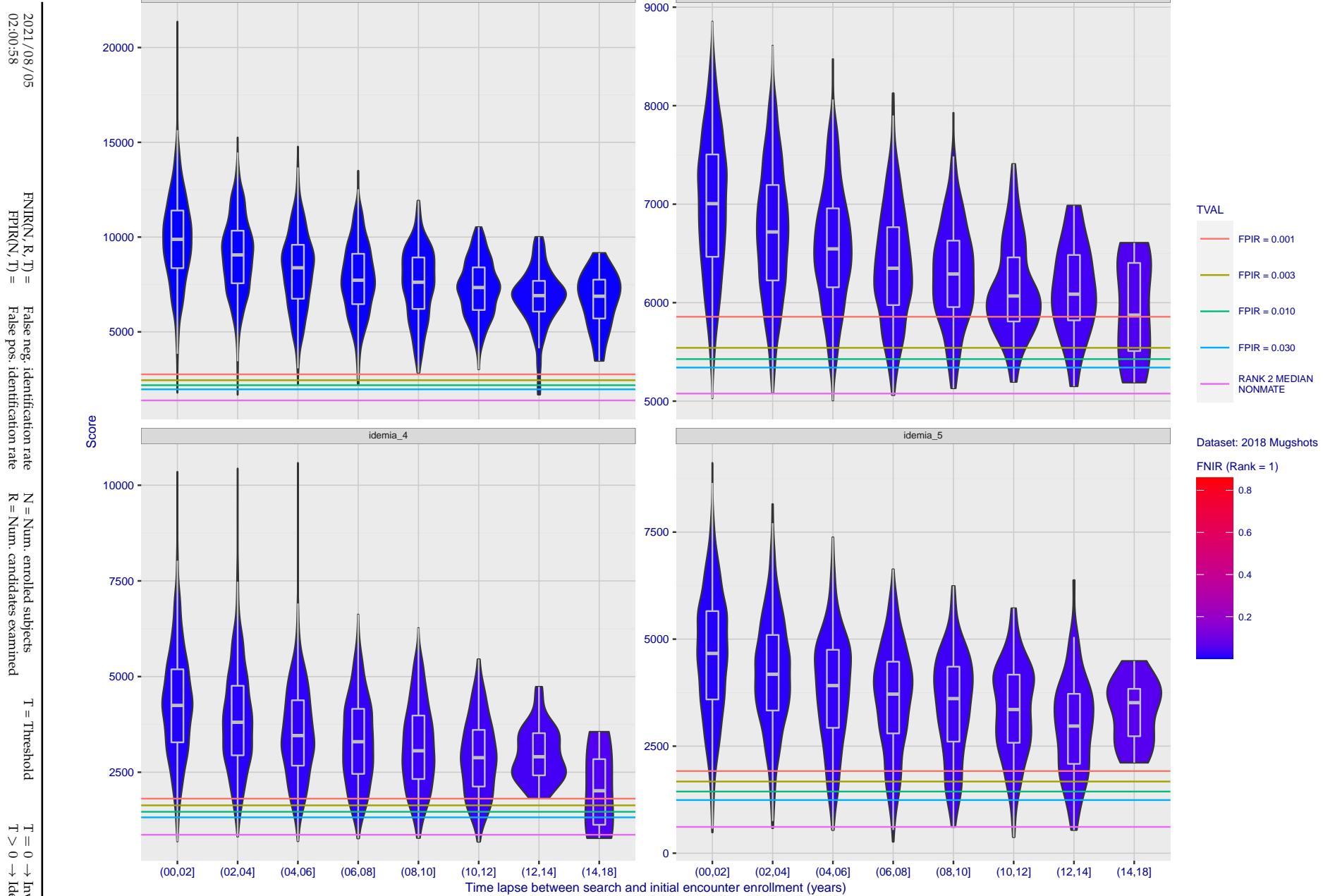


Figure 129: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

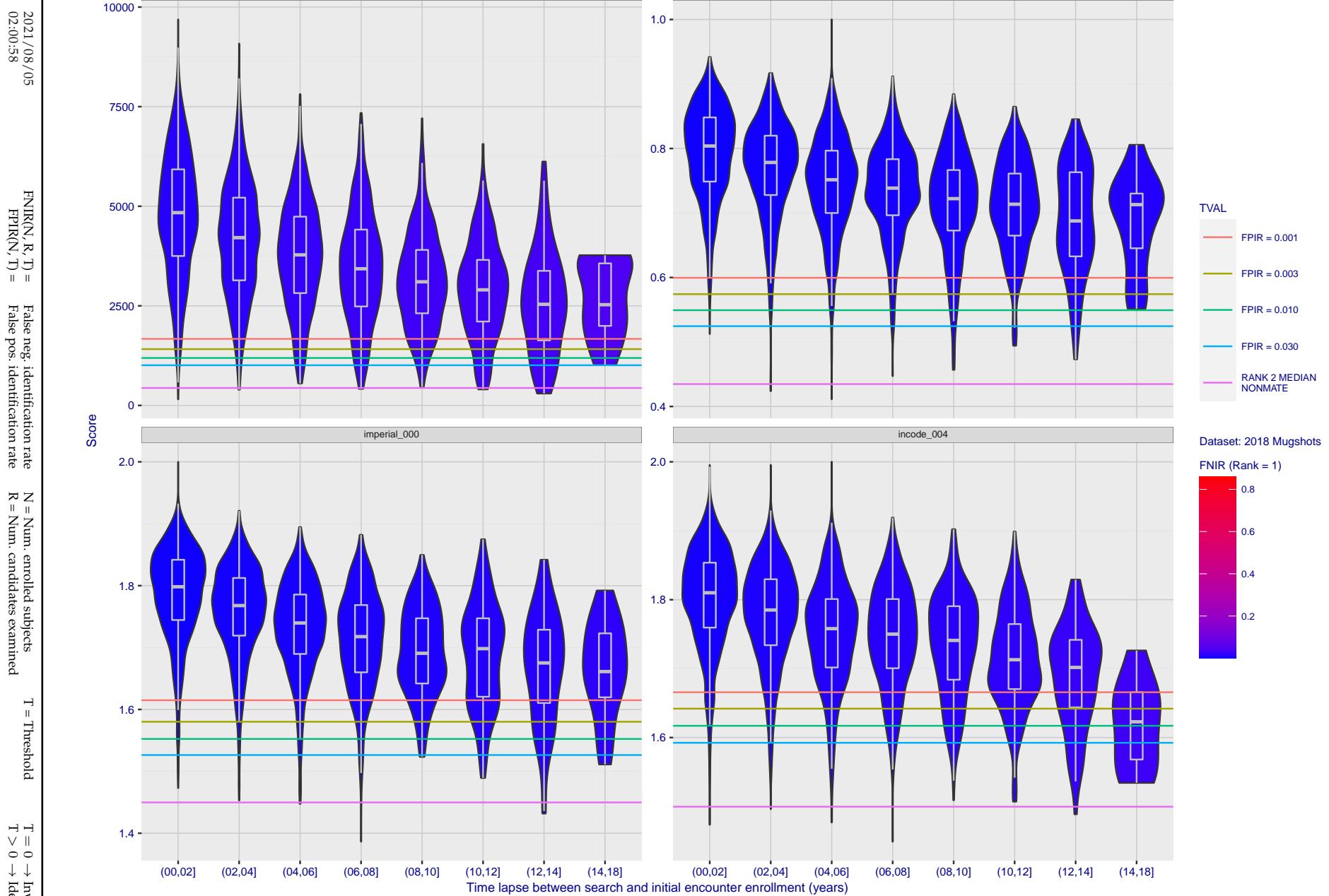


Figure 130: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

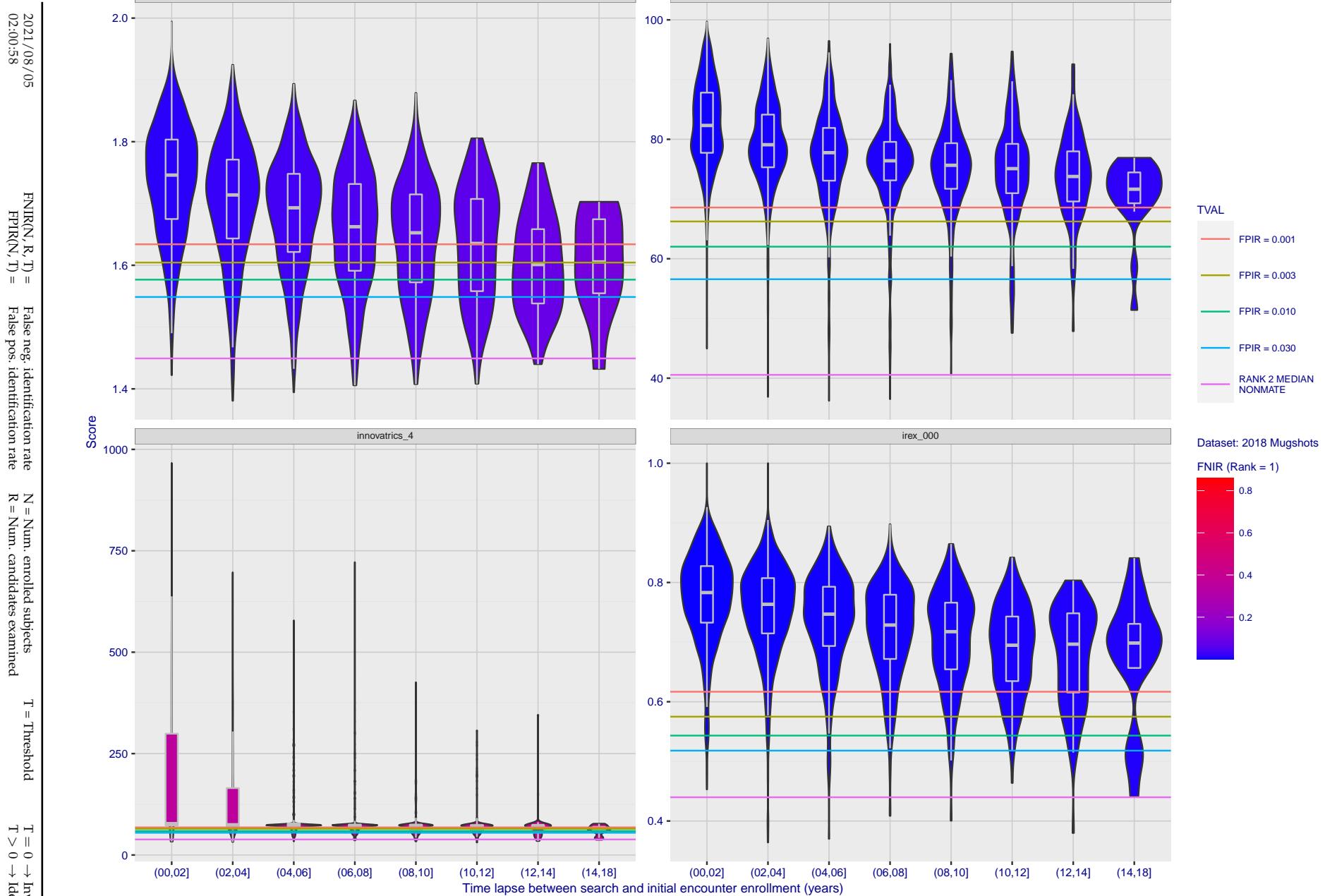


Figure 131: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

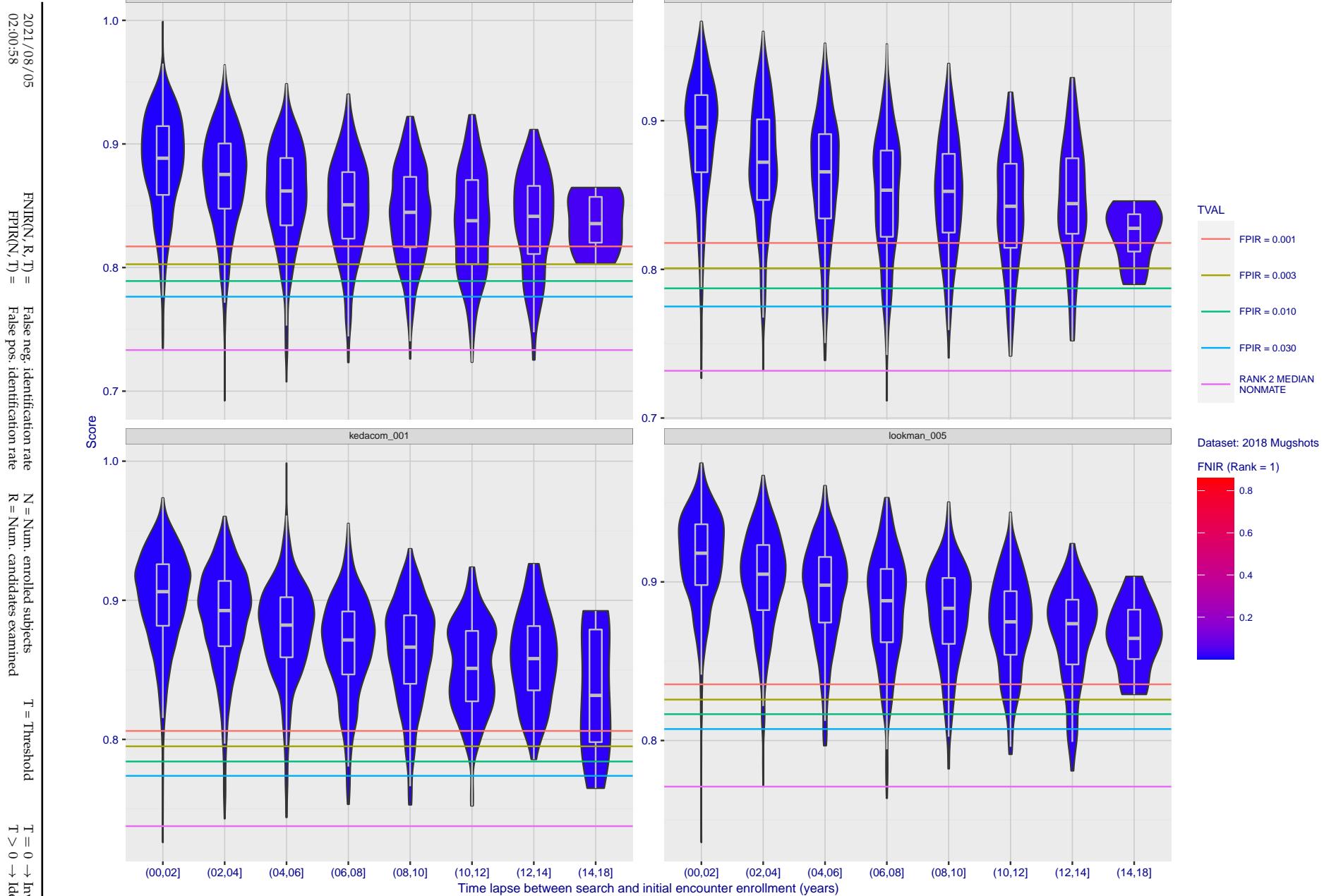


Figure 132: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

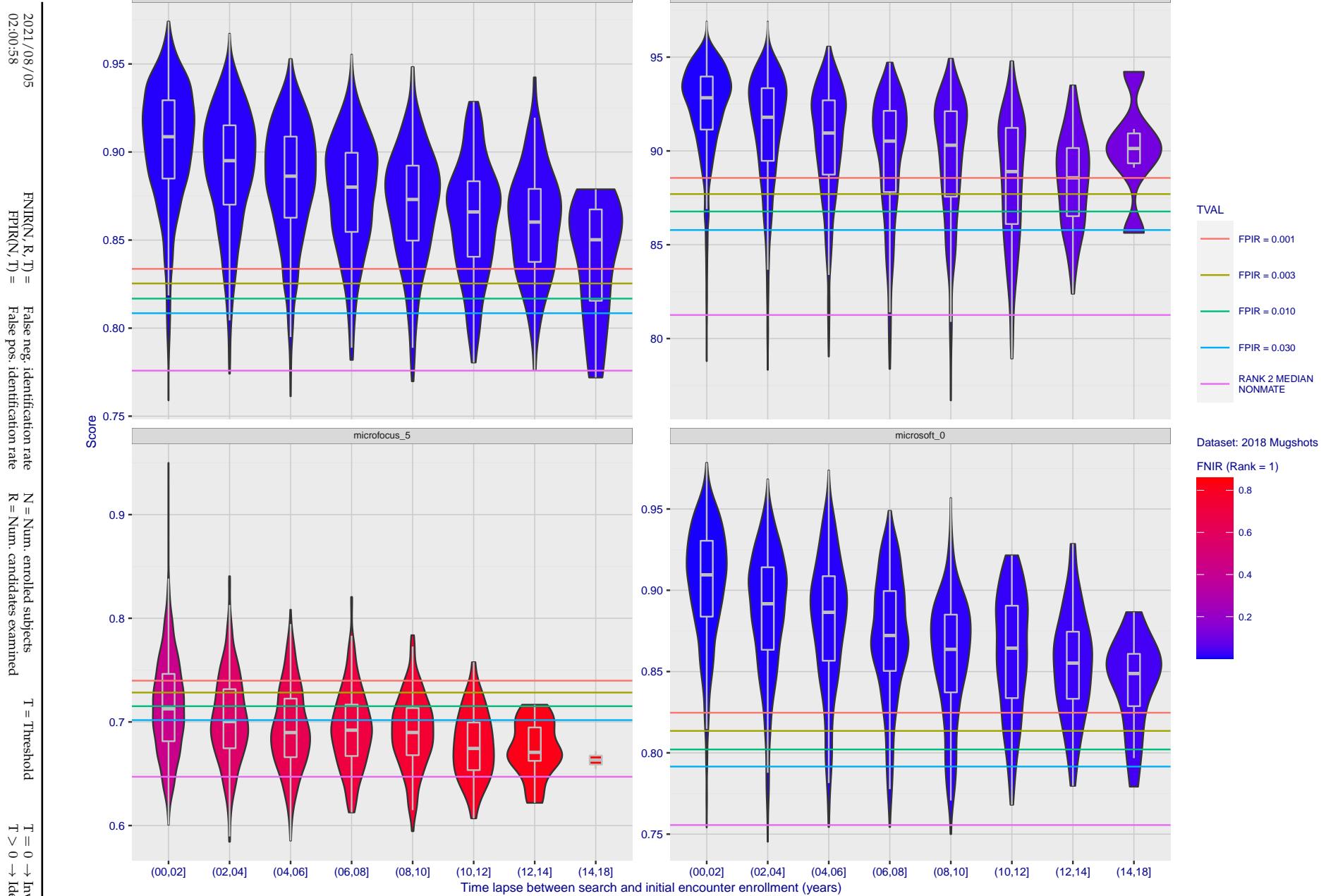


Figure 133: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

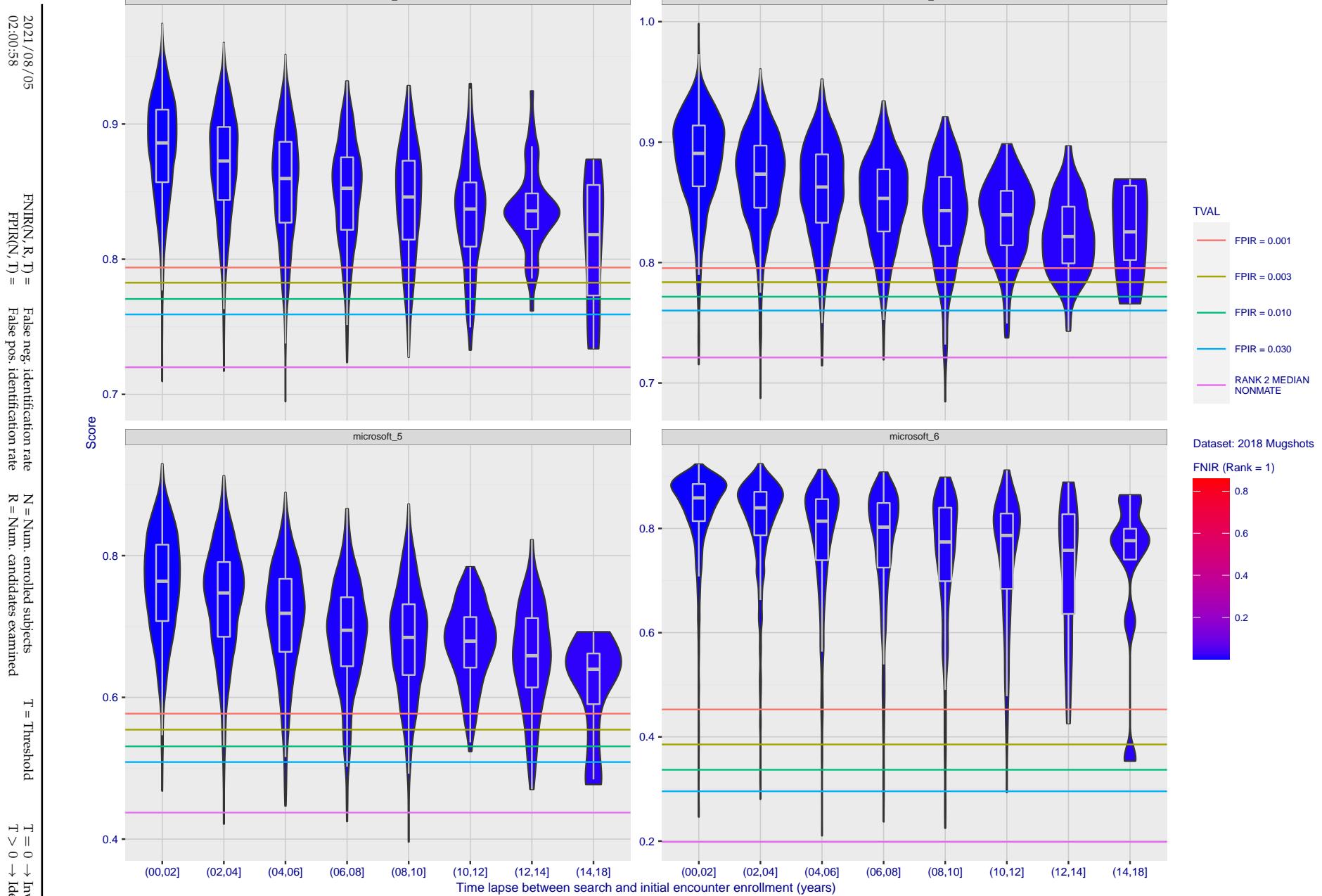


Figure 134: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

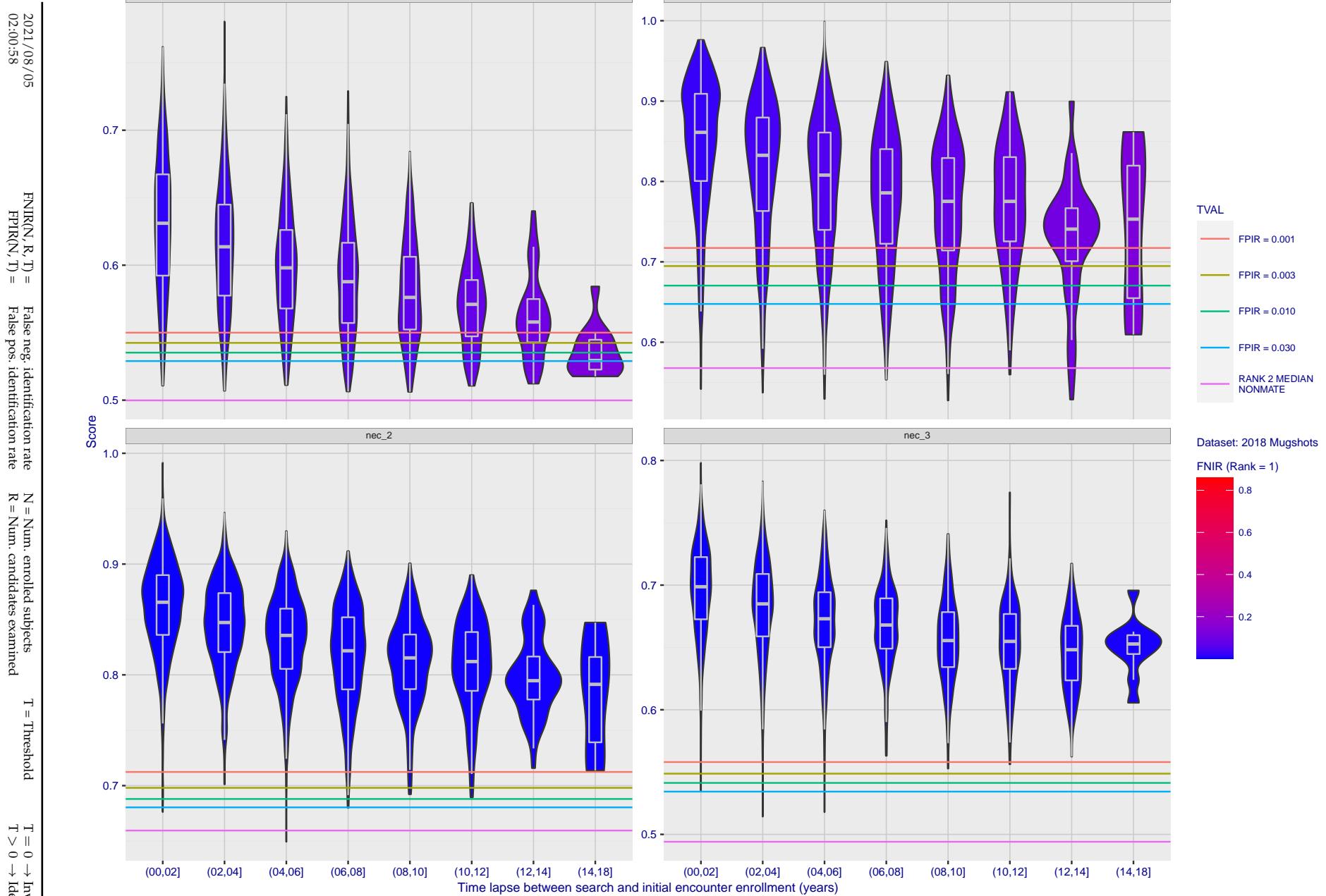


Figure 135: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

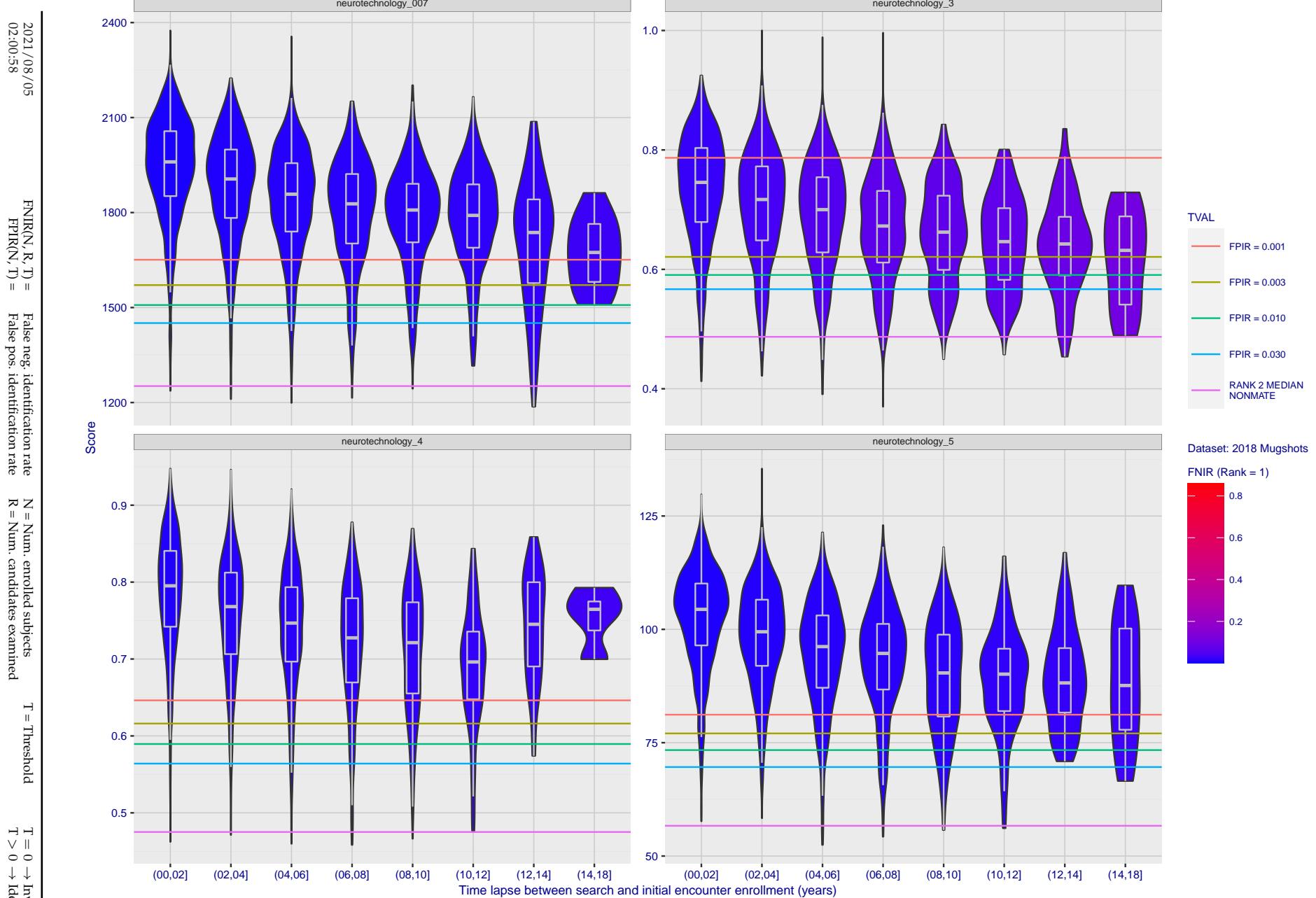


Figure 136: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

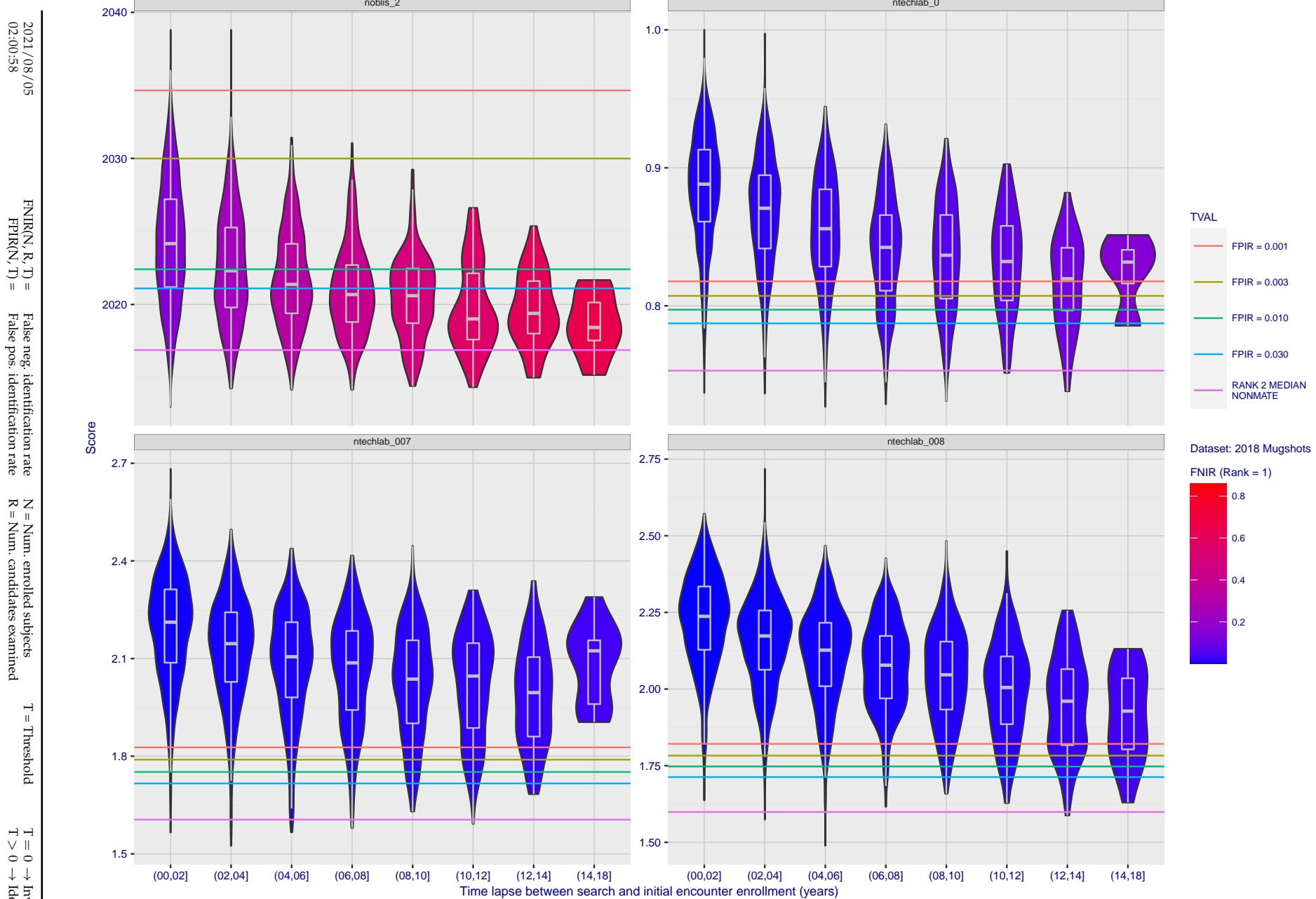


Figure 137: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

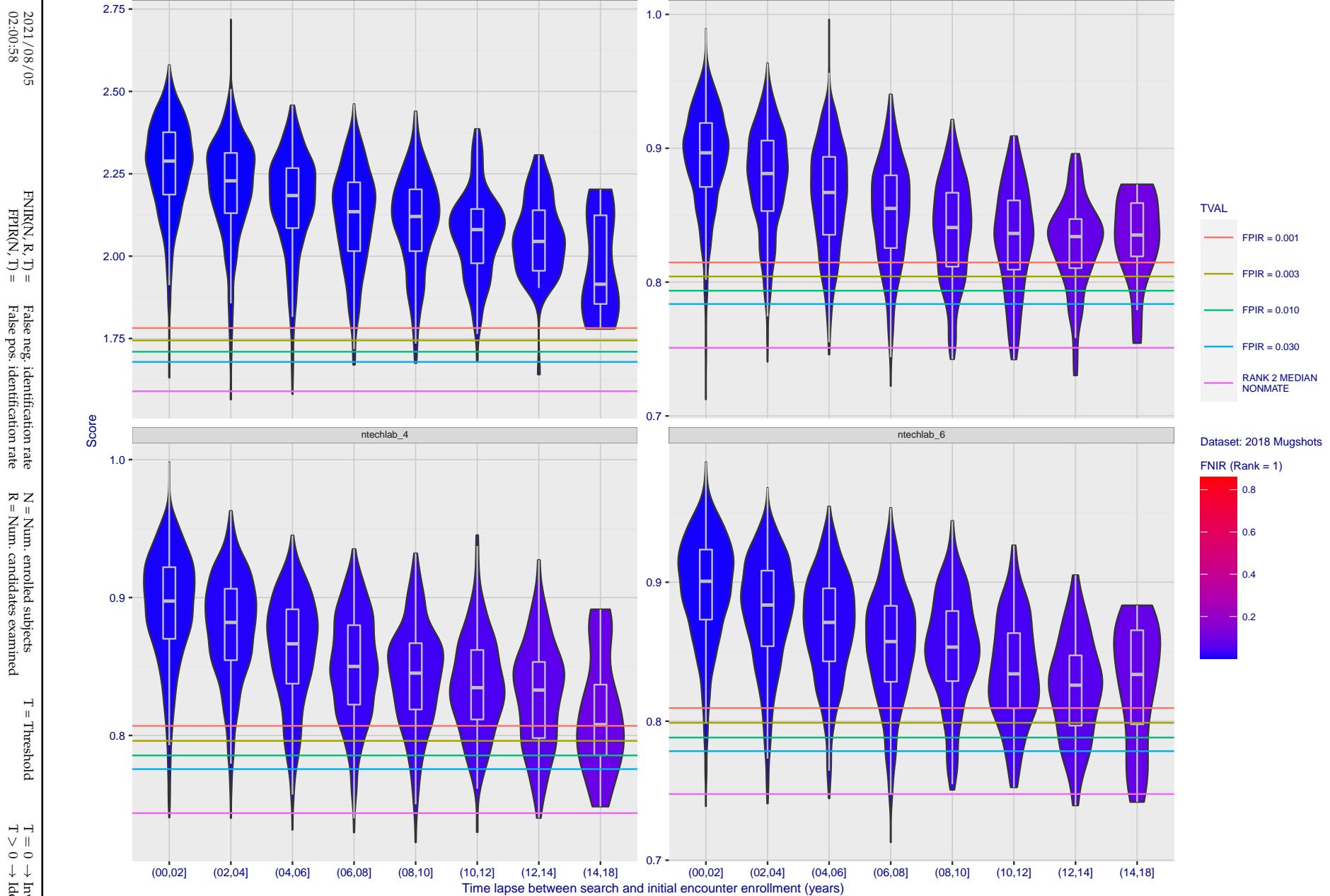


Figure 138: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

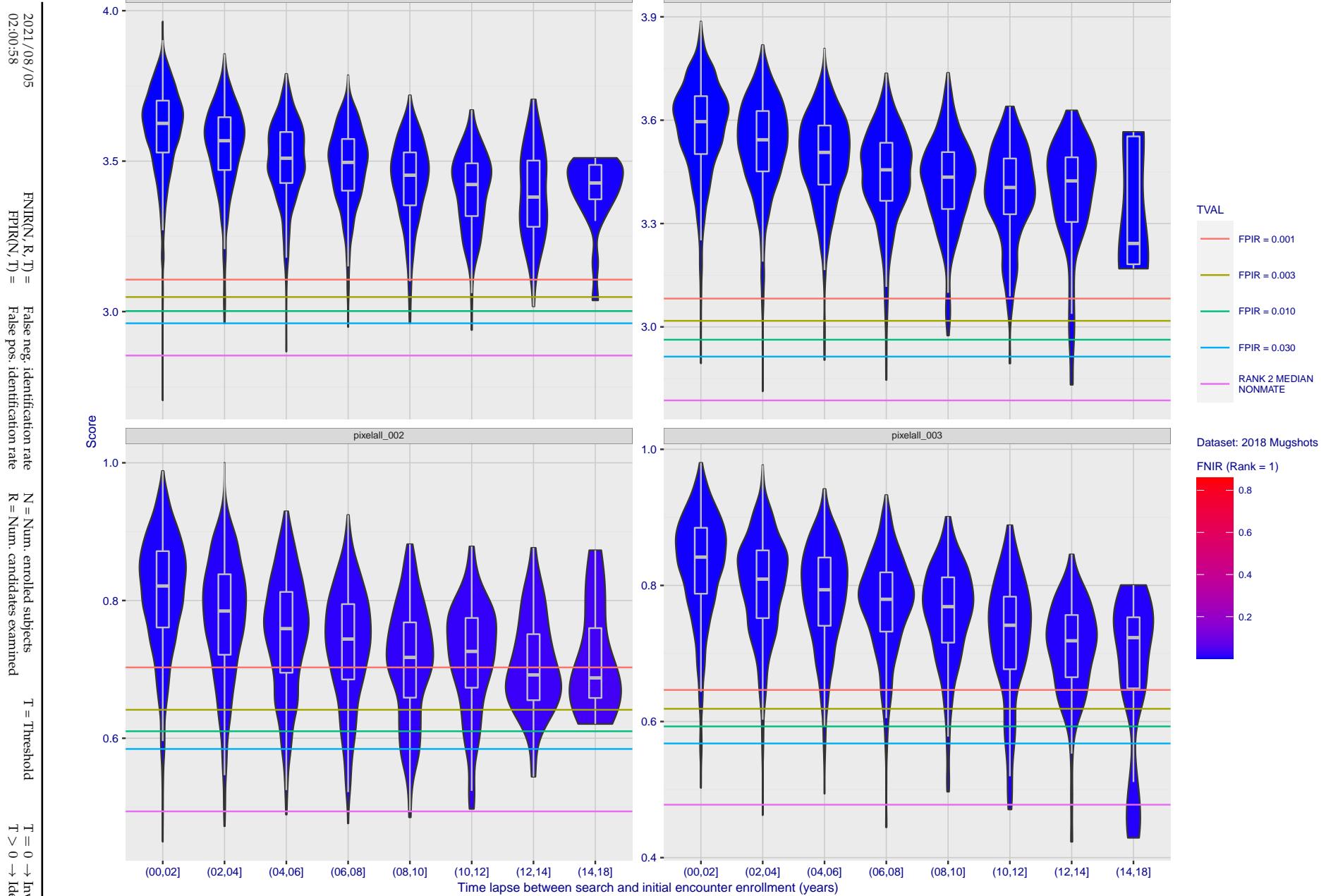


Figure 139: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

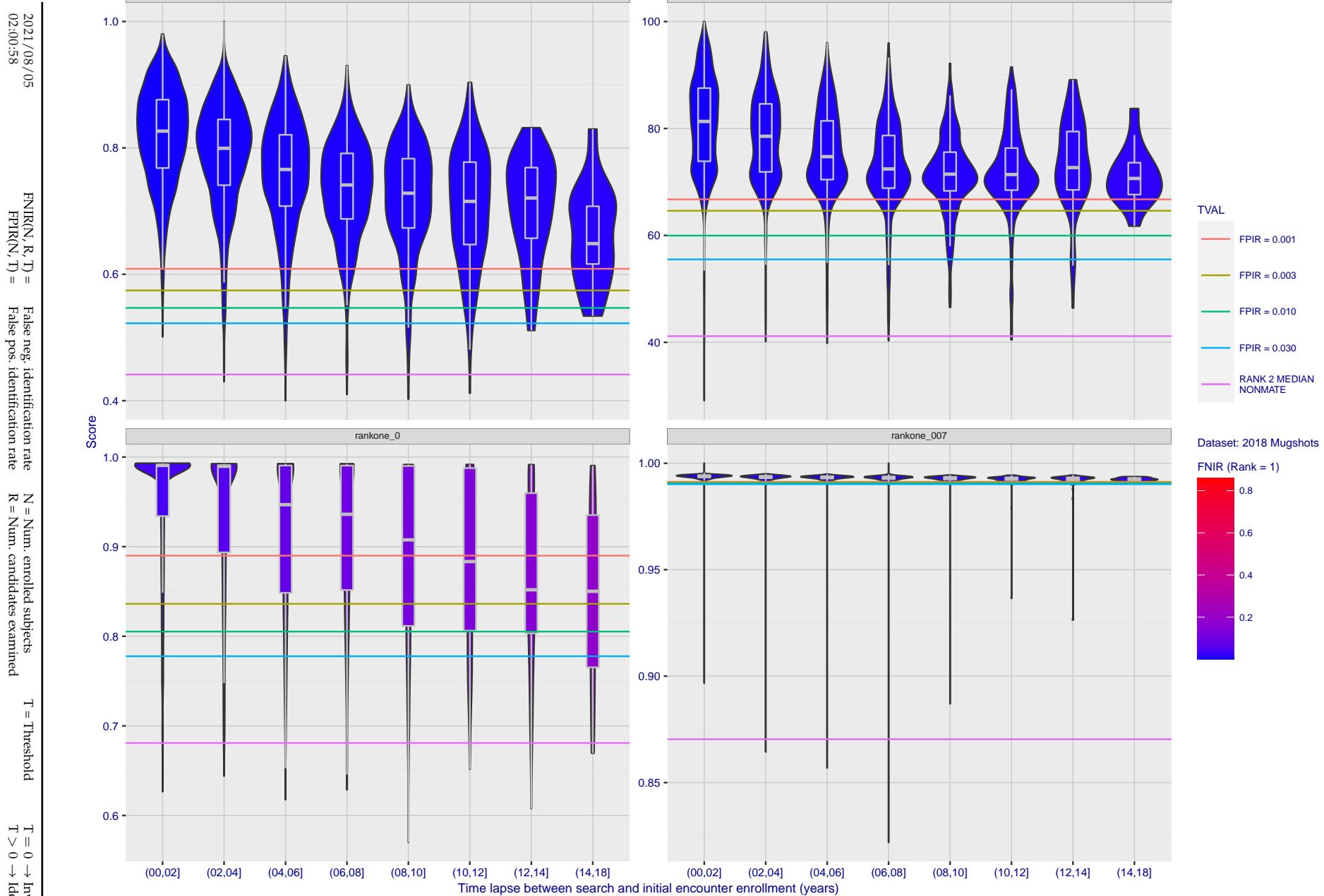


Figure 140: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

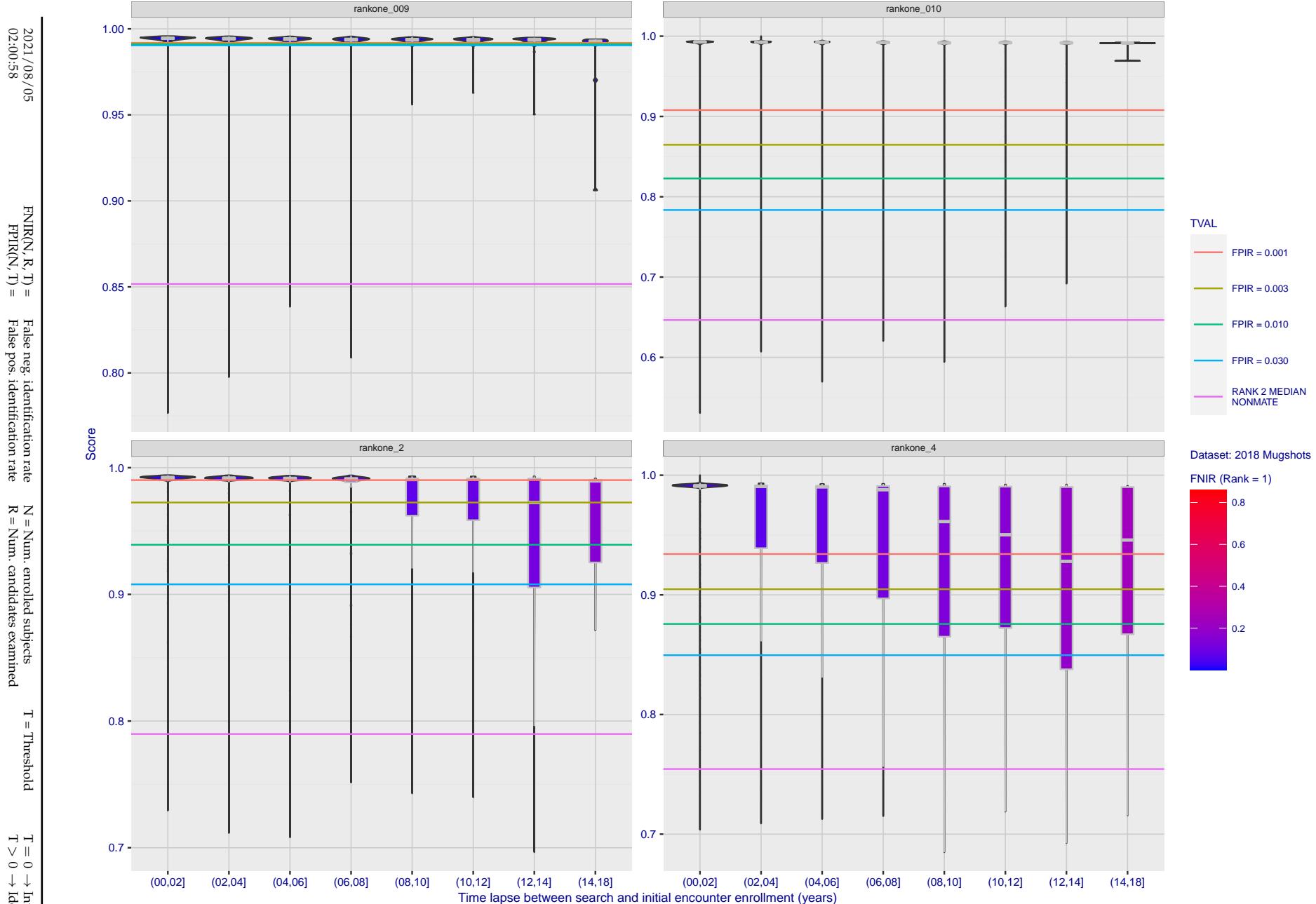


Figure 141: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

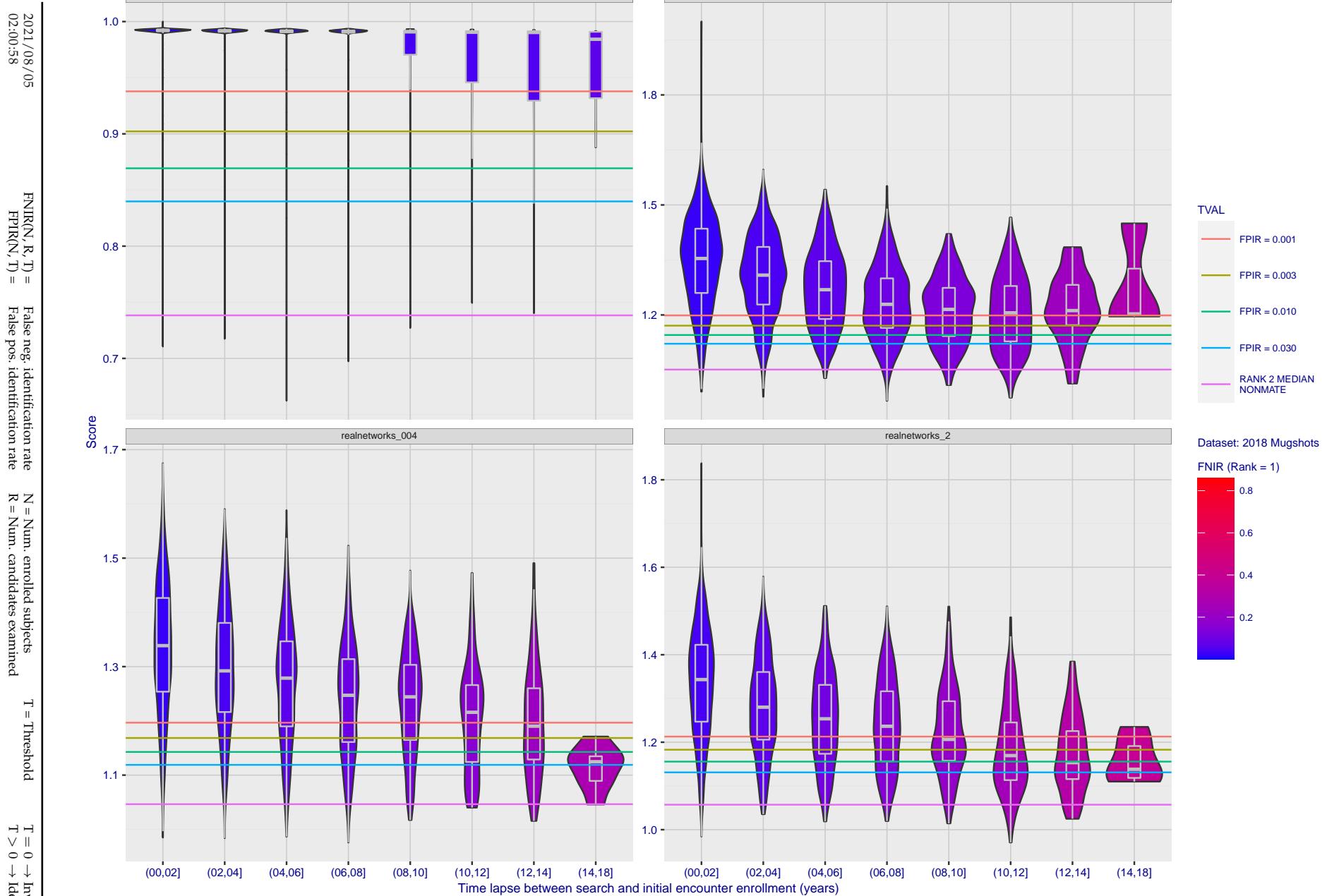


Figure 142: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

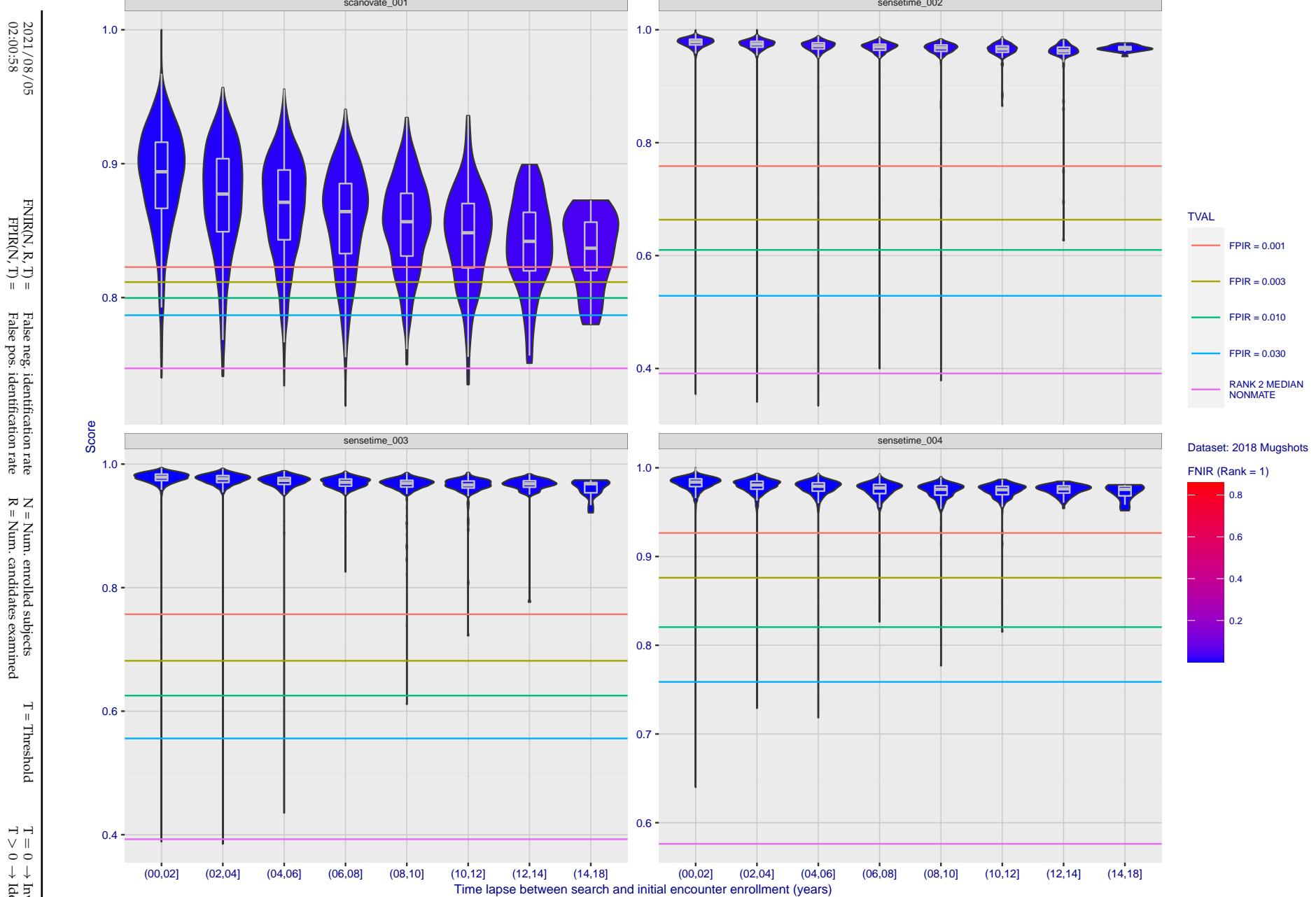


Figure 143: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

2021/08/05
02:00:58

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate

N = Num. enrolled subjects
R = Num. candidates examined

T = Threshold

T = 0 → Investigation
T > 0 → Identification

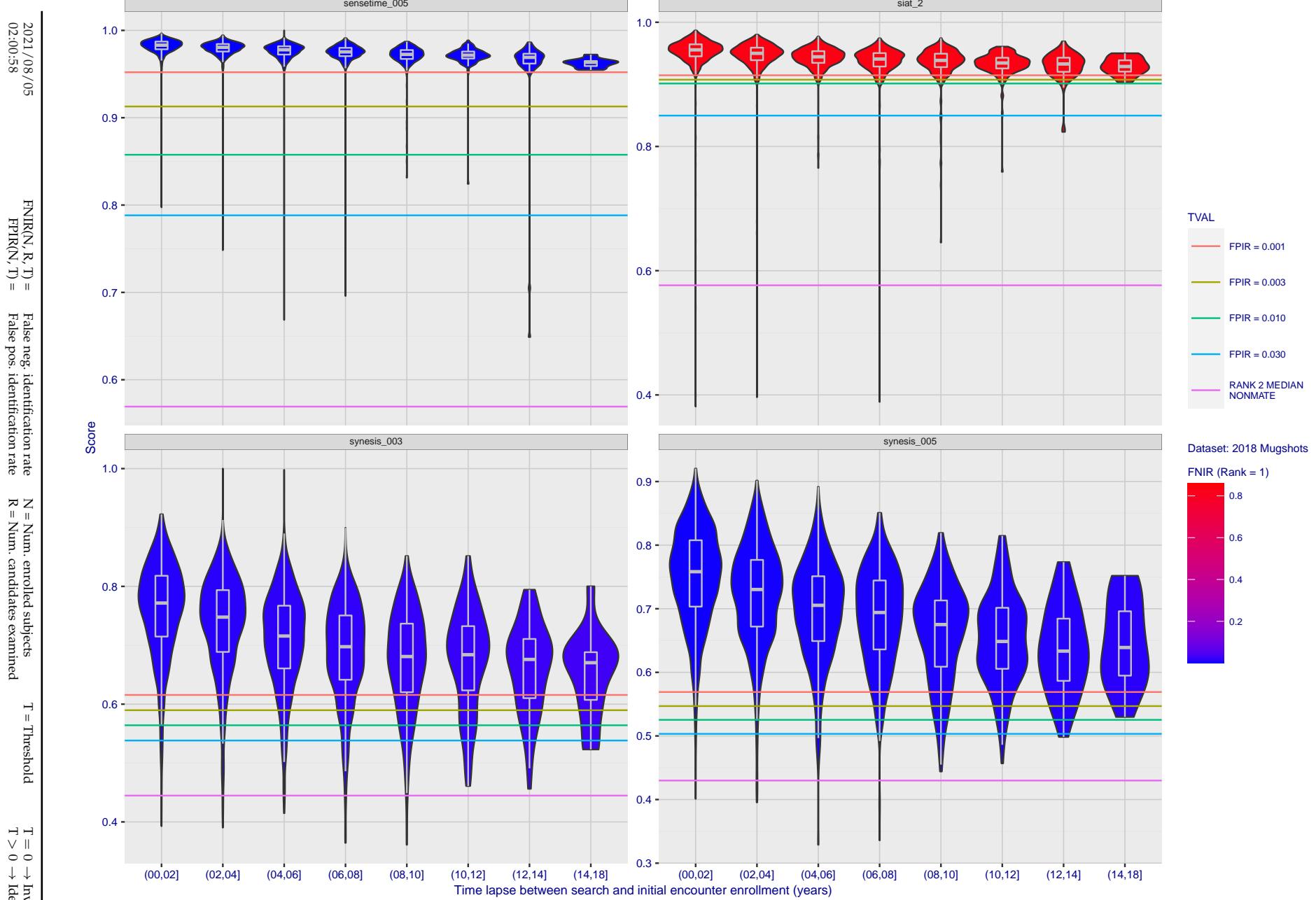


Figure 144: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

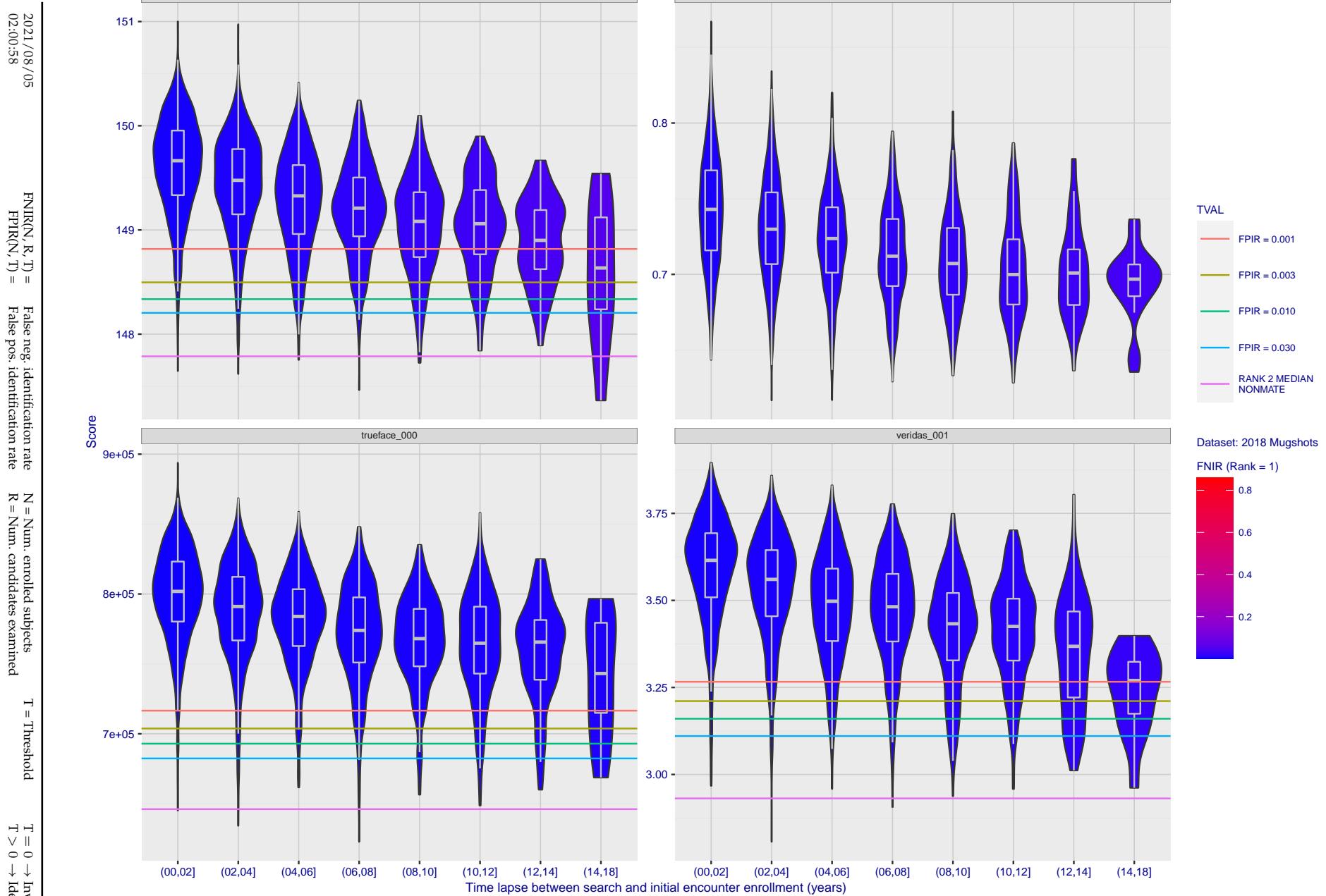


Figure 145: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

2021/08/05
02:00:58FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rateN = Num. enrolled subjects
R = Num. candidates examinedT = Threshold
T = 0 → Investigation

T > 0 → Identification

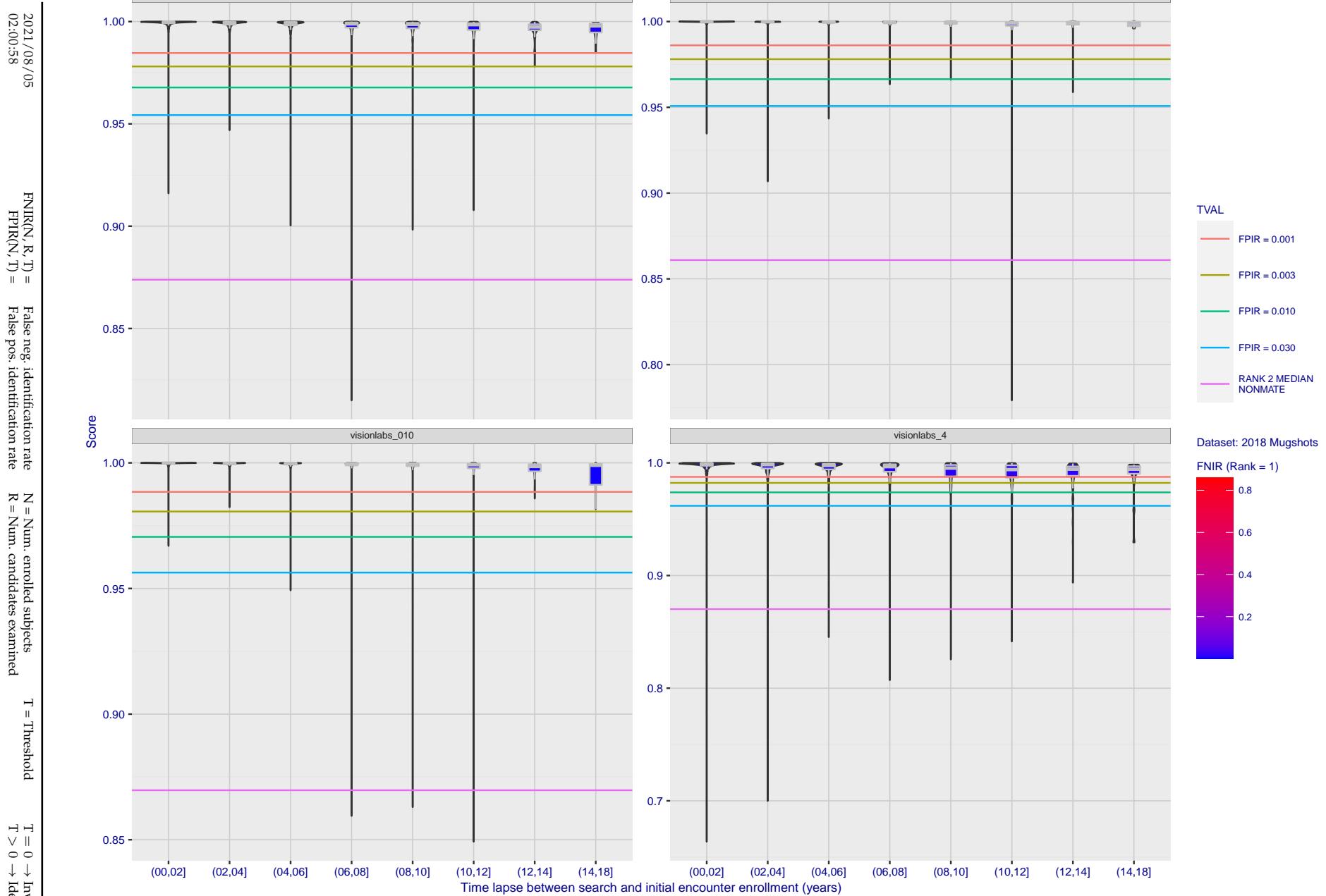


Figure 146: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

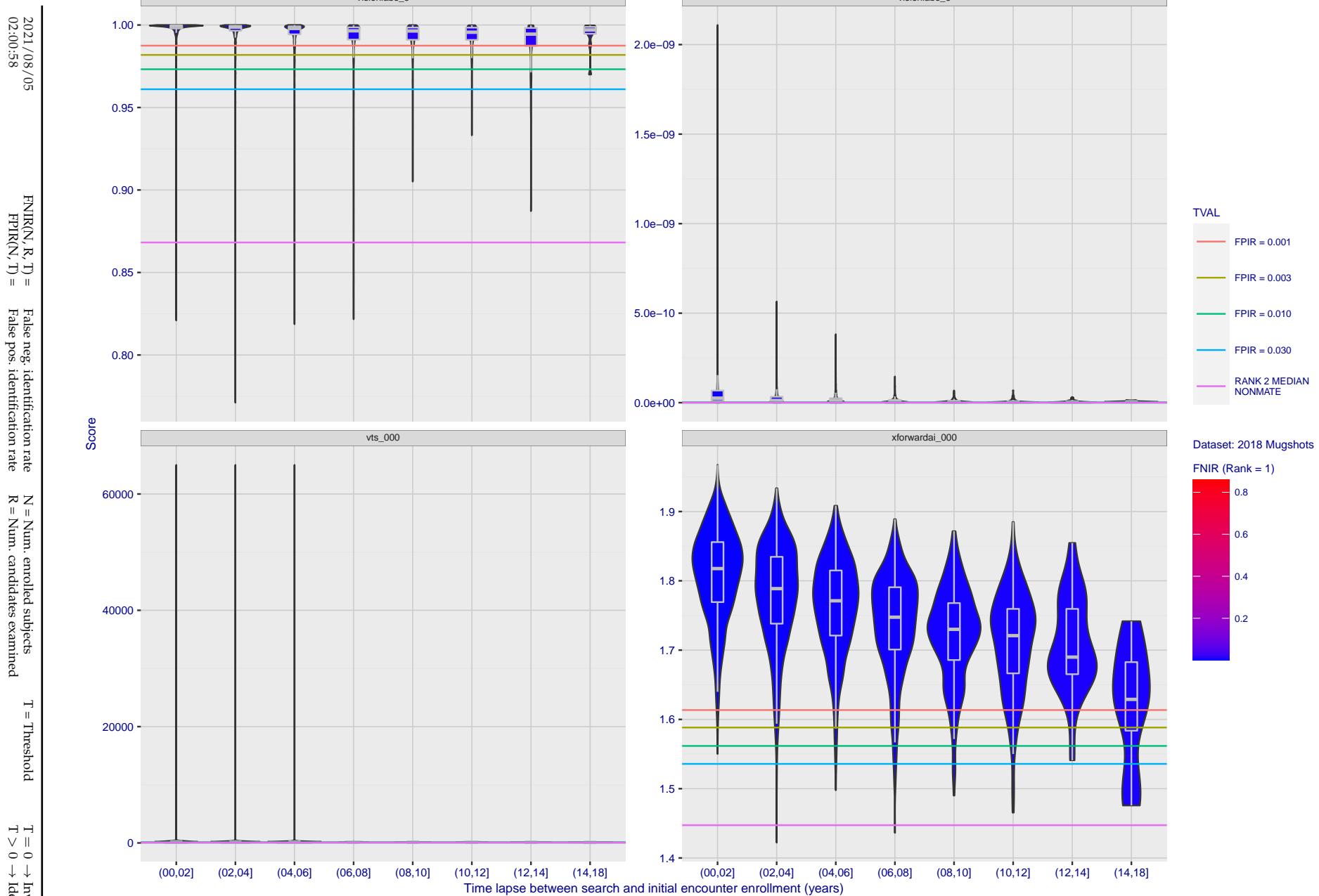


Figure 147: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

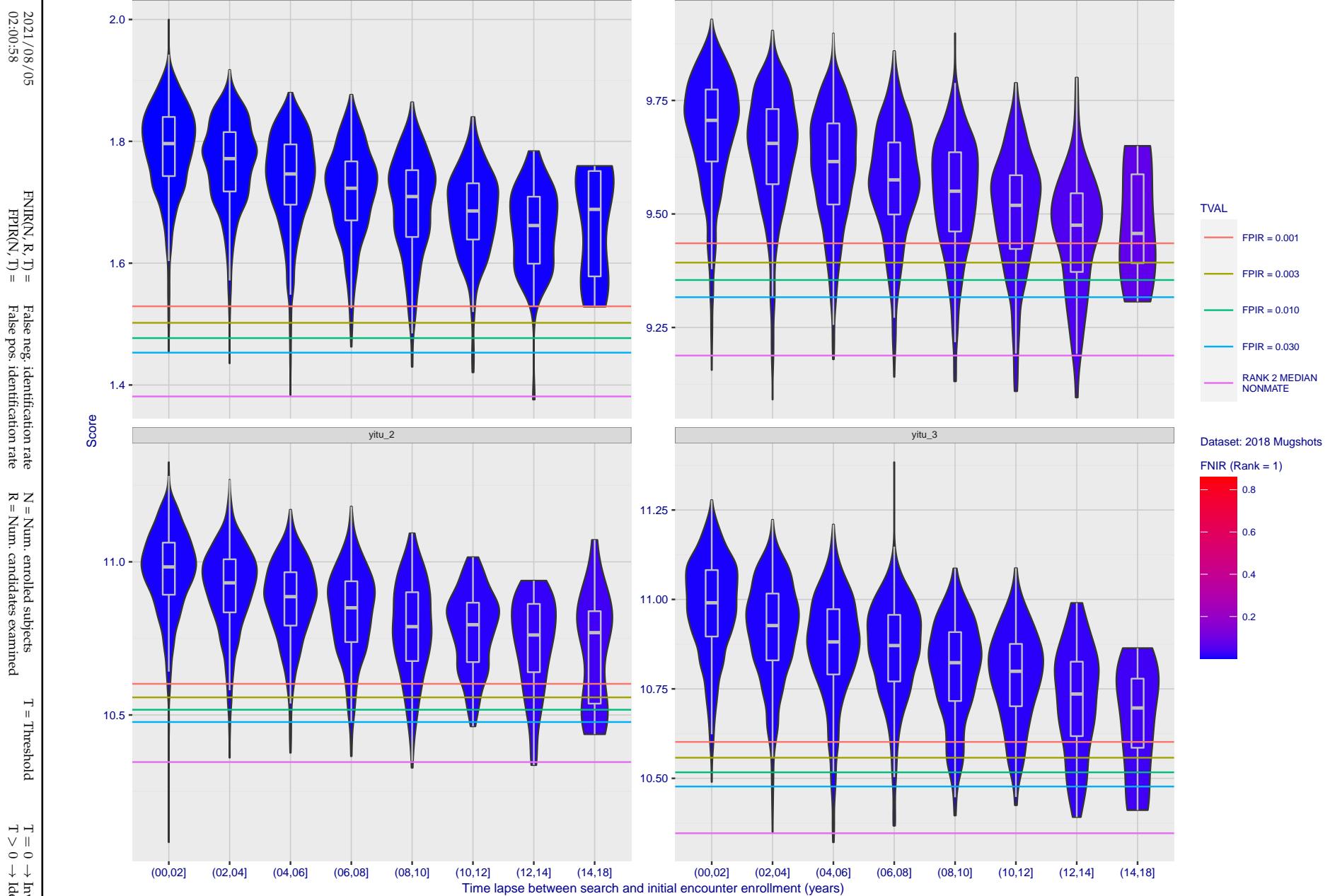


Figure 148: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

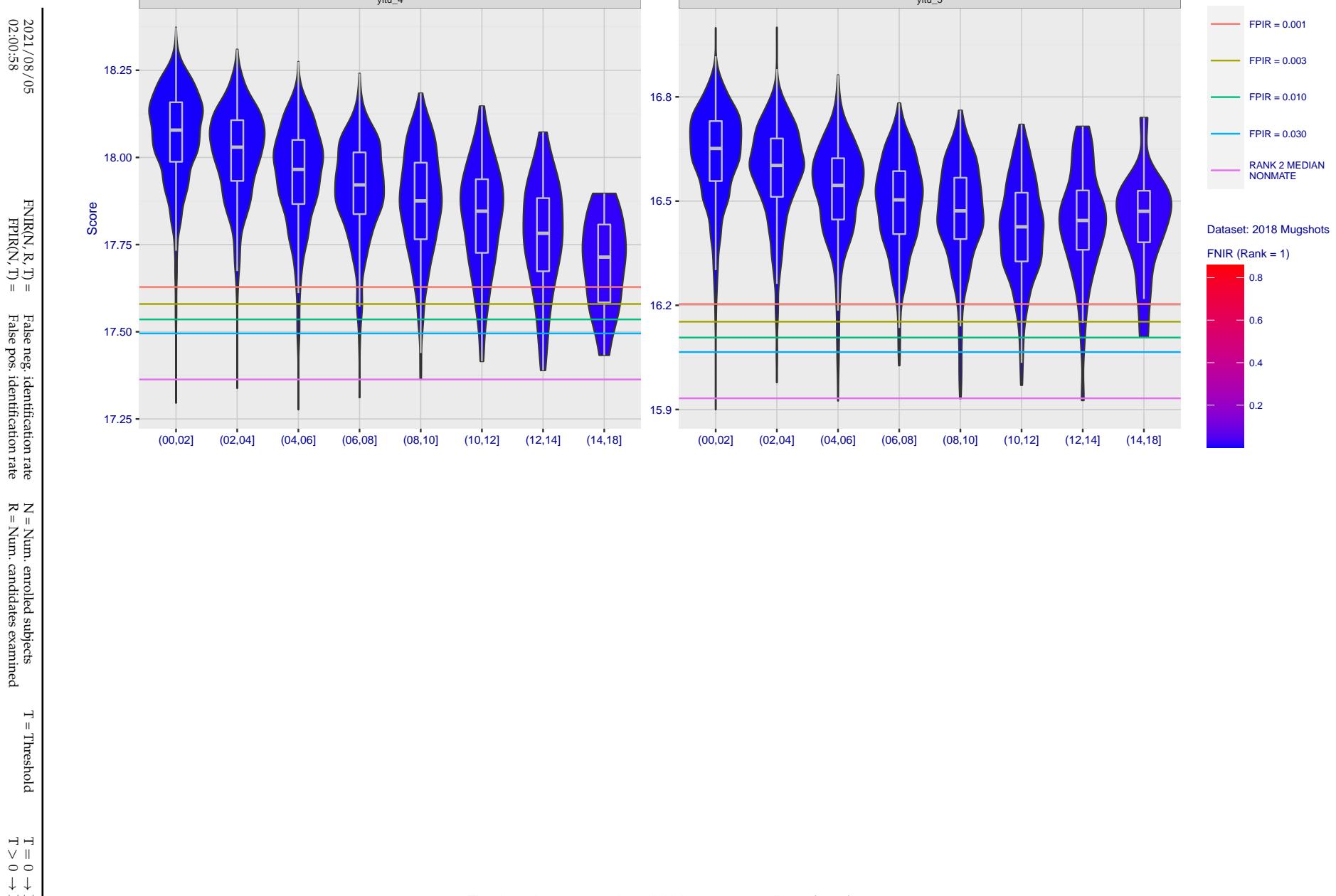


Figure 149: [FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed. The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 1 binned by number of years between search and initial enrollment.

Appendix C Effect of enrolling multiple images

2021/08/05
02:00:58

 $FNIR(N, R, T) =$
False neg. identification rate
 $FPIR(N, T) =$
False pos. identification rate

 $N =$ Num. enrolled subjects
 $R =$ Num. candidates examined

 $T =$ Threshold
 $T = 0 \rightarrow$ Investigation
 $T > 0 \rightarrow$ Identification

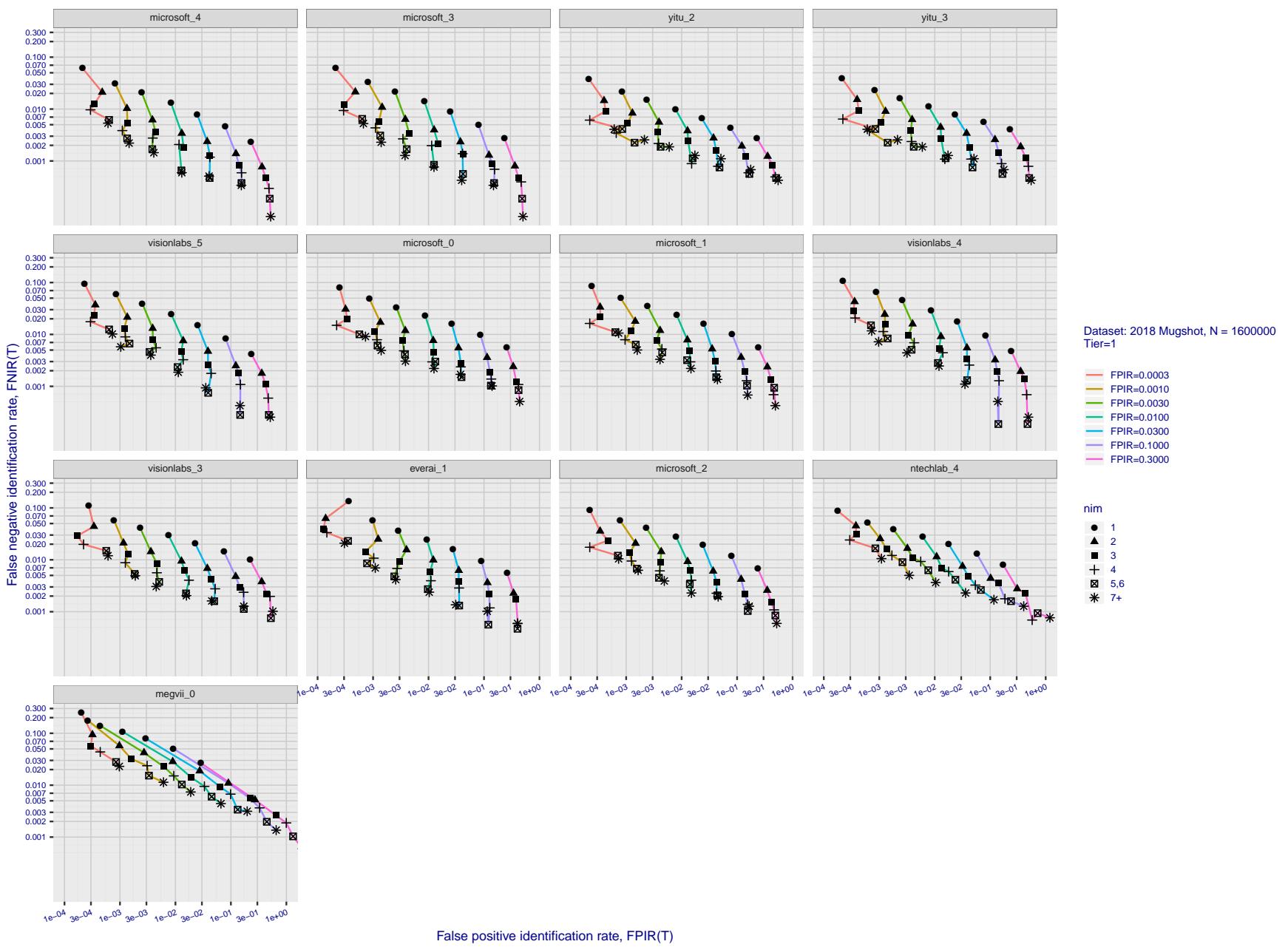


Figure 150: [FRVT-2018 Mugshot Dataset] Effect of enrolling multiple images for each identity. The plot shows an identification miss rates vs. false positive rates, at seven operating thresholds. The enrolled population size is fixed. The images are enrolled with lifetime-consolidation - see section 2.3.

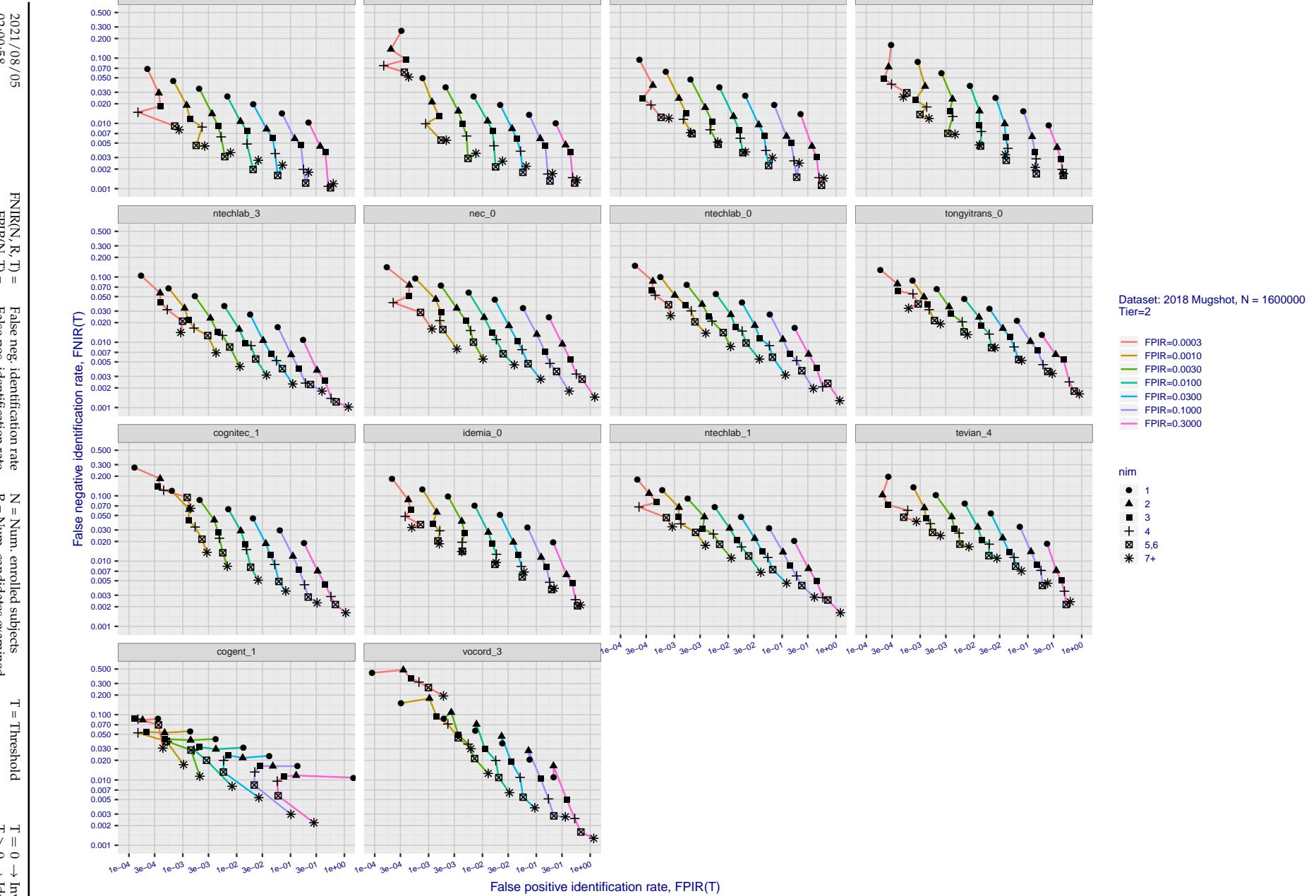


Figure 151: [FRVT-2018 Mugshot Dataset] Effect of enrolling multiple images for each identity. The plot shows an identification miss rates vs. false positive rates, at seven operating thresholds. The enrolled population size is fixed. The images are enrolled with lifetime-consolidation - see section 2.3.

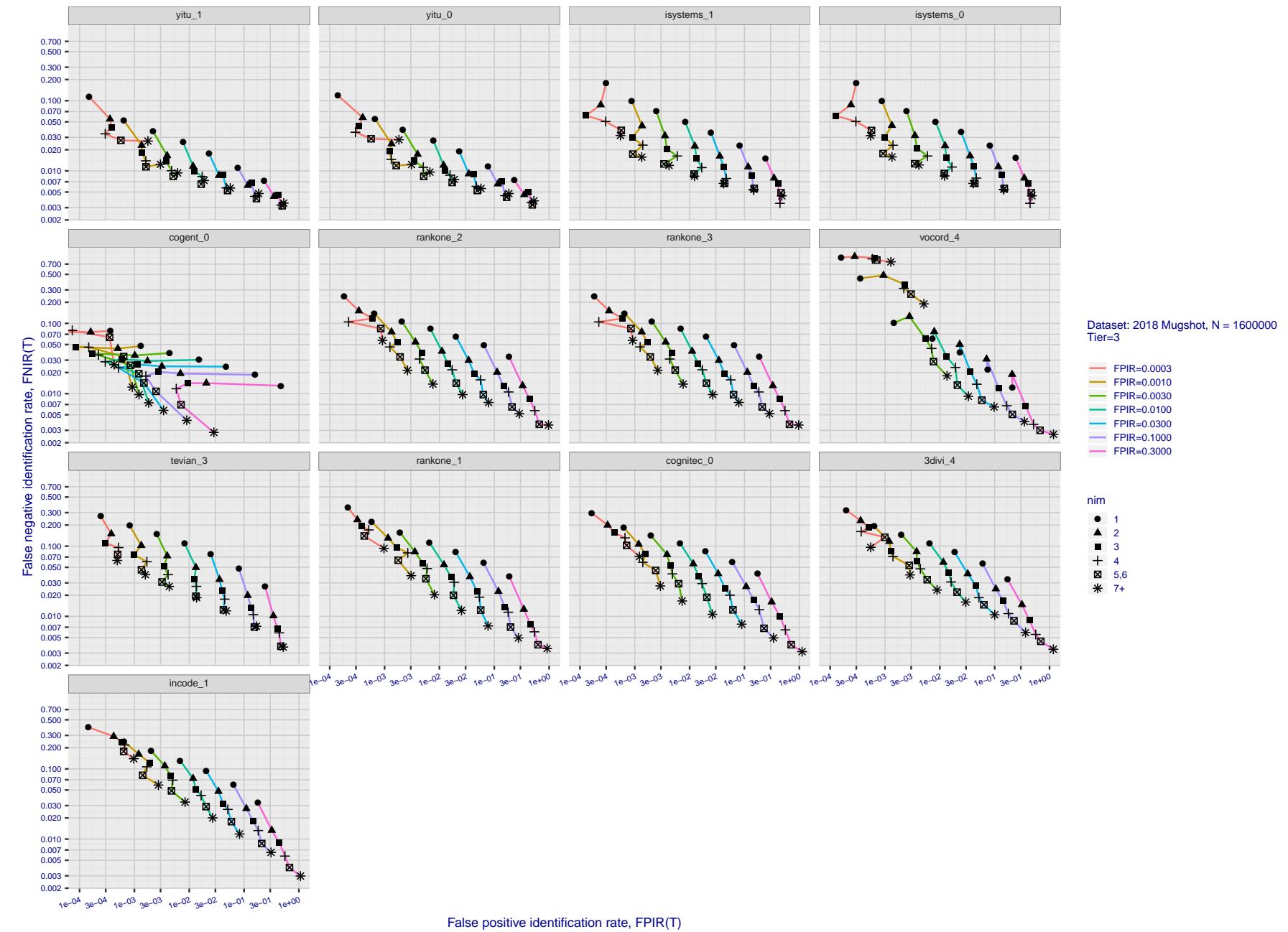


Figure 152: [FRVT-2018 Mugshot Dataset] Effect of enrolling multiple images for each identity. The plot shows an identification miss rates vs. false positive rates, at seven operating thresholds. The enrolled population size is fixed. The images are enrolled with lifetime-consolidation - see section 2.3.

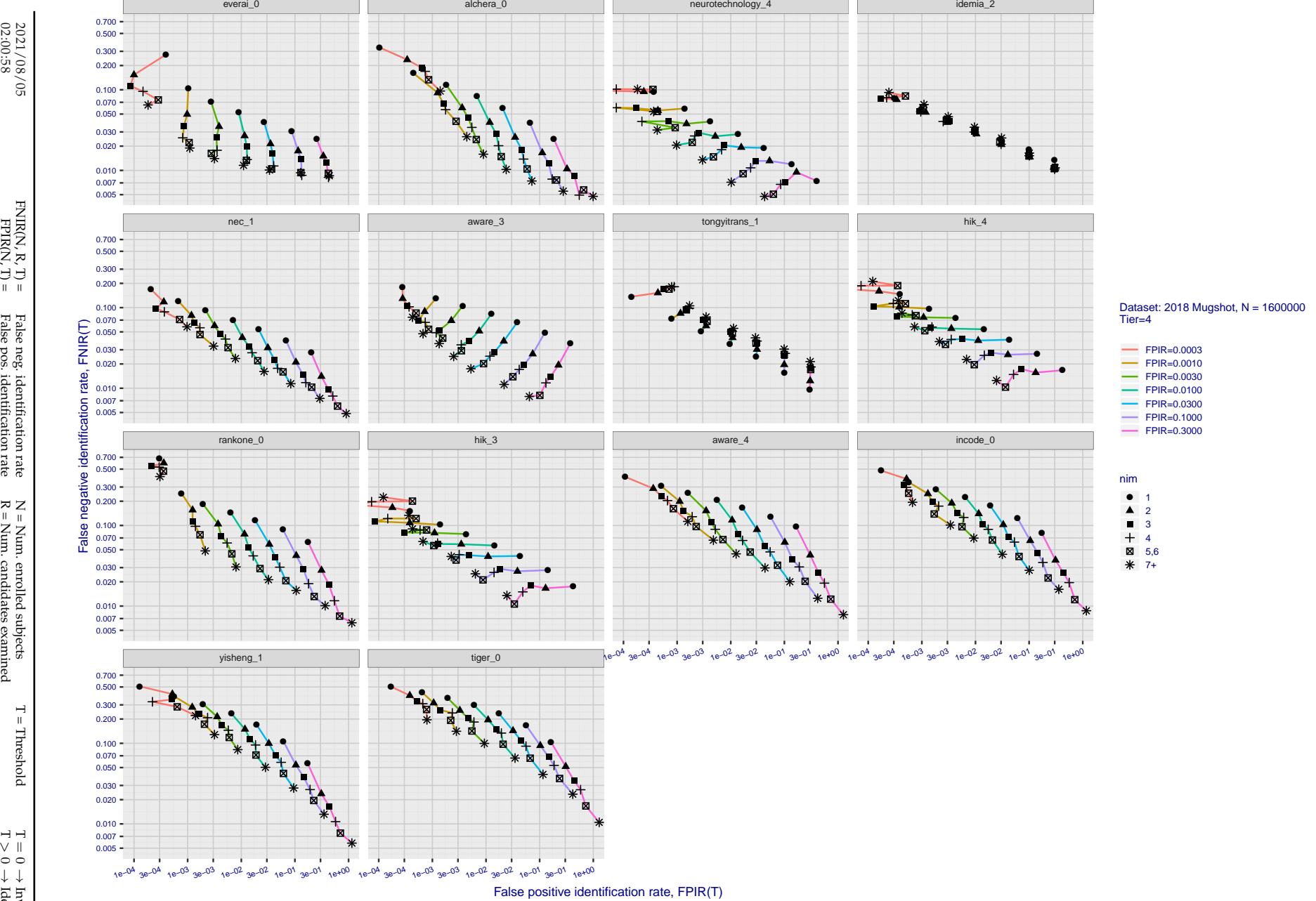


Figure 153: [FRVT-2018 Mugshot Dataset] Effect of enrolling multiple images for each identity. The plot shows an identification miss rates vs. false positive rates, at seven operating thresholds. The enrolled population size is fixed. The images are enrolled with lifetime-consolidation - see section 2.3.

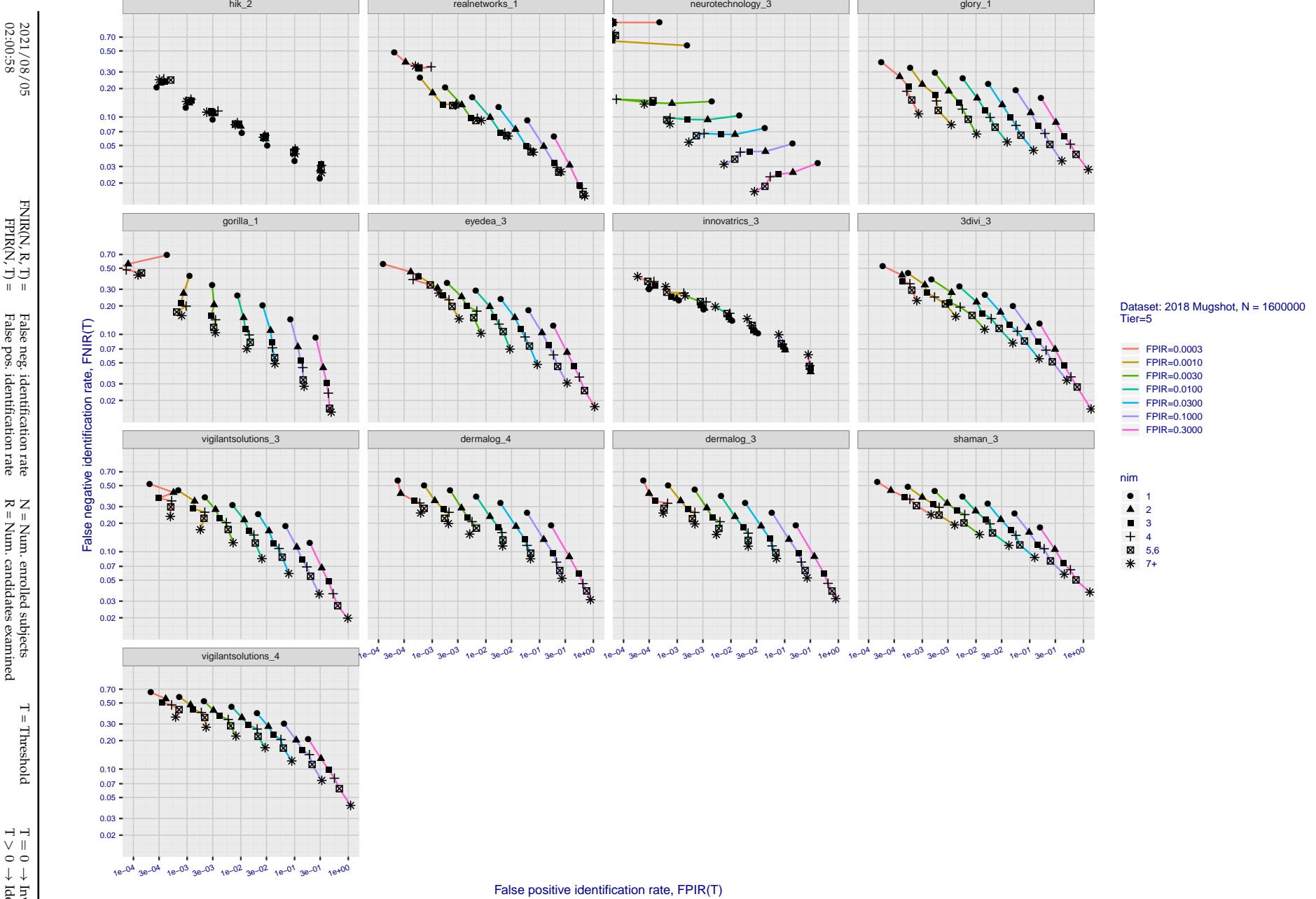


Figure 154: [FRVT-2018 Mugshot Dataset] Effect of enrolling multiple images for each identity. The plot shows an identification miss rates vs. false positive rates, at seven operating thresholds. The enrolled population size is fixed. The images are enrolled with lifetime-consolidation - see section 2.3.

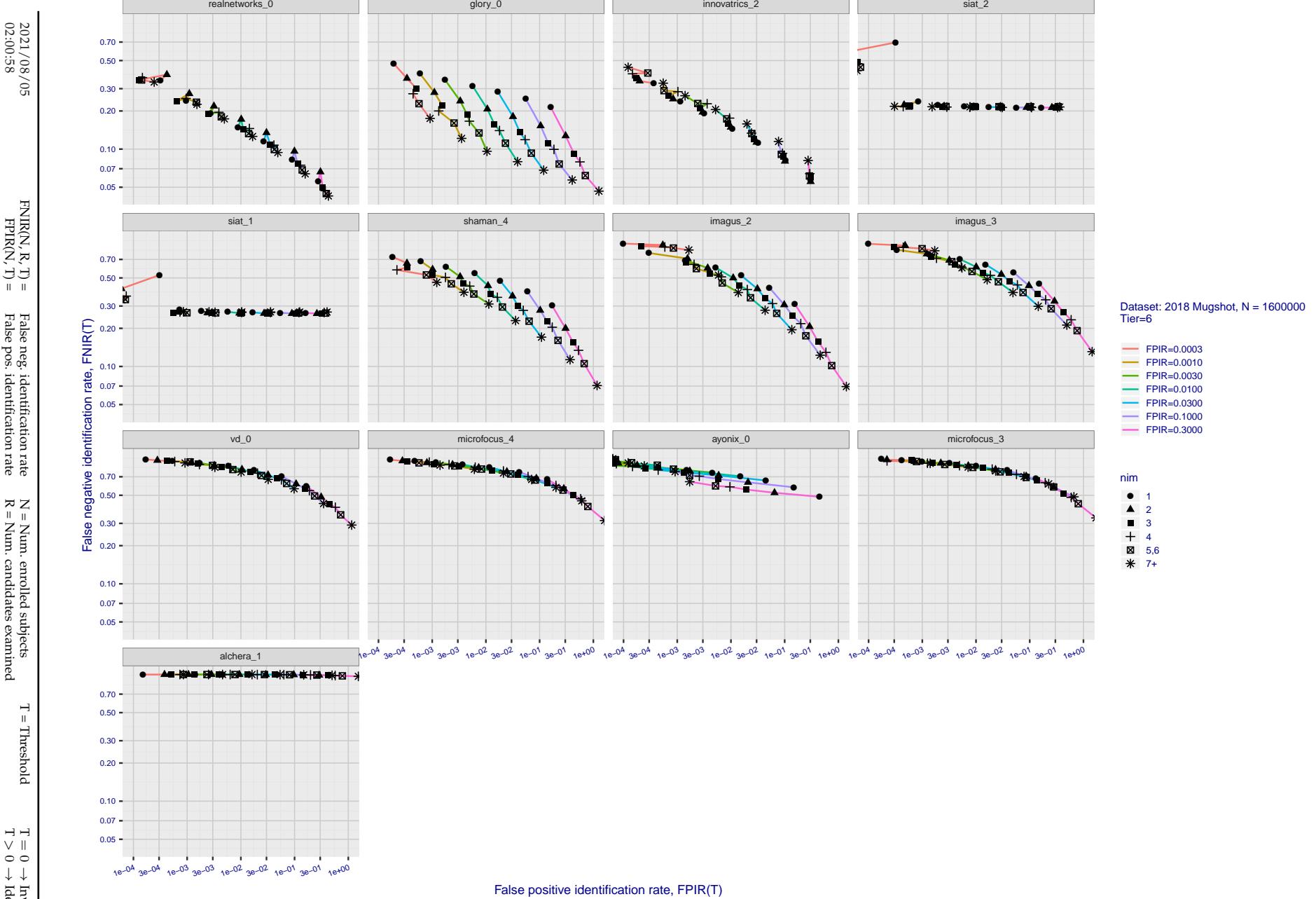
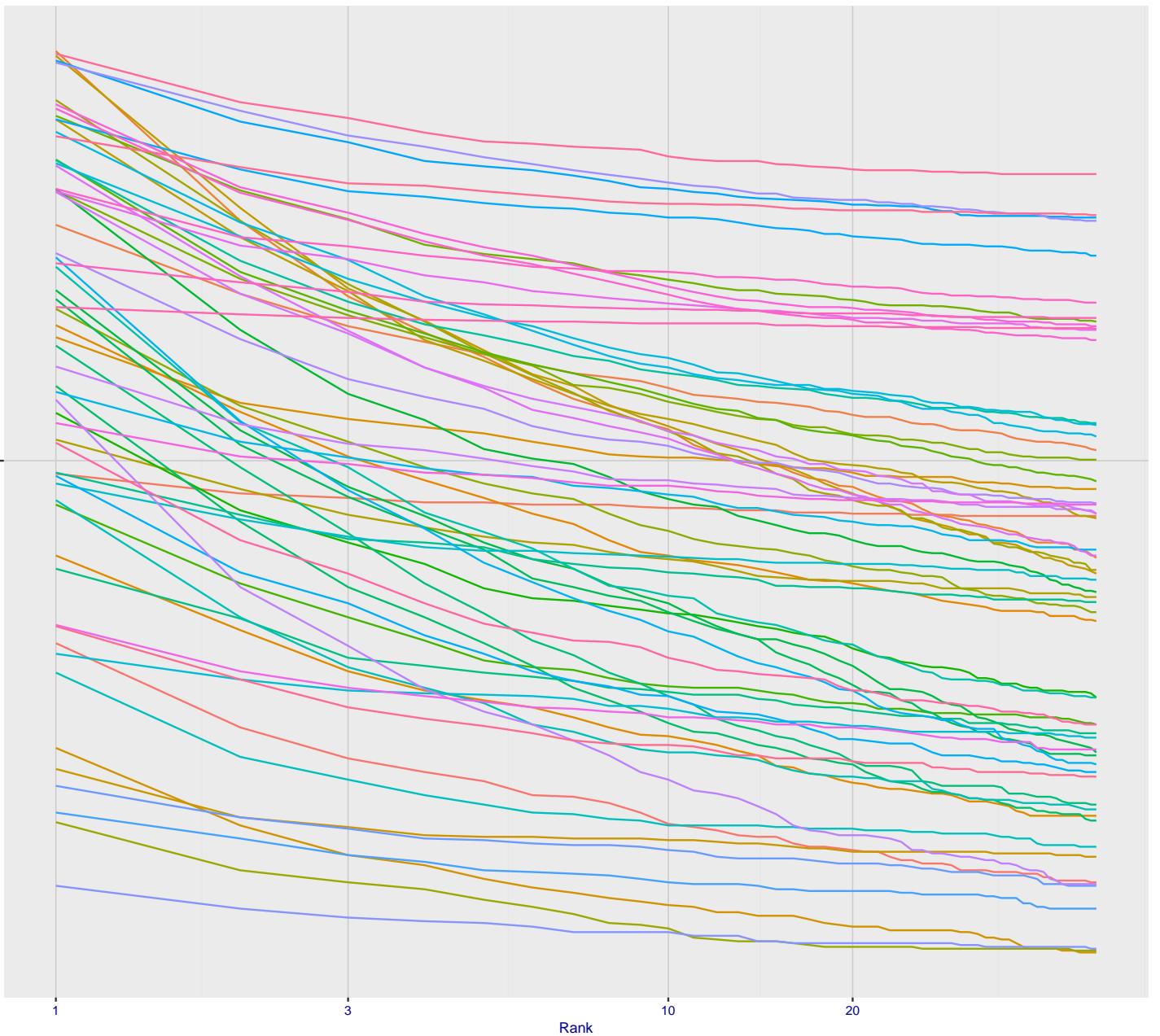


Figure 155: [FRVT-2018 Mugshot Dataset] Effect of enrolling multiple images for each identity. The plot shows an identification miss rates vs. false positive rates, at seven operating thresholds. The enrolled population size is fixed. The images are enrolled with lifetime-consolidation - see section 2.3.

Appendix D Accuracy with poor quality webcam images

2021/08/05
02:00:58 FNIR(N, R, T) = False neg. identification rate
 FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

2021/08/05 02:00:58
 FNIR(N, R, T) = False neg. identification rate
 FPR(N, T) = False pos. identification rate
 N = Num. enrolled subjects
 R = Num. candidates examined
 T = Threshold
 $T = 0 \rightarrow$ Investigation
 $T > 0 \rightarrow$ Identification



Dataset: Webcam
 Tier: 1
 FNIR(R=1, N=1600000)
 and Algorithm
 0.016 cognitec_004
 0.016 yitu_3
 0.016 deepsea_001
 0.016 sensetime_0
 0.016 sensetime_1
 0.015 idemia_007
 0.015 visionlabs_6
 0.015 visionlabs_7
 0.015 imperial_000
 0.015 dermalog_008
 0.015 rendip_000
 0.015 pixelall_004
 0.014 yitu_5
 0.014 innovatrics_005
 0.014 neurotechnology_008
 0.014 pixelall_003
 0.014 veridias_001
 0.014 line_000
 0.014 xforwardai_000
 0.014 imagus_006
 0.014 visionlabs_008
 0.014 trueface_000
 0.013 cogent_004
 0.013 synesis_005
 0.013 rankone_009
 0.013 xforwardai_001
 0.012 techlab_007
 0.012 microsoft_3
 0.012 microsoft_4
 0.012 xforwardai_002
 0.012 imagus_005
 0.012 cyberlink_002
 0.012 dahua_002
 0.011 microsoft_6
 0.011 tevian_006
 0.011 microsoft_5
 0.011 pixelall_005
 0.011 tech5_002
 0.011 kakao_000
 0.010 visionlabs_010
 0.010 everai_paravision_004
 0.010 yitu_2
 0.010 nec_3
 0.010 cloudwalk_hr_000
 0.010 rankone_010
 0.010 paravision_005
 0.010 techlab_008
 0.010 irex_000
 0.009 cyberlink_003
 0.009 nec_2
 0.008 visionlabs_009
 0.008 yitu_4
 0.008 cib_000
 0.008 paravision_007
 0.008 ntechlab_009
 0.007 dahua_003
 0.007 deepgint_001
 0.007 sensetime_004
 0.007 sensetime_003
 0.007 idemia_008
 0.006 sensetime_005

Figure 156: [Webcam Dataset] Identification miss rates vs. rank. The results apply to cross-domain recognition in which webcams are searched against enrolled mugshots. The FNIR values are higher than those for mugshot-mugshot identification due to low image resolution, lighting and less constrained subject pose in webcam images - see Figure 6.

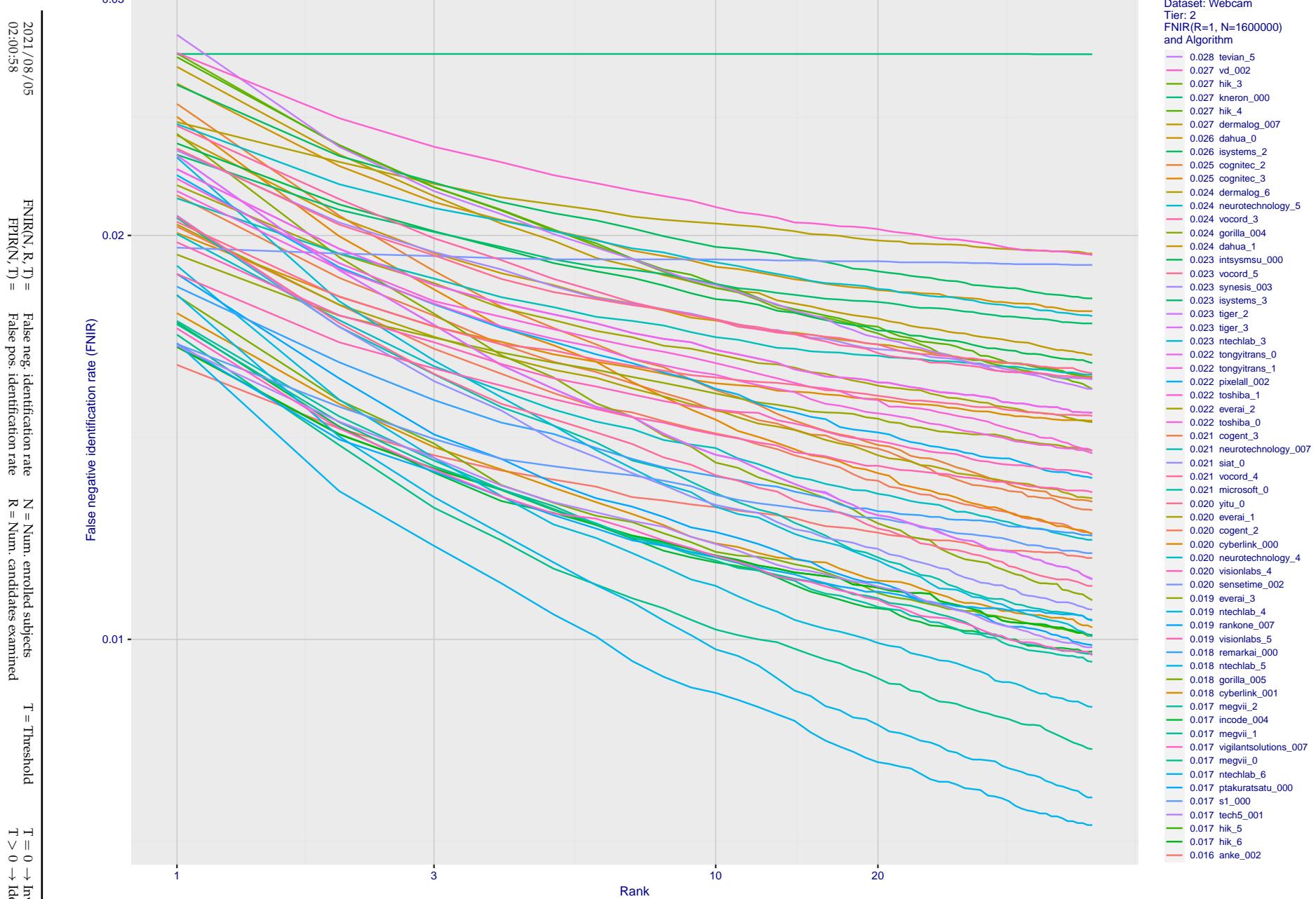


Figure 157: [Webcam Dataset] Identification miss rates vs. rank. The results apply to cross-domain recognition in which webcams are searched against enrolled mugshots. The FNIR values are higher than those for mugshot-mugshot identification due to low image resolution, lighting and less constrained subject pose in webcam images - see Figure 6.

2021/08/05
02:00:58

$\text{FNIR}(N, K, I) =$ False neg. identification rate
 $\text{FPIR}(N, T) =$ False pos. identification rate
 $N =$ Num. enrolled subjects
 $R =$ Num. candidates examined

N = Num. enrolled subjects
R = Num. candidates examined

$I = 0 \rightarrow$ Investigation
 $T > 0 \rightarrow$ Identification

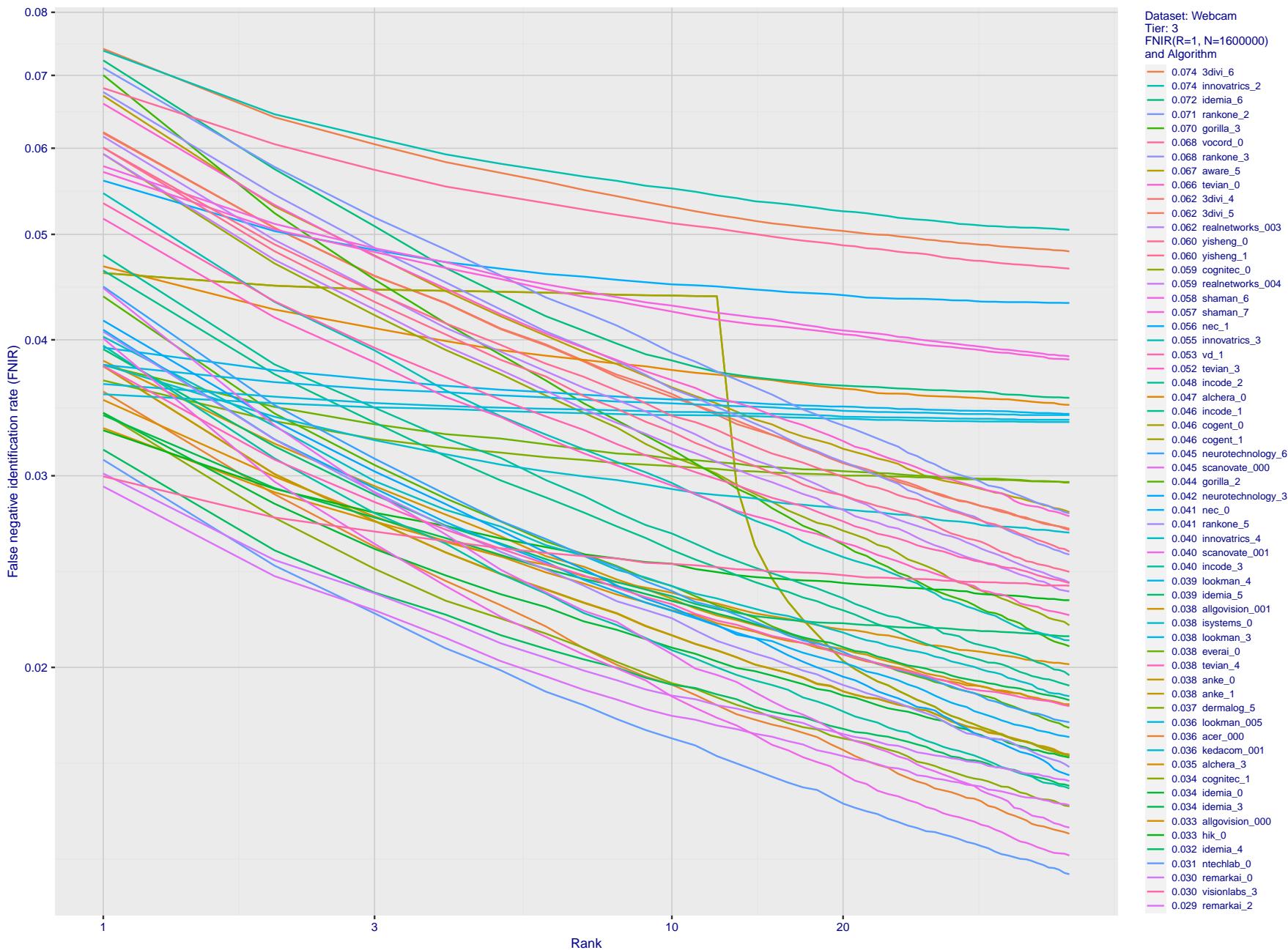


Figure 158: **[Webcam Dataset] Identification miss rates vs. rank.** The results apply to cross-domain recognition in which webcams are searched against enrolled mugshots. The FNIR values are higher than those for mugshot-mugshot identification due to low image resolution, lighting and less constrained subject pose in webcam images - see Figure 6.

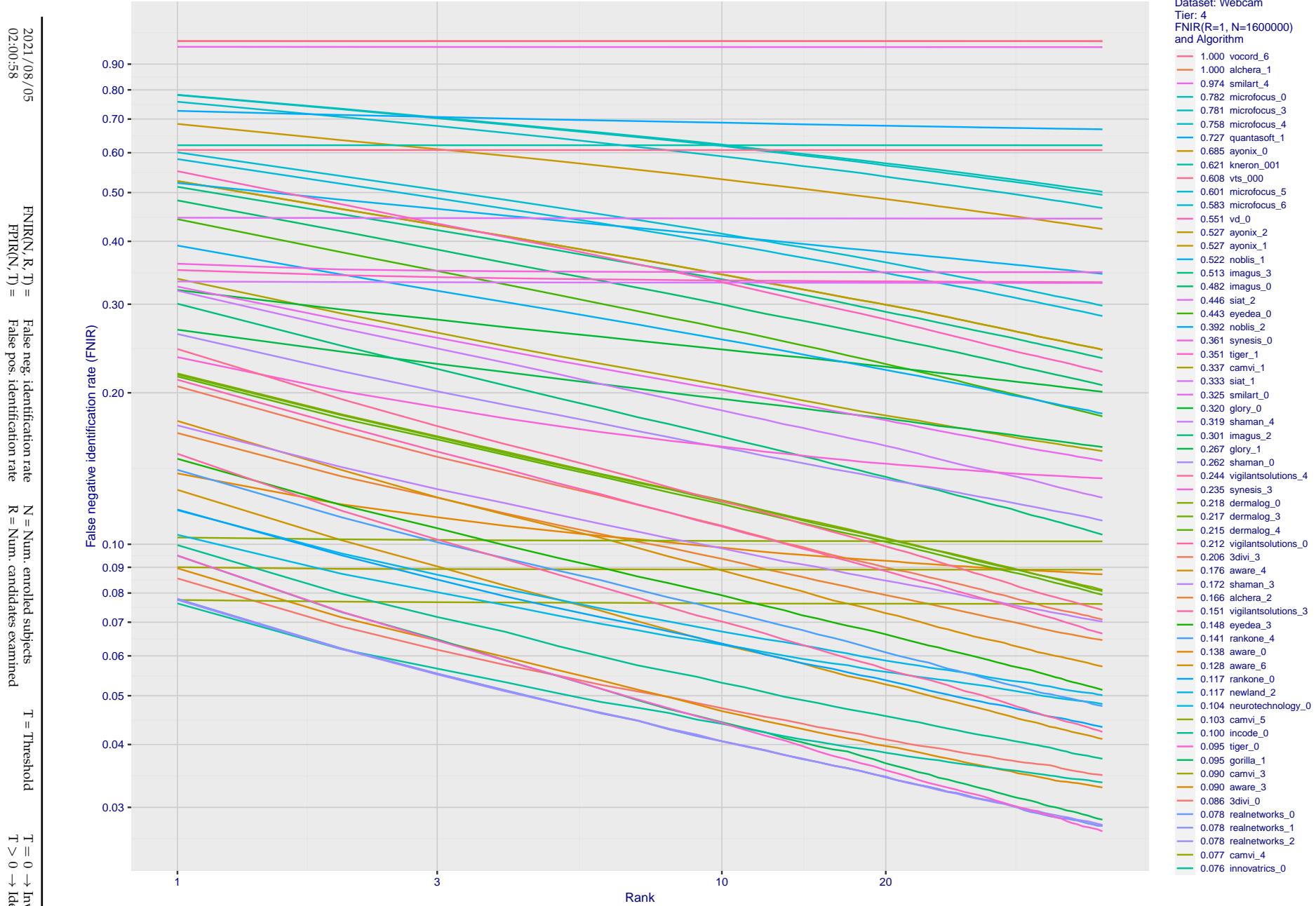


Figure 159: [Webcam Dataset] Identification miss rates vs. rank. The results apply to cross-domain recognition in which webcams are searched against enrolled mugshots. The FNIR values are higher than those for mugshot-mugshot identification due to low image resolution, lighting and less constrained subject pose in webcam images - see Figure 6.

2021/08/05
02:00:58 FNIR(N, R, T) = False neg. identification rate
 FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

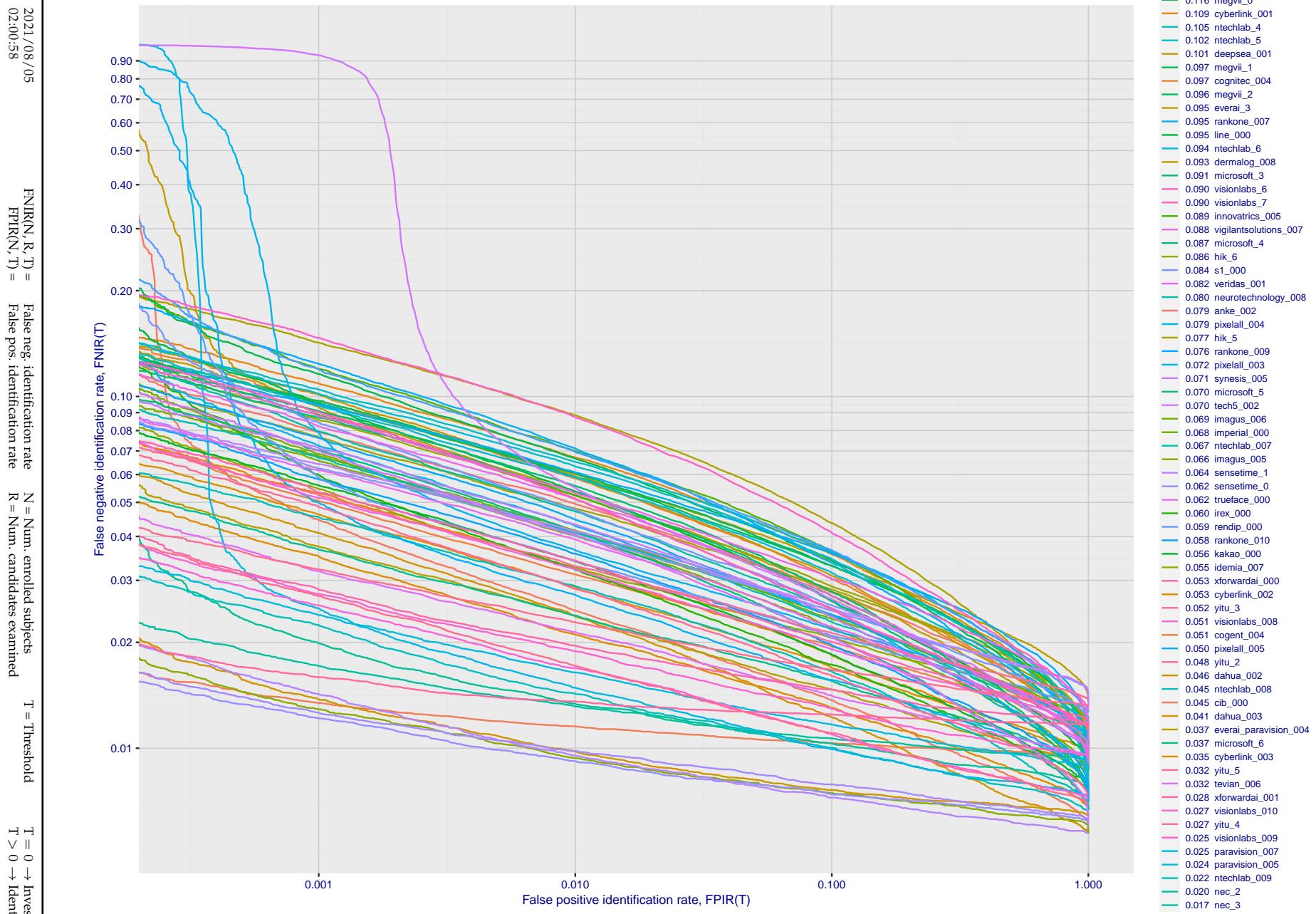


Figure 160: [Webcam Dataset] Identification miss rates vs. false positive rates. The results apply to cross-domain recognition in which webcams are searched against enrolled mugshots. The FNIR values are higher than those for mugshot-mugshot identification due to low image resolution, lighting and less constrained subject pose in webcam images - see Figure 6.

2021/08/05
02:00:58

 $\text{FNIR}(N, R, T) =$
False neg. identification rate
 $\text{FPIR}(N, T) =$
False pos. identification rate
 $N = \text{Num. enrolled subjects}$
 $R = \text{Num. candidates examined}$
 $T = \text{Threshold}$
 $T = 0 \rightarrow \text{Investigation}$
 $T > 0 \rightarrow \text{Identification}$

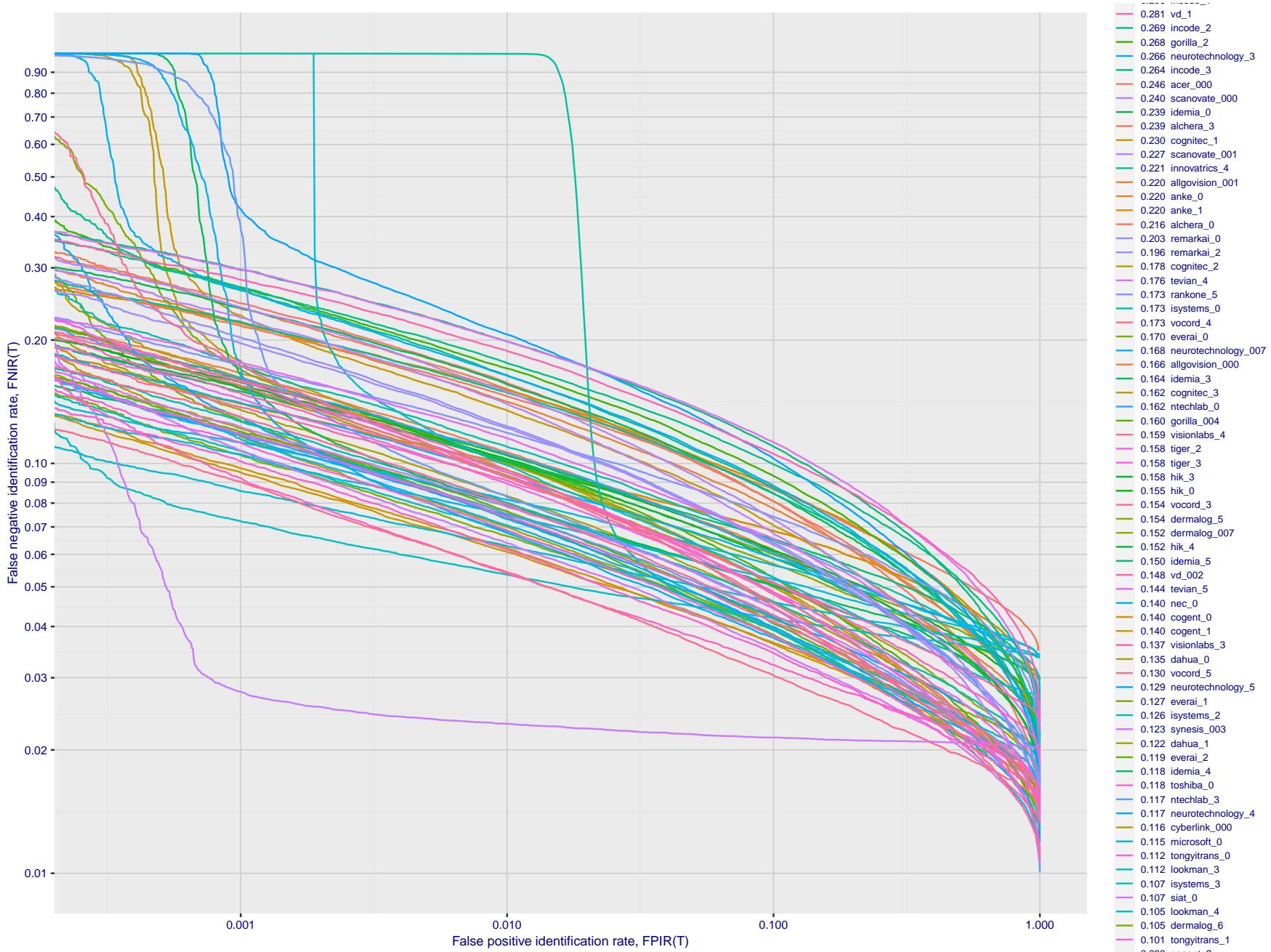


Figure 161: [Webcam Dataset] Identification miss rates vs. false positive rates. The results apply to cross-domain recognition in which webcams are searched against enrolled mugshots. The FNIR values are higher than those for mugshot-mugshot identification due to low image resolution, lighting and less constrained subject pose in webcam images - see Figure 6.

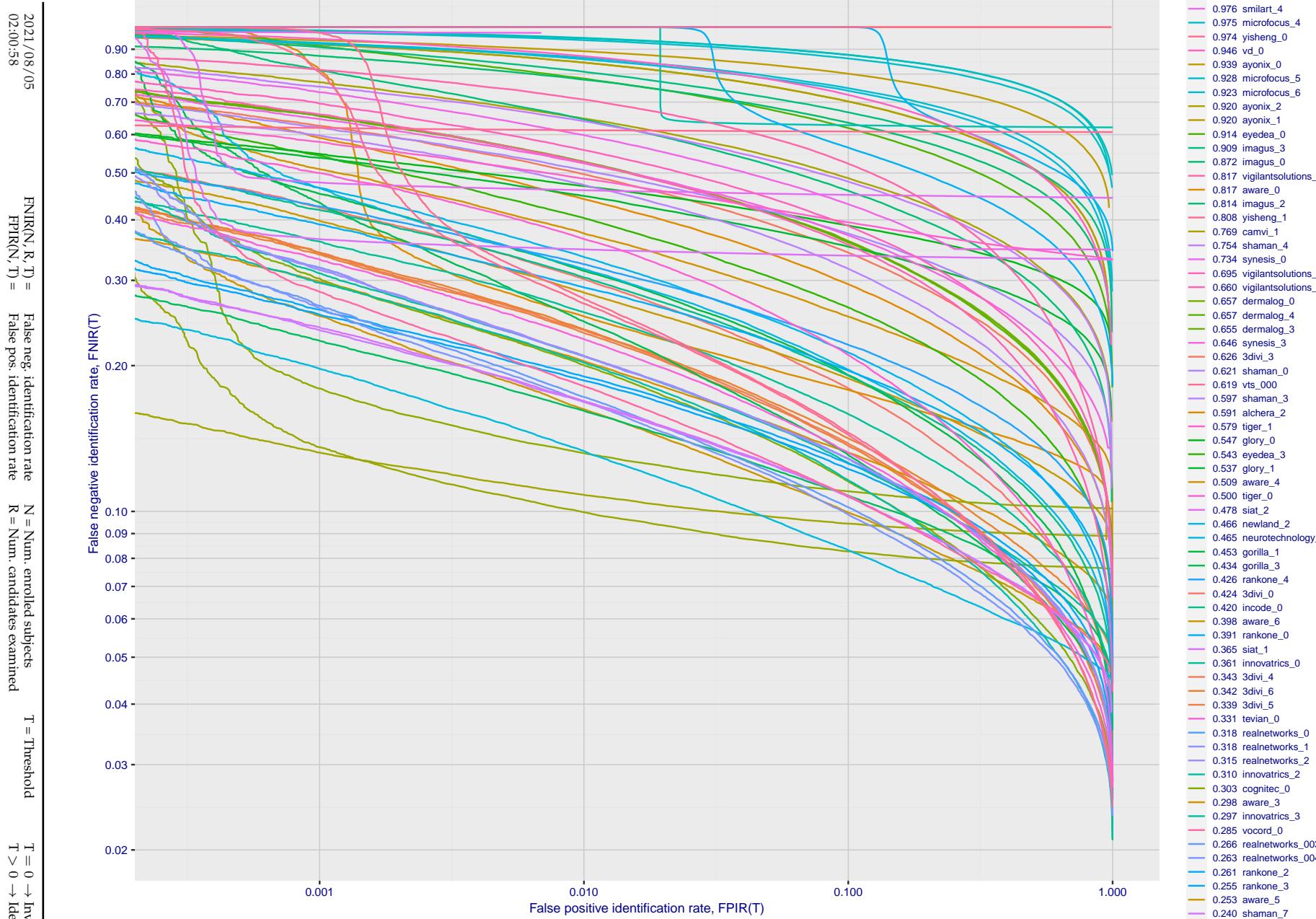


Figure 162: [Webcam Dataset] Identification miss rates vs. false positive rates. The results apply to cross-domain recognition in which webcams are searched against enrolled mugshots. The FNIR values are higher than those for mugshot-mugshot identification due to low image resolution, lighting and less constrained subject pose in webcam images - see Figure 6.

Appendix E Accuracy for profile-view to frontal recognition

Figures 163 - 165 gives accuracy results for searching 100 000 mated and 100 000 non-mated profile-view images against the same FRVT 2018 frontal enrollment dataset, $N = 1\,600\,000$, used in the main mugshot trials. This experiment corresponds to row-13 of Table 1. An example of profile-view image is given in Figure 7.

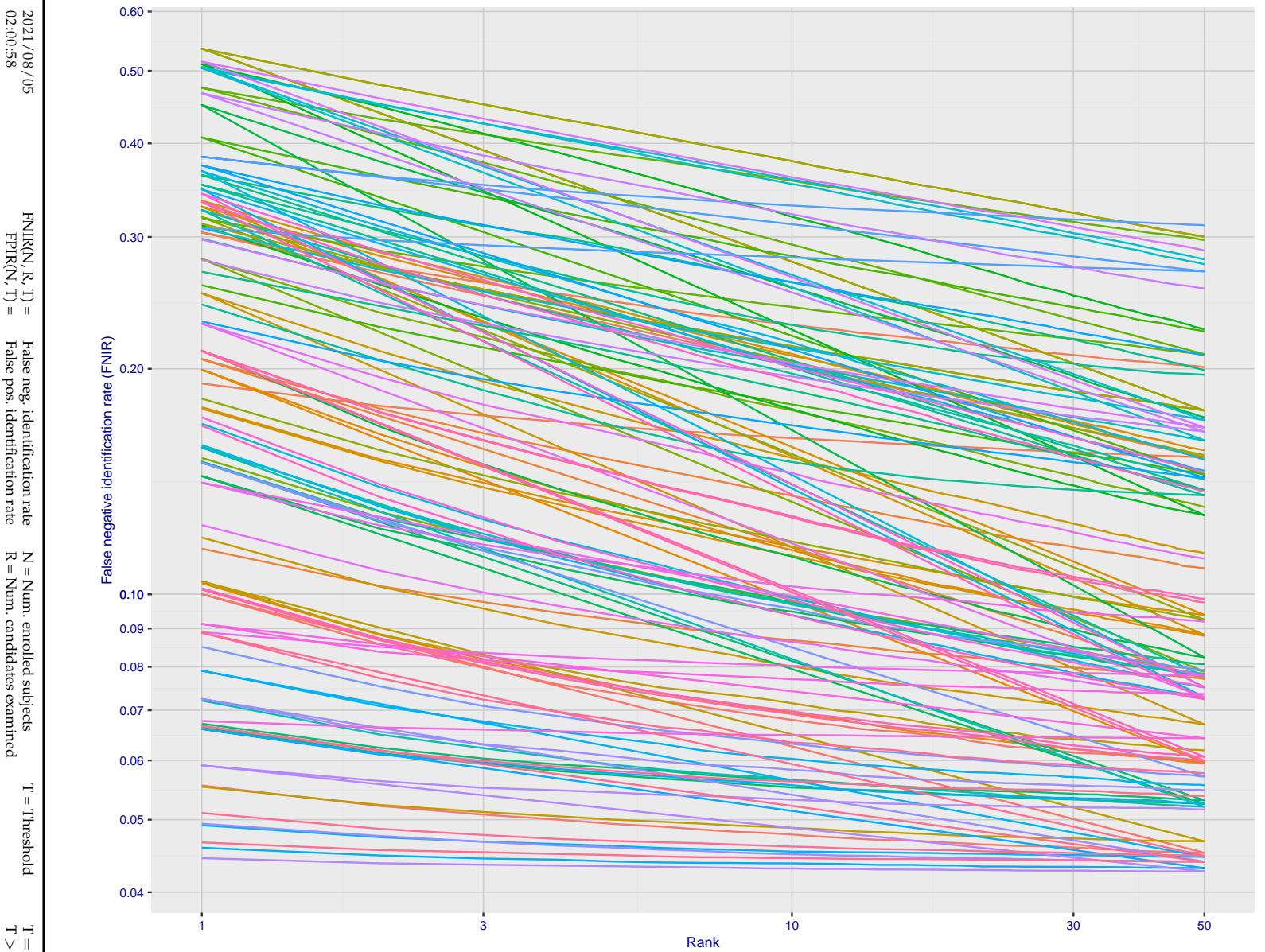


Figure 163: [Mugshot and profile-view dataset] Rank-based accuracy. For some of the more accurate Phase 3 algorithms the figure plots error tradeoff characteristics for frontal and profile-view searches into an enrolled set of $N = 1600\,000$ frontal images. Note that some algorithms fail on profile-view images with $\text{FNIR} \rightarrow 1$ - this evaluation did not ask developers to provide profile-view capability. Some algorithms, on the other hand, give FNIR approaching that for frontal-view searches using c. 2010 algorithms. The best result is that 91% of profile-view searches yield the correct mate at rank 1, and better than 94% in the top-50 candidates.

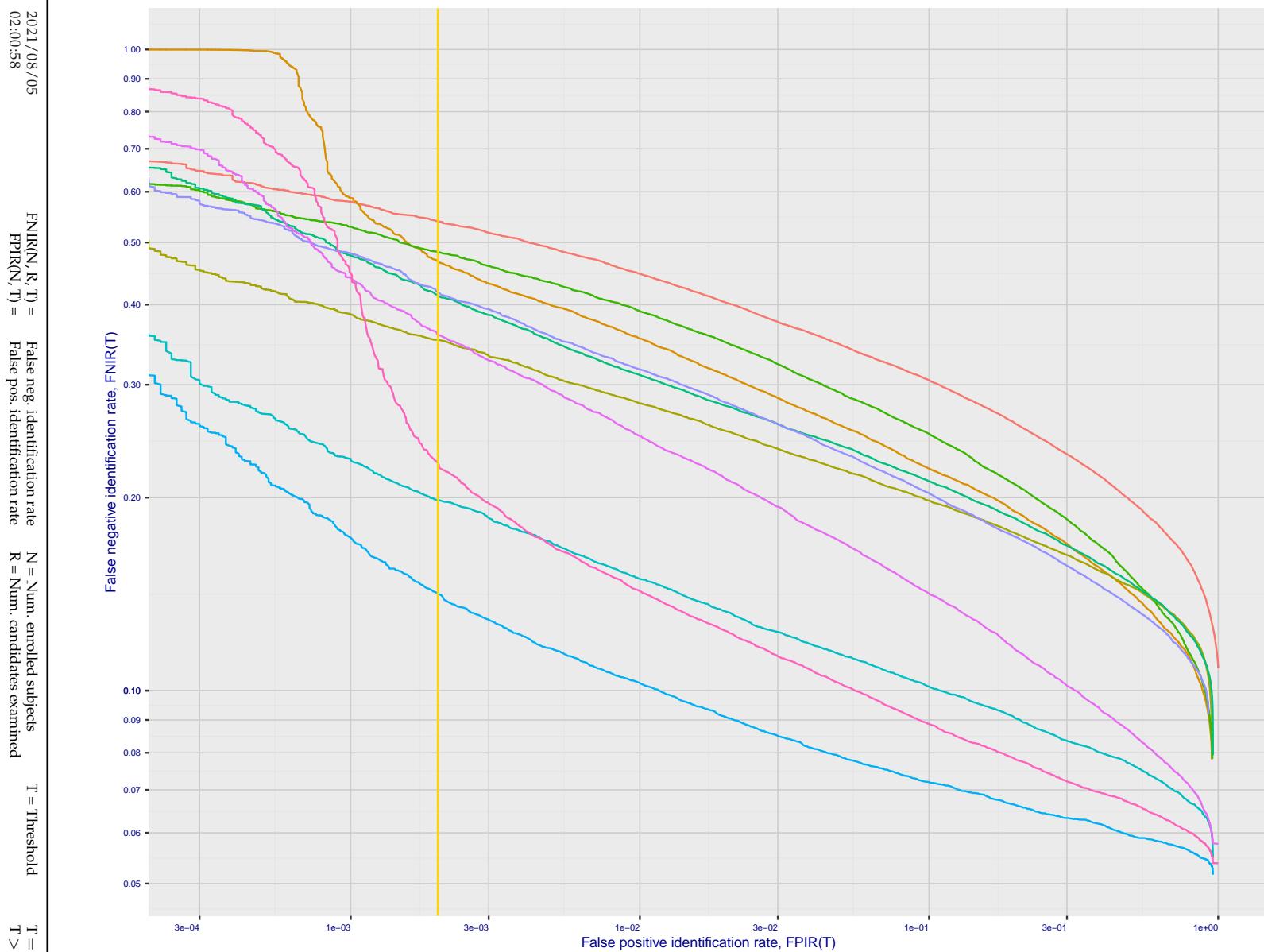


Figure 164: [Mugshot and profile-view dataset] Threshold-based accuracy. For some of the more accurate Phase 3 algorithms the figure plots error tradeoff characteristics for frontal and profile-view searches into an enrolled set of $N = 1\,600\,000$ frontal images. Note that some algorithms fail on profile-view images with $FNIR \rightarrow 1$ - this evaluation did not ask developers to provide profile-view capability. Some algorithms, on the other hand, give $FNIR$ approaching that for frontal-view searches using c. 2010 algorithms.

2021/08/05
02:00:58FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rateN = Num. enrolled subjects
R = Num. candidates examined

T = Threshold

T = 0 → Investigation
 $T > 0 \rightarrow$ Identification

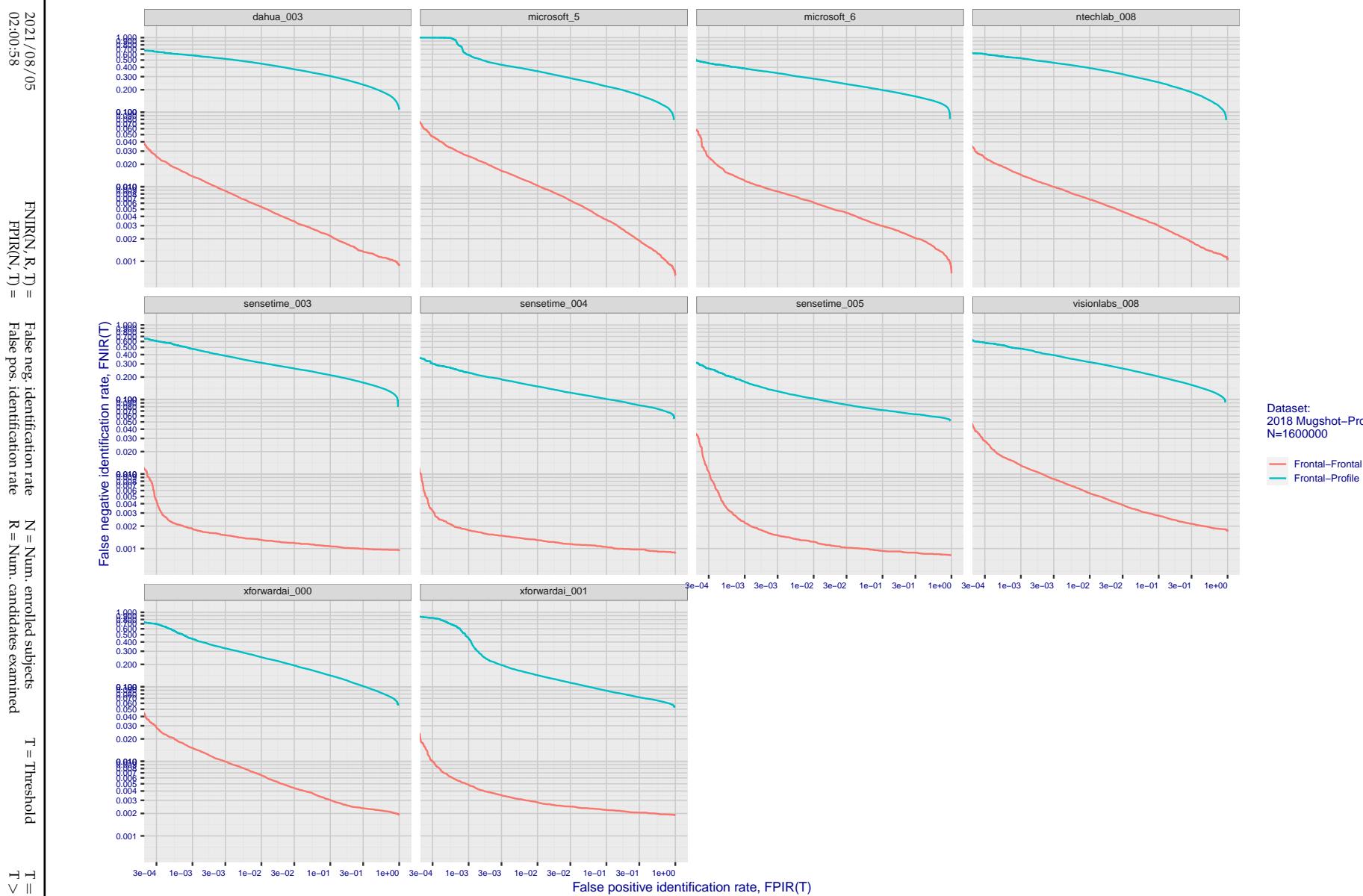


Figure 165: [Mugshot and profile-view dataset] Speed-accuracy tradeoff. For some of the more accurate Phase 3 algorithms the figure plots error tradeoff characteristics for frontal and profile-view searches into an enrolled set of $N = 1\,600\,000$ frontal images. Some algorithms fail on profile-view images with $\text{FNIR} \rightarrow 1$ - this evaluation did not ask developers to provide profile-view capability. Some algorithms, on the other hand, give FNIR approaching that for frontal-view searches using c. 2010 algorithms. Blue lines connect points of equal threshold from which it is evident that some algorithms would give markedly higher false positive outcomes if profile-view images were searched in a system configured for frontal searches. This would be a vulnerability in an access control system.

Appendix F Search duration

As in and prior tests, this section documents search speeds spanning three orders of magnitude. In applications where search volumes are high enough, this will have implications for hardware requirements especially for large N or when search duration is appreciably larger than the time it takes to prepare a template from the search image(s). Further, given very large (and growing) operational databases, the scalability of algorithms is important. It has been reported previously [8] that search duration can scale sublinearly with enrolled population size N. Further there has been considerable recent research on indexing, exact [13] and approximate nearest neighbor search [1,13] and fast-search [14,16].

Figure 166 charts the search duration measurements presented earlier in Tables 2 - 4.

- ▷ Most algorithms scale linearly. For those in that category, there is a wide range in speed with search durations ranging from 82 milliseconds for a 12 million gallery (for NEC-3) to more than 40 seconds (for Yitu-3, Toshiba-2) and even higher for less accurate algorithms.
- ▷ Some developers (Camvi, Dermalog, EverAI, Innovatrics, and Visionlabs) provide algorithms whose template search durations grow approximately logarithmically i.e. $T(N) \sim \log N$ with the constant a varying between implementations. In the figure this model is fit using the point $T(1) = 0$, and $T(640\,000)$. This very sublinear behaviour affords extremely fast search times in very large galleries. One caveat for the sublinear algorithms is that their fast-search data structures can require considerable computation time - on the order of hours - for N in the millions, and this scales mildly super-linearly, i.e. $O(N^b)$, $b > 1$. There are exceptions: the Camvi algorithms take minutes; and Innovatrics' scale sublinearly.

2021/08/05 02:00:58	$\text{FNIR}(N, R, T) =$ $\text{FPTR}(N, T) =$	False neg. identification rate False pos. identification rate	$N =$ Num. enrolled subjects $R =$ Num. candidates examined	$T =$ Threshold $T > 0 \rightarrow$ Identification	$T = 0 \rightarrow$ Investigation
------------------------	---	--	--	---	-----------------------------------

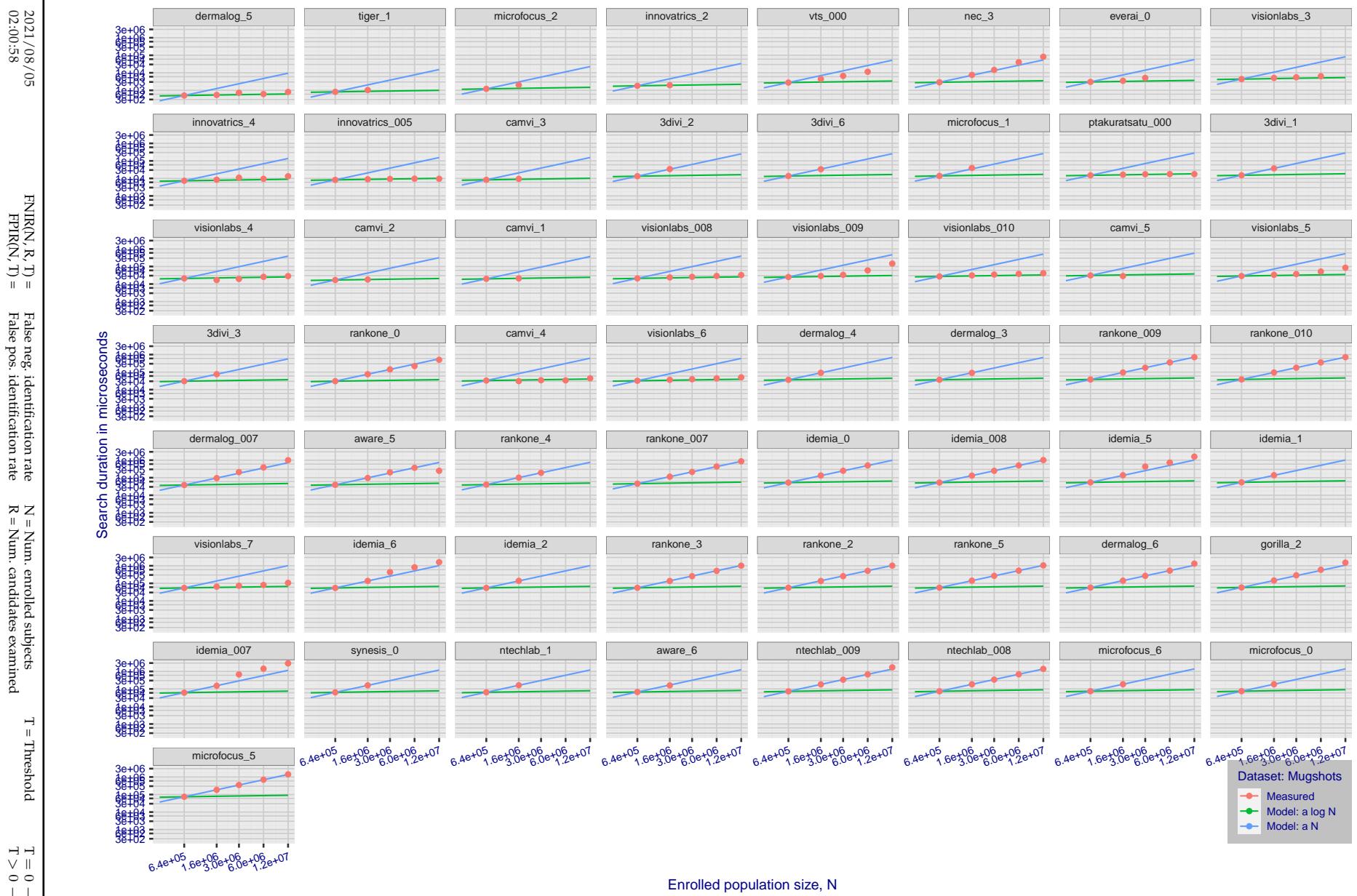


Figure 166: [Mugshot Dataset] Search duration vs. enrolled population size. In red are the actual point durations measured on a single c. 2016 core. The blue shows linear growth from $N = 640\,000$. The green line shows logarithmic growth from that point to $N = 1\,600\,000$. Note the sublinear growth from algorithms from Camvi, Dermalog, EverAI, Innovatrics, and Visionlabs. The tiger_1 algorithm is also sublinear, but inaccurate and inoperable at $N \geq 3000000$. This capability sometimes comes at the additional expense of converting a linear gallery data structure into whatever fast-search data structure is used. Note that search times are sometimes dominated by the template generation times shown in Table 20.

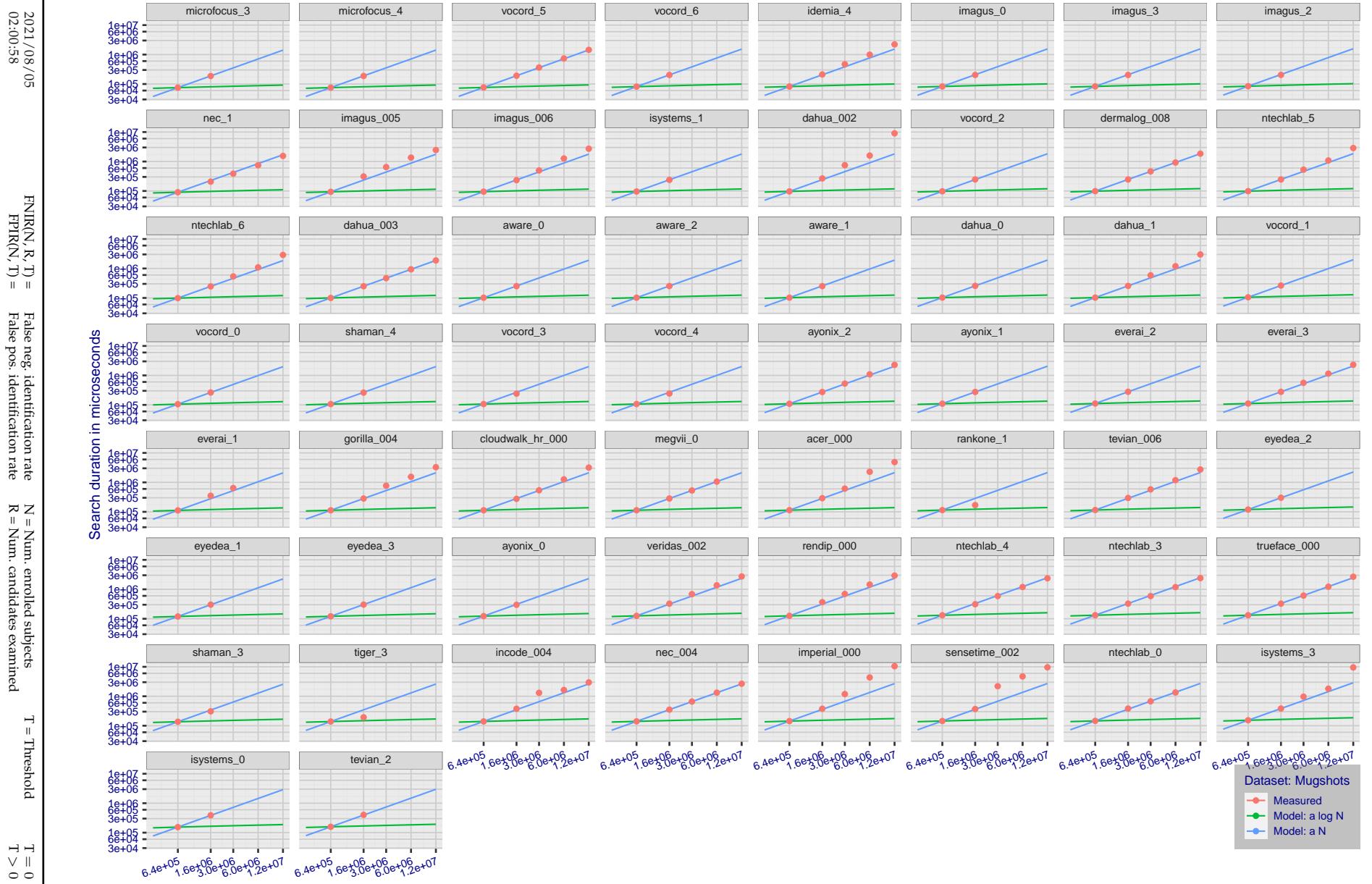


Figure 167: [Mugshot Dataset] Search duration vs. enrolled population size. In red are the actual point durations measured on a single c. 2016 core. The blue shows linear growth from $N = 640\,000$. The green line shows logarithmic growth from that point to $N = 1\,600\,000$. Note the sublinear growth from algorithms from Camvi, Dermalog, EverAI, Innovatrics, and Visionlabs. The tiger_1 algorithm is also sublinear, but inaccurate and inoperable at $N \geq 3000000$. This capability sometimes comes at the additional expense of converting a linear gallery data structure into whatever fast-search data structure is used. Note that search times are sometimes dominated by the template generation times shown in Table 20.

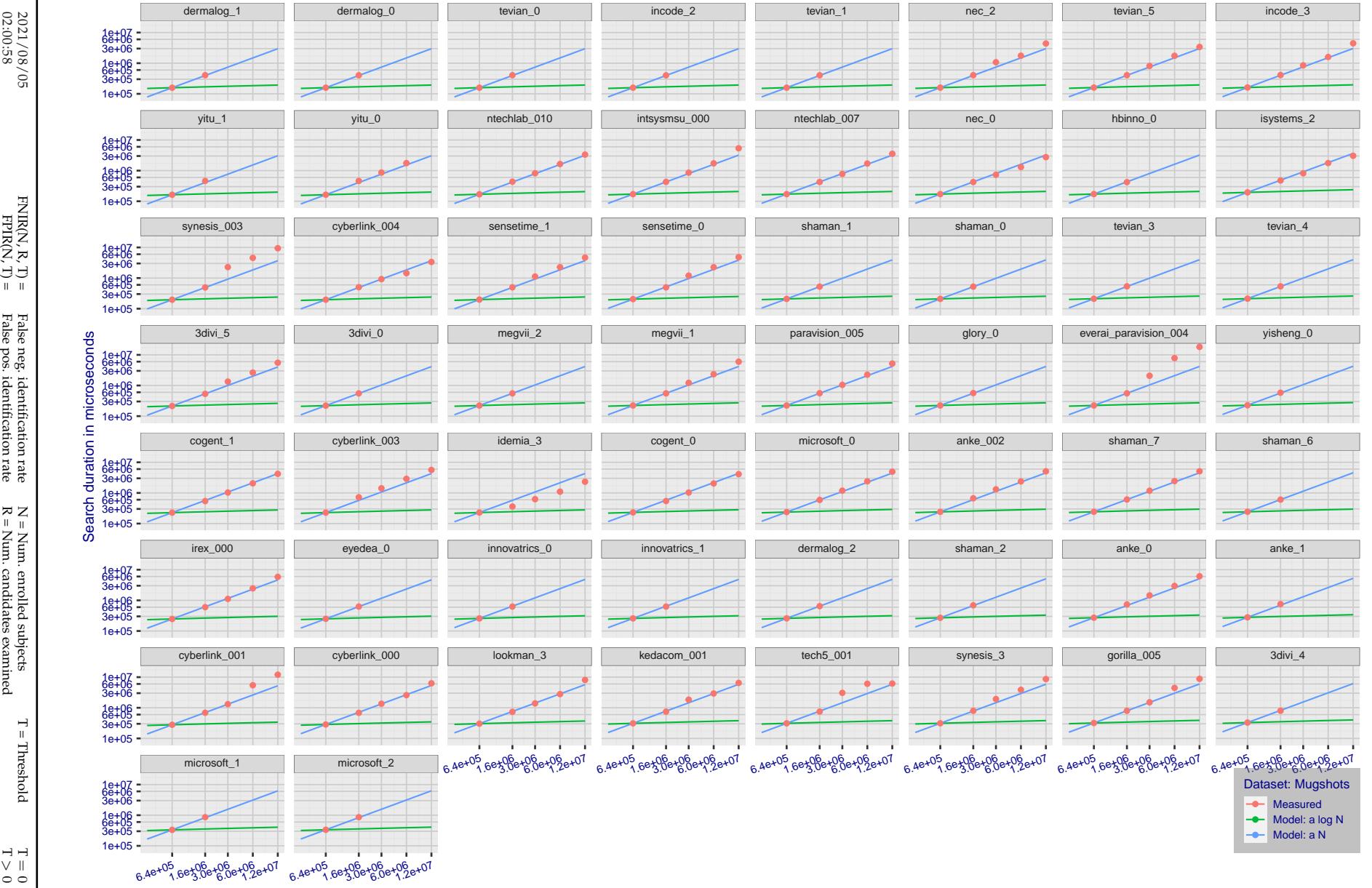


Figure 168: [Mugshot Dataset] Search duration vs. enrolled population size. In red are the actual point durations measured on a single c. 2016 core. The blue shows linear growth from $N = 640\,000$. The green line shows logarithmic growth from that point to $N = 1\,600\,000$. Note the sublinear growth from algorithms from Camvi, Dermalog, EverAI, Innovatrics, and Visionlabs. The tiger_1 algorithm is also sublinear, but inaccurate and inoperable at $N \geq 3000000$. This capability sometimes comes at the additional expense of converting a linear gallery data structure into whatever fast-search data structure is used. Note that search times are sometimes dominated by the template generation times shown in Table 20.

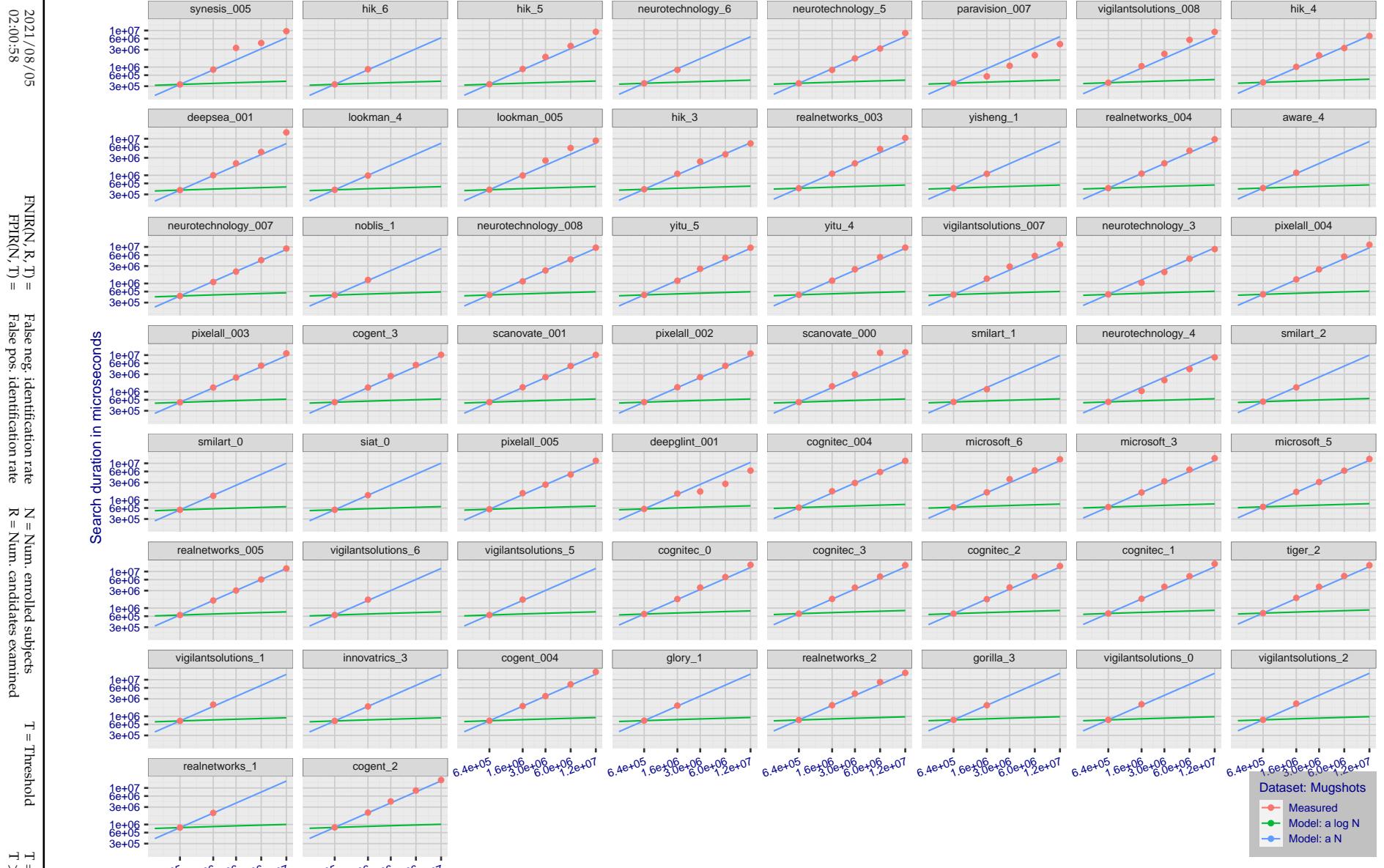


Figure 169: [Mugshot Dataset] Search duration vs. enrolled population size. In red are the actual point durations measured on a single c. 2016 core. The blue shows linear growth from $N = 640\,000$. The green line shows logarithmic growth from that point to $N = 1\,600\,000$. Note the sublinear growth from algorithms from Camvi, Dermalog, EverAI, Innovatrics, and Visionlabs. The tiger_1 algorithm is also sublinear, but inaccurate and inoperable at $N \geq 3000000$. This capability sometimes comes at the additional expense of converting a linear gallery data structure into whatever fast-search data structure is used. Note that search times are sometimes dominated by the template generation times shown in Table 20.

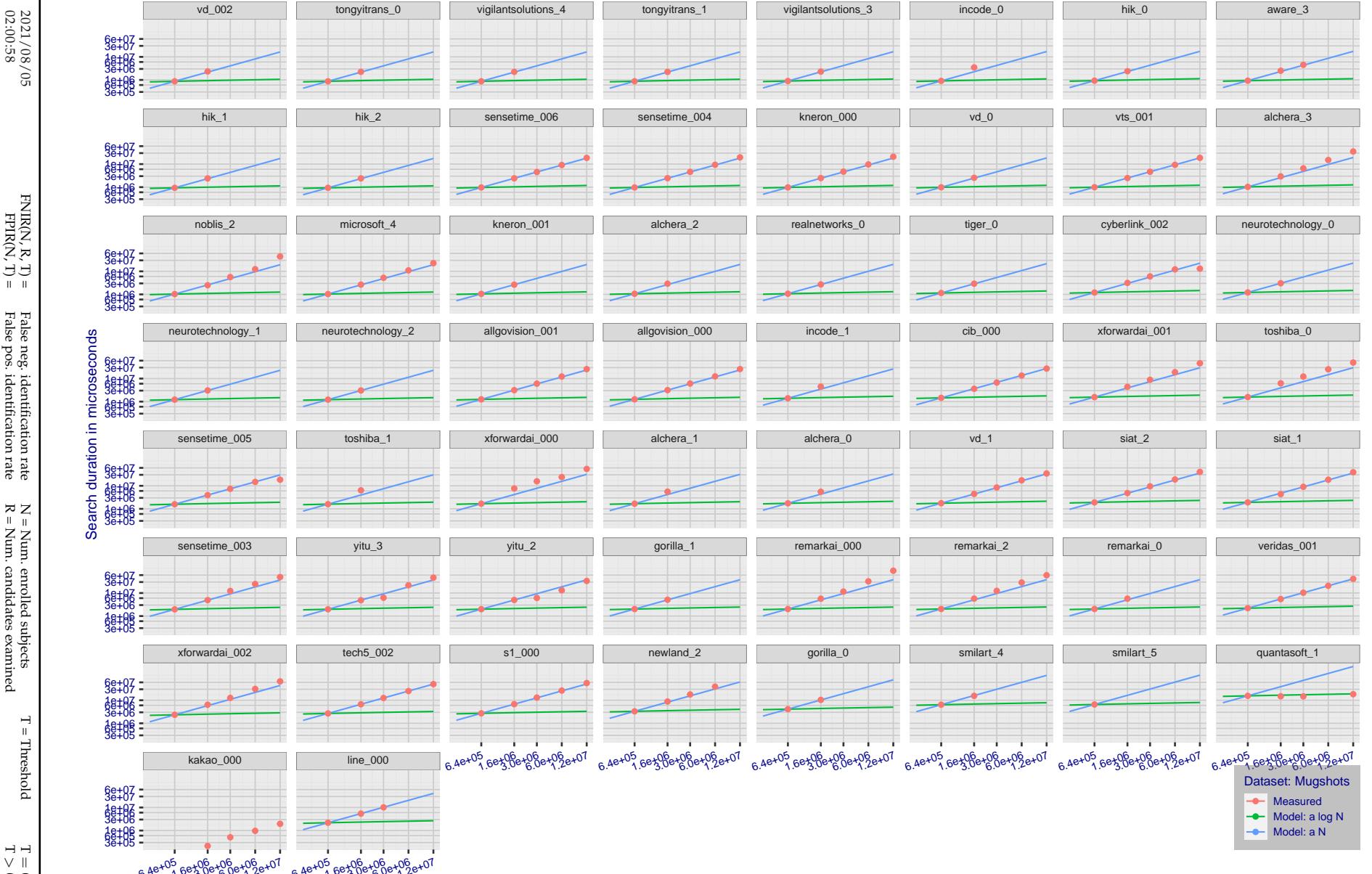


Figure 170: [Mugshot Dataset] Search duration vs. enrolled population size. In red are the actual point durations measured on a single c. 2016 core. The blue shows linear growth from $N = 640\,000$. The green line shows logarithmic growth from that point to $N = 1\,600\,000$. Note the sublinear growth from algorithms from Camvi, Dermalog, EverAI, Innovatrics, and Visionlabs. The tiger_1 algorithm is also sublinear, but inaccurate and inoperable at $N \geq 3000000$. This capability sometimes comes at the additional expense of converting a linear gallery data structure into whatever fast-search data structure is used. Note that search times are sometimes dominated by the template generation times shown in Table 20.

2021/08/05
02:00:58FNIR(N, R, T) = False neg. identification rate
FPFR(N, T) = False pos. identification rateN = Num. enrolled subjects
R = Num. candidates examined

T = Threshold

T = 0 → Investigation
 $T > 0 \rightarrow$ Identification

Appendix G Gallery Insertion Timing

2021/08/05
02:00:58FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rateN = Num. enrolled subjects
R = Num. candidates examined

T = Threshold

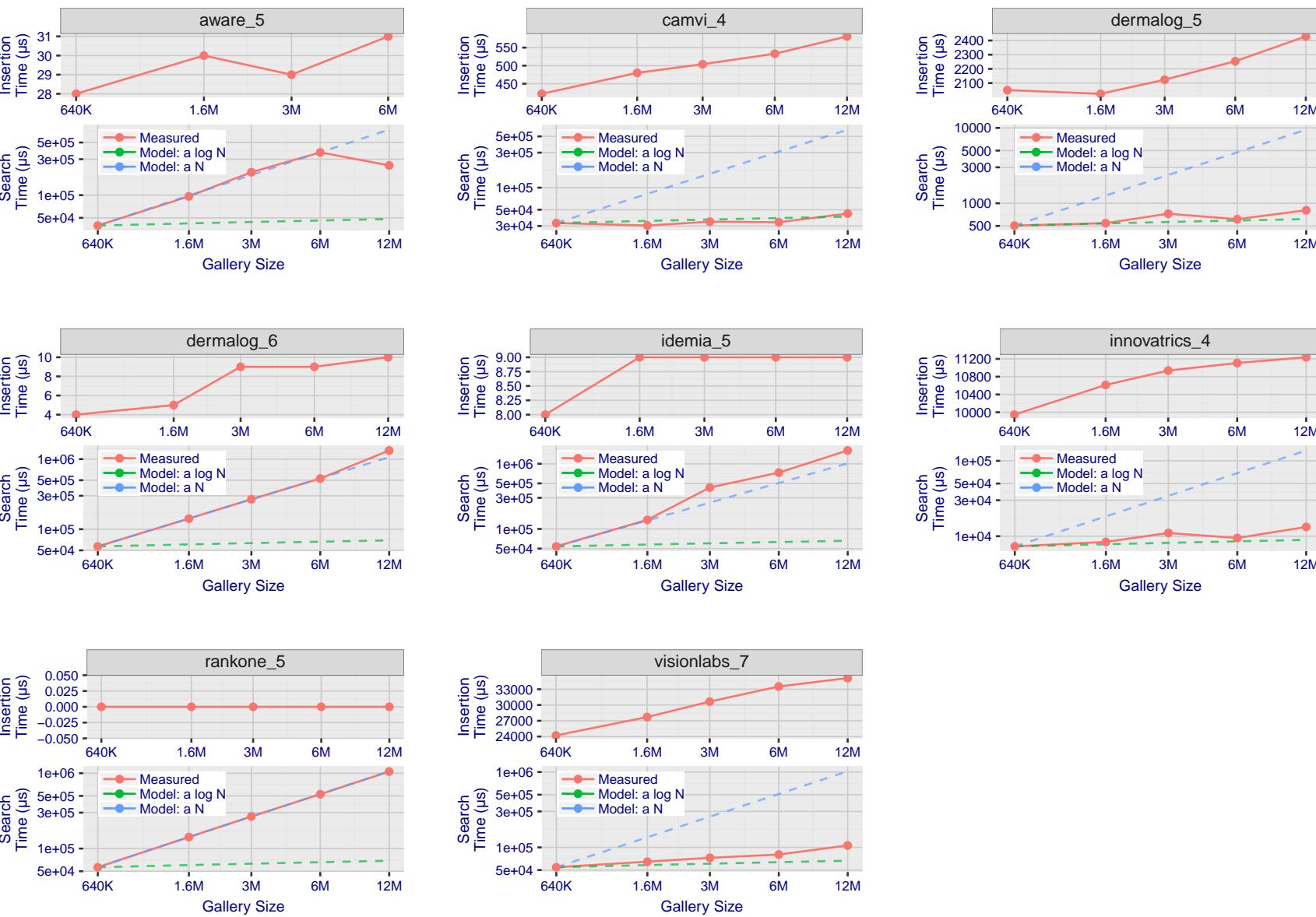
T = 0 → Investigation
T > 0 → Identification

Figure 171: [Mugshot Dataset] Gallery insertion duration vs. enrolled population size. This chart plots the time it takes to insert a single template into a finalized gallery, illustrated over increasing gallery sizes. For reference, search times on finalized galleries of corresponding sizes are plotted right underneath. Gallery insertion time plots were generated on algorithms that 1) successfully implemented gallery insertion with no errors and 2) that were run on galleries with N up to 12 000 000. Generally, only the more accurate algorithms were run on galleries with N up to 12 000 000.

2021/08/05
02:00:58FNIR(N, R, T) = False neg. identification rate
FPFR(N, T) = False pos. identification rateN = Num. enrolled subjects
R = Num. candidates examinedT = Threshold
T = 0 → Investigation

T > 0 → Identification

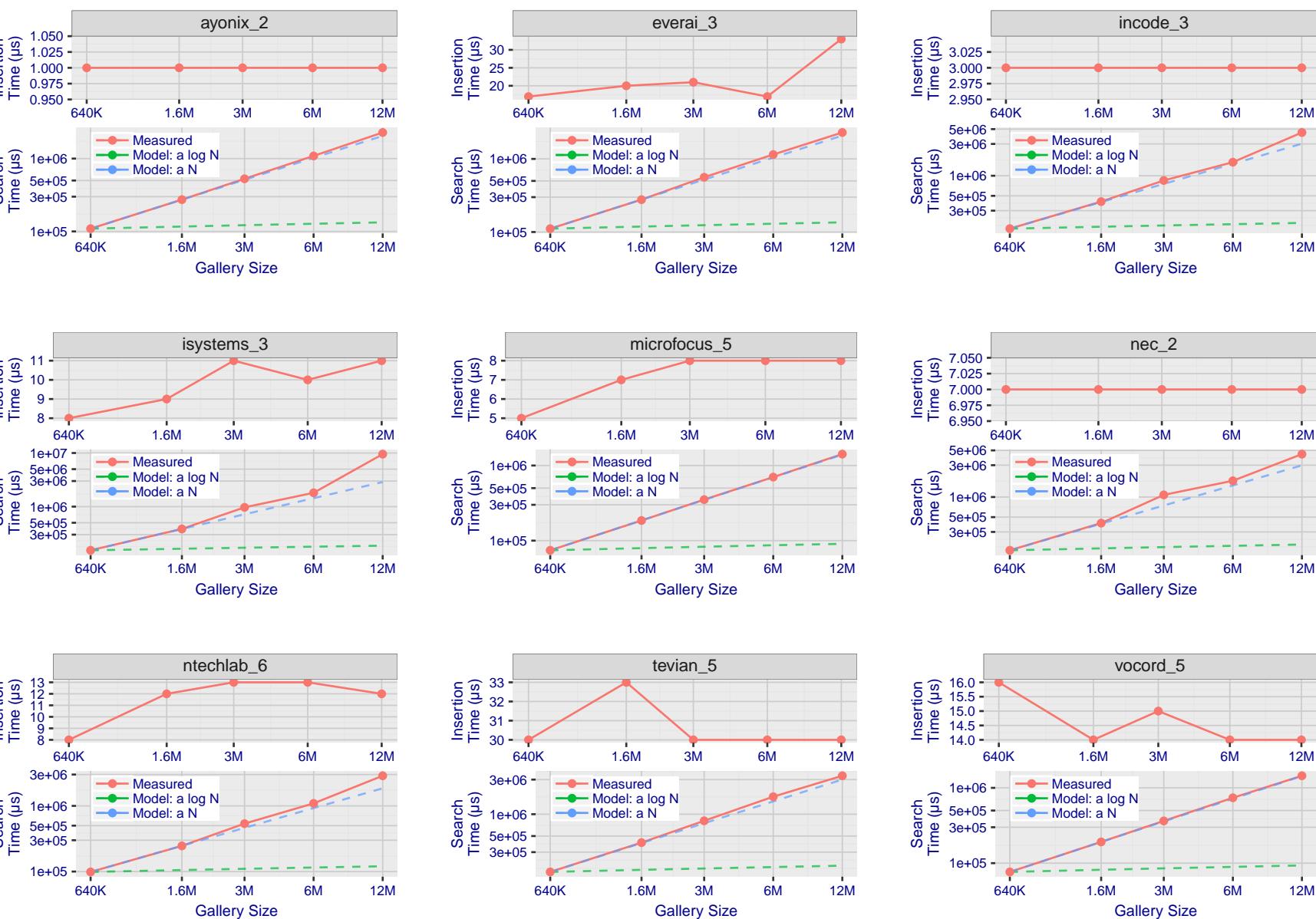


Figure 172: **[Mugshot Dataset] Gallery insertion duration vs. enrolled population size.** This chart plots the time it takes to insert a single template into a finalized gallery, illustrated over increasing gallery sizes. For reference, search times on finalized galleries of corresponding sizes are plotted right underneath. Gallery insertion time plots were generated on algorithms that 1) successfully implemented gallery insertion with no errors and 2) that were run on galleries with N up to 12 000 000. Generally, only the more accurate algorithms were run on galleries with N up to 12 000 000.

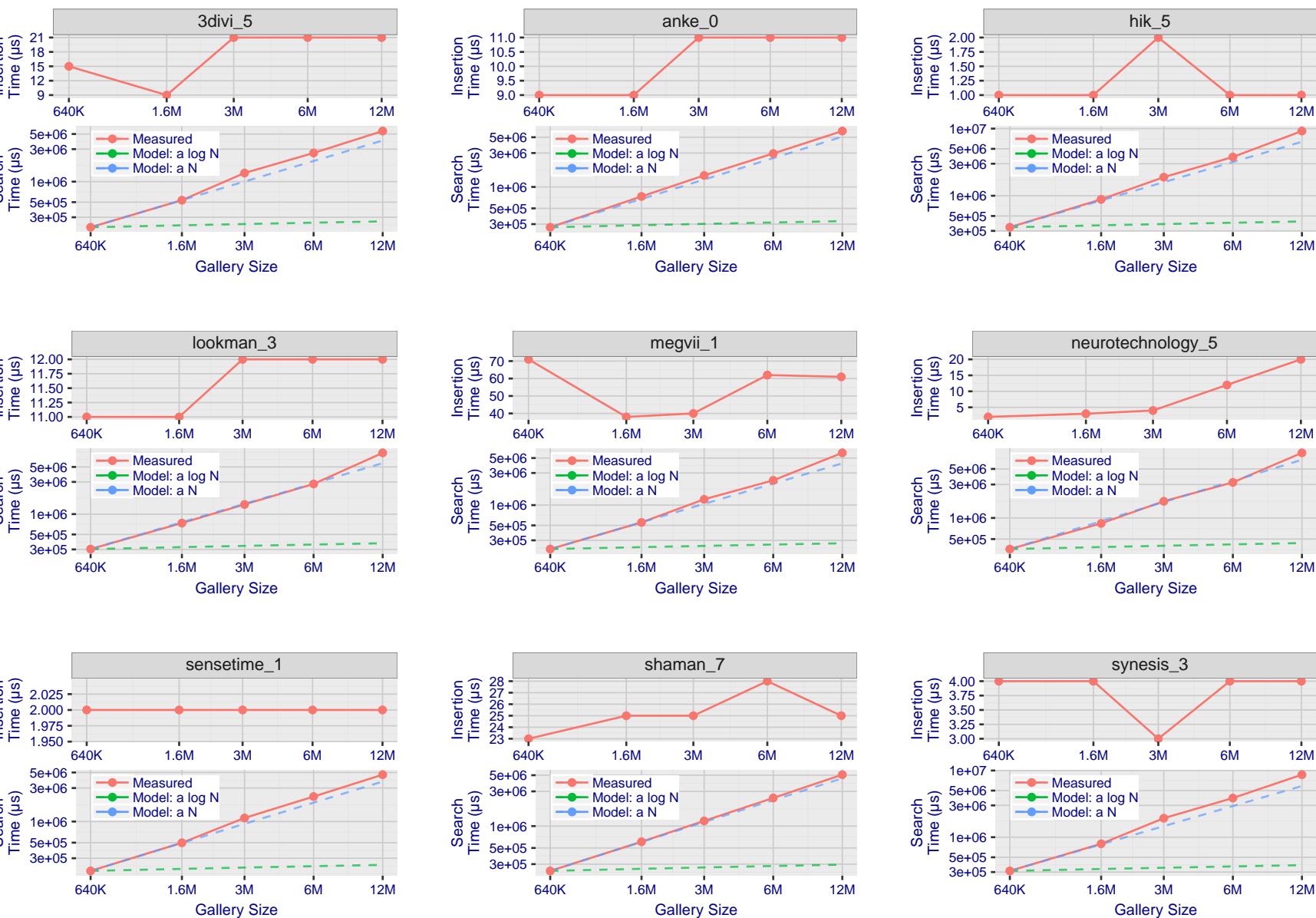
2021/08/05
02:00:58FNIR(N, R, T) = False neg. identification rate
FPTR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examinedT = Threshold
T = 0 → Investigation
 $T > 0 \rightarrow$ Identification

Figure 173: [Mugshot Dataset] Gallery insertion duration vs. enrolled population size. This chart plots the time it takes to insert a single template into a finalized gallery, illustrated over increasing gallery sizes. For reference, search times on finalized galleries of corresponding sizes are plotted right underneath. Gallery insertion time plots were generated on algorithms that 1) successfully implemented gallery insertion with no errors and 2) that were run on galleries with N up to 12 000 000. Generally, only the more accurate algorithms were run on galleries with N up to 12 000 000.

2021/08/05
02:00:58FNIR(N, R, T) = False neg. identification rate
FPTR(N, T) = False pos. identification rateN = Num. enrolled subjects
R = Num. candidates examinedT = Threshold
T = 0 → Investigation

T > 0 → Identification

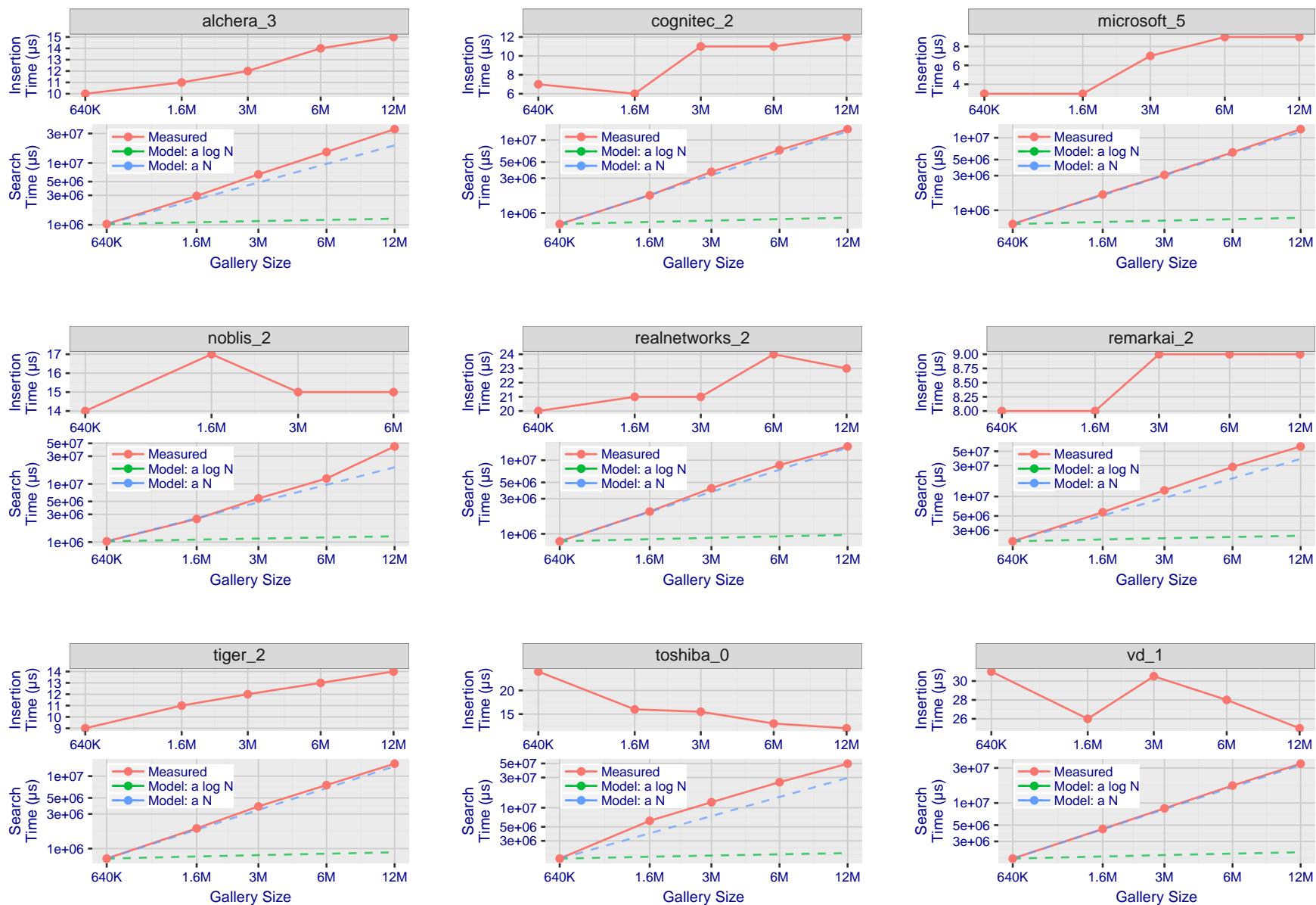


Figure 174: [Mugshot Dataset] Gallery insertion duration vs. enrolled population size. This chart plots the time it takes to insert a single template into a finalized gallery, illustrated over increasing gallery sizes. For reference, search times on finalized galleries of corresponding sizes are plotted right underneath. Gallery insertion time plots were generated on algorithms that 1) successfully implemented gallery insertion with no errors and 2) that were run on galleries with N up to 12 000 000. Generally, only the more accurate algorithms were run on galleries with N up to 12 000 000.

References

- [1] Artem Babenko and Victor Lempitsky. Efficient indexing of billion-scale datasets of deep descriptors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [2] L. Best-Rowden and A. K. Jain. Longitudinal study of automatic face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1):148–162, Jan 2018.
- [3] Blumstein, Cohen, Roth, and Visher, editors. *Random parameter stochastic models of criminal careers*. National Academy of Sciences Press, 1986.
- [4] Thomas P. Bonczar and Lauren E. Glaze. Probation and parole in the united statesm 2007, statistical tables. Technical report, Bureau of Justice Statistics, December 2008.
- [5] White D., Kemp R. I., Jenkins R., Matheson M, and Burton A. M. Passport officers' errors in face matching. *PLoS ONE*, 9(8), 2014. e103510. doi:10.1371/journal.pone.0103510.
- [6] P. Grother, G. W. Quinn, and P. J. Phillips. Evaluation of 2d still-image face recognition algorithms. NIST Interagency Report 7709, National Institute of Standards and Technology, 8 2010. <http://face.nist.gov/mbe> as MBE2010 FRVT2010.
- [7] P. J. Grother, R. J. Micheals, and P. J. Phillips. Performance metrics for the frvt 2002 evaluation. In *Proceedings of Audio and Video Based Person Authentication Conference (AVBPA)*, June 2003.
- [8] Patrick Grother and Mei Ngan. Interagency report 8009, performance of face identification algorithms. *Face Recognition Vendor Test (FRVT)*, May 2014.
- [9] Patrick Grother, George Quinn, and Mei Ngan. Face in video evaluation (five) face recognition of non-cooperative subjects. Interagency Report 8173, National Institute of Standards and Technology, March 2017. <https://doi.org/10.6028/NIST.IR.8173>.
- [10] Patrick Grother, George W. Quinn, and Mei Ngan. Face recognition vendor test - still face image and video concept, evaluation plan and api. Technical report, National Institute of Standards and Technology, 7 2013. http://biometrics.nist.gov/cs.links/face/frvt/frvt2012/NIST_FRVT2012.api_Aug15.pdf.
- [11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.
- [12] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.
- [13] Masato Ishii, Hitoshi Imaoka, and Atsushi Sato. Fast k-nearest neighbor search for face identification using bounds of residual score. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 194–199, Los Alamitos, CA, USA, May 2017. IEEE Computer Society.
- [14] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *CoRR*, abs/1702.08734, 2017.

- [15] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. *CoRR*, abs/1512.00596, 2015.
- [16] Yury A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *CoRR*, abs/1603.09320, 2016.
- [17] Joyce A. Martin, Brady E. Hamilton, Michelle J.K. Osterman, Anne K. Driscoll, , and Patrick Drake. National vital statistics reports. Technical Report 8, Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System, Division of Vital Statistics, November 2018.
- [18] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.
- [19] P. Jonathon Phillips, Amy N. Yates, Ying Hu, Carina A. Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G. Cava-zos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa, David White, and Alice J. O'Toole. Face recognition accuracy of forensic examiners, superrecognitioners, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018.
- [20] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015.
- [21] Jeroen Smits and Christiaan Monden. Twinning across the developing world. *PLOS ONE*, 6(9):1–5, 09 2011.
- [22] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '14*, pages 1701–1708, Washington, DC, USA, 2014. IEEE Computer Society.
- [23] A. Towler, R. I. Kemp, and D White. *Unfamiliar face matching systems in applied settings*. Nova Science, 2017.
- [24] Working Group 3. Ed. M. Werner. *ISO/IEC 19794-5 Information Technology - Biometric Data Interchange Formats - Part 5: Face image data*. JTC1 :: SC37, 2 edition, 2011. <http://webstore.ansi.org>.
- [25] David White, James D. Dunn, Alexandra C. Schmid, and Richard I. Kemp. Error rates in users of automatic face recognition software. *PLoS ONE*, 10:1–14, October 2015.
- [26] Bradford Wing and R. Michael McCabe. Special publication 500-271: American national standard for information systems data format for the interchange of fingerprint, facial, and other biometric information part 1. Technical report, NIST, September 2015. ANSI/NIST ITL 1-2015.
- [27] Andreas Wolf. Portrait quality - (reference facial images for mrtd). Technical report, ICAO, April 2018.
- [28] D. Yadav, N. Kohli, P. Pandey, R. Singh, M. Vatsa, and A. Noore. Effect of illicit drug abuse on face recognition. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–7, Los Alamitos, CA, USA, mar 2016. IEEE Computer Society.