U.S. Department of Homeland Security

SCIENCE AND TECHNOLOGY DIRECTORATE

Fairness, Demographic Differentials, and ISO 19795-10 Updates

Science and Technology

John J. Howard
Principal Data Scientist
Identity and Data Sciences Laboratory at
the Maryland Test Facility

Arun Vemury
Lead
Biometric and Identity Technology Center
DHS Science & Technology Directorate

November 2022

Disclaimer

- This research was funded by the U.S. Department of Homeland Security, Science and Technology Directorate on contract number 70RSAT18CB0000034.

- This work was performed by the Identity and Data Sciences Laboratory team at the Maryland Test Facility.

- The views presented here are those of the authors and do not represent those of the Department of Homeland Security, the U.S. Government, or their employers.

- The data used in this research was acquired under IRB protocol or is non-PII data that is publicly available.

Science and Technology

Biometric & Identity Technology Center

Vision

- **Drive biometric and identity innovation** at DHS through RDT&E capability
- **Facilitate and accelerate understanding of biometrics and identity technologies** for new DHS use cases
- Follow "**Build once, use widely**" approach

Goals

- **Drive efficiencies** by supporting cross cutting methods, best practices, and solutions across programs
- **Deliver Subject Matter Expertise** across the DHS enterprise
- **Engage Industry** and provide feedback
- **Encourage Innovation** with Industry and Academia

Science and Technology

Background – Fairness

- Fairness models in the broader AI community is an area of active research
  - Verma and Rubin – 20 fairness models (2018)
  - Barocas, Hardt, Narayanan – 3 fairness classes,15 fairness models (2019)
  - Mehrabi, et. al – 18 kinds of bias, 10 fairness models (2021)

- Demographic fairness in face recognition is inherently complex:
  - Multi-disciplinary (computer science, sociology, psychology, neuroscience, law)
  - Multiple error conditions (false positive, false negative)
  - The frequency of each error is weighed by some social cost that differs depending on use case
  - Across multiple, possibly intersectional, groups
  - With a final binary outcome (regulatory) or a continuous outcome (test & evaluation)
  - Not the only parameter to optimize around, i.e., accuracy

Background – The Need

- Numerous regulations already adopted or being proposed across the EU, US, Australia, and UK regarding AI generally, face recognition specifically.
  - Generally prohibit “discrimination” based on demographic category
  - Or require demographic differential performance assessments

- As of 2022, there is no standard on how to measure discrimination or fairness in biometric systems (ISO/IEC 19795-10 under development)

- However, in 2021 two fairness models were proposed by prominent research groups:
  - Fairness Discrepancy Rate (FDR) from the Swiss Idiap Institute[1]
  - Inequity Rate (IR) from the U.S. National Institute of Standards and Technology[2]

Background - Fairness

- Having two competing models in biometrics prompts several questions:
  - What are the pros and cons of each?
  - When should each one be used or not used?
  - Are there generalizable characteristics of a good fairness measure?
  - How do researchers interpret of otherwise make use of the numeric output of a fairness model?
  - What data should we use to answer these questions?

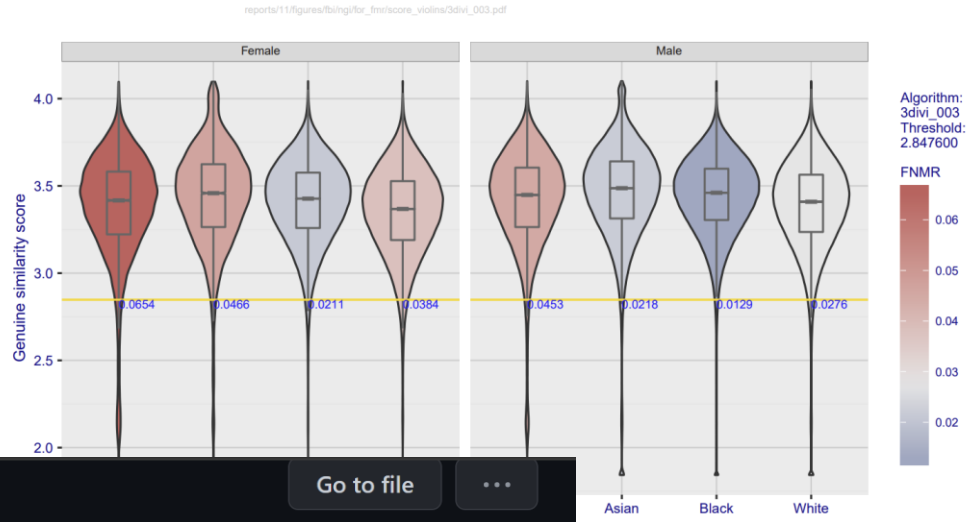
How can we answer these questions?

- Requirements of **data** to evaluate face recognition fairness models:
  - False match rates
  - False non-match rates
  - At a single threshold per algorithm (e.g. FMR = 1e-5)
  - Broken down by demographic group
  - Across a representative number of algorithms.

These data did not exist in a readily accessible manner in 2021

How can we answer these questions?

- However, it did exist
  - Annex 15 of the NIST FRVT Part 3
  - We hand transcribed these values into a machine readable dataset
  - Available on the MdTF GitHub Page:



master mdtf-public / datasets / nist-frvt-annex15 / nist-frvt3-annex15-data-flat.csv Go to file

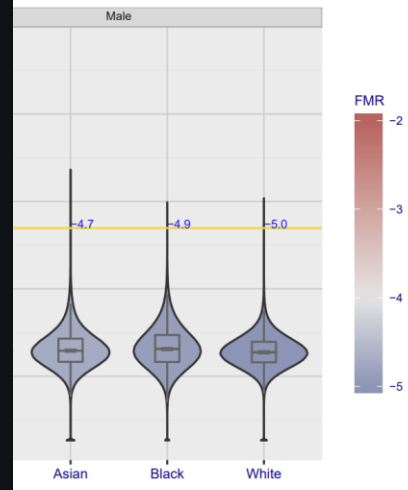
JH minor edits, lower case folder name, README updates Latest commit 095d4a6 on Mar 9 History

0 contributors

Executable File | 127 lines (127 sloc) | 23.5 KB Raw Blame

Search this file...

	Algorithm	FNMR.F.AmIndian	FMR.F.AmIndian	FNMR.F.Asian	FMR.F.Asian	FNMR.F.Black	FMR.F.Black	FNMR.F.White	FMR.F.White	FNMR.M.AmIndian	FI
2	cyberextruder-002	0.0909	0.000398107170553497	0.0709	7.94328234724282e-05	0.0699	0.000199526231496888	0.0824	1.25892541179417e-05	0.0793	6.
3	didiglobalface-001	0.0033	0.000501187233627273	0.0052	3.16227766016838e-05	0.0011	7.94328234724282e-05	0.0021	1e-05	0.005	0.
4	everai.paravision-003	0.0072	0.000501187233627273	0.0066	1.58489319246111e-05	0.0012	1e-04	0.0028	1e-05	0.0066	0.



Study the Behavior of Fairness Models

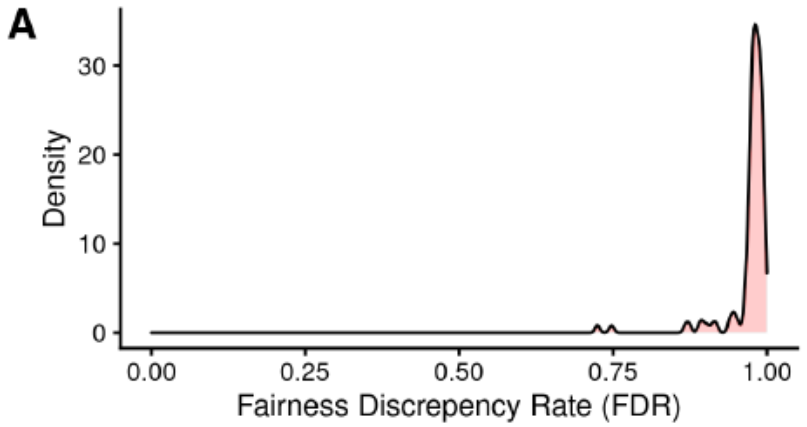
- Idiap Fairness Discrepancy Rate:

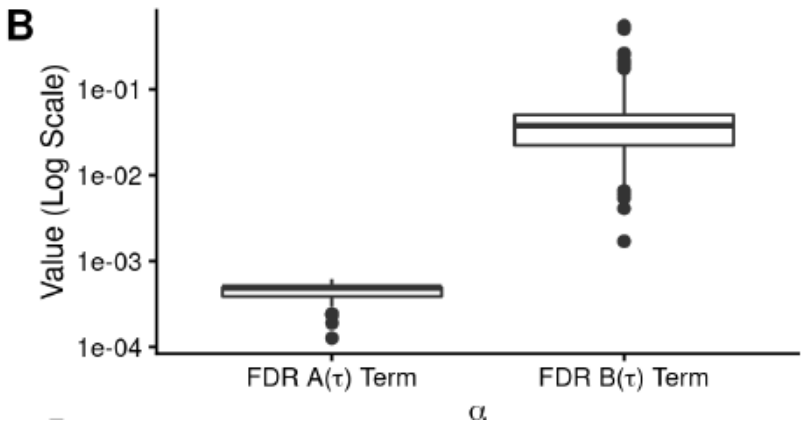
$$A(\tau) = max(|FMR_{d_i}(\tau) - FMR_{d_j}(\tau)|) \;\; \forall d_i, d_j \in D \qquad (1)$$

$$B(\tau) = max(|FNMR_{d_i}(\tau) - FNMR_{d_j}(\tau)|) \;\; \forall d_i, d_j \in D \qquad (2)$$

$$FDR(\tau) = 1 - (\alpha A(\tau) + (1 - \alpha)B(\tau)) \qquad (3)$$

- Straightforward, bounded from 0-1
- But in real world applications, FMR and FNMR exist on very different scales:





Science and Technology

Study the Behavior of Fairness Models

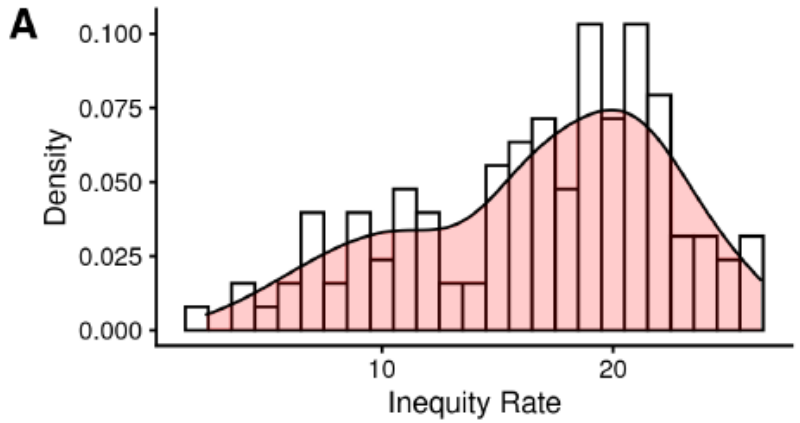
- NIST Inequity Rate:

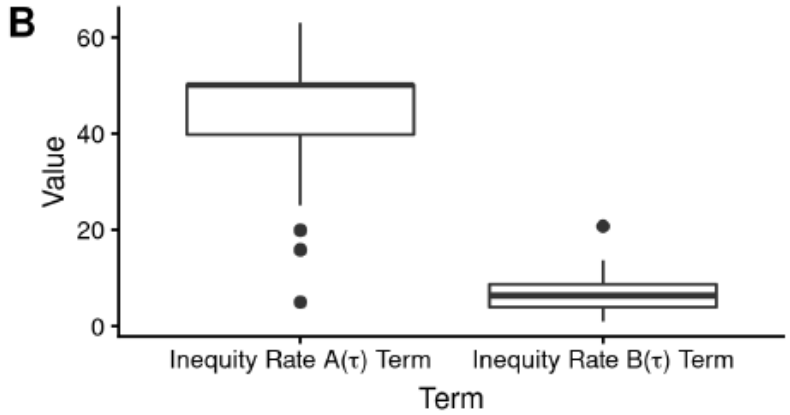
$$A(\tau) = \frac{\max_{d_i} FMR_{d_i}(\tau)}{\min_{d_j} FMR_{d_j}(\tau)} \quad \forall d_i, d_j \in D \qquad (4)$$

$$B(\tau) = \frac{\max_{d_i} FNMR_{d_i}(\tau)}{\min_{d_j} FNMR_{d_j}(\tau)} \quad \forall d_i, d_j \in D \qquad (5)$$

$$IR = A(\tau)^{\alpha} B(\tau)^{1-\alpha} \qquad (6)$$

- Intuitive, straightforward, overcomes many of the issues with FDR





- Issues: 1) unbounded 2) undefined in the presence of 0% error rate

Science and Technology

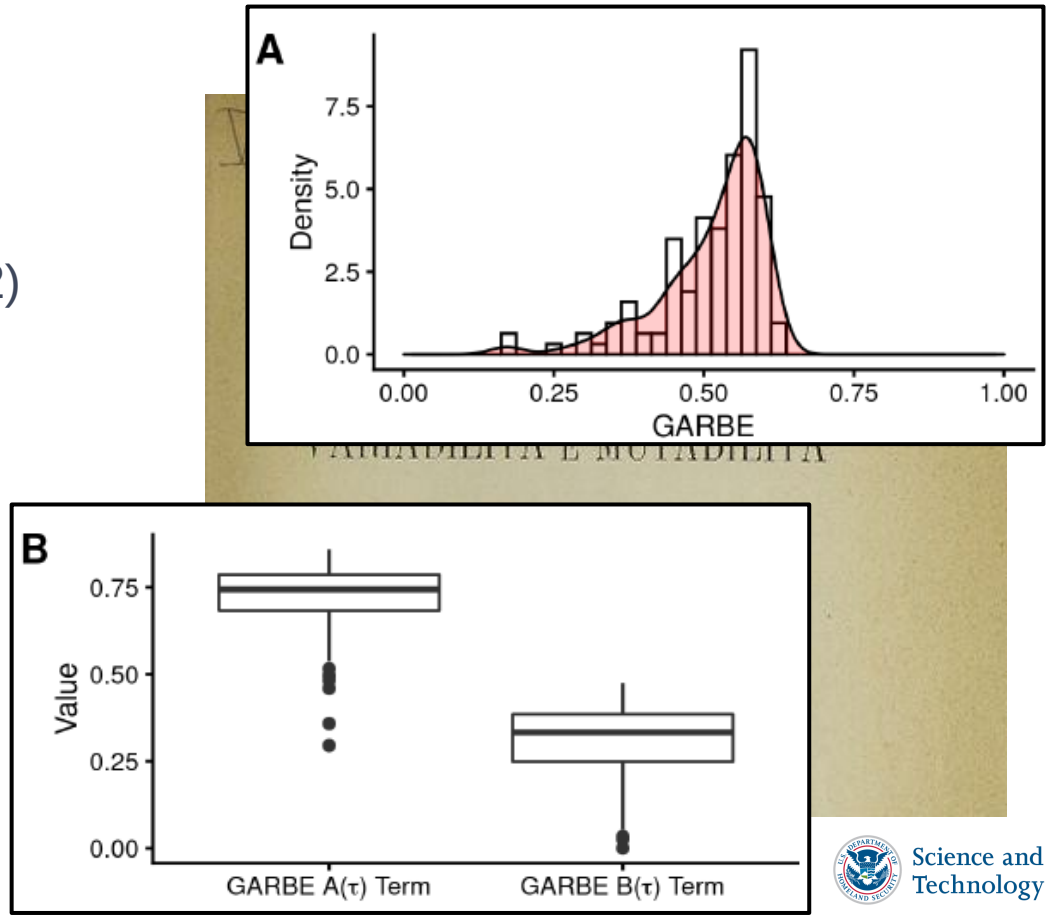
Study the Behavior of Fairness Models

- Can we develop a fairness model that is the best of both approaches?
  - Bounded, like FDR
  - Defined when FNMR or FMR = 0%, like FDR.
  - Intuitive range, like INEQ
- Enter the Gini Coefficient
  - Long standing measure of statistical dispersion (1912)
  - Often applied to income disparity (UN, OECD, WB)
  - Also used in other fields, biodiversity, dating apps

$$G_x = \left(\frac{n}{n-1}\right)\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mid x_i - x_j \mid}{2n^2\bar{x}}\right) \forall d_i, d_j \in D$$

$$A(\tau) = G_{FMR_\tau};\ B(\tau) = G_{FNMR_\tau}$$

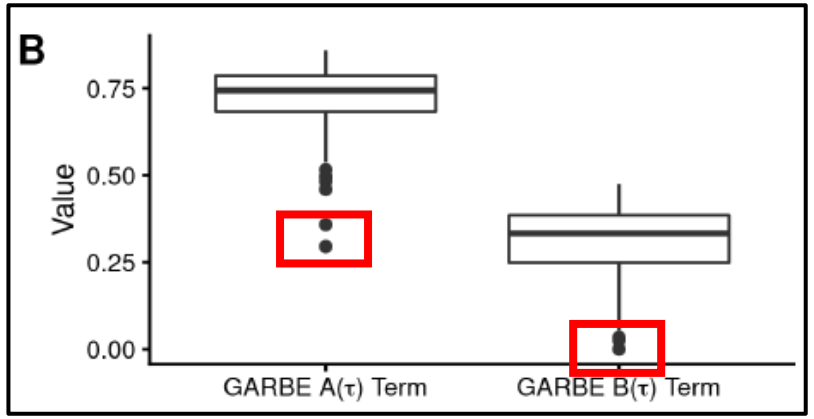
$$GARBE(\tau) = \alpha A(\tau) + (1-\alpha)B(\tau)$$

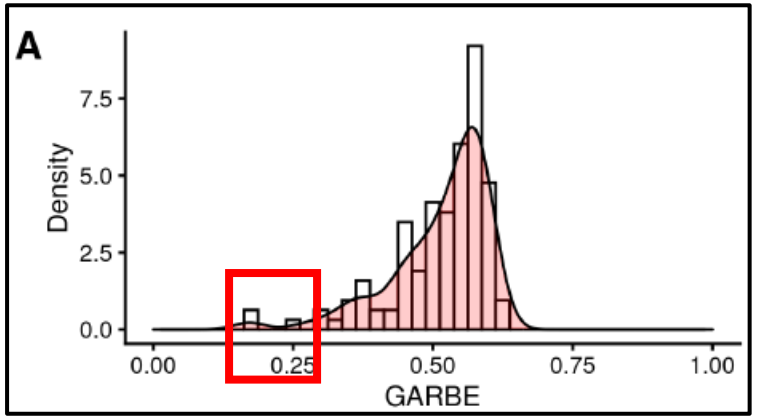


Science and Technology

What to do with a “fairness” model?

- Select algorithms that are “most fair”





- But fairness often isn’t the only consideration
- A FR algorithm can be perfectly fair by saying every face pair is a match:

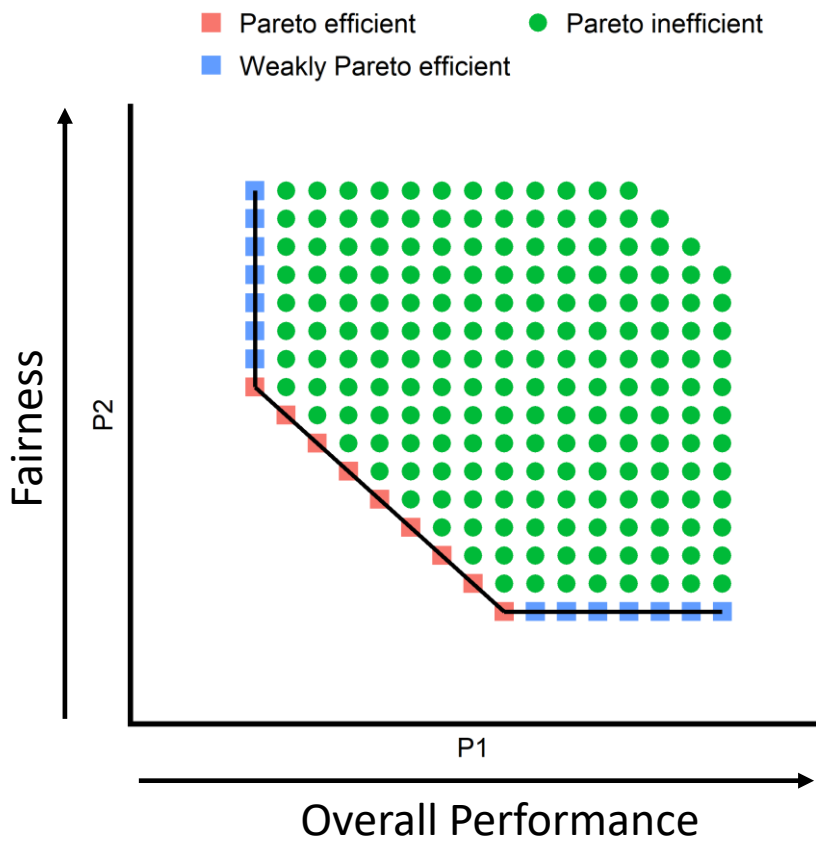
$$FMR_{d_i} == FMR_{d_j} == 100\%$$

Very fair

Science and Technology

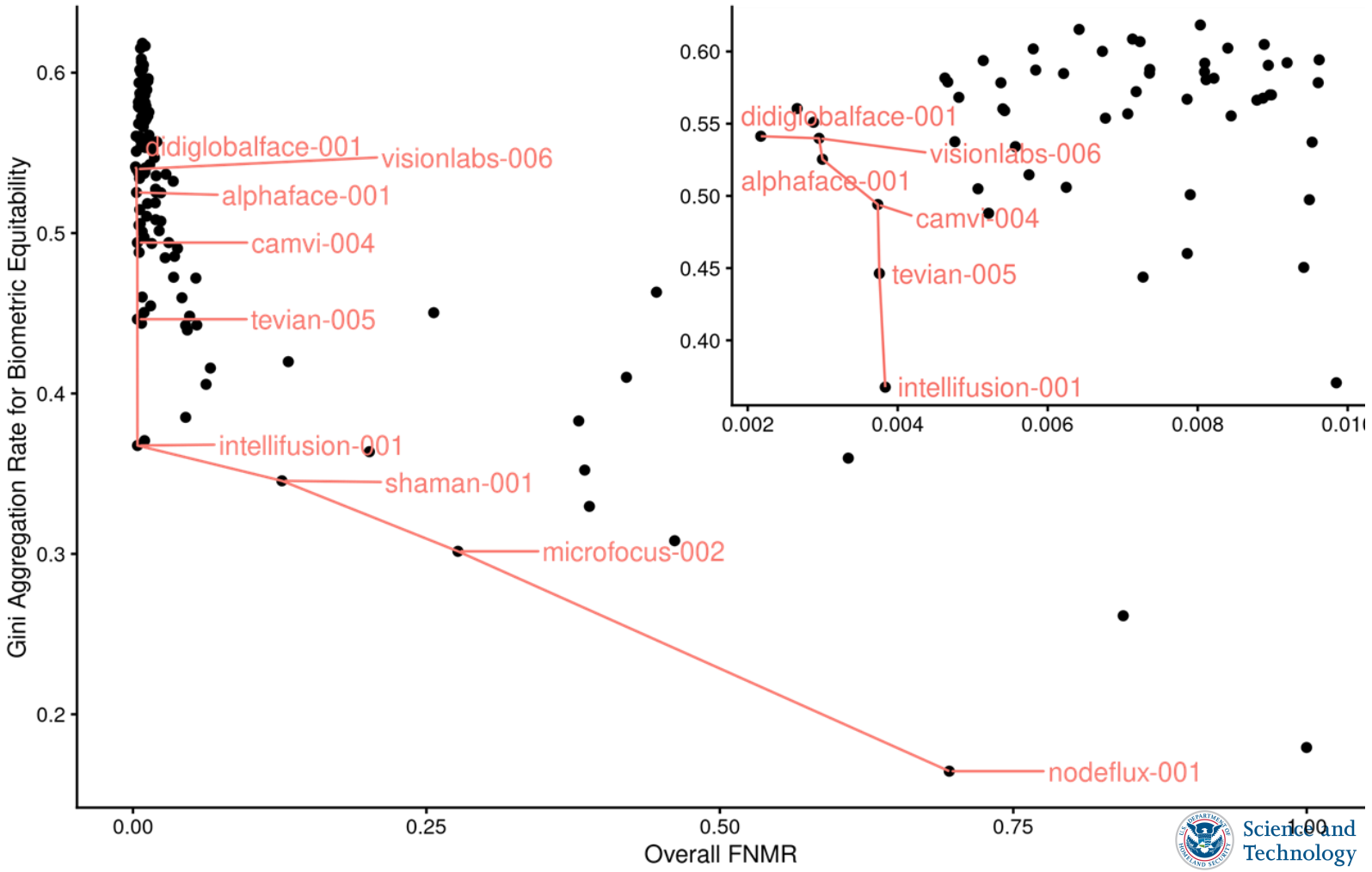
Pareto Optimizations

- Fairness is often part of a trade space with other parameters, namely accuracy
- In engineering, this is often called multi-objective optimization
- Pareto efficiency is one technique to reduce the search space in a multi-objective space



Pareto Optimization of NIST FRVT P3 Numbers

- Only need to consider options on the Pareto boundary
- Reduces candidates to consider from 127 -> 9
- Intellifusion – FNMR of 0.38%, GARBE of 0.37
- Didglobal – FNMR of 0.22% GARBE of 0.54
- Trade 0.16 performance for 0.15 fairness?



Conclusions

- Specifically -- False positive and negative differentials based on the Gini coefficients have desirable properties. Multi-objective optimization using the Pareto front also helps.

- Generally -- **Audit the audit** -- Fairness is important, properties of fairness models need to be understood at the time of their release. Please do this.

- Generally -- **Call for Data** – It's not surprising this was only done in a limited fashion before – we had no data! We hope we've taken a small step to rectifying this situation but more can be done.
  - Range of thresholds
  - Cross group FMRs

Howard, Laird, Sirotin, Rubin, Tipton, and Vemury. **"Evaluating Proposed Fairness Models for Face Recognition Algorithms",** *International Conference on Pattern Recognition* (2022).

Science and Technology

Updates on ISO/IEC **19795-10**: Biometric performance testing and reporting – Part 10: Quantifying biometric system performance variation across demographic groups

Homeland Security
Science and Technology

A Brief History

New Work Item Registered
2020-08

New Work Item Approved for Working Draft
2021-01

Committee Draft Expected
2023-Q1

Draft International Standard
2023-09

Publication
2024-09

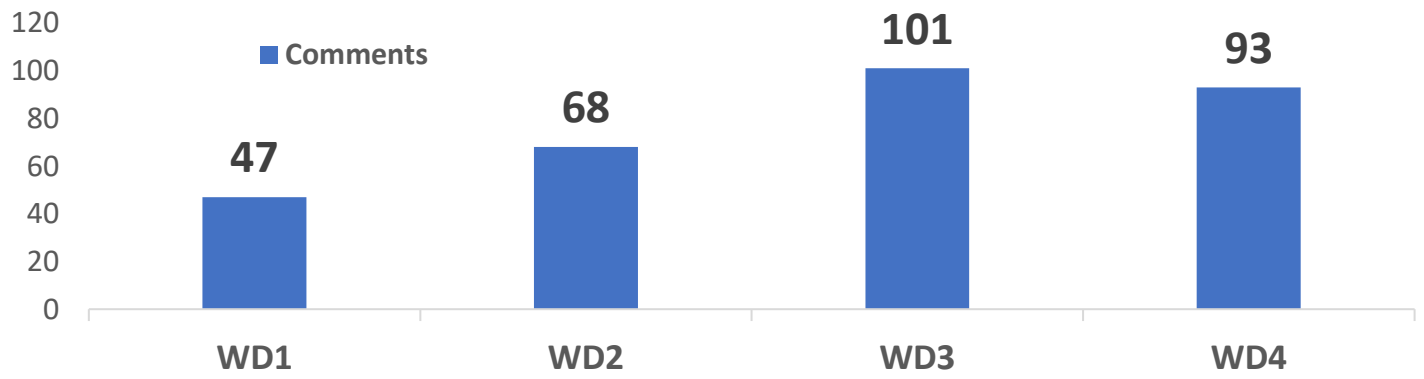
WD1
2021-05

WD2
2021-12

WD3
2022-05

WD4
2022-08

WD5
2023-01



Science and Technology

A Brief History

- Calls for Comment over Time:

WD3 -> WD4

WD3 Editor's Notes and Call for Contributions:

- **Section 5.2** - [EDITOR'S NOTE: Should experiments involving the full population be considered separately from experiments involving sampling? For instance, differentials in a time and attendance system for a specific organization can be computed for all members of the organization. These estimates would carry no error bars.]
- **Section 5.2.1** - [CALL FOR CONTRIBUTION – selecting a threshold for use in the calculation of demographic differential performance]
- **Section 5.2.1** - [CALL FOR CONTRIBUTION – calculating differential performance using continuous variables]
- **Section 5.2.1** - [EDITOR'S NOTE – The editors are aware that the current draft contains five methods for calculating a demographic differential. Comments are welcome regarding whether it is adviseable to group and/or downselect these techniques to further standardize the methodology presented below]
- **Section 5.2.2** - [CALL FOR CONTRIBUTION – pros and cons of different approaches to aggregate equitability measures, including challenges in interpretability]
- **Section 5.2.3.1** - [CALL FOR CONTRIBUTION – describe the bootsrap algorithm for the reader]
- **Section 5.2.3.2** - [CALL FOR CONTRIBUTION – techniques for calculating uncertainty in aggregate equitability measures]
- **Section 5.2.5** - [CALL FOR CONTRIBUTIONS: Contributions are welcome regarding how to formalize response variables E.g.Timing, performance, user feedback .Identify factors and levels. • Build factor list (use stakeholders & SMEs) • Identify factor categories: ▪ Device, Scenario, Operator, Subject, Process, Environment • Identify manipulated, fixed, or blocked factors; Include counterbalancing factors. Can everything be tested at once? • Treatment = tested factor/level combination • Use fractional factorial designs to reduce number of treatments needed • Use separate sub-experiments to reduce design complexity.]
- **Section 5.3.8** - [EDITOR'S NOTE: The editor's are attempting to determine if there is enough content regarding exception handling to warrant an entire section.]
- **Section 6** - [EDITOR'S NOTE: The editor's are attempting to determine if enough contribution/information for types of evaluation to warrant its own sections or whether it can be merged earlier in the document]
- **Section 6.1** - [CALL FOR CONTRIBUTION: NIST and other organizations have vast experience performing technology evaluations. How should we handle biometric subsystems as part of the general biometric model to identify demographic differentials in technology evaluations?]
- **Section 6.2** - [CALL FOR CONTRIBUTION: Looking for contribution on types of analyses and differing requirements for a differential evaluation of demographics within enrollment processes, verification systems, and identification systems.
- **Section 6.3** - [CALL FOR CONTRIBUTION: How does an operational evaluation differ from a scenario evaluation for a demographic differential evaluation, in regards to: establishing ground truth, cooperative systems/non-cooperative systems, variation in performance regarding policy and practice

WD4 -> WD5 (tentative)

WD4 Editor's Notes and Call for Contributions:

- **Section 5.2.2.3** - [CALL FOR CONTRIBUTION – regarding the issues with using other demographic factors as a proxy to ethnicity.]
- **Section 5.2.4.2** - [CALL FOR CONTRIBUTION – descriptions and best practices for collection of friction ridge pitch and eyelid palperbral aperature across demographic groups, along with samples from the literature are needed to retain Sections 5.2.4.1 and 5.2.4.1. Also, is a reference to TR 22116 appropriate?]
- **Section 5.3.3.1** - [CALL FOR CONTRIBUTION – regarding how the choice of epsilon impacts measured differentials when substituting for zero error rates in calculation of ratio-based demographic differentials, including epsilon values in geometric means]
- **Section 5.3.3.3** - [CALL FOR CONTRIBUTION – regarding when analysis of similarity score differentials is appropriate]
- **Section 5.3.5.1** - [CALL FOR CONTRIBUTION – describe the bootstrap algorithm for the reader]
- **Section 5.2.5** - [CALL FOR CONTRIBUTIONS: Contributions are welcome regarding how to formalize response variables E.g.Timing, performance, user feedback .Identify factors and levels. • Build factor list (use stakeholders & SMEs) • Identify factor categories: ▪ Device, Scenario, Operator, Subject, Process, Environment • Identify manipulated, fixed, or blocked factors; Include counterbalancing factors. Can everything be tested at once? • Treatment = tested factor/level combination • Use fractional factorial designs to reduce number of treatments needed • Use separate sub-experiments to reduce design complexity.]
- **Section 5.3.8** - [EDITOR'S NOTE: The editor's are attempting to determine if there is enough content regarding exception handling to warrant an entire section.]
- **Section 6 -** [EDITOR'S NOTE: The editor's are attempting to determine if enough contribution/information for types of evaluation to warrant its own sections or whether it can be merged earlier in the document]
- **Section 6.1** - [CALL FOR CONTRIBUTION: NIST and other organizations have vast experience performing technology evaluations. How should we handle biometric subsystems as part of the general biometric model to identify demographic differentials in technology evaluations?]
- **Section 6.2** - [CALL FOR CONTRIBUTION: Looking for contribution on types of analyses and differing requirements for a differential evaluation of demographics within enrollment processes, verification systems, and identification systems.
- **Section 6.3** - [CALL FOR CONTRIBUTION: How does an operational evaluation differ from a scenario evaluation for a demographic differential evaluation, in regards to: establishing ground truth, cooperative systems/non-cooperative systems, variation in performance regarding policy and practice

Steady coalescing of the technical content in the standard

Content

- Terms:
  - **Differential Performance** – differences in final system results between different demographic groups
  - **False Negative Differential Performance** – a difference in false negative error rates within multiple demographic groups
  - **False Positive Differential Performance** – a difference in false positive error rates within multiple demographic groups
  - **Score Differential Measures** – differences in system measures between different demographic groups not represented in biometric system outcomes.

Differential Performance

- Between two groups:
  - Based on a difference:

$$B(\tau) = FNMR_{d_i}(\tau) - FNMR_{d_j}(\tau); d_i, d_j \in D$$

False Negative Differential

  - Based on a ratio:

$$A(\tau) = \frac{FMR_{d_i}(\tau)}{FMR_{d_j}(\tau)}; d_i, d_j \in D$$

False Positive Differential

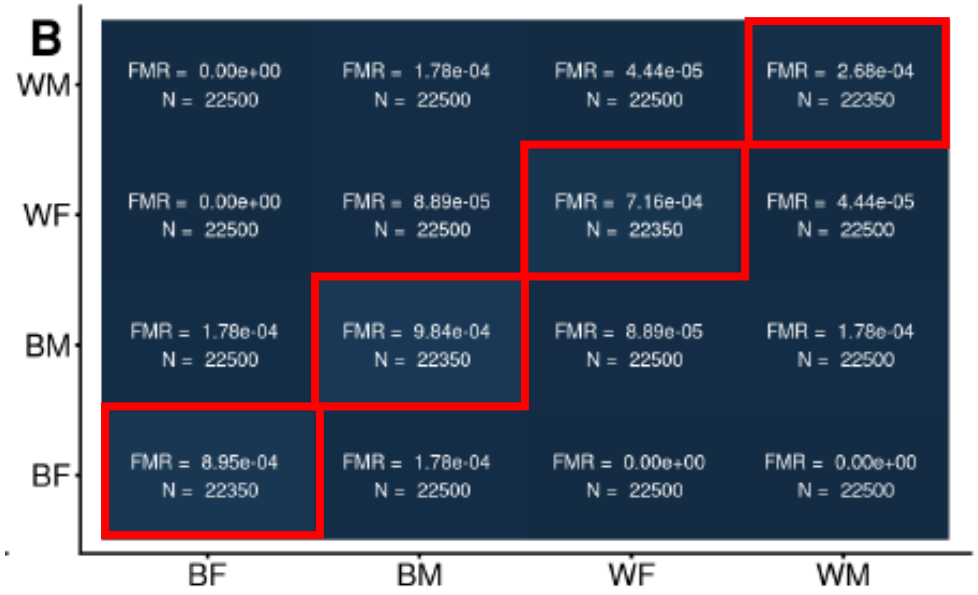
$$B(\tau) = \frac{FNMR_{d_i}(\tau)}{FNMR_{d_j}(\tau)}; d_i, d_j \in D$$

False Negative Differential

Differential Performance

- Between more than two groups:

- Worst case error rate divided by the geometric mean:
  - Outstanding questions on what to do in the presence of 0% error rates



$$A(\tau) = \frac{max_{d_i}(FMR_{d_i}(\tau))}{\widehat{FMR}(\tau)} \forall\, d_i \in D$$

$$B(\tau) = \frac{max_{d_i}(FNMR_{d_i}(\tau))}{\widehat{FNMR}(\tau)} \forall\, d_i \in D$$

$$\hat{x}(\tau) = \left(\prod_{d_i \epsilon D} x_{d_i}\right)^{\frac{1}{n}}$$

Science and Technology

Differential Performance

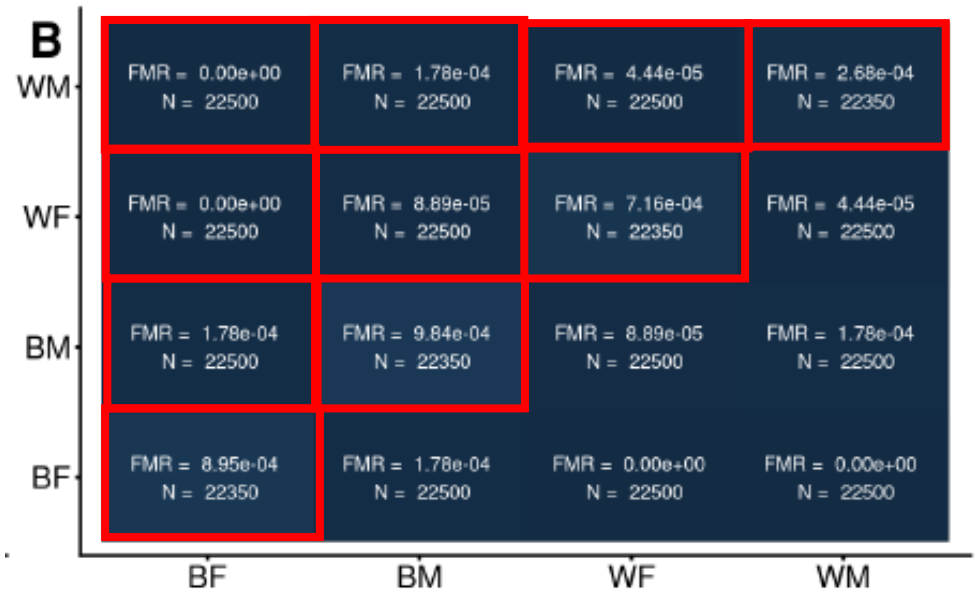
- Between more than two groups:

- Gini based error rate “spread”:

$$A(\tau)=\left(\frac{n}{n-1}\right)\frac{\sum_i\sum_j\left|FMR_{d_i}(\tau)-FMR_{d_j}(\tau)\right|}{2n^2\overline{FMR}(\tau)}\forall\, d_i,d_j\in D$$

$$B(\tau)=\left(\frac{n}{n-1}\right)\frac{\sum_i\sum_j\left|FNMR_{d_i}(\tau)-FNMR_{d_j}(\tau)\right|}{2n^2\overline{FNMR}(\tau)}\forall\, d_i,d_j\in D$$

$$\overline{x}(\tau)=\frac{1}{n}\sum_{d_i\epsilon D}x_{d_i}$$



Differential Performance

- Aggregate measures now deprecated

$$A(\tau) = \frac{\max_{d_i} FMR_{d_i}(\tau)}{\min_{d_j} FMR_{d_j}(\tau)} \quad \forall d_i, d_j \in D \tag{4}$$

$$B(\tau) = \frac{\max_{d_i} FNMR_{d_i}(\tau)}{\min_{d_j} FNMR_{d_j}(\tau)} \quad \forall d_i, d_j \in D \tag{5}$$

$$IR = A(\tau)^{\alpha} B(\tau)^{1-\alpha} \tag{6}$$

$$G_x = \left(\frac{n}{n-1}\right)\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mid x_i - x_j \mid}{2n^2\bar{x}}\right) \forall d_i, d_j \in D$$

$$A(\tau) = G_{FMR_\tau};\ B(\tau) = G_{FNMR_\tau}$$

$$GARBE(\tau) = \alpha A(\tau) + (1-\alpha) B(\tau)$$

Differential Treatment

- Terms
  - **Differential treatment** – taking a set of actions for a biometric enrollee or biometric capture subject based on their demographic characteristics

5.5.4.3 Reporting for systems that set thresholds for each demographic group

For systems that use different thresholds for different demographic groups, the evaluation report shall tabulate, for each demographic group, the threshold value, and the false negative and false positive rates.

NOTE Studies that report differentials across cohorts where the match threshold is allowed to vary between cohort groups are not representative of how differentials would be experienced in real world operations. For example, reporting false negative identification rate for cohort A and B where the false positive rate across those groups is constant masks the fact that to achieve a constant false positive rate, the threshold for the two groups would have to be changed, and this is often impossible to implement.

NOTE To evaluate a system where a threshold is set per demographic group, the tester must report on the demographic group misclassification error rate of that system. Since demographic group classification is outside the scope of this standard (see Introduction), it is not possible to test such systems against this standard. The only guidance this standard provides for such systems is that they meet the definition provided here of a system that exercises a differential treatment policy. This is not a recommended practice.

Identification Evaluations

5.4.2.3 Identification (1:N)

Measurement of demographic differentials for evaluations of identification systems is more complex. These shall consider the demographic composition of both the probe image and the full identification gallery. Similar to 1:1 non-mate performance, a simplifying approach is to constrain analysis to cohorts where demographics of non-mated samples are matched (e.g. both Female or both Male.. However, in identification scenarios, these differentials are sufficient only for systems where comparisons are only made between individuals of the same gender. If this is not the case, the tester should investigate whether FMR might increase when comparing samples from different demographic groups.

NOTE Some face recognition algorithms have been shown to produce false matches across pose angles. This can occur if tall and short subjects were imaged with a fixed camera.

Another simplifying approach is to use a constant demographically diverse set of reference samples or gallery samples with demographic composition selected based on the context of use. In this case differentials would

be computed only based on probe demographics. However, in some cases differentials for each combination of demographic pairings will have to be computed (e.g. Female-Male, Female-Female, Male-Male).

The specific test approach will depend on the biometric modality and type of evaluation. The test design shall balance external validity, which requires keeping true to the context of use, and interpretability which requires better control. The evaluation plan and evaluation report shall state the rationale for cohort selection for generation of non-mated transactions, and describe its relation to the intended context of use.

Questions & Answers

- Contact information
  - arun.vemury@hq.dhs.gov
  - jhoward@idslabs.org
  - peoplescreening@hq.dhs.gov

- Visit our websites for additional information
  - To see additional work DHS S&T supports, visit www.dhs.gov/science-and-technology
  - Detailed application instructions will be available in a separate document on https://mdtf.org
  - To view additional information about this year and prior Rallies, visit https://mdtf.org