International Standard

ISO/IEC 19795-10

First edition
2024-10

Information technology — Biometric performance testing and reporting —

Part 10: Quantifying biometric system performance variation across demographic groups

Technologies de l'information — Essais et rapports de performance biométriques —
Partie 10: Quantification de la variation des performances du système biométrique selon les groupes démographiques

Reference number
ISO/IEC 19795-10:2024(en)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

ISO and IEC draw attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO and IEC take no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO and IEC had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents and https://patents.iec.ch. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 37, *Biometrics.*

A list of all parts in the ISO/IEC 19795 series can be found on the ISO and IEC websites.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html and www.iec.ch/national-committees.

Introduction

As the use of biometric technology increases, so too does public interest in establishing whether the technology performs similarly for all individuals. Stakeholders are asking government and industry organizations that use biometric technology to establish whether these technologies vary in performance for different demographic groups. The intention of this document is to provide guidance on how to measure and report performance variation across demographic groups.[2]

This document is intended to help organizations evaluate demographic performance in biometric systems and report their results. Specifically, this document outlines how to measure and report biometric performance variations across demographic groups. It provides a set of metrics and best practices to facilitate such testing. However, this document does not provide guidance on how to establish specific causes for the observed variations. The following demographic variables are explicitly discussed in this document:[7][10][12]

— biological characteristics, such as:

  — sex, age, weight, height and skin lightness;

— social constructs, such as:

  — ethnicity, gender and language.

Many other variables can cause systematic changes in biometric characteristics or in how individuals interact with biometric systems. The following demographic variables are relevant although not explicitly discussed in this document:

— performance variations based on temporary states, such as:

  — self-styling (e.g. makeup, eyewear, mask-wearing, clothing, hairstyles),

  — behavioural or emotional states (e.g. intoxication),

  — behaviours (e.g. smiling, closing eyes, varying pose);

— performance variation caused by diseases or injuries, such as:

  — eye surgery, cataracts, vision correction,

  — stroke, cleft lip, Apert's syndrome,

  — missing digits;

— performance variation caused by disabilities.

Demographic performance variation for applications other than biometric recognition, such as emotion, gender or age estimation, are not considered in this document.

Information technology — Biometric performance testing and reporting —

Part 10: Quantifying biometric system performance variation across demographic groups

1 Scope

This document establishes requirements for estimating and reporting on performance variations observed when cohorts belonging to different demographic groups engage with biometric enrolment and recognition systems. In this context, performance refers to failure-to-enrol rate, failure-to-acquire rate, shifts in comparison score, recognition error rates, and aspects of response and processing time (throughput).

This document is applicable to the following:

— demographic group membership;

— using phenotypic measures;

— reporting on tests;

— stating statistical uncertainty estimates;

— operational thresholds settings;

— equitability;

— procurement agency activities.

This document also provides terms and definitions to be used when reporting performance variation across demographic groups.

This document is applicable to:

— technology evaluations of algorithms, subsystems and systems;

— scenario evaluations of systems;

— operational evaluations of fielded systems.

Application of this document does not require detailed knowledge of a system's algorithms but it does require specific knowledge of the demographic characteristics for the population of interest.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 2382-37, *Information technology — Vocabulary — Part 37: Biometrics*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 2382-37 and the following apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
differential performance measure
DPM
difference in biometric system measures across different demographic groups

EXAMPLE Differences in error rates [e.g. False Match Rate (FMR), False Non-Match Rate (FNMR)] between different demographic groups.

Note 1 to entry: ISO/IEC 2382-37:2022 term 37.09.28 defines "demographic differential" as the difference in "outcome of a biometric system". This definition is equivalent to this document's "differential performance measure". This document also recognizes other kinds of demographic differentials, such as *differential treatment* (3.7) and *comparison score differential measure* (3.4).

3.2
false negative differential performance
FND
difference in false negative error rates calculated across multiple demographic groups

EXAMPLE If Group A's false non-match rate is 10 %, and Group B's false non-match rate is 20 %, the false negative differential is 10 percentage points if viewed as a mathematical difference or a factor of 2 if viewed as a mathematical ratio (see 6.4).

3.3
false positive differential performance
FPD
difference in false positive error rates calculated across multiple demographic groups

EXAMPLE If Group A's false match rate is 1 %, and Group B's false match rate is 3 %, the false positive differential is 2 percentage points if viewed as a mathematical difference or a factor of 3 if viewed as a mathematical ratio (see 6.4).

3.4
comparison score differential measure
difference in system measures across different demographic groups represented through comparison score analysis

EXAMPLE Differences in mean comparison scores for different demographic groups (see 6.9).

3.5
mated comparison score differential measure
difference in the statistics of mated score distributions observed for different demographic groups

EXAMPLE If the mean mated comparison score for subjects in Group A is 10 and the mean mated comparison score for subjects in Group B is 5, then the mated comparison score differential measure is a mean difference of 5 (see 6.9).

3.6
non-mated comparison score differential measure
difference in the statistics of non-mated score distributions observed for different demographic groups

EXAMPLE If the mean non-mated comparison score for subjects in Group A is 10 and the mean non-mated comparison score for subjects in Group B is 5, then the non-mated comparison score differential measure is a mean difference of 5 (see 6.9).

3.7
differential treatment
different set of actions for a biometric enrolee or biometric capture subject based on their demographic group

EXAMPLE Implementing a system in which one machine learning model recognizes male faces and a different machine learning model recognizes female faces.

3.8
categorical demographic variable
demographic variable of an individual that is nominally or ordinally described

EXAMPLE A data subject's gender or ethnicity.

3.9
continuous demographic variable
demographic variable of an individual that is observable, measurable and not necessarily constrained to discrete categories

EXAMPLE An individual's age or the measurement of a phenotypic trait, such as an individual's skin lightness.

3.10
intersectional demographic variable
demographic group that is the combination of multiple categorical demographic variables.

EXAMPLE A data subject's gender-ethnicity.

3.11
demographic group
value of a continuous, categorical or intersectional demographic variable associated with a data subject

EXAMPLE A data subject that has self-reported their gender as female has a demographic group of female for the categorical demographic variable of gender.

3.12
demographic reference database
database comprising biometric references annotated with demographic variables and groups

3.13
aggregate equitability measure
AEM
performance measure that combines multiple measures of differential performance into an aggregate measure of overall differential performance

3.14
confidence interval
interval estimator (T_0, T_1) for the parameter θ with the statistics T_0 and T_1 as interval limits and for which it holds that $P[T_0 < \theta < T_1] \geq 1 - \alpha$

Note 1 to entry: Unless otherwise stated, the threshold for statistical significance, α, is 0.05, which equates to a 95 % probability that the parameter is within the interval limit.

[SOURCE: ISO 3534-1:2006, 1.28, modified — original Notes to entry have been removed and replaced by a new Note 1 to entry.]

3.15
effect magnitude
statistical measure of the size of an observed differential

EXAMPLE 1 A mathematical difference of 20 percentage points in false non-match rates between two demographic groups (e.g. 5 % vs. 25 %).

EXAMPLE 2 A mathematical ratio of 5 between false non-match rates between two demographic groups (e.g. 5 % vs. 25 %).

4 Conformance

To conform to this document, a biometric evaluation assessing performance variation across demographic groups shall be planned, executed and reported in accordance with the requirements contained in Clauses 5 to 7.

5 Planning the evaluation

5.1 Identifying the scope of the evaluation

This subclause establishes the experimental methods for designing evaluations to measure demographic differences in the performance of biometric systems. In general, experimental design includes setting the objectives of an evaluation and determining the statistical properties and design of the evaluation to match the objectives. This document applies specifically to evaluations in which one of the objectives is to calculate differential performance measures (DPMs) or calculate comparison score differential measures in biometric systems across different demographic groups. Prior to executing the evaluation, the tester shall prepare a test plan describing the evaluation.

— Test plans shall describe the objectives as well as any models or hypotheses of the evaluation, including:

  — demographic variables and groups of interest;

  — biometric performance measures of interest;

  — demographic differential performance measure(s) and/or comparison score differential measure(s) of interest;

  — effect magnitude(s) of interest.

— Test plans shall describe the data that will be gathered to test these models or hypotheses, including:

  — how demographic variables are to be measured or otherwise collected;

  — manipulated, fixed or blocked factors, including counterbalancing factors where appropriate;

  — controls for non-tested factors;

  — target sample size requirements, including the rationale for cohort selection when generating mated and non-mated trials and the target type I and type II error.

— Test plans should describe what analyses will be performed on these data and what inductions will be attempted, including:

  — the expected uncertainty around differential performance measures;

  — statistical tests to be performed.

The balance between the internal and external validity of the evaluation should be considered and explained. Evaluations with high internal validity, such as technology tests, focus on specific components of biometric systems and are well controlled: many factors are considered, documented and manipulated in a controlled fashion or fixed at pre-determined levels. Evaluations with high external validity, such as operational tests, are not necessarily able to attribute the observed differentials uniquely to the factors of interest due to uncontrolled variation in the test environment. These evaluations therefore have lower internal validity. Any deviations between the test design and the envisioned operational conditions for the system shall be noted and reported as these efforts to control variation may change the effect magnitude observed relative to the target environment within which the biometric system operates (see 7.2).

This document does not specify what constitutes an acceptable amount of differential performance. To inform the design of the evaluation, regulators or procurement guidelines can specify allowable differential performance, where appropriate, calculated according to at least one of the methods described in 6.4 to 6.9.

When specifying that a level of differential performance is not acceptable, regulators or procurement guidelines can utilize benchmarks as described in 7.4.3.

5.2 Demographic variables

5.2.1 Ground truth requirements

Evaluations of biometric performance have strict requirements for establishing the ground truth identity of data capture subjects. This is to ensure the validity of any metrics derived from these classifications, such as false match and false non-match rates. Evaluations of biometric performance across demographic groups have three additional constraints:

— Demographic evaluations shall specify the demographic variables of interest.

— Each demographic variable shall be comprised of defined demographic groups which shall be associated to individual data capture subjects.

— Demographic group membership for demographic variables of interest and other metadata should be collected at the same time as the corpus samples to avoid errors in inference.

Evaluations to measure performance variation across demographic groups should involve focused data collection where demographic groups are recorded and where ground-truth identity information is established.

Demographic group membership should not be inferred directly from biometric samples. An example of this is assigning the value of ethnicity or the value of gender from a face sample. Demographic groups are properties of a data subject or a data subject's biometric characteristic. They are not properties of a biometric sample. Estimating demographic groups from biometric samples can introduce spurious correlations between biometric performance and demographic variables. For example, if the width of a face or eyelid palpebral aperture used to estimate the demographic group is measured from the same sample used for biometric comparison, any lens distortion can affect both the biometric and the demographic outcomes. If it is not possible to establish demographic group membership independent of the biometric sample, other techniques should be applied (see 5.2.2 and 5.2.3). In this case, the tester should carefully consider and shall document any correlations and impacts between demographic variables and the biometric sample collection technique.

Many demographic variables are categorical. Categorical demographic variables are those that take a distinct, limited number of possible values, such as gender and ethnicity. Other demographic variables are continuous and have an infinite number of possible values. These can be combined into demographic groups for the purpose of analysis. In some practical applications, continuous demographic variables such as age and height are bound by natural limits and should be reported in appropriate granularity.

5.2.2 Categorical demographic variables

5.2.2.1 Sex

Sex is defined as the state of being male or female as it relates to biological factors such as DNA, anatomy and physiology. Sex typically consists of two categories, "male" and "female". Female individuals generally possess two copies of the X chromosome. Male individuals generally possess one copy each of an X and a Y chromosome. Important exceptions do occur and complicate binary classification. The tester should establish appropriate categories for sex. If necessary, the tester can extend the general binary classification model of male/female.

When sex is included in the evaluation, it shall be determined through the collection and analysis of DNA or by self-report. In evaluations that include sex, the tester shall prepare a statement that documents how sex was determined (see 7.3).

NOTE If sex was determined by self-report, gender can also be recorded.