ISO/IEC JTC 1/SC 27/WG 5 "Identity management and privacy technologies"
Convenorship: **DIN**
Convenor: **Rannenberg Kai Mr Prof. Dr.**

ISO/IEC AWI 27566-2

Document type	Related content	Document date	Expected action
Ballot / Reference document	Ballot: ISO/IEC AWI 27566-2 (restricted access)	2023-12-05	**COMMENT/REPLY** by 2024-01-30

ISO 27566-2:2025(X)

ISO/IEC JTC1/SC27/WG5

Secretariat: DIN

Information technology, cybersecurity and privacy protection – Age assurance systems –

Part 2: Benchmarks for benchmarking analysis

Sécurité de l'information, cybersécurité et protection de la vie privée - Systèmes d'assurance de l'âge –

Partie 2: Points de repère pour analyses comparatives

WD1

Warning for WDs and CDs

This document is not an ISO International Standard. It is distributed for review and comment. It is subject to change without notice and may not be referred to as an International Standard.

Recipients of this draft are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.

To help you, this guide on writing standards was produced by the ISO/TMB and is available at https://www.iso.org/iso/how-to-write-standards.pdf

A model manuscript of a draft International Standard (known as "The Rice Model") is available at https://www.iso.org/iso/model_document-rice_model.pdf

Contents

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

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 ISO documents 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).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

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.

This document was prepared by Technical Committee *[or Project Committee]* ISO/TC *[or ISO/PC]* ###, *[name of committee]*, Subcommittee SC ##, *[name of subcommittee]*.

This second/third/… edition cancels and replaces the first/second/… edition (ISO #####:####), which has been technically revised.

The main changes compared to the previous edition are as follows:

— xxx xxxxxxx xxx xxxx

A list of all parts in the ISO ##### series can be found on the ISO website.

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.

Introduction

This document sets out the approach to benchmarking of age assurance systems deployed for the purpose of enabling age-related eligibility decisions established in accordance with [Part 1 of this document] [ISO/IEC 27566-1 (currently in development) Age Assurance Systems – Framework].

Age assurance is a declaration that provides an indication of confidence in the length of time that a person has lived.

This document aims to establish consistent approaches to benchmarking of age assurance systems.

This document is intended to:

- Establish a common approach to benchmarking of age assurance systems that could be utilised to support age assurance interoperability, age attribute exchanges and, with standardised certification, intelligent monitoring and testing of age assurance systems
- Enable attributes attestation providers or identity attribute providers to position themselves against internationally recognised principles in relation to age assurance.

This document provides for the requirements, analysis, testing and certification of different approaches to age assurance systems.

This document does not:

- Establish or hinder the establishment of any methodologies (called assurance components in this standard) for age assurance systems – it is technology agnostic
- Establish or recommend the age assurance thresholds or determine the required levels of assurance for different products, content or services – these are matters for policy makers
- Deal with financial or commercial models for age assurance systems – these are a matter for economic operators in the age assurance process
- Address, save for some high level principles specifically applicable to age assurance systems, the requirements for securing data protection and privacy of persons – these are a matter for data controllers
- Establish the detailed requirements for interoperability, age assurance trust frameworks, age assurance exchanges or communities of interest for age assurance systems – these could be a matter for future standards, technical specifications or technical reports
- Establish the detailed test methodologies for assurance components, other than adopting the benchmarking framework, as set out in the ISO 29155 series of standards.

The International Organization for Standardization (ISO) *[and/or] International Electrotechnical Commission (IEC)* draw[s] attention to the fact that it is claimed that compliance with this document may involve the use of a patent.

ISO *[and/or] IEC* take[s] no position concerning the evidence, validity and scope of this patent right.

The holder of this patent right has assured ISO *[and/or] IEC* that he/she is willing to negotiate licences under reasonable and non-discriminatory terms and conditions with applicants throughout the world. In this respect, the statement of the holder of this patent right is registered with ISO *[and/or] IEC*. Information may be obtained from the patent database available at www.iso.org/patents.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights other than those in the patent database. ISO *[and/or] IEC* shall not be held responsible for identifying any or all such patent rights.

Information technology, cybersecurity and privacy protection – Age assurance systems – Part 2: Benchmarks for benchmarking analysis

Sécurité de l'information, cybersécurité et protection de la vie privée - Systèmes d'assurance de l'âge – Partie 2: Points de repère pour analyses comparatives

1 Scope

This document establishes benchmarks for specifying, differentiating and comparing characteristics of age assurance methods and components.

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 2382-1:2012, *Information technology – Vocabulary – Part 1 :*

ISO/IEC 29155-1:2016, *Systems and software engineering — Information technology project performance benchmarking framework — Part 1: Concepts and definitions*

ISO/IEC 29155-2:2013, *Systems and software engineering — Information technology project performance benchmarking framework — Part 2: Requirements for benchmarking*

ISO/IEC 29155-3:2015, *Systems and software engineering — Information technology project performance benchmarking framework — Part 3: Guidance for reporting*

ISO/IEC 29155-4:2016, *Systems and software engineering — Information technology project performance benchmarking framework — Part 3: Guidance for data collection and maintenance*

... [There are more to include here]

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

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

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

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

The terms and definitions in Part 1 shall apply.

4 Age Assurance Systems Benchmarking

4.1 General

4.1.1 Application of the benchmarking framework

This document adopts the ISO/IEC 29155 series approach to systems and software engineering – information technology project performance – benchmarking for the purposes of evaluating the efficacy, security and reliability of Age Assurance Systems. This is an existing, widely adopted, methodology.

The benchmarking framework (ISO/IEC 29155-1, ISO/IEC 29155-2, ISO/IEC 29155-3 and ISO/IEC 29115-4) are relevant to comparing “objects of interest” to each other, or against a benchmark, to evaluate characteristic(s). In the context of the ISO/IEC 27566 – Age Assurance Systems - Framework, the “object of interest” is the performance of age assurance system, and the characteristics relative to its efficacy to deliver reliable indicators of confidence and the meeting of security and privacy objectives.

The ISO/IEC 29155 series contains multiple parts:

- Part 1 provides the overall framework model for IT project performance benchmarking. It consists of activities and components that are necessary to successfully identify, define, select, apply, and improve benchmarking. It also provides definitions for IT project performance benchmarking terms;
- Part 2 describes the required tasks in individual benchmarking activities that are necessary to execute various activities to conduct and/or support successful benchmarking in an organization;
- Part 3 provides general requirements and guidance for reporting processes and contents of typical reports;
- Part 4 provides general requirements and guidance for the activities to collect data of IT project to be entered into and maintained in a benchmarking repository.

This document applies the Benchmarking Framework to the assessment, analysis and performance reporting of Age Assurance Systems.

4.1.2 Overview

Age assurance systems can be subjected to benchmarking as with any other IT project.

Age assurance systems have certain characteristics that need special consideration during an benchmarking, including the following:

- Analysis of the output of age assurance systems against expected performance to achieve the different indicators of confidence set out in ISO/IEC 27566 (zero, basic, standard, enhanced or strict) or against other measures/sub-categories as needed for the uses that a particular age assurance system is put to
- Capture, analysis, resolution and communication of contra-indicators, such as mis-matches in claimed and identified attributes, complicity, trust framework attacks and attempts to circumvent the age assurance process
- For remote age assurance systems, there are performance error rates: Biometric authentication does not work as deterministically as other means for authentication or identification of users. Some performance error rates (e.g. according to ISO/IEC 19795-1) have an impact on the security of the system and need to be considered during a benchmarking process.
- Presentation attack detection (PAD): It is well known that some biometric systems (e.g. PAD subsystem, data capture subsystem, or full system) may be vulnerable against presentation attacks. The benchmarking of the capability to detect and defeat these attacks may form part of benchmarking.
- Vulnerability assessment: Age assurance systems in general may be subject to special kind of attacks (such as credential stuffing) that will need consideration during benchmarking.
- Privacy objectives: ISO/IEC 27566-1 sets out a series of privacy objectives that Age Assurance Systems are required to meet.

For these areas, special guidance is required in order to facilitate a comparable benchmarking in all conformity assessment activity worldwide. Special characteristics of age assurance systems in benchmarking are dealt with in form of guidance for the benchmarking analyst performing an benchmarking and the developer of a age assurance system.

4.1.3 General benchmarking aspects

A benchmark analyst (either internal to an organisation or provided externally, perhaps through a benchmark service provider or conformity assessment body) may develop benchmarking methods to produce a core report. Typically, a core report may consist of an executive summary and a detailed report for each instance of benchmarking.

Most aspects can be applied to age assurance systems as to any other IT product. However, in some areas, specific guidance is given to the benchmarking analyst on how to evaluate these aspects. For example, the description of the design of a age assurance system refers to specific aspects of the technology.

4.2 Benchmarking the age check practice statement

The core benchmarking activity in the context of an Age Assurance System or component shall be against the Age Check Practice Statement (see clause X.X of ISO/IEC 27566-1).

In the terms of ISO/IEC 29155-1, this forms the business layer, including the responsibility for organisational business decision making and commitment.

The benchmarking analysis report shall include an assessment of the performance of the Age Assurance System or component in delivering the commitments expressed by the service provider in their Age Check Practice Statement. This is referenced in ISO/IEC 29155-3, 7.2.1 as Reports of “Conduct benchmarking” activity.

4.3 External benchmarks for age assurance systems

4.3.1 Statement of headline accuracy

Part 1 establishes five indicators of confidence that may be utilised in Age Assurance Systems. It does not prescribe that all five shall be used, nor that a particular case may require more than five, but it does establish a framework based around five indicators of confidence – these are asserted, basic, standards, enhanced and strict age assurance.

In benchmarking the performance of age assurance systems a headline statement of overall accuracy of the age assurance measure should be provided aligned to indicators of confidence.

This enables a quick, easy, and readily accessible indication of accuracy to be provided to an unfamiliar audience.

The headline statement of accuracy shall indicate the observed proportion of correctly classified subjects by the technology in the benchmarking instance. In addition, it shall be accompanied by a description of the Age Assurance component (if the benchmarking is of a single component) or the sequence or combination of components in an Age Assurance System (if the benchmarking is of a system of multiple, sequential or combined components).

Correct classification shall be the output of the balanced accuracy of the age assurance component or system under benchmarking analysis.

The balanced accuracy is the average of the true positive rate (TPR) and the true negative rate (TNR). It is the same as the accuracy if the test data set is balanced.

$$Balanced\ Accuracy = \frac{1}{2}(TPR + TNR)$$

The external benchmarks for age assurance systems shall be established as:

- For **asserted age assurance**, there shall be no benchmark for correct classification (it is assumed that asserted age assurance is not subject to benchmarking)
- For **basic age assurance**, the classification accuracy shall be at least 90%
- For **standard age assurance**, the classification accuracy shall be at least 99%
- For **enhanced age assurance**, the classification accuracy shall be at least 99.9%
- For **strict age assurance**, the classification accuracy shall be at least 99.99%

Classification accuracy is a measure of how likely the age assurance component or system is to deliver an accurate response to an age-related eligibility prompt that has two possible answers (i.e. a binary response). An example could be, ‘is this person over 18?’; to which the answer would either be ‘yes’ or ‘no’.

4.3.2 Statement of headline accuracy qualified by age buffering

For some types of age assurance system or component, the indicator of headline accuracy can be affected by a process of age buffering – or effectively creating a safety buffer between the ‘age of interest’ – that is the age that the relying party must establish to make an age-related eligibility decision – and the ‘challenge age’ – that is the age at which the age assurance system or component is confident in its output without requiring the relying party to deploy a further type of age assurance component.

This is more easily explained by considering age estimation as a process. An age estimation system may be very confident that a subject is (say) over 35 years old, but far less confidence that a subject is (say) 18.5 years old. In such circumstances, it may be configured to respond to the question, is this person over 18 with, ‘yes’ if it thinks they are over a challenge age of (say) 25, and ‘maybe’ if it thinks they are perhaps between 18 or 25 (configured as a ‘no’ in system responses); or ‘no’ if it thinks they are clearly under 18. Those indicated as ‘maybe’, could be subject to a secondary age assurance process.

It is likely that the misclassification rate will be higher for those persons who are closest to any age thresholds. For example, if the technology is estimating whether a person is over 13 or not, it is likely to be more accurate at classifying people who are 10 years or younger or 16 years and over, compared to someone who is 12. Therefore, it is not uncommon for users to apply an age buffer to a threshold. For example, if the age at which a person has access to services is 13, the application of an age threshold of 16 will increase

confidence that those who are identified as above 16, are indeed over 13. This is illustrated in the Challenge Age scheme that asks individuals to prove they are over 18 if they look under 21 or 25.

In some circumstances, it may be appropriate to implement an age buffer in relation to electronic age assurance, particularly when applying the tolerance levels. In our view, however, the setting of an age buffer is only really relevant where there is a statutory penalty for non-compliance, such as for the sale of alcohol, weapons, tobacco, etc to under 18s. In these circumstances, the law requires retailers to take all reasonable precautions and exercise all due diligence to avoid the commission of the offence.

For those cases where the technology estimates a person's age and age threshold is applied, it is possible to further explore how close to the age threshold an incorrect classification is. For example, in scenario 1, one possible incorrect classification is classifying someone who is under the age threshold as over. If the age threshold is 18 and the person who is incorrectly assigned as over 18 is 17, this may be deemed as less of a failure than someone who is 13 and incorrectly misclassified as over 18.

To measure the size of failures in these instances, the measures defined to assess age estimation are also appropriate here, but rather than comparing the predicted age with the true age, we compare the predicted age with the missed age threshold to better understand how close to this threshold the technology was.

For example, given the following parameters:

- The true (or observed) age of sample *i*:.
- The predicted age of sample *i*:.
- The age threshold: T_A.

The false positive absolute error for sample *i*can be calculated as:

$$FPAE_i = \begin{cases} |p_i - T_A| & if\ p_i > T_A\ and\ o_i < T_A \\ 0 & otherwise \end{cases}$$

It is possible to then calculate, for example, the mean false positive absolute error over all false positive results.

In such circumstances, a statement of headline accuracy can be qualified by a description of the age buffering policy deployed by the age assurance system or component.

4.3.3 Transparency obligations for age assurance systems

A benchmarking report shall include:

(a) Data indicating the results of continuous measurement analysis, where applicable to the type of age assurance system or component subject to benchmarking. Examples of types of continuous measurement analysis that may be applicable are provided in Annex A (informative).
(b) Data indicating the results of binary measurement analysis, where applicable to the type of age assurance system or component subject to benchmarking. Examples of types of binary measurement analysis that may be applicable are provided in Annex B (informative).
(c) Data indicating classification or outcome error parity with respect to protected characteristics, such as gender or race. Examples of approaches to securing classification or outcome error parity are provided in Annex C (informative).

The benchmarking shall be reported in accordance with ISO/IEC 29155-3:2015. In particular, an Age Assurance Service provider shall ensure that:

(a) The executive summary of a core report is made publicly available alongside their Age Check Practice Statement;
(b) The detailed report of a core report is structured in accordance with ISO/IEC 29155-3:2015 and made available to interested parties and stakeholders subject to such commercial confidentiality and non-disclosure agreements the Age Assurance Service Provider considers appropriate; and
(c) An explanatory report, if necessary, be produced for providing complementary information in order to assist understanding and to avoid inappropriate usage of the age assurance system or component.

Note 1: A benchmarking report usually consists of various formats (e.g. textual descriptions, numeric values, statistical charts and tables), and is exchanged via various methods (e.g. electronic documents, electronic data set, printed document, and embedded data within specific computer software).

4.4 Principles applicable to benchmarking reporting

4.4.1.1 Age assurance efficacy

Efficacy is the ability to perform a task (such as age estimation) to a satisfactory degree. In this context, efficacy is examined via measures of accuracy.

Note: Measurement accuracy is defined as the closeness of agreement between a measured quantity and a true quantity value of a measurand (i.e., the quantity intended to be measured). (ISO-JCGM 200, 2008 International Vocabulary of metrology- Basic and general concepts and associated terms (VIM))

The guidance in Annex A and B includes two approaches to the measurement of efficacy:

- Annex A provides possible measures applicable to continuous age assurance outputs, where the Age Assurance System or component provides an estimation of age based on algorithms or assessments; and

- Annex B provides possible measures applicable to binary age assurance outputs, where the Age Assurance System or component provides a positive declaration with only two possible options: – ‘yes’ or ‘no’.

Note: Age assurance systems can contain multiple components It is also possible that a measure could start as continuous (i.e. this person is likely to be between 55 and 65), but when applied to an age assurance threshold, it becomes binary (i.e. is that same person over 18: yes). This conversion to binary provides the headline statement of accuracy.

4.4.1.2 Age assurance equality

Equality involves ensuring that technologies treat different people fairly and equally with respect to protected characteristics such as gender and race. Whilst there is no single definition of fairness, potential assessment measures could include:

- **Anti-classification**: The Age Assurance System or component is fair if it does not use protected characteristics (except age itself in this context) or proxies from which protected characteristics can be inferred (i.e., a protected characteristic is not used to predict age).

- **Classification or outcome error parity**: The Age Assurance System or component is fair if protected groups receive an equal proportion of positive outcomes, or an equal proportion of errors.

- **Calibration**: The Age Assurance System or component is well-calibrated if the predicted ages reflect the actual ages in real life for the observations given those predictions. Equal calibration definitions of fairness say that a model should be equally calibrated between protected attribute groups

4.4.1.3 Comparability

Comparability is the extent to which differences between statistics from different age assurance technology testing, or over time, can be attributed to differences between the true values or the statistical analysis and testing.

Comparability could be more easily described as how to discuss the differences and similarities between 'apples' and 'pears'. This is an important aspect that underpins a well-functioning competitive marketplace. If economic decision makers (i.e., those procuring age assurance technologies for implementation) are not able to compare one product effectively and efficiently with another, the market for age assurance technology will be deficient.

Testing techniques should result in metrics that users are able to use in a comparable manner to either rank or distinguish their service from others that are operating in the marketplace. It is important that, in an open fair market, age assurance technology descriptions are not misleading.

4.4.1.4 Repeatability

Repeatability is a measure of precision which quantifies the degree to which repeated measurements under the same operating conditions show the same results. This is in contrast to reproducibility which is where a test environment can reproduce the results found in-house, for example.

Knowledge of the uncertainty associated with measurement results is essential to the interpretation of the results. Without quantitative benchmarkings of uncertainty, it is impossible to decide whether observed differences between results reflect more than experimental variability, whether test items comply with specifications, or whether laws based on limits have been broken. Without information on uncertainty, there is a risk of misinterpretation of results.

4.4.2 Error rates in testing

When it comes to benchmarking an age assurance system, the relevant error rates are an important aspect of the functionality to be considered. According to ISO/IEC 29155-1, the benchmarking analyst will perform the following steps:

- Identify the relevant benchmarking approach: Various approaches are available starting from a database based technology test of an age assurance system to an benchmarking of the performance of the system under operation. The correct approach highly depends on the content of the Age Check Practice Statement.
- Identify the age assurance relevant error rates:
- Plan the execution of the benchmarking: The actual execution has to be planned and described within the benchmarking planning stage documentation in advance.

It is essential to develop an idea about the amount of test data that is required before starting the actual process of test data acquisition.

- Document the test plan: It is essential to plan the required documentation for the test in advance of the test itself.
- Acquire test crew: For the quality of results, it is essential that the benchmarking analyst utilizes a test crew not known to the developer of the system beforehand.
- Perform test: The test is carried out under the sole control and responsibility of the benchmarking analyst.
- Evaluate test results: After testing, results will be evaluated and reported according to defined metrics.

4.4.3 Presentation attack detection benchmarking

The requirement for PAD mechanisms is dependent on the intended environment of the age assurance system.

For example, an age verification system under the strong and constant control of a premises entry control attendant may not require PAD, while an online age estimation system that uses biometrics as the only means for gaining age assurance would typically require PAD. The guidelines for the benchmarking of biometrics, however, specify that PAD mechanisms, if existing, belong to the security functionality of the system and therefore are to be evaluated.

PAD mechanisms can be viewed from two perspectives:

- PAD mechanisms belong to the security functionality of the age assurance system and are functionally tested. Guidelines direct the benchmarking analyst on how to plan, conduct, document, and evaluate such a functional test.
- PAD mechanisms also fall into the area of vulnerability assessment, as the use of a Presentation Attack Instrument (PAI) (such as false or altered documents) against the age assurance system is an attempt to circumvent the security functionality of the Age Assurance System or component.

The differences between the two perspectives can best be visualized using a concrete example. In the area of functional testing, the benchmarking analysts' concern regarding PAD is to verify that the TOE meets certain performance requirements. The PAD mechanism has to perform within a certain range of performance.

Testing can be achieved by the use of a standardized toolbox. Beside some dedicated requirements on testing and documentation, this situation is very close to the situation in classical performance testing. Having passed the test from a functional perspective is a prerequisite to start the vulnerability assessment. If the PAD mechanisms would not work within sufficient performance limitations, any kind of vulnerability assessment would be useless. In the vulnerability assessment, the benchmarking analyst will then try to circumvent the PAD mechanism, working within the limitations of the attack potential of the current benchmarking. This can lead to a situation in which a TOE passes the functional test but where the benchmarking analyst can build a so-called "golden fake" that reproducibly breaches the security functionality of the TOE. If this happens, the TOE fails the security benchmarking even though it showed good performance during functional testing.

As a basic rule, it can be said that one successful attack against a TOE (always under consideration of the maximum attack potential) will make the security benchmarking fail. This is one of the major differences of a security benchmarking compared to a pure performance test.

4.4.4 Vulnerability assessment

4.4.4.1 Typical attack scenarios

Specific kinds of attacks against age assurance systems exist. Presentation attacks are only one very prominent example. Also, for example, a biometric system can be vulnerable against a hill-climbing attack or an age verification system can be vulnerable to a credential-stuffing attack.

It is important that the benchmarking analyst considers typical and well-known presentation attacks during the benchmarking of an age assurance system. While the system is not necessarily vulnerable to all attacks, as a starting point for a vulnerability analysis, it is important that all typical attacks are considered. These can be seen as a minimum list of attacks to be considered. They do not claim to be complete and the benchmarking analyst will, in any case, develop additional attack scenarios during benchmarking.

4.4.4.2 Rating attacks

Guidance for the security benchmarking of age assurance systems introduces a dedicated scheme to rate the attack potential of attacks against age assurance systems as minimal, basic, enhanced-basic, moderate, high, or beyond high. The level chosen for the vulnerability analysis is one of the most important aspects of the chosen EAL. This decision basically answers the question against which attack potential a TOE is expected to be resistant.

The benchmarking analyst will perform their vulnerability assessment and penetration testing “only” up to the chosen level. Common Criteria uses a dedicated list of criteria to classify an attack in general. To reflect the dedicated characteristics of attacks against age estimations systems, an extension and interpretation of the standard attack rating scheme are necessary.

Annex A - Approaches to Measurement of Continuous Age Assurance (Informative)

A1. About Continuous Age Assurance

Continuous age assurance technologies provide an estimate of a person's age. The closer this estimate is to the true age of that person, the more accurate the estimate. In the following table a set of measures are defined that can be used to assess the accuracy of the age technology given a set of samples or testing data. Since the outcome, age, is a continuous outcome, the measures below can all be applied to continuous data. Each measure is defined using the following parameters:

- The true (or observed) age of sample i: o_i.
- The predicted age of sample i: p_i.
- The number of samples tested: n.

Table 1 – Measures Applicable to Age Estimation Technologies

Measure	Definition	Meaning/Notes
Error	$E_i = p_i - o_i$	The error is the difference between the predicted and true age of sample i. It is impacted by whether the prediction is an over or underestimate of the true age (it will be positive for the former and negative for the latter). The distribution of errors across all n samples can be visualised by a histogram, which will highlight the shape of the distribution (is it symmetrical or skewed) and the range of errors across the full sample.
Absolute Error	$AE_i = \lvert p_i - o_i \rvert$	The absolute error is the absolute difference between the predicted and true age of sample i. The error is the magnitude of the size of the difference between the predicted and observed ages (i.e., it is positive irrespective of whether the prediction is an over or underestimate). The absolute error is a useful overall measure of accuracy, and we will focus on it below when defining measures of central tendency and spread over the sample distribution of absolute errors. Note that there are cases when understanding whether an age estimate is over or underestimating the true age is informative; particularly for model performance improvements and checking for differences between protected characteristics, for example.

Measure	Definition	Meaning/Notes
		As above, the distribution of absolute errors across all n samples can be visualised by a histogram, which will highlight the shape of the distribution (is it symmetrical or skewed) and the range of absolute errors across the full sample.
Mean Absolute Error	$MAE = \frac{\sum_{i=1}^{n} \left\lvert (p_i - o_i) \right\rvert}{n}$	The mean absolute error is the central value of the absolute errors; it is the average value of the sample. There is another measure of central tendency that can be useful, particularly if the distribution of the errors suffers from outliers, known as the median (note the mean and median are identical in symmetric distributions).
Median Absolute Error	The median error ($MEDAE$ ¿ is the middle number in the sorted (ascending or descending) list of absolute errors.	The median is sometimes used rather than the mean when the distribution of absolute errors is heavily skewed. In this instance, the mean may be influenced by outliers (i.e., a small number of samples with particularly large errors) and not be a reliable measure of central tendency. The mean is the most frequently used measure of central tendency and will continue to be the focus here, but the median is worth consideration in these specific circumstances.
AE Standard Deviation	$SD_{AE} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(AE_i - MAE \right)^2}$	The standard deviation is a measure of the amount of variation or spread over the distribution of absolute errors. A low standard deviation indicates that the values are close to the MAE and a higher value indicates a larger spread. Other measures of spread can be calculated for example, the range (the maximum minus the minimum absolute errors) and the interquartile range.
MAE 95% Confidence Interval	$CI_{MAE} = MAE \pm 1.96 \frac{SD_{AE}}{\sqrt{n}}$ Note at a minimum the lower bound is 0 and should be truncated if needed	A confidence interval quantifies the uncertainty associated with an estimate, such as the MAE. The interval is calculated from the sample and is the range of values in which we estimate the MAE to lie with 95% confidence. A 95% confidence level is recommended as this is what is used most in ISO standards and the statistical community. All estimates such as the MAE should be

Measure	Definition	Meaning/Notes
		reported with a confidence interval to understand and quantify their uncertainty. Without this additional measure, they are not very informative. In this example, the number 1.96 is the critical value of the Normal distribution based on a 95% confidence level. It is dependent on the data meeting the Central Limit Theorem which establishes that for a large enough sample, the sample average tends to a normal distribution. Typically, a sample size of more than 30 is deemed large enough.
AE 95% Prediction Interval	$PI_{AE} = MAE \pm 1.96\, SD_{AE}$ Note at a minimum the lower bound is 0 and should be truncated if needed	A prediction interval or predictive confidence interval quantifies the uncertainty associated with the absolute error of a single individual. It is the range of values in which we estimate the absolute error of the individual to lie with 95% confidence.

A2. Observations on Age Estimation Measurement

There are several key points to bear in mind:

1. The MAE is a useful overall measure to summarise the accuracy of an age estimation technology on average. The MAE is a measure of central tendency of the sample. An age technology with low MAE tells you that you have a good “average” performance over the sample or training data set.

2. However, the MAE should not be looked at in isolation and, on its own, is not sufficiently informative. Reporting the absolute error standard deviation quantifies the spread of the distribution. If the standard deviation is low as well as the MAE, then the performance is not only good on average, but also across the entire dataset. For example, if two different technologies have both been assessed with a MAE of 2.5 years they could be assessed as having the same level of accuracy, but this is not the case if one has an AE standard deviation of 0.25 years and the other has an AE standard deviation of 1.5 years. Looking at the MAE on its own would not have highlighted that the performance of the technology with the lower standard deviation is better overall.

3. The spread of the distribution can be quantified further by producing a 95% absolute error prediction interval. For example, for a standard deviation of 2, an individual is predicted to lie within +/- 3.92 years of the MAE with 95% confidence, compared to +/- 1.96 years with a standard deviation of 1.

4. The distribution of absolute errors should be visualised using a histogram to understand its shape and the spread or range of absolute errors. If the distribution contains outliers, the MEDAE should also be reported.

5. To understand whether an age technology is over or under predicting ages, the distribution of the errors (rather than absolute errors) will help. This can be useful for identifying areas to improve performance and for investigating differences between protected characteristics, for example).

6. The MAE should not be reported without its associated 95% CI to quantify its uncertainty. The smaller the 95% CI the more precise the estimate.

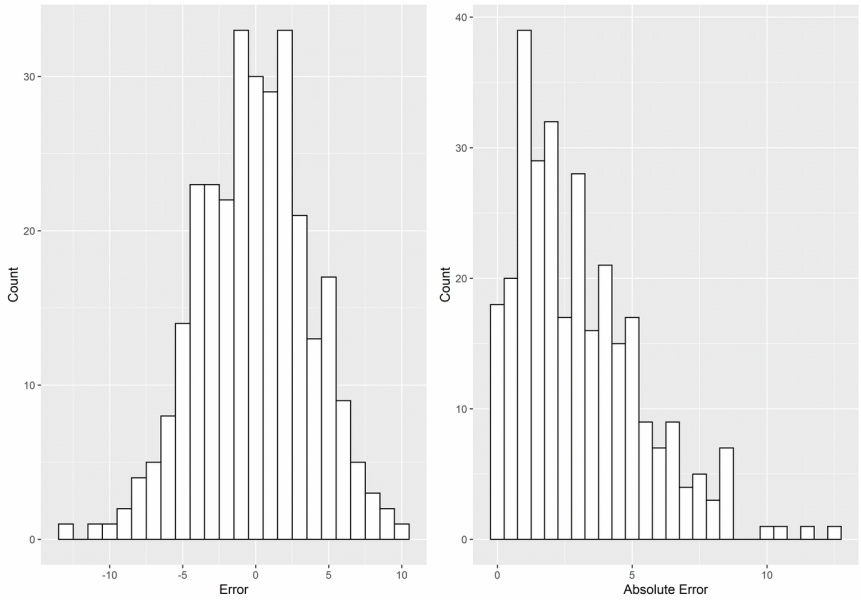
Identifying what is an acceptable level of MAE and AE standard deviation (i.e., how low does the MAE need to be for the accuracy of the age estimation technology to be deemed acceptable) is a decision for regulators.

Worked Example for Age Estimation Measurement

A worked example of these measures is given below based on a pseudo data set made up of 300 samples aged between 14 and 18.

Firstly, the errors and absolute errors are calculated for each of the three samples and plotted using histograms below (with the errors on the left and absolute errors on the right).

Figure 1 **- Histogram of Errors for Age Estimation measurement**



The histogram of the errors illustrates that they are reasonably symmetrical (indicating that there is unlikely to be a bias towards over or under prediction) and the errors range between –12.7 and 10.34.

The histogram of the absolute errors illustrates that these range between 0 and 12.7 with the peak of the distribution greater than 0, but less than 5.

The mean absolute error (MAE) is calculated to be 3.0 years and the median absolute error (MEDAE) 2.6. These two measures are similar as there are no large outliers within the data set.

To illustrate the impact of outliers on the mean if we added another three observations with errors greater than 25 the MAE changes to 3.3 but the MEDAE remains 2.6. The impact of the outlier is therefore not too large, and the data set would have to suffer from very large outliers to suggest that the MAE was not reliable.

The 95% confidence interval of the MAE is [2.8, 3.3] indicating that the MAE estimate is reasonably precise with a margin of error of 0.35 years. The standard deviation of the absolute errors is 2.3 years and 95% of the absolute errors lie between 1.2 and 8.5 years. The 95% prediction interval for the absolute error of an individual is [0, 7.5].

Test protocols shall be in place to secure effective, repeatable testing of the target of benchmarking – the system that is under test. Test protocols should describe the capture methodology setting out the subjects, devices and environmental circumstances that will be used to present the test to the target of benchmarking.

The capture subject describes the individual who is going to be subject to the age verification, categorisation or estimation process.

Note: ISO/IEC 2382-37:2017 - Information technology — Vocabulary — Part 37: Biometrics, 3.7.3

The test protocols may use members of a test crew, who are real people with true identities – called bona fide identity subjects - or they could use a series of simulated identities which have existed over time (i.e., may have been used in tests previously) – called Avatars. In International Standards they are referred to as subversive capture subjects. It may not be necessary to utilise a real or simulated identity depending on the Target of Benchmarking.

Note: ISO/IEC 2382-37:2017 - Information technology — Vocabulary — Part 37: Biometrics, 3.7.17

The presentation of a capture subject should also record the facial orientation – typically at indices up to 15° of centre, between 15° and 30° of centre and greater than 30° of centre. By default, <15° of centre should be used as test methodology. It is important to note, however, that the operational capability of age assurance technologies may need testing at much wider orientations – for instance, some are still designed to be effective at 90°orientation (i.e., a profile shot of the subject).

The capture device is the equipment or system that we are going to utilise to collect the signal from the capture subject to perform the test.

Note: ISO/IEC 2382-37:2017 - Information technology — Vocabulary — Part 37: Biometrics, 3.4.1

A capture device could be:

- Integrated/Purpose Built in an age assurance technology
- A smart device or connected device (like a mobile phone)
- A Web Camera
- A Microphone/Telephone (Audio Only Testing)
- A Scanner

A3. Presentation Attack Detection

Presentation attack detection is the process of determining if an Age Assurance system is susceptible to being ‘spoofed’.

This can involve the presentation of attack instruments such as:

- Pseudo Identities
- Mannequins
- Masks
- False Identity Documents
- False Instruments

- Tamper Evident Instruments
- Genuine Instruments that have been amended
- Disfigured Instruments

Biometric presentation attack is set out in international standards BS ISO/IEC 30107-3:2017 - Information technology — Biometric presentation attack detection – Part 3: Testing and Reporting.

When a non-living object that exhibits human traits (an "artifact") is presented to a camera or biometric sensor, it is called a "spoof." Photos, videos on screens, masks, and dolls are all common examples of spoof artifacts. When biometric data is tampered with post-capture, or the camera is bypassed altogether, that is called a "bypass." A deepfake puppet injected into the camera feed is an example of a bypass. There are no NIST/NLVAP lab tests available for PAD Level 3, or Levels 4 & 5 bypasses, as those attack vectors are missing from the ISO 30107-3 standard and thus all associated lab testing.

- **Table** 2 **- Presentation Attack Detection - Artefact Types**

Artefact Type	**Description**
Level 1	Hi-res paper & digital photos, hi-def challenge/response videos and paper masks.
Level 2	Commercially available lifelike dolls, and human-worn resin, latex & silicone 3D masks
Level 3	Custom-made ultra-realistic 3D masks, wax heads, etc
Level 4	Decrypt & edit the contents of a 3D FaceMap™ to contain synthetic data not collected from the session, have the Server process and respond with Liveness Success.
Level 5	Take over the camera feed & inject previously captured video frames or a deepfake puppet that results in the AI responding with "Liveness Success."

As age assurance systems become more broadly deployed through information society services, it will be necessary to continuously review and address threats associated with both simple presentation attack, but also much more sophisticated attacks which will become prevalent and easily accessible to young people seeking to circumvent age assurance systems.

Annex B - Approaches to Measurement of Binary Age Assurance (Informative)

B1. About Binary Age Assurance

This informative annex identifies the measures applicable to binary age assurance outputs, where there is a positive declaration with only two possible states – 'yes' or 'no'.

Binary age assurance techniques are the output of posing a question to which there are only two possible answers to a question– e.g. Is this person aged over 18? Yes, or No. These approaches are more generally associated with age verification methods using access to information, data, documents or records to gain a level of confidence in the truthfulness of the binary outcome. It includes statistical analysis of age verification techniques.

Measurement of Age Verification Techniques

The objective of age verification is to identify whether a person is:

- Scenario 1: Over an age threshold (e.g., 13 or 18) to stop access to age-inappropriate products/materials/services.
- Scenario 2: Under an age threshold to access safe places where no adults are allowed for safeguarding issues (except for appointed safeguarding monitors).
- Scenario 3: Between one specified age and another. In the ICO's Children's Code, these are pre and early-literacy (0-5), core primary school years (6-9), pre-teen years (10-13) and transition to adulthood years (14-17), and adults (18+) to access services in each age group, but they could be any categorisation of age.

A technology may simply provide verification alone or could be an age estimation technology that applies the age threshold to the estimated age. In either case, the outcome is binary as follows:

- Scenario 1: A person is identified as over the age threshold (positive) or under (negative).
- Scenario 2: A person is identified as under the age threshold (positive) or over (negative).
- Scenario 3: A person is identified as within the specified age range (positive) or outside (negative).

As such, the measures to assess accuracy must be tailored to a binary outcome. For those technologies that produce a continuous outcome, the accuracy measures defined in the age estimation section can be applied.

Measures that can be used to assess the accuracy of the age verification technology are defined below. Scenario 1 is used to define and illustrate these measures, but they are applicable to all three scenarios described above. The measures are all based around the matrix below that gives the frequency of the results according to the observed and predicted age thresholds of a sample or training data set.

Table 3 – Matrix Describing The Performance of the Age Verification Technology

		Predicted	
		Positive: Over Threshold	Negative: Under Threshold
Observed	Positive:	True Positives (TP)	False Negatives (FN)

Over Threshold		
Negative: Under Threshold	False Positives (FP)	True Negatives (TN)

- True Positives: the number of samples that are both observed and predicted to be over the threshold (i.e., the number of samples correctly classified as over the threshold).
- True Negatives: the number of samples that are both observed and predicted to be under the threshold (i.e., the number of samples correctly classified as under the threshold).
- False Positives: the number of samples that are observed to be under the threshold but predicted to be over it (i.e., the number of samples incorrectly classified as being over the threshold).
- False Negatives: the number of samples that are observed to be over the threshold but predicted to be under it (i.e., the number of samples incorrectly classified as being under the threshold).

In an ideal scenario, all samples would either be true positives or true negatives, which means that no sample had been incorrectly classified.

Possible measures to assess accuracy are defined below.

Measures Applicable to Age Verification Technologies

Measure	Definition	Meaning/Notes
True Positive Rate (TPR) **Also known as: Sensitivity, Recall, or Probability of Detection**	$TPR = \frac{TP}{TP + FN}$	The sensitivity is the technology's ability to correctly detect people who are over the age threshold. It is the proportion of the sample who have been predicted as being over the age threshold among those who are over the age threshold.
True Negative Rate (TNR) **Also known as: Specificity or Selectivity**	$TNR = \frac{TN}{FP + TN}$	The specificity is the technology's ability to correctly detect people who are not over the age threshold. It is the proportion of the sample who have been predicted as being under the threshold among those who are under the age threshold.
False Positive Rate (FPR) **Also known as: Fall-Out or Probability of False Alarm**	$FPR = \frac{FP}{FP + TN}$	The false positive rate is the technology's probability of false alarm (i.e., incorrectly identifying someone as being over the age threshold). It is the proportion of the sample who have been predicted as being over the

Measure	Definition	Meaning/Notes
		threshold among those who are not over the age threshold.
False Negative Rate (FNR) **Also known as: Miss Rate**	$FNR = \frac{FN}{TP + FN}$	The false negative rate is the technology's miss rate (i.e., incorrectly identifying someone as being under the age threshold). It is the proportion of the sample who have been predicted as being under the threshold among those who are over the age threshold.
Accuracy	$ACC = \frac{TP + TN}{TP + TN + FP + FN}$	The accuracy is the proportion of the sample that have been correctly classified as being over or under the age threshold. Note assumes that the balance between samples of over and under the age threshold is reasonable.
Positive Predictive Value (PPV) **Also known as: Precision**	$PPV = \frac{TP}{TP + FP}$	The PPV is the proportion of the sample correctly identified as being over the age threshold given that they have been predicted as being over the age threshold.
Negative Predictive Value (NPV)	$NPV = \frac{TN}{TN + FN}$	The NPV is the proportion of the sample correctly identified as under the age threshold given that they have been predicted as being under the age threshold.
False Discover Rate (FDR)	$FDR = \frac{FP}{FP + TP}$	The FDR is the proportion of the sample incorrectly identified as over the age threshold given that they have been predicted as being over the age threshold.

Measure	Definition	Meaning/Notes
False Omission Rate (FOR)	$FOR = \frac{FN}{FN + TN}$	The FOR is the proportion of the sample incorrectly identified as under the age threshold given that they have been predicted as being under the age threshold.
Positive Likelihood Ratio (LR+)	$LR+¿\frac{TPR}{FPR}$	The positive likelihood ratio is the value in performing the test. It is the ratio of the true positive and false positive rates. The greater the value over 1 indicates the greater the probability that a positive test result is evidence that the person is over the age threshold.
Negative Likelihood Ratio (LR-)	$LR-¿\frac{FNR}{TNR}$	The negative likelihood ratio test is the value in performing the test. It is the ratio of false negative and true negative rates. The closer the value to 0 the greater the probability that a negative test result is evidence that the person is under the age threshold.

Ideally, a technology would correctly classify all persons (i.e., 100% accuracy).But this is unrealistic. It is important that, based on the implications of an incorrect classification, technology minimises false positives which are defined as follows for each scenario:

- Scenario 1 False Positives: those under the age threshold are incorrectly classified as over it allowing access to age-inappropriate content.
- Scenario 2 False Positives: those over the age threshold are incorrectly classified as under it allowing adult access to safe spaces causing safeguarding issues.
- Scenario 3 False Positives: those outside of the age range incorrectly classified as within it allowing access to content tailored to a different age group.

In all cases false positives have the potential to cause harm (particularly in scenarios 1 and 2). False negatives should be minimised where possible (e.g., in scenario 1 where someone over the age threshold has been identified as under), but these are less critical since they result in inconvenience (and potential economic consequences if it results in users abandoning the technology) rather than harm. Therefore, when assessing the above measures, it is important to note that false positives are more critical to minimise.

It is worth noting that in all the metrics above, 95% confidence intervals could be calculated to quantify their uncertainty. However, if the sample size has been correctly calculated (with inputs that are aligned to the

deployment and expected outcomes of the technology) then the confidence intervals of the metrics should be close to the margin of error defined in the sample size calculation.

B2. Document Authenticity

Presentation attacks utilising false identity documentation or records are affected by the assessment of the capability to detect documents and extract age attributes.

For authentication, documents should be classified (scored) according to their inherent features that are designed to provide detectable security features.

- **Table** 4 **- Classification of Document Authenticity Security Features**

Document Type	Description
Tier 1	No material security features available, and no fraud evaluation can be performed. Extraction only. Documents in this tier sometimes include hand-written documents.
Tier 2	A low security document where only basic fraud checks can be performed and confidence in authenticity (based on a digital photo) is low. • No cross-comparison possible due to missing Machine-Readable Zone (MRZ) or barcode; and/or • The documents may not have consistent template format and/or fonts.
Tier 3	Documents in this tier lack advanced security features and are easier to execute fraud attacks, but still carry sufficient security features to enable automated verification using data cross-comparison, checksums, and other logical checks. Documents have a consistent template format and font within a version. Documents in this tier SHALL Include one or more of the following features: • machine readable zone (MRZ) • barcode Tier 3 also SHOULD meet requirements for Tier 2.

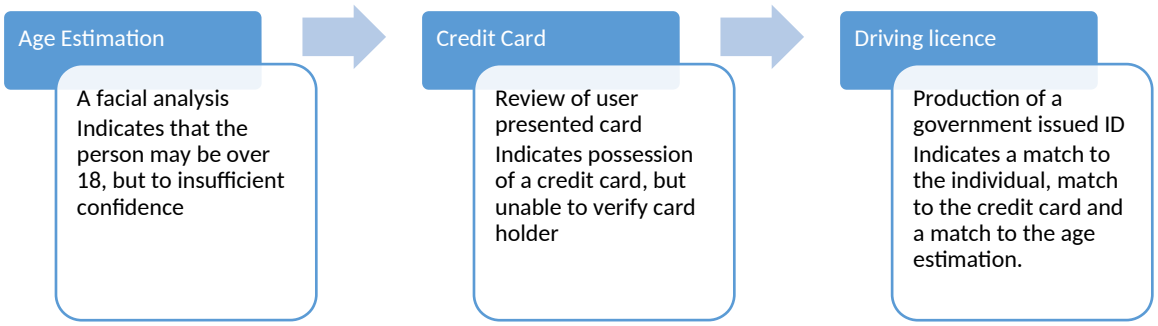
Document Type	Description
Tier 4	Documents of this tier are highly secured documents with state-of-the-art security features. Documents in this tier SHALL include one or more of the following technologies: • optically variable ink (OVI), holograms, watergrams • guilloche (e.g., intricate and subtle patterns of thin interwoven lines) • tactile laser engraving • micro printing • ghost image Tier 4 also SHOULD meet requirements for Tier 3.
Tier 5	Documents of this tier are highly secured documents with state-of-the-art security features. Documents in this tier SHALL include one or more of the following technologies: ● embedded chip technology (e.g., contact card, RFID, NFC)

B3. Age Verification: Waterfall Technique

The waterfall technique for age verification is a breakdown of age assurance activities into linear sequential phases, where each phase depends on the output of the previous one and corresponds to a series of decisions providing greater or lesser levels of confidence in the age assurance gained from the process.

Some technologies rely on multiple gateways to assess whether a person is, for example, over an age threshold such as 18. At each gateway a new source of information or database is added (for example, electoral register, credit card reference data, mobile phone data etc.). At each gateway a person is assigned as over 18 (positive) or insufficient evidence to identify as over 18 (negative). The technology passes a certain number of gateways until they are confident that those have not been assigned as being over 18 are under 18.

The approach to a 'waterfall technique' is that the cumulative results of the age assurance components are greater than the individual results of each component on its own. The whole is greater than the sum of its parts. This presents a statistical difficulty, which needs to be explored further when considering Trust Frameworks and interoperability. Whilst, in theory, the whole is greater than the sum of its parts, in statistical theory, this propagation of uncertainty results in the errors associated with each part being multiplied together. This fails to recognise the cumulative knowledge gained by the multiple components, so a method of statistically recognising this is required.



Like any other age verification technique, the same binary measures can be applied to the final classifications after a person has reached the last gateway. To assess the accuracy of each individual gateway, the technology can also calculate the overall accuracy at each (assuming those who have not been assigned as over 18 are under 18 as there is no evidence to the contrary) with the expectation that this overall accuracy will improve with each additional gateway added.

A well-designed waterfall technique is privacy protecting, as the sequence of data gathering is directly tailored to the level of confidence sought before the process is completed. However, a poorly designed sequencing can lead to the collection of unnecessary data. It could also potentially be more intrusive and could breach the data minimisation principles.

B4. Observations on Age Verification Measurement

The accuracy measure gives a good indication of the overall accuracy of the technology. However, on its own, it does not provide additional information on whether the technology's misclassifications are because of false positives or false negatives (and we know that here false positives are more problematic).

Key points to note are:

1. The results of an age verification assessment can be summarised by a confusion table, which details the four different combinations of possible results (true positives, true negatives, false positives and false negatives).

2. The overall accuracy (proportion of correctly classified samples) is a useful overall measure and should be reported. But in isolation, it does not provide any information on the type of errors that are present (false positives or false negatives).

3. Reporting both the sensitivity (TPR) and specificity (TNR) informs the user about the prevalence of different errors. The greater the sensitivity, the fewer the false negative errors and the greater the specificity, the fewer the false positive errors.

4. Maximising the TNR/minimising the FPR may be more of a priority than the TPR/FNR since false positive errors have the potential to cause harm.

5. Ultimately, the system will be judged on its False Positive Rate (FPR), but this should not be considered on its own without also considering the sensitivity and specificity of the age assurance system.

6. Predictive values are helpful to users of technology. Given that the technology has predicted a result, what is the probability that it is right? In this case maximising the PPV, minimises the FDR (the more critical errors).

Sensitivity & Specificity

The sensitivity and specificity of the age assurance component is a crucial element of understanding the overall efficacy of the system

- A high sensitivity (TPR) means that the technology will rarely misclassify those who are over the age threshold. The false negative rate is 1-TPR.

- A high specificity (TNR) means that the technology will rarely misdiagnose those who are under the age threshold. The false positive rate is 1-TNR.

Based on the above, the primary aim of the technology is to maximise TNR (and therefore minimise FPR). Of course, one way to have a 100% TNR and 0% FPR, is to assign everyone as under the age threshold (using scenario 1 as an example), but of course this is not practical. Therefore, there must be a trade-off between sensitivity and specificity, but the weighting to specificity is higher.

Predictive Values

The predictive values are likely to be helpful to users of the technology. Sensitivity and specificity condition on the true outcome, e.g., given the true outcome, what is the probability that the technology got the classification right? However, when the technology is being used, the true age of the person is unknown and therefore we need to ask: given that the technology says the person is over the age threshold, what is the probability that is correct? Both PPV and NPV are important, but maximising PPV is imperative; the probability that someone who is classified as being over the age threshold is over the age threshold. Maximising PPV by default minimises the False Discover Rate (FDR) since PPV = 1 – FDR and we want to reduce the chance of a false discovery (the probability that someone who is identified as being over the age threshold but is in fact under it).

Information Retrieval

Information retrieval/AI often focus on precision (sensitivity) and recall (PPV), but these measures do not consider true negatives and therefore could bias predictions if they are the only focus. In information retrieval, the number of true negatives is unknown and much larger than the true positives; this does not hold in this application.

Annex C – Approaches to outcome error parity and fairness (informative)

C1. About outcome fairness

Outcome fairness is the best measure to quantifiably assess how a technology owner has implemented all four forms of fairness and one method to do so is to ensure that error rates are equitably distributed across different subgroups of the population.

For continuous (age estimation) techniques that produce a continuous outcome, error parity is similarly the focus but in this case the measures include:

- Mean Absolute Error Parity: ensuring that the overall accuracy of the technology is equivalent between different population subgroups.

- Mean Error Parity: ensuring that the technology is not biased towards over or under prediction for different population subgroups.

For binary (age verification) techniques that produce a binary outcome, measures include:

- True Positive Parity: ensuring that the accuracy of the technology is equivalent between different population subgroups. Also known as ‘equal opportunity’ fairness.

- False Positive Parity: ensuring that the error rate of the technology is equivalent between different population subgroups.

- Positive Predictive Value Parity: ensuring that the precision of the technology is equivalent between different population subgroups.

- In practice the accuracy or error rates for a technology will never be the same across different population subgroups due to the inherent variability of the technologies. Defining what an acceptable difference between these measures for subgroups to accept parity between the subgroups is one that must be defined by regulators.

It must be identified which protected characteristics are at risk of bias or discrimination and therefore error parity examined for these chosen characteristics. While it is relatively simple to examine protected characteristics individually, it is important to acknowledge the potential for intersectional biases where there are biases within combinations of protected characteristics (such as race and gender in combination). Investigating intersectionality is more difficult since there are likely to be many combinations to consider and the sample size within each combination will be small.

To investigate error parity fully, ideally there would be the equivalent sample size in each population subgroup as the size recommended for the full subgroup so that the estimate for each subgroup is estimated with the same level of confidence and margin of error. This is unlikely to be possible, but it is important to ensure that each subgroup has a reasonable sample size.

C2. Ambient Lighting

It is important to note that the performance of electronic detection devices, such as smartphone cameras, webcams or scanners, are susceptible to diminished performance in different ambient lighting conditions.

The ambient lighting can have a significant impact on the efficacy of the data capture, so tests should be carried out under controlled lighting conditions. The lighting can be directed ambient to the presentation object (i.e., the person being age estimated) or the detection device (i.e., the camera) or both.

The following ambient lighting choices should be considered:

- Bright LED Gantry (such as may be found in a retail shop) – around 700 lux
- Sodium Low Level (such as may be found in a pub or restaurant) – around 70 lux
- Strobe Lighting (such as may be found in an entertainment venue)
- Ultraviolet Lighting (such as may be used in a scanner detection devices)
- Multi Colour Lighting (such as may be emitted by a gaming machine)
- Outdoor Daylight
- Outdoor Nightlight

In addition to the effect of lighting on the presentation object, there can be adverse effects of lighting on the detection device, caused by issues like:

- Glare - which occurs when one part of the visual field is much brighter than the average brightness to which the detection device is adapted
- Colour effects – which occurs when the detection device is lit by different artificial light sources, or by daylight under changing sky conditions, may appear to vary in colour
- Under monochromatic light sources - such as low-pressure sodium discharge lamps, colours will not be identifiable a detection device may not perform properly
- Stroboscopic effects – can confuse detection devices. When the magnitude of these oscillations is great, Presentation Attack Instruments will appear to be stationary or moving in a different manner. This is called the stroboscopic effect.
- Flicker - Light modulation at lower frequencies (about 50 Hz or less) which is visible to most people, is called flicker. Detection Devices can be sensitive to flicker, and it is especially detectable at the edges of the visual system's field of view.
- Veiling reflections - are high luminance reflections which overlay the detail of the Presentation Attack Instrument. Such reflections may be sharp-edged or vague in outline, but regardless of form they can affect Detection Device performance.
- Infrared and ultraviolet radiation - Some lamp designs also produce significant emissions at infrared and ultraviolet wavelengths, both of which are invisible; some Detection Devices also rely upon Infrared and ultraviolet radiation.

We do not believe that ambient temperature, humidity, pressure or other climatic conditions have a material impact on the efficacy of the Target of Benchmarking.

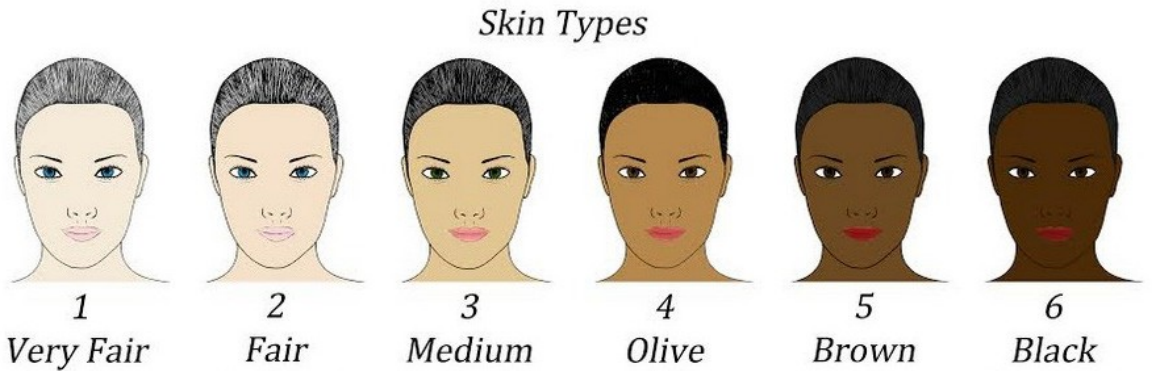
Note: ISO 8995-1:2002 - Lighting of workplaces — Part 1: Indoor

C3. Data subject skin tone

Biometric age estimation systems can be adversely affected by inherent skin tone bias. This is all dependent on the range of training images that are used. The Fitzpatrick Scale[1] 1 – 6 is used to determine the skin tone of our presentation attack assets. All PAI assets should be assigned a skin tone score.

[1] Fitzpatrick, T. B. (1975). "Soleil et peau" [Sun and skin]. *Journal de Médecine Esthétique* (in French): 33–34

- **Table** 5 **- Fitzpatrick Scale of Skin Tone Types**



C4. Sample size and breakdown

To calculate a sufficient sample size when testing an age estimation or verification technology, the objective of the assessment must be defined. This would typically reflect how the technology would be deployed and what metric is being used to assess its accuracy. Some illustrative examples are given below for both an age estimation and verification technology.

Age Estimation Technology

If the technology is being deployed to estimate the ages of teenagers, for example, the objective of the test would be:

What is the MAE of an age estimation technology for those who are 13-18 years old?

Here, the primary accuracy measure is MAE to a sample size formula for estimating a population mean can be used to calculate the sample size. The formula is as follows:

$$N = \frac{N \cdot X}{(N + X - 1)},$$

where,

$$X = \frac{Z_{\alpha/2}^2 \cdot \sigma^2}{MO\,E^2},$$

and $Z_{\alpha/2}$ is the critical value of the Normal distribution at $\alpha/2$(e.g., for a confidence level of 95%, $\alpha$is 0.05 and the critical value is 1.96), MOE is the margin of error, σ^2 is the population variance, and N is the population size. Note that a Finite Population Correction has been applied to the sample size formula.

This sample size calculation provides the recommended number of samples required to estimate the true population mean (in this case the MAE) with the required margin of error and confidence level.

The margin of error is the level of precision required. This is the plus or minus number that is often reported with an estimated mean and is also called the confidence interval. It is the range in which the true population mean is estimated to be. Note that the actual precision achieved after you collect your data will be more or less than this target amount, because it will be based on the population variance estimated from the data and not your expected variance.

The confidence level is the probability that the margin of error contains the true mean. If the study was repeated and the range calculated each time, you would expect the true value to lie within these ranges on 95% of occasions. The higher the confidence level the more certain you can be that the interval contains the true mean.

The population size is the total number of distinct individuals in your population. In this formula we use a finite population correction to account for sampling from populations that are small. If your population is large, but you do not know how large, you can conservatively use 100,000. The sample size does not change much for populations larger than 100,000.

The population variance tells you how the data points in a specific population are spread out. It is the average of the distances from each data point in the population to the mean, squared. An estimate of the expected variance is required for the calculation and may be obtained from previous tests carried out on the technology.

The table below shows how the sample size changes as the inputs change (assuming a population size of 100,000). The larger the sample size, the more certain you can be that the estimates reflect the population, so the narrower the confidence interval. However, the relationship is not linear, e.g., doubling the sample size does not halve the confidence interval.

Confidence Level	**Margin of Error**	**Population Variance**		
		4	**9**	**16**
90%	**0.25**	173	389	688
95%		246	551	974
99%		423	947	1671
90%	**0.5**	44	98	173
95%		62	139	246
99%		107	239	423
90%	**1.0**	11	25	44
95%		16	35	62
99%		27	60	107

For example, for a technology in this deployment setting that has an expected MAE variance of 9 years (or standard deviation of 3 years), a sample size of 139 would be needed to achieve a margin of error of 0.5 years with 95% confidence (i.e., to estimate the MAE within plus or minus half a year with 95% confidence), but the sample size would need to increase to 551 for a margin of error of 0.25 years.

Note that if the population variance was underestimated, for example, then for the same sample size, the actual margin of error calculated from the sample would then be larger (the confidence interval would be greater than plus or minus the margin of error stated in the sample size calculation).

Once a sample size has been calculated, the test subjects it is made up with should reflect its deployment and therefore, in the above example, be made up of 13- to 18-year-olds and the breakdown of characteristics

should be representative of the population in relation to age, gender and skin tone (e.g., the proportion of females to males should be approximately 50/50).

Bibliography

[1] ISO #####-#, *General title — Part #: Title of part*

[2] ISO #####-##:20##, *General title — Part ##: Title of part*