The Impact of Age-Related Variables on Facial Comparisons with Images of Children: Algorithm and Practitioner Performance

Dana Jaclyn Michalski

B.Psych (Hons) B.BehSc.

School of Psychology

University of Adelaide

November 2017

THE UNIVERSITY of ADELAIDE

Thesis submitted for the degree of Doctor of Philosophy

Contents

List of Figures

List of Tables

Abstract

Determining the identity of children is critical for many national security agencies for example, to aid in the fight against child exploitation, trafficking, and radicalised minors, as well as for passport control and visa issuance purposes. Facial comparison is one method that may be used to achieve this. Facial comparison can be conducted using an algorithm (within a facial recognition system), manually by a facial comparison practitioner, or by a combination of the two. Much of the previous research examining facial comparison performance of both algorithms and practitioners has been conducted using images of adults. Due to the substantial amount of age-related facial growth that occurs in childhood, compared to adulthood, it is likely that performance will be poorer with images of children. The overarching aim of the research therefore, was to determine the impact of age-related variables, namely chronological age and age variation (the age difference between images) on facial comparison performance of algorithms and practitioners with images of children.

Study 1 involved consultation with national security agencies and algorithm vendors to identify the key requirements to examine in this thesis. After reviewing the literature to identify research gaps, five empirical studies were conducted. To ensure the studies were as operationally relevant as possible, a large database containing several million controlled images of children and adults was sourced, and five state-of-the-art facial recognition algorithms were employed. In addition, facial comparison practitioners from a government agency participated in the practitioner studies. Study 2A compared algorithm performance with images of children to performance with images of adults. Study 2B compared practitioner performance with images of children to performance with images of adults. Study 3A examined algorithm performance with images of children at each chronological age in childhood (0–17 years) and age variations ranging from 0–10 years apart. Study 3B examined

practitioner performance on the same age-related variables examined in Study 3A. Study 4 demonstrated how the data collected in Study 3A and 3B could be used to answer agency specific questions.

This thesis concludes with a series of recommendations for both the algorithm and practitioner domains, as well as future research directions designed to improve knowledge and performance regarding facial comparisons with images of children.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

In addition, I certify that no part of this work will, in the future, be used in a submission for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968. I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

……

Dana Michalski

November 2017

Acknowledgements

I would like to thank my PhD supervisors, Dr Carolyn Semmler and Dr Rebecca Heyer for supporting me during my candidature. Dr Carolyn Semmler, thank you for agreeing to take me on and your words of encouragement over the last few years. Your enthusiasm and advice has made this experience more rewarding. Dr Rebecca Heyer, thank you for your support as a PhD and work−based supervisor. Your work ethic and passion to research keeps me motivated when things get tough. I hope we can continue to work together well into the future.

This PhD would not have been possible without the encouragement of Dr Brett McLindin. Without your ability to convince me that "doing a PhD is easy" (ha!), I would never have considered undertaking a PhD that crosses two distinct disciplines. Once again, you threw me in the deep end and because of that, I have become a stronger researcher.

I would also like to acknowledge the Defence Science and Technology Group (DST Group) and members of the Biometrics Team. Thank you to all team members who have provided support along the journey. A special mention must go to John Stwein and Martyn Hole for providing assistance through the development of a special plotting program that could handle the large amounts of algorithm data thrown at it. The constant back and forth to get the plotting program developed just right seemed like an additional burden at a time when there was already chaos. The final product is something you should both be proud of and will be used in future research by the team. To Peter Aquilina and Dr Kathy Hanton, thank you for taking the time to view the 23,760 image pairs selected for Study 3B. Although not a fun task, this helped to ensure rigour in the dataset selected. Sau Yee Yiu and Dr Veneta MacLeod, thank you for taking the time to read my thesis during the final draft. Your editorial comments helped to make this a more succinct thesis. A special mention also goes to Chris Malec. Thank you for

your support as I grappled with learning new terminology, methodology, and analyses techniques used in the biometrics world that completely differs to that from the psychology domain. Also, for your support throughout my PhD by persistently asking “are you broken yet?” I hope to return the favour to you one day soon! To Dr Robin Nicholson, thank you for supporting my research and providing me the opportunity to present at Biometrics 2016 in London (even if it felt like you didn’t have a choice).

The operational validity of this research was made possible by the generosity of agencies and industry. Thank you to each Australian and international agency who took the time to meet with me and provide requirements for this research. Your insights have made this an area that I have become passionate about. The studies would not have been possible on such a large scale without access to an already established database. Thank you to the providers of this database as it has allowed a depth of knowledge not possible before. To each of the practitioners who participated, thank you for taking the time to make this research possible. I hope the research will be used to make your jobs easier in the future. To the Westbourne Group, thank you for providing assistance in the development of an experimental application for the practitioner studies. Also, thank you to the algorithm vendors: 3M, Aware, Cognitec, Morpho, and NEC who provided their facial recognition algorithms for research purposes under Material Transfer Agreements. Your continued support helps agencies and researchers better understand the current state of facial recognition.

To all my family and close friends, thank you for your support and understanding along the journey. Remember all those times you asked what I was doing? Well now you can just read the next 300+ pages to find out! Finally, a special thank you goes to Jaimee Spurr. Without your encouragement, compassion, and food I would not have made it.

Chapter 1.

Research Overview

The child's face is not merely a miniature of the adult (Enlow, 1990, p.17).

1.1 *Introduction*

Determining the identity of children is necessary in a range of national security agencies for the purposes of investigating missing persons, exploited children, child trafficking, and radicalised minors as well as for passport control and visa issuance. Facial comparison is one method that may be used to determine a child's identity. It involves comparing two or more images or an image to the bearer of that image (such as an individual presenting a passport to a Customs Officer), to determine if they are the same person or not (Facial Identification Scientific Working Group [FISWG], 2012).

In recent times, with the advancement of biometric technology, and the improvements of algorithms used within facial recognition systems, there has been an increasing interest in using these systems to conduct facial comparisons with images of children (Interpol, 2015; Wei & Li, 2017). In some applications, these systems are fully automated, such as SmartGate, an automated border control system used by the Department of Immigration and Border Protection (DIBP) at Australian international airports (DIBP, n.d.-a). Others require a 'human-in-the-loop' (i.e., a facial comparison practitioner) to make the final decision from the

output of a facial recognition system, such as during passport application processing (White, Dunn, Schmid, & Kemp, 2015). Although research interest continues to grow in the facial comparison space with regards to algorithm (Grother, Ngan, & Hanaoka, 2017; Yiu, McLindin, Malec, Bourn, & Hanton, (2015) and practitioner performance (Calic, 2012; Heyer, 2013; White, Kemp, Jenkins, Matheson, & Burton, 2014), the focus has predominately been on images of adults.

From the anthropometric literature, it is known that facial ageing during the stages of birth to adulthood and adulthood to old age are different (Albert, Ricanek, & Patterson, 2007; Jayasinghe & Dharmatne, 2009). There is a considerable amount of craniofacial growth and development occurring in childhood (Kozak, Ospina, & Cardenas, 2015; Ricanek, Mahalingam, Albert, & Vorder Bruegge, 2013). Thus, the performance of algorithms and facial comparison practitioners with images of children is likely to be poorer than with images of adults. Anecdotal evidence from practitioners supports this theory (Heyer, 2013; Heyer, MacLeod, Hopley, Semmler, & Ma-Wyatt, 2017). Consequently, the results of studies examining performance with images of adults cannot simply be extrapolated to inform performance with images of children (Wen, Fang, Ding, & Zhang, 2010).

A better understanding of the capabilities of algorithms and facial comparison practitioners with images of children is important for several reasons, including to:

1) determine if there are specific ages and age variations between images (i.e., age differences) that are more difficult for facial comparisons by algorithms and/or practitioners than others;
2) inform practices, policies, and procedures within national security agencies;
3) determine if mitigating strategies need to be implemented into existing work processes to minimise risk;
4) tailor facial comparison training for practitioners with images of children to focus on problem areas;
5) provide empirical support that could help an agency determine the feasibility of purchasing a facial recognition system for their operational application;
6) enable algorithm vendors to focus future development of their algorithms on deficient areas when employed with images of children, and

7) facilitate the removal of children from danger and ultimately save lives.

An overview of the chapters and key requirements examined in this thesis are discussed next.

1.2 *Overview of the Thesis*

Figure 1 provides an overview of the studies conducted and the nine chapters in this thesis.

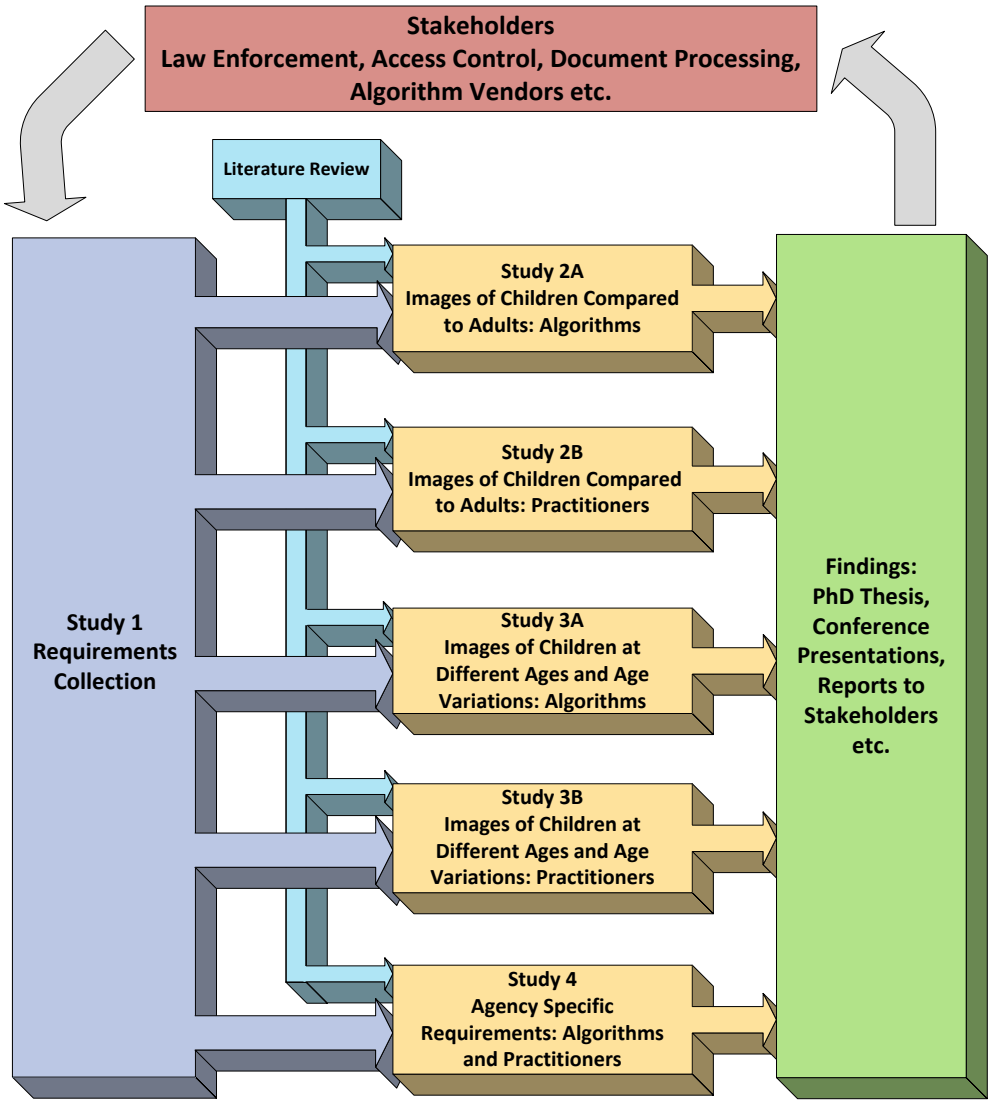


Figure 1. Overview of the thesis.

The current chapter (Chapter 1) introduces several key areas to provide the reader with a general understanding of the importance of this research. This includes an overview of:

- applications that require the determination of a child's identity;
- biometrics and facial recognition systems;
- the difference between familiar face recognition, unfamiliar face recognition, and face matching (as conducted by humans in general);
- the different levels of facial comparison (as conducted by practitioners);
- the definition of age and age variation as defined for this thesis;
- how the face ages throughout the lifespan;
- suggested approaches in the literature to minimise the impact of age-related facial changes on performance; and
- how this research aims to fill the knowledge gaps for operational applications within the national security context.

Chapter 2 describes the exploratory study (Study 1) that was conducted by consulting with national security agencies and algorithm vendors to determine key areas requiring empirical research. This study was used to shape the five empirical studies (as shown in Figure 1) contained in this thesis. Chapter 3 presents information in relation to the large database of operational images provided for this research, along with a high level overview of the methodology and analyses employed in the algorithm and practitioner studies. Chapter 4 examines algorithm performance with images of children compared to adults to determine if anecdotal evidence that came from the exploratory study holds true when tested on a large database of images (Study 2A). Chapter 5 empirically tests practitioner performance on the same variables examined in the previous chapter (Study 2B). Chapter 6 presents an extensive large-scale study conducted to determine algorithm performance with images of children at each chronological age in childhood (0–17 years) and age variations ranging from 0–10 years apart (Study 3A). Chapter 7 is a large-scale practitioner study (Study 3B) designed to determine performance across childhood, similar to that conducted in the previous chapter. Chapter 8 provides a mock operational example to demonstrate how the algorithm and practitioner data from previous studies can be used to inform operational applications (Study 4). Chapter 9 provides a summary of the key findings and concludes with a number of recommendations for agencies and future research.

1.3 *Determining the Identity of Children in National Security Agencies*

Many national security agencies that determine the identity of adults also need to determine the identity of children. In Australia, the *Migration Amendment (Strengthening Biometrics Integrity) Bill 2015* (Cth) expands on existing powers of collection of personal information. This Bill allows personal identifiers (i.e., biometrics) to be collected from children from five years of age without parental consent. The rationale provided for this is to help identify cases of child smuggling/trafficking and radicalised minors. Biometrics that can be collected from children under this Amendment include: facial images, fingerprints, handprints, iris scans, height and weight measurements, audio or video recordings, signatures, or any other identifier prescribed by the regulations that does not involve an intimate forensic procedure. Similarly, the identity of children for visa purposes also needs to be determined. In Australia, biometrics acquired from children for visa applications include face and fingerprints (DIBP, n.d.-b).

In the passport context, any Australian child exiting the country is required to have their own passport (including babies). Passports for children under 16 years are valid for 5 years and require a facial image of that child (Australian Passport Office, n.d.), as shown in Figure 2.

Figure 2. Image of a child holding their current passport.[1]

[1] Copyright © 2016. Image used with signed parental consent.

In investigative applications, a range of different tools and techniques may be used to try and determine one's identity for cases involving missing or exploited children. For example, in missing child cases, investigators may rely on the public notifying them when they recognise a missing child's picture displayed on news programs or noticeboards and match it to a child's face. An investigator is then required to compare images provided by the public with original images of the missing child (Anser Analytic Services, 2011). Biometric technology may also be used. For example, in 2010, there was interest in AmberVision, a system designed to augment AMBER Alerts across the United States. The aim of the system was to aid law enforcement by comparing an image of a missing child to that taken by an officer in the field and provide an automated facial comparison directly on the officer's hand-held device (Biometric Technology Today, 2010). However, no information could be found regarding its current use within law enforcement.

The examples provided here demonstrate that children are not exempt from biometric collection procedures designed for identity purposes. While many forms of biometric identifiers may be used and/or collected, the face seems to be the most common within national security agencies for child identity purposes. Algorithms within facial recognition systems and/or practitioners are required to make the facial comparison.

1.4 *Facial Comparisons – Algorithms*

The implementation of biometric systems throughout national security agencies has increased since 9/11 (Goldman & Maret, 2016; Meek, 2016; Rose & Lacher, 2017). The following section provides an overview of biometric systems followed by more specific information in relation to facial recognition systems.

1.4.1 *Introduction to Biometric Systems*

Biometric systems can be used to recognise an individual based on their physiological or behavioural characteristics in an automated way (Martin, 2013). Several subsystems make up a biometric system. These subsystems and their role are presented in Table 1.

Table 1. *Subsystems in a Biometric System and their Role (Martin, 2013)*

Subsystem	**Role**
Data capture	Collects biometric data from a user.
Signal processing	Converts the biometric data into a template (usually a string of numbers representing the distinctive characteristics of the collected biometric) suitable for matching and storing.
Data storage	Stores the template.
Matching	Receives a new reference template from the signal processing subsystem and compares it to template data stored on the data storage subsystem. The output from the matching subsystem is a comparison score that indicates how closely the two templates match.
Decision	The comparison score is compared to a set threshold. A comparison score greater than the sepcified threshold value will be considered a match, otherwise it will be considered a non-match.

There are three main operations a biometric system can perform: enrolment, verification, and identification. Enrolment occurs when biometric data (e.g., a face image) has been acquired and successfully converted to a template for use in later verification or identification operations. One-to-one comparisons (verification) determine whether captured biometric data matches the stored template for that identity (Martin, 2013). One-to-many comparisons (identification) is when captured biometric data is compared to all of the stored templates and ranked in order of comparison score (Petrov, 2012). One-to-one and one-to-many facial comparisons will be discussed further in Sections 1.4.2.1 and 1.4.2.2 respectively.

Figure 3 presents the subsystems within a biometric system as described in Table 1 as well as the different processes for enrolment, verification, and identification.

Figure 3. Components of a generic biometric system (ISO/IEC 19795-1:2006, 2006).[2]

Common biometric modalities include face, fingerprint, iris, voice, and palm prints (Rose & Lacher, 2017; Vacca, 2007). In national security agencies, face is one of the fastest growing biometric technologies (Jordan, 2016; U.S. Department of Homeland Security, 2015).

1.4.2 *Facial Recognition Systems*

Facial recognition technology is used on a range of databases of facial images in the national security context including: passports, visas, driver's licences, mug shots, and missing persons (Jain, Ross, & Prabhakar, 2004; Mann & Smith, 2017). In Australia, the National Facial Biometric Matching Capability is currently being established to allow government agencies to share and compare facial images from their existing databases when there is a lawful need to do so (Attorney General's Department, 2015).

[2] Copyright © ISO and Standards Australia Limited. Copied by Dana Michalski with the permission of ISO and Standards Australia under Licence.

A number of mathematical techniques have been applied to automate facial recognition from principal components analysis (Sirovich & Kirby, 1987) and eigenfaces (Turk & Pentland, 1991) to the currently popular convolutional neural network approach (Lawrence, Giles, Tsoi, & Back, 1997; Parkhi, Vedaldi, & Zisserman, 2015). It has been claimed, including by some commercial state-of-the-art algorithm vendors, that facial recognition algorithms measure different nodal points on the inner regions of the face. This can include various peaks and valleys of the face such as the depth of the eye sockets and the width of the nose as well as the distance between the eyes and the dimensions of features themselves such as the width of the mouth (De Silva, Roberts, & Dowden, 2014; Joseph, 2017). Given that commercial algorithms are proprietary in nature, it is unknown what methods or combination of methods are actually used by each commercial vendor.

Depending on the application, facial recognition systems may be used to conduct one-to-one or one-to-many comparisons. The empirical studies conducted in this thesis are conducted at a one-to-one level based on findings from Study 1 that suggest this is the most common task within agencies (see Section 2.4.1.1). This thesis evaluates five state-of-the-art algorithms used within these facial recognition systems to determine their error rates under different conditions. As such, when an algorithm incorrectly rejects a mated pair (i.e., same person) of images (templates) at a predefined threshold, this is referred to as a false non-match. When an algorithm incorrectly accepts a non-mated pair (i.e., different people) at a predefined threshold, it is referred to as a false match. Throughout the algorithm studies, 'templates' will be referred to as 'images' for simplicity and to keep consistency between the algorithm and practitioner studies.

1.4.2.1 *One-to-One Facial Recognition*

One-to-one facial recognition is used to determine if a person is who they *claim* to be (Chellappa, Sinha, & Phillips, 2010). The query facial image is compared against an enrolment image whose identity is being claimed. An example of this type of application is traveller self-service through immigration at the airport using an ePassport (Li & Jain, 2011). One-to-one facial recognition is also used to determine if a person is who they are *believed* to be. For example, comparing an image taken from social media to an image of a missing child.

1.4.2.2 *One-to-Many Facial Recognition*

One-to-many facial recognition is used to identify a person from a database (Chellappa et al., 2010). This involves comparing the query facial image to images stored in a database to associate the identity of the query facial image to one of those already collected. This process may be conducted to find the most similar face, or to return a candidate list of the most similar faces with the associated comparison scores above a certain threshold (Li & Jain, 2011). This process may be conducted, for example, to develop a list of potential suspects to investigate (Spaun, 2011).

1.4.2.3 *Performance of Facial Recognition Systems*

Facial recognition systems can be optimised for various applications through algorithm and threshold selection. Systems that are designed to provide convenience, such as entry to Disney World for pass holders using their fingerprint, require less tolerance for incorrectly denying legitimate matches and more tolerance for the acceptance of some incorrect matches, as the aim is to keep customers satisfied (Newton, 2007; Partington, 2013; Vrankulj, 2013). Therefore, false match rates in these scenarios can be set higher. National security agencies require less tolerance of accepting people who are not who they claim to be due to security concerns. Therefore, the false match rate needs to be set as low as possible (Newton, 2007; Partington, 2013). In some national security contexts and algorithm evaluations, a threshold is set so that the false match rate is at 0.001 (Ferguson, 2015; Grother, Quinn, & Phillips, 2011). This means that the operating threshold is set to incorrectly match a pair of non-mated images (i.e., different person) on average 0.1% of the time.

The Face Recognition Vendor Test (FRVT) conducted by the National Institute of Standards and Technology (NIST) is an internationally recognised large-scale evaluation that has been conducted every few years (although under different names) since 2000 (Blackburn, Bone, & Phillips, 2001). It is designed to determine the progress of algorithms including state-of-the-art algorithms used in commercial facial recognition systems. In 2017, the approach changed to an ongoing format whereby vendors can submit their algorithms for evaluation when they choose. The algorithms are then tested as they are submitted and NIST reports the results to the public on an ongoing basis (Grother et al., 2017). Under optimal conditions, these tests show that accuracy levels for algorithms have improved in recent years (Grother & Ngan, 2014; Grother et al., 2011; Phillips et al., 2003; Phillips et al., 2007). However,

in operational settings, the conditions are not likely to be so favourable for example, when the age variation between images is larger than those typically encountered in these tests (which is often around 2–3 years). In addition, the data is grouped by age and/or age variation differently among tests and may not be grouped in ways that are relevant to inform agency objectives.

The best algorithm for each agency is dependent upon a number of factors including the paradigm being used (one-to-one or one-to-many), the type of images being compared (controlled or uncontrolled[3]), and the search speed required (for one-to-many comparisons) (Grother & Ngan, 2014). Variables within the images, such as age, are also known to impact on performance (Jain, Klare, & Park, 2012; Tistarelli, Yadav, Vatsa, & Singh, 2013). Although some research has been conducted to determine the performance of algorithms with images of children, they too have been conducted on relatively limited size datasets (Grother et al., 2017; Grother & Ngan, 2014), or using uncontrolled images where age may not be the only variable impacting on performance (Ferguson, 2015; Ricanek, Bhardwaj, & Sodomsky, 2015; Zeng, Ling, Latecki, Fitzhugh, & Guo, 2012). Therefore, more extensive empirical studies are necessary as it is still not clear to what extent age-related variables impact on the performance of algorithms, particularly with images of children. This is despite algorithms increasingly being adopted by agencies and used with images of children, as well as new and novel applications being considered (Chung, Christoudias, Darrell, Ziniel, & Kalish, 2012; Grother & Ngan, 2015).

The widespread adoption of facial recognition systems by various government agencies has warranted the need to better understand the performance of algorithms with images of children, particularly how performance is impacted by age (Fairhurst, 2013; Jain, Nandakumar, & Ross, 2016; Otto, Han, & Jain, 2012). Implementation of these systems without consideration for their performance is concerning, given they are becoming increasingly automated (Calic, 2012). This is problematic as it is not well understood whether the implementation of technology is enhancing security or potentially putting citizens at

[3] Controlled images refers to images taken in controlled settings of a person front on, depicting a neutral expression and neutral background such as a passport image. Uncontrolled images can vary greatly in terms of lighting, pose, expression, background, and distance from the camera, such as in surveillance images (Spaun, 2009).

additional risk. Furthermore, a facial comparison practitioner is still often required to carefully interpret and verify the results given that they are the final decision maker in the process (Jain et al., 2012; National Policing Improvement Agency, 2009; Spaun, 2011). Despite this, facial comparison practitioner performance with images of children that have been returned by a facial recognition system has not been appropriately evaluated.

1.5 *Facial Comparisons – Practitioners*

For the purposes of this thesis, the process of humans determining the identity of a person using the face are segregated into processes used by the general population (i.e., novices) and those commonly used in agencies (i.e., practitioners). These are discussed next followed by the limitations of past research in this space.

1.5.1 *Face Recognition and Face Matching conducted by the General Population*

In our day-to-day lives, humans conduct familiar face recognition, unfamiliar face recognition, and face matching. Familiar face recognition involves those faces we have previously seen before such as that of a family member or friend (Bruce, 2012). Unfamiliar face recognition involves faces we have only seen once or twice, such as witnessing a stranger commit a crime (Hancock, 2012). Familiar and unfamiliar face recognition rely heavily on memory to recall a face that has been previously seen and compare it to a physically present stimulus (e.g., live individual, image, CCTV footage) in order to make a recognition decision. In contrast, face matching involves at least two physical representations (e.g., live individual, image, CCTV footage) of a face being present at the same time to compare. An example of this is comparing an image of a missing child to an image of a child from a social media website. Therefore, memory nor prior exposure to the stimuli is required in this process (Hillstrom, Sauer, & Hope, 2011). Research examining the overall accuracy of novices conducting face matching under optimal conditions, such as images taken on the same day, typically ranges from 70–80% (Megreya, Sandford, & Burton, 2013; Burton, White, & McNeill, 2010).

Facial image comparison is the type of face matching most often used in operational contexts and is more common than face recognition (Hancock & McIntyre, 2012; Jenkins & Burton, 2008). Despite this, it has received less attention in the literature than face recognition

(Johnston & Bindemann, 2013). For the purposes of this thesis, the abbreviated term 'facial comparison' is used. The role of facial comparison practitioners can vary and is discussed next.

1.5.2 *Facial Comparison Practitioners*

According to the Training Standards Working Group (TSWG) in Australia, there are four categories of facial comparison practitioner (Moss, 2015; 2016), as presented in Table 2.

Table 2. *Categories of Facial Comparison Practitioners and their Role*

Category	**Role**
Facial assessor	One-to-one comparisons that involve comparing an image to a person or an image to another image.
Facial image reviewer	One-to-one comparisons conducted relatively quickly with limited notes required on the decision or one-to-many, either conducted manually or to evaluate output of a facial recognition system.
Facial image examiner	One-to-one comparisons conducted with more rigour. Additional resources including time, tools, and other staff (to provide a second opinion) are available. This role may also involve conducting tasks at the facial image reviewer level to obtain images from a facial recognition system or preparing information for forensic facial image examiners.
Forensic facial image examiner	Performs the same roles as a facial image examiner but has the additional role of conducting comparisons for court purposes.

The research conducted in this thesis examined the performance of people working at the reviewer level as it is the most common among agencies, although performance is also likely

to inform the facial assessor level as well. The results may also be a valuable guide for examiners, for example, to help identify ages that are more difficult than others to compare and therefore require more caution when making decisions.

Research examining the performance of facial comparison practitioners on one-to-one image tasks is rare and almost non-existent with images of children. The one study that could be found that was primarily conducted to examine performance with images of children by practitioners (as well as participants with no experience) was Ferguson (2015). In her study, 76 participants (18 experts, 11 limited experience, 43 no experience, 4 other non-face experience) viewed 20 uncontrolled image pairs of children, a severely restricted dataset. Performance was similar across the experience levels. Overall accuracy for these participants was 65.79%.

Calic (2012) conducted a one-to-one image study with 90 practitioners from five different government agencies. These practitioners were presented with a short video (2–4 seconds) of a person walking towards a camera and the last close-up image from the video remained on the screen. This was compared to a still image also displayed on the screen. A total of 100 one-to-one trials with images of adults were completed by each practitioner. Practitioner accuracy was 93.18%.

White, Kemp, Jenkins, Matheson, et al. (2014) conducted a one-to-one image study with 27 practitioners and 38 novices (students). In this study, images were scanned photos from IDs of students, images taken in relatively controlled conditions, and images taken from video two years prior (from their person-to-image study). Mated accuracy of practitioners and novices combined was 70.9% and 89.4% for non-mated pairs. This study was limited by the severely restricted dataset containing only 17 images of people of the same gender to select non-mated pairs from which may explain the large difference in accuracy between the pair types (i.e., mated and non-mated).

White, Phillips, Hahn, Hill, and O'Toole (2015) also conducted research examining practitioners as well as controls (people with knowledge about facial comparisons but who did not conduct them) and novices (students). Unfortunately, the majority of performance data from this research is not directly comparable to other research as they measured performance

differently using the area under the curve (for an explanation see Doty, 1996). This is not directly comparable to accuracy/error data presented as a percentage that is typically provided to examine practitioner performance. However, it was clear from this research that practitioners were the most accurate, followed by controls, and then novices, although performance between examiners and controls was not statistically significant for two out of the six studies conducted.

Some of the research discussed and other relevant research will be presented in more detail in subsequent chapters prior to introducing each study.

1.5.3 *Limitations of Facial Comparison Research*

Some research evaluating facial comparison performance is often dissimilar to how practitioners conduct their job operationally, leading to questions regarding whether results can be extrapolated to operational settings (Burton, 2013; Dick, 2015; Megreya et al., 2013). Many of the studies in the psychological literature on facial comparisons utilise cropped greyscale images, images taken on the same day, students as participants, and/or deadlined time limits in which to make decisions (Lanitis, 2008; Megreya & Burton, 2007; White, Kemp, Jenkins, Matheson, et al., 2014). These are generally not reflective of operational applications.

Typically in operational applications, coloured images are used that are not cropped directly around the face. Burton (2013) highlighted that using artificial or edited images has become so common in research that acknowledgement of this image type is often omitted from published titles and abstracts. It can be argued that cropped greyscale images that are generally not of any operational relevance have also become so common that they too are not acknowledged as such. While there are instances where using greyscale images is appropriate, this is an important experimental decision, and thus it is reasonable to expect it should be clearly mentioned (Burton, 2013). In addition, results should be caveated to ensure the reader understands performance is based on this image type and may not be representative of performance with coloured operational (i.e., agency collected) images.

Similarly, images taken on the same day or within a few days is common in past research examining facial comparison performance (Megreya & Burton, 2007; Burton et al., 2010; White, Kemp, Jenkins, Matheson, et al., 2014). This is not reflective of what can be expected

in operational applications given that adult driver's licences and passports are valid for 10 years and children's passports are valid for 5 years or if for example, a child has been missing for several years.

Psychological studies that have evaluated facial comparison performance have also typically done so with undergraduate students as participants (Burton et al., 2010; Megreya, Bindemann, & Havard, 2011; Zeng et al., 2012). Research has shown that practitioners at the examiner level outperform students (White, Dunn, et al., 2015; White, Phillips, et al., 2015; Wilkinson & Evans, 2009) but students and reviewers perform similarly and quite poorly at around 48% on a one-to-eight task (White, Dunn, et al., 2015). However, other research examining practitioners at the reviewer level have found practitioners perform well at around 93% on a one-to-one task (Calic, 2012) and 94% on a one-to-ten task (Heyer, 2013). One reason for the discrepancy amongst research may be the deadlined methodology that tends to be present in poorer performing practitioner studies with reviewers. In these circumstances, results may be more a reflection of the task they were presented with in the study rather than their abilities when performing their role as a practitioner. Although reviewers are expected to work quickly and accurately in their profession, they are not deadlined (as determined during Study 1, see Section 2.4.1.1). Therefore, in studies where practitioners are not deadlined, they have more opportunity to use the skills acquired during training and through on the job experience, which often involves a feature-by-feature approach (FISWG, 2012; White, Kemp, Jenkins, Matheson, et al., 2014). This is in contrast to students who are likely to make decisions based on the face as whole, which is a strategy practitioners also adopt when they are not allocated enough time to conduct feature-by-feature comparisons (FISWG, 2012). Deadlining, therefore, is likely to put additional pressure on practitioners to conduct the facial comparison task differently to how they would typically and as a result, may degrade performance.

Results from studies examining student performance or examining practitioners but with methodological restrictions are often generalised by researchers or the media to infer performance of facial comparison practitioners conducting this role in operational applications. This is something that agencies find particularly frustrating (Dick, 2015). These limitations may be acceptable when developing theories, but become problematic when

methods adopted are not ecologically valid, yet are being extrapolated to infer real-world performance (Connor, 2014; Weule, 2014).

More specifically, the limited research that has tried to determine performance based on age-related variables with images of children has been constrained by the size of the databases available (Ferguson, 2015; Zeng et al., 2012). Publicly available databases are relatively small, containing uncontrolled and in some instances, greyscale images (Panis, Lanitis, Tsapatsoulis, & Cootes, 2015; Phillips, Moon, Rauss, & Rizvi, 1997). Yet results from studies using these databases are often generalised to infer that the age variable is the cause of degradations in performance even though a multitude of other variables could be to blame (Akhtar, Rattani, Hadid, & Tistarelli, 2013; Guo, 2013).

The issues discussed raise questions regarding the generalisability of such results to the operational context. Therefore, access to a large controlled operational facial image database was obtained for this research, to ensure the results were operationally valid. A controlled database also ensures that performance is more likely to be based on age-related variables rather than other variables that may degrade performance. In addition, facial comparison practitioners from a government agency participated in this research, rather than students.

The age-related variables as they relate to the research conducted in this thesis are discussed next.

1.6 *Age Defined in the Context of this Thesis*

As the use of facial recognition systems becomes more accepted and implemented into national security agencies for longer periods of time, understanding the impact of age on the accuracy of algorithms within these systems becomes increasingly important. Similarly, the impact that age has on facial comparison practitioner performance is also critical to understand. This is not only in terms of chronological age, but also in terms of age variation. Although closely related, these two age variables need to be considered independently as their impact on performance is often differential. For the purposes of this thesis, they will be defined as:

1) **Age**

   Age refers to the chronological age of a person. This variable is concerned with performance based on the age of a person in an image.

2) **Age Variation**

   Age variation refers to the age difference between the person or people in images being compared. In mated pairs, this is also the elapsed time between images[4]. This variable is concerned with the amount of ageing that may have occurred between images.

Age and age variation will be examined throughout this thesis in terms of their impact on algorithm and practitioner performance. Age variation is a problem that impacts all biometric modalities (Rebera & Mordini, 2013) and many approaches have been suggested to deal with this issue.

1.7 *Approaches to the Issue of Age Variation*

Age variation has been highlighted as one of the main factors that compromises algorithm accuracy (Jain et al., 2012). It is also expected to impact on practitioner performance (Heyer et al., 2017). In regards to algorithms, there are currently four approaches to help resolve the problem of age variation (Rebera & Mordini, 2013).

1) Update templates

   One of the easiest ways to counteract the impact of ageing on biometric systems is to regularly update the templates (Fairhurst, 2013; Ortega, Brodo, Bicego, & Tistarelli, 2009). However, re-enrolment is not always possible, can be time consuming and expensive, and is often seen as an inconvenience to the users (Fairhurst, 2013). This approach is used in national security agencies where possible, for example, to update passport and driver's licence images. However, the extent of ageing that can occur

[4] Although it was considered to use the term 'elapsed time between images,' this term is imprecise when referring to non-mated pairs as the elapsed time between when images were taken, for example, of a person 3 years of age and another person 7 years of age could be the same day rather than 4 years apart as with mated pairs.

over the time an ID document remains valid and the impact on algorithm performance has not been thoroughly investigated.

2) Simulate the effects of ageing

Simulating the ageing effects and re-creating biometric templates accordingly has been suggested as a way to improve algorithm performance (Du, Zhai, & Ye, 2013). This approach requires a detailed understanding of the effects of ageing as well as accurate age progression methods, both of which are currently lacking to a level that would be considered valuable by national security agencies (Rebera & Mordini, 2013). Furthermore, as the ageing process can be influenced by a number of factors, ageing would be difficult to predict and model without establishing person specific ageing patterns that take all factors into consideration (Lanitis, 2009).

3) Use age-invariant biometric features

Age-invariant systems may involve using age-invariant features, such as some parts of the face that are more stable than others for instance, the orbital region (Hughes, Lichter, Oswald, & Whitfield, 2009). A range of different approaches have been proposed including discriminative methods (Li, Gong, & Tao, 2016; Li, Park, & Jain, 2011), data-driven methods (Chen, Chen, & Hsu, 2015), and deep learning models (Khiyari, & Wechsler, 2016; Wen, Li, & Qiao, 2016). Some of these age-invariant approaches show promise but are still being developed and explored. Alternatively, age-invariant systems can involve using an emerging biometric modality such as vein pattern recognition that is more stable across time rather than the face (Kandhasamy, 2017). Unfortunately, access to another more stable biometric may not be an option in many contexts (Best-Rowden, Hoole, & Jain, 2016).

4) Multi-modality

The multi-modality approach involves the combination of several modalities to improve performance which is tailored for different ages. For example, if the faces of children are considered unreliable (as discussed in more detail in Section 1.8), the iris or another more appropriate modality may be used instead (Rebera & Mordini, 2013) or a fusion of these modalities to improve performance. Again, this approach is not

likely to be possible for many national security applications as they may be restricted to only information from the face.

As the majority of these approaches are not yet feasible and require further research before being implemented operationally, the best way to currently deal with age-related issues is to understand the nature of the changes (Fairhurst, 2013). This will help to identify whether implementing mitigating strategies is necessary to ensure that national security agencies can function as effectively as possible.

The main approach used by practitioners to address age variation is to compare facial features that are known to remain stable over time such as the ears. FISWG is currently drafting a document to provide information on the relative physical stability of facial features of adults for facial comparison purposes (FISWG, 2016). The stability of facial features are rated as low, medium, or high stability as a function of expression, ageing (short and long term), significant weight change, health changes, and intentional alteration. The next section discusses how the face changes throughout the lifespan.

1.8 *Age-Related Facial Changes throughout the Lifespan*

The rate of age-related change in the human face is not continuous from birth to death (Partington, 2013). Therefore, age-related changes are likely to impact on performance differently at different stages of life. Age-related changes can impact both the shape and texture of the face (We & Li, 2017). In childhood, facial changes are extensive, particularly in the younger years (Feik & Glover, 1998) and the appearance of babies and children's faces differ quite considerably from those of adult's (Kozak et al., 2015) as shown in Figure 4. The features of a baby's face are underdeveloped and diminutive. As such, they will resemble each other more than they resemble adult faces (Wilkinson, 2012). The growth of the face during childhood is not simply an enlargement of existing structures, but rather, a series of changes occurring to different structures at different rates, at different times, and in different directions (Enlow, 1990).

Figure 4. Age-related facial changes in siblings from birth to adulthood. Age of female sibling in top row: birth, 10 months, 2.5 years, 6 years, 10 years, 14 years, 18 years, and 25 years. Age of male sibling in bottom row: birth, 7 months, 2 years, 5 years, 9 years, 15 years, 19 years, and 25 years.[5]

Bone growth is the largest form of change that occurs to the face in childhood as shown in Figure 5 and this will impact the shape of the face. Over time, a child's face grows down and forward (Taylor, 2001). The bones below the eyes elongate until this portion of the face grows from half the vertical length to two-thirds of the face (Gibson, 2010).

Figure 5. Bone growth across childhood. Skull at age: 4 months (left), 1.5 years (middle), and 13 years (right).

[5] Copyright © 2015 by Frederick K. Kozak, Juan Camilo Ospina, & Marcela Fandiño Cardenas. Image used with permission.

Some craniofacial developmental changes that occur in childhood and into young adulthood are similar for all individuals, regardless of gender or population (Ricanek et al., 2013). Females and males generally experience similar growth patterns in height, width, depth, and volume of facial features until around 11 years of age. A growth spurt occurs in females at around 11 years of age with minimal growth after this time. Growth is almost complete relative to the adult head by around 14 years of age. Males can have a growth spurt expanding from ages 11 to 16 years (Ferrario, Sforza, Poggio, & Schmitz, 1998), although most of the characteristics of a face are defined by around 11 years of age (Hunter, Tidderman, & Perret, 2012).

Facial anthropometric studies have shown that some facial features attain growth saturation earlier than others (Bishara, Peterson, & Bishara, 1984; Farkas & Hreczko, 1994; Sforza, Gandi, Binelli et al., 2009; Sforza, Grandi, Catti et al., 2009; Sforza, Grandi, Binelli et al., 2010; Sforza, Grandi, De Menezes, Tartaglia, & Ferrario, 2010). During facial growth, there is an increase in the size of the nose, mandible, and maxilla. The outcome is the larger and coarser features of the adult face compared to those of the child's (Kozak et al., 2015).

In adulthood, the largest change is due to texture modifications. Facial skin eventually becomes slack, darker, less flexible, and rougher. As a result of the reduced skin elasticity, lines, wrinkles, folds, and blemishes gradually appear (Lanitis, Tsapatsoulis, & Maronidis, 2013). Taylor (2001) and Albert et al. (2007) offer a general decade-by-decade account of how the face changes in adulthood. The appearance of the face is subjected to many factors that can cause changes in the shape and/or texture of the face as it ages. These factors can include: sun exposure, diet, health conditions, drug use, cosmetic surgery, and lifestyle (Albert et al., 2007; Lanitis, 2009; Lanitis et al., 2013).

Age-related facial changes are highly non-linear with different facial features growing and changing at different rates and at different times, particularly in the younger (0–17) and older (>65) years. Despite this, the age distribution of the people in existing facial databases for research purposes is typically 18–60 (Lanitis et al., 2013). Publicly available facial image databases that do contain images of children are restricted in terms of the: number of children, ages of the children, short age variations, and/or images of an uncontrolled nature (Panis et al., 2015; Ricanek et al., 2015; Ricanek & Tesafaye, 2006; Somanath, MV, & Kambhamettu, 2011). Therefore, there is a lack of knowledge as to how age-related changes impact on

algorithm and practitioner performance, particularly with images of children. This lack of knowledge will be addressed in this thesis.

1.9 *Thesis Aims*

This thesis aims to evaluate facial comparison performance with images of children by algorithms and practitioners for different ages and age variations. This will be evaluated through conducting algorithm studies with five state-of-the-art commercial facial recognition algorithms and conducting human studies with facial comparison practitioners. These studies incorporate a database of controlled operational (i.e., agency collected) images containing several million images of children and adults. This will allow age and age variation variables to be examined in a more operationally valid way.

Understanding the impact that age has on algorithm and practitioner performance can lead to more appropriate tools, methods, and strategies being adopted to help minimise any negative impact; enabling national security agencies to exploit modern techniques and technologies with increased confidence. It also enables algorithm vendors to make improvements on their algorithms that they may not have identified before due to lack of access to appropriate images for testing purposes.

The following chapter describes Study 1, which was designed to collect requirements from agencies and insights from algorithm vendors to help shape the empirical studies presented later in this thesis.

Chapter 2.

Study 1: Requirements Collection from System Administrators, Facial Comparison Practitioners, and Algorithm Vendors

2.1 *Introduction*

Facial recognition systems have been integrated into many government agencies throughout the world in order to support a range of applications including those for investigative, identity document processing, and access control purposes. In Australia, facial recognition systems are used by the Department of Immigration and Border Protection, Australian Federal Police, Department of Foreign Affairs and Trade, Department of Defence, several state policing agencies, and state motor vehicle licencing authorities (Mann & Smith, 2017; Prince, 2013). The Australian Criminal Intelligence Commission (ACIC) is scheduled to have a multi-modal biometric capability by the end of 2017, which includes facial recognition, for the purposes of identification, crime investigation, and border security (ACIC, 2017; CrimTrac, 2016).

System administrators of facial recognition systems are located at each agency. Their role is to make decisions regarding business workflow, including how facial comparisons will be conducted. Decisions are based on their business objectives and information obtained from a variety of sources including other agencies, research reports, performance evaluations (both human and system), in-house testing, and facial comparison practitioners. Even with these

sources available, information deemed useful for agency requirements is scarce. This scarcity is particularly true for information regarding the performance of commercially available state-of-the-art facial recognition algorithms with images of children. This was made evident by the current ISO/IEC Report – Information Technology – Cross Jurisdictional and Societal aspects of Implementation of Biometric Technologies – Biometrics and Children (ISO/IEC 30110, 2015) suggesting facial recognition systems should not be used on children younger than five years of age. This suggestion was based on information originally obtained from an algorithm vendor's FAQs webpage that is over 13 years old and no longer available. Clearly, current and more reliable algorithm performance data with images of children is required.

Past research has identified that agencies are frustrated with the lack of information available to guide decisions regarding the use of facial recognition systems (Heyer, 2013). Although these facial recognition systems are increasingly relied upon, facial comparison practitioners are still generally required to make a final decision and in instances where a facial recognition system is not available or feasible, only a facial comparison practitioner will be involved in the facial comparison task (Calic, 2012; Prince, 2013). Again, little research exists to understand practitioner performance with images of children. Heyer et al. (2017) conducted a survey of facial comparison practitioners from Australian government agencies and found that identifying children was considered challenging, particularly when comparing a baby photo to a toddler. Further research is therefore necessary to provide agencies with empirical data to make more informed decisions.

The aim of this study was to explore how facial comparisons are being conducted in agencies and what specific areas are of concern in regards to facial comparisons with images of children for various stakeholders (system administrators, facial comparison practitioners, and facial recognition algorithm vendors). Requirements were collected in order to shape the research conducted in this thesis, aimed at examining the impact of age-related variables on facial comparisons with images of children conducted by algorithms and practitioners. This was to ensure the research was as operationally relevant as possible. To achieve this, information was gathered from agencies in regards to the paradigm commonly used within that agency (i.e., one-to-one verification or one-to-many identification), type of images being compared (controlled or uncontrolled), and time allocated per comparison (deadlined or self-paced). Consulting agencies at this early stage helped to ensure that the empirical studies were

beneficial to help guide agencies in their practices, policies, and procedures relating to facial comparisons with images of children. The results also helped to determine what aspects of facial comparisons with images of children were most difficult and therefore identify where further research and training may be warranted.

Algorithm vendors were also consulted to provide insight into the performance of their algorithms in regards to the agencies requirements. Including algorithm vendors at this early stage also determined that their involvement in the empirical studies could be of benefit by providing them with the performance results of their algorithms on a large database containing several million images. The intention was to provide results that could help to improve the algorithms that agencies have access to in the future. With these factors in mind, the questions this study aimed to answer were:

1) How are algorithms and facial comparison practitioners currently being used to make facial comparisons in various operational settings?

2) How often are facial comparison practitioners required to make facial comparisons with images of children, at what ages in childhood, and at what age variations between images?

3) At what ages do system administrators and facial comparison practitioners believe that performance degradations occur in childhood?

4) What are the main age-related aspects of facial comparisons with images of children as conducted by algorithms and facial comparison practitioners that require further research?

2.2 *Method*

This section provides the sampling, design, and data collection methods used for this study.

2.2.1 *Sampling*

A stratified purposeful sampling method was used for this study. Stratified purposeful sampling involves drawing samples from various subgroups (Babbie, 2010). In this study, the subgroups were different types of stakeholders selected for their knowledge and involvement

in facial comparisons. These included: system administrators, facial comparison practitioners, and algorithm vendors. This approach yields information rich cases that are selected to gain insights and in-depth knowledge about an area under investigation (Patton, 2002). This method ensures a more credible study as it incorporates various subgroups and therefore can capture major variations in opinions even though a common core may emerge in the analysis (Patton, 2002; Suri, 2011).

System administrators and facial comparison practitioners were required to be from Australian or international agencies that conduct facial comparisons for investigative, processing, or access control purposes. This was to ensure that the requirements collected were gathered from national security applications where the common goal is geared towards the safety and protection of people, rather than commercial applications where the goal may be to sell more products. Algorithm vendors were asked to complete a survey based on information obtained from system administrators and facial comparison practitioners to better understand the user demand and performance of their algorithms with images of children. The various stakeholders were selected to ensure a holistic understanding of the current areas of concern from different perspectives, to ensure all areas were acknowledged, and to help shape further studies in this thesis (Creswell, 2014; Patton, 2002; Walker, 2004).

Participants were recruited via emails sent to current contacts at each agency and algorithm vendor, and through word of mouth by other government agencies on behalf of this research. Participants were eight system administrators (together with four people from higher management), 17 facial comparison practitioners (facial reviewers, facial examiners, forensic facial examiners), and eight algorithm developers from:

- four Australian government agencies (1x investigative, 2x processing, 1x access control);
- one international government agency (1x investigative);
- one international non-profit agency (1x investigative); and
- five state-of-the-art facial recognition algorithm vendors.

Ethics approval for the research conducted in this thesis was granted by the Ethics Review Panel at the Defence Science and Technology Group (DST Group, reference number

NSID 03-13) and the University of Adelaide's School of Psychology Human Research Ethics Subcommittee (reference number 13/97).

2.2.2 *Design*

An exploratory qualitative design was chosen for this research. An exploratory qualitative design is useful at the preliminary stage of investigation, when the focus is on gaining insight and familiarity of the area, and to identify and refine research problems for further investigation (Kothari, 2004; Krishnaswamy, Sivakumar, & Mathirajan, 2006; McNabb, 2010). A range of qualitative data collection methods were used, which is common in qualitative research (Maxwell, 2013). The rationale for using qualitative methods in this initial study was that the topic was ill-defined and not well theorised, thereby precluding the use of quantitative methods (Walker, 2004).

2.2.3 *Data Collection*

Three methods of data collection were chosen to gather information from the three types of stakeholders. These methods were: focus groups (for system administrators), observations (for facial comparison practitioners), and surveys (for algorithm vendors). Each of these methods is discussed in more detail in the following sections.

2.2.3.1 *Focus Groups (System Administrators)*

Focus groups are a highly accepted method for collecting requirements (Martin, Murphy, Crowe, & Norris, 2006). Focus group interviews involve unstructured and generally open-ended questions intended to elicit views and opinions from the participants (Creswell, 2014). This method helps the participants be involved at an early stage, explore what they see as important factors and to identify contextual problems (Martin et al., 2006). Focus groups are a useful method when participants cannot be directly observed (Creswell, 2014). This is the case with system administrators who typically do not conduct facial comparisons but still provide a level of knowledge and authority as to how facial comparison processes will be implemented within their agency.

Prior to the focus group sessions, an email was sent out to each agency containing a brief agenda along with planned duration (about one to one and a half hours per session). The agenda explained that facial comparisons with images of children was the main aspect under

investigation. This was to make clear that other potential areas of concern that have been identified and studied in previous research were not under investigation (Hassaballah & Aly, 2015; Heyer et al., 2017; O'Toole et al., 2007; Sinha, Balas, Ostrovsky, & Russell, 2006). This email also ensured that the participants had the opportunity to think about the topic prior to the focus group session, to have any relevant material or figures available on hand, and to invite other subject matter experts to be present (as suggested by Robertson & Robertson, 2013). One focus group was conducted at each of the four Australian government agencies. A focus group session was not conducted at the international government agency as the section of the agency participating in this research is currently without facial recognition algorithm capability. A focus group was conducted at the international non-profit agency that also does not have algorithm capability but wanted to participate to discuss their interest in acquiring this technology and learn more about the current research. Due to the sensitive nature of the information provided, similarities and differences between the agencies are discussed in general terms only and all agencies will remain anonymous. For confidentiality reasons, these agencies are referred to as Agency A (processing agency), Agency B (processing agency), Agency C (access control agency), Agency D (investigative agency), Agency E (investigative agency), and Agency F (investigative agency).

Focus groups were conducted in a conference room at each participating agency. These sessions were attended by system administrators, but also management and other experts (e.g., forensic facial examiners) who had been invited by system administrators based on their knowledge and experience in facial comparisons. Table 3 provides a list of the attendees at each agency.

Table 3. *Attendees at each Focus Group by Agency*

Agency	System Administrators	Management	Other Experts
A	2	1	2
B	2	1	1
C	1	0	4
D	1	1	3
F	2	1	4

Participants were provided with an Information Sheet to read (see Appendix A) and a Consent Form to sign (see Appendix B). Participants were informed that they would remain anonymous. Questions were open-ended and were aimed to elicit any problems that system administrators could identify in regards to facial comparisons with images of children. For example: *"What ages are believed to be the most difficult to conduct facial comparisons on?"* (See Appendix C for a list of questions).

Notes were taken throughout these sessions. A recording device was not used as it was considered inappropriate for topics of a sensitive nature (Bachiochi & Weiner, 2002). During the sessions, participants equally contributed to the discussions, their comments were reiterated, and they were asked to clarify any points necessary. At the end of the session, participants were thanked for their time and reminded how the results were to be used.

Focus groups are often restricted to the immediate area of the people being interviewed. Therefore, it is best to avoid them as the sole method to collect requirements and instead, use them in conjunction with other techniques (Robertson & Robertson, 2013). This is particularly true for focus groups conducted in this study with system administrators who may not perform facial comparison tasks, but play an integral role in the agency in regards to implementing policies regarding facial comparison tasks. Martin et al. (2006) suggest focus groups be complimented by observations of relevant situations, such as practitioners performing facial comparison tasks.

2.2.3.2 *Observations (Facial Comparison Practitioners)*

Observations enable a firsthand view of the activities of the participants under study and involve collecting data based on these observations (Schwandt, 2007). This provides an opportunity to evaluate actions in practice, rather than relying on people's opinions and self-interpretations of their behaviours and attitudes (Gray, 2009). Patton (2002) identifies several advantages of directly observing a setting. These include allowing the researcher: a better understanding and ability to capture the context of the issue under study, the opportunity to see things such as important nuances that may be missed by the people in that setting due to the routine of the participants' work, and a more comprehensive view of the area under investigation rather than relying on second-hand interviews.

For this study, business processes used by practitioners in various agencies to conduct facial comparison tasks, either with the support of facial recognition algorithms or completely manually, were observed. Notes taken of these observations were combined with information collected from informal interviews and participants' descriptions during observation sessions (as suggested by Patton, 2002).

The role of the observations was to allow more insight and familiarisation into various operational contexts that require facial comparisons which would help to inform the methodology for the remaining studies in this thesis. This time was also used to ask facial comparison practitioners similar questions asked of system administrators. Observation sessions were conducted at five agencies, three in Australia and two internationally, with one practitioner from each agency. Sessions were scheduled immediately after the focus group session at each agency where appropriate, including at two agencies where demonstrations were provided by forensic facial examiners who were present in the focus group session.

Prior to the pre-planned observation sessions, facial comparison practitioners were sent an email to inform them of the purpose of the study and provided with an Information Sheet to read (see Appendix A) and a Consent Form to sign (see Appendix B). Practitioners were also informed that notes would be taken to record observations and any additional information provided by them. They were informed that notes would not identify them or the agency for which they worked. Again, no recording device was used due to the sensitivity of the discussions. As notes were taken at the time of investigation, this allowed immediate

interaction with data, thereby exposing gaps in knowledge and allowing for further investigation in particular areas to be made. As recommended by Gray (2009), questions were asked throughout this process for further clarification where required. Observation sessions took between one and two hours each, depending on the processes at each agency and the additional questions it generated. Sessions concluded once a solid understanding of the processes was attained.

2.2.3.3 *Open-Ended Surveys (Algorithm Vendors)*

Open-ended surveys are a useful way to collect data from companies (Bachiochi, & Weiner, 2002). Surveys were deemed the most appropriate method to gather information from algorithm vendors as they are located in various parts of the world, making face-to-face methods impractical. Furthermore, using open-ended surveys is advantageous as it provides a greater variety of information to be captured, allows unanticipated answers to be obtained, and reduces investigator bias (Kumar, 2011).

The survey was emailed to the main point of contact at each of the six algorithm vendors who had Material Transfer Agreements with DST Group (one of these vendors chose not to participate). The survey was then forwarded on to the most relevant person(s) in the company that could answer the questions. The aim was to not necessarily collect a substantial amount of data from various people within the company, but to gather rich data by the most appropriate person(s) to answer questions about specific areas under investigation.

Information about the research was provided on the front of the survey to ensure the person answering the questions knew the reason behind why the survey had been developed. The front page also contained information to reassure participants that the answers they provided in the survey would not personally identify themselves or the company they were representing. Informed consent was assumed if they filled in the survey and returned it. The survey contained 20 questions (some with two parts as shown in the subsequent example), divided into five different sections based on findings resulting from the focus groups and observations. Questions were open-ended, however closed questions were used to determine if participants were in a position to comment about a particular area of questioning. For example:

"Have customers mentioned their need or potential requirement to use automated facial recognition on children?"

"If yes, in what contexts? (i.e., for what applications)."

As suggested by Bachiochi and Weiner (2002), a final catch-all question was provided at the end of the survey to ensure that the survey respondents had the opportunity to share any opinions they may not have had the appropriate place to share elsewhere:

"Please share any additional comments about any of the questions above or other thoughts."

A copy of the survey questions is provided in Appendix D. Five algorithm vendors completed one survey each and returned it via email. A sixth algorithm vendor decided not to complete the survey but did provide their facial recognition algorithm for the empirical studies conducted as part of this thesis.

2.3 *Analysis*

Exploratory data analysis is a set of steps used when open-ended data has been collected to explain a new area under investigation. This analysis begins as soon as data collection commences and directs subsequent sessions and data collection during the same research program (Stebbins, 2008). For instance, written notes from focus groups were taken at each session and this knowledge was then used to direct subsequent sessions during the focus groups with system administrators, the observation sessions with facial comparison practitioners, and the survey questions for the algorithm vendors.

As the data was recorded, ideas such as possible concepts and generalisations that emerged were written as notes. Each record and accompanying notes were compared soon after each session of data collection for similarities and differences (constant comparative analysis). Once all of the data was gathered, the second phase of data analysis began. This involved a closer examination of the notes taken to identify common patterns relevant to the operational context and then merging these into requirements that would become the focus of the empirical studies in this thesis. Similar to Heyer (2013) who gathered qualitative data from

government agencies to identify key variables to focus empirical research, this was conducted at the higher level of meaning (uncovering key requirements), rather than on any underlying or latent meaning.

The requirements from each focus group were placed into matrices in Microsoft Excel™ so that focus group sessions and each concept could be clearly viewed. This ensured that specific features of the data were focused on, informed by the comments made by the system administrators as well as the research agenda. Narrative summaries were constructed to preserve the context and the story. Combining narrative summaries with coding and matrices allowed a deeper understanding of the focus groups that neither could have provided alone (Maxwell, 2013). A similar process was also possible for the observational data and the surveys. Extensive overlap was identified between system administrators and facial comparison practitioners so these were combined into the same matrices.

2.4 *Findings*

The data collected was analysed to gain an understanding of how facial comparisons are conducted in operational settings and to identify requirements from stakeholders focusing on facial comparisons with images of children. These requirements were developed by the categories emerging from the data, based on the information provided in the focus groups, observation sessions, and to some extent, the surveys.

Facial comparison practitioners were from different levels of the facial comparison realm (facial reviewers, facial examiners, and forensic facial examiners). A general overview of the business processes and four key requirements that emerged relating to the impact of age-related variables on facial comparisons with images of children that require further investigation are discussed. In addition, two areas that could potentially improve facial comparison performance with images of children were considered. These are briefly mentioned as the information gathered about these solutions has been incorporated into other research currently underway and will be discussed in the final chapter as part of other potential future research directions.

The survey results from the algorithm vendors provide insight into the level of knowledge and testing that has been conducted on five state-of-the-art facial recognition algorithms. They

are also used to identify whether empirical evidence using their algorithms on the requirements collected from agencies provides any additional benefit to vendors, in terms of being able to improve their products and therefore improve the performance of algorithms in future versions available to agencies.

2.4.1 *Focus Groups and Observation Sessions conducted at Agencies*

As the information gathered from focus groups and observation sessions complemented each other, the key findings from system administrators and facial comparison practitioners at each agency were able to be collated and are presented next at the agency level. This information helped to guide the methodology for the empirical studies.

2.4.1.1 *General Overview of Business Processes, Resources, and Images used in Agencies*

The findings showed that facial comparisons conducted by algorithms and/or practitioners varied in terms of the type of comparisons being conducted, resources available, the quality of images involved, and the amount of time available for each comparison. Although some agencies conduct one-to-many comparisons, the dominant paradigm used by participating agencies is one-to-one comparisons, particularly when a substantial amount of care is required in the decision. Agencies varied on what resources they had available to make facial comparison decisions; algorithms, practitioners, or both. Table 4 shows the resources used by each agency to make facial comparisons.

Table 4. *Resources used by each Agency to make One-to-One Facial Comparisons*

Agency	Practitioner	Algorithm
A	✔	✔
B	✔	✔
C		✔
D	✔	✔
E	✔	
F	✔	

Although two agencies (Agency E and F) at the time this study was conducted did not have algorithm technology, they wanted to be informed about algorithm performance in order to determine whether implementation of such technology was pertinent to their business objectives. Agency E did not have a facial recognition system but had accessed one to test its efficacy on a limited number of operational cases. Agency F however, had been supplied with several algorithms and conducted their own small in-house tests with them but were yet to implement any into their workflow. Information obtained from Agency C was specific to one application that was considered fully automated (even though a facial comparison practitioner was required to resolve any issues). As such, the information gathered from Agency C was based solely on this automated application.

One system administrator mentioned using a different algorithm threshold setting for images of children and adults. Another from a different agency, mentioned that they do not use a different threshold setting with images of children and that the threshold was always set with images of adults. The reason provided was that the youngest age of children being compared by this agency was young teenagers and therefore they did not believe this warranted a separate operational threshold.

The quality of the images used when making facial comparisons varied depending on the purpose of the task in each agency. These ranged from uncontrolled images (e.g., an image from CCTV) up to standardised controlled images (e.g., a passport photo). The types of images discussed included images from CCTV, scanned Polaroids, magazines, web cameras, social media sites, smart phones, mug shots, visas, drivers' licences, and passports. Overall, there was no one image quality standard or type of image that was used more than others.

Facial comparison practitioners varied with regards to how long they had to make a facial comparison decision. This depended on the skill level of the practitioner (facial reviewer, facial examiner, or forensic facial examiner) in conducting facial comparisons. Facial reviewers in Agency A and B were expected to make facial comparisons quickly and accurately due to the large workloads experienced but were not deadlined per comparison. If reviewers in these agencies had difficulties making a decision, they had the ability to defer the images to facial examiners who could take longer to compare the images and had more tools available. Other agencies did not feel time restricted with Agency E claiming to take up to several hours to

conduct a facial comparison if necessary. This agency however worked up to the forensic facial examiner level. Agency C used a fully automated process rather than facial comparison practitioners so the facial comparison component was almost instantaneous.

The following sections provide the main requirements identified by agencies as critical areas where empirical research is needed.

2.4.1.2 *Requirement 1: Determining Facial Comparison Performance with Images of Children and Adults*

Although the main focus of this research was on facial comparisons with images of children, the findings showed that facial comparisons with images of adults were also conducted by five participating agencies. Facial comparison practitioners in these five agencies were expected to conduct facial comparisons regardless of whether the images contained children or adults. These five agencies were aware that facial comparisons with images of children were more difficult but needed to know the extent of this degradation so that more caution could be exercised and, if necessary, determine whether further training with images of children was warranted.

The caseload containing children's facial images varied depending upon the objectives of the agency and ranged from approximately 20% to 100%. Agencies discussed both the current ages examined when conducting facial comparisons (i.e., at the time this study was conducted) as well as the ages that were required for business objectives (see Table 5).

Table 5. *Ages Currently Examined by each Agency and the Ages Required to be Examined for Business Objectives*

Agency	Ages Currently Examined	Ages for Business Objectives
A	≥ 6 years	All
B	≥ 12 years	All
C	≥ 10 years	All
D	≥ 12 years	All
E	All	All
F	0–17 years	0–17 years

Four participating agencies did not make facial comparisons on the entire age range that their business objectives required. All agencies wanted to conduct facial comparisons on young children but anecdotal evidence from facial comparison practitioners suggested that their performance, as well as algorithm performance, was too poor. One practitioner stated that *"all babies look alike"* and that practitioners can only *"look at skin colour"* of babies to try and make a determination. Agency A did not make facial comparisons with images of children less than six years of age because the document renewal process would return images of the child, potentially as a baby. Thus, they did not attempt to make a comparison. Instead, this agency (and others) relied on a range of other information available to the practitioners to guide their decisions as to whether the images belonged to the same person or not.

Although Agency E and Agency F compared facial images of the youngest ages of children, these agencies only used highly skilled forensic facial examiners who were provided with the time, tools, training, and experience necessary to make informative decisions. Often facial comparisons could be challenging in these agencies due to the nature of the images and thus other sources rather than the face were sometimes relied upon to make a decision pertaining to a child. However, one participant from an agency without algorithm technology stated:

"I would spend more time on face if I had the capability [technology]."

All agencies mentioned that they needed to compare facial images taken up to 10 years apart, with one agency suggesting possibly even 20 years. Examples provided of this included drivers' licences and passports that were a valid form of identification for adults for up to 10 years and passports for children that were a valid form of identification for up to 5 years, but which may not be renewed consecutively. An example from an investigative application was of missing children cases where a child may be missing for several years and the public sends in images of children they believe to be the missing child. These images may initially be compared to an image of the missing child that the family supplied to an investigative agency to determine whether further investigation is warranted.

2.4.1.3 *Requirement 2: Determining Facial Comparison Performance with Images of Children at Different Ages and Age Variations*

As has already been discussed, agencies are not conducting facial comparisons on all ages that business objectives require. As such, the most dominant request to come from agencies was the need to know how age impacts performance of algorithms and facial comparison practitioners across childhood.

A facial comparison practitioner from Agency B believed that algorithms were not very useful with images of children by stating that: *"algorithms generally don't match with kids."* This practitioner was interested to see if their beliefs were accurate or if different algorithms besides the one being used in the agency would perform better with images of children.

There was also an interest in this requirement to help determine if there were certain ages or groups of ages that would only be worth providing to an algorithm (due to practitioners performing so poorly) in order to help reduce the workload on facial comparison practitioners and vice versa. Agencies were also interested to know the lowest age possible in childhood that was feasible to present to an algorithm or facial comparison practitioner to conduct facial comparisons on, although what was considered feasible varied for different applications. Again, agencies were interested in this for age variations of up to 10 years to see how performance changes.

Agencies D and E specifically mentioned the need to know how performance changed over the child/adult border where a person may be a child in one image and an adult in another.

An example provided of a situation where this was necessary was when a child had been abused in child exploitation material but did not come forward until he/she was an adult. Another example was of a driver's licence obtained in adolescence that was a valid form of identification for 10 years. There was a need to know if this identity document could be used to identify a person 10 years later or if too much facial growth had occurred in the teen years, rendering the identity document ineffective.

2.4.1.4 *Requirement 3: Determining Facial Comparison Performance for Agency Specific Requirements*

General requirements that were consistent with all agencies have been discussed as Requirements 1 (see Section 2.4.1.2) and 2 (see Section 2.4.1.3), but unique requirements that were agency specific also emerged. These requirements were considered sensitive so will not be discussed in detail here however, a mock example will be provided in Study 4 (Chapter 8) to demonstrate how the data collected in this thesis could be tailored for specific agencies to answer their own unique requirements.

Agency specific questions focused on a need to know how performance would change if facial comparison tasks were conducted on all ages that their business objectives require (i.e., all ages in childhood) and comparing this to the ages that facial comparisons are currently conducted on within their agency. This was to better understand how current and potential future procedures could impact performance. There was also an interest to group data into specific age groups in childhood to ascertain if specific procedural decisions are or would be effective for certain ages.

For agencies currently without algorithms, there was interest in whether algorithms were feasible to use with different age groups of children. An example provided was from an investigative application where children were loosely grouped into ages that were categorised as baby, toddler, prepubescent, pubescent, and adolescent. These age groups could vary between agencies and be based on the specific requirements within the agency.

2.4.1.5 *Requirement 4: Determining Facial Comparison Performance with Mated and Non-Mated Image Pairs*

It was found that depending on the type of agency (investigative, processing, or access control), the amount of mated compared to non-mated facial comparisons that were required to be conducted varied. In processing and access control applications, the number of non-mated comparisons practitioners were presented with was expected to be a lot smaller than in investigative applications, although exact numbers could not be provided as the nature of the work made it almost impossible for the agency to know. For processing applications, facial comparisons were required to detect non-mated images that could be the result of someone trying to obtain a fraudulent document or from human error somewhere in the process. In access control, an example of a situation when an automated system may be presented with a non-mated image was when someone was deliberately trying to gain access when they should not (e.g., by using someone else's identity document or a fraudulent document to try and fool the system). Another example was due to user error where family members had accidentally swapped identity documents and tried to gain access through the automated system with the wrong document. In investigative applications, the number of mated to non-mated images varied and depended on individual cases. In all applications, there were consequences for making the wrong decision. These could range from a simple inconvenience to the client, to a radicalised minor entering or fleeing the country, or a child being continually abused. As such, agencies wanted to know if performance varied based on whether facial comparisons involved mated or non-mated images and if so, how it varied.

As this requirement related to the three previous requirements mentioned, it was not considered a separate requirement for this thesis but rather, was incorporated into each of the previous three requirements. Therefore, rather than just providing data at an overall level of performance, results were also divided into mated and non-mated performance.

Although the aim of this thesis was to evaluate the impact of age-related variables on algorithm and practitioner performance with images of children, two potential solutions emerged to help improve performance; age estimation and age progression. These are briefly discussed as Requirements 5 (see Section 2.4.1.6) and 6 (see Section 2.4.1.7) respectively as they have been used to shape other current research and should also be considered as part of potential future research.

2.4.1.6 *Requirement 5: Determining Age Estimation Performance with Images of Children and Adults*

Age estimation algorithms are different to facial recognition algorithms, although many vendors specialise in both. The ability of practitioners to estimate the age of a person in an image was an issue of importance for three agencies. This was often required in situations when it was necessary to determine whether the person in the image was a child or an adult. Understanding how well algorithms and practitioners can estimate age was also deemed important for scenarios where someone under 18 years needed to be categorised, for example, as a toddler, pre-pubescent, or adolescent. The performance of age estimation algorithms was of interest to determine if they were accurate enough to be used as a filtering tool to help improve algorithm and practitioner performance. If age estimation algorithms perform accurately, they could reduce the number of images an algorithm or practitioner may need to examine, therefore potentially increasing the speed of examination, and reducing the number of errors made.

2.4.1.7 *Requirement 6: Determining the Performance of Age Progressed Images for Facial Comparison*

Three agencies were interested in whether algorithm and practitioner performance would improve if artificially aged images produced by automated software were used to reduce the age variation between images, rather than using an image that is several years old. This interest of using age progressed images predominantly focused around the identification of children due to how much their faces change in the early years. An example provided of this application was for missing person cases. Agencies were also interested in whether there would be a difference in performance using age progressed images designed by forensic artists compared to automated age progression software. One participant from Agency F stated a potential benefit if the performance of both automated software and forensic artists was known:

> *"Possibly in the future we could use automated systems then have a human [forensic artist] tweak the final picture with family photos."*

2.4.2 *Surveys Completed by Algorithm Vendors*

The aim of the survey was to collect information from the algorithm vendors focused on the requirements collected from agencies as it pertained to their own algorithms. This was to determine what information vendors could provide in these requirement areas and whether addressing these requirements would also benefit the vendors to help improve their algorithms, thereby providing a better product for agencies in the future. A general overview of the types of applications their algorithms are used for and the testing conducted on their algorithms is provided. This is then followed by any information they could offer about the performance of their algorithms in regards to the requirements collected from agencies. However, given that Requirement 3 (determining facial comparisons performance for agency specific requirements) was sensitive, specific questions were not asked regarding this requirement. The opportunity to gain insights from vendors was also taken to gather details regarding age estimation and age progression to feed into other research currently underway and potential future research.

2.4.2.1 *General Overview of Algorithm Vendor Testing with Images of Children*

Four algorithm vendors had conducted some form of in-house testing or independent testing of their facial recognition algorithms with images of children. However, results from in-house testing were confidential so no specific details could be provided. Nevertheless, vendors were happy to provide general information where they had knowledge of the performance of their algorithms under certain conditions asked in the survey. In addition, three vendors had previously been involved in the FRVT 2013 (Grother & Ngan, 2014) and the NIST Multiple-Biometric Evaluation (MBE) 2010 (Grother et al., 2011).

No vendor was satisfied with the amount of testing that had been conducted on their facial recognition algorithms with images of children. Reasons provided for this were that:

- access to databases with annotated ages was not possible;
- there were more technical difficulties associated with algorithm development with children;
- litigious reasons made testing with images of children difficult;
- there was always scope for more testing in different environments;
- further research identifies more information and improves algorithms; and
- more time was needed on testing with images of adults due to higher market demand.

Technical difficulties with images of children was mentioned as one reason for requiring further testing. This was due to the nature of the algorithms currently being used. Pattern and shape recognition were used to detect consistent facial features. Therefore, any growth or changes in these shapes reduce the efficacy of the algorithms to detect the same face again. Thus, more testing is necessary to improve algorithms with this problem in mind.

Four vendors acknowledged their interest to supply their algorithms for the empirical studies conducted as part of this thesis as the results could help improve the performance of their algorithms by identifying specific areas requiring further investigation. One algorithm developer stated:

> *"The more research and data we can get will help to identify and improve any weak areas."*

One vendor did not wish to participate in the empirical studies as their algorithm was tailored towards very low resolution images rather than the higher quality images used in this thesis.

2.4.2.2 *Facial Comparison Performance with Images of Children and Adults*

All algorithm vendors have facial recognition algorithms that are predominantly used to conduct facial comparisons on adults. Four algorithm vendors stated they had been approached by customers or potential customers requiring their algorithms for the purpose of conducting facial comparisons on children. The use of their algorithms with children was predominantly for investigative and access control applications, however annual theme park access and time and attendance at schools were also mentioned. In addition, an example provided by vendors for an investigative application was to aid in the fight against child exploitation. Access control applications included one-to-one verification at borders, schools and other youth institutions.

Three vendors believed that there were degradations in their algorithms when used on images of children compared to adults but were unclear as to the extent of this degradation nor were they able to pinpoint specific ages. One vendor believed the reason for this was that there was typically more stability in the head and facial structures of adults, resulting in higher performance with greater age variations between images of the same person.

2.4.2.3 *Facial Comparison Performance with Images of Children at Different Ages and Age Variations*

Vendors were unable to provide a definitive age or ages in childhood where their algorithm performed best. One vendor mentioned that performance varies depending on the dataset. Another had conducted a small-scale study (with no information provided on how small) and found that age 14 and above was best, but cautioned that a larger dataset was required to make any conclusions. The reason provided for better performance from 14 years of age by this vendor was that children under 14 had faster changing facial features and structure. One vendor stated that they had not conducted sufficient testing to provide a specific age. Another vendor suggested that the older the better, with closer to 18 years of age resulting in better performance. The reason for this, as indicated by the vendor, was that faces of people closer to adulthood have more pronounced features.

Similarly, vendors were unable to provide an age or ages where their algorithm performed worst in childhood. One vendor commented that very young children were particularly challenging, suggesting that the cause may be due to all babies looking alike and having less pronounced features in their facial structures.

Vendors stated that their algorithms were more accurate when comparing images with minimal age variation. The reason provided for this was because facial features typically changed more during longer periods of time. The same vendor also noted that:

> *"For a given ageing gap (e.g. 5 years), we see bigger changes in facial features and structure in younger (0–19) age groups."*

Four vendors were aware of limitations of their algorithms due to age variation between images. Similar answers were provided by three vendors for what these limitations were with one vendor stating:

> *"The performance of the algorithms in general degrades continuously with larger temporal differences as the changes in faces tend to increase at larger time differences."*

To improve algorithm performance with images of children, vendors provided several suggestions including:

- continuing to work on improving their algorithms;
- further study and research into the ageing process of younger faces;
- specific training with images of children;
- dynamic selection;
- threshold variation for different age groups; and
- making renewal periods more frequent for younger ages.

Dynamic selection was one suggestion and refers to filtering based on appropriate parameters. For example, images may be filtered based on demographic information thereby reducing the number of returned images that are not likely to be a match. Threshold variation was also suggested. This refers to altering the threshold used by algorithms with specific ages to improve performance. For example, if algorithm performance is poorer with younger children, the threshold for younger children could be altered to make the system more secure.

One vendor mentioned that their algorithms were relatively insensitive to the age variable due to their algorithms being typically used on very low resolution images.

2.4.2.4 *Facial Comparison Performance with Mated and Non-Mated Images*

Specific questions regarding differences in performance based on pair type (i.e., mated or non-mated) were not asked in the survey as this difference was not being considered a separate requirement but rather, being investigated as part of other requirements (i.e., Requirements 1–3). However, one vendor did highlight that the time between when two mated images were taken would result in higher false non-match rates in images of children than in images of adults due to the larger changes in facial features in childhood.

2.4.2.5 *Age Estimation Performance with Images of Children and Adults*

As discussed in Section 2.4.1.6, vendors had also developed automated age estimation algorithms. Algorithm vendors' customers or potential customers had expressed an interest in adopting age estimation algorithms for profiling, marketing, retail, security, surveillance, and to identify potential errors in demographic data. Some age estimation algorithms showed

an age range (e.g., 20–25) rather than an actual age (e.g., 25). Again, the amount of research that had been conducted to determine the performance of these commercially available algorithms was limited by the datasets available to them. Two vendors have had their age estimation algorithms evaluated as part of the FRVT 2013 (Ngan & Grother, 2014).

Vendors provided different age ranges that their age estimation algorithm performed best on. For one vendor, ages 20–50 was best but they cautioned that a new algorithm was due to be released with significant improvements so this age range could change. Children and elderly were the best for another vendor, believed to be due to the richer and more discriminative features. Ages 0–17 were highlighted by another vendor as being the best. The reason provided was that the appearance of the face changes the fastest during childhood thus, age-relevant patterns are easier to detect in children than in adults.

2.4.2.6 *Performance of Age Progressed Images for Facial Comparison*

Although in existence, no algorithm vendor surveyed in this study had developed automated age progression algorithms. One vendor believed their algorithms had been used on age progressed facial images but had no more information that they could provide.

2.5 *Discussion*

Facial comparisons are conducted by algorithms and/or practitioners in many agencies around the world in order to support a number of objectives (Prince, 2013). Previous surveys and research have identified the need to understand how age-related variables impact facial comparison performance by algorithms and practitioners (Biometrics Institute, 2014; Grother et al., 2011; Hassaballah & Aly, 2015; Heyer et al., 2017). This is particularly of interest in relation to facial comparisons conducted with images of children where age is more likely to impact performance.

This study was conducted to gain an in-depth understanding of the processes used in agencies and identify key requirements that need further empirical investigation. Inviting system administrators, facial comparison practitioners, and algorithm vendors to participate at such an early stage of the research ensured that relevant operational input was sourced, that a holistic approach involving multiple perspectives was considered, and that the empirical

studies in this thesis would be as beneficial as possible to stakeholders (Creswell, 2014; Martin et al., 2006; Patton, 2002; Walker, 2004).

It was identified that the most common paradigm when conducting facial comparisons was to perform a one-to-one comparison. This was consistent with findings from past research (Heyer et al., 2017). Evaluating performance using the one-to-one paradigm is useful to determine the upper bound levels of performance prior to evaluating one-to-many comparisons where the 'many' can vary considerably between agencies.

The quality of the images (e.g., controlled or uncontrolled) and the type of images (e.g., smart phone photo) varied depending on the objectives of each agency. As there was no consensus on image type and the aim was to evaluate performance based on age-related variables, it was decided that standardised controlled images would be used for the empirical studies. That way, the results would be informative for several agencies who currently work with controlled images, while ensuring that the performance results were based on age-related variables and not other variables that are often present in uncontrolled images such as pose, illumination, and expression.

It was also found that practitioners do not work with specific time restrictions per image comparison, although many were expected to work as quickly and accurately as possible. Deadlining the amount of time images are available for facial comparison has been common in past research. Examples include: 500 milliseconds (O'Toole et al., 2007), 2 seconds (Fletcher, Butavicius, & Lee, 2008; O'Toole et al., 2007; White, Phillips, et al., 2015), 6 seconds (Fletcher et al., 2008), 18 seconds (White, Dunn, et al., 2015), and 30 seconds (White, Phillips, et al., 2015). Others have opted for a self-paced approach (Calic, 2012; Ferguson, 2015; Heyer, 2013; Valladares, 2012; White, Burton, Kemp, & Jenkins, 2013). Although research has shown that two seconds is optimal with novices (Ozbek & Bindemann, 2011), other research with facial comparison practitioners has shown that performance improves as facial comparison practitioners have more time to view images (White, Phillips, et al., 2015). Hence, there is evidence of a trade-off between speed and accuracy. As such, the decision was made not to deadline participants in the empirical studies in this thesis, but instead advise them to work as quickly and as accurately as possible, reflective of instructions provided to them during their work duties.

A main finding from this study was the extent to which facial comparisons with images of children were required, which ranged from approximately 20 to 100%. Agencies expressed a requirement to conduct facial comparisons on images of people as young as possible. Due to a lack of research, agencies were not able to rely on empirical evidence to make their decisions on what ages were deemed feasible for algorithms and facial comparison practitioners. Instead, anecdotal evidence from facial comparison practitioners was typically used to decide what age was the appropriate minimum. Each agency had a different minimum age that they considered appropriate. This highlighted the need to understand facial comparison performance when comparing images of children so that decisions can be made based on empirical evidence.

All agencies indicated that an important requirement was to know the performance of algorithms and facial comparison practitioners when comparing facial images with an age variation of up to 10 years (or, for one agency, up to 20 years). In these scenarios, it was also important to know if performance for images taken up to 10 years apart differed depending on whether images were of children or of adults. This can help to determine the extent of the performance difference between facial comparisons conducted on images of children compared to images of adults and whether more caution and/or training needs to be incorporated into the agencies' business processes.

Facial comparisons are conducted to identify a person or ensure that a person is who they claim to be (or who they are believed to be). Therefore, the aim of algorithms and practitioners conducting facial comparison tasks is to accurately determine whether images are of the 'same' person or 'different' people (i.e., mated or non-mated). The number of mated to non-mated images that agencies are exposed to varies considerably depending on the type of application, and the consequences of a wrong decision could vary substantially. Regardless, the performance of algorithms and facial comparison practitioners based on the images being mated or non-mated also warrants investigation to determine if differences in performance can be expected. If differences exist, this could mean that further research and training is required to improve performance on specific pair types (e.g., more training that focuses on similarities and/or differences in facial features in childhood over time).

It was also determined that one agency used a different algorithm threshold setting for images of children and adults while another did not. This is an important decision that can have an impact on algorithm performance and thus will be examined as part of the algorithm studies where appropriate. Results from these studies will provide a better level of understanding as to how this procedural decision can impact performance.

A method suggested by agencies as a way to improve performance was by using age estimation. Age estimation was often required in applications where there was a need to specify if a person in an image was a child or an adult or to help categorise the person in the image (e.g., baby, toddler, adolescent). Often this was required so that agencies could determine how a specific case needed to be handled. Furthermore, it may help to reduce the workload by narrowing down the number of images facial comparison practitioners manually need to compare. Determining practitioner performance on age estimation is also beneficial for agencies where they are required to categorise a child based on age. Age estimation algorithm performance has been investigated in concurrent research by this author on the same images provided for this research with the aim to also evaluate practitioner performance on age estimation in the future. The results aim to inform a range of profiling, security, and surveillance applications.

Determining whether age progressed images enhance the performance of algorithms and practitioners compared to using an image that is several years old was suggested as one way to potentially improve performance with images of children. This was suggested for circumstances where only longer age variations between images were available or where no images were available (e.g., for missing person cases). Research in this area would help to determine whether reducing the amount of time between images helps to make facial comparisons easier or whether automated age progression software and/or forensic artists hinder algorithm and/or practitioner performance. Conducting this research is currently in the consultation stage with multiple agencies as a separate program of research to this thesis.

Based on the information provided by agencies and vendors it was evident that vendors understand their customers' needs and requirements. All vendors stated a need for further research on their algorithms to better understand their performance and to focus on deficient areas. One reason provided by vendors as to why they were not satisfied with current levels

of testing was due to the lack of access they had to images with annotated ages. A lack of access to appropriate facial images to test the age variable is a consistent finding in the literature (Akhtar et al., 2013; Fu, Guo, & Huang, 2010; Hassaballah & Aly, 2015; Lanitis, Tsapatsoulis, & Maronidis, 2013; Ramanathan, Chellappa, & Biswas, 2009). Vendors understand that there are many potential customers requiring reliable algorithms to be used on all ages including children. Providing their algorithms for use in the empirical studies in this thesis is a valuable way for these vendors to improve the performance of their algorithms. For example, if there are specific areas where degradations are identified, developers can work specifically in these areas to strengthen them, thereby improving their products, and in doing so, improving the products potentially used by various agencies.

2.6 *Summary*

This study has provided unique insight into facial comparison processes adopted in various agencies by the system administrators and facial comparison practitioners. It also captured rare insight from the algorithm vendors' perspective. Key areas that require further investigation in regards to the impact of age on facial comparison performance with images of children by algorithms and facial comparison practitioners have been extensively identified by these stakeholders. These findings complement existing research but also expand on this knowledge by identifying specific areas and ages that were crucial to agency objectives. The first three requirements were evaluated as separate studies with the fourth requirement being incorporated into each of these. Based on these requirements, the following five empirical studies were designed:

1) **Study 2A (Chapter 4): Facial Comparison Performance with Images of Children and Adults – Algorithm Study.**
   This study incorporated several million images to examine state-of-the-art algorithm performance with images of children compared to images of adults (Requirement 1, see Section 2.4.1.2).

2) **Study 2B (Chapter 5): Facial Comparison Performance with Images of Children and Adults – Practitioner Study.**
   This study incorporated 200 image pairs to examine the performance of 35 facial comparison practitioners under similar conditions to Study 2A. The aim was to

determine practitioner performance with images of children compared to images of adults (Requirement 1, see Section 2.4.1.2).

3) **Study 3A (Chapter 6): Facial Comparison Performance with Images of Children at each Age and Age Variation - Algorithm Study.**
This study incorporated several million images to determine state-of-the-art algorithm performance with images of children (0–17 years) and age variations ranging from 0–10 years (Requirement 2, see Section 2.4.1.3).

4) **Study 3B (Chapter 7): Facial Comparison Performance with Images of Children at each Age and Age Variation - Practitioner Study.**
This study incorporated 23,760 image pairs to examine the performance of 120 facial comparison practitioners from one agency under similar conditions to Study 3A. This extensive study determined practitioner performance when conducting facial comparisons with images at every age in childhood (0–17 years) and age variations ranging from 0–10 years (Requirement 2, see Section 2.4.1.3).

5) **Study 4 (Chapter 8): Facial Comparison Performance for Agency Specific Requirements – Algorithm and Practitioner Study.**
This study examined algorithm and facial comparison practitioner performance on a mock operational example to show how the data can be used to answer agency specific questions (Requirement 3, see Section 2.4.1.4).

In order for these empirical studies to be conducted, access to a large database containing several million images of children and adults, state-of-the-art algorithms, and facial comparison practitioners was necessary. The next chapter introduces the database acquired for the empirical studies and the image preparation undertaken, followed by a high level overview of the methods and analytical techniques adopted for the algorithm and facial comparison practitioner studies reported in Chapters 4–8 of this thesis.

Chapter 3.

A Methodological Primer: Image Preparation, Justification for the Methods Adopted, Performance Measures, and Analytical Techniques

3.1 *Introduction*

The success of facial image comparisons in operational contexts is reliant upon the underlying performance of the algorithms and/or the facial comparison practitioners employed for this task. The literature has thoroughly examined a number of variables that impact on performance such as pose, illumination, and expression (Hole & Bourne, 2010; O'Toole et al., 2007). However, age-related variables have received less empirical attention and are expected to impact on performance with children's faces even more than with adult's due to an extensive amount of growth occurring in childhood (Kozak et al., 2015; Ricanek et al., 2013; Yadav, Singh, Vatsa, & Noore 2014). One potential reason for the lack of research in evaluating how age-related variables impact on performance is the lack of access to appropriate databases that contain standardised images where other variables are controlled (Akhtar et al., 2013; Guo, 2013). Furthermore, databases have not been large enough to test age-related variables in much detail, often grouping data into arbitrary age groups that differ between studies, making it problematic when trying to compare results (Fairhurst, 2013). Acquiring images of children at various ages is also especially difficult, thus research conducted in this area is rare despite its critical need. A better understanding of how age-related variables

impact on performance could help inform a range of operational applications including the identification of: missing and exploited children, victims of human trafficking, illegally adopted children, and child soldiers, as well as input into visa and passport verification processes.

Past research has often failed to consider the requirements of operational applications and the empirical evidence needed to make more informed policy and procedural decisions. This has resulted in agencies relying on anecdotal evidence from practitioners conducting facial comparison tasks or research that may not be representative of operational applications.

To address this issue, Study 1 (Chapter 2) first sought to gather requirements from various agencies to determine specifically where further research is required. These requirements have helped shape the empirical studies conducted in this thesis, with the aim to rigorously investigate the performance of algorithms and facial comparison practitioners with images of children under a range of age-related conditions.

Based on Study 1, five empirical studies were designed. Each of these studies examine aspects of the facial comparison task that required further investigation. To ensure that the empirical studies were as operationally relevant as possible, a large database of controlled operational facial images was used to empirically examine performance as it pertains to age-related variables. Furthermore, the studies have incorporated either state-of-the-art facial recognition algorithms (Study 2A, 3A, and 4) or facial comparison practitioners from a government agency that conduct facial comparisons as part of their regular duties (Study 2B, 3B, and 4).

This chapter will describe in more detail the operational database that was acquired for this research including the rigorous approach to establish a groundtruth for the database and the selection of images for the empirical studies. It also provides a high level overview of the methodologies and analytical techniques applied for the algorithm studies and the facial comparison practitioner studies, including justifications as to why specific approaches were adopted.

3.1.1 *A Controlled Operational Facial Image Database*

A controlled operational facial image database was acquired from an Australian government agency for this research. The FRVT 2002 (Phillips et al., 2003) defined an 'extremely large database' as one that contains at least 3,300,000 individuals and at least 10,000,000 stored samples. Although reporting the total size was not permissible in this thesis, the database acquired for this research was 'extremely large' and the number of images used in each study is provided where appropriate.

The database contained several million, front-on, neutral expression, coloured facial images. Each image was between 17 kB to 451 kB in size (mean = 74 kB, SD = 22 kB). Image quality standards ensured variables such as pose, illumination, and expression were controlled as much as possible (International Civil Aviation Organization, 2016). These standards, as well as the technology (e.g., cameras and scanners), have improved over time, thereby improving the quality of the images within the database. Figure 6 provides an example of the types of images that were and were not accepted into the database.

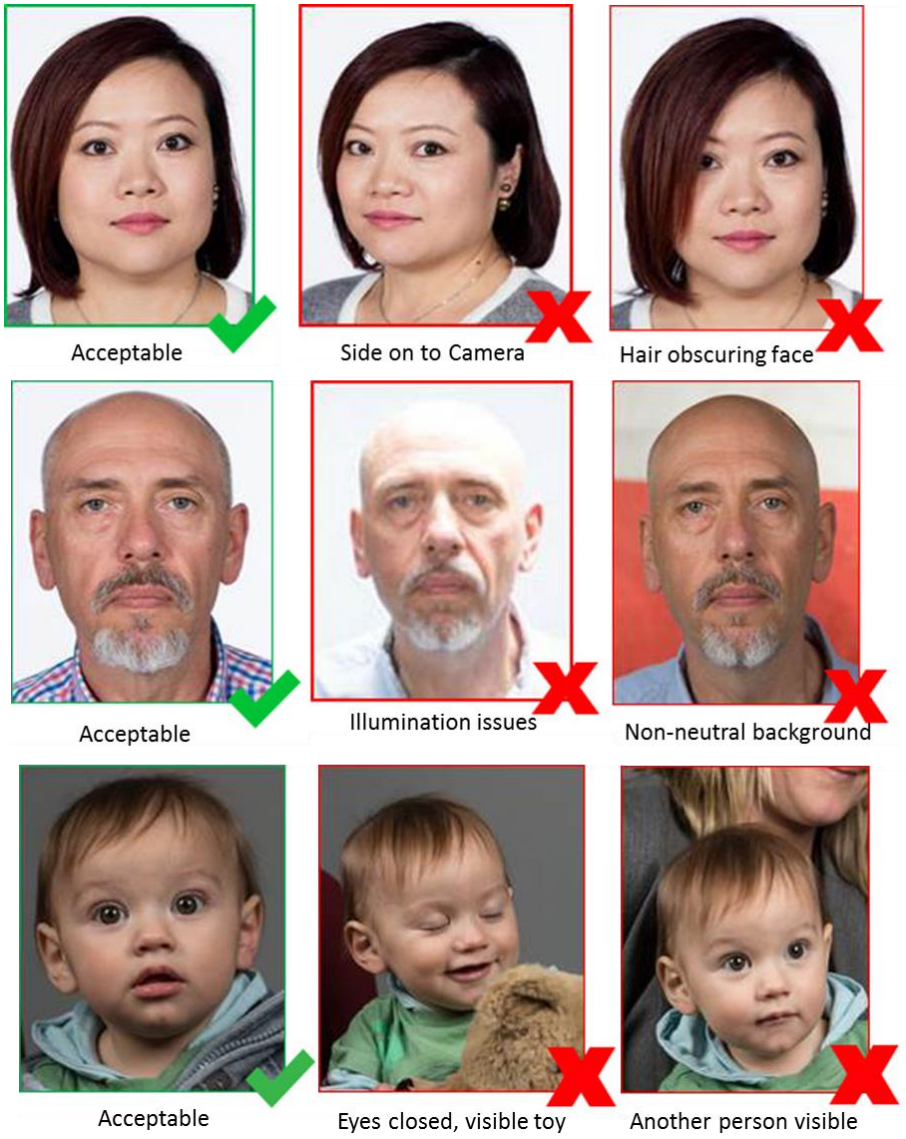


Figure 6. Example of the types of images that were and were not accepted into the database.[6]

The ages of people in the database ranged from less than one month to 105 years of age. The number of facial images per person in the database ranged from 1–14 images. For those with multiple images, the age variation ranged from one month to 13 years.

People who supplied facial images to the government agency had the option to have their images excluded for research purposes by contacting the agency or ticking a box on the application form. No identifying information, such as name or address of the person in each

[6] Copyright © 2016 by the Department of Foreign Affairs and Trade. Images used with permission.

image, was supplied with the images for this research. Instead, a random ID number for each person in the database was provided, which was then converted to an ID number consistent with the naming convention used on other projects by the Biometrics Team at DST Group, and to suit the naming convention used by the state-of-the-art algorithms evaluated in this thesis. Metadata supplied for each image included the date of birth of the person, the date when the image was supplied, and gender. No information regarding ethnicity was available, however, the ethnicity of the people in the database reflected the Australian population. The date of birth and the date that the image was provided to the agency were used to determine the age of the person in the image (i.e., date provided to agency down to the month minus date of birth down to the month).

A limitation of this database was that the researcher did not personally capture the images and therefore had no control over when the images were taken. Therefore, it cannot be certain that the people in the images were the exact age calculated to be. It is possible that an image of a person could have been taken when they were younger than when they submitted the image. However, regulations are in place by the agency to minimise this concern, as the applicant, as well as a witness, must sign a declaration stating that the image was no more than six months old. Furthermore, this type of image is generally not taken until needed. Therefore, this issue was likely to be minimal.

Past research has typically examined age-related variables using databases that have other variables present that are known to impact on performance. This often includes pose, illumination, expression, and blur which makes it problematic to determine what variable is impacting on accuracy and to what extent (Akhtar et al., 2013; Guo, 2013; Lanitis et al., 2013; Ramanathan & Chellappa, 2006b). The database used in this thesis is indispensable as it ensured that an operationally valid dataset was used, but it also minimised the presence of other variables commonly known to impact on facial comparison performance. Due to the size of the database provided and the standardisation of the images in the database, large-scale studies of algorithm and facial comparison practitioner performance that explore age and age variation with operational images were able to be conducted in this thesis. This has not been possible on this scale in previous research.

3.1.2 *Failure-to-Enrol Rate of the Database*

The failure-to-enrol (FTE) rate is when the system fails to create a template of a person's biometric that is of suitable quality for subsequent comparisons. The failure-to-enrol rate is expressed as the proportion of enrolments that do not succeed (Schuckers, 2010). Appendix E contains the failure-to-enrol rates by algorithm for the images used in the studies, based on whether images contain children or adults (Study 2A) and at each individual age in childhood (Study 3A). As vendors can tune the enrolment parameters to suit different types of images, failure-to-enrol rates should not be considered as the most suitable metric when comparing algorithms. It is also important to note that these images have already been accepted into an operational database and so have already been successfully enrolled by a state-of-the-art algorithm.

3.1.3 *Data Integrity Checking of the Database and Pre-Testing of the Algorithms*

Consistent with past research by Yiu et al. (2015), the following two steps were undertaken on the database prior to conducting the five empirical studies:

1) Data Integrity Checking

   An operational database can contain several possible issues surrounding data integrity that may not be found in publicly available research databases. Thus, preliminary assessment of the database was conducted to ensure the integrity of the database prior to the empirical studies. This was achieved by considering the possible scenarios that can result in errors in the database:

   a) one identity, multiple people (identity fraud);
   b) multiple identities, one person (assumed identities); and
   c) identical images that are used more than one year apart (regulations are in place that instruct people to provide images no more than six months old).

   Determining the extent of these possible errors required manually viewing a subset of low matching mated images (error a), high matching non-mated images (error b), and mated image pairs with a perfect match score when images were submitted to the agency over a year apart (error c). As very few image pairs were considered errors (and

removed from testing), it was determined any additional errors in the database would be minimal and not likely to impact on performance due to the size of the database. Therefore, manually searching beyond this subset was not conducted due to the lack of errors identified and the amount of human resources that would be required.

2) Set-Up and Pre-Testing of the Algorithms

Each vendor assisted with the instalment of their algorithm by attending DST Group (apart from one vendor who was available to assist via email). Where necessary, algorithms were tested on images from the FERET database (Phillips, Moon, Rauss, & Rizvi, 1997) at DST Group and compared to results obtained by vendors to ensure results were replicable.

3.1.4 *Secure Storage of the Database*

The database was stored on the Biometrics Team's High Performance Computing Cluster (see Yiu et al., 2015) in a secure laboratory. The standalone cluster is a high performance computing environment used by the Biometrics Team at DST Group to conduct large-scale biometrics research in a secure environment on images of various classifications. It is located in a secured laboratory on a secure DST Group facility in Edinburgh, South Australia. By distributing the workload over 13 nodes (108 cores), biometrics research can be effectively conducted using several different state-of-the-art commercial algorithms on extremely large databases of images. In addition to the data integrity checking and pre-testing of the algorithms, this cluster computing environment was used to store and test each algorithm used in the algorithm studies (Study 2A, 3A, and 4) and to select images for the facial comparison practitioner studies (Study 2B, 3B, and 4). Access to this database was restricted to researchers with appropriate security clearances and a need to access the database for this research as agreed with by the agency that supplied the database.

3.1.5 *Structure of the Database*

For the purposes of the empirical studies conducted in this thesis, the database was divided into mated images (more than one image per person) and non-mated images (only one image per person). Images were then divided into gender and age. The age of the person in an image was grouped by chronological age in years for Study 2A and 2B and then down to months for

Study 3A and 3B when a finer level of examination was required. Figure 7 provides a diagram of the structure of the database for Study 2A and 2B.

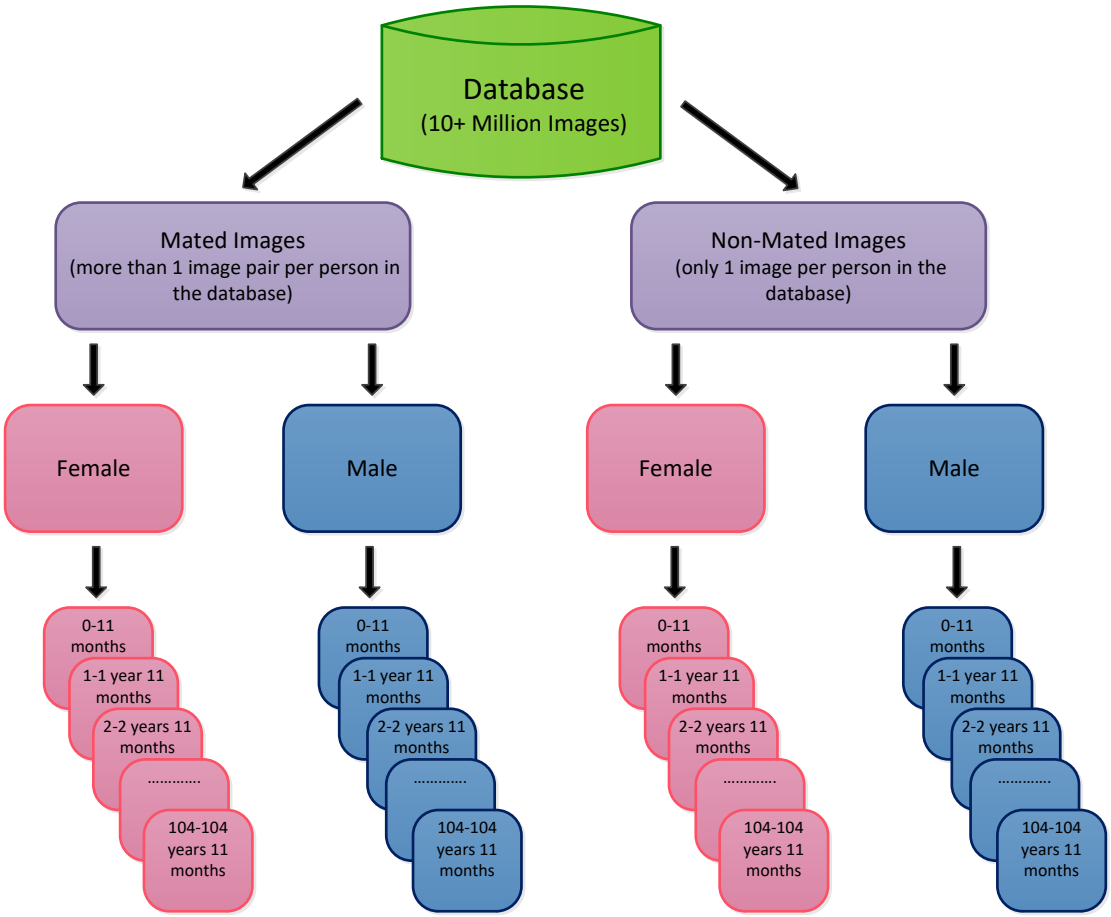


Figure 7. The structure of the controlled facial image database for Study 2A and 2B.

3.1.5.1 *Ethnicity Considerations*

Own-ethnicity bias (sometimes referred to as the other-race effect or own-race effect) is a well-known phenomenon that refers to the tendency to better recognise people from one's own ethnicity than people from another (Hugenberg, Young, Berstein, & Sacco, 2010). Algorithm evaluations have shown that an ethnicity bias exists in some facial recognition algorithms (Grother et al., 2017). The MBE 2010 (Grother et al., 2011) found that five of the six algorithms evaluated performed better on images of people of African descent than on Caucasians. Three of the algorithms also performed better on images of American Indians and Asians compared to another algorithm. This ethnicity bias is likely to be a result of the different training datasets used by the various vendors. This was evident in a study by Phillips, Jiang,

Narvekar, Ayyad, and O'Toole (2010) where it was found that Caucasian faces were easier to recognise than Asian faces by Western-developed algorithms and Asian faces were easier to recognise than Caucasian faces by Asian-developed algorithms. Similarly, human studies have also shown that own-ethnicity bias exists amongst humans conducting facial comparison tasks (Hole & Bourne, 2010; Rhodes, 2013). The bias is believed to be a result of humans tending to have more experience with people from their own ethnicity or race than with people from others.

Although removing other ethnicities and focusing solely on Caucasians to remove this ethnicity bias was considered for this research, it was decided not to adopt this approach for several reasons. This included wanting to ensure that:

1) the empirical studies represented the Australian population and not a subgroup of it;
2) variables (e.g., ethnicity) that could not be controlled at the time of image acquisition were not removed (pose, illumination, expression, and glasses are variables that can be removed in some national security contexts);
3) due to some low level categories in Study 3B not containing enough image pairs of Caucasians, incorporating other ethnicities ensured statistical analyses could be conducted and on an even number of image pairs per category;
4) algorithm results were not skewed and provided an accurate representation of what could be expected on a database containing images from the Australian population;
5) all facial comparison practitioners could participate, not just Caucasian facial comparison practitioners;
6) a resource intensive and subjective task of manually sorting through and labelling every image with an ethnicity was not necessary; and
7) the methodology was more representative of how government agencies operate (i.e., they must make facial comparisons regardless of ethnicity) and the results reflected their business processes.

3.2 *Conduct of Algorithm Studies and Facial Comparison Practitioner Studies*

The aim of this thesis was to better understand how algorithms and facial comparison practitioners perform with images of children as a function of age and age variation. Algorithm and practitioner studies were conducted independently of each other as they required different methodologies to be adopted. Best practices from algorithm evaluations (Grother et al., 2011; Yiu, et al., 2015) were adopted and expanded on for the algorithm studies, while best practices from psychology (Burton, 2013; Calic, 2012; Heyer, 2013; White, Dunn, et al., 2015) were adopted and expanded on for the facial comparison practitioner studies. Because of this, there were methodological differences that would make parallel comparisons of results between algorithms and practitioners potentially misleading. It should be noted that the aim of this research was not to compare algorithm and practitioner performance directly, but rather, to see how algorithms and practitioners performed under various age-related conditions that may be expected in operational applications using methods appropriate from the two distinct fields. Instead, trends in performance between algorithm and practitioner studies are beneficial to provide a general sense as to how similar or different they perform under the same conditions. Similarly, as this was an examination of two age variables (chronological age and age variation) rather than a direct examination between algorithms, vendor results are not provided on the same plots.

In addition, as the aim of this research was to examine the age-related variables, gender results were combined. Although providing performance data based on the gender of the person in an image was considered, this was viewed as a deviation from the intent of this research, particularly as examining gender independently was not identified as a requirement during Study 1 (Chapter 2) by agencies. This may be because agencies are not likely to make different procedural decisions based on whether images contain a male or female. In addition, the amount of data that would need to be reported based on gender as well was considered too extensive for this one thesis alone (i.e., instead of 50 pages of plots in the appendices, there would be 150). As such, age and age variation were strictly the only variables under investigation in this thesis. However, a recommendation provided in the final chapter (Chapter 9) is for future research to examine performance based on gender.

It should also be noted that the intention of the research reported in this thesis was to investigate the more applied aspects of performance that were of benefit to agencies rather than investigate them from a theoretical perspective. However, possible theoretical explanations are provided to explain some results and these are discussed in more detail in the final chapter (Chapter 9) as potential areas to explore in future research. The remainder of this chapter outlines the methodology adopted for the algorithm and facial comparison practitioner studies, followed by a summary.

3.3 *Algorithm Studies*

Six facial recognition algorithm vendors who had Material Transfer Agreements with DST Group were invited to participate in this research via email or face-to-face meetings with a representative from each vendor. Five chose to participate; four of these vendors participated in Study 1. Vendors each supplied a Software Development Kit (SDK) of their most current facial recognition algorithm free of charge for this research. These vendors are currently used by various government agencies worldwide, thus results provide algorithm data that is valuable to a range of agencies as well as to the vendors themselves. Additionally, an open source algorithm OpenBR (Klontz, Klare, Klum, Jain, & Burge, 2013) was used to provide performance results from a freely available algorithm.

The six algorithms were used to perform similar age-related facial comparison tasks to the facial comparison practitioners, but on a much larger dataset. Similar to the NIST algorithm evaluations (Grother & Ngan, 2014; Grother et al., 2011), the algorithm studies conducted as part of this thesis used a black box testing method. This means that there was no interest in or exposure to how the algorithms actually work. Rather, the interest was in determining results for these algorithms under age-related variables with a particular focus on images of children to provide results that are of benefit to both agencies and the vendors.

3.3.1 *Selection of Images*

One advantage of facial recognition algorithms over facial comparison practitioners is that they can deal with an extremely large number of faces (Patterson, Sethuram, Albert, Ricanek, & King, 2007). As such, the empirical studies incorporated more images for each of the algorithm studies than was possible during the facial comparison practitioner studies. Due to the several million images being compared in the algorithm studies, it was not possible to

remove other variables as conducted for the facial comparison practitioner datasets (see Table 6) as the database was too large. However, as the database was controlled, these variables were likely to 'washout', particularly due to the large number of images available. As mated and non-mated images were selected differently and varied between studies, they are discussed in more detail in each study chapter.

The results for the algorithm (and practitioner) studies were based on people at younger ages (Image A) being compared to people at older ages (Image B). Although it was expected that algorithms would perform similarly regardless of whether an older or younger image was being used to select the second image in a pair, the algorithm data was evaluated both ways to be certain and consistent results were found. However, for ease of reporting, results are presented based on a younger image (Image A) being used to select an older image (Image B).

Typically, algorithm studies have focused on a false match rate that has been based on zero effort impostors (Grother, 2004; Grother et al., 2017; Wayman, 2016). Zero effort implies that an imposter is not making any effort to be incorrectly recognised (Wayman, Possolo, & Mansfield, 2010). In real applications where someone is deliberately being fraudulent, zero effort impostors are not likely to represent the real amount of false acceptances. People deliberately trying to circumvent the algorithm could use a range of physical and/or technological techniques. Physical techniques could include make-up, plastic surgery or facial masks (Anjos, Komulainen, Marcel, Hadid, & Pietikainen, 2014). Technological techniques could include altering information about an applicant for a visa (Fladsrud, 2005).

Although a more realistic scenario on which to base a study is when an impostor has deliberately gone to great lengths to avoid detection, this is something that could not be achieved in this thesis as methods employed by an impostor are too varied and developing effort imposters was out of scope for this thesis. Therefore, it was decided to follow past algorithm evaluations and conduct research where non-mated pairs were based on zero-effort (Grother & Ngan, 2014; Grother et al., 2011) but with an extra step. DST Group algorithm evaluations typically involve selecting non-mated pairs based on similar parameters as those used for the mated pairs (Yiu et al., 2015). For example, the non-mated pairs selected in this thesis contained the same gender and the same ages as those selected in the mated pairs. This method was chosen as many of the operational applications involving one-to-one comparisons

are likely to be based on people of the same gender and age. For example, if a six year old girl went missing two years ago, her missing person's photo is likely to be compared by investigative agencies to a current image of an eight year old girl rather than a current image of a 30 year old man. Therefore, comparing mated and non-mated pairs based on the same gender, ages, and age variations provides more precise results for many operational applications. This approach has also recently been included as part of NIST FRVT evaluations (Grother et al., 2017). It is worth reminding the reader, however, that false match rates could still be higher in operational applications if people have made an effort to avoid detection. Although, not all instances are necessarily fraud as some situations may also arise due to human error, for example, during data entry.

3.4 *Explanation of Variables used to Measure Algorithm Performance*

In one-to-one facial comparisons conducted by an algorithm in an operational context, an image is compared to its matching corresponding image in that person's record, for example. A single similarity score is produced and compared against a threshold to confirm or deny the verification of the person (Leonard, 2016). Testing of the algorithms during each study was conducted in a threshold independent manner (i.e., a match threshold setting of zero) allowing all match scores to be used for analysis. This is typical in algorithm evaluations (Yiu et al., 2015).

When facial comparisons are conducted, two types of error can occur; a false match and a false non-match, as discussed in the following sections. These two metrics were used to measure algorithm performance in this thesis and are commonly used in one-to-one algorithm evaluations (Grother et al., 2011; Yiu et al., 2015).

3.4.1 *False Match Rate (FMR)*

The false match rate is the proportion of times an algorithm incorrectly declares a pair of non-mated images as belonging to the same person at a given threshold (i.e., the algorithm returns a score that is above the set threshold suggesting that the image pair contains the same person when in fact they are different people). This is known as a Type I error (Poh et al., 2012).

3.4.2 *False Non-Match Rate (FNMR)*

The false non-match rate is the proportion of times an algorithm incorrectly rejects a mated pair of images at a given threshold (i.e., the algorithm returns a score that is below the set threshold suggesting that the image pair contains different people when in fact they are the same person). This is known as a Type II error (Poh, Chan, Kittler, Fierrez, & Galbally, 2012).

3.5 *Data Analysis Techniques for Algorithm Studies*

Detection Error Trade-off (DET) curves, cumulative probability plots, and heat map data matrices were used to measure algorithm performance, as discussed in the following sections. Rather than conducting statistical analyses between variables, the data is presented in ways that are typical for algorithm evaluations (Grother et al., 2017; Grother et al., 2011; Yiu et al., 2015) and that are of relevance to agencies and vendors.

3.5.1 *Detection Error Trade-off (DET) Curves*

The DET curve (Martin, Doddington, Kamm, Ordowski, & Przybocki, 1997) is used to plot error rates for all match thresholds and is commonly used in algorithm evaluations (Grother et al., 2011; Yiu et al., 2015). The false match rate and false non-match rate are plotted on separate axes. This gives uniform treatment to both types of error making it beneficial in applications where trade-off between the two error types is required. The DETs have a logarithmic scale that spreads out the plot and distinguishes similarly performing conditions more clearly.

In this thesis, algorithm scores were binned in up to 10,000 bins. This is because 10,000 points per DET line was considered sufficient to demonstrate algorithm performance while reducing computational expense when generating the large number of plots required in this thesis.

An example of a DET plot is shown in Figure 8. The closer the data is to the bottom left, the better the performance. To provide an example of how to read this plot, the circle at point A shows a false match rate of 0.001 (0.1%), giving a corresponding false non-match rate of 0.104 (10.4%). This means that, on average, if the algorithm was set to incorrectly match a non-mated pair of images 0.1% of the time, 89.6% (100% - 10.4%) of mated pair comparisons would be successfully verified by the algorithm.

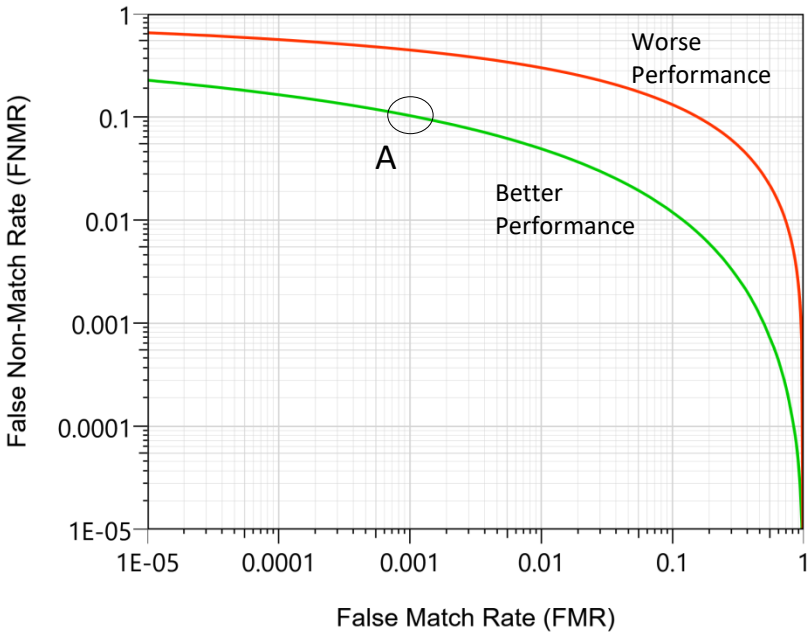


Figure 8. Example of a Detection Error Trade-off (DET) curve.

3.5.2 *Cumulative Probability Plots*

Cumulative probability plots were also used to report performance. Cumulative probability plots display error rates for the mated (i.e., false non-match rate) and non-mated (i.e., false match rate) image pairs separately. This allows for easy identification of causes for differences in performance between different conditions. Figure 9 provides an example of a cumulative probability plot displaying false non-match rates and false match rates with images of children (red lines) and images of adults (green lines). It shows that the difference in performance is predominantly due to the false non-match rate (i.e., there is a difference in performance between mated images of children and mated images of adults). False non-match rate lines that are further to the left indicate poorer performance with mated pairs whereas false match rate lines further to the right indicate poorer performance with non-mated pairs. Better performing algorithms or datasets will have minimal overlap between the false non-match rate and false match rate so that a threshold can be set that minimises both errors (Yiu et al., 2015).

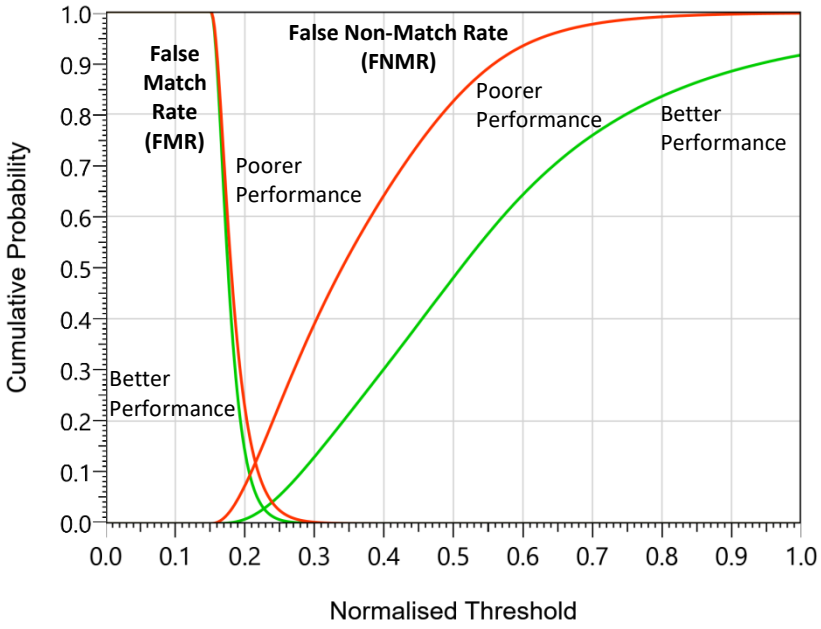


Figure 9. Example of a cumulative probability plot.

3.5.3 *Heat Map Data Matrices*

The aim of a heat map data matrix is to display performance measures for each age in childhood (0–17 years) across age variations ranging from 0–10 years to quickly distinguish areas that are better or poorer performing than others. This graphical representation was chosen as it has previously been shown to several agencies and they acknowledged that the results could be easily read and interpreted. Heat map data matrices have also recently been used to present this data to international audiences (Michalski, McLindin, Heyer, & Semmler 2016) and have been incorporated into NIST evaluations to present large amounts of data (Grother et al., 2017).

Figure 10 provides an example of a heat map data matrix. This particular heat map data matrix reports the false non-match rates at a false match rate of 0.001 (0.1%) for one algorithm at every age in childhood (0–17 years) across age variations ranging from 0–10 years. Typically in algorithm evaluations, a false non-match rate at a false match rate of 0.001 is reported to compare differences in performance between different algorithms or different conditions using the one-to-one paradigm (Grother et al., 2011). Thus, in this thesis, a false match rate of 0.001 was used, unless otherwise specified.

Youngest Age of Child in Image Pair (Years)	Age Variation (Years)										
	0	1	2	3	4	5	6	7	8	9	10
0	0.736	0.752	0.756	0.782	0.785	0.785	0.778	0.782	0.833	0.846	0.848
1	0.335	0.397	0.420	0.510	0.511	0.504	0.519	0.541	0.576	0.610	0.645
2	0.278	0.216	0.316	0.348	0.375	0.361	0.372	0.392	0.424	0.499	0.541
3	0.202	0.186	0.270	0.313	0.276	0.273	0.300	0.341	0.388	0.441	0.473
4	0.167	0.195	0.197	0.224	0.210	0.223	0.247	0.280	0.346	0.389	0.415
5	0.128	0.160	0.172	0.163	0.164	0.183	0.208	0.244	0.308	0.324	0.352
6	0.138	0.098	0.107	0.118	0.143	0.154	0.189	0.211	0.267	0.280	0.317
7	0.114	0.109	0.063	0.129	0.117	0.130	0.167	0.202	0.213	0.236	0.238
8	0.048	0.054	0.075	0.085	0.105	0.123	0.148	0.174	0.190	0.206	0.211
9	0.025	0.038	0.053	0.079	0.095	0.106	0.126	0.141	0.155	0.175	0.179
10	0.041	0.033	0.041	0.064	0.082	0.091	0.111	0.110	0.123	0.136	0.157
11	0.012	0.027	0.027	0.077	0.074	0.078	0.085	0.104	0.108	0.126	0.134
12	0.045	0.028	0.038	0.038	0.052	0.059	0.069	0.082	0.094	0.103	0.119
13	0.008	0.021	0.031	0.039	0.040	0.041	0.056	0.065	0.066	0.073	0.095
14	0.011	0.012	0.019	0.024	0.026	0.032	0.037	0.041	0.049	0.057	0.069
15	0.005	0.011	0.008	0.018	0.020	0.024	0.029	0.035	0.041	0.046	0.050
16	0.008	0.009	0.011	0.014	0.015	0.017	0.021	0.022	0.027	0.034	0.031
17	0.005	0.005	0.006	0.012	0.011	0.013	0.018	0.018	0.024	0.028	0.031

Figure 10. Example of a heat map data matrix of false non-match rates at a false match rate of 0.001.

To provide an example of how to read this matrix, age 0 (displayed in the left column) represents images of children aged 0–11 months, age 1 represents images of children aged 1 year 0 months to 1 year 11 months, and so on. A 0 year age variation (displayed across the top) represents an age difference of 0–11 months, a 1 year age variation represents an age difference of 1 year 0 months to 1 year 11 months, and so on. The red circle shows the false non-match rate for images where the youngest child in a pair is 3 years old and the age variation between images is 6 years. The false non-match rate is 0.300 (30%). This means that on average, when an algorithm is set to incorrectly match a pair of non-mated images 0.1% of the time (FMR = 0.001), the algorithm will correctly match mated pairs 70% (100% - 30%) of the time (FNMR = 0.300).

It is important to note that the heat map colours are arbitrary. Poorer performance in the algorithm heat maps is represented in blue while the better performing areas are presented

in purple and midpoint as yellow. However, blue may not necessarily mean "bad" for a particular agency, nor purple mean "good". The colouring is simply to identify the trends and performance changes over different conditions. The colouring varies between the algorithm and practitioner studies to ensure readers do not directly compare results as data between studies is not directly comparable. In addition, algorithm data is presented using error rates and practitioner data is presented using accuracy rates (as well as other performance metrics). This suits common approaches used from the two distinct fields, which also helps to minimise direct comparisons. However, trends in the data can be easily compared and can help to provide a general understanding as to whether the algorithms and practitioners perform similarly under the same conditions.

3.6 *Facial Comparison Practitioner Studies*

A large number of facial comparison practitioners from a government agency in Australia were selected as participants for the facial comparison practitioner studies (Study 2B and 3B). This provided a rare opportunity to evaluate people currently conducting facial comparisons as part of their daily role in an operational setting, ensuring a more accurate representation was obtained.

During Study 1 it was identified that one-to-one comparisons where a 'same' or 'different' decision is required was the most common paradigm used by the participating agencies. It was determined that baselining the age variable by evaluating it on a one-to-one paradigm was also essential as an initial empirical investigation. This was rather than beginning with a one-to-many paradigm where the number of images per candidate list can vary between agencies and are likely to result in poorer performance due to additional distractors.

Other decisions had to be made in regards to the selection of images for the facial comparison practitioner studies. This included the number of mated to non-mated images to present in each study and how to select non-mated image pairs. These issues are discussed in more detail with the justification of methods adopted for each.

3.6.1 *Ratio of Mated to Non-Mated Image Pairs*

Past research has been divided on whether a higher amount of non-mated image pairs in research compared to operational applications provides different performance results. For

example, Papesh and Goldinger (2014) found that the infrequency with which people encounter fraudulent images (i.e., non-mated images) will have a large impact on their ability to detect when there are fraudulent images. This suggests that in operational applications where non-mated prevalence is lower compared to the research, the ability to detect non-mated pairs in operational applications may also be lower. This is known as the *low-prevalence effect* (Papesh & Goldinger, 2014). However, others have found no evidence of a low-prevalence effect when making facial comparison decisions (Bindemann, Avetisyan, & Blackwell, 2010; Calic, 2012). Regardless, it is worth cautioning the reader about the potential of a low-prevalence effect and that accuracy results for non-mated pairs presented in this thesis could potentially be lower than what could be expected in operational applications. However, if a low-prevalence effect was to occur, the trends found in the data that are based on age-related variables are not likely to change as a result.

Although it would have been preferable to show facial comparison practitioners less non-mated pairs than mated pairs to mimic typical operational applications, it would be impossible to satisfy all applications where each has a different mated-to-non-mated ratio. It is also not feasible due to the extensive amount of time and resources that would be required to make these studies more realistic. For example, identity fraud in passport applications has been estimated as 1 in 400 applications (BBC News, 2007). Therefore, to make Study 3B specifically relevant to passport applications, approximately 24,060 hours of facial comparison practitioner time would be required to conduct the same study. Furthermore, it was found in Study 1 that the number of fraudulent or non-mated images also varied for each agency and in most instances, was only an estimate. As such, it was decided to adopt the 50% mated (i.e., same person in both images) and 50% non-mated (i.e., different people in each image) approach as it is consistent with the majority of past research examining human performance (Burton et al., 2010; Heyer, 2013; Megreya & Burton, 2006; Megreya & Burton, 2008; White, Kemp, Jenkins, & Burton, 2014).

3.6.2 *Pre-Selection of Image Pairs*

The selection of image pairs to use in the facial comparison practitioner studies is discussed in each empirical chapter where appropriate. However, the criterion used to select pairs along with an explanation of mated image pair selection is provided here, followed by a more in-depth exploration and justification of the methods adopted to select non-mated pairs.

Although a controlled image database was provided for the empirical investigations in this thesis, there were instances where other variables were present, for example, some expression differences at younger ages in childhood and image quality differences between images. As such, a list of criteria was devised that each image pair needed to comply with in order to be part of the practitioner studies. This list is presented in Table 6 along with the justification for each criterion.

Table 6. *Criteria and Justification for Selection of Image Pairs*

Criteria	**Justification**
No to minimal pose, illumination, or expression issues	To ensure the age variable was being tested in isolation as much as possible from other variables known to impact on performance and to keep consistent with ID document standards.
No blur	Blur may impact on performance, but can also cause eye strain.
No occlusions	No occlusions on the face such as glasses that could be removed at the time of acquisition.
Neutral background	To remove distractions in the background (such as a mother in the background of a baby photo).
Similar image quality between images in a pair	To reduce the possibility of image quality impacting on performance, particularly over longer age variations on mated pairs.
Loosely similar appearance e.g., ethnicity and gender	To ensure image pairs were not too easy and to keep consistency between and within pairs.

Mated images were selected based on the variables under investigation (age and age variation), but also had to comply with the criterion presented in Table 6. For example, mated pairs also had to be selected based on similar image quality. This was to ensure that results on mated pairs were not impacted by image quality differences, particularly at longer age variations when one image in the pair was taken several years earlier when image quality standards and technology may have been less precise. Not taking this step could make it difficult to determine whether any degradations at longer age variations were due to age-related variables or due to image quality. Other variables that were removed were ones that could be controlled at the time of acquisition (in some applications). For example, pose, illumination, expression, blur, glasses, and background noise. This was to ensure consistency throughout image selection and to remove other variables that are known to impact on performance. Ethnicity is a variable also known to impact on performance, but cannot be controlled at the time of acquisition and so remained in the datasets with a conscious effort to keep differences in ethnicity consistent throughout the various conditions that were selected as part of the image datasets.

Non-mated image pair selection has varied in past research and may involve panel choice, algorithm selection, or random choice. Performance can vary based on which method is selected (Calic, 2012). In this thesis, a more novel approach was adopted to make selection of non-mated images more realistic to operational applications, but to also ensure that other variables apart from age were controlled as much as possible. Firstly, this involved selecting an image based on the variables under investigation. Secondly, each of these images were compared to other images in the database based on the appropriate gender and age categories. This was achieved by using a state-of-the-art algorithm from the vendor that the agency supplying the operational database currently uses.

It was considered that the second image that should be included in the image pair be the top scoring non-mated image produced by the algorithm. However, algorithms can return images that are easily detected by a facial comparison practitioner as being different (e.g., a Caucasian could be the top scoring image returned for an Asian). This could lead to inconsistencies throughout the dataset selected and results based on variables other than age. Furthermore, having some very easy pairs may also change a facial comparison practitioner's decision threshold. For example, if a facial comparison practitioner is presented with a pair containing

a Caucasian face and an Asian face they may infer that all non-mated pairs will be easy in the studies.

In operational applications, the top 10 or 100 scoring images (or any other number decided by the agency), or any images over a certain threshold may be returned by a facial recognition system for evaluation. It was decided that the top 10 scoring images would be returned based on 10 being a figure similar to that used by more than one agency interviewed in Study 1. However, non-mated images selected were also required to fulfil several criteria as presented in Table 6 and under some conditions, 10 images was not adequate for this down-sampling stage. This was due to circumstances such as differences in image quality or ethnicity. Therefore, to keep consistency throughout pairs and reduce other non age-related variables, the top 30 scoring non-mated images were returned for each query image selected and an appropriate image from this 30 was selected to be used as part of the pair in the study.

Once images had been returned by the algorithm and an appropriate image from the 30 available had been manually selected, two independent judges inspected the selected images to ensure they complied with the list of criteria presented in Table 6. This approach was used as the interest was not in selecting the easiest or most difficult pair as decided by a panel but rather, to maintain consistency throughout the dataset by ensuring pairs selected were based on the same criteria. This approach preserved rigour in the methodology and ensured similarity levels were maintained, while reducing non age-related variables.

Typically, in operational applications such as document renewal, the latest image of a person (Image B) is compared to the database of images that the agency has acquired over time. This database will likely contain a younger image of the person, for example, from their previous outdated document (Image A). Therefore, the oldest (Image B) is being compared to the database to return an image of the applicant at a younger age (Image A). However, in other applications, such as missing and exploited children cases, it may be that an earlier image (Image A) is being compared to an older image (Image B). For example, if a child has been missing for a long period of time, an old image from before the child went missing (Image A) may be provided to an investigative agency and used to compare to images of children posted on social media at the older age the child would be (Image B). Thus, for ease of understanding and interpretation, the more intuitive approach where images are selected based on the

youngest person in a pair (Image A) being compared to a database of images where the second image contains the older individual (Image B) was used throughout this thesis.

As the studies conducted in this thesis involved the one-to-one paradigm, this change was not expected to impact on facial comparison practitioner performance. This is because images were selected from the same database regardless of whether Image A was used to select Image B or vice versa. Although different images would have been selected for the datasets, the ages and age variations remained the same, and all images were required to fulfil the criteria presented in Table 6, keeping consistency throughout the selection process.

3.7 *Explanation of Variables used to Measure Facial Comparison Practitioner Performance*

A range of variables were used to measure facial comparison practitioner performance. An overview of these performance variables has been adapted from Heyer (2013) to suit this research and is provided in Table 7. These will be explained in the following section.

Table 7. *Variables used to Measure Facial Comparison Practitioner Performance*

Variable	Definition	Units	Range
Accuracy	The percentage of correct responses overall and based on pair type (i.e., mated and non-mated image pairs)	%	0 – 100
Confidence	Self-reported confidence in the decision made (in 10% increments)	%	0 – 100
Response Time	The time taken to make a decision (measured from when images appeared on the screen to when the decision was made)	Seconds	> 0
A' Discrimination	Ability to distinguish between mated and non-mated images	-	0 – 1
B" Bias	Tendency to respond 'same' or 'different'	-	-1 – 1

3.7.1 *Accuracy*

Accuracy is one of the most common variables used to measure practitioner performance in facial comparison studies and is based on the percentage of correct decisions made (Calic, 2012; Ferguson, 2015; Heyer, 2013; White, Dunn, et al., 2015). Accuracy for mated and non-mated image pairs was also presented independently as a percentage (rather than presenting accuracy on mated pairs and errors on non-mated pairs as is sometimes reported). This was to ensure easy comparison between the large amounts of data presented based on image pair type (i.e., mated or non-mated), particularly in Study 3B (Chapter 7).

3.7.2 *Confidence*

Confidence is a measure that represents insights into the decision making process of facial comparison practitioners. Research has shown that practitioners have good

confidence-accuracy calibration with images of adults (Semmler, Heyer, Ma-Wyatt, & MacLeod, 2013; Stephens, Semmler, & Sauer, 2017). For example, high confidence but low accuracy may suggest unwarranted confidence in a practitioner's ability to make accurate decisions (Graves et al., 2011). Confidence ratings are not typically provided by facial reviewers in operational settings, however, some facial examiners and forensic facial examiners have adopted the use of opinion scales to provide a degree of confidence to their decisions (Bromby, 2003; Prince, 2013). An example is 'no match', 'image rejection', 'possible match' (Prince, 2013). Confidence ratings were collected on a continuum scale ranging from 0% to 100% in 10% increments. This scale has been used by others in similar research (Heyer, 2013; Valladares, 2012).

3.7.3 *Response Time*

Response time is another common performance measure and is used to calculate the time it takes for practitioners to make a decision on a facial comparison task when they have not been deadlined (Calic, 2012; Heyer, 2013). It was identified in Study 1 that facial comparison practitioners were not deadlined in operational applications so it is of benefit for agencies to know how long it takes practitioners to make facial comparison decisions, particularly across different conditions.

3.7.4 *Discrimination and Bias*

Discrimination and bias were reported as they provide a deeper understanding of the sources of poor performance. Discrimination refers to how well someone can distinguish between targets (mated images) and non-targets (non-mated images). Bias refers to a person's tendency to respond 'same' or 'different' and can be considered as a type of decision threshold (McNicol, 2005).

As the data was significantly skewed (see Section 3.8.3), non-parametric measures of discrimination (A' or A prime) and bias (B" or B double prime) were chosen (Stanislaw & Todorov, 1999). These measures do not make assumptions about the underlying response distributions and have been used in previous facial comparison studies (Bucks, Garner, Tarrant, Bradley, & Mogg, 2008; Butavicius et al., 2011; Fletcher, Butavicius, & Lee, 2008; Heyer, 2013). As indicated by Abdi (2009), scores for discrimination vary from 0 to 1, with 0.5 indicating chance performance and 1 indicating perfect discrimination performance. Scores

for bias vary from -1 to 1 with values less than 0 indicating liberal bias (i.e., respond 'same' more often) and values greater than 0 indicating conservative bias (i.e., respond 'same' less often).

3.8 *Data Analysis Techniques used to Measure Facial Comparison Practitioner Performance*

A range of techniques, such as notched boxplots, heat map data matrices, and statistical tests were used to analyse and visualise the performance measures from the facial comparison practitioner studies (Study 2B, 3B, and 4), as discussed in the following sections.

3.8.1 *Notched Boxplots*

Descriptive statistics for overall accuracy, confidence, and response time were presented in Study 2B using notched boxplots. Notched boxplots are a non-parametric way to graphically present data (Dunstone & Yager, 2009). They provide a distributional summary in a compact format and are useful for comparing distributions across groups. They also enable comparisons between centres of the distributions (Salkind, 2006), something that is not possible with the commonly used bar charts.

Notched boxplots can be used to indicate whether the medians of two groups are significantly different. If the corresponding notches do not overlap, there is a statistically significant difference at the 0.05 level (Wicklin, 2010). Notched boxplots have been used in previous biometric research to present accuracy, confidence, and response time of facial comparison practitioners (Heyer, 2013). Appendix F provides an example and explanation of a notched boxplot.

3.8.2 *Heat Map Data Matrices*

Heat map data matrices were used in this thesis as a graphical representation for the large amounts of data acquired as part of Study 3A and 3B. As part of Study 3B, they were used to display results for each of the performance measures (accuracy, confidence, and response times). Heat map data matrices were explained in further detail in Section 3.5.3.

3.8.3 *Statistical Analyses*

Data from each study was screened prior to analyses to check for missing data and to assess normality. A large amount of data was significantly skewed and therefore violated the assumption of normality. Past research shows that skewness on variables such as accuracy and response time is typical (Burton et al., 2010; Heyer, 2013; Valladares, 2012), thus, transformations of the data was not undertaken. Instead, non-parametric tests were chosen. When more than two conditions were compared, the Friedman ANOVA was used. If significant results were found, a Wilcoxon Signed Rank test was performed on groups of interest to determine which groups were statistically significant from each other (Field, 2013). To protect against Type I error, the standard Bonferroni adjustment was used where necessary (Shaffer, 1995). Effect sizes for the Wilcoxon Signed Rank tests were calculated using $r = z/\sqrt{N}$ (i.e., dividing the z value (test statistic) by the square root of N, with N being the number of observations over the two conditions) and interpreted using the criteria outlined by Cohen (1988) of large (r ≥ 0.5), medium (r ≥ 0.3), and small (r ≥ 0.1) effects. Formulas for the non-parametric measures of discrimination and bias were sourced from Stanislaw and Todorov (1999).

3.9 *Summary*

This chapter has provided a high level overview of the database acquired for this research, including the rigorous and methodological approach adopted to groundtruth the database and select images for the empirical studies to be reported in Chapters 4–8. It has also provided an overview of the methods and analytical techniques used for the algorithm studies (2A, 3A, and 4) and the facial comparison practitioner studies (2B, 3B, and 4), including justifications for the approaches taken.

The following five chapters describe the empirical studies that were conducted as part of this thesis. This commences with the first algorithm study (Study 2A), designed to provide empirical evidence to feed into Requirement 1 (conducting facial comparisons with images of children and adults). This study examined the impact that age has on performance by conducting facial comparisons with images of children compared to images of adults under the same conditions.

Chapter 4.

Study 2A: Facial Comparison Performance with Images of Children and Adults — Algorithm Study

4.1 *Introduction*

Study 1 (Chapter 2) aimed to better understand how agencies conduct facial comparisons and to collect requirements that warrant further empirical investigation. Four requirements relating to both algorithm and facial comparison practitioner performance that were identified in Study 1 are examined in this thesis. This chapter reports on the first requirement, to determine facial comparison performance with images of children and adults, from an algorithm perspective (Chapter 5 reports on the practitioner perspective).

Facial comparisons with images of children have typically been neglected in the literature, training, and testing of algorithms but have fast become an essential portion of some government agencies business processes. During Study 1, it was identified that many agencies use algorithms to conduct facial comparisons with images of children and images of adults. Typically, agencies conduct facial comparisons on images with up to a 10 year age variation, although in some instances, it can be longer. In addition, facial recognition algorithms could potentially become an important investigative tool. For example, one algorithm vendor identified that they had been approached by an investigative agency with an interest in using

the technology for the purpose of identifying children in child exploitation images. If current algorithms are effective with images of children, they could be implemented into business processes to reduce the burden on investigating officers viewing such horrendous images. However, data to inform these decisions is scarce.

One agency that participated in Study 1 indicated that they were considering purchasing a facial recognition system dependent on its performance with images of children. Algorithm vendors were aware that their algorithms had been implemented into some applications that require facial comparisons with children, particularly for access control purposes such as at borders and in schools. Despite this, algorithm vendors surveyed in Study 1 mentioned that their algorithms were typically designed and used for the purposes of comparing adult faces and little research or testing has been conducted on the performance of children in their systems. Furthermore, vendors were unclear as to the extent of the degradation of their algorithm's performance when used with images of children compared to images of adults, but acknowledged that a degradation was likely to exist. As such, a study investigating the performance of state-of-the-art algorithms on operational images of children and adults was deemed of value not only to agencies, but also to vendors to better understand the performance of their own algorithms. Such performance data provides vendors with an indication of the upper bounds of performance for images of children, compared to images of adults. It may also be useful to provide insights to vendors enabling a better understanding of whether the algorithms used with adults can be improved for use with children too, or if the degradation is so extreme that the development of separate algorithms for the purpose of comparing facial images of children may be a better solution. These results may also assist agencies to determine whether they need to put contingencies in place to mediate any concern. For example, whether stricter human adjudication processes with images of children, relying on other available data, or postponing implementation of systems for use with children until algorithm performance has improved should be considered.

The physical ageing process is expected to impact algorithm performance with both images of children and adults. However, given the amount of craniofacial growth that occurs in childhood (Ricanek et al., 2013), the ageing process is likely to have more of an impact with images of children. Although researchers have investigated the physical ageing process and its impact on facial comparisons by algorithms, this has predominantly been with facial images of

adults (Albert, Sethuram, & Ricanek, 2011; Lui, Bolme, Draper, Beveridge, Givens, & Phillips, 2009; Ricanek & Tesafaye, 2006; Singh, Vatsa, Noore, & Singh, 2007). Studies that have incorporated images of children have all found that the younger a person is in an image, the poorer the algorithm performance (Ferguson, 2015; Grother & Ngan, 2014; Mahalingam & Kambhamettu, 2012).

Due to different growth patterns in children compared to adults (Kozak et al., 2015; Ricanek et al., 2013), performance may differ on mated and non-mated image pairs between these two groups. For example, one vendor in Study 1 suggested that algorithm performance with images of children would result in higher false non-match rates than images of adults over the same age variation (i.e., performance with mated images of children would be worse than performance with mated images of adults). This concern has been echoed by the Biometrics Institute (Lee, 2015). When the Australian Government announced it would introduce new legislation allowing the use of biometrics with children, the Biometrics Institute made a statement warning that biometrics work less accurately on children whose bodies had not fully developed. The concern was that the use of biometrics with children could lead to high incidences of both false accept (synonymous with false match) and false reject (synonymous with false non-match) rates. This concern is supported by the findings of the FRVT 2013 (Grother & Ngan, 2014) which evaluated one-to-many facial comparisons of state-of-the-art algorithms. The results found that younger children were simultaneously difficult to recognise as themselves and difficult to tell apart from others. This is likely due to children having less discriminating facial features. However, to date, no study could be found that has been conducted to determine algorithm performance on controlled images of children compared to adults with up to a 10 year age variation, as typically expected in operational applications.

The FRVT 2002 (Phillips et al., 2003), FRVT 2006 (Phillips et al., 2007), and MBE 2010 (Grother et al., 2011) evaluations conducted by NIST included one-to-one testing on a dataset of visa applicant images that deviated from standardised quality and contained predominantly images of adults (15+ years old). The results over this eight year period showed that algorithm performance had improved. At a false match rate of 0.001, false non-match rates reduced from 0.2 (Phillips et al., 2003), to 0.026 (Phillips et al., 2007), and 0.003 (Grother et al., 2011) respectively.

Ricanek et al. (2015) evaluated the performance of Cognitec's FaceVacs-SDK 8.3, a state-of-the-art system and OpenBR (Klontz et al., 2013), among other facial recognition techniques such as principal components analysis (PCA) and linear discriminant analysis (LDA). A dataset containing 1,705 uncontrolled images of 304 child celebrities 2–16 years of age was collected. Cognitec reported the best results with a false non-match rate of 0.63 at a false match rate of 0.01. OpenBR produced a false non-match rate of 0.75 at a false match rate of 0.01. Although the NIST and Ricanek et al. (2015) evaluations were conducted with some state-of-the-art algorithms, performance differences between the studies were large. These differences may be explained in part by the different quality image sets employed between the studies, but it may also be a result of the difficulty algorithms have with images of children compared to adults. Therefore, a detailed study is required to examine performance of state-of-the-art algorithms with images of children and adults that are of the same image quality to establish whether, and to what extent, a performance difference with images of children compared to images of adults exists.

4.2 *Research Questions*

Past research in this space has been limited due to a lack of access to: controlled images of children, controlled images of adults, and state-of-the-art algorithms. This has resulted in a limited understanding as to how algorithm performance differs with images of children compared to images of adults, something critical for many investigative, processing, and access control applications, as identified during Study 1. Thus, this study aimed to provide empirical data to inform Requirement 1 via three research questions.

Question 1. Is there a difference in algorithm performance when conducting facial comparisons with images of children compared to images of adults?

Question 2. Is there a difference in algorithm performance when conducting facial comparisons with images of children compared to images of adults based on the type of image pair presented (i.e., mated or non-mated)?

Question 3. Does the false non-match rate differ between images of children and adults based on set false match rates?

Considering the anecdotes from agencies and algorithm vendors from Study 1, as well as past research (Ferguson, 2015; Grother et al., 2017; Grother & Ngan, 2014), it was hypothesised that algorithms would be less accurate with images of children than with images of adults.

It was also hypothesised that since there is considerable facial change in childhood (Lampinen, Erickson, Frowd, & Mahoney, 2017) and based on the results shown in the Ongoing FRVT 2017 (Grother et al., 2017) and FRVT 2013 (Grother & Ngan, 2014), there would be higher false non-match rates with images of children than with images of adults. As younger children have less discriminating facial features and therefore are more difficult to distinguish from each other (Wilkinson, 2012), it was hypothesised that false match rates would also be higher with images of children than with images of adults. Again, this finding was present in the FRVT 2013 (Grother & Ngan, 2014) on less controlled images and based on a one-to-many paradigm.

4.3 *Methodology*

This section contains the methodology applied to test algorithm performance with images of children compared to images of adults based on a one-to-one paradigm.

4.3.1 *Design*

A technology evaluation design was employed for the algorithm studies conducted in this thesis. A technology evaluation design aims to measure the performance of a biometric system by focusing on a particular component of that system. For the purposes of this thesis, the component of interest was the algorithm. This approach is repeatable, provides performance data, and is used to identify specific areas that require further investigation (Poh et al., 2012). A technology evaluation design has been consistently used by NIST (Grother & Ngan, 2014; Ngan & Grother, 2014; 2015) and by DST Group (Yiu et al., 2015) to evaluate facial recognition algorithm performance.

4.3.2 *Participants*

Five state-of-the-art facial recognition algorithm vendors supplied SDKs of their most current facial recognition algorithm for testing purposes (designated Algorithm A–E). OpenBR (Klontz et al., 2013) (designated Algorithm F), was sourced from the internet to provide results from a free and publicly available algorithm.

4.3.3 *Materials*

Materials used in this study included the controlled operational facial image database (see Section 3.1.1), the Biometrics High Performance Computing Cluster (see Section 3.1.4), five state-of-the-art algorithms (see Section 3.3), one open source algorithm (see Section 3.3), and a plotting program developed in-house to enable plotting of large amounts of data. As the database of images and algorithms have been discussed previously, only the plotting program is discussed here, followed by the methods used to select the images for this study.

4.3.3.1 *Plotting Program*

Due to the extensive number of scores returned in this study (154,232,240), an in-house plotting program was designed that could plot the large amounts of data quickly and effectively. The plotting program was tested several times against the plotting functions used in Microsoft Excel™ to ensure the plots were consistent and that there were no bugs within the program. The plotting program can be used to plot DETs, receiver operating characteristic (ROC) curves, and cumulative probability plots (both rank order and match score threshold), which allow users to evaluate the one-to-one and/or one-to-many performance of different variables. Comma separated values files can be uploaded into the program and different variables under investigation can easily be selected and plotted. Figure 11 shows a screen shot of the program interface.

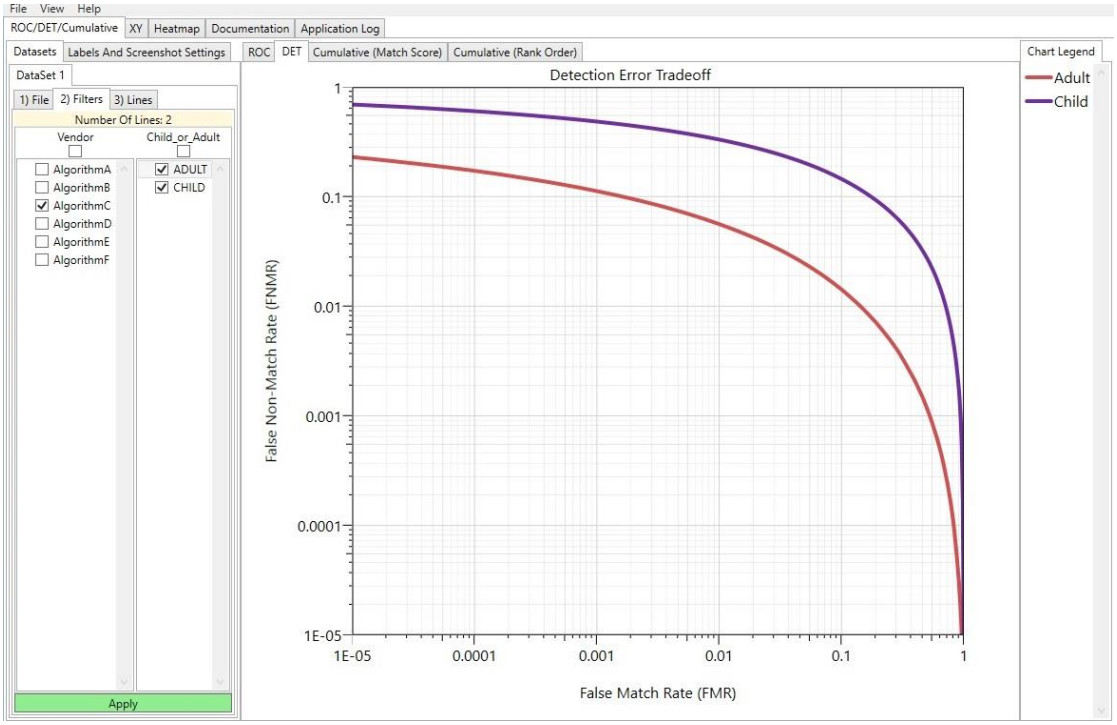


Figure 11. Screenshot of plotting program. Different variables (filters) can be applied to create the required plots.[7]

4.3.4 *Image Pair Selection*

In total, 4,628,323 images of children and 6,363,154 images of adults were used in this study. Slight variations in the final numbers that were matched varied between algorithms due to differences in failure-to-enrol rates as presented in Appendix E.

Mated image pair selection involved obtaining match scores from every mated image pair in the database that had an age variation of up to and including 10 years (i.e., up to 10 years 11 months). Images were then filtered based on the age of the youngest image in a pair to determine whether they should be in the Child group (less than 18 years) or Adult group (18+ years) as defined by Australian law. In instances where a person had more than two images in the database, match scores were collected on all possible combinations of pairs where the younger image in a pair was the query image. For example, if Image A < Image B <

[7] Note that the plotting program was designed to plot the axes using scientific notation when five decimal places or more are required. As such, the results in this thesis were plotted with the axes formatted as shown in Figure 11.

Image C, then all possible combinations of pairs would be Image A and Image B, Image B and Image C, and Image A and Image C. Non-mated image pair selection was conducted by comparing the younger image from all mated pairs to 30, other randomly selected people of the same age and gender as the older image in a mated pair. Therefore, the number of non-mated images was 30 times higher than the number of mated image pairs in the Child and Adult groups.

The reason for selecting up to and including a 10 year age variation for the algorithm studies was because agencies mentioned in Study 1 that they typically conducted facial comparisons with images up to 10 years apart and ID documents for adults are typically valid for 10 years. Although children's passports are currently only valid for 5 years, there are times when a child's passport may not be consecutively renewed. Furthermore, in investigative applications, there may be an interest in using algorithms for cold cases where a child has been missing for an extended period of time. Thus, using images with up to a 10 year age variation reflected real-world applications and helped to identify how algorithm performance differed based on whether the images contained children or adults under the same conditions.

It is important to note that due to the nature of images in the database, images of children and adults were not evenly distributed in regards to age variation. The mean age variation for the mated Child group was 5.57 years and for the non-mated Child group was 5.94 years. For the mated Adult group, the mean age variation was 6.99 years and for the non-mated Adult group was 7.11 years. It is possible therefore, that performance with images of children in this study could actually be reflecting better performance than if the database for the Child and Adult groups at larger age variations were even. However, given the limited previous research in this space, and that the images were from an operational database, the current study was still deemed to be reflective of operational applications.

4.3.5 *Procedure*

Five state-of-the-art algorithms and one open source publicly available algorithm were independently evaluated on the Biometrics High Performing Computing Cluster. Section 3.1.3 discusses the data integrity checking that was initially conducted on the database of images as well as the methodology chosen to ensure each algorithm vendor's SDK was providing

appropriate and reproducible scores. Each algorithm was tested on the same dataset. It took approximately six weeks to complete both the mated and non-mated matching for this study.

Once each algorithm had completed matching all appropriate image pairs, the match score data was sorted into files based on the algorithm and age group (Child or Adult). This data was then plotted using the in-house developed plotting program (discussed in Section 4.3.3.1).

4.4 *Results*

This section provides the results for each vendor evaluating algorithm facial comparison performance for images of children and adults (Requirement 1). Approval to report all vendor names with results was not received from all vendors before submission of this thesis. Therefore, the five state-of-the-art algorithms were reported as Algorithm's A–E and OpenBR as Algorithm F.

4.4.1 *Algorithm Performance with Images of Children and Adults*

Figure 12 shows the DET plots for each of the six algorithms on their performance with the Child and Adult groups.

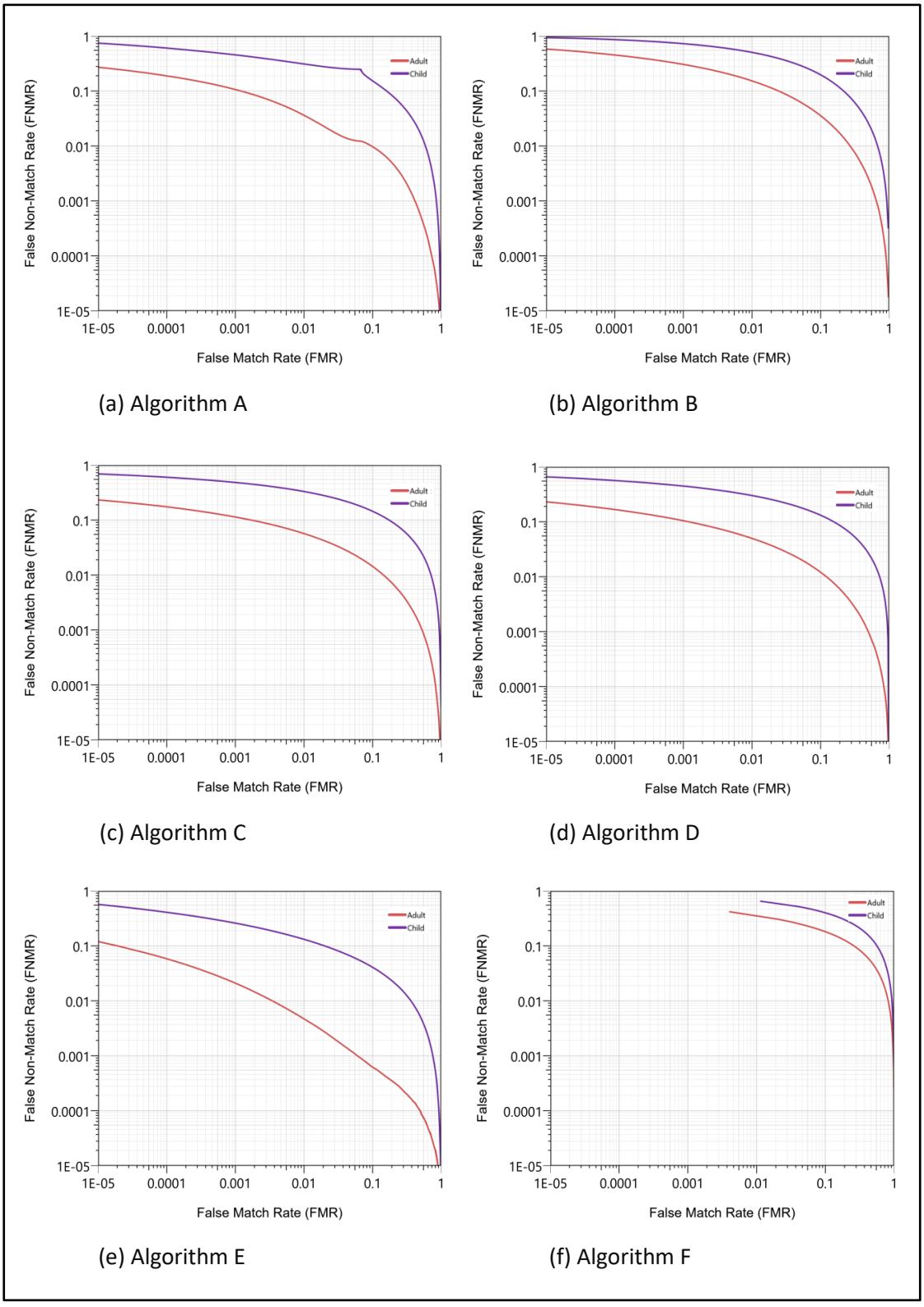


Figure 12. DETs displaying performance of each algorithm across Child and Adult groups.

As can be seen in Figure 12, the performance of each algorithm was poorer with images of children than it was with images of adults. The lack of data on the OpenBR plot was due to there being no false matches returned at high thresholds resulting in false match rates of 0, which could not be plotted on a logarithmic scale.

Algorithm E had the largest difference in performance between the Child and Adult groups, however, this algorithm also performed best both with images of children and adults than any other algorithm.

4.4.2 *Algorithm Performance with Images of Children and Adults on Mated and Non-Mated Image Pairs*

Figure 13 shows the cumulative probability plots of the Child and Adult groups for each algorithm. These plots show how the false non-match rate and false match rate performance differed across the Child and Adult groups based on pair type (i.e., mated and non-mated pairs).

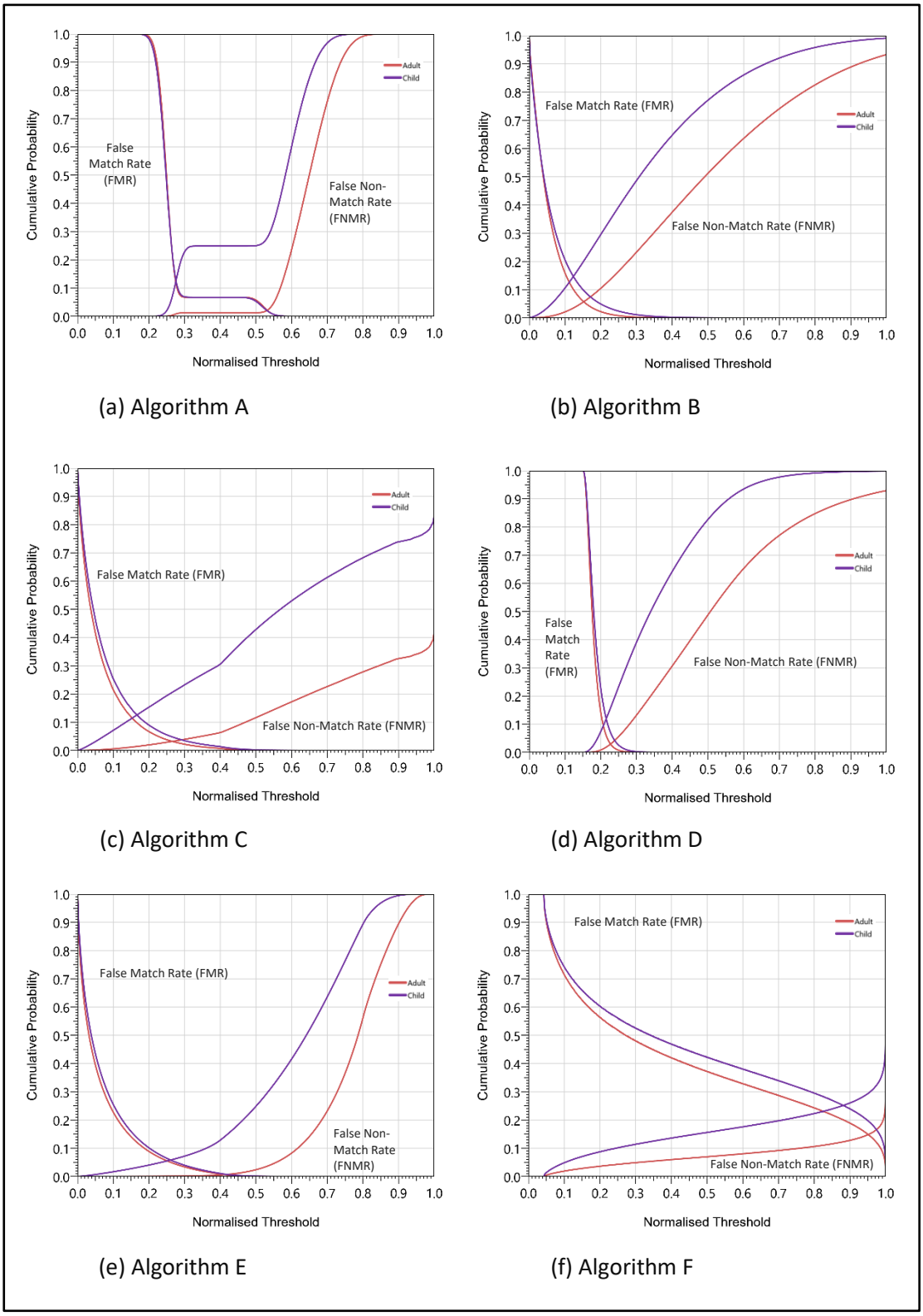


Figure 13. Cumulative probability plots displaying performance of each algorithm across Child and Adult groups.

For all algorithms, the overlap between the false match rate and false-non-match rate was higher for the Child group than the Adult group indicating that algorithms were less accurate with images of children than with images of adults. The results of Algorithms A–E also show that the false match rate (performance for non-mated pairs) was similar for the Child and Adult groups. The primary variation was in terms of the false non-match rate performance, with the Child group having higher false non-match rates at all thresholds. These results show that algorithms performed poorly with mated pairs of children, which means recognising images that are of the same child was relatively difficult for these algorithms.

4.4.3 *Algorithm Performance with Images of Children and Adults based on set False Match Rates*

Figure 14 displays a bar chart for each of the six algorithms to show how the false non-match rate varied for the Child and Adult groups when their respective false match rates were 0.1, 0.01, and 0.001. Although a false match rate of 0.001 is generally of most interest to security contexts, data was also presented in this study at set false match rates of 0.01 and 0.1 to demonstrate how performance differed and for comparison with past algorithm studies. For example, Ricanek et al. (2015) used a false accept rate of 1% (synonymous with a false match rate of 0.01) with images of children.

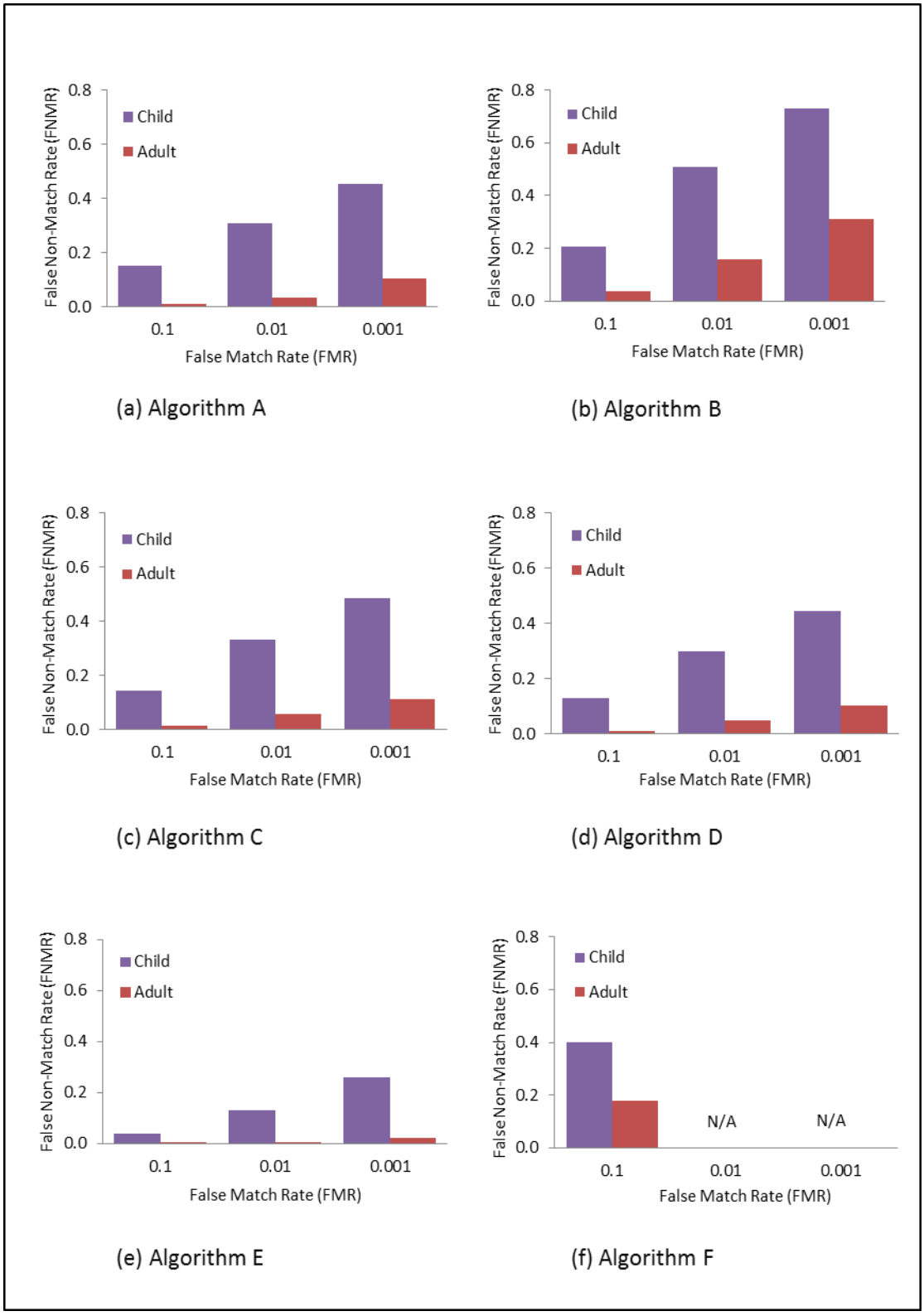


Figure 14. False non-match rates at false match rates of 0.1, 0.01, and 0.001 by Child and Adult groups.

The charts show that the false non-match rate was consistently higher for the Child group than it was for the Adult group for each algorithm at these false match rate settings. For instance, at a false match rate of 0.1% (0.001), Algorithm D was 34.2% more accurate in correctly matching mated pairs of adults (FNMR = 0.105) than mated pairs of children (FNMR = 0.447).

Algorithm F (OpenBR) had no data at false match rates of 0.01 or 0.001 due to low performance of the algorithm. It was considered to use Mathematica's interpolation function to find these values, but as this was an applied thesis examining performance for operational environments, it was decided to only present data based on the millions of images examined, rather than based on theoretical estimations.

It may be typical in an operational application to use one threshold regardless of whether images contain children or adults as identified during Study 1. As such, Table 8 provides examples of how performance would vary in this regard. Data was based on systems set at the thresholds where false match rates of adults were 0.1, 0.01, and 0.001. Again, Algorithm F (OpenBR) had no data at an adult false match rate of 0.01 or 0.001 because the algorithm performance was so poor.

Table 8. *False Match Rate and False Non-Match Rate of the Child group when the Adult False Match Rate was set at 0.1, 0.01, and 0.001*

Algorithm	**FMR Adult**	**FNMR Adult**	**FMR Child**	**FNMR Child**
Algorithm A	0.100	0.010	0.110	0.142
	0.010	0.036	0.010	0.312
	0.001	0.106	0.001	0.429
Algorithm B	0.100	0.036	0.146	0.148
	0.010	0.153	0.028	0.375
	0.001	0.306	0.005	0.574
Algorithm C	0.100	0.014	0.128	0.125
	0.010	0.057	0.018	0.287
	0.001	0.114	0.003	0.424
Algorithm D	0.100	0.012	0.176	0.092
	0.010	0.050	0.032	0.217
	0.001	0.105	0.006	0.340
Algorithm E	0.100	0.001	0.114	0.037
	0.010	0.005	0.012	0.122
	0.001	0.021	0.002	0.235
Algorithm F	0.100	0.179	0.143	0.346
	0.010	N/A	N/A	N/A
	0.001	N/A	N/A	N/A

By way of example, when the algorithm threshold was set so that the false match rate of adults was 0.001, Algorithm E correctly matched pairs of mated adult images on average 97.9% of the time (FNMR = 0.021). If an agency was to use the same threshold for images of children (i.e., where the threshold was set so the FMR = 0.001 for adults), Algorithm E would incorrectly match pairs of non-mated images of children on average 0.2% of the time (FMR = 0.002) and correctly matched mated pairs of children on average 76.5% of the time (FNMR = 0.235). This

may result in a higher workload for facial comparison reviewers if the same threshold was used for both Child and Adult groups. In fully automated settings, it would also double the amount of non-mated pairs of children that are falsely accepted by the system.

4.5 *Discussion*

The discussion section is divided into the three questions this study aimed to answer in determining facial comparison performance with images of children and adults (Requirement 1).

4.5.1 *Algorithm Performance with Images of Children and Adults*

Consistent with previous research (Grother et al., 2017; Grother & Ngan, 2014) this study showed that algorithms perform less accurately with images of children compared to images of adults. This confirms the information obtained in the vendor surveys during Study 1 (Chapter 2) that facial comparisons with images of children were more difficult. As mentioned in those surveys, vendors had less access to images of children for training and testing purposes, and due to customer demand, their main focus had been on testing and improving algorithms for use with an adult population. One vendor also mentioned that their algorithm incorporated pattern and shape information. As children's faces vary in shape considerably compared to adult's due to craniofacial growth (Ricanek et al., 2013), it could be that this hindered performance on images of children for this particular algorithm (and any others that rely on pattern and shape information). Past research shows that training algorithms on images of people at the same age as those that the algorithm would be used on improves performance (Klare, Burge, Vorder Bruegge, & Jain, 2012). Therefore, it may be more suitable for vendors to consider a separate algorithm for images of children that is trained on images of children and/or one that relies less on pattern and shape information.

Given the high cost of systems for agencies, a separate algorithm for children may not be feasible. Furthermore, some agencies already have algorithms in place for use with both children and adults. Therefore, it is important to know how state-of-the-art algorithms perform with both images of children and adults to inform agencies and vendors of the current state-of-the-art algorithm performance. Another option is for agencies to consider more stringent criteria or policies to help mitigate these differences in performance. For example, the passport renewal period in Australia, New Zealand, Canada, United States, and the United

Kingdom for children less than 16 years of age is five years and for adults it is 10 years (Passport Canada, 2012). In this example, a conscious effort has been made to reduce the degradation in performance with children by having different policies in place for children than those for adults. Given the deviations in algorithm performance with images of children compared to adults presented in this study, such an approach is supported to try and mitigate these differences. However, more research is necessary to better understand what algorithm performance can be obtained with images of children at different age variations such as 5 or 10 year renewal periods, to be more relevant for agencies. This will be explored with algorithms in Study 3A (Chapter 6) and Study 4 (Chapter 8).

4.5.2 *Algorithm Performance with Images of Children and Adults on Mated and Non-Mated Image Pairs*

It was hypothesised that the false non-match rate would be higher with images of children than with images of adults. This hypothesis was supported anecdotally in Study 1 by an algorithm vendor, but was also empirically supported in this study. This suggests that children are more difficult to match against themselves than adults. This result is likely due to more facial growth and change occurring in childhood than in adulthood (Kozak et al., 2015; Ricanek et al., 2013), as discussed during Chapter 1 (Section 1.8).

It was also hypothesised that the false match rate would be higher for images of children compared to adults due to children having less discriminating facial features (Wilkinson, 2012). The cumulative probability plots (see Figure 13) show that false match rates for the Child and Adult groups were similar (or slightly higher for the Child group) and this was consistent for all algorithms, apart from OpenBR. Table 8 outlined the results for the Child and Adult groups when assessed at the same threshold at three different operating points (corresponding to three different thresholds). The table showed that the false match rate for the Child and Adult groups varied when using the same threshold and the amount of variance was dependent upon the algorithm and chosen operational setting. For example, when the false match rate was set for adults at 0.001, the false match rate for children varied from 0.001 (Algorithm A) to 0.006 (Algorithm D). When the false match rate was set for adults at 0.1, the false match rate for children varied from 0.110 (Algorithm A) to 0.176 (Algorithm D). This highlights that agencies should consider (if they do not already) implementing different threshold settings for images of children and adults to optimise performance. In some applications, this could be

operationalised by using biographical data to conduct an age calculation which could then be used to adjust the threshold settings in the system.

4.5.3 *Algorithm Performance with Images of Children and Adults based on set False Match Rates*

In the MBE 2010 (Grother et al., 2011) the best performing algorithm with visa images (predominantly of adults) had a false non-match rate of 0.003 at a false match rate of 0.001. In the current study, the best performing algorithm with images of adults at the same false match rate had a false non-match rate of 0.021 (Algorithm E). Although it would be expected that the algorithms would have improved over time, it is possible that this difference in performance can be explained by the larger age variations between images in pairs presented in the current study (up to 10.92 years), compared to the MBE 2010 (up to 3.7 years), which may have degraded performance. Furthermore, non-mated image pairs selected in this study were of the same age and gender as those used in the mated pairs. This is likely to make non-mated pairs more difficult and as a result, the algorithms in the current study had lower overall performance compared to those in the MBE 2010 (Grother et al., 2011).

Ricanek et al. (2015) found that Cognitec achieved a false non-match rate of 0.630 at a false match rate of 0.01 with images of children. Based on the data in Table 8, the best performing algorithm in the current study with images of children achieved a false non-match rate of 0.121 at a false match rate of 0.012. This large difference in performance between the studies was likely due to the current study sourcing controlled images of children, as well as testing multiple state-of-the-art algorithms, some of which are known to exceed Cognitec's performance (Grother et al., 2011). This once again reinforces the benefit of conducting research with operational images and multiple state-of-the-art algorithms to ensure results are operationally relevant and can inform agency requirements appropriately.

Given that the Child group in this study included images where the youngest age in a pair could range from 0–17 years and the age variation could range from 0–10 years, it is likely that some ages and age variations across this span were more difficult for algorithms than others. For example, image pairs containing babies may have increased the false non-match rate for the Child group. Therefore, facial comparisons with images of children should not be considered problematic based on this one study alone. It is possible that algorithm performance with

images of children at various ages in childhood would be considered acceptable for various operational applications. Further research is required to examine algorithm performance with images of children at individual ages across childhood. This will provide agencies and vendors with a better understanding of whether performance varies and if so, to what extent. This will be explored in Study 3A (Chapter 6).

4.5.4 *Summary*

This study was conducted to provide algorithm data to determine comparative facial comparison performance between images of children and adults (Requirement 1). This study was unique in that performance of five state-of-the-art algorithms (Algorithms A–E) and one open source algorithm (Algorithm F) were tested on controlled operational images of both children and adults up to the same 10 year age variation. This makes this research highly valuable to both agencies and vendors alike. Agencies can see how much variance there is in an algorithm based on whether images contain children or adults, and vendors can use this same data to inform further in-house training and testing with the aim of improving their algorithms in future versions.

As demand increases in this space for various applications requiring facial comparisons with children, the desire to improve performance with images of children is likely to grow. Although this study found large differences in performance between the Child and Adult groups, differences in performance throughout childhood are also likely to exist. As such, a finer level investigation will be conducted to determine facial comparison performance with images of children at different ages and age variations (Requirement 2) for both algorithms (Study 3A; Chapter 6) and facial comparison practitioners (Study 3B; Chapter 7). But first, the following chapter (Chapter 5) will present results from Study 2B, which examined facial comparison practitioner performance with images of children compared to images of adults, fulfilling Requirement 1 from a practitioner perspective, and complementing the current study.

Chapter 5.

Study 2B: Facial Comparison Performance with Images of Children and Adults — Practitioner Study

5.1 *Introduction*

Study 1 identified that practitioners conduct facial comparisons on both images of children and adults. The majority of facial comparison practitioners from Study 1 were required to conduct facial comparisons on images taken up to 10 years apart regardless of whether an image contained a child or an adult. In Study 1 (Chapter 2) anecdotally, it was determined that practitioners found facial comparisons with images of children more difficult than with images of adults. However, the extent of this degradation was unknown, and requires an empirical study (similar to the algorithm study conducted in Chapter 4) to investigate further.

The lack of research in this space is alarming considering that several agencies are conducting such comparisons on a daily basis. An extensive search of the literature found remarkably few empirical studies that had evaluated human performance with images of children compared to images of adults.

A study conducted by Zeng et al. (2012) aimed to determine accuracy on one-to-one facial comparison tasks when participants were provided with images of people at various ages. The

MORPH Album I database was used and 31 students were exposed to 300 image pairs, 30% of which were mated. Overall accuracy on this task was found to be 78.8%. The results were then grouped based on the age of people in the images (i.e., <18 years, 18–29, 30–39, and 40+). Their results showed that the worst performing age group was 40+ (75%), followed by <18 years (77.5%), 18–29 years (78%), and 30–39 years (82%) on images taken up to five years apart. However, no statistical analyses were conducted on these groups. They also found that response time had a significant impact on accuracy but again, little information was provided in regards to this. Furthermore, as the ages of people contained in the MORPH Album 1 database only start at 16 years of age, the <18 year age group is not likely to be reflective of facial comparison performance across childhood.

Lanitis (2008) presented 30 participants with 100 image pairs from the FG-Net database, which contains more images of children than the MORPH Album 1 database. No additional information was provided on the ages of people presented in the image pairs nor the age variations between images. It is known, however, that this database contains images of newborns through to 69 year old subjects. Overall accuracy in this study was 66.9%.

Zeng et al. (2012) compared their study to that of Lanitis (2008) and concluded that the results for the Lanitis (2008) study were poorer than theirs. The cause of this degradation, they claimed, was due to Lanitis (2008) incorporating images of children that were considered more difficult. Although these studies were similar in that they both used the one-to-one paradigm, two different databases were used, both of which contained uncontrolled images. Using publicly available databases has been suggested as useful for being able to compare results across studies on the same database (Gross, 2005), however, comparisons in this example were made on different databases with different participants and therefore may have been misleading. Depending on what other variables were present in the images and how severe the differences were between databases, these variables could have impacted on performance. Furthermore, it is unclear if the age variation was consistent between the two studies.

In addition, from an operational perspective, the results in the Zeng et al. (2012) study were based on the performance of students and therefore may not have been representative of results of facial comparison practitioners, who have more experience and training with facial

images. The Lanitis (2008) study failed to mention who their participants were. Another difference from operational applications is that images presented to participants in both studies were greyscale, with images in Lanitis (2008) also cropped directly around the face and below the hairline. Greyscale and cropped images are less common in operational settings and thus, are additional variables which may also alter performance compared to that expected in operational applications. In fact, Lanitis (2008) presented the same cropped greyscale images to the same participants in a second session but uncropped and in colour. They found over a 10% improvement in accuracy (77.8%) compared to viewing the images in cropped greyscale (66.9%). Although it is hard to determine whether this was due to a test–retest bias or due to coloured images being used, it is worth considering that coloured images from an operational database with practitioners may yield higher results than those provided in both the Zeng et al. (2012) and Lanitis (2008) studies.

White, Dunn, et al. (2015) conducted a more operationally relevant study by using images of children, teenagers, and adults selected from a subset of the Australian passport database (containing 20,000 identities). Participants were required to complete 306 one-to-eight trials (102 trials for each age group). Images were selected for the study by using Cognitec's system to search this subset and return the eight highest ranking images for each of the 306 trials. The ages represented for each group were: 6–13 years (children), 14–22 years (teenagers), and 40–47 years (adults). As child passports in Australia are only valid for 5 years and adults are valid for 10 years, the average age variation differed for the three groups under investigation (children mean = 6.2 years, teenager mean = 6.3 years, and adult mean = 9.7 years). Trials were presented in three blocks with short breaks between blocks. Participants had a maximum time of 1 hour to complete the study with six out of the 42 participants not completing the study in the allocated time. Overall accuracy for the three age groups was: 39% children, 41.1% teenagers, and 45% adults. Although operational images were used, some considerations need to be made about the methodology employed. For example, 306 one-to-eight trials deadlined to an hour was excessive. This would not be expected in an operational setting and this was likely to have impacted on performance. An indication of this is that six of the participants could not complete the study in time. Furthermore, participants were students who were not trained in facial comparison. It is possible that if the participants were practitioners and they were allocated more time, performance would have improved, as evidenced in other research (White, Phillips, et al., 2015).

A further limitation to the White, Dunn, et al. (2015) study was that images presented were selected from a small age range for the children (6–13 years) and adult (40–47 years) groups and the entire age range for the teenager (14–22 years) group. It is unclear if the age range of 6–13 years and 40–47 years was a representative age range for both children and adults respectively. For example, in adulthood, there are periods of significant facial change and other periods with no noticeable changes at all (Ricanek & Boone, 2005). Therefore, images only selected from the 40–47 year age span may have provided artificially high or low performance results depending on the amount of facial change that is typical over this age range. As such, datasets should incorporate multiple images over a wide span of ages. The average age variation also differed between groups. Furthermore, an additional group (teenagers) was incorporated into this study that contained images from both childhood and adulthood. Agencies from Study 1 highlighted the need to know the difference in performance when conducting facial comparisons on images of children compared to images of adults. Therefore, an accurate representation of facial comparison practitioner performance with images of children and adults over the same conditions is required.

Depending on the agency's business processes, the repercussions for making the wrong decision can lead to anything from a minor inconvenience to a serious national security incident and this can differ depending on whether the images being compared are mated or non-mated. For example, from a passport renewal context, if a passport application is put on hold because there is uncertainty whether the renewal image is the same person as that from a previous passport image under the same passport number, it can become an inconvenience to that passport holder to send in more images and prove they are who they claim to be (i.e., mated pair). If two images are believed to be the same person and they are in fact different people (i.e., non-mated pair), this could result in, for example, a kidnapped child being taken to another country to be sold into slavery. Typically, in operational applications, the critical need is to ensure that people who are not who they claim to be are accurately identified. Thus, facial comparison practitioners need to be as accurate as possible with both mated and non-mated images.

Determining how facial comparison practitioners perform on mated and non-mated images may highlight where more training is required. For example, if performance is significantly poorer with non-mated images of children, agencies may consider more training that focuses

on ways to identify facial feature dissimilarities between children. Conversely, if mated images of children are the most difficult for facial comparison practitioners, agencies may consider training that focuses on specific facial features that stay more constant over time in childhood.

A study by White, Kemp, Jenkins, Matheson, et al. (2014) was conducted with 27 facial comparison practitioners from the Australian Passport Office and 38 university students using a one-to-one paradigm with images of adults taken two years apart. The results showed that participants were more accurate with non-mated images (89.4%) than mated images (70.9%) and concluded that this difference in performance may highlight that ID documents are less representative of the person over time, but that the person still does not look more like others. Participants in this study also took longer to make decisions on mated than non-mated image pairs. Limitations of the study, as mentioned by the authors, included that images were taken over short periods of time and that non-mated images were selected from a very small sample, both of which were likely to impact on performance and result in different performance levels than that expected in operational applications. Non-mated images were selected from a sample of 17 people of the same gender and this group was considered very diverse. Therefore, it makes sense that non-mated images took less time and were easier to compare than mated image pairs. This provides further justification that research examining practitioner performance with images pairs that have been selected in a more operationally realistic manner is required. This will help to determine how performance on mated and non-mated pairs may vary in an operational setting.

Locating research that has examined facial comparison performance with images of children compared to adults proved difficult. The previous studies mentioned have been limited in terms of the type of images, methodologies, and/or participants used that do not reflect typical operational applications. This study attempted to address these limitations by incorporating operational images, a methodology that was more operationally realistic, and facial comparison practitioners as participants.

5.2 *Research Questions*

The aim of this study was to provide empirical research to determine the performance of practitioners when conducting facial comparisons with images of children compared to images

of adults (Requirement 1) so that agencies can make more informed decisions. This requirement was addressed via two research questions.

Question 1. To what extent does facial comparison practitioner performance differ when conducting facial comparisons with images of children compared to images of adults?

Question 2. To what extent does facial comparison practitioner performance vary with images of children and adults based on the type of image pair presented (i.e., mated or non-mated)?

Given the anecdotes provided by agencies (Study 1, Chapter 2), the limited research available, and the differences in craniofacial growth between children and adults (Kozak et al., 2015; Ricanek et al., 2013), it was hypothesised that facial comparison practitioners would perform poorer with images of children than with images of adults.

It was also expected that non-mated images of children would result in the poorest performance due to children having less discriminating facial features than adults and therefore being more difficult to distinguish from each other (Wilkinson, 2012).

5.3 *Methodology*

This section contains the methodology applied to evaluate facial comparison practitioner performance.

5.3.1 *Design*

A repeated measures design was conducted for the three empirical facial comparison practitioner studies. A repeated measures design is used when the same participants take part in each condition of the independent variable (Verma, 2016). Therefore, each condition in an experiment includes the same group of participants. One issue highlighted regarding repeated measures is that there could be order effects. This means that the order of the conditions could have an effect on decisions made by the facial comparison practitioners (Verma, 2016). To combat this issue, the conditions in each study were randomised (Wilson & Sharples, 2015).

For the current study, Study 3B (Chapter 7) and Study 4 (Chapter 8), the independent variable was the type of images supplied to the participants (e.g., child image or adult image) and the dependent variables were the performance measures of the participants (e.g., accuracy, response time, and confidence).

5.3.2 *Participants*

A total of 35 facial comparison practitioners (26 females, 9 males) from an Australian government agency that conducts facial comparisons as part of its processing role completed the study (mean age = 38.97 years, SD = 11.01). The practitioners were experienced in making facial comparisons with both images of children and adults (experience in facial comparisons ranging from 3 months to 30 years). Practitioners were from the agency that participated in Study 1 (Chapter 2) that conducts facial comparisons on the largest age range of people at the facial review level (i.e., 6 years and above). All participants had normal or corrected to normal vision. The majority of participants (80%) were Caucasian.

5.3.3 *Materials*

Materials used in this study included the controlled operational facial image database, the Biometrics High Performance Computing Cluster, Comparer (an image comparison software tool), the experimental application software, work computers, and one state-of-the-art face recognition algorithm. The database and computing cluster have been described in detail previously (see Section 3.1.1 and 3.1.4 respectively). The remainder of the materials will be described next, followed by an overview of the methods used to select the images from the database for this study.

5.3.3.1 *Comparer - Image Comparison Software Tool*

Comparer is an image comparison software tool designed in-house by DST Group (Hole et al., 2015). It is used to assist the manual selection of images for experiments as well as for data integrity checking purposes. Figure 15 provides a screenshot of this software. The columns in this screenshot from left to right show the: ID number for Image A, ID number for Image B, a checkbox to select the appropriate image pair to use in the study, Image A (probe), and Image B (gallery). This screenshot provides an example for a non-mated image pair where the top 30 highest scoring images were returned and the first appropriate image that fulfilled the

criterion presented in Table 6 was selected. Comparer can be used to show one image pair or multiple image pairs at a time.

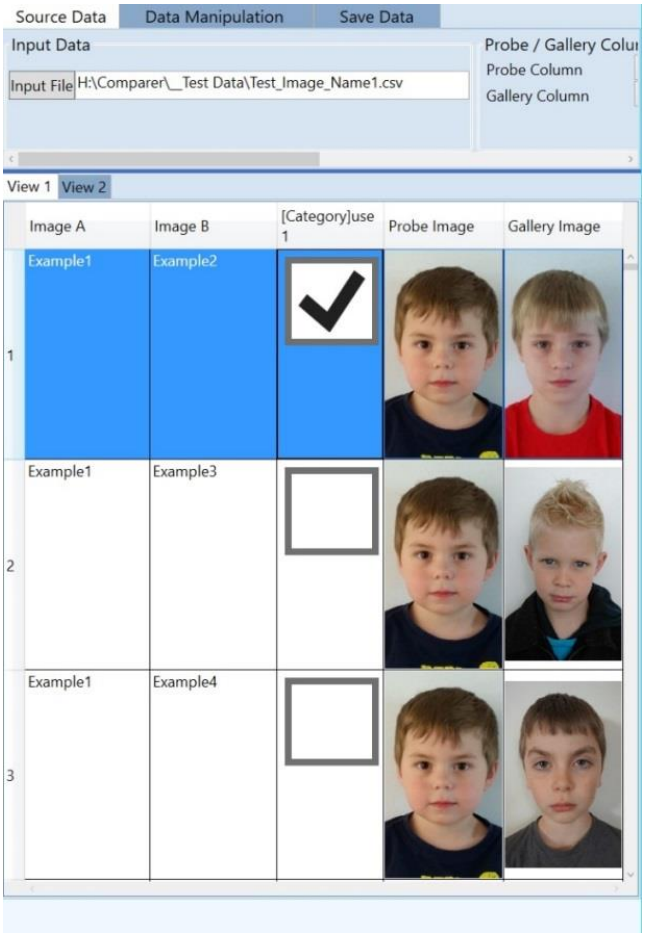


Figure 15. Screenshot of Comparer. Images are for illustration purposes only.[8]

5.3.3.2 *Experimental Application Software*

As the facial comparison practitioner studies were computer-based, experimental application software was required to run the studies. This application was designed based on the computer screen layouts used by facial comparison practitioners during their work duties at various agencies, as observed during Study 1, as well as past research evaluating human performance (Valladares, 2012; Heyer, 2013; Calic, 2012; Zeng et al., 2012). Each page of the experimental application was designed by the researcher using corresponding slides

[8] Copyright © 2017. Images used with signed parental consent.

developed in Microsoft PowerPoint™. These slides were then sent to the Westbourne Group (Westbourne Group, 2017), a third party with access to the participating government agency's test server. The Westbourne Group created the software using the provided slides as a reference. Correspondence continued until both parties were satisfied with the layout of the software.

The experimental application contained the following screens:

1) **Login** - practitioners provided their unique ID and password to gain access to the experiment;
2) **Information** - information provided about the experiment and their consent to participate in the study, consistent with a regular Information Sheet;
3) **Demographic questions** - demographic questions to gain background information about participants (e.g., age, gender, and ethnicity);
4) **Instructions** - instructions provided on what was expected of the practitioner during the experiment;
5) **Practise trial** - practitioners completed two trials to get familiar with the experiment;
6) **Experiment** - each pair of images was presented on a separate screen (as shown in Figure 17); and
7) **End of experiment** - informed practitioners that the experiment was over. This screen also offered practitioners the opportunity to provide feedback about the experiment and advised them that they could receive their results by contacting the researcher.

At the end of development phase, the experimental application was first tested on two DST Group colleagues to ensure it was in perfect working order (using publicly available images). The application was then tested at the local office of the participating agency on one facial comparison practitioner. The practitioner was then questioned following the completion of the study to determine if anything needed to be changed in the experimental application, such as wording or layout. This step was conducted to ensure that a facial comparison practitioner was involved in the process so that they could provide any additional insight that may have been overlooked by the researcher that could ultimately bias or alter responses of practitioners in any way. The practitioner acknowledged that they were satisfied with the experimental application.

5.3.3.3 *A State-of-the-Art Facial Recognition Algorithm*

A state-of-the-art facial recognition algorithm was used to conduct one-to-many searches of the database, which provided the top scoring images per search consistent with many operational applications. The top 30 scoring non-mated matches were used to select an appropriate image for non-mated pairs as discussed in Chapter 3 (see Section 3.6.2) and in more detail in the next section (Section 5.3.4). The algorithm chosen for this process was from the vendor currently used by the agency that supplied the operational database of images and practitioners for testing purposes. Therefore, images returned were images that may be expected during typical work duties.

5.3.3.4 *Work Computers*

Facial comparison practitioners used their own work computers, with screens typically around 19 inches in size, for the studies. As the studies were conducted at each facial comparison practitioner's own PC workstation, any differences in monitor size were consistent with their regular facial comparison duties.

5.3.4 *Image Pair Selection*

The experiment contained 200 image pairs. This included 100 image pairs of adults (Adult group) and 100 pairs where at least one image in a pair was of a child (Child group). An equal number of males and females were selected for the Child and Adult groups, as well as an equal number of mated and non-mated image pairs in each group. See Figure 16 for a visual representation of the image pairs.

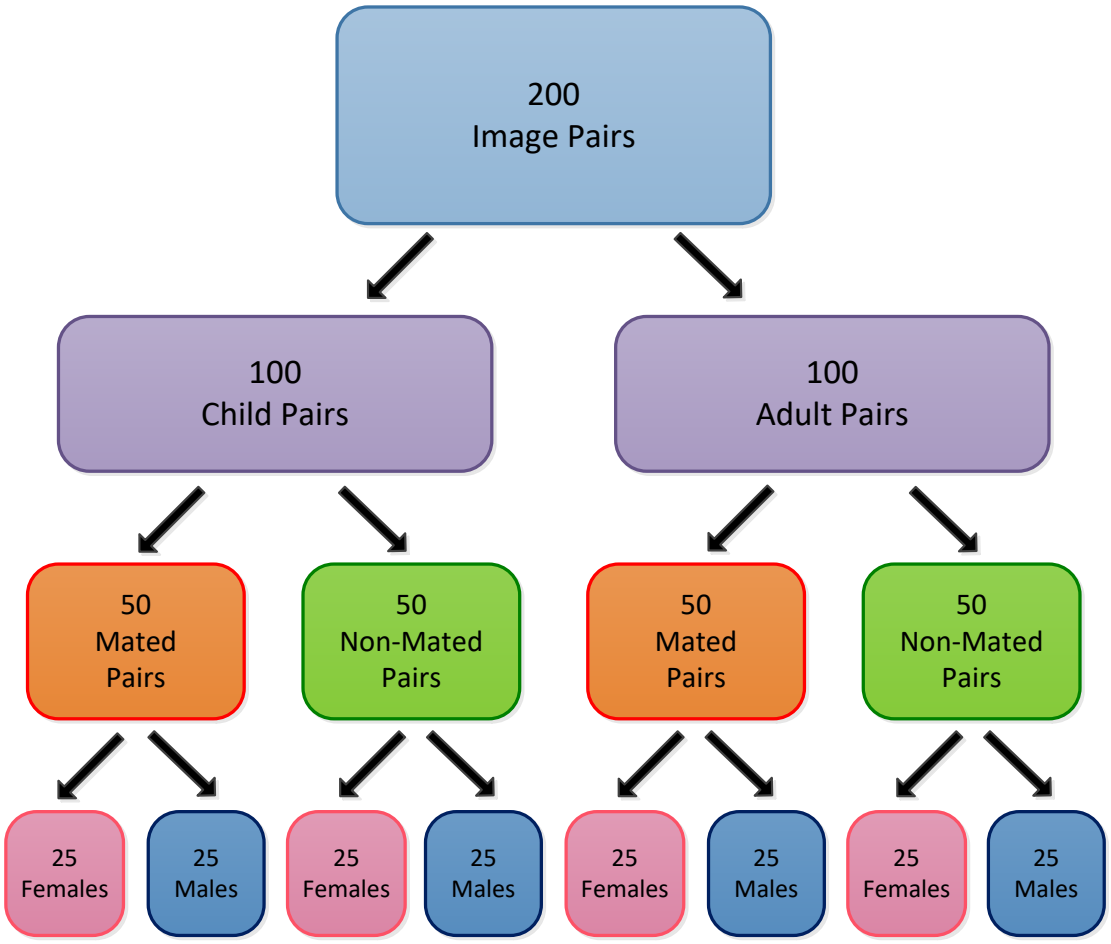


Figure 16. Image pairs presented to the facial comparison practitioners.

The group that each image pair was assigned to (i.e., either Child group or Adult group) was based on the age of the person in the youngest age image. Therefore, depending on the age variation, some images that were randomly selected for the Child group were matched with the image of an individual 18 years of age or older. These image pairs were retained as this situation can be expected in operational applications. It also ensured a proper randomised representation of the child ages, with randomised age variations, was conducted that would not have been possible otherwise. For example, if a 10 year age variation was randomly chosen for any images of children above 7 years of age, these image pairs would have not been able to be used, thereby skewing the images presented in the Child group. Allowing these image pairs to be included also ensured that the average age variation for image pairs was similar for each group: mated child image pairs (5.54 years), non-mated child image pairs (5.78 years), mated adult image pairs (5.34 years), and non-mated adult image pairs (5.42 years). Image

pairs were presented in random order to each participant, once again using Microsoft Excel™ for the randomisation. This was to ensure that there were no order effects (Wilson & Sharples, 2015). The process taken to select the mated and non-mated images pairs are discussed next.

5.3.4.1 *Mated Image Pair Selection*

To select the 50 child mated image pairs for this study, mated pairs were listed in spreadsheets in Microsoft Excel™ along with their corresponding age, age variation, and gender information. Several spreadsheets were required to achieve this as Microsoft Excel™ has a maximum of 1 million lines and there were over 1 million mated image pairs available to randomly select from. A random number generator in Microsoft Excel™ was used to randomly select an age from 0–17 years for the youngest image, an age variation ranging from 0–10 years, and a corresponding image pair based on this data. This was controlled by gender so that an equal number of male and female pairs were selected for each group as shown in Figure 16. This was conducted 80 times for the 50 child mated pairs in case any of the first 50 pairs did not fit the criteria presented in Table 6. This provided opportunity for substitution without having to conduct the process again. The ID numbers for these 80 pairs were placed in a comma separated values file and added into Comparer (see Section 5.3.3.1) to view the images and select the first 50 appropriate pairs based on criteria from Table 6.

Image selection for the 50 mated adult image pairs followed the same process, but the youngest age of the person in a pair was required to be at least 18 years of age.

5.3.4.2 *Non-Mated Image Pair Selection*

Non-mated images were selected using the non-mated portion of the database rather than the entire database. This was to ensure that images were not accidentally used twice during this study (i.e., once as a mated image pair and once as a non-mated image pair). Seven steps were taken to randomly select each of the 50 non-mated child image pairs (separated by gender):

1) an age from ages 0–17 years was randomly selected for the youngest image;
2) an image from this age was randomly selected;
3) an age variation ranging from 0–10 years was randomly selected;

4) for the images selected in step 2, an algorithm conducted a one-to-many search on images that were the appropriate age (based on steps 1 and 3) and same gender as the selected image;
5) the top 30 highest scoring non-mated images were returned;
6) the returned images were sorted from highest scoring to lowest scoring as would be expected in an operational application; and
7) the first image pair from this age and age variation that fulfilled the criteria presented in Table 6 was selected.

This was repeated until all 50 non-mated image pairs of children were selected. Non-mated image pair selection was conducted in the same way for the Adult group and with an even number of males and females.

5.3.5 *Procedure*

Approval by higher management, followed by section managers in an Australian government agency was sought prior to asking facial comparison practitioners if they would like to participate in this study. This was achieved via Project Arrangements with higher management and memos emailed to section managers. Section managers provided email addresses for all facial comparison practitioners interested in participating.

Facial comparison practitioners were sent an email the week prior to the study opening to give them ample time to plan when they would complete the study the following week, so as to minimise impact on workflow. Facial comparison practitioners logged onto the experiment via a link sent to them in an email. This approach was invaluable as it provided access to a large number of practitioners for this study and particularly for Study 3B (Chapter 7) that were separated over several states in Australia and also internationally (but who were working for the same agency). Each facial comparison practitioner was provided with their own unique ID and password to gain access. Email addresses were not kept with username and password information. Facial comparison practitioners were informed that the experiment would be open for one week. This gave practitioners the opportunity to assign a block of time of around one hour to complete the study during work hours. They were also informed that participation was voluntary, they were to work through the study alone, and that their individual results would not be provided to management. No financial incentives were provided.

Practitioners were encouraged to complete the study in one session, but had the option to stop at any time and go back to it if necessary. This was to ensure that they were not rushing through the experiment and could go back to regular duties if required. Practitioners read through a consent screen, information screen, and several screens requiring them to provide demographic details. They were then provided with the following vignette to explain their role during the experiment with the aim to set a similar decision threshold amongst participants:

> *"You will be presented with pairs of facial images. Your role is to look at the pairs of images from the perspective of an employee processing passport applications and decide whether the images are of the SAME person or of DIFFERENT people. Please work as quickly and accurately as possible. Once you have made your decision, you will not be able to change it. You will then be required to rate your confidence in your decision."*

Participants then completed two practise examples before beginning the actual experiment in order to become familiar with the layout. They were presented with the vignette again immediately before proceeding to the actual experiment containing 200 one-to-one trials. An example of the experimental screen layout is provided in Figure 17.

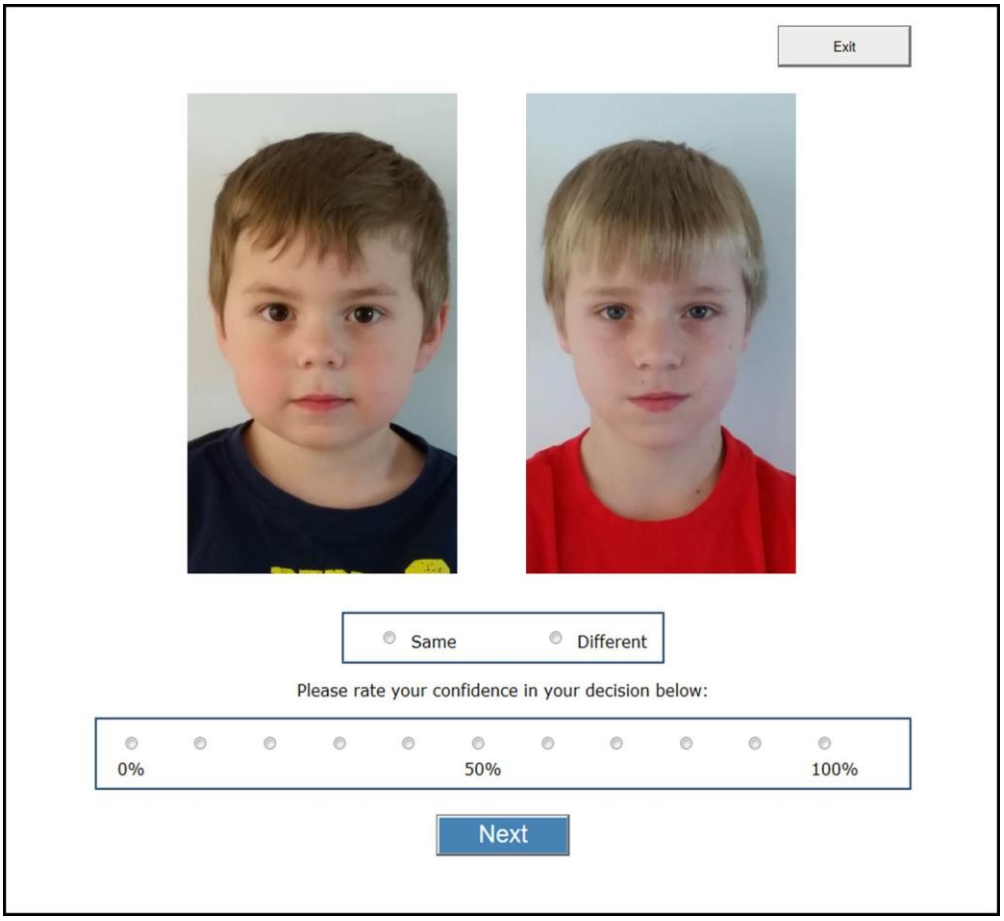


Figure 17. An example of the experimental screen layout for the facial comparison practitioner studies. Images are for illustration purposes only.[9]

The final screen allowed facial comparison practitioners the opportunity to provide feedback and reminded them to send an email with their unique ID if they wanted their results once analysed. They were informed that this would be the only time that their results would be linked to them.

5.4 *Results*

Prior to analysis, the data was screened to check for any missing data and to assess normality.

[9] Copyright © 2017. Images used with signed parental consent.

5.4.1 *Data Screening and Assumption Checking*

It was identified that six of the 35 facial comparison practitioners experienced technical difficulties due to issues with the experimental application software. This occurred on occasions when a facial comparison practitioner clicked 'Next' twice in a row. The software registered the 'Next' on the second occasion as a response to the following image pair, thereby skipping this image pair altogether. These six facial comparison practitioners saw between 196 and 199 of the expected 200 image pairs, depending on how many times the glitch occurred during their experiment (and was not likely noticeable). After discussions with the software engineer at the Westbourne Group, as well as consulting with a DST Group software engineer on the Biometrics Team, the issue was identified and it was determined with confidence that the rest of the data was reliable and could be used and would be rectified for the following facial comparison practitioner study. As a result, the data for these participants was analysed taking into consideration the missing pairs.

Testing for normality found that the data was significantly skewed for some of the variables. As such, non-parametric tests robust to non-normal data were used (as rationalised in Section 3.8.3). The Friedman test was conducted when more than two groups were being compared. The Wilcoxon Signed Rank test was conducted with a Bonferroni correction applied, resulting in a significance level set at $p < .013$ when conducting a series of pairwise comparisons between the groups on each of the performance measures. Effect sizes for each comparison were calculated and Cohen's (1988) criterion was used to determine the size of these effects (i.e., large: $r \geq 0.5$, medium: $r \geq 0.3$, and small: $r \geq 0.1$). Although the mean (*M*) is not typically reported with non-parametric data, it is reported here with the median (*Mdn*) to provide average performance data that can be directly compared to studies that typically use parametric tests (and thus only report the mean).

Inspection of response time data showed that some people had either become distracted or had opened the experiment but forgot to close it again, resulting in extremely high timings recorded for some image pairs. Response time data can be problematic particularly when participants are not deadlined or in uncontrolled testing (Field, 2013). However, this was considered a small cost as it made the process more operationally realistic and the response times were still valuable to provide a better understanding of how long practitioners take to

make decisions, as most studies have used a deadlined methodology (e.g., White, Dunn, et al., 2015; White, Phillips, et al., 2015).

A common way to deal with this issue is to remove any timings that are 2.5 standard deviations from the mean (Ratcliff, 1993; Gagné & Spalding, 2014; Lowie, Verspoor, & Seton, 2010). Although some have argued that there is some bias in this approach (Field, 2013). Upon manual inspection of the data and consulting with a facial comparison practitioner in the area, it was agreed that the resulting timings from this method would be appropriate. As such, timings above 2.5 standard deviations for both the Child and Adult groups were removed and the results were recalculated.

The following sections (Section 5.4.2 and 5.4.3) report results for the two research questions this study aimed to answer.

5.4.2 *Facial Comparison Practitioner Performance with Images of Children and Adults*

Figure 18 presents the descriptive statistics for overall accuracy and confidence for the Child and Adult groups.

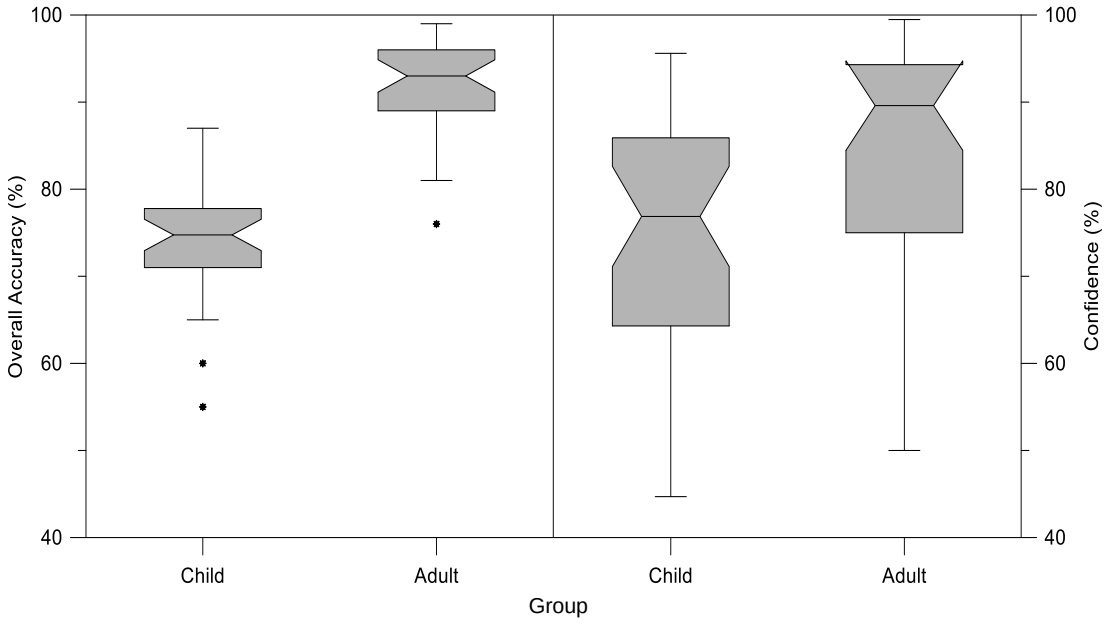


Figure 18. Overall accuracy and confidence for the Child and Adult groups.

The Wilcoxon Signed Rank test revealed a statistically significant decrease in overall accuracy for the Child group (*M* = 73.86%, *Mdn* = 74.75%) compared to the Adult group (*M* = 92.05%, *Mdn* = 93%), *z* = -5.16, *p* < .001, *r* = -.62. One participant was an outlier in both the Child and Adult groups, performing lower than the rest of the participating facial comparison practitioners. The confidence of facial comparison practitioners was also statistically significant, with facial comparison practitioners being less confident in their decisions with image pairs from the Child group (*M* = 73.72%, *Mdn* = 76.87%) than with image pairs from the Adult group (*M* = 84.81%, *Mdn* = 89.60%), *z* = -5.07, *p* < .001, *r* = -.61.

Figure 19 presents the descriptive statistics for the response times for the Child and Adult groups.

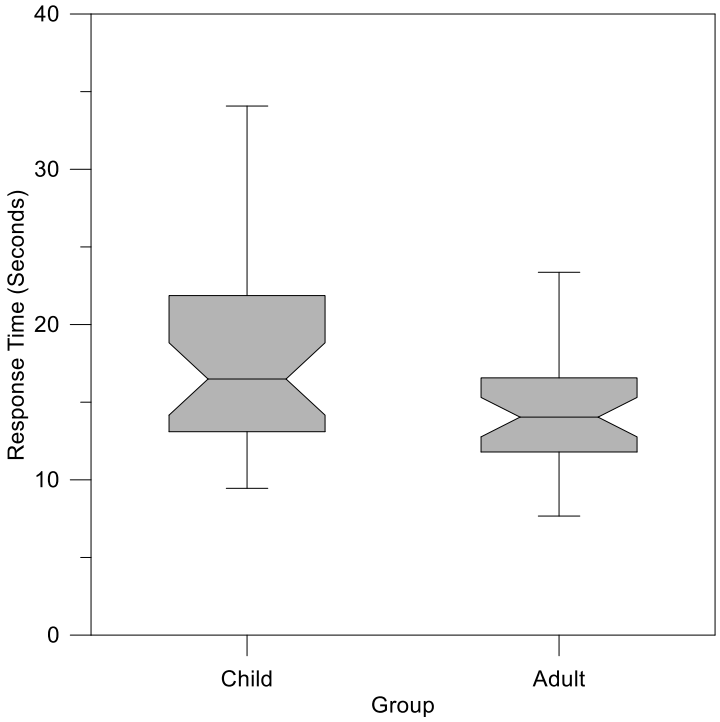


Figure 19. Response times for the Child and Adult groups.

As hypothesised, facial comparison practitioners were also significantly slower at making decisions on images from the Child group (*M* = 17.46 seconds, *Mdn* = 16.50 seconds) compared to images from the Adult group (*M* = 14.07 seconds, *Mdn* = 14.04 seconds), $z = -5.00$, $p < .001$, $r = -0.60$.

Table 9 provides the descriptive statistics and pairwise comparisons for the measures of discrimination and bias.

Table 9. *Discrimination and Bias for the Child and Adult Groups*

	Mean (SD)	Median	Min – Max	Pairwise Comparisons
A' Discrimination				
Child group	.61 (.06)	.60	.51 – .74	z = -5.14, p < .001, r = -.61
Adult group	.84 (.09)	.85	.61 – .98	
B" Bias				
Child group	-.20 (.29)	-.15	-.85 – .37	z = -3.19, p < .001, r = -.38
Adult group	-.49 (.55)	-.64	-.1 – 1	

A significant difference was found in discrimination between the Child group (M = .61, Mdn = .60) and the Adult group (M = .84, Mdn = .85), z = -5.14, p < .001, r = -.61 suggesting that facial comparison practitioners are better at discriminating between 'same' (i.e., mated) and 'different' (i.e., non-mated) image pairs of adults than they are with image pairs of children.

Although facial comparison practitioners showed a liberal bias towards both the Child and Adult groups, claiming images were of the 'same' person, there was a significant difference in bias between the Child group (M = -.20, Mdn = -.15) and the Adult group (M = -.49, Mdn = -.64), z = -3.19, p < .001, r = -.38, suggesting that facial comparison practitioners were more biased to claim images of adults were of the 'same' person than they were with images of children.

In summary, these results support the hypothesis that conducting facial comparisons with images of children is more difficult than conducting facial comparisons with images of adults. When the facial comparison practitioners made decisions on images of children, they were less accurate, less confident, and slower than with images of adults and less able to distinguish between signal and noise.

5.4.3 *Facial Comparison Practitioner Performance with Images of Children and Adults on Mated and Non-Mated Image Pairs*

Figure 20 presents the descriptive statistics for the accuracy of the Child and Adult groups based on whether images were mated or non-mated.

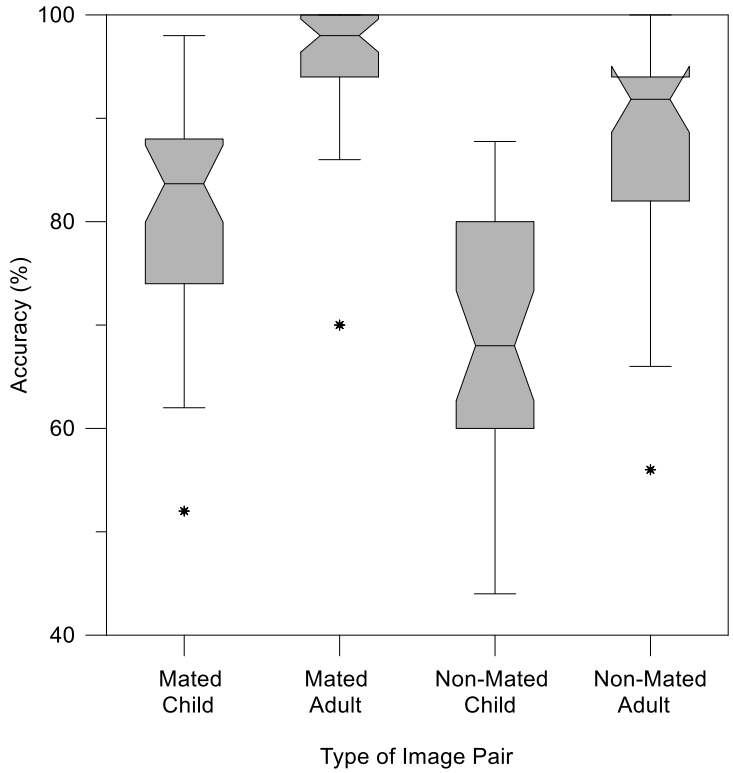


Figure 20. Accuracy for mated or non-mated Child and Adult groups.

The Friedman test showed that there was a statistically significant difference in accuracy based on the type of image pair presented to the participants, χ^2 (3) = 75.28, p < .001.

Post hoc analysis with Wilcoxon Signed Rank tests was conducted, which showed that facial comparison practitioners were significantly less accurate with mated images from the Child group (M = 80.71%, Mdn = 83.67%) compared to mated images from the Adult group (M = 95.99%, Mdn = 98%), z = -5.16, p < .001, r = -.62. Practitioners were also significantly less accurate with non-mated images from the Child group (M = 67.04%, Mdn = 68%) compared to non-mated images from the Adult group (M = 88.11%, Mdn = 91.84%), z = -5.16, p < .001, r = -.61. Finally, practitioners were significantly less accurate with non-mated images from the

Child group (*M* = 67.04%, *Mdn* = 68%) compared to mated images for the Child group (*M* = 80.71%, *Mdn* = 83.67%), *z* = -3.31, *p* = .001, *r* = -.40, and significantly less accurate with non-mated images from the Adult group (*M* = 88.11%, *Mdn* = 91.84%) compared to mated images from the Adult group (*M* = 95.99%, *Mdn* = 98%), *z* = -3.71, *p* < .001, *r* = -.45.

The outlier for both the mated Child and mated Adult groups was from the same facial comparison practitioner, however, they scored above average on non-mated Child and non-mated Adult image pairs. This suggests that this facial comparison practitioner (based on the vignette provided to practitioners) was more likely to err on the side of caution and say that image pairs were of different people.

Figure 21 presents the descriptive statistics for the confidence of the mated and non-mated Child and Adult groups.

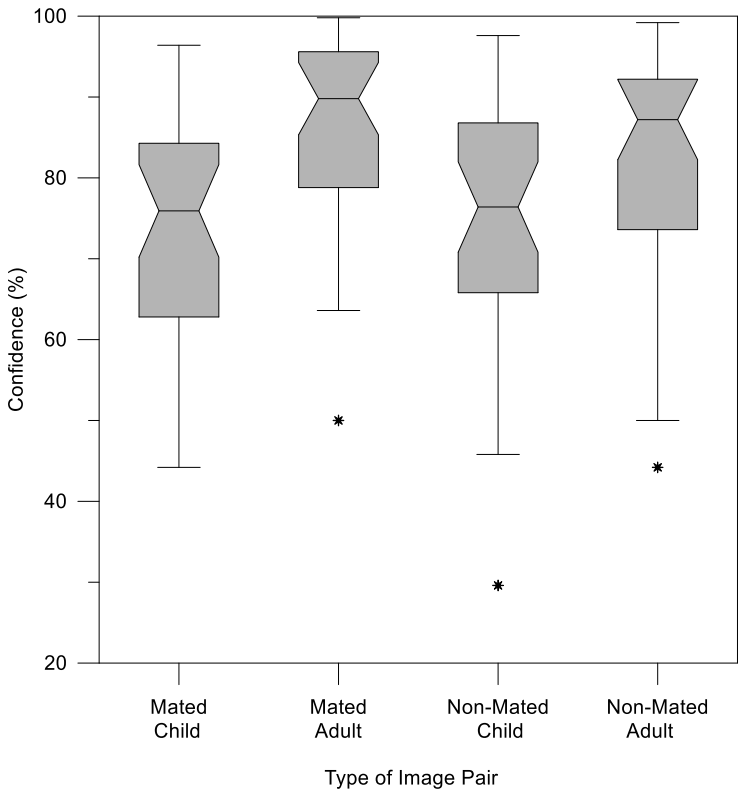


Figure 21. Confidence for mated and non-mated Child and Adult groups.

The Friedman test showed that there was a statistically significant difference in confidence based on the type of image pair presented to the participants, χ^2 (3) = 80.56, p < .001.

The results from the Wilcoxon Signed Rank test showed that there was a statistically significant difference between the mated Child group (*M* = 73.17%, *Mdn* = 75.92%) compared to the mated Adult group (*M* = 86.95%, *Mdn* = 89.80%), *z* = -5.09, *p* < .001, *r* = -.60, the non-mated Child group (*M* = 74.26%, *Mdn* = 76.40%) compared to the non-mated Adult group (*M* = 82.67%, *Mdn* = 87.20%), *z* = -5.09, *p* < .001, *r* = -.60, the mated Child group (*M* = 73.17%, *Mdn* = 75.92%) compared to the non-mated Child group[10] (*M* = 74.26%, *Mdn* = 76.40%), *z* = -2.60, *p* < .001, *r* = -.30, and the mated Adult group (*M* = 86.95%, *Mdn* = 89.80%) compared to the non-mated Adult group (*M* = 82.67%, *Mdn* = 87.20%), *z* = -3.76, *p* = .009, *r* = -.40. The outliers for the non-mated Child group and non-mated Adult group were from the same facial comparison practitioner.

Figure 22 presents the descriptive statistics for the response times of the mated and non-mated Child and Adult groups.

[10] Although it may seem unusual for the confidence of the mated Child group compared to the non-mated Child group to be statistically significant even though the median is very similar, the Wilcoxon Signed Rank test is a rank sum test, not a median test. Although rare, it is possible for groups to have different rank sums and have very similar or equal medians. Upon further inspection of the data it was identified that the significant difference was due to the different rank sums.

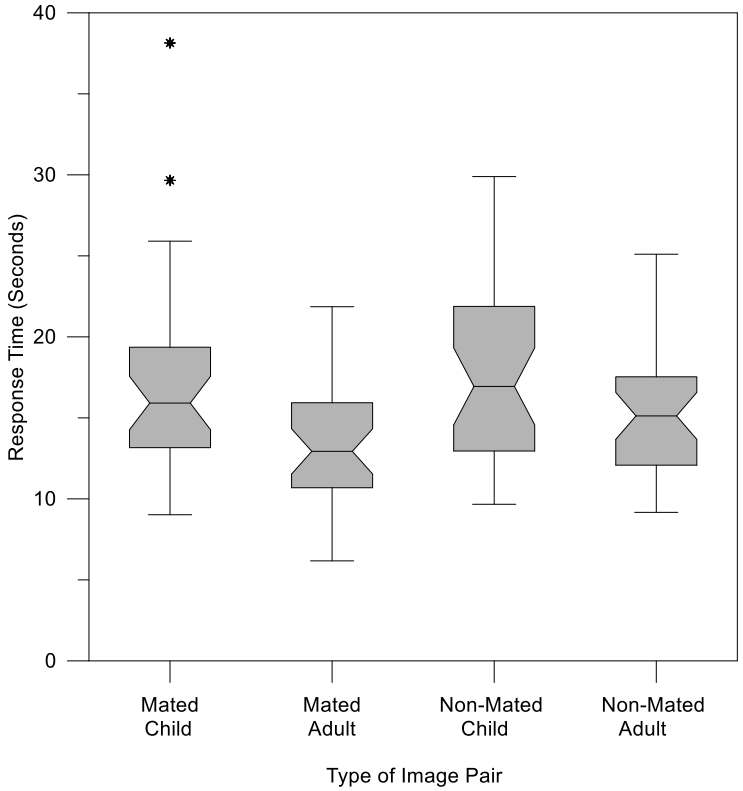


Figure 22. Response times for mated and non-mated Child and Adult groups.

The Friedman test was also performed on the response time data, which showed a statistically significant difference in response time based on the type of image pair, χ^2 (3) = 43.35, $p < .001$.

The Wilcoxon Signed Rank tests revealed statistically significant differences in response times between the mated Child group (*M* = 17.04 seconds, *Mdn* = 15.92 seconds) and the mated Adult group (*M* = 13.06 seconds, *Mdn* = 12.93 seconds), *z* = -4.77, *p* < .001, *r* = -.60, the non-mated Child group (*M* = 17.90 seconds, *Mdn* = 16.94 seconds) and the non-mated Adult group (*M* = 15.14 seconds, *Mdn* = 15.12 seconds), *z* = -3.98, *p* < .001, *r* = -.50, and also the mated Adult group (*M* = 13.06 seconds, *Mdn* = 12.93 seconds) and the non-mated Adult group (*M* = 15.14 seconds, *Mdn* = 15.12 seconds), *z* = -3.55, *p* < .001, *r* = -.40. However, there was not a significant difference between the mated Child group (*M* = 17.04 seconds, *Mdn* = 15.92 seconds) and the non-mated Child group (*M* = 17.90 seconds, *Mdn* = 16.94 seconds), *z* = -1.38, *p* = .169.

In summary, these results support the hypotheses that facial comparisons with non-mated images of children were the most difficult group resulting in the lowest accuracy and confidence, and highest response times.

5.5 *Discussion*

The aim of this study was to gather empirical data from facial comparison practitioners to determine facial comparison performance with images of children and adults to feed into Requirement 1, as identified during Study 1 (Chapter 2) of this thesis. The following two sections (Section 5.5.1 and 5.5.2) discuss the results based on the two research questions this study aimed to answer.

5.5.1 *Facial Comparison Practitioner Performance with Images of Children and Adults*

Study 1 found that practitioners conduct facial comparisons on images of both children and adults, yet were unaware, empirically, of the extent to which performance differed based on whether image pairs contained a child or an adult. This study supports the anecdotal claims provided by agencies in Study 1, that facial comparisons with images of children are more difficult than facial comparisons with images of adults. Using the medians (and means), the overall accuracy in the current study was 18% lower for images of children. It also showed that facial comparison practitioners were less confident and took longer to make their decisions with images of children compared to images of adults.

The overall performance results were in line with past research (Lanitis, 2008; White, Dunn, et al., 2015; Zeng et al., 2012) that facial comparisons with images of children are more difficult than with images of adults. In this study, accuracy and confidence of practitioners were lower and response times were higher with images of children. This trend is consistent with past research despite a range of differences in methodology between studies including with participants, images, time allocations, and task paradigm. The facial comparison practitioners in the current study also outperformed in terms of accuracy compared to participants in these previous studies. However, the overall accuracy with images of adults was on par with results from Calic (2012) of 93%, a study which was also conducted with images of adults and practitioners as participants (as discussed in Section 1.5.2). The findings suggest that the

amount of growth and change occurring in childhood compared to adulthood does have an impact on practitioner performance. Agencies should keep this in mind when developing practices, policies, procedures, and training for facial comparison practitioners.

The Department of Foreign Affairs and Trade is an example of an agency that has put policies in place to help deal with differences in performance with images of children compared to images of adults. These policies limit the validity of a child's passport to only five years and an adult's passport to 10 years. Unfortunately for other applications, such as those in law enforcement, the luxury of being able to implement such policies and reduce the age variation is not afforded due to the nature of the work. In such circumstances, it is important for agencies to at least be aware of the extent of the difference in performance when practitioners are conducting facial comparisons with images of children compared to images of adults so they can manage this in a way that suits their particular needs.

In general, practitioners were liberal at deciding both images of children and adults were the 'same' person. This liberal bias may be due to facial comparison practitioners being exposed to more images of the 'same' person than 'different' people during their regular duties.

5.5.2 *Facial Comparison Practitioner Performance with Images of Children and Adults on Mated and Non-Mated Image Pairs*

Performance measures of accuracy, confidence, and response time were also used to analyse results by image pair type (i.e., mated and non-mated) on each group (i.e., Child and Adult). Accuracy was higher in both the mated Child group and mated Adult group compared to their non-mated counterparts. Past studies that have examined mated and non-mated performance have sometimes found that accuracy is higher for non-mated image pairs than mated images pairs (White, Kemp, Jenkins, Matheson, et al., 2014; Ferguson, 2015). However, this is likely due to the limited images available to select from in past research (e.g., White, Kemp, Jenkins, Matheson, et al., 2014). Others have selected pairs using a randomisation process which is known to result in higher accuracy than other selection methods, likely due to non-mated pairs not being selected based on any similarity (Calic, 2012). Because of this, results from these studies may not reflect operational performance. In the current study, a facial recognition algorithm was used and an image pair of similar quality and appearance was

selected, making the task more difficult compared to much of the past research, but also more realistic to operational applications.

In the current study, practitioners performed the poorest with non-mated image pairs of children than any other group. This may be due to children having less discriminating facial features, therefore making them harder to tell apart (Wilkinson, 2012), as hypothesised at the beginning of the chapter. As practitioners performed most poorly with non-mated pairs, agencies may consider further training with images of children that focuses on distinguishing children from each other. For example, training with images of adults focuses on facial features that stay stable over time to determine if similarities and differences between images can be explained. In childhood, it could be that different facial features or strategies would be more appropriate than what is typically taught with images of adults. This will be discussed further in Study 3B (Section 7.5.3) and the final chapter (Section 9.6). In addition, practitioner confidence was higher than accuracy with non-mated pairs of children. This indicates that practitioners were also overconfident in their ability to make accurate decisions with non-mated pairs of children.

Confidence and response times were relatively consistent with images of children regardless of pair type and were significantly different to their adult counterparts. This once again highlights the difficulty of facial comparisons with images of children compared to images of adults and supports the accuracy findings.

This study also highlights a need to examine at a finer level what ages in childhood are most difficult. For example, since babies have the least discriminating facial features, do practitioners perform poorest on images of babies (i.e., age 0)? Does performance slowly improve as age increases and more distinguishing facial features develop?

5.6 *Summary*

This study was conducted at a high level to determine the difference in practitioner performance with images of children compared to images of adults. A finer level study focusing on images of children at all ages in childhood is necessary to determine how performance varies over childhood with age variations that may be expected in operational applications. This was identified in Study 1 and will be addressed as Requirement 2

(determining facial comparison performance on images of children at different ages and age variations) for algorithms (Study 3A) and facial comparison practitioners (Study 3B). The next chapter presents Study 3A.

Chapter 6.

Study 3A: Facial Comparison Performance with Images of Children at Different Ages and Age Variations — Algorithm Study

6.1 *Introduction*

The aim of Study 2A was to provide algorithm performance data for Requirement 1 (determining facial comparison performance with images of children and adults). Although it was clear that there are algorithm performance deficiencies with images of children compared to adults, a more granular examination of performance was required across childhood (0–17 years) and over different age variations (0–10 years). This was identified as Requirement 2 (determining facial comparison performance with images of children at different ages and age variations).

This chapter reports on Study 3A, which aimed to address Requirement 2 by testing the performance of five state-of-the-art algorithms and one open source algorithm. Results provided from this study are of value to both agencies and algorithm vendors. Similar to the previous studies, research in this space is scarce, particularly with controlled operational images of children.

As was evident during Study 1, many agencies currently use algorithms for facial comparisons with images of children. However, no large-scale one-to-one study has been conducted to evaluate algorithm performance with images of children at each individual age and across individual age variations. Some agencies require algorithms to conduct facial comparisons with images of children in conjunction with a practitioner who makes the final decision, such as in document processing applications. In other applications, such as access control, the process may be fully automated. For example, SmartGate systems are installed in Australian airports which allow children as young as 10 years of age to use the system when arriving back in Australia (DIBP, n.d.-a) and any age when departing (DIBP, n.d.-c). In law enforcement agencies, algorithms may be used as an investigative tool. For example, to compare an image of a missing child to images on escort service websites (Sadwick, 2017).

As different agencies conduct facial comparisons with images of children at different ages and age variations, it is necessary to conduct research on a large-scale database of controlled images to determine how performance changes throughout childhood. Not only does a controlled database ensure the age-related variables are tested in isolation as much as possible from other variables that can impact on performance, it also provides an upper bound level of algorithm performance to use as a guide for operational applications. Vendors during Study 1 mentioned that they were not sure how their algorithms performed based on the age of children in images. However, one vendor did expect that the older the children in images, the better the algorithm performance would be. This anecdote highlights that more research is warranted.

The FRVT 2013 (Grother & Ngan, 2014) conducted one-to-many testing with images of children. The age variable was tested by dividing the dataset of visa images (N = 19,972) into seven age groups: baby (0–3), kid (3–8), pre (8–13), teen (13–19), young (19–30), parents (30–55), and older (55–101). The results show that performance does vary based on the age of the children in images. In their report, the most accurate algorithm produced false negative identification rates[11] of 0.7 for babies, 0.4 for kids, 0.29 for preteens, and 0.057 for teens at a

[11] False negative identification rate is a term often used when conducting one-to-many (identification) tasks. It refers to the proportion of identification attempts in which a user is enrolled in a database but their identity is not returned by the system (Jain, Ross, & Nandakumar, 2011).

fixed false positive identification rate[12] of 0.005 (203,082 non-mated searches in total). This shows that identification is progressively easier with advancing age across childhood and this was consistent for each algorithm tested. This finding is similar to previous research (Ferguson, 2015; Mahalingam & Kambhamettu, 2012). Younger children were simultaneously more difficult to recognise as themselves and also more difficult to tell apart from others. Identification of the baby group (0–3 years) failed more often than it succeeded. The report highlights that their sample of 57 subjects in this group was small, yet their error rates were so high that the finding was still significant. The FRVT 2013 (Grother & Ngan, 2014) highlighted that longitudinal ageing affects were not quantifiable as age variations were less than 4 years and only around 1.6 years for the 0–3 year age group. However, the impact of ageing presented in the 0–3 year age group was believed to be a result of the considerable amount of craniofacial growth occurring.

Of strong practical interest is the impact age variation has on algorithm performance. Several studies have examined the impact of age variation (Guo, Mu, & Ricanek, 2010; Ling, Soatto, Ramanathan, & Jacobs, 2010; Lui et al., 2009; Ramanathan & Chellappa, 2006a; Ricanek & Tesafaye, 2006). All studies in a meta-analysis by Lui et al. (2009) consistently found that as the age variation between images increased, the accuracy of facial recognition algorithms decreased. This is consistent with anecdotes from vendors in Study 1. Although these studies employed different algorithms, image sets, and methodologies, studies assessing the age variations between images produced a similar pattern of results. However, Grother et al. (2011) found counterintuitive results for many algorithms when the age variation was between 30 months (2.5 years) and around 71 months (5.92 years) with false non-match rates decreasing when a set false match rate of 0.001 was used. The researchers suggested this may be the result of hidden factors within this range (not directly related to face ageing) that makes them easier to recognise. The researchers acknowledge that due to confounding factors such as subject age, further research is necessary. Ling et al. (2010) evaluated discriminative methods with a dataset of passport images containing 1824 mated and 9492 non-mated image pairs. They found that performance based on age variation did not change after four years

[12] False positive identification rate is also a term often used when conducting one-to-many tasks. It refers to the proportion of identification attempts in which a user is not enrolled in the database but an identity is returned by the system (Jain et al., 2011).

(and up to 10 years). The majority of the studies discussed so far did not test algorithm performance exclusively on images of children nor did they use state-of-the-art algorithms.

Ferguson (2015) however did conduct one-to-one algorithm testing with images of children using an NEC facial recognition algorithm. The research focused on White European children aged 0–15 years, but also included 80 images of adults (above 15 years). The dataset predominantly contained uncontrolled images and were grouped into four age groups: 0–4, 5–9, 10–14, and 15+ years. However, at a false match rate of 0.001, the overall false non-match rate was 0.73. Ferguson found that similarity scores were higher for mated image pairs with no age variation between the images (i.e., the age of the person in the images were the same). As the age variation increased, the similarity scores for mated and non-mated pairs decreased. However, as the age of the person in an image increased there was more stability in similarity scores across age variations in mated image pairs. Ferguson also found that the 0–4 year age group had considerable overlap in mated and non-mated image pair scores. Ferguson hypothesised that this may either be due to children in this age group yet to develop individuating features or due to the age variation spanning up to five years apart. Non-mated pairs in the 0–4 year age group showed slightly higher similarity scores than older age groups. Ferguson did however caution that there was a substantial lack of images in the two youngest age groups to make any concrete conclusions and that the 5–9 year age group contained high quality images which may have skewed the data. As a result, Ferguson recommended that more research be conducted on a larger and more controlled database of children's images.

The lack of research in testing algorithm performance with images of children is alarming given that agencies have purchased and implemented such biometric systems into their procedures. Furthermore, there is very little information that vendors can provide to agencies due to their lack of access to large databases of images for testing purposes, particularly with images of children. This study aimed to address these limitations.

6.2 *Research Questions*

The aim of the current study was to test algorithm performance with images of children at a finite level to provide information that feeds into Requirement 2. This was achieved via three research questions.

Question 1. To what extent is algorithm performance impacted by the age of children (0–17 years) in images and age variations ranging from 0–10 years?

Question 2. To what extent does algorithm performance vary by age and age variation based on the type of image pair presented (i.e., mated or non-mated)?

Question 3. To what extent does performance vary at a set false match rate of 0.001 based on images of adults (a typical operating point used for evaluating facial recognition algorithms) as well as at each age and age variation?

As there are less discriminating facial features at younger ages in childhood, it was hypothesised that the younger the child, the poorer the algorithm performance would be. Due to considerable facial growth occurring in childhood (Fiek & Glover, 1998), it was also hypothesised that as the age variation increased, algorithm performance would decrease.

In regards to performance based on image pair type, as found by Grother and Ngan (2014), it was hypothesised that the younger the children were in images, the more difficult they would be to recognise as themselves, resulting in higher false non-match rates. Younger children would also be more difficult to discriminate from others, resulting in higher false match rates. The considerable amount of facial growth occurring over time in childhood was also expected to increase the false non-match rate as the age variation increased.

6.3 *Methodology*

This section contains the methodology used to test algorithm performance across age and age variation.

6.3.1 *Participants*

The same five state-of-the-art facial recognition algorithm vendors that participated in Study 2B participated in this study. OpenBR (Klontz et al., 2013) was also tested to provide results from a free and publicly available algorithm.

6.3.2 *Materials*

Materials used in this study included the controlled operational facial image database (see Section 3.1.1), the Biometrics High Performance Computing Cluster (see Section 3.1.4), five state-of-the-art facial recognition algorithms (see Section 4.3.2), one open source algorithm (see Section 4.3.2), and a plotting program developed in-house (see Section 4.3.3.1). As each of these has been explained in previous chapters they will not be described in detail here.

6.3.3 *Image Pair Selection*

The total number of images used in this study was 4,652,868. The number of image pairs that each algorithm conducted matching on varied slightly due to different failure-to-enrol rates returned by each algorithm as presented in Appendix E.

For each algorithm, the false non-match rate was obtained by matching each mated pair of a child in the database. Similarly, the false match rate was obtained by matching each of the youngest age images in a mated pair, to 30 randomly selected non-mated images that were of a person the same age as the second image in that mated pair (i.e., each person's earliest mated image was matched with 30 randomly selected non-mated images based on gender and age of the mated person's second, alternate image). Appendix G contains the number of mated pairs used in this study by age and age variation.

6.3.4 *Procedure*

The procedure of this study was similar to Study 2A (see Section 4.3.5) but images were filed by age and age variation in months and then grouped back up into appropriate years. This was to ensure age variation data was correctly categorised. In addition, the results were sorted based on the age of the youngest child in a pair and the age variation between images rather than based on whether the image pairs belonged to the Child or Adult group. This resulted in 198 different categories as originally shown in Figure 10 (ages 0–17 = 18 categories and age variations 0–10 = 11 categories, 18 x 11 = 198 categories).

6.4 *Results*

The results are divided into three sections based on the three questions designed to provide data for Requirement 2. DET and cumulative probability plots showing performance at a high

level are provided for each algorithm in the results section. Algorithm E is used throughout the results section to demonstrate performance at a more granular level as it is the best performing algorithm. The results for the remaining algorithms are presented in the appropriate appendices. The colours for the lines in these plots were selected to slightly change based on age or age variation and spread out as much as possible based on the RBG values. This was to make it easier to demonstrate the pattern in performance across these age-related variables.

6.4.1 *Algorithm Performance with Images of Children at Different Ages and Age Variations*

Figure 23 shows the DET plots that display how performance changed based on the age of the youngest child in an image pair for each of the six algorithms (i.e., all age variations grouped by age).

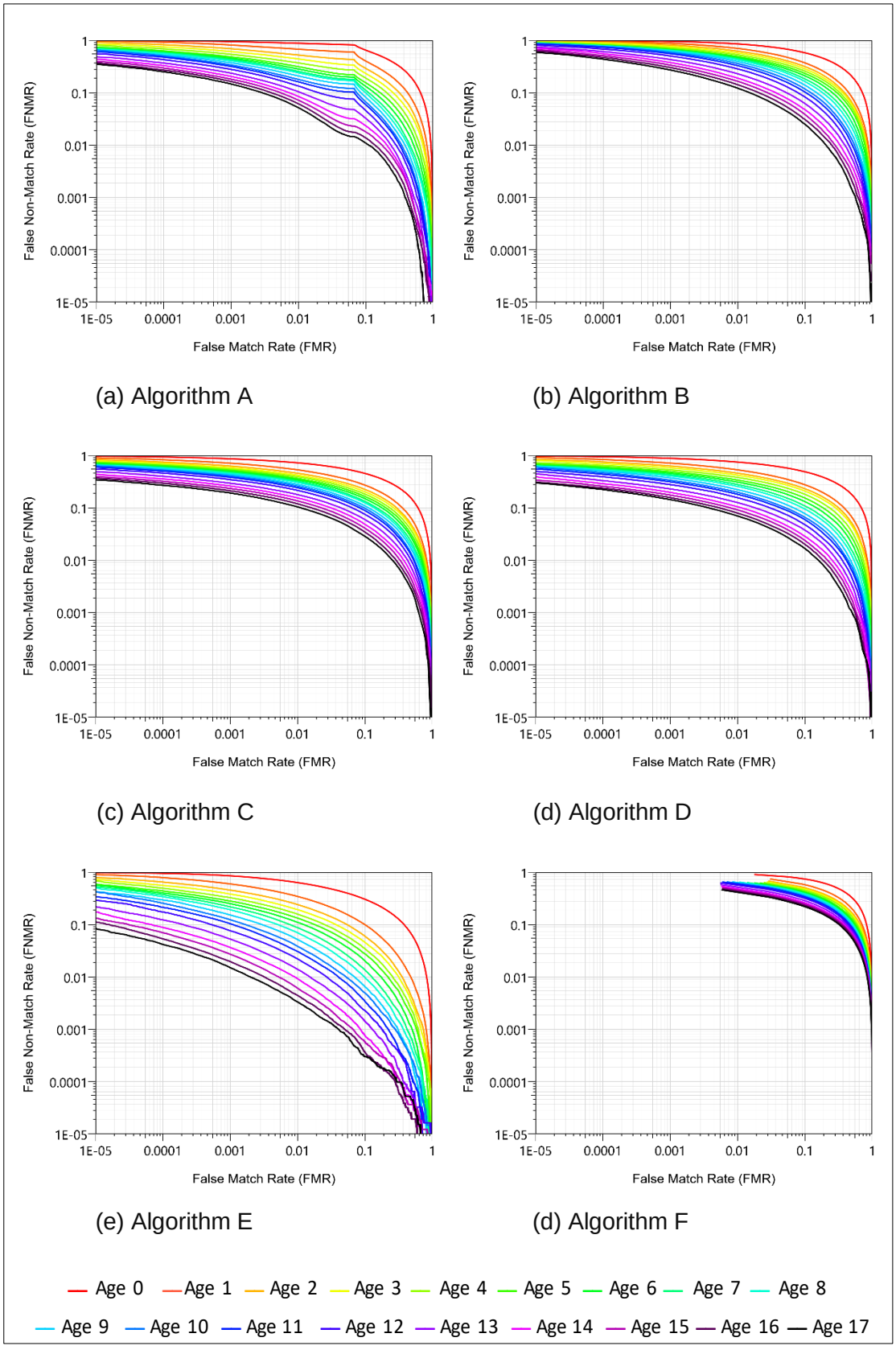


Figure 23. DETs for each algorithm based on age.

The plots show that performance consistently increased with advancing age. This performance increase was consistent within algorithms with a slightly larger change in performance obtained between age 0 and age 1. In terms of algorithm performance, Algorithm E appeared to offer the best performance across all age groups and a larger increase in performance was also observed for this algorithm as children got older, compared to any other algorithm. OpenBR (Algorithm F) appeared to be the poorest performing with the least change in performance across age.

Figure 24 and Figure 25 show DETs for Algorithm E with each plot displaying how performance changed based on the age of the youngest child in a pair ranging from 0–17 years at a set age variation (0–10 years). Appendix H presents the DETs for the five remaining algorithms.

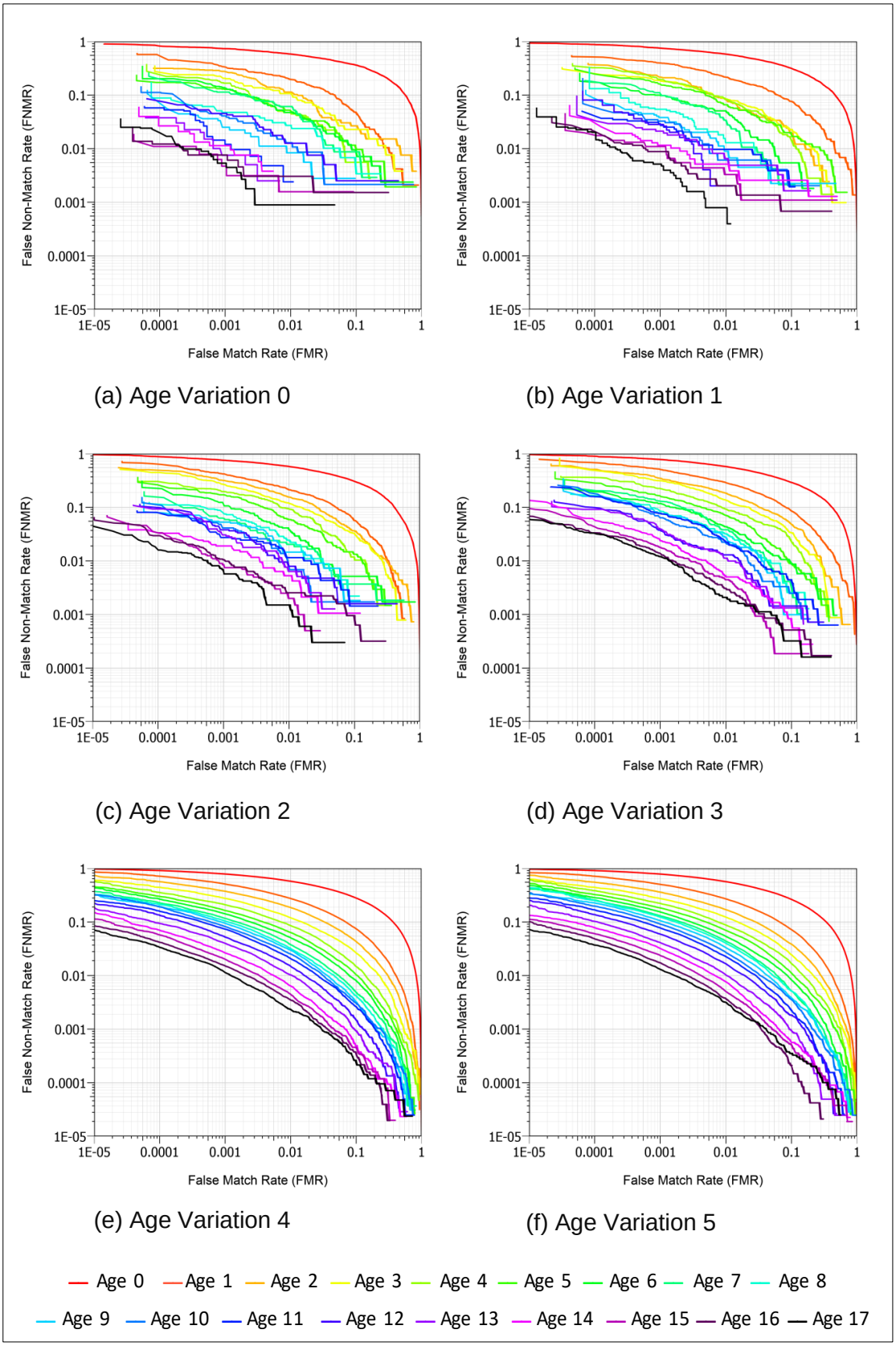
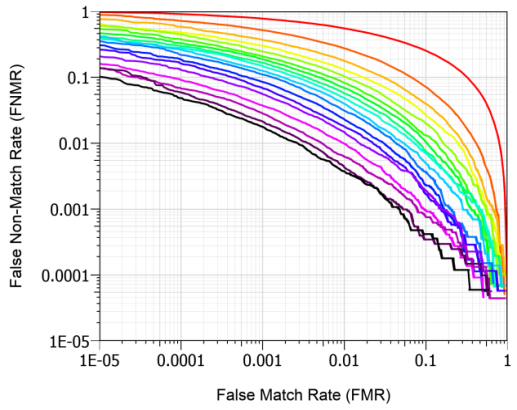
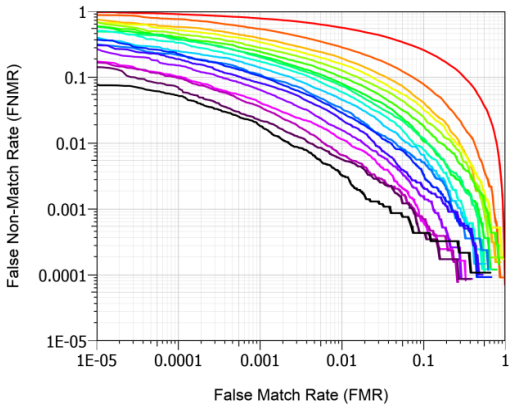


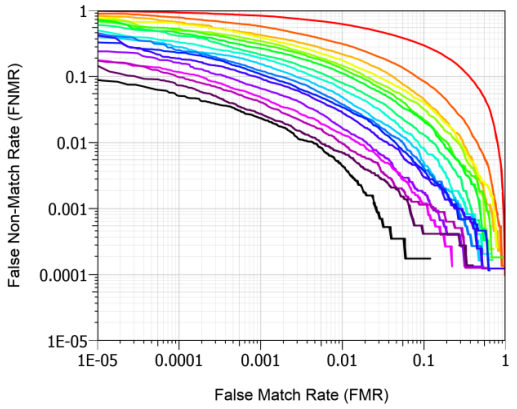
Figure 24. DETs for Algorithm E displaying how age impacts on performance for age variations spanning 0–5 years.



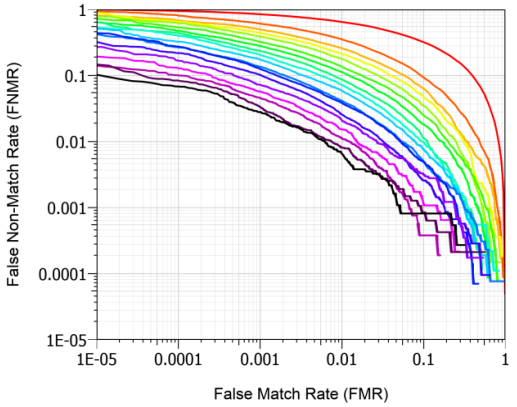
(g) Age Variation 6



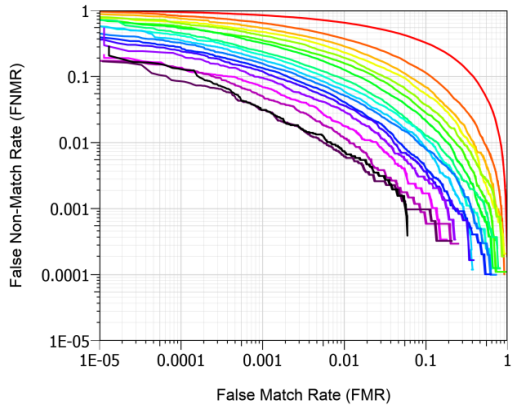
(h) Age Variation 7



(i) Age Variation 8



(j) Age Variation 9



(k) Age Variation 10

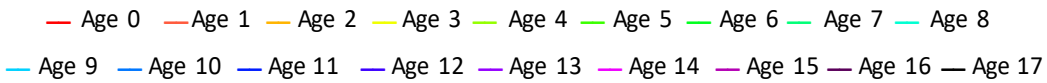


Figure 25. DETs for Algorithm E displaying how age impacts on performance for age variations spanning 6–10 years.

The DETs show that lines on each plot are spread out regardless of age variation, indicating that the youngest age of a child in an image pair impacted on performance regardless of age variation. The DETs show an almost identical pattern in performance across age variation. That is, image pairs were always poorest performing when the youngest age in a pair was age 0 and then progressively improved with advancing age. Similar to the results presented in Figure 23, the difference in performance between age 0 and age 1 was considerably larger than any other consecutive ages. The DETs also show that as the age variation increased, performance at each age worsened.

Figure 26 shows the DET plots that display how performance changed based on age variation for each of the six algorithms (i.e., all ages grouped by age variation).

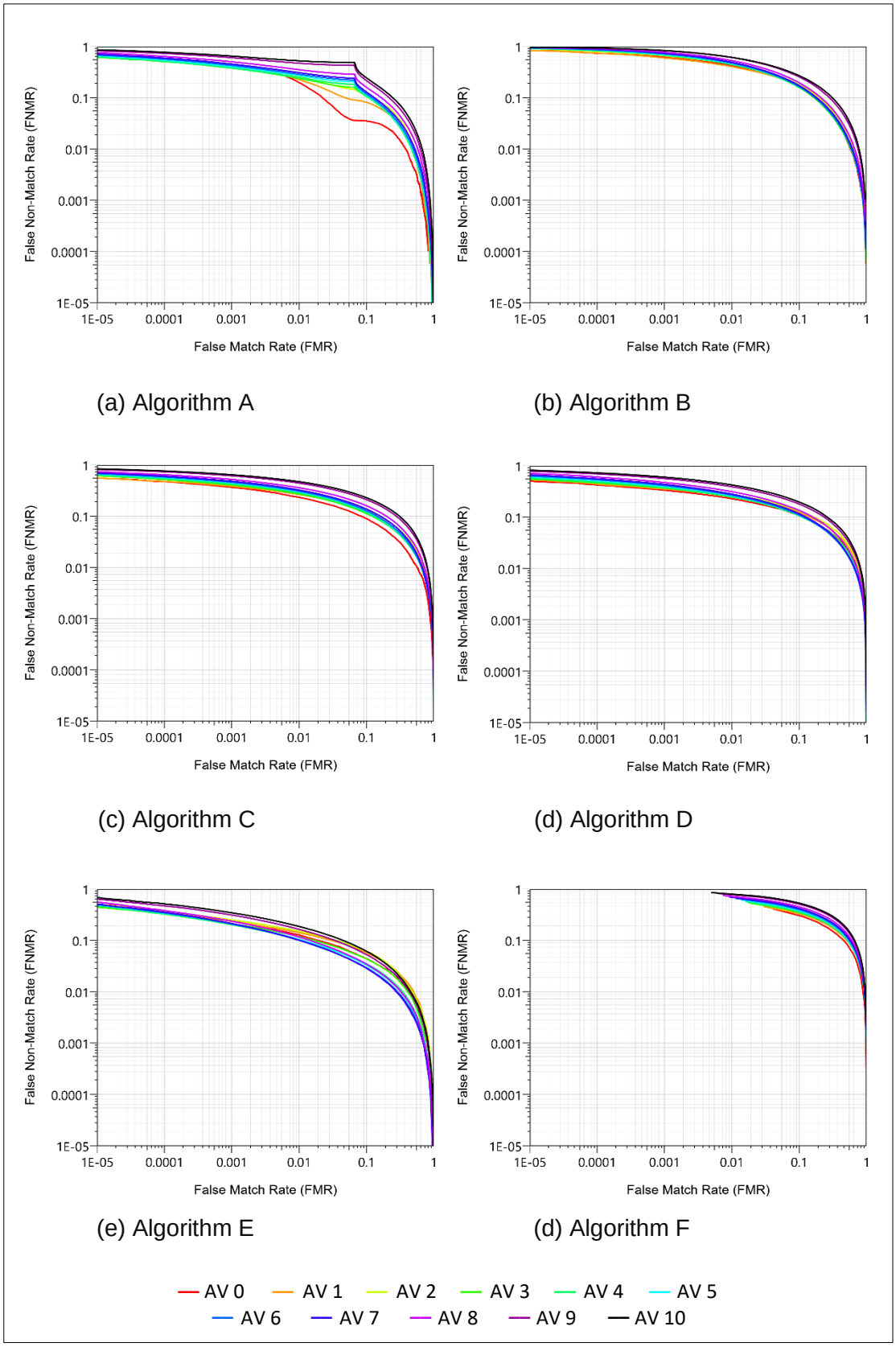


Figure 26. DETs for each algorithm based on age variation (AV = age variation in years).

The plots show that for all algorithms, each age variation had relatively similar performance to each other. This suggests that age variation between images did not impact considerably on algorithm performance compared to the age of the child in the images as demonstrated in Figure 23. In these plots, it is hard to determine if performance consistently declined as the age variation increased. Algorithm A, however, shows that performance appeared to be better as the age variation between images decreased.

Figure 27 to Figure 29 show DET plots for Algorithm E that display how performance changed based on age variations between images ranging from 0–10 years for each individual age in childhood (0–17 years). Appendix I contains the results for the remaining five algorithms.

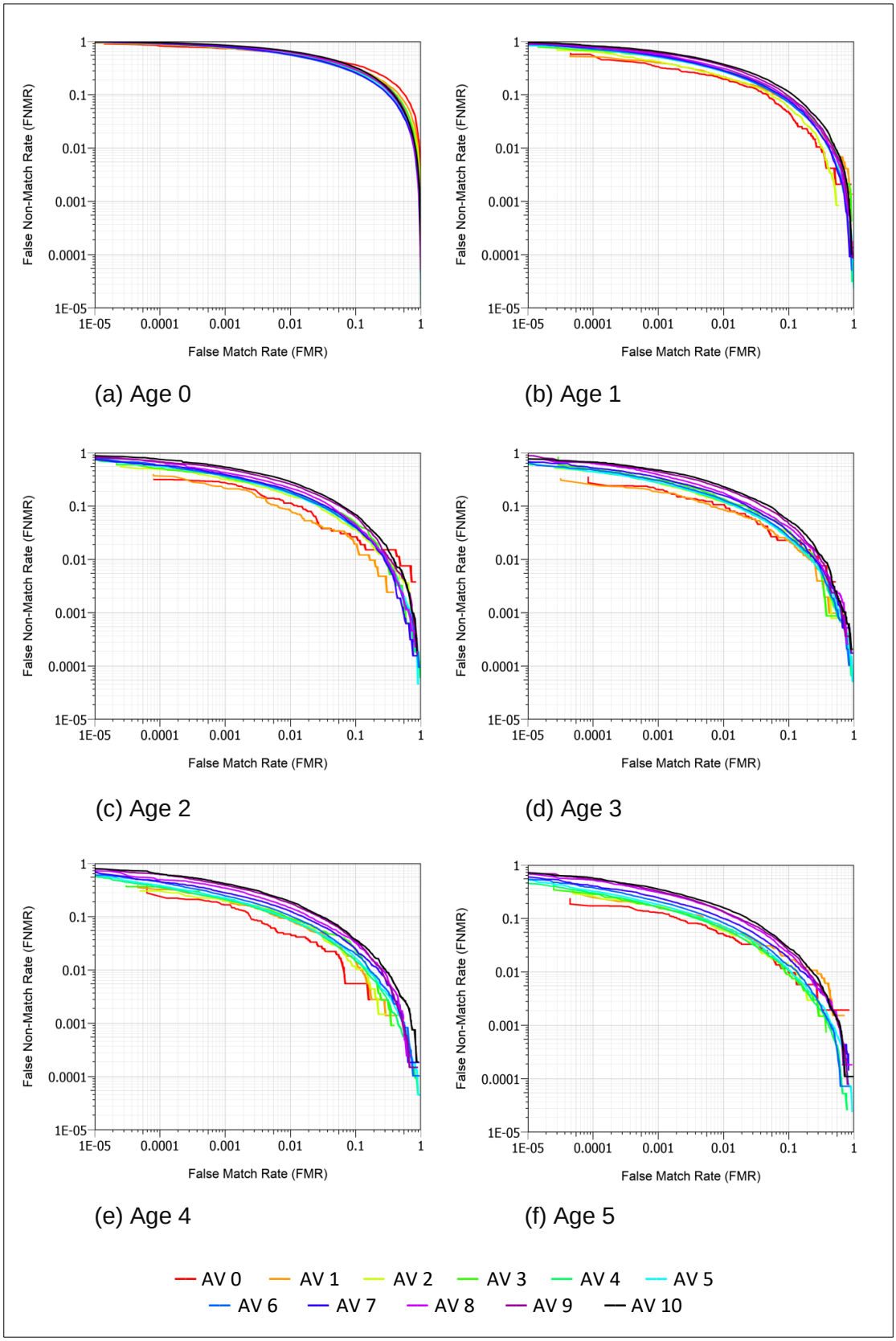


Figure 27. DETs for Algorithm E displaying how age variations impact on performance for ages 0–5 (AV = age variation in years).

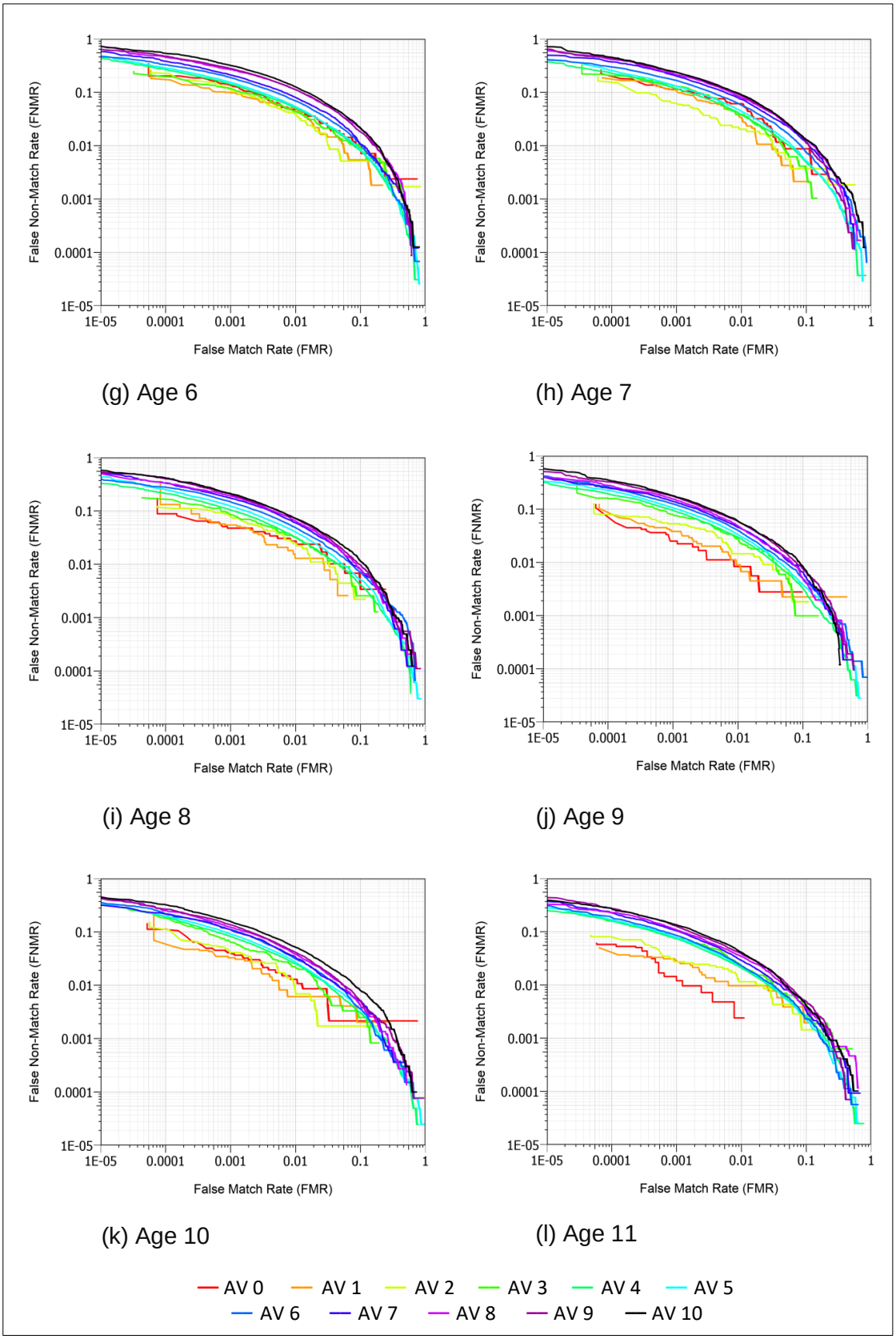


Figure 28. DETs for Algorithm E displaying how age variations impact on performance for ages 6–11 (AV = age variation in years).

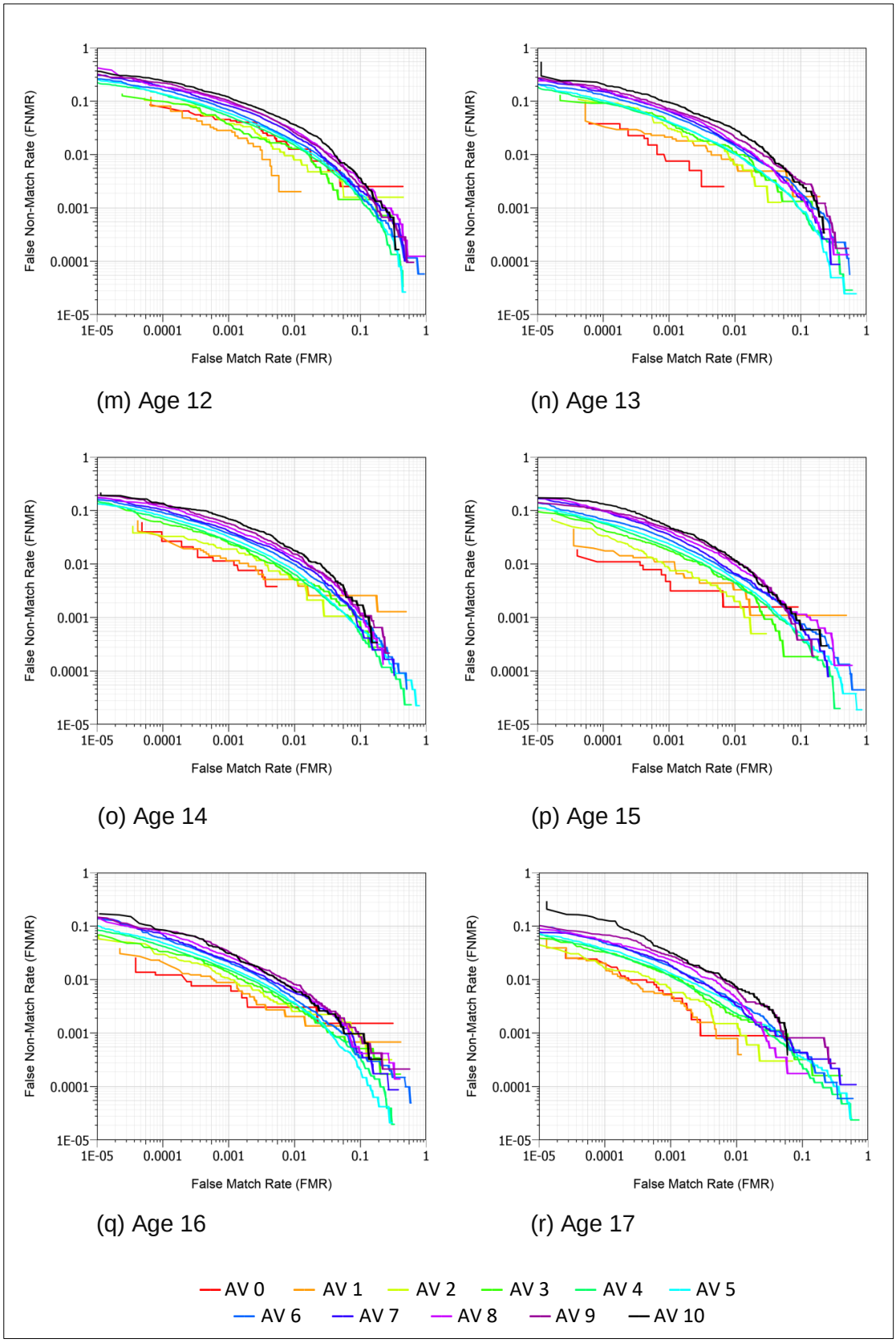


Figure 29. DETs for Algorithm E displaying how age variation impacts on performance for ages 12–17 (AV = age variation in years).

The plots show that performance for Algorithm E increased as the youngest age of the child in a pair increased. At age 0, performance across age variations ranging from 0–10 years was indistinguishable from each other and each age variation also performed poorer than the same age variation examined for ages 1–17. As the age of the child in the youngest image increased, performance across age variation spread out more, indicating that there was more variance in performance across age variation as age increased. The DETs also show that shorter age variations, such as 0–2 years, were typically best performing regardless of age, likely due to less facial changes occurring over shorter periods.

6.4.2 *Algorithm Performance with Images of Children at Different Ages and Age Variations on Mated and Non-Mated Image Pairs*

Cumulative probability plots were used to better distinguish performance over mated and non-mated image pairs (i.e., false non-match rate and false match rate respectively). Figure 30 shows the cumulative probability plots that display how performance changed based on the age of the youngest child in an image pair for each of the six algorithms (i.e., all age variation data grouped by age).

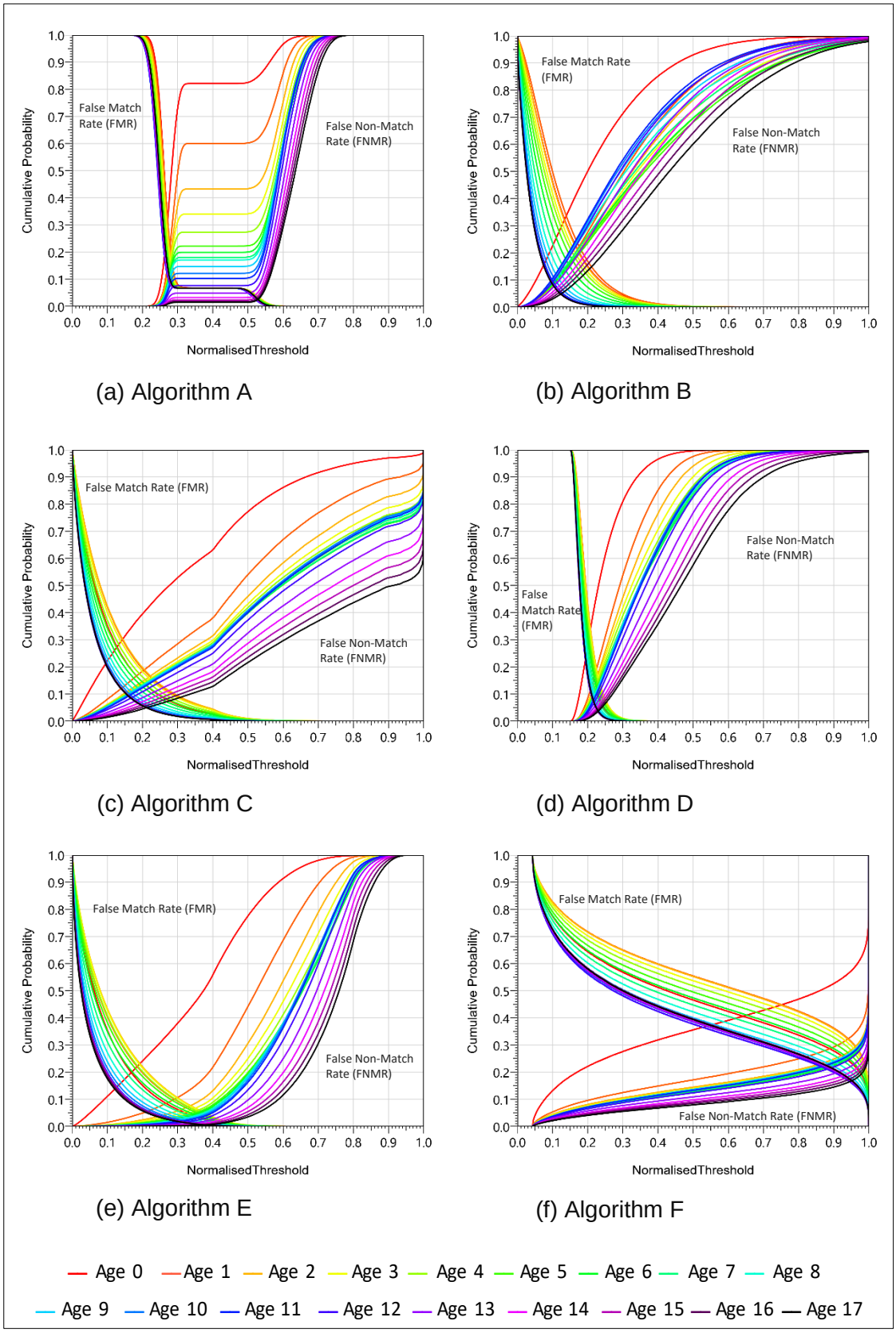


Figure 30. Cumulative probability plots for each algorithm based on age.

The plots show that there was less difference in performance for each algorithm with non-mated pairs compared to mated pairs across age. That is, there was greater variation between the different ages in terms of false non-match rate performance with the false match rate being minimally impacted by age. Although there was little variation in performance between different ages for non-mated pairs, the false match rate performance indicated that a child at an older age was generally easier to distinguish from other children. However, the false match rate at age 0 was lower than some other ages on Algorithm C, E, and F. This may perhaps be due to differences in how these algorithms compared faces or due to a much larger set of data at age 0 than other ages (as indicated in Appendix G), producing a lower false match rate than some other age groups. The plots also show that for all algorithms, the false non-match rate was remarkably higher with images at age 0 indicating that babies were harder to recognise than any other ages, and the large overlaps between the false match rate and the false non-match rate indicate that all algorithms performed poorest with images at age 0.

Figure 31 and Figure 32 show the performance of Algorithm E at each age variation and how performance changed based on the age of the child in the youngest image spanning 0–17 years of age. Appendix J contains the results for the remaining five algorithms.

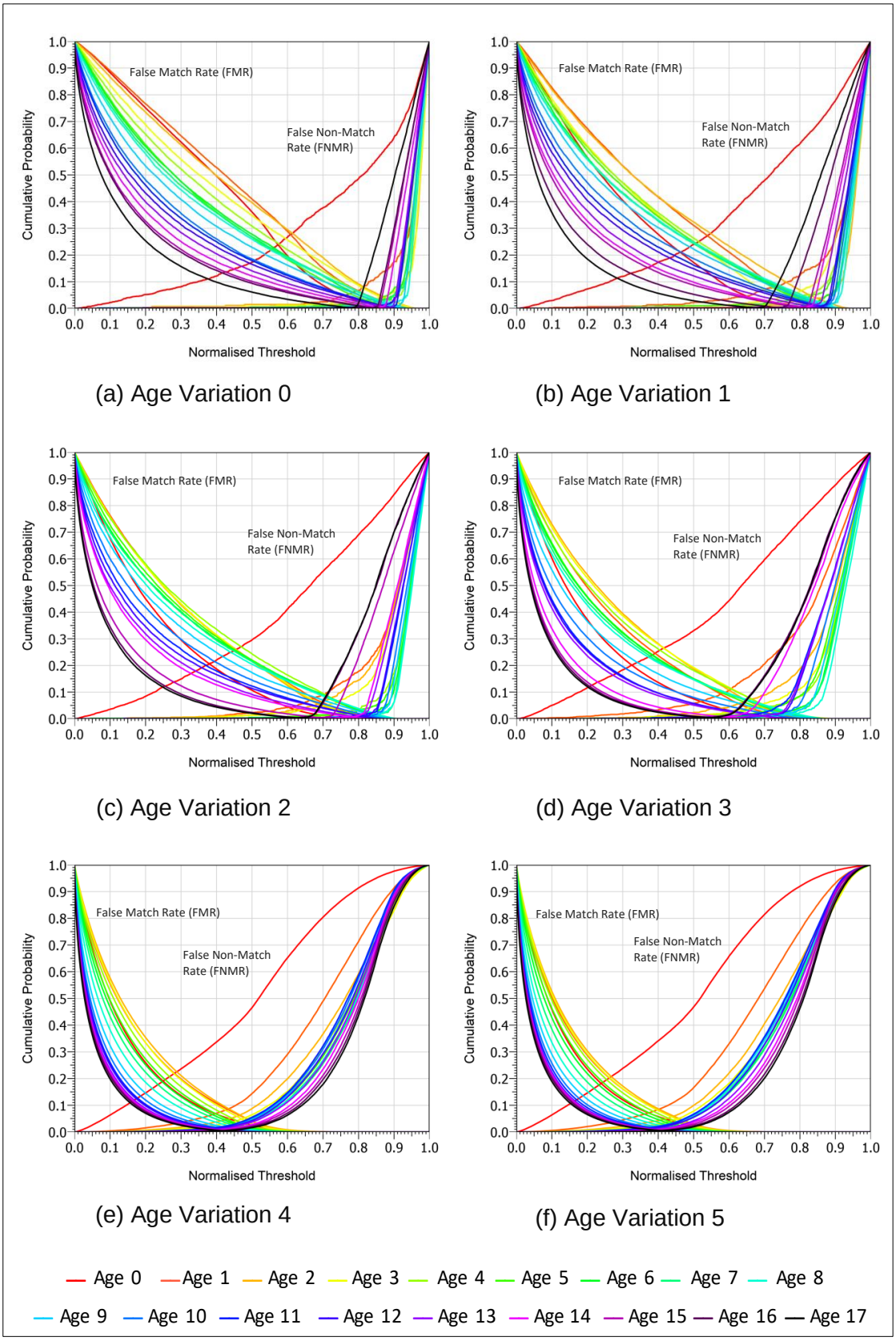


Figure 31. Cumulative probability plots for Algorithm E displaying how age impacts on performance for age variations spanning 0–5 years.

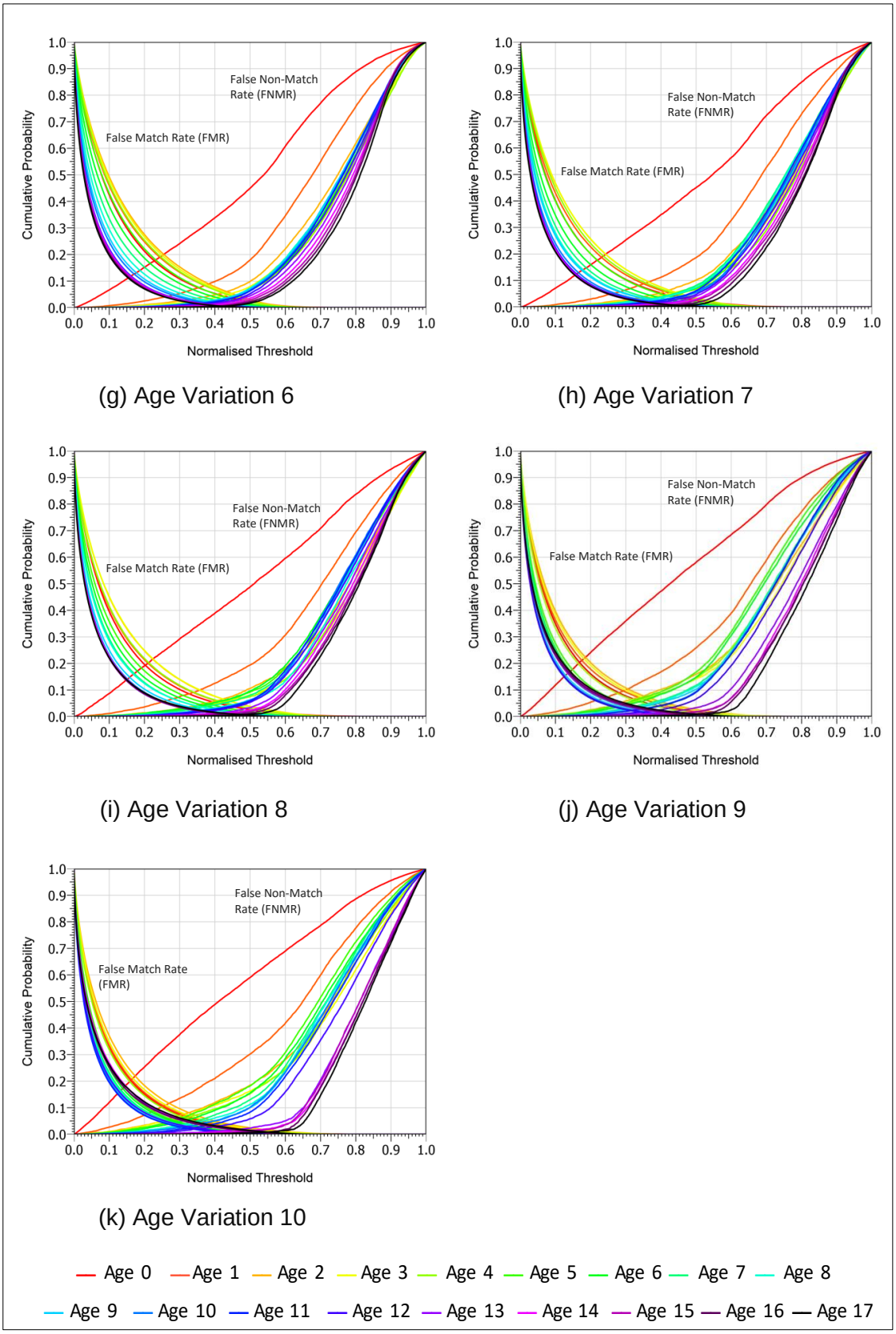


Figure 32. Cumulative probability plots for Algorithm E displaying how age impacts on performance for age variations spanning 6–10 years.

The false non-match rate curves clearly show that for each age variation, Algorithm E performed considerably worse at age 0 than any other age with mated pairs. That means a child was harder to be recognised when the youngest age image in a pair was age 0. Performance of Algorithm E was better with mated pairs at shorter age variations (i.e., 0–3 years) than larger age variations.

The false match rate curves show that Algorithm E was more accurate with non-mated pairs at a 10 year age variation than at a 0 year age variation indicating that it was harder to tell two children apart if they were of a similar age. There was also a considerable difference in performance with non-mated pairs across different ages when the 0 year age variation was compared to 10 years, indicating that there was less difference and more stability in performance at larger age variations.

Figure 33 shows the cumulative probability plots that display how performance changed based on age variation for each of the six algorithms (i.e., all ages grouped by age variation).

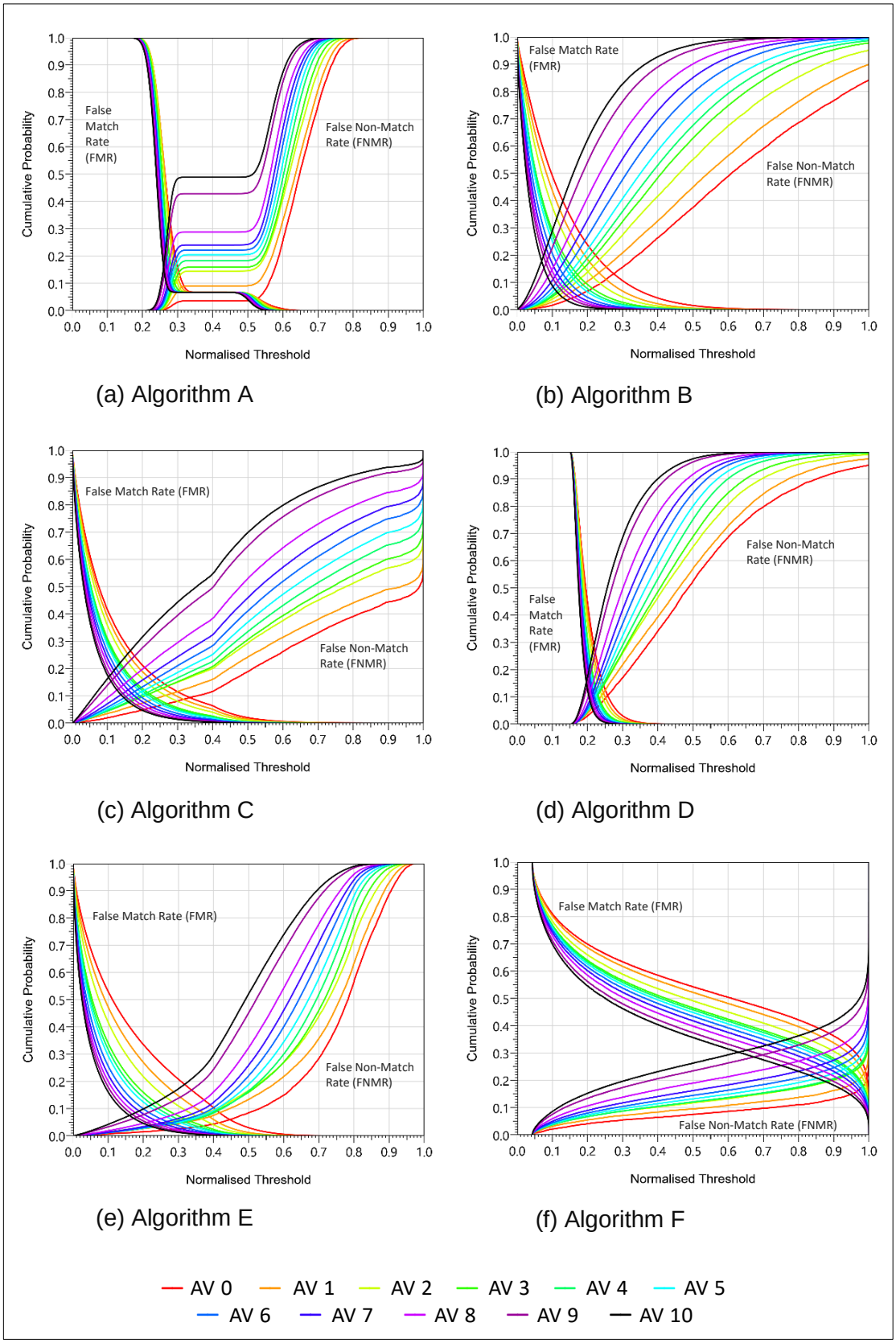


Figure 33. Cumulative probability plots for each algorithm based on age variation (AV = age variation in years).

The plots show that there was a larger difference in performance across age variation with mated pairs compared to non-mated pairs, suggesting that age variation played a larger role in performance on mated pairs with false non-match rate performance decreasing as the age variation increased. Therefore, a child was harder to recognise when the age difference between their mated images increased.

In terms of performance on non-mated pairs, although the difference between the age variations was not considerable compared to the mated pairs, the false match rate curves do show that children were easier to identify as separate individuals when the age variation between two images increased.

Figure 34 to Figure 36 show how the performance of Algorithm E changed based on the age variation spanning from 0–10 years for each individual age (0–17 years). Appendix K contains the cumulative probability plots for the remaining five algorithms.

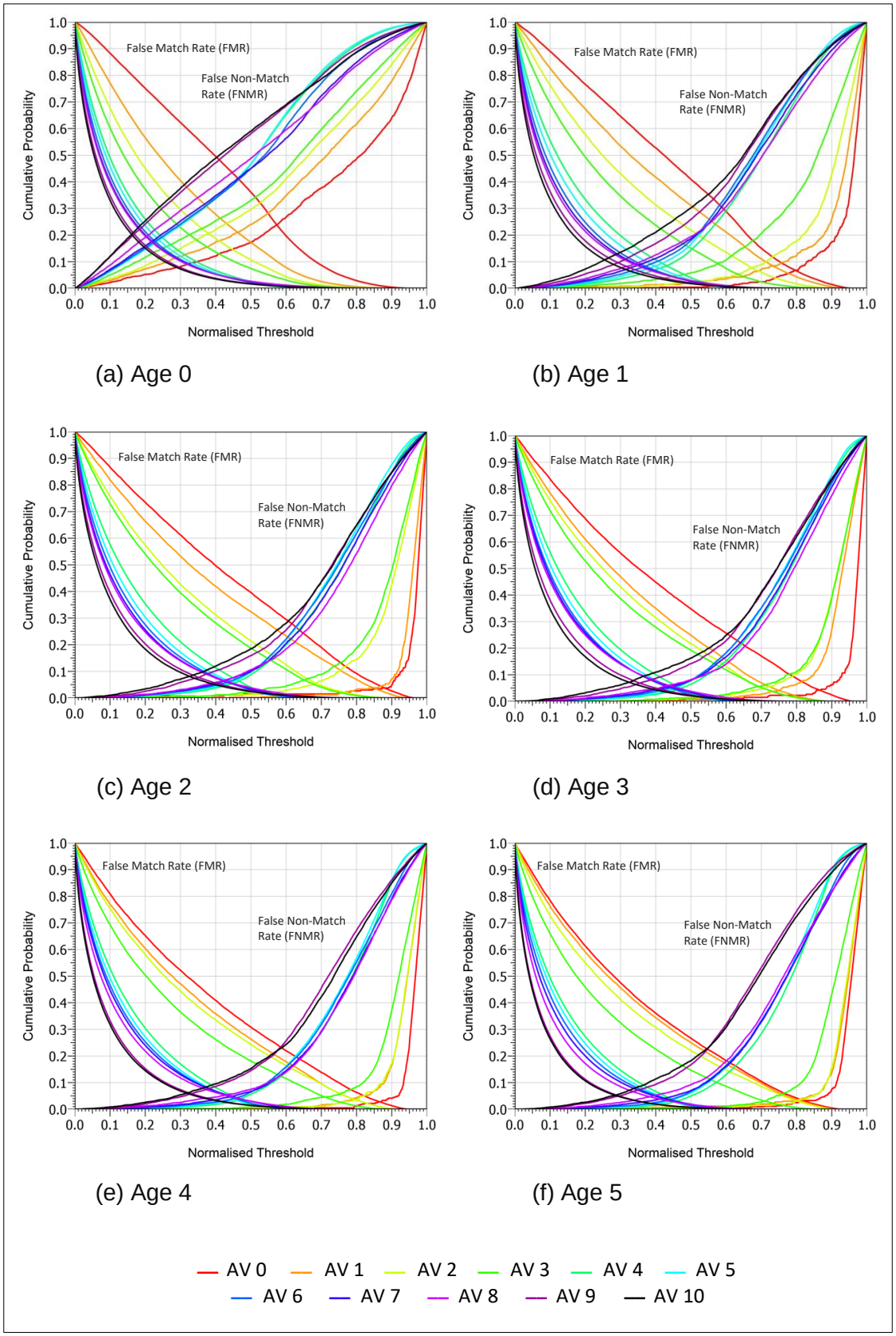


Figure 34. Cumulative probability plots for Algorithm E displaying how age variation impacts on performance for ages 0–5 (AV = age variation in years).

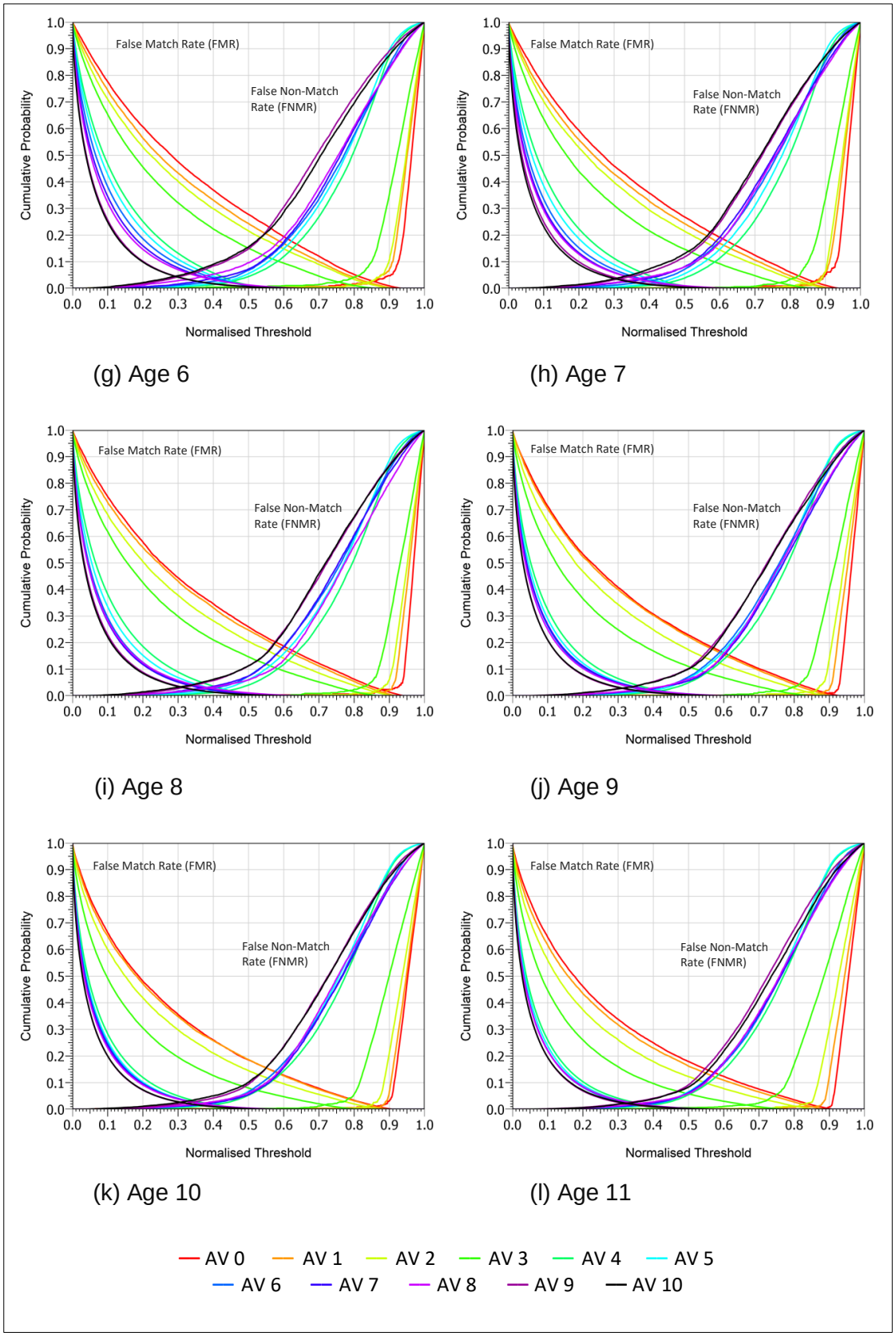


Figure 35. Cumulative probability plots for Algorithm E displaying how age variation impacts on performance for ages 6–11 (AV = age variation in years).

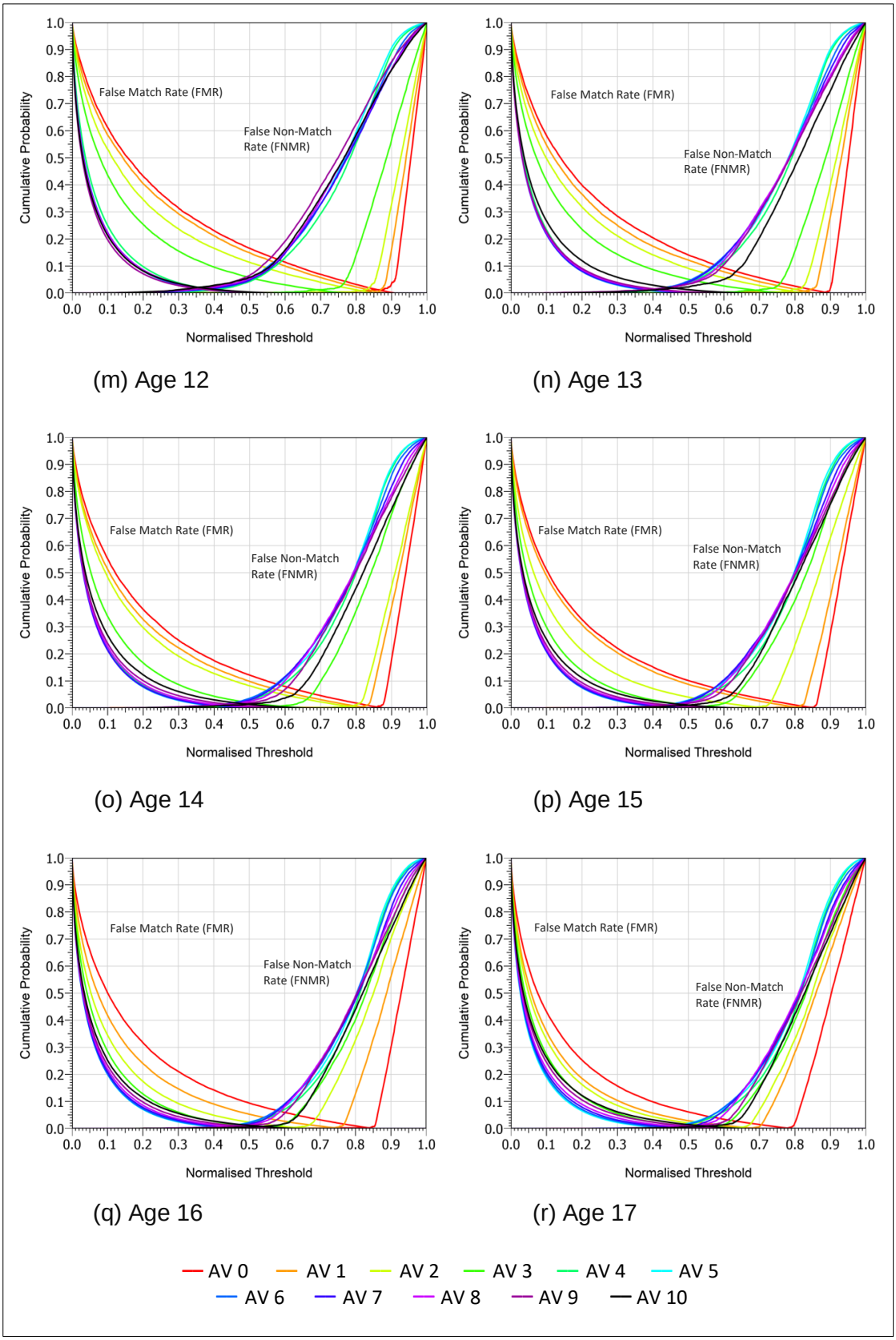


Figure 36. Cumulative probability plots for Algorithm E displaying how age variation impacts on performance for ages 12–17 (AV = age variation in years).

The plots show that the performance for both the false match rates and false non-match rates were predominately impacted by short age variations of 0–2 years regardless of age. That is, the false non-match rate was lower and the false match rate was considerably higher at these shorter age variations. Age variations greater than 3 years had minimal impact on performance across the different ages. In the teenage years, the overlap between the false match rate and the false non-match rate for each age variation was minimal, which suggests that performance based on age variation for teenagers was better than younger children. The performance between different age variations was also less spread out for both mated and non-mated pairs in the teenage years. This indicates that there was more stability in performance during the teenage years compared to younger ages.

6.4.3 *Algorithm Performance with Images of Children at Different Ages and Age Variations based on set False Match Rates*

In some operational contexts, a facial recognition system may already be set up with an operating threshold based on images of adults with a false match rate of 0.001. As shown in Study 2A (Chapter 4), this can have an impact on performance with images of children as the false match rate can be different from 0.001 when the same operating threshold is used on children. Therefore, heat map data matrices for the best performing algorithm (Algorithm E) are provided in Figure 37 to demonstrate the false match rates and false non-match rates for images of children at every age (0–17 years) and age variation (0–10 years) when the operating threshold was set for a false match rate based on images of adults at 0.001. The false match rate and false non-match rate heat map data matrices were coloured independently based on the values in the two matrices. They were coloured so that poor performance was blue, yellow was the midpoint, and purple was best performing. Appendix L contains the heat map data matrices for the remaining four state-of-the-art algorithms. The performance for OpenBR was not available at a false match rate of 0.001 as the algorithm performed too poorly.

FMR

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	0.1164	0.0266	0.0111	0.0048	0.0028	0.0019	0.0009	0.0004	0.0002	0.0002	0.0001
1	0.0867	0.0465	0.0218	0.0158	0.0088	0.0052	0.0026	0.0015	0.0009	0.0004	0.0003
2	0.0508	0.0305	0.0264	0.0200	0.0104	0.0062	0.0034	0.0020	0.0011	0.0006	0.0003
3	0.0325	0.0329	0.0277	0.0174	0.0094	0.0060	0.0035	0.0023	0.0013	0.0005	0.0002
4	0.0303	0.0279	0.0199	0.0138	0.0081	0.0055	0.0034	0.0019	0.0008	0.0003	0.0001
5	0.0221	0.0225	0.0149	0.0108	0.0070	0.0049	0.0027	0.0013	0.0005	0.0002	0.0001
6	0.0235	0.0155	0.0115	0.0079	0.0056	0.0036	0.0019	0.0007	0.0003	0.0001	0.0001
7	0.0156	0.0138	0.0092	0.0070	0.0041	0.0025	0.0011	0.0006	0.0003	0.0001	0.0001
8	0.0142	0.0104	0.0066	0.0051	0.0026	0.0015	0.0006	0.0003	0.0001	0.0002	0.0001
9	0.0100	0.0078	0.0038	0.0036	0.0013	0.0008	0.0004	0.0002	0.0001	0.0001	0.0001
10	0.0063	0.0051	0.0025	0.0015	0.0008	0.0005	0.0003	0.0001	0.0001	0.0001	0.0001
11	0.0040	0.0032	0.0023	0.0011	0.0006	0.0004	0.0002	0.0002	0.0001	0.0001	0.0001
12	0.0039	0.0019	0.0014	0.0006	0.0005	0.0003	0.0002	0.0002	0.0001	0.0001	0.0001
13	0.0015	0.0011	0.0008	0.0008	0.0004	0.0003	0.0002	0.0002	0.0001	0.0001	0.0001
14	0.0011	0.0014	0.0007	0.0005	0.0004	0.0003	0.0002	0.0002	0.0002	0.0001	0.0001
15	0.0009	0.0010	0.0005	0.0004	0.0003	0.0003	0.0002	0.0002	0.0002	0.0001	0.0001
16	0.0007	0.0006	0.0005	0.0004	0.0003	0.0003	0.0002	0.0002	0.0001	0.0001	0.0001
17	0.0004	0.0004	0.0005	0.0004	0.0003	0.0002	0.0002	0.0002	0.0001	0.0001	0.0002

FNMR

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	0.344	0.477	0.568	0.652	0.704	0.731	0.783	0.837	0.897	0.933	0.951
1	0.052	0.119	0.166	0.248	0.287	0.335	0.418	0.497	0.590	0.698	0.768
2	0.034	0.041	0.100	0.134	0.183	0.205	0.266	0.324	0.417	0.553	0.662
3	0.050	0.053	0.073	0.106	0.124	0.155	0.200	0.266	0.361	0.505	0.615
4	0.028	0.051	0.065	0.082	0.096	0.118	0.168	0.237	0.360	0.505	0.599
5	0.033	0.048	0.049	0.069	0.076	0.101	0.149	0.224	0.361	0.476	0.561
6	0.029	0.031	0.033	0.048	0.069	0.092	0.151	0.232	0.360	0.453	0.525
7	0.032	0.023	0.020	0.046	0.065	0.093	0.160	0.249	0.337	0.412	0.456
8	0.024	0.013	0.033	0.046	0.070	0.105	0.174	0.242	0.312	0.369	0.422
9	0.008	0.011	0.038	0.054	0.084	0.118	0.172	0.227	0.273	0.328	0.370
10	0.015	0.012	0.028	0.052	0.089	0.119	0.166	0.199	0.240	0.284	0.325
11	0.005	0.014	0.024	0.076	0.091	0.116	0.143	0.187	0.219	0.260	0.286
12	0.025	0.020	0.037	0.056	0.076	0.093	0.121	0.151	0.187	0.217	0.254
13	0.008	0.021	0.038	0.051	0.061	0.070	0.098	0.116	0.141	0.167	0.202
14	0.011	0.012	0.020	0.036	0.043	0.055	0.070	0.085	0.106	0.126	0.152
15	0.005	0.011	0.013	0.026	0.034	0.043	0.056	0.069	0.083	0.101	0.129
16	0.008	0.010	0.018	0.024	0.027	0.034	0.043	0.053	0.068	0.089	0.095
17	0.009	0.006	0.011	0.018	0.023	0.028	0.037	0.043	0.057	0.069	0.088

Figure 37. False match rate and false non-match rate data for Algorithm E based on a threshold set at a false match rate of 0.001 with images of adults.

Heat maps in Figure 37 and Appendix L suggest that performance between algorithms varied considerably as well as within algorithms, based on age and age variation. The heat map data matrices in Figure 37 show that if images of children were compared using a facial recognition system set at a threshold to falsely accept non-mated image pairs of adults 0.1% (i.e., FMR = 0.001) of the time, Algorithm E's performance on non-mated image pairs of children would be poorer for younger children with shorter age variations. For example, if the algorithm was set at a false match rate of 0.001 with images of adults, Algorithm E would falsely accept a non-mated image pair of babies to be the same person 11.64% of the time, compared to only 0.01% when a baby is matched to another child that is 10 years old.

In terms of the false non-match rate, Algorithm E's performance with mated image pairs also varied when the system threshold was set to incorrectly match non-mated pairs of adults 0.1% of the time. Better performance was shown with mated images of children with shorter age variations between images. Performance was poorest with babies and also young children with a large age variation between images. For example, Algorithm E would almost always fail to recognise a mated image pair of a baby and a 10 year old to be the same child (FNMR = 0.951) at the threshold chosen for use with images of adults.

Some agencies may be considering purchasing a facial recognition system just for use with children, as mentioned during Study 1. Alternatively, there may be a need to know algorithm performance with images of children when an algorithm is employed to perform at its peak. For example, by using threshold variation so that the false match rate is always 0.001 with non-mated images of children, regardless of age in childhood and the age variation between images. Figure 38 presents the false non-match rate of Algorithm E when the false match rate was set at 0.001 for each age and age variation. The heat map data matrices for the remaining four state-of-the-art algorithms can be found in Appendix M.

Youngest Age of Child in Image Pair (Years)	Age Variation (Years)										
	0	1	2	3	4	5	6	7	8	9	10
0	0.736	0.752	0.756	0.782	0.785	0.785	0.778	0.782	0.833	0.846	0.848
1	0.335	0.397	0.420	0.510	0.511	0.504	0.519	0.541	0.576	0.610	0.645
2	0.278	0.216	0.316	0.348	0.375	0.361	0.372	0.392	0.424	0.499	0.541
3	0.202	0.186	0.270	0.313	0.276	0.273	0.300	0.341	0.388	0.441	0.473
4	0.167	0.195	0.197	0.224	0.210	0.223	0.247	0.280	0.346	0.389	0.415
5	0.128	0.160	0.172	0.163	0.164	0.183	0.208	0.244	0.308	0.324	0.352
6	0.138	0.098	0.107	0.118	0.143	0.154	0.189	0.211	0.267	0.280	0.317
7	0.114	0.109	0.063	0.129	0.117	0.130	0.167	0.202	0.213	0.236	0.238
8	0.048	0.054	0.075	0.085	0.105	0.123	0.148	0.174	0.190	0.206	0.211
9	0.025	0.038	0.053	0.079	0.095	0.106	0.126	0.141	0.155	0.175	0.179
10	0.041	0.033	0.041	0.064	0.082	0.091	0.111	0.110	0.123	0.136	0.157
11	0.012	0.027	0.027	0.077	0.074	0.078	0.085	0.104	0.108	0.126	0.134
12	0.045	0.028	0.038	0.038	0.052	0.059	0.069	0.082	0.094	0.103	0.119
13	0.008	0.021	0.031	0.039	0.040	0.041	0.056	0.065	0.066	0.073	0.095
14	0.011	0.012	0.019	0.024	0.026	0.032	0.037	0.041	0.049	0.057	0.069
15	0.005	0.011	0.008	0.018	0.020	0.024	0.029	0.035	0.041	0.046	0.050
16	0.008	0.009	0.011	0.014	0.015	0.017	0.021	0.022	0.027	0.034	0.031
17	0.005	0.005	0.006	0.012	0.011	0.013	0.018	0.018	0.024	0.028	0.031

Figure 38. False non-match rate data for every age (0–17 years) and age variation (0–10 years) for Algorithm E based on a false match rate of 0.001 based on images of children at each of these ages and age variations.

This heat map data matrix clearly shows that as age increased, performance increased at every age variation and as the age variation increased, performance decreased at every age. For example, at age 6, mated pairs were accurately matched, on average, 89.3% (FNMR = 0.107) of the time when there was a 2 year age variation between images, compared to only 68.3% (FNMR = 0.317) for images with a 10 year age variation. As another example, for a 5 year age variation, mated pairs were correctly matched, on average, 77.7% (FNMR = 0.223) of the time when the youngest child in a pair was 4 years of age, and 98.7% (FNMR = 0.013) of the time when the youngest child in a pair was 17 years of age.

6.5 *Discussion*

The aim of this study was to provide data for Requirement 2 (determining facial comparison performance with images of children at different ages and age variations). A large database of images was acquired for this research that enabled algorithm performance to be captured at every age in childhood (0–17 years) with age variations ranging from 0–10 years. This was important because grouping data by multiple ages and/or age variations can make the results of minimal relevance to agencies if not specifically tailored based on their needs. Conclusions that are made can also be impacted by how ages and/or age variations have been grouped (Erbilek & Fairhurst, 2012).

Five state-of-the-art algorithms were also provided to make this the largest study with operational data and state-of-the-art algorithms with images of children to date. This is something that was critically required in a wide variety of processing, access control, and investigative applications. It is also valuable for algorithm vendors who have minimal access to images of children to conduct large-scale performance tests of their algorithms. Vendors can use the results of their algorithm to determine if an age effect exists and whether this aspect of their algorithm requires further development (Grother et al., 2011). This study was conducted with controlled images so that age-related variables could be tested in isolation as much as possible from other variables known to impact on algorithm performance (Phillips et al., 2005; Tian, Kanade, & Cohn, 2005; Yui et al., 2009; Zhang & Gao, 2009).

Discussions of the findings from this study are divided into the three sections based on the research questions this study aimed to answer.

6.5.1 *Algorithm Performance with Images of Children at Different Ages and Age Variations*

It was hypothesised that the younger the child was in an image, the poorer the algorithm performance would be. Performance in this study was shown to be progressively easier for all algorithms as age increased and supported the hypothesis and past research (Ferguson, 2015; Grother & Ngan, 2014; Mahalingam & Kambhamettu, 2012).

All algorithms performed poorly with images of babies. Wen et al. (2010) proposed that such degradations could be due to faces of babies rarely being incorporated into the training sample and/or the difference between the faces of babies being much smaller than between older faces. Vendors during Study 1 also mentioned that all babies look alike and have less pronounced facial features. As such, some researchers are developing algorithms specifically for use on newborns (Bharadwaj, Bhatt, Vatsa, & Singh; 2016; Goyal, Nagar, & Kumar, 2014; Tiwari, Singh, & Singh, 2012).

Algorithm performance was also shown to decrease as the age variation increased. Vendors during Study 1 also expected that this was due to the amount of facial change occurring throughout childhood. Again, past research has found similar results (Lui et al., 2009; Mahalingam & Kambhamettu, 2012). However, this study has shown the extent of performance degradation at each year with age variation increases of up to 10 years and at each individual age in childhood.

Although the aim of this research was to provide relevant information for agencies based on the requirements collected in Study 1, the next step would be to test algorithms using a one-to-many paradigm as this is also relevant to many operational applications. For example, an image of a child may be searched against a missing person's database. Therefore, understanding performance at the one-to-many level for agencies that are required to search databases would also be extremely valuable.

Although the aim of this study was to test algorithm performance on controlled images of children, a similar test using uncontrolled images would also be valuable to understand how much performance can deviate when other variables are present. NIST (2015) tested algorithm performance on uncontrolled images of children as part of their Child Exploitation Image

Analytics Recognition Evaluation (CHEXIA – FACE). Both one-to-one and one-to-many evaluations were conducted. Of the algorithms evaluated at the time this thesis was submitted, false non-match rates ranged from 0.47 to 1.00 based on a false match rate of 0.0001 (Grother et al., 2017). Ferguson (2015) has also conducted a study with uncontrolled images of children, however, larger datasets such as that used in the current study would also be valuable to determine performance at individual ages rather than performance at a high level that incorporates all images of children as one group. As the current study has demonstrated, there is considerable variability in performance throughout childhood which can result in misleading findings if performance is only reported at a high level.

6.5.2 *Algorithm Performance with Images of Children at Different Ages and Age Variations on Mated and Non-Mated Image Pairs*

It was hypothesised that since there are less discriminating facial features in younger children, they will be harder to recognise as themselves and to discriminate from others (Wilkinson, 2012). As such, it was expected that both the false non-match rates and false match rates would be worse at younger ages. The cumulative probability plots showed that the false non-match rates and false match rates were typically higher with younger children, supporting this hypothesis. This finding is also consistent with the evaluations conducted in the Ongoing FRVT 2017 (Grother et al., 2017) and the FRVT 2013 (Grother & Ngan, 2014). Ferguson (2015) also found there was more stability in match scores at older ages across age variation for mated pairs in her study.

It was also found that false match rates were higher for non-mated pairs with shorter age variations. This suggests that algorithms find it difficult to tell children apart if they are of a similar age rather than at different ages. This was consistent with findings from the Ongoing FRVT 2017 (Grother et al., 2017). This suggests that children of similar age are more similar in appearance than children of different ages. This is likely due to the craniofacial developmental changes that occur in childhood being similar for all individuals regardless of population (Ricanek et al., 2013). Therefore, children at the same age and gender will look like other children of the same age and gender rather than children of a different age because of these facial changes. Algorithm performance was also found to be similar on age variations of greater than 3 years (and up to 10 years) particularly at older ages. Ling et al. (2010) also found

this ‘saturation’ effect with age variations greater than 4 years and up to 10 years with passport images.

6.5.3 *Algorithm Performance with Images of Children at Different Ages and Age Variations based on set False Match Rates*

The heat map data matrices showed how performance varied when a system was already operationalised using a threshold set at a false match rate of 0.001 based on images of adults (Figure 37) as well as one that used threshold variation (Figure 38). Comparisons between these matrices highlight that threshold variation could improve accuracy of algorithms and make them more effective for comparisons with children for national security contexts (i.e., by keeping the false match rate consistently low). This is not only important for agencies considering purchasing a facial recognition system for use with children, but those that already have one but may not be using it to its full potential.

The heat map data matrices using threshold variation showed that the false non-match rate varied considerably for each algorithm. The performance of each algorithm also varied considerably based on the age of the youngest child in an image pair and the age variation between images. This data is valuable for agencies to begin to understand how drastically performance can change within childhood.

For the best performing algorithm (Algorithm E) in Study 2B of this thesis, a false non-match rate of 0.021 for adults and 0.235 for children was achieved, when both groups operated on the same threshold where the false match rate for adults was set at 0.001. The heat map data matrix in Figure 38 showed the results for the same algorithm using threshold variation approaches. As can be seen from that heat map, most of the ages at various age variations displayed false non-match rates that were lower than 0.235. This highlights issues with testing algorithm performance of children as just one overall group. Performance varies considerably and out of the 198 different categories in the heat map, only 56 (28.28%) had false non-match rates higher than 0.235 at a set false match rate of 0.001, 33 of which were from age 0–3 year categories. This is important to note because evaluating just overall performance of children would skew the data due to the extremely poor algorithm performance with babies and toddlers. As some agencies may not even make comparisons on images of younger children,

this could mislead them to believe that algorithms would not be beneficial for their particular needs.

There have been dramatic improvements in algorithm performance over the years with the best algorithm achieving a false non-match rate of 0.2 in the FRVT 2002 (Phillips et al., 2003) down to 0.003 in the MBE 2010 (Grother et al., 2011) when the false match rate was set at 0.001 with images of adults. Given these improvements and that many of the results from this study were already better than those of 0.2 originally reported in the FRVT 2002 (Phillips et al., 2003), there is hope that performance with images of children could also be improved. However, given the nature of how much the face grows and changes in childhood (Fiek & Glover, 1998; Ricanek et al., 2013), this is likely to be a much more difficult task. One way to improve performance is to conduct specific training of algorithms on images of people at the same age as those that would be inputted into the system (Klare et al., 2012). However, the lack of access to large databases of children's images is problematic for algorithm vendors. Algorithm vendors during Study 1 suggested a range of other approaches to improve performance. These included to: continue to work on their algorithms, further research into exploring the ageing of young faces, dynamic selection, threshold variation for different groups, and shortening renewal periods for younger ages.

As facial growth, particularly at younger ages is substantial, it would also be valuable to examine algorithm performance with images of children taken on the same day to gain a true baseline of performance. Although this has less operational value, it is important to understand algorithm baseline performance and how it changes over time. Jain et al. (2015) evaluated the performance of two commercial-off-the-shelf (COTS-A and COTS-B) automated systems with facial images of 206 newborns and toddlers aged 0–4 years. Images were taken in a hospital setting in India in uncontrolled conditions and each child had multiple images taken less than a minute apart with 810 images taken in total. Of these, 699 images were enrolled by both systems. When the threshold was set to incorrectly match non-mated pairs of images on average 0.1% of the time, COTS-A would correctly match mated pairs of images on average 87.57% of the time and COTS-B 81.99% of the time. A follow-up paper (Best-Rowden et al., 2016) highlighted that if the age of the child at enrolment (i.e., the age of the child in the youngest image) was less than one year of age, performance became much worse than for children one year of age or older, with almost a 30% difference in the true

accept rate. This was also found in the current study with a 40% difference in performance between age 0 with a 0 year age variation (FNMR = 0.736) and age 1 with a 0 year age variation (FNMR = 0.335). This highlights once again that, unless required for specific operational requirements, grouping data can be misleading. In the current study (using threshold variation), based on Algorithm E performance, mated pairs at age 0 with a 0 year age variation on average would only be correctly matched 26.4% (FNMR = 0.736) of the time for a false match rate of 0.001. As the age increased, the performance increased and by 4 years of age, 83.3% (FNMR = 0.167) of the mated image pairs with a 0 year age variation were correctly matched.

The heat map data matrix in Figure 38 also showed that there was a considerable reduction in the false non-match rate with images at age 8 compared to images at age 7. At age 7 with a 0 year age variation Algorithm E, on average, correctly matched pairs of mated images 88.6% of the time (FNMR = 0.114) whereas at age 8 with a 0 year age variation, Algorithm E correctly matched the majority of mated pairs (95.2%, FNMR = 0.048). Correctly measuring the distance between the eyes is an integral part of facial recognition algorithms (Riopka & Boult, 2003; Wechsler, 2007). However, the intercanthal width (i.e., the distance between the eyes) that is used by some vendors does not mature until 8 years of age in Caucasian females and 11 years of age for Caucasian males (Farkas & Hreczko, 1994). Therefore, some of this improvement in performance may be explained by the stability in the distance between the eyes of females once they reach 8 years of age. Further research examining algorithm performance based on gender across these ages would help determine if this explanation is probable.

One state-of-the-art vendor, NEC, claims that their algorithm measures approximately 80 facial points including the: distance between the eyes, width of nose, and shape of cheekbones (De Silva, Roberts, & Dowden, 2014). According to Farkas and Hreczko (1994), the width of the nose does not mature in Caucasian females until 12 years of age and 14 years of age for Caucasian males. Algorithms measuring the shape of cheekbones are also problematic for younger children who have thicker facial tissue due to the presence of subcutaneous fat. This fat along with the position of the mandible creates a more convex profile (Spalding, 2004) compared to older children. Therefore, these changes in the face over childhood are likely to impact on algorithm performance if they rely on facial points (and/or shape information as suggested in Study 1) to remain stable between image captures.

As facial recognition has traditionally been created for use on adults, facial change across childhood has largely not been taken into consideration. Future development of algorithms for use with children should consider facial change in childhood and perhaps weight different portions of the face based on their stability over time and at different ages. Age invariant systems, simulating the effects of ageing, and using a multi-modality approach are all options that may become more feasible as advancements in these areas continue (see Section 1.7 for more information).

6.6 *Summary*

Although there are many possible technological solutions that may perhaps be implemented in the future, many of these are still in their infancy and require further advancements before being operational. Currently, updating templates of children through re-enrolment on a regular basis would be the easiest solution, however, it is not often possible due to the time, costs, and nature of various applications. Therefore, it is imperative to understand performance degradations of what is currently available so that agencies can mitigate any issues they believe may impact on their current practices, policies, and procedures and help guide decisions in the future.

This study has made a significant contribution to facial recognition research. It has incorporated 4,652,868 controlled operational images of children making it the largest study testing algorithm performance with images of children known to date. Five state-of-the-art algorithms and one open source algorithm were tested on every age in childhood (0–17 years) and age variations ranging from 0–10 years. Although not all agencies use controlled images, the benefit of this research still extends to those who work with uncontrolled images as it provides data at the upper bound levels of performance. Furthermore, research at this granular level on large amounts of data not only provides data for Requirement 2 (determining facial comparison performance with images of children at different ages and age variations), it also ensures that data can be provided to answer agency specific requirements. This can be achieved by grouping the data up into specific groups based on age and/or age variation. An example of this is presented in Study 4 (Chapter 8). But first, the next chapter discusses the practitioner study conducted to collect data for Requirement 2 (determining facial comparison performance with images of children at different ages and age variations) from a practitioner perspective.

Chapter 7.

Study 3B: Facial Comparison Performance with Images of Children at Different Ages and Age Variations — Practitioner Study

7.1 *Introduction*

Study 2B (Chapter 5) was designed to provide empirical evidence for determining facial comparison performance with images of children and adults from a facial comparison practitioner perspective (to address Requirement 1). Study 2B empirically confirmed anecdotes by agencies provided in Study 1 (Chapter 2) and other studies (Heyer et al., 2017) that facial comparisons are more difficult with images of children than with images of adults, even despite there being similar image quality and age variations between the two groups.

These findings suggest that the age of a person in an image plays an important role in facial comparison practitioner performance. This is to be expected particularly in childhood when an extensive amount of facial growth is occurring (Fiek & Glover, 1998). The results may also suggest that age variations could have more of an impact on practitioner performance with images of children than images of adults. Again, this may be due to how much the face develops and changes over time (Kozak et al., 2015; Ricanek et al., 2013). Although it is critical for agencies to understand how age and age variation can impact on practitioner performance

with images taken across childhood, empirical research in this space is scarce and often anecdotal.

According to anecdotes from Study 1, some agencies were reluctant to conduct facial comparisons on images of young children (i.e., babies and toddlers) because there was a view that such a task was too difficult and unreliable. In Study 1 it was identified that there had been instances where facial comparison practitioners had relied upon skin colour when trying to determine if a baby was the same person in two images. Practitioners from all participating agencies in Study 1 were expected to make facial comparisons with images of children. Agencies typically conduct facial comparisons on images of children with age variations ranging from 0–10 years, with one agency, up to 20 years. As such, research examining practitioner performance with age variations ranging from 0–10 years with images of children is essential, although to date, such research has been limited.

Ferguson (2015) evaluated the performance of humans and algorithms conducting facial comparisons with images of children. The human study consisted of 76 participants ranging in level of experience from: expert (n = 18), limited experience (n = 11), no experience (n = 43), and other non-face experience (n = 4). These participants were exposed to 20 one-to-one trials and 10 one-to-ten trials. Age variation ranged from 0–5 years, but the majority were between 0–3 years (with no images with at a 4 year age variation and only one at 5 years). Images were acquired from a range of sources including previous studies, historical photos, internet images (including celebrity childhood images), and retrospective images which resulted in the dataset of images varying in quality. Overall performance was 65.79% for one-to-one trials and 50.72% for one-to-ten trials. Unfortunately, performance for individual ages or age groups was not attainable due to the minimal numbers of images presented to participants. Age variation only accounted for 6% of the total variance in accuracy, suggesting other variables were responsible for the variation in performance. This is perhaps not surprising given the short age variations and the large variability amongst the uncontrolled images presented in the study. Ferguson (2015) highlighted the need for a much larger study to be conducted using images of children where images were available for each age in childhood and that a study using controlled images would be beneficial in determining the impact of age and age variation.

In another study, Yadav et al. (2014) exposed 482 participants (novices) to 1 of 10 sets of images acquired from the FG-Net database and the IIIT-Delhi Facial Ageing database. A total of 54 different individuals were represented in the images, half Caucasian, half Indian. As part of this set, 5 pairs of images were presented that belonged to different age categories (0–5, 6–10), (6–10, 11–20), (11–20, 21–30), (31–50, 51–70), and (51–70, >70). Participants were required to determine if the pairs of images belonged to the same person. Images contained either the full face or a portion of the face (T-region only, T-region masked, chin region only, or binocular region only). Participants performed the highest on the (11–20, 21–30) age category when the full face was available for comparisons (87%). The poorest performing pairs were from the (0–5, 6–10) age category with accuracy of 33% for images where the T region of the face was masked. For this age category, the highest accuracy was for images displaying the binocular region (67.02%), followed by the whole face (60.41%). Yadav et al. (2014) hypothesised that as the face was undergoing the most significant amount of change during these younger years, facial comparisons were most difficult during these younger ages. It could also be evidence that the binocular region is one of the most important facial areas for comparisons with images of younger children and this may suggest that different strategies and facial features are suitable for different ages.

Although there is a lack of studies examining the impact of age variation with images of children, several studies exist with images of adults. For example, Valladares (2012) conducted a one-to-one image study with 240 novices (DST Group staff) who were presented with 120 image pairs of adults selected from the MORPH database. Images in a pair were separated by an age variation of either: 0, 5, or 10 years. The findings showed that accuracy was best for the a 0 year age variation (87.59%), followed by 10 years (85.20%), with worst performance on image pairs with a 5 year age variation (77.56%). A reason provided for the unexpected lower performance in the 5 year age variation group compared to the 10 year group was that perhaps images in the 5 year age variation group were more difficult due to differences observed in image quality. Valladares (2012) highlighted the criticality of using controlled images for age-related testing purposes to avoid such issues.

Megreya et al. (2013) conducted a one-to-one study with 80 novices (students) who were presented with cropped greyscale images that were either taken at the same time or several months apart. Accuracy rates declined more drastically than in Valladares (2012), by

approximately 20%, when images were taken months apart (Time 1/Time 2 = 67.5%, Time 2/Time 1 = 72.1%) compared to at the same time (Time 1/Time 1 = 87.5%, Time 2/Time 2 = 92.3%). Deviations were once again likely due to considerable differences in methodologies employed, for example the image types used.

Overall accuracy is useful to provide a high level overview of performance, but agencies also need to understand how much of a role the actual image type plays in performance. For example, is performance consistent for both mated (i.e., same person) and non-mated images (i.e., different people) or does difficulty vary more for one image type than the other over age and age variation in childhood? Agencies require this information because errors in facial comparison decisions can have varying consequences depending on whether errors occur on mated or non-mated image pairs. For example, in processing applications, practitioners usually conduct more facial comparisons on mated pairs since non-mated pairs would only be present due to fraud or human error. If practitioners find it difficult to determine that a pair of images is mated, it could inconvenience customers by expecting them to provide more information to prove they are who they claim to be. If practitioners find it difficult to determine that a pair of images is non-mated, there are potential issues with fraudulent images being accepted and this could have severe implications for national security depending on the particular circumstances. In investigative applications, low accuracy rates on mated image pairs could result in victims being unidentified, whilst low accuracy rates on non-mated pairs could result in a loss of valuable resources due to following false leads.

Very few studies have considered the difference in performance of facial comparison practitioners between mated and non-mated image pairs, particularly with images of children. Ferguson's (2015) study found that performance with mated images of children was 62.44% compared to 70.72% with non-mated pairs. Unfortunately, this was only conducted at a high level with uncontrolled images and not able to be broken down into ages (or age groups) or different age variations due to the limited size of the study. Thus, it is still unclear if practitioner performance varies for mated and non-mated image pairs over childhood with controlled operational images. Understanding performance based on pair type may help agencies to better direct their resources and determine where more training with facial comparison practitioners would be valuable. If results for the current study show a consistent pattern with Study 2B in that non-mated images are more difficult, this would reinforce the need to focus

on trying to discriminate between facial features in images of children. Discriminating between facial features is also useful in applications where only exclusion is required rather than identification or verification. For example, in investigative agencies, all that may be required is to exclude potential child victims from a case.

The current study was designed to follow on from the previous facial comparison practitioner study conducted in this thesis (Study 2B, Chapter 5). It is important for agencies to understand how performance changes as a function of age and age variation. The aim of this study was to empirically determine the performance of facial comparison practitioners with images of children in regards to age and age variation at a finer level, and to address Requirement 2 captured during Study 1. To date, the few studies that have been conducted in this area have been severely restricted due to a lack of access to a reasonable number of controlled images at every age throughout childhood, as well as the need for multiple images of the same child. The database of images acquired for this research (see Section 3.1.1) provided a unique opportunity to utilise data at this low level. Understanding how facial comparison performance changes as a factor of age and age variation can aid agencies in:

- determining where more caution may be required when making facial comparison decisions;
- providing a general awareness of deficient performance for any ages and age variations that may be evident in the results;
- designing training tailored for more difficult ages in childhood and/or age variations;
- selecting particular images to be automatically referred to a higher level or specialist group where more tools, expertise, and time are available to make a decision; and
- considering alternative methods apart from the face to make a decision for ages and/or age variations that are more difficult.

7.2 *Research Questions*

The aim of the current study was to acquire as much data as possible, at the finest level possible, to provide information to address Requirement 2. Two research questions were answered to achieve this.

Question 1. To what extent is facial comparison practitioner performance impacted by the age of children (0–17 years) in images and the age variation between images in a pair ranging from 0–10 years?

Question 2. To what extent does facial comparison practitioner performance vary over age and age variation based on the type of image pair presented (i.e., mated or non-mated)?

Anecdotes from facial comparison practitioners in Study 1 and previous research (Heyer et al., 2017) suggest that facial comparisons with younger children are difficult. It is also well-known that the younger a child is, the more underdeveloped their facial features are (Wilkinson, 2012). Thus, it was anticipated that overall accuracy would be poorest at younger ages and gradually increase as age increased. As a considerable amount of facial growth occurs in childhood (Kozak et al., 2015; Ricanek et al., 2013), it was also expected that as age variation between images in a pair increased, overall performance would decrease.

Similar to what was hypothesised in Study 2B (Chapter 5), it was also hypothesised here that as children tend to look alike and have less discriminating facial features making them more difficult to distinguish from each other (Wilkinson, 2012), non-mated performance would be lower than mated performance throughout childhood. Due to the extensive amount of change and development over time (Kozak et al., 2015; Ricanek et al., 2013), it was also expected that mated and non-mated accuracy would decrease as the age variation increased. This is because the amount of facial growth and change in children may make them look less like themselves and more like others.

7.3 *Methodology*

This section describes the methodology applied to evaluate facial comparison practitioner performance.

7.3.1 *Participants*

A total of 120 facial comparison practitioners (90 females, 30 males) completed the study (mean age = 41.49 years, SD = 10.65). Practitioners were from the same government agency that participated in Study 2B. There was approximately four months between Study 2B and

the current study and facial comparison practitioners had the opportunity to participate in either, both, or none of the studies but were only allowed to participate in either study once. Practitioners had experience in making facial comparison decisions ranging from 1 month to 36 years. All practitioners had normal or corrected to normal vision. The majority of participants (84.17%) were Caucasian. A total of 115 (95%) practitioners had some form of facial comparison training (77.5% on the job, 40% formal training).

7.3.2 *Materials*

Materials required for this study were previously discussed and include the controlled operational facial image database (Section 3.1.1), the Biometrics High Performance Computing Cluster (Section 3.1.4), Comparer (Section 5.3.3.1), the experimental application software (Section 5.3.3.2), work computers (Section 5.3.3.4), and one state-of-the-art facial recognition algorithm (Section 4.3.2).

7.3.3 *Image Pair Selection*

A total of 23,760 image pairs were manually selected for this facial comparison practitioner study. This consisted of 120 image pairs (60 male, 60 female) being selected for every age in childhood (0–17 years) and every age variation in years ranging from 0–10 years. This resulted in 198 different categories (198 categories x 120 pairs per category = 23,760 pairs). This is illustrated in Figure 39 followed by the process employed to select mated and non-mated image pairs.

Youngest Age of Child in Image Pair (Years)	Age Variation (Years)										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	120	120	120	120	120	120	120	120	120	120	120
1	120	120	120	120	120	120	120	120	120	120	120
2	120	120	120	120	120	120	120	120	120	120	120
3	120	120	120	120	120	120	120	120	120	120	120
4	120	120	120	120	120	120	120	120	120	120	120
5	120	120	120	120	120	120	120	120	120	120	120
6	120	120	120	120	120	120	120	120	120	120	120
7	120	120	120	120	120	120	120	120	120	120	120
8	120	120	120	120	120	120	120	120	120	120	120
9	120	120	120	120	120	120	120	120	120	120	120
10	120	120	120	120	120	120	120	120	120	120	120
11	120	120	120	120	120	120	120	120	120	120	120
12	120	120	120	120	120	120	120	120	120	120	120
13	120	120	120	120	120	120	120	120	120	120	120
14	120	120	120	120	120	120	120	120	120	120	120
15	120	120	120	120	120	120	120	120	120	120	120
16	120	120	120	120	120	120	120	120	120	120	120
17	120	120	120	120	120	120	120	120	120	120	120

Figure 39. The 198 different categories in this study based on age and age variation with 120 image pairs selected per category. The circled cell indicates 120 image pairs contain a child at age 6 and an image of a child 4 years older.

7.3.3.1 *Mated Image Pair Selection*

A total of 11,880 mated image pairs were required for this study. This consisted of 60 mated image pairs (30 male, 30 female) at every age in childhood (0–17 years) and every age variation ranging from 0–10 years. The IDs of all appropriate mated pairs of children from the database were separated in Microsoft Excel™ by the 198 required categories and gender.

A random number generator in Microsoft Excel™ was used to randomly order these pairs by category and gender. A manual inspection of the pairs was conducted using Comparer to ensure that pairs fulfilled the criteria presented in Table 6 (in Chapter 3). If an image pair did not fulfil the criteria, the pair was discarded and the next appropriate image pair was selected. This process continued until all 30 female mated image pairs and all 30 male mated image pairs were selected for each of the 198 categories. Images were checked to ensure the same images were not used twice (e.g., if multiple images were available of a person, the different images of that person could have been selected multiple times for different categories, but each individual image was only used once).

7.3.3.2 *Non-Mated Image Pair Selection*

A total of 11,880 non-mated image pairs were also required for this study. Five steps were taken to select the non-mated pairs:

1) one state-of-the-art facial recognition algorithm was used to conduct a one-to-many search on each of the 11,880 youngest age images used in the mated pairs (the same age variation and gender was used from those mated pairs);
2) the top 30 highest scoring non-mated images for each of the 11,880 images were returned;
3) the 30 image pairs were sorted from highest scoring to lowest scoring for each of the 11,880 images;
4) the first image pair from the 30 pairs that fulfilled the criteria presented in Table 6 was selected; and
5) step 4 was repeated another 11,879 times.

In summary, as shown in Figure 40, the youngest age image from every mated pair was used as the youngest age image in every non-mated image pair. The same age variation was used for the mated pairs as the corresponding non-mated pairs. This approach was taken to keep some consistency between the mated and non-mated pairs and provide a better indication of whether children matched better to themselves or to someone else as originally hypothesised.

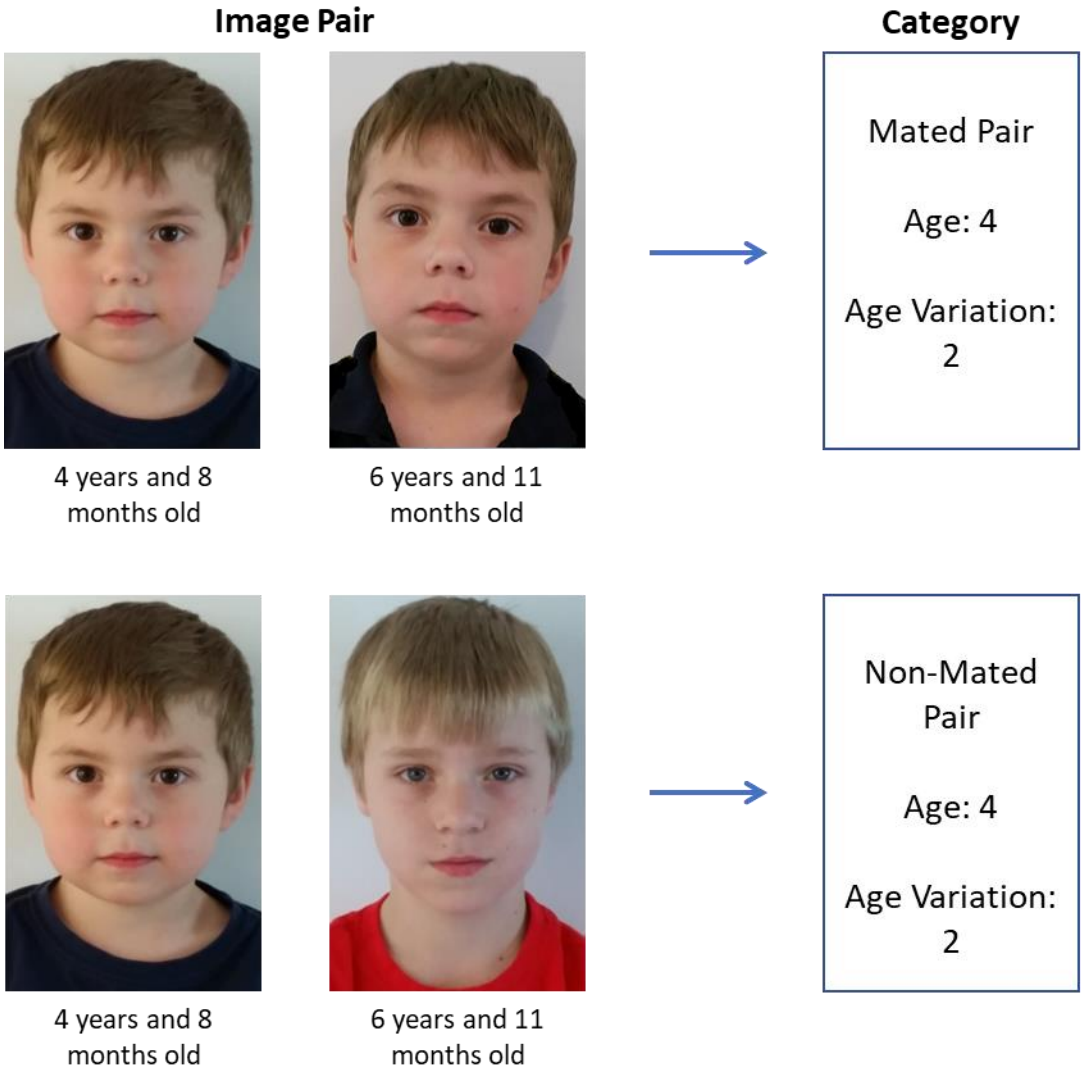


Figure 40. Example of the type of images used for the mated and non-mated pairs. The image of the child at the youngest age in the mated pair is also used as the youngest child in the non-mated pairs. Images are for illustration purposes only.[13]

Once the total 23,760 image pairs were manually screened to ensure that non age-related variables were removed as per Table 6, a check in Microsoft Excel™ was conducted to ensure that the older image in a non-mated pair was only ever used once. Inspection of the images showed that 130 images (1.1%) were selected more than once. Therefore, some image pairs needed to be reselected. This required returning to each appropriate category and selecting a different second image from the 30 available. Once this was complete, another check was

[13] Copyright © 2017. Images used with signed parental consent.

conducted that identified 10 images that were still used more than once. Once these 10 were replaced, another check was conducted finding only one image that needed replacing. One final check was conducted and ensured that all the older images in a non-mated pair were unique.

Once this process was complete, two independent judges with security clearances manually screened the 23,760 image pairs. This was to ensure consistency in image quality within the image pairs, particularly as the age variation increased because image quality standards changed over the 10 year period. Judges were provided with Table 6 to ensure that each image pair adhered to the criteria. One last final check over the 23,760 image pairs was then conducted to ensure the researcher was satisfied with the consistency throughout the dataset. This image selection process took four months.

7.3.4 *Procedure*

A similar process to Study 2B was followed where approval by higher management, followed by section managers was sought prior to seeking interest from facial comparison practitioners. This occurred approximately four months after Study 2B had been conducted. Information regarding the study was distributed by higher management to inform facial comparison practitioners that a study was going to be conducted and to seek volunteers to participate. Section managers collected email addresses from practitioners who chose to nominate.

Similar to Study 2B, practitioners were sent an email the week prior to the study opening to give them an opportunity to plan when they would set aside a block of time of around one hour to conduct the study. This was in order to minimise impact on workflow and to ensure a higher rate of participation as practitioners could choose a time that suited them best. Facial comparison practitioners were then sent out an email which included a link to the study available on the agency's intranet, along with a unique ID and password. The email also informed practitioners that they could withdraw from the study at any time, that they were to work on the study alone, and that their individual results would not be provided to management.

The layout of the experimental application was consistent with Study 2B (see Section 5.3.3.2). Practitioners were asked a range of demographic questions and then presented with 198

image pairs and asked to decide if the pairs were of the 'same' person or 'different' people and to rate their confidence in their decision. Following the image comparison component participants were asked a final question:

> *"Please comment on any methods you used to help make a decision. For example, looking at the whole face overall, gut feeling, specific facial features etc."*

This question was asked to provide qualitative information that may help to determine if there are particular facial features that practitioner's compare to aid in performance and whether a rather homogenous group (i.e., practitioners from the same agency predominantly with training) all use the same strategies. This information was also collected to compare with studies in the future to try and ascertain what facial features in childhood are beneficial for discriminating between children.

A total of 120 practitioners were presented with a unique set of 198 image pairs. One image pair was selected from each of the 198 categories based on age (0–17 years) and age variation (0–10 years) as shown in Figure 39. This consisted of 99 mated and 99 non-mated image pairs, half the pairs were male, half female. Thus, 120 image pairs per category were each seen once by one of the 120 practitioners, as would be typically expected in operational applications where it is likely that only one practitioner will view the images and make a decision. This approach was chosen as the focus here was on performance of different ages in childhood over different age variations, rather than individual differences of practitioners. As each participant saw only one image pair per category, there was no concern that they would view the same youngest image twice, once in a mated image pair and once in a non-mated image pair as they would represent the same age category.

Since each facial comparison practitioner was assigned a unique set of 198 image pairs, several weeks were given to the original 120 practitioners to complete the study. Two email reminders were sent out to practitioners over this time. As some of these 120 practitioners decided not to participate, their unique set of 198 pairs were provided to another practitioner that was available on standby and had not already participated. These practitioners were given a different ID and password to open the specific dataset that had not yet been completed (in

case the original practitioner changed their mind and decided to participate, this approach ensured they would not see each other's responses). Three iterations of this process were required before 120 facial comparison practitioners completed one of the 120 unique datasets containing 198 unique image pairs from the 23,760 unique image pairs available. This process took four months.

7.4 *Results*

The results section is divided into the two research questions that this study aimed to provide empirical data for from a facial comparison practitioner perspective. Discrimination and bias data is also presented in Section 7.4.1. The results for each research question contain performance data based on accuracy, confidence, and response times using the same data screening, assumption checking, and data analysis approach as Study 2B (Chapter 5). Unless otherwise stated, the heat map data matrices were coloured so that green was better performance, red was worse and yellow was the midpoint. This is followed by a list of strategies/facial features that practitioners indicated that they used when making their decisions.

7.4.1 *Practitioner Performance with Images of Children at each Age and Age Variation*

Figure 41 provides the overall accuracy for facial comparison practitioners with images of children at every age in childhood (0–17 years) with age variations ranging from 0–10 years.

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	70.83	69.17	61.67	65.00	68.33	59.17	63.33	63.33	59.17	64.17	61.67
1	80.00	80.00	76.67	70.00	73.33	69.17	73.33	66.67	63.33	64.17	61.67
2	87.50	80.83	78.33	70.83	77.50	75.00	72.50	81.67	72.50	67.50	70.00
3	81.67	85.83	83.33	83.33	84.17	80.83	74.17	80.00	75.00	75.83	70.83
4	83.33	80.00	85.83	85.00	76.67	83.33	76.67	78.33	78.33	69.17	75.83
5	90.00	79.17	80.83	85.00	77.50	79.17	79.17	78.33	75.83	78.33	70.83
6	90.83	86.67	80.00	85.00	82.50	83.33	81.67	84.17	80.83	75.83	73.33
7	88.33	85.00	87.50	94.17	84.17	83.33	77.50	80.83	79.17	79.17	79.17
8	89.17	88.33	92.50	79.17	85.00	85.83	83.33	83.33	81.67	75.00	81.67
9	93.33	89.17	88.33	85.00	85.00	83.33	80.00	78.33	79.17	75.83	80.83
10	90.83	94.17	86.67	81.67	86.67	84.17	83.33	87.50	80.00	75.00	76.67
11	89.17	89.17	92.50	84.17	89.17	80.00	84.17	84.17	81.67	78.33	74.17
12	90.83	88.33	90.00	83.33	81.67	80.83	81.67	85.00	85.00	83.33	80.83
13	89.17	83.33	88.33	84.17	90.83	84.17	79.17	87.50	82.50	81.67	84.17
14	93.33	90.00	89.17	92.50	90.83	87.50	85.83	88.33	84.17	82.50	85.00
15	90.00	94.17	90.00	85.00	85.83	85.83	88.33	85.00	85.00	80.00	83.33
16	90.00	90.00	95.00	86.67	92.50	85.00	82.50	85.83	90.83	90.00	81.67
17	92.50	92.50	90.00	94.17	92.50	88.33	88.33	85.83	85.00	85.83	87.50

Figure 41. Overall accuracy for each age and age variation (%).

Overall accuracy ranged from 59.17% to 95% with an overall mean of 81.81% (*Mdn* = 83.33%, *SD* = 7.68). Overall patterns can be seen in the heat map data matrix. For example, overall accuracy was best in the lower left corner of the matrix indicating that facial comparison practitioners performed more accurately with images of children that were older and which had less age variation between images. Performance was best with image pairs at age 16 with a 2 year age variation (95%). It can also be seen that generally the longer the age variation, the poorer the performance for each age. Accuracy was poor with images of babies, regardless of the age variation. Performance was poorest with image pairs at age 0 with a 5 and 8 year age variation (59.17%).

Figure 42 shows the overall confidence for facial comparison practitioners with images of children at every age in childhood (0–17 years) with age variations ranging from 0–10 years.

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	71.00	67.42	63.83	65.75	64.08	63.08	64.08	63.75	63.58	63.92	62.17
1	73.25	73.50	73.50	69.92	69.33	69.92	69.25	71.08	69.17	65.58	68.25
2	73.25	73.67	74.33	74.92	70.42	72.83	74.25	72.08	73.00	69.50	67.92
3	74.58	76.00	77.92	74.17	75.25	76.00	74.25	71.83	72.83	71.83	69.75
4	77.17	76.00	75.58	76.58	73.17	73.08	72.58	72.25	70.58	72.25	70.92
5	78.08	76.08	76.25	78.08	75.00	77.42	75.00	74.75	73.92	70.75	70.00
6	78.00	78.42	76.83	76.83	76.33	76.50	76.08	73.33	75.33	72.67	70.92
7	79.08	77.33	76.58	76.92	77.75	76.42	77.17	74.33	73.67	73.00	73.17
8	78.08	77.42	80.08	76.42	77.67	76.17	75.67	76.67	73.08	74.75	74.67
9	76.67	79.00	77.17	78.92	73.83	76.92	72.00	77.00	74.58	73.75	73.08
10	78.67	78.00	78.00	76.83	75.00	77.83	76.67	74.92	75.50	74.92	73.67
11	80.67	76.67	79.92	77.75	77.42	75.92	76.83	75.67	79.67	74.50	77.25
12	80.00	80.33	77.75	76.83	77.33	77.00	74.67	77.67	77.42	77.08	77.58
13	83.58	80.00	81.00	77.00	77.58	76.83	76.50	77.33	77.50	76.08	75.67
14	79.08	78.83	77.33	77.42	77.92	76.00	76.50	77.00	79.08	78.08	77.42
15	81.00	78.67	79.50	78.08	78.08	75.83	76.08	79.75	78.00	79.08	77.92
16	79.50	80.33	77.08	79.25	79.42	77.75	78.67	78.08	77.17	77.17	76.25
17	77.92	81.42	77.25	79.75	79.17	77.50	78.00	78.00	78.33	76.00	76.08

Figure 42. Overall confidence for each age and age variation (%).

Overall confidence ranged from 62.17% to 83.58% with an overall mean of 75.32% (*Mdn* = 76.46%, *SD* = 3.89). Overall trends in confidence over age and age variation were reasonably consistent with overall accuracy data, in that, the older the child and the shorter the age variation, the more confident the facial comparison practitioners felt in their decisions. Practitioners were most confident with image pairs at age 13 with a 0 year age variation (83.58%) and were the least confident with babies (age 0) across all age variations, particularly with images at age 0 with a 10 year age variation (62.17%).

Figure 43 shows the overall timings (in seconds) for facial comparison practitioners with images of children at every age in childhood (0–17 years) with age variations ranging from 0–10 years.

Youngest Age of Child in Image Pair (Years)	Age Variation (Years)										
	0	1	2	3	4	5	6	7	8	9	10
0	15.87	16.54	15.45	17.28	16.00	16.64	16.16	17.37	16.46	16.48	17.62
1	16.43	14.86	14.46	15.29	17.20	15.75	15.20	15.82	15.95	16.37	16.83
2	17.42	14.35	14.91	14.25	15.73	13.93	14.99	16.58	16.55	16.07	17.25
3	16.08	15.35	15.48	14.99	14.69	14.36	13.54	17.75	16.51	14.95	16.02
4	15.54	14.28	13.95	13.87	16.17	16.19	15.42	15.98	14.72	16.87	16.15
5	15.10	14.66	14.83	14.46	14.52	15.55	15.76	16.41	15.28	16.78	16.99
6	13.32	13.02	14.66	14.26	13.91	14.46	13.47	16.54	16.06	15.00	17.19
7	15.13	13.98	14.24	13.13	14.86	15.11	15.04	15.05	15.54	14.88	15.53
8	14.12	14.34	15.14	14.12	14.68	15.18	16.60	14.31	15.70	17.36	15.56
9	13.38	14.46	14.47	14.33	16.15	14.83	16.59	15.74	18.18	16.34	15.29
10	14.24	14.87	13.82	14.88	16.60	14.73	15.69	16.73	15.93	16.19	15.43
11	13.11	13.93	14.34	15.01	14.87	15.41	15.04	15.77	14.08	16.86	15.12
12	13.16	14.99	16.56	16.24	16.47	16.22	15.79	15.63	15.01	16.47	15.13
13	13.07	15.34	14.56	15.33	15.37	16.59	14.80	14.81	14.99	15.27	16.06
14	14.01	16.25	14.58	16.61	15.74	16.55	16.05	16.16	15.95	15.84	13.43
15	14.01	14.54	15.07	14.21	15.08	15.52	15.79	15.68	15.19	15.64	14.83
16	15.03	14.73	16.00	14.53	14.35	14.84	15.12	16.23	15.01	13.78	16.14
17	15.24	15.06	14.56	13.70	15.35	14.27	16.21	14.69	14.74	14.92	14.30

Figure 43. Overall response times for each age and age variation (seconds).

Overall response times ranged from 13.02 to 18.18 seconds with an overall mean of 15.33 seconds (*Mdn* = 15.22, *SD* = 1.03). Trends in overall response times showed that practitioners were generally faster when age variations were under 5 years compared to above 5 years. The shortest response time was with image pairs of children at age 6 with a 1 year age variation (13.02 seconds). The longest average response time was with image pairs of children at age 9 with an 8 year age variation (18.18 seconds).

Figure 44 contains discrimination and bias data for every age (0–17 years) and age variation (0–10 years) examined. As there were 198 conditions, providing descriptive statistics and pairwise comparisons for each one was considered disproportionate to the amount of valuable information this would provide. As such, it was decided Figure 44 should present the most valuable data that could be obtained from discrimination and bias metrics and so the mean data for each age and age variation was presented to provide data that could be compared to other research that typically does not conduct non-parametric testing and therefore only reports the mean. Discrimination was colour coded: 0 as red, 0.5 as yellow (chance performance), 1 as green (perfect discrimination). Bias was colour coded: -1 as red, 0 as yellow, 1 as green, with less than 0 indicating liberal bias (i.e., say ‘same’ more) and more than 0 indicating conservative bias (i.e., say ‘same’ less).

A'
Discrimination

Youngest Age of Child in Image Pair (Years)	Age Variation (Years)										
	0	1	2	3	4	5	6	7	8	9	10
0	.57	.56	.53	.54	.55	.52	.53	.53	.52	.54	.53
1	.65	.65	.62	.57	.59	.56	.59	.55	.53	.54	.53
2	.75	.66	.63	.57	.63	.60	.58	.67	.59	.55	.57
3	.67	.73	.69	.69	.70	.66	.60	.65	.60	.61	.57
4	.69	.65	.73	.71	.62	.69	.62	.64	.63	.56	.61
5	.79	.64	.66	.71	.63	.64	.64	.64	.61	.64	.57
6	.81	.74	.65	.71	.68	.69	.67	.70	.66	.61	.59
7	.76	.71	.75	.87	.70	.69	.63	.66	.64	.64	.64
8	.78	.76	.84	.64	.71	.73	.69	.69	.67	.60	.67
9	.85	.78	.76	.71	.71	.69	.65	.63	.64	.61	.66
10	.81	.87	.74	.67	.74	.70	.69	.75	.65	.61	.61
11	.78	.78	.84	.70	.78	.65	.70	.70	.67	.63	.59
12	.81	.76	.79	.69	.67	.66	.67	.71	.71	.69	.66
13	.78	.69	.76	.70	.81	.70	.64	.75	.68	.67	.70
14	.85	.79	.78	.84	.81	.75	.73	.76	.70	.68	.71
15	.79	.87	.79	.71	.72	.73	.76	.71	.71	.65	.69
16	.79	.79	.89	.74	.84	.71	.68	.72	.80	.79	.67
17	.84	.84	.79	.87	.84	.76	.76	.73	.71	.73	.75

B" Bias

Youngest Age of Child in Image Pair (Years)	Age Variation (Years)										
	0	1	2	3	4	5	6	7	8	9	10
0	.12	-.14	-.16	-.09	-.37	-.12	-.22	-.08	-.10	-.22	-.10
1	-.13	.00	-.26	-.20	-.40	-.28	-.30	-.25	-.14	-.19	-.08
2	.29	-.24	-.42	-.27	-.13	-.45	-.26	-.37	-.13	-.12	-.32
3	.00	-.05	-.24	-.24	.04	-.03	-.02	-.13	-.18	-.27	-.27
4	-.16	.00	.15	.00	.10	-.33	-.15	-.11	-.42	-.01	-.45
5	.00	.21	-.03	-.18	-.08	-.15	.03	.00	-.27	-.11	-.32
6	-.08	.00	-.06	.18	-.19	.00	.07	-.22	-.39	-.39	-.35
7	-.12	-.09	-.29	.69	-.13	-.33	-.08	-.17	-.34	-.21	-.57
8	-.07	-.12	-.31	-.15	-.09	-.05	.08	-.24	.00	-.39	-.22
9	-.23	-.49	.12	-.09	-.38	-.08	-.26	-.29	-.27	-.39	-.47
10	-.25	.13	-.32	-.22	-.44	-.40	-.24	-.29	-.40	-.09	-.67
11	-.20	-.65	-.31	-.13	-.34	-.40	-.22	-.22	-.29	-.67	-.40
12	-.25	-.12	-.30	-.42	-.29	-.39	-.22	-.28	.00	-.24	-.55
13	-.07	.00	-.25	-.22	-.42	.13	-.21	-.41	-.27	-.45	-.50
14	-.47	-.46	-.20	-.52	-.25	-.29	-.25	-.25	.04	-.44	-.85
15	-.15	-.41	-.15	-.18	-.46	.15	-.12	-.28	-.38	-.40	-.24
16	-.46	.15	-.32	.00	.10	.00	-.19	-.46	-.79	-.80	-.22
17	.31	-.31	-.30	-.41	.10	.00	-.25	-.25	-.18	-.25	-.29

Figure 44. Discrimination and bias for each age and age variation.

The discrimination heat map data matrix shows that practitioners performed better than chance for every age and age variation examined. Performance was closest to chance with images of babies regardless of age variation and with toddlers at longer age variations.

Practitioners were better at discriminating whether image pairs were of the same person or different people when the age variation was shorter as well as with image pairs containing older children.

The bias heat map data matrix shows that for a large majority of the different ages and age variations, practitioners were liberal in deciding that image pairs contained the same person. Practitioners were always liberal with image pairs containing babies or toddlers with age variations of 5 years or greater. Practitioners were most liberal in selecting 'same' for image pairs containing older teenagers with longer age variations.

7.4.2 *Practitioner Performance with Images of Children at each Age and Age Variation on Mated and Non-Mated Image Pairs*

Figure 45 presents the heat map data matrices for practitioner accuracy on the 11,880 mated pairs and accuracy on the 11,880 non-mated pairs respectively. The mated and non-mated heat map data matrices used the same colouring format rules to show how performance collectively varied based over the two pair types (i.e., lowest performance in either matrix was coloured red and highest in either was coloured green with yellow being the midpoint of the highest and lowest scores). Mated and non-mated heat map data matrices are provided together for easy comparison followed by an explanation of the data for the two matrices.

Mated Accuracy

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	65.00	76.67	76.67	71.67	86.67	73.33	80.00	70.00	71.67	80.00	71.67
1	83.33	80.00	85.00	80.00	88.33	83.33	85.00	81.67	75.00	78.33	70.00
2	83.33	86.67	90.00	83.33	81.67	90.00	83.33	90.00	78.33	75.00	85.00
3	81.67	86.67	88.33	88.33	83.33	81.67	75.00	83.33	81.67	85.00	83.33
4	86.67	80.00	83.33	85.00	73.33	90.00	81.67	81.67	90.00	70.00	90.00
5	90.00	73.33	81.67	88.33	80.00	83.33	78.33	78.33	85.00	81.67	85.00
6	91.67	86.67	81.67	81.67	86.67	83.33	80.00	88.33	90.00	88.33	86.67
7	90.00	86.67	91.67	90.00	86.67	90.00	80.00	85.00	88.33	85.00	93.33
8	90.00	90.00	95.00	83.33	86.67	86.67	81.67	88.33	81.67	88.33	86.67
9	95.00	95.00	86.67	86.67	91.67	85.00	86.67	86.67	86.67	88.33	91.67
10	93.33	93.33	91.67	86.67	93.33	91.67	88.33	91.67	90.00	78.33	95.00
11	91.67	96.67	95.00	86.67	93.33	90.00	88.33	88.33	88.33	95.00	88.33
12	93.33	90.00	93.33	91.67	88.33	90.00	86.67	90.00	85.00	88.33	93.33
13	90.00	83.33	91.67	88.33	95.00	81.67	85.00	93.33	88.33	91.67	93.33
14	96.67	95.00	91.67	96.67	93.33	91.67	90.00	91.67	83.33	91.67	98.33
15	91.67	96.67	91.67	88.33	93.33	83.33	90.00	90.00	91.67	90.00	88.33
16	95.00	88.33	96.67	86.67	91.67	85.00	86.67	93.33	98.33	98.33	86.67
17	90.00	95.00	93.33	96.67	91.67	88.33	91.67	90.00	88.33	90.00	91.67

Non-Mated Accuracy

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	76.67	61.67	46.67	58.33	50.00	45.00	46.67	56.67	46.67	48.33	51.67
1	76.67	80.00	68.33	60.00	58.33	55.00	61.67	51.67	51.67	50.00	53.33
2	91.67	75.00	66.67	58.33	73.33	60.00	61.67	73.33	66.67	60.00	55.00
3	81.67	85.00	78.33	78.33	85.00	80.00	73.33	76.67	68.33	66.67	58.33
4	80.00	80.00	88.33	85.00	80.00	76.67	71.67	75.00	66.67	68.33	61.67
5	90.00	85.00	80.00	81.67	75.00	75.00	80.00	78.33	66.67	75.00	56.67
6	90.00	86.67	78.33	88.33	78.33	83.33	83.33	80.00	71.67	63.33	60.00
7	86.67	83.33	83.33	98.33	81.67	76.67	75.00	76.67	70.00	73.33	65.00
8	88.33	86.67	90.00	75.00	83.33	85.00	85.00	78.33	81.67	61.67	76.67
9	91.67	83.33	90.00	83.33	78.33	81.67	73.33	70.00	71.67	63.33	70.00
10	88.33	95.00	81.67	76.67	80.00	76.67	78.33	83.33	70.00	71.67	58.33
11	86.67	81.67	90.00	81.67	85.00	70.00	80.00	80.00	75.00	61.67	60.00
12	88.33	86.67	86.67	75.00	75.00	71.67	76.67	80.00	85.00	78.33	68.33
13	88.33	83.33	85.00	80.00	86.67	86.67	73.33	81.67	76.67	71.67	75.00
14	90.00	85.00	86.67	88.33	88.33	83.33	81.67	85.00	85.00	73.33	71.67
15	88.33	91.67	88.33	81.67	78.33	88.33	86.67	80.00	78.33	70.00	78.33
16	85.00	91.67	93.33	86.67	93.33	85.00	78.33	78.33	83.33	81.67	76.67
17	95.00	90.00	86.67	91.67	93.33	88.33	85.00	81.67	81.67	81.67	83.33

Figure 45. Accuracy for mated and non-mated image pairs for each age and age variation (%).

Figure 45 shows that the colouring in the mated heat map data matrices does not change much due to the lower accuracy on non-mated pairs and highlights the superior performance of practitioners on mated image pairs.

Accuracy on mated pairs ranged from 65% to 98.33% with a mean 86.97% (*Mdn* = 88.33%, *SD* = 6.21). Practitioners had the highest accuracy on mated pairs with images at age 15 with a 10 year age variation and at age 16 with an 8 and 9 year age variation (98.33%). Lowest accuracy on mated pairs was with images at age 0 with a 0 year age variation (65%).

A visual inspection of the data suggests that mated pairs were impacted by age but not age variation. A Friedman ANOVA was run to determine if these observations were statistically significant by grouping the data by age (i.e., all results for age 0 with age variations of 0–10 years combined, all results for age 1 with age variations of 0–10 years combined etc.). A Friedman ANOVA indicated that performance significantly varied across age, χ^2 (17) = 184.99, $p < .001$.

Follow-up pairwise comparisons were not of interest to agencies (i.e., statistically significant results at individual ages were not of interest at this level). However, it was decided to report this data to help highlight significant differences in performance across childhood that may be of interest to other agencies and/or researchers. Writing out a considerably long and detailed list of descriptive statistics for the 153 comparisons that were conducted with mated image pairs at every age (compared to every other age) was considered arduous and of little value on this scale. Instead, Appendix N presents the data showing all ages that were statistically significant in practitioner accuracy from other ages on mated image pairs. Data reported is at the $p < .001$ level (as required due to the Bonferroni adjustment) using the Wilcoxon Signed Rank test.

Separately, agency specific questions can be answered and provided to agencies where more appropriate. In addition, a mock example that is relevant to passport processing and border control applications to determine statistical significance amongst groups is provided as Study 4 (Chapter 8).

A Friedman ANOVA was also conducted to determine if age variations for mated pairs were significant (i.e., all results for a 0 year age variation with ages 0–17 years combined, all results for a 1 year age variation with ages 0–17 years combined etc.). The test indicated that performance did not differ significantly based on age variation, χ^2 (10) = 14.15, $p = .166$. Thus,

the youngest age of a child in a mated image pair impacted performance, but the age variation between mated images in a pair did not.

A Wilcoxon Signed Rank test revealed a statistically significant decrease in accuracy with non-mated image pairs (*M* = 76.66%, *Mdn* = 78.33%) compared to mated image pairs (*M* = 86.97%, *Mdn* = 88.33%), $z = -10.67$, $p < .001$, $r = -.54$.

Non-mated accuracy as presented in Figure 45 ranged from 45% to 98.33% with a mean of 76.66% (*Mdn* = 78.33%, *SD* = 11.41). The best performing category for non-mated pairs was age 7 with a 3 year age variation (98.33%), while the worst performing category was age 0 with a 5 year age variation (45%). In general, performance was poorer when an image pair contained a baby regardless of age variation and there was a noticeable decline in accuracy for age variations above 7 years on non-mated pairs.

The non-mated heat map data matrices suggest that practitioner performance was impacted by both age and age variation. Friedman tests were run once again to confirm whether these observations were statistically significant. As expected, the Friedman ANOVA indicated that performance significantly varied across age, χ^2 (17) = 301.46, $p < .001$. Appendix O presents the data showing all ages that were statistically significant in practitioner accuracy from other ages on non-mated image pairs.

A Friedman test also indicated that performance on non-mated pairs differed significantly based on age variation, χ^2 (10) = 199.81, $p < .001$. Appendix P presents the data showing all age variations that were statistically significant in practitioner accuracy from each other on non-mated image pairs.

In summary, facial comparison practitioner performance with mated pairs of children differed as a function of age but not as a function of age variation. However, accuracy with non-mated pairs of children differed as a function of age and as a function of age variation.

Figure 46 presents the heat map data matrices for practitioner confidence on mated and non-mated image pairs respectively.

Mated Confidence

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	69.33	64.50	63.83	67.67	64.17	62.50	66.67	62.17	67.67	59.50	61.83
1	76.33	74.50	77.17	72.67	70.83	74.17	70.00	74.00	72.33	68.50	67.00
2	71.50	76.50	74.33	78.17	71.17	73.67	74.83	73.17	74.83	66.67	68.83
3	76.83	76.67	81.83	75.50	74.00	78.17	77.67	71.83	73.67	72.00	66.83
4	75.67	73.50	77.33	75.50	73.67	77.33	72.83	72.83	71.50	70.50	74.50
5	78.67	75.50	78.17	79.17	75.17	79.67	75.00	75.50	75.00	74.00	72.50
6	79.67	77.67	76.50	75.17	75.33	78.33	76.67	73.67	76.50	72.67	76.83
7	79.33	76.00	77.83	77.00	79.17	76.83	79.33	75.83	74.83	72.50	74.00
8	78.50	76.33	78.17	79.33	78.67	72.83	78.17	76.67	75.00	75.83	74.50
9	80.83	81.50	75.17	76.83	75.83	78.33	75.33	81.00	77.83	74.83	73.83
10	81.67	80.50	79.33	76.83	71.67	79.17	73.83	76.67	76.83	78.00	78.33
11	83.17	76.83	83.33	78.33	77.50	79.00	76.17	73.17	81.67	79.33	81.67
12	81.67	85.33	81.33	77.33	80.67	77.83	78.33	78.67	77.67	77.17	79.17
13	84.50	81.17	82.00	77.67	76.83	78.00	77.83	79.17	77.00	75.83	80.67
14	83.00	81.17	81.17	78.00	78.00	77.33	78.17	78.00	81.17	79.00	80.33
15	82.50	82.00	82.83	80.67	80.83	76.17	77.83	80.83	80.17	82.67	78.50
16	85.33	79.67	79.00	78.17	78.50	79.00	79.17	82.17	78.83	79.67	76.17
17	78.83	84.83	77.83	85.33	81.83	78.17	83.83	80.83	80.00	77.00	76.83

Non-Mated Confidence

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	72.83	71.67	64.00	64.00	63.83	64.00	61.33	64.50	59.83	68.33	62.83
1	69.83	74.00	70.33	68.00	68.67	66.17	69.00	68.67	66.17	63.50	69.83
2	75.33	71.83	74.50	72.33	69.83	73.50	74.17	71.83	70.17	72.50	67.00
3	73.17	75.50	74.50	73.17	77.33	73.67	71.33	72.17	72.50	72.67	73.67
4	79.33	79.00	74.33	78.17	73.67	69.67	73.00	72.50	70.33	74.00	67.67
5	78.33	77.00	75.33	77.67	75.83	76.00	75.67	74.50	73.33	68.17	67.83
6	77.33	79.50	78.00	78.83	77.33	75.67	75.33	73.33	74.17	73.17	65.17
7	79.67	79.00	76.00	77.00	76.83	76.17	76.00	72.67	72.83	74.50	73.00
8	77.67	78.67	82.17	74.17	77.67	80.00	74.00	77.50	71.50	74.33	75.33
9	73.50	77.33	79.67	81.67	72.67	76.50	68.83	73.50	71.50	72.50	73.00
10	76.33	76.17	76.67	77.17	79.17	77.50	79.83	73.33	74.83	71.50	70.17
11	78.33	77.33	77.50	77.67	78.17	73.50	78.33	78.33	78.50	69.83	73.67
12	78.00	76.17	75.00	77.17	74.33	76.17	72.17	77.50	77.83	77.67	76.67
13	82.83	78.67	80.33	77.00	78.83	76.50	75.50	76.33	78.67	76.33	71.33
14	75.50	77.33	74.50	77.67	78.00	75.33	75.50	76.33	77.67	77.17	75.50
15	80.33	76.00	77.00	76.00	75.33	76.50	75.00	78.33	76.50	75.33	78.17
16	74.33	81.50	75.83	80.67	80.67	77.17	79.00	75.00	76.00	75.33	76.33
17	77.83	78.50	78.00	75.00	77.00	77.83	72.17	75.83	78.17	75.50	75.50

Figure 46. Confidence for mated and non-mated image pairs for each age and age variation (%).

The heat map data matrix shows that the confidence of facial comparison practitioners with mated image pairs ranged from 59.50% to 85.33% with a mean of 76.55% (*Mdn* = 77.25%,

SD = 4.57). It can be seen that the older a child is in the youngest image, the more confident practitioners were in their decisions. Practitioners were the most confident with image pairs at age 17 with a 3 year age variation (85.33%). Practitioners were least confident with image pairs at age 0 regardless of age variation, but they were least confident with image pairs at age 0 with a 9 year age variation (59.50%).

Friedman ANOVAs were run to determine if there was any difference in practitioner's confidence with mated image pairs based on both age or age variation. There was a statistically significant difference in confidence with mated pairs based on age, χ^2 (17) = 400.48, p < .001 and age variation, χ^2 (10) = 77.98, *p* < .001.

A Wilcoxon Signed Rank test revealed a statistically significant decrease in confidence with non-mated image pairs (*M* = 74.61%, *Mdn* = 75.50%) compared to mated image pairs (*M* = 76.55%, *Mdn* = 77.25%), *z* = -6.59, *p* < .001, *r* = -.33.

Practitioner's confidence with non-mated image pairs ranged from 59.83% to 82.83% with a mean 74.61% (*Mdn* = 75.50%, *SD* = 4.14). Similar patterns can be seen in both the mated and non-mated heat map data matrices, albeit generally slightly lower in confidence with non-mated pairs. The top right corner of the non-mated heat map data matrix indicates lower confidence, suggesting practitioners were typically less confident with younger ages at larger age variations. Practitioners were the least confident when non-mated image pairs were at age 0 with an 8 year age variation (59.83%), but were most confident when non-mated image pairs were at age 13 with a 0 year age variation (82.83%). Friedman ANOVAs indicated that confidence with non-mated pairs varied significantly based on age, χ^2 (17) = 266.41, p < .001 and age variation, χ^2 (10) = 125.29, *p* < .001.

Figure 47 presents the heat map data matrices for practitioner response times (in seconds) on mated and non-mated image pairs respectively.

Mated Timings

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	16.23	16.02	15.73	18.31	14.86	16.66	17.00	16.62	15.23	18.39	15.74
1	17.17	16.00	14.13	15.42	16.23	15.94	14.88	15.77	15.19	14.91	15.59
2	18.58	13.34	14.17	14.20	13.77	14.05	12.92	17.14	15.47	16.66	16.53
3	16.49	15.54	14.51	15.41	14.38	15.10	14.40	17.84	14.92	15.66	14.18
4	17.13	15.36	15.20	11.90	16.50	15.64	14.81	15.86	13.98	16.84	15.08
5	15.05	13.99	14.60	13.10	15.54	15.91	15.85	17.35	14.82	14.88	17.14
6	14.44	13.58	14.51	14.06	15.29	16.27	12.29	15.69	15.36	15.58	16.22
7	14.32	13.52	15.03	12.82	14.85	15.49	14.66	13.84	13.91	15.02	14.65
8	15.57	13.89	16.36	14.41	16.09	16.22	14.62	15.89	14.33	16.94	16.24
9	14.33	14.09	13.68	14.42	14.75	15.26	16.08	14.87	17.05	17.02	12.91
10	16.13	13.72	13.71	14.10	16.39	13.68	15.35	15.25	15.92	14.76	13.94
11	12.01	14.55	13.29	14.03	14.98	15.78	14.01	14.37	13.80	15.65	13.91
12	12.89	14.54	17.70	14.09	15.48	15.67	15.28	14.75	15.88	17.21	14.03
13	13.82	13.91	13.00	15.10	13.33	15.08	13.88	13.71	14.45	14.98	14.91
14	12.98	15.03	13.70	16.60	17.19	16.13	16.52	14.93	16.25	15.23	12.03
15	13.82	14.68	14.58	14.44	15.83	15.42	16.56	15.20	13.77	15.11	14.94
16	13.85	14.25	16.43	14.40	15.85	13.33	15.70	15.23	13.41	13.31	14.31
17	15.37	15.07	13.02	13.62	14.05	13.94	14.13	13.35	13.77	13.24	13.97

Non-Mated Timings

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	15.51	17.06	15.18	16.24	17.14	16.61	15.32	18.11	17.70	14.57	19.50
1	15.70	13.71	14.79	15.17	18.16	15.57	15.53	15.86	16.71	17.83	18.06
2	16.27	15.37	15.65	14.29	17.69	13.81	17.06	16.02	17.64	15.48	17.97
3	15.67	15.17	16.45	14.57	14.99	13.62	12.68	17.66	18.10	14.24	17.87
4	13.95	13.21	12.69	15.85	15.84	16.74	16.02	16.10	15.45	16.90	17.22
5	15.15	15.33	15.06	15.82	13.49	15.19	15.66	15.46	15.73	18.69	16.84
6	12.21	12.45	14.82	14.45	12.54	12.64	14.66	17.40	16.75	14.41	18.16
7	15.93	14.44	13.45	13.43	14.87	14.73	15.42	16.26	17.17	14.73	16.40
8	12.66	14.79	13.92	13.83	13.26	14.15	18.58	12.73	17.07	17.77	14.89
9	12.44	14.82	15.26	14.23	17.56	14.39	17.11	16.62	19.31	15.66	17.68
10	12.35	16.02	13.94	15.65	16.80	15.78	16.03	18.20	15.93	17.62	16.93
11	14.22	13.30	15.38	15.99	14.76	15.03	16.08	17.16	14.35	18.07	16.32
12	13.43	15.43	15.42	18.38	17.46	16.77	16.31	16.51	14.14	15.72	16.22
13	12.31	16.76	16.11	15.55	17.40	18.11	15.72	15.90	15.53	15.56	17.22
14	15.04	17.47	15.45	16.61	14.29	16.96	15.57	17.39	15.64	16.45	14.84
15	14.21	14.40	15.56	13.98	14.34	15.62	15.01	16.16	16.61	16.16	14.71
16	16.20	15.22	15.57	14.65	12.84	16.34	14.54	17.23	16.61	14.24	17.98
17	15.11	15.06	16.10	13.78	16.66	14.60	18.30	16.03	15.72	16.60	14.63

Figure 47. Response times for mated and non-mated image pairs for each age and age variation (seconds).

A visual inspection of the two heat map data matrices shows that practitioners generally took longer on non-mated than mated image pairs. The average response times for facial

comparison practitioners for mated image pairs ranged from 11.90 to 18.58 seconds with a mean of 14.99 seconds (*Mdn* = 14.94 seconds, *SD* = 1.26). Generally, practitioners were faster at making decisions on image pairs with shorter age variations, although there is variation across the matrix. Practitioners were faster at making decisions when image pairs were at age 4 with a 3 year age variation (11.90 seconds). Practitioners took the longest to make decisions on image pairs at age 2 with a 0 year age variation (18.58 seconds).

Friedman ANOVAs were run to determine if there was any difference in practitioner's response times with mated image pairs based on age or age variation. These results indicated that response times were statistically significant based on age, χ^2 (17) = 35.64, p = .005 but not based on age variation, χ^2 (10) = 13.51, p = .197. However, when age 0 was removed, response times were not statistically significant across ages 1–17 years, χ^2 (16) = 25.19, p = .066.

A Wilcoxon Signed Rank test revealed a statistically significant increase in how long it took to make decisions with non-mated image pairs (*M* = 15.66 seconds, *Mdn* = 15.65 seconds) compared to mated image pairs (*M* = 14.99 seconds, *Mdn* = 14.94 seconds), z = -4.69, p < .001, r = -.24.

The average response times for facial comparison practitioners for non-mated image pairs ranged from 12.21 to 19.50 seconds with a mean of 15.66 seconds (*Mdn* = 15.65 seconds, *SD* = 1.53). Practitioners were fastest at making decisions on image pairs at age 6 with a 0 year age variation (12.21 seconds). Practitioners took the longest to make decisions on image pairs at age 0 with a 10 year age variation (19.50 seconds).

Friedman ANOVAs indicated that response times with non-mated pairs varied significantly based on age, χ^2 (17) = 36.66, p = .004 and age variation, χ^2 (17) = 61.93, p < .001.

7.4.3 *Strategies Adopted by Practitioners to make Facial Comparison Decisions*

As part of the study, practitioners were asked an open-ended question on what strategies they used and/or facial features they compared when making their facial comparison decisions. Practitioners responses to this question were categorised. For example, if a practitioner mentioned one or multiple components of the nose (e.g., nostrils, bridge), this was grouped

into the 'nose' category and recorded once (as one practitioner mentioned components of the nose). All responses for each practitioner were categorised and the frequency was calculated. Table 10 provides a list of the facial features and strategies adopted by practitioners when making facial comparison decisions and the number of practitioners adopting these approaches.

Table 10. *Strategies/Features Adopted to make Facial Comparison Decisions*

Strategy/Feature	Number of Practitioners (%)
Ears e.g., shape, size, position	63 (52.50%)
Nose e.g., shape, nostrils, bridge	37 (30.83%)
Eyes e.g., shape, inner canthus, eyelid shape, outer area of eyes	37 (30.83%)
Mouth e.g., lip shape, size, gap between lips, cupids bow	28 (23.33%)
Whole face	26 (21.67%)
Gut feeling	21 (17.50%)
Markings e.g., freckles, moles, blemishes	21 (17.50%)
Individual facial features (not specified)	20 (16.67%)
Face shape	13 (10.83%)
Chin e.g., shape, distance from other features	7 (5.83%)
Eye pupil distance e.g., relative to each other and rest of face	6 (5.00%)
Hair e.g., hairline, hair patterns	6 (5.00%)
Jaw e.g., jawline and shape	4 (3.33%)
Eyebrows	2 (1.67%)
Forehead e.g., size, forehead-face ratio	2 (1.67%)
Philtrum	1 (0.83%)
6 FR points i.e., ears, eyes, nose, mouth, shape of face, facial marks	1 (0.83%)
Compared lighting angles	1 (0.83%)
Took into account pose, lighting, expression	1 (0.83%)

Strategies adopted amongst a relatively homogenous group (trained facial comparison practitioners from the same agency) varied considerably. Some practitioners started by

viewing the whole face then breaking it down into facial features if necessary. However, while some practitioners said they looked at the face as a whole for at least part of their strategy, others were insistent that they would not look at the face as a whole at all.

Over half of the practitioners (52.50%) mentioned that they compared the ears, although it was also mentioned that the ears were not always visible. Another 20 practitioners (16.67%) mentioned that they focussed on facial features, but did not specify which ones. Comparing face shape was another strategy adopted by 12 practitioners (10.83%). Comparing the eyes and nose were typically mentioned together. Neck and facial markings also tended to be used if other common facial features did not provide enough information for practitioners to come to a decision. One participant (0.83%) mentioned using the six FR points. This refers to the six face segments; ears, eyes, nose, mouth, shape of face, and facial marks which are sometimes taught in facial comparison training programs (DIBP, 2009). Some practitioners mentioned specific components within common facial features, such as the inner canthus (corner of eye) and cupids bow (curve of upper lip) highlighting their knowledge of different facial features.

A total of 21 practitioners (17.50%) mentioned making decisions using their 'gut feeling', particularly when other strategies were not useful. One practitioner (0.83%) mentioned that they compared lighting angles between images in a pair. This is a concern given that it was known to practitioners that images were taken at different times and in different locations which would result in different lighting angles. Furthermore, differences in lighting would also have been minimal given the controlled nature of the images and similar image quality selected within pairs.

7.5 *Discussion*

The aim of this study was to collect data to provide empirical evidence for determining facial comparison performance with images of children at different ages and age variations (Requirement 2, identified in Study 1). A unique opportunity to evaluate performance at a fine level across childhood at every age (0–17 years) and age variation (0–10 years) with facial comparison practitioners from a government agency using a controlled operational facial image database was made possible for this study. This makes this research extremely important as it has not been possible in past studies and is critical research required for many operational applications.

A novel, but more operationally appropriate methodology than has been typically used in previous research, was adopted for this study. In this study, each participant was presented with a unique set of images. Showing the same set of images to each practitioner was considered, however this would have resulted in only one pair of images being selected for each of the 198 categories, rather than 120 different pairs. Therefore, each category would have contained either a male or female, mated or non-mated pair. Furthermore, selecting just one pair for that category may not have been representative of children for that particular age and age variation category. Therefore, it was decided to use an approach that reflected real operational applications where practitioners are working on unique caseloads and to evaluate performance based on the group as a whole.

The methodology was also unique in that image selection involved a state-of-the-art algorithm to return the highest scoring non-mated images. An image was then selected that followed strict criteria including similar appearance and image quality as presented in Table 6. Given that this was a unique opportunity to test 120 practitioners from a government agency, the aim was to select images in a pair that may be more realistic in operational settings rather than randomising the images used in each pair. Achieving this was extremely resource intensive, but was considered a more operationally representative approach and therefore more ecologically valid.

It is worth reminding the reader, that practitioners from the participating agency do not typically conduct facial comparisons with images of children under 6 years of age. Therefore, performance data for these lower ages only demonstrates that they found these ages more difficult. This could reflect the difficulty of this task as anecdotally identified in Study 1 (Chapter 2) and in Heyer et al. (2017), but it could also be due to these practitioners having less experience conducting facial comparisons with images of children at these younger ages.

Next, a discussion based on the two questions this study aimed to answer and the strategies adopted by practitioners to aid in their facial comparison decisions is provided.

7.5.1 *Practitioner Performance with Images of Children at each Age and Age Variation*

As hypothesised, practitioners performed poorer with younger images of children. This can be seen in the discrimination heat map data matrix presented in Figure 44, that showed that practitioners performed at just above chance level with images of babies and gradually increased as age increased. The overall accuracy heat map data matrix presented in Figure 41 also showed that as the age of the child in the youngest image increased, performance of practitioners increased. This finding is also consistent with anecdotes provided by facial comparison practitioners in Study 1 and in Heyer et al. (2017). Even when the age variation stayed constant at 0 years, performance was lower for younger ages in childhood. This was depicted in Figure 41 where the first column of data in the heat map data matrix displayed the 0 year age variation for each age. This column showed that there was up to a 22.5% difference in performance over the different ages at a 0 year age variation (0–11 months). This demonstrates that performance was not perfect even when there were short age variations between images. From an operational perspective, Figure 41 highlights that even if there was less than a year between when images were taken, it will be more difficult to determine if younger children (i.e., babies and toddlers) are the same person or different people, compared to if the images contained a child at 5 years of age or older when performance becomes more consistent.

However, it is worth remembering that facial growth is a nonlinear process and a considerable amount of growth can occur within a 0 year age variation, particularly at younger ages. As such, it would be beneficial for future research to test practitioner performance with images of children taken on the same day to ascertain the upper bound level of practitioner performance.

Overall accuracy in the heat map data matrix differed by up to 35.83% based on age and age variation. There was up to a 33% difference based on age alone (i.e., when the age variation was the same). This was for age 0 compared to age 16, both with a 2 year age variation. Poorer performance at younger ages may be due to these children having less discriminating facial features, but also due to the amount of facial change occurring early in life (Kozak et al., 2015; Ricanek et al., 2013; Wilkinson, 2012).

Megreya et al. (2013) showed that when presented with images containing the faces of adults with an age variation of several months apart, performance decreased by about 20%. In this current study, overall accuracy with images of children did not decrease as drastically, with only a 20% decrease in performance being reached once the age variation was expanded to 9 years (this was for a 0 year age variation compared to a 9 year age variation for age 2). Valladares (2012) found a 10% decrease in performance for a 5 year age variation compared to a 0 year. However, these other studies used different types of datasets containing images of adults and the participants were novices, making comparisons with previous research problematic. A study by Ferguson (2015), which was more similar to the current study, as it used images of children and included practitioners as participants, found that age variations of up to five years with image pairs of children had minimal impact on overall performance compared to other variables that may have impacted on performance such as image type (i.e., uncontrolled images).

The heat map data matrix providing the confidence of practitioners (Figure 42) showed relatively similar findings to the patterns displayed in the overall accuracy matrix (Figure 41). That is that practitioners were the least confident with pairs containing younger children and most confident when images were of older children with shorter age variations. Practitioners were generally slightly more accurate than confident at each age and age variation. The patterns across the two heat maps changed relatively consistently, suggesting that practitioners were generally aware of their abilities across the age-related conditions. Although, there were larger differences between accuracy and confidence for some individual categories. For example, practitioners were considerably under confident with image pairs at age 7 with a 3 year age variation (17.25% different between accuracy and confidence). More noticeable differences between accuracy and confidence can be seen once the data is divided into pair type. This is discussed in Section 7.5.2.

Overall response times are notoriously problematic (Field, 2103) and the response times collected in this study were no exception. However, it could be seen that practitioners were typically faster when age variations were under five years. One-to-one studies using images of adults and that have not deadlined participating practitioners show that response times can vary. In Calic's (2012) study, practitioners averaged around 3 seconds per pair and in the photo-to-photo test in White, Kemp, Jenkins, Matheson, et al. (2014), about 8 seconds. The

time taken to make decisions on image pairs in the current study was a lot higher, at around 15 seconds, although this was similar to images from the Child group in Study 2B. This could be an indication of the difficulty of conducting facial comparisons with images of children compared to adults. The response time data also provides agencies with a better understanding of how long a facial comparison review decision takes practitioners when not deadlined.

7.5.2 *Practitioner Performance with Images of Children for each Age and Age Variation on Mated and Non-Mated Image Pairs*

It is critical for agencies to understand facial comparison practitioner performance based on image pair types to better understand whether practitioners perform differently when making facial comparison decisions on mated and non-mated images. This is because the consequences for inaccurate decisions can vary depending on the image pair type. Once the data was divided into pair type, it was evident that practitioners were generally less accurate, less confident, and took longer to make decisions on non-mated than mated image pairs.

Practitioners were 89.87% accurate with mated image pairs and 76.66% accurate with non-mated pairs. This is consistent with the hypothesis that non-mated pairs would be more difficult than mated pairs and the findings from Study 2B. This may be due to children portraying less discriminating facial features, making them more difficult to distinguish from each other (Wilkinson, 2012). Performance on non-mated pairs also decreased as the age variation increased. The bias heat map data matrix presented in Figure 44, showed that practitioners were selecting 'same' more often at longer age variations. This could suggest that practitioners were changing their decision threshold criteria as the age variation increased based on knowledge that the face can drastically change in childhood. Thus, they may have been convincing themselves that the change in appearance was based on a growth spurt rather than it actually being two different people. Further evidence of this can be seen in the heatmap data matrix presented in Figure 45 that shows a distinct decline in accuracy after a seven year age variation, with minimal decline in confidence (as shown in Figure 46). After a seven year age variation, practitioners in this study were more confident than accurate.

Vernon (1952), who was interested in the cognitive aspects of visual perception, stated that "possible and permissible variations in appearance and behaviour are always taken into

account" (p. 14). Bruce (1994) posited that 'possible and permissible' variations of appearance helped to determine the boundaries between one face and another; and that our general knowledge of faces and how they change allows us to anticipate how they may appear with a different expression, or at another viewpoint. As suggested previously, it could also be that practitioners used their general knowledge of how faces change to anticipate how they may appear at a different point in time. This becomes problematic when comparing non-mated image pairs of children over long age variations because what is believed to be 'possible and permissible' becomes a lot broader and, as a result, accuracy decreases but confidence remains relatively stable. It may be that more training on how the face actually changes across childhood would improve accuracy if practitioners were provided with information of what is morphologically 'possible and permissible'.

Another hypothesis was that accuracy with mated image pairs would also decrease as the age variation increased due to the amount of facial growth occurring, making children look more like others than themselves over time. However, accuracy for mated pairs was relatively stable across age variation. This contrasting difference in results to what was hypothesised may also be explained by practitioner's perceptions as to what they believe is 'possible and permissible'. For example, at shorter age variations, practitioners may see the similarities within the two images being compared, but as the age variation increases, there is a shift in their decision criterion as they need to rely more on what they believe is 'possible and permissible'. A better understanding of what facial features are most stable in childhood for comparison purposes would ensure that practitioners were basing their decisions on knowledge rather than on assumptions.

Nevertheless, the accuracy rates of facial comparison practitioners with mated (89.87%) and non-mated (76.66%) images of children that were obtained here is promising. This is particularly given the accuracy rates in past research (Ferguson, 2015; White, Kemp, Jenkins, Matheson, et al., 2014) and that image pairs in the current study included babies and toddlers (that are not typically compared by these practitioners) and age variations of up to 10 years. However, there was a lot of variance across childhood based on age and age variation. As shown in the heat map data matrices in Figure 45, accuracy on mated image pairs varied by up to 31.67% based on age and by up to 21.67% based on age variation. Accuracy on non-mated image pairs varied by up to 46.67% based on age and by up to 36.67% based on

age variation. The diverse results over age and age variation suggest that agencies should exert extreme caution when conducting facial comparisons with images of children, particularly at lower ages. Given that the images in this study were controlled, it is likely that performance would degrade even further with uncontrolled images over the same ages and age variations.

The findings show that generally, the longer the age variation, the more difficult it was for practitioners to determine that the images were of different people (i.e., non-mated). There are several possible explanations for this. Practitioners may be:

- able to mentally visualise the difference between two faces over the age variation based on what they believe is 'possible and permissible';
- selecting 'same' as a default option since their work duties would consist more of comparing mated than non-mated image pairs. Thus, they may only select 'different' if they can actually see something dissimilar between images in a pair that cannot be explained; and
- focusing on facial features that they expect to be stable but are not in childhood (e.g., the nose changes considerably in early childhood).

To determine the potential reason for this finding, it would be beneficial to conduct a study with novices or a different group of practitioners who are exposed to more non-mated than mated pairs to see if the pattern in results is similar to those presented here. It would also be valuable for future research to identify facial features that become stable in childhood to help discern children from each other.

Due to practitioners performing poorer with non-mated image pairs that have longer age variations, it may be useful to consider age progressing the younger image to the same age as the older image in a pair and looking for any resemblance. This was identified in Study 1 as another requirement of interest to agencies (Requirement 6: determining the performance of age progressed images for facial comparisons). Future research should consider testing for any benefits to both mated and non-mated accuracy when age progressing images. Research should consider age progression both by forensic artists who have detailed knowledge of how the face changes in childhood and also by automated algorithms that have been trained on

large numbers of images. Currently, such research is being planned with multiple agencies as a separate research program from this thesis.

When the confidence of practitioners was divided into mated and non-mated pairs, it was evident that practitioners were slightly more confident on mated pairs throughout the different ages and age variations. Again, practitioners were the least confident with babies (age 0) over the different age variations. Confidence with babies was highest with a 0 year age variation, regardless of whether pairs were mated or non-mated. However, mated accuracy with image pairs at age 0 with a 0 year age variation was 65%, lower than any other 0 year age variation with mated pairs. Non-mated accuracy with image pairs at age 0 with a 0 year age variation was 76.65%, which was higher than any other age variation at age 0 with non-mated pairs. This may be explained by babies going through a lot of facial change and practitioners not expecting their appearance to change so drastically, therefore assuming the images were of different people and not realising otherwise (i.e., they do not believe the amount of change is 'possible and permissible'). Confidence slightly declined further on non-mated pairs with large age variations. This may also reflect practitioners making decisions on what they believe is 'possible and permissible' at these longer age variations but being less sure if they have made accurate assumptions than they were on shorter age variations.

Response times also showed practitioners were slower at making decisions on non-mated pairs than mated pairs, particularly at longer age variations. This was expected due to the nature of the non-mated image selection and the amount of facial growth that occurs in childhood, making it difficult for practitioners to distinguish these faces. Previous research by Megreya et al. (2013) also showed that participants took longer on one-to-ten trials when the target was not present compared to when they were. They also found that participants were faster at making decisions when the age variation was short (i.e., same day) rather than over longer periods (i.e., months apart). These patterns are consistent with the findings from the current study. However, in their one-to-one trials, participants took longer with mated pairs with an age variation of several months compared to an age variation of the same day, but no significant difference was found on non-mated pairs over the two age variations. This difference in performance between the current study and Megreya et al's. (2013) one-to-one study may be due the different approaches used to select non-mated pairs in the studies with Megreya et al. (2013) using randomisation resulting in easier images pairs than the current

study. Furthermore, the Megreya et al. (2013) study used cropped greyscale images of adults and students as participants, whereas the current study used coloured non-cropped images of children and practitioners as participants, which could also be contributing factors to these different findings.

7.5.3 *Strategies Adopted by Practitioners to make Facial Comparison Decisions*

Facial comparison practitioners typically reported using multiple strategies throughout the study to help inform their decisions. Ferguson (2015) also found that her participants used different strategies and divided this into experience level of the participant. Many practitioners in the current study suggested they may make a holistic comparison first and then examine specific facial features if necessary. This strategy was also mentioned by practitioners in Ferguson's (2015) study. The main facial feature practitioners in the current study mentioned they used the most was the ears. Experts, participants with limited experience, and novices in Ferguson's (2015) study also mentioned using the ears for comparisons. Ears are known to be one of the most stable facial features for facial comparisons, particularly with images of adults (FISWG, 2016). However, the ears are not always visible in facial images, particularly with images of babies as taking images of babies generally requires taking the image very close up and, as a result, the ears cannot be seen. The importance of making the ear visible in document identification has been recognised in countries such as Kenya, where it is a requirement to have the ear visible in passport photos (Kenya Mission, 2015).

Some practitioners mentioned that they would compare markings such as moles or freckles. Facial marks were reported only by participants with limited experience in facial comparisons in Ferguson's (2015) study. This is another feature that is advantageous to use in comparisons with adults, however, becomes more problematic with images of children. Not all markings are present from birth and many may form over childhood and between image captures (Paller & Mancini, 2016). For example, it was noted during image selection for this study that an abundance of freckles may appear between image captures and that moles would also appear over time. It was also noted during image pair selection that eye colour could not help to discriminate images at younger ages as the eye colour sometimes changed. Research has

shown that changes in eye colour can occur, particularly before the age of 4 and at a reduced rate after this time (Matheny & Dolan, 1975).

Face shape was also mentioned as a way to help make facial comparison decisions by practitioners in the current study. Some facial comparison methods encourage practitioners to characterise images based on class characteristics such as the face shape (Spaun, 2011). However, research suggests this method does not work as there is low agreement when classifying face shape, both within and between novices even on the same images (Towler, 2014). Therefore, this strategy may not be effective for operational settings, although further research with practitioners should be considered. Face shape was also identified as a strategy to compare facial features in Ferguson's (2015) study, but only by participants who had no experience in facial comparisons.

Not enough is currently known about the stability of facial features in childhood. Facial comparison training in agencies typically involves a morphological approach that requires comparing individual facial features and forming conclusions. This approach is recommended by FISWG over other methods such as photo-anthropometry (FISWG, 2012). However, a morphological approach is likely easier with images of adults compared to children as adult's faces are more stable over time, whereas children's faces are still growing and developing. Furthermore, there are no international standards as to what specific facial features should be examined. FISWG are now addressing this issue for images of adults (FISWG, 2016). However, standards for adults are not likely to be relevant to inform standards involving images of children. The researcher is currently developing a stability of facial features in childhood document to address this lack of knowledge.

Typically, training focuses on images of adults with a small portion of training that involves images of children. According to C. Thomas who developed and trained practitioners in the Department of Immigration and Border Protection, training with images of children is mainly focused on highlighting those facial features used to compare images of adults that are less useful with images of children (personal communication, 16 September, 2016). More research is required to understand what facial features become stable the earliest in childhood so practitioners have a better chance of making accurate decisions by being able to identify similarities and differences in stable facial features. For example, the nose becomes

recognisable in its adult morphology by around 8 years of age (Wilkinson, 2012). Therefore, the nose may only become reliable as a facial feature at this time. Similarly, there may be ages when facial features should not be used for comparison purposes at all. For example, the nose undergoes a growth spurt between the ages of 1–6 years and can change considerably during this time (Farkas & Hreczko, 1994).

As previously mentioned, Yadav et al. (2014) found that it was not the whole face but just the binocular region presented to participants that resulted in the best performance for images in the age category (0–5, 6–10). This may suggest that this region of the face is most valuable for discriminating younger children. They also found that participants performed best (87% accuracy) on the (11–20, 21–30) age category when the full face was available. This may suggest that the strategy needs to change depending on whether images are of younger or older children.

Although collecting information regarding strategies adopted by facial comparison practitioners is useful, it is important to note that it is subjective. Therefore, a further recommendation is to conduct a study to empirically investigate whether practitioners are using the strategies they claim and to correlate which strategies result in higher performance. This can be achieved by conducting an eye tracking study to determine the strategies employed (Havard, 2007).

7.5.4 *Summary*

Even though facial comparison practitioners from many agencies conduct facial comparisons on images of children, very little was known about how practitioner performance is impacted by the age of children in images and the age variation between images. This study was conducted to provide empirical data to feed into Requirement 2 collected from agencies in Study 1. This study incorporated 120 facial comparison practitioners from a government agency and 23,760 unique image pairs carefully selected from an operational database. This makes this study the largest of its kind known to date. The aim was to gather data at the finest level possible, rather than arbitrarily grouping children into age groups of age variations that are not relevant to a wide audience. Conducting this study at such a low level has enabled trends and patterns in the data to be easily compared over the 198 different categories in an easy to understand reference table (i.e., a heat map data matrix). Data has also been separated

into pair types (i.e., mated and non-mated). This was so that agencies can better understand what pair type facial comparison practitioners find more difficult so that appropriate mitigating strategies and/or training can be implemented where necessary.

The data presented in this study can help agencies determine if performance is reliable enough to continue with current business process or additional contingencies need to be considered. These decisions will vary based on the agency, the application, and the feasibility of implementing new procedures.

The vast differences in performance with controlled images over age and age variation show that changes in the face throughout childhood have a significant impact on practitioner performance. Improvements in accuracy may be achievable with a better understanding of which facial features are most appropriate to compare across childhood. It is possible that different facial features should be compared with images of children than those currently taught for facial comparisons with images of adults or that different strategies may need to be adopted (Yadav et al., 2014). Identifying any stable facial features in childhood, or a better understanding of how the face changes in childhood, would be particularly useful and this information should be incorporated into facial comparison training programs in the future.

The next chapter of this thesis provides a mock example to demonstrate how the large amounts of data collected during Study 3A (Chapter 6) and 3B (current chapter) can be grouped to answer agency specific requirements. This was identified in Study 1 as Requirement 3 (determining facial comparison performance for agency specific requirements).

Chapter 8.

Study 4: Facial Comparison Performance for Agency Specific Requirements — Practitioner and Algorithm Study

8.1 *Introduction*

Study 2A (Chapter 4) and 3A (Chapter 6) provided a general examination of algorithm performance. Study 2B (Chapter 5) and 3B (Chapter 7) were designed to provide a general examination of facial comparison practitioner performance. These studies have yielded some interesting insights and provided information for Requirements 1 and 2 that were identified in Study 1. However, each agency also has their own unique requirements in regards to facial comparisons with images of children at specific ages and age variations. Past research examining age-related variables has analysed data by grouping it into several ages and/or age variations (Megreya et al., 2013; Valladares, 2012; White, Dunn, et al., 2015; Yadav et al., 2014; Zeng et al., 2012). Although grouping data is practical, these groups are often arbitrary leading to results that are difficult to compare or apply to operational applications. Agencies in Study 1 identified their own unique requirements in need of examination, with some ages and age variations in childhood requiring grouping to answer their specific questions.

As Study 3A and 3B were conducted at a low level by individual ages in childhood and age variations spanning 0–10 years on a large number of images, the data can be grouped to

provide data specific to agencies. Although each agency is different in regards to training, resources, and the type of images facial comparisons are conducted on, the data provided in this thesis and to agencies can be used to understand performance regardless of their particular composition. For example, as the images used in this study were controlled, results are highly relevant to processing and access control applications. Although investigative agencies in some instances may compare controlled images, the majority of images for comparison may be uncontrolled. The data can be still used to demonstrate the upper bound levels of performance and be used to understand how performance varies based on the age-related variables under investigation.

It is not the intention of this thesis to provide information to answer every agency requirement. Doing so would require discussing agency processes in more detail than permissible in this thesis. However, it seems prudent to demonstrate how the data collected in Study 3A and 3B can be used to answer agency specific questions that can be provided to agencies on a case-by-case basis. Thus, for the purposes of illustration for this thesis, a mock example is provided to demonstrate how the algorithm and facial comparison practitioner data can be tailored for agency specific requirements (Requirement 3).

8.1.1 *Mock Example*

The mock example involves the renewal period of children's passports. Currently in Australia, New Zealand, Canada, United States, and the United Kingdom, renewal of passports for children under 16 years of age occurs every five years (Passport Canada, 2012). However, a 10 year renewal period, consistent with adult renewals may be suitable for some ages in childhood, such as during the teenage years as some facial characteristics are defined by age 11 (Hunter et al., 2012). This includes the intercanthal width that is mature in Caucasian children by age 11 (Farkas & Hreczko, 1994) and integral to the performance of some facial recognition algorithms (Riopka & Boult, 2003; Wechsler, 2007). The aim of this mock example therefore, is to determine whether performance of algorithms and facial comparison practitioners differ when images of children from three age groups (0–4, 5–10, 11–15 years) are examined over a 5 and 10 year renewal period.

Much of the relevant research in this space has already been discussed in the previous algorithm study chapters (e.g., Grother et al., 2014; Grother et al., 2017; Ricanek et al., 2015;

Yiu et al., 2015) and facial comparison practitioner study chapters (e.g., Heyer, 2013; Lanitis, 2008; White, Kemp, Jenkins, Matheson, et al., 2014; White, Dunn et al., 2015; Zeng et al., 2012). As such, the following sections present information to describe the operational contexts that this mock example aims to provide data for, rather than repeating past research. Information regarding algorithms and facial comparison practitioners will be discussed separately.

8.1.1.1 *Algorithms*

In many countries, electronic gate (eGate) systems, such as SmartGate, are implemented in international airports in order for travellers to self-process through passport control. The SmartGate system relies on the information in ePassports and a facial recognition algorithm to conduct checks that would normally be done by a facial comparison practitioner, which in this application, would be a Border Force officer (DIBP, n.d.-a). In Australia, the minimum age to use this system was originally 18 years (Sharma, 2008), however over time, this age has reduced and currently Australian children as low as 10 years of age, who are accompanied by at least two adults, can access SmartGate when arriving back into Australia (DIBP, n.d.-a). The age is even lower when departing Australia, as children of any age can use the system as long as they can follow the process without help (DIBP, n.d.-c).

Data from this study is valuable to understand how these state-of-the-art algorithms would differ by the age groups selected (0–4, 5–10, 11–15 years) if they were implemented into an eGate system based on current policies (5 year renewal period) and if policies changed in the future (10 year renewal period). Determining the performance of state-of-the-art algorithms is also important because even though policies now allow children to use SmartGate at younger ages, this could result in higher false non-match rates and unsuccessful attempts to enter through the gate. This would lead to children and their parents having to queue in the manual processing line. Alternatively, a child may successfully pass through the gate on someone else's passport, either due to human error (i.e., a parent handed out the passports to the wrong family members) or due to fraud, as in the case of child trafficking. It is also important to determine algorithm performance for each age group (0–4, 5–10, 11–15 years) and renewal period (5 and 10 years) when the false match rate is set at 0.001, as may be experienced in operational applications.

Some airports hope to rely on technology and eradicate manual processing as much as possible (Planet Biometrics, 2017). Although this would eliminate human errors that may occur by border staff, such as the reported incident where the passport of a toy unicorn was stamped rather than the passport of a child (Robinson, 2013), a lack of human interaction has others concerned. Facial comparison practitioners have an inherent ability to pick up on things that automated systems currently cannot. For example, in 2016 there was concern in the United Kingdom about children using eGate systems because automated systems cannot pick up on children travelling under duress, such as those being trafficked, which may be evident to a facial comparison practitioner (Barrett, 2016).

8.1.1.2 *Facial Comparison Practitioners*

Facial comparison practitioners in a passport processing role make image pair comparisons when, for example, a traveller renews their passport (White, Kemp, Jenkins, Matheson, et al., 2014). The purpose is to check for any fraudulent applications or human errors. Identity fraud can occur if a person assumes the identity of someone else to apply for a passport. This may be in order to conceal their identity (such as terrorists or fugitives), illegally enter a country or avoid deportation, or to facilitate a crime such as drug trafficking or alien smuggling (U.S. Department of State, n.d.). Human error in relation to passport images may occur during the application process, for example, if a mother completes passport applications for two of her children, but accidentally attaches the images of one child on the other child's application.

Due to the purpose of conducting facial comparisons in a passport processing role, it is important to understand how practitioner performance differs by pair type (i.e., on mated and non-mated pairs). For example, if a facial comparison practitioner incorrectly determined that mated images were of different people, this could result in inconveniencing genuine customers who may need to provide additional information to prove they are who they claim to be. If a practitioner incorrectly determined that non-mated images were of the same person, this could result in cases of identity fraud or human error being processed.

Facial comparison practitioners may make incorrect decisions on mated pairs due to how much an individual child has changed over time. The implications of such an incorrect decision to genuine travellers can vary on a case-by-case basis and although they are generally considered just a nuisance, they can be more serious in some circumstances. For example in

2015, French police detained a six year old girl over suspicions she was carrying a fake passport. She was carrying all necessary documentation with her at the time and her mother was waiting to collect her at the airport, however, the mother ended up waiting for three days because police did not believe the child resembled her passport image. The concern was that since she did not resemble her image, she may be a victim of kidnapping or human trafficking (The Guardian, 2015).

Although inconveniencing genuine customers would be a nuisance, fraudulent images being processed could have severe implications for the children involved. For example, Zhang (2007) highlighted that human traffickers may move children across borders and through countries by purchasing expired documents, such as outdated passports from the black market. These traffickers purchase documents where the images resemble the child they wish to transport. A new passport is then submitted for renewal using the expired documents of someone else but using a recent image of the trafficked child. If this is not picked up during the passport application processes, it could result in a child being sold into slavery and never being found again. As such, a passport processing agency may consider it necessary to understand if there was a higher risk of such potential occurrences if there was contemplation to extend the passport renewal period for children.

The facial comparison practitioner data in this study can also be used to inform potential performance of practitioners working at the border. Calic (2012) found that image-to-image and image-to-live-face comparisons result in the same overall performance when conducted by practitioners (although accuracy on image-to-image comparisons was slightly higher with mated pairs and slightly lower with non-mated pairs). Therefore, the results for this requirement would also be beneficial to inform border processing staff who, rather than comparing two images, need to compare the passport image of a child to the child bearing that passport. In the case of identity fraud at the border, this can occur if the trafficked child is using another child's passport that is similar looking in appearance.

8.2 *Research Questions*

The aim of the current study was to demonstrate an applied use of the data captured during Study 3A and 3B to answer agency specific requirements (Requirement 3). This was achieved via five research questions.

Question 1. Does algorithm performance with images of children at ages 0–4, 5–10, and 11–15 differ when the passport renewal period is extended from 5 to 10 years?

Question 2. Does algorithm performance with images of children at ages 0–4, 5–10, and 11–15 differ when the passport renewal period is extended from 5 to 10 years based on the type of image pair presented (i.e., mated or non-mated)?

Question 3. Does the false non-match rate of algorithms with images of children at ages 0–4, 5–10, and 11–15 differ at a set false match rate of 0.001 when the passport renewal period is extended from 5 to 10 years?

Question 4. Does facial comparison practitioner performance with images of children at ages 0–4, 5–10, and 11–15 differ when the passport renewal period is extended from 5 to 10 years?

Question 5. Does facial comparison practitioner performance with images of children at ages 0–4, 5–10, and 11–15 differ when the passport renewal period is extended from 5 to 10 years based on the type of image pair presented (i.e., mated or non-mated)?

It was hypothesised that algorithm and practitioner performance would worsen as the renewal period increased from 5 to 10 years. In addition, the poorest performing age group would be 0–4 years. This was based on the results from previous chapters in this thesis which were likely due to the amount of change occurring across longer age variations in childhood, as well as the lack of distinguishing facial features in the 0–4 year age group compared to the 5–10 and 11–15 year age groups (Kozak et al., 2015).

It was also expected that algorithm performance would be impacted more by mated than non-mated pairs and that the false non-match rates for each algorithm would increase if the renewal period was extended.

It was also anticipated that practitioners would perform poorer at a renewal period of 10 years, possibly due to their assumptions as to what were 'possible and permissible' facial changes due to age variations.

8.3 *Methodology*

The data for this study was collected during Study 3A (Chapter 6) and Study 3B (Chapter 7) and was grouped where necessary to provide data for this mock example. As such, the methodology for collecting the data has already been discussed (see Sections 6.3 and Section 7.3) and will not be repeated here.

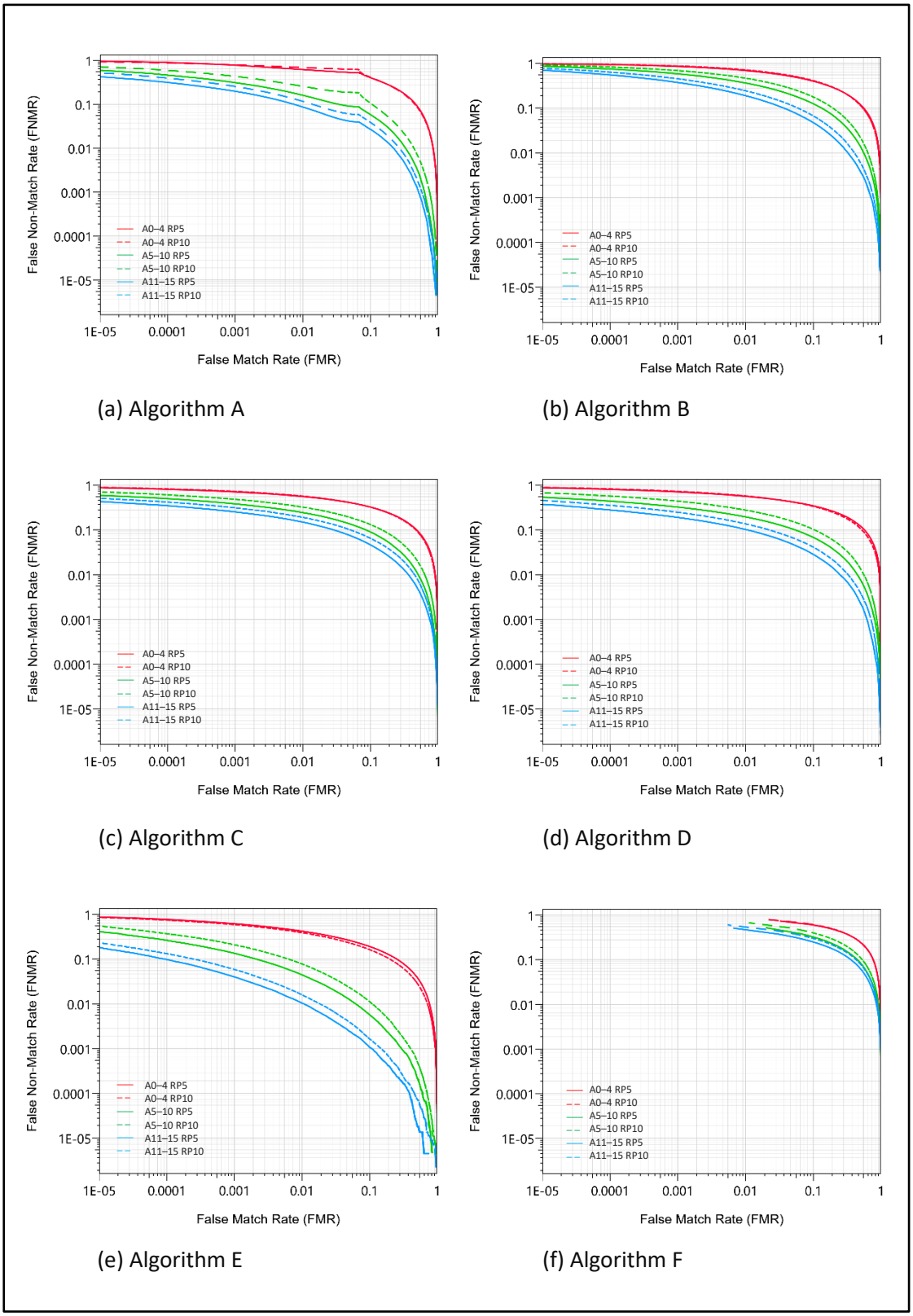
Children in the 0–4 year age group included any image pairs where the youngest child in a pair was aged between 0 years 0 months and 4 years 11 months. The 5–10 year age group included any image pairs where the youngest child in a pair was aged between 5 years 0 months and 10 years 11 months. The 11–15 year age group included any image pairs where the youngest child in a pair was aged between 11 years 0 months and 15 years 11 months. Renewal periods are synonymous with age variations. In this example, a renewal period of 5 years included all image pairs where the age variation was less than 5 years (i.e., between 0 years 0 months and 4 years 11 months). A renewal period of 10 years included all image pairs where the age variation in an image pair was from 5 years and above (i.e., between 5 years 0 months and 9 years 11 months).

8.4 *Results*

The results are divided into five main sections based on the five research questions: algorithm performance, algorithm performance based on pair type, algorithm performance based on a set false match rate of 0.001, overall facial comparison practitioner performance, and facial comparison practitioner performance based on pair type.

8.4.1 *Algorithm Performance by Age Group and Renewal Period*

Figure 48 shows the DET plots which display the overall performance of each of the six algorithms on the six different groups examined (i.e., images of children from 0–4, 5–10, and 11–15 year age groups with passport renewal periods of 5 and 10 years).



(a) Algorithm A

(b) Algorithm B

(c) Algorithm C

(d) Algorithm D

(e) Algorithm E

(f) Algorithm F

Figure 48. DETs displaying performance of each algorithm across the six groups (A = age, RP = renewal period).

For all algorithms, the performance was similarly poor for the youngest age group of 0–4 years, regardless of the renewal period. However, algorithm performance with the older two age groups of 5–10 and 11–15 years was better when the renewal period was 5 years, rather than 10 years. These patterns were consistent for all algorithms. The performance of OpenBR (Algorithm F) on the two renewal periods was relatively similar and was also the worst performing across all age groups due to the poor matching capability of this algorithm. The DETs also show that algorithm performance is impacted more by the age group than the renewal period.

8.4.2 *Algorithm Performance by Age Group and Renewal Period on Mated and Non-Mated Image Pairs*

Figure 49 shows the cumulative probability plots that display for each of the six algorithms, the individual false match rate and false non-match rate performance across the six different groups.

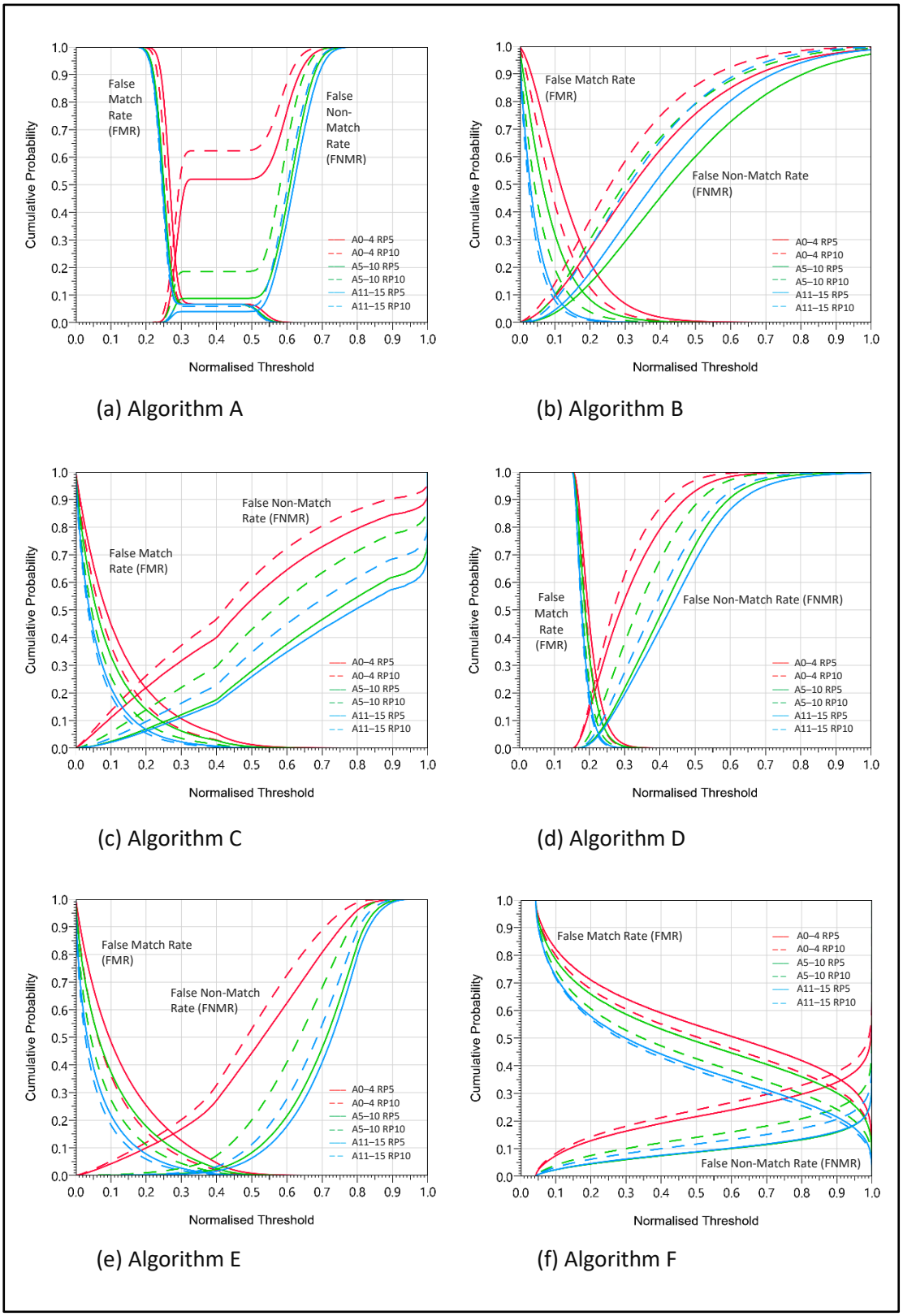


Figure 49. Cumulative probability plots displaying performance of each algorithm across the six groups (A = age, RP = renewal period).

The plots show substantial differences in algorithm performance between algorithms and within algorithms for both mated and non-mated pairs. For example, there was minimal difference in false match rate performance with non-mated pairs between the six groups for Algorithms A and D. This suggests that non-mated pairs processed using Algorithm A or D would be minimally impacted by the age group and renewal period. Algorithms B, C, E, and F showed more variance between the false match rates for the six groups. For each age group (0–4, 5–10, 11–15 years), the false match rate performance improved when the renewal period was extended from 5 to 10 years. In many instances, the performance in terms of false match rate was highest for the youngest age group (0–4 years) with a 5 year renewal period, followed by the youngest age group with a 10 year renewal period, then the middle age group (5–10 years) with a 5 then a 10 year renewal period. Finally, the oldest age group (11–15 years) with a 5 then a 10 year renewal period had the lowest false match rates, indicating that algorithms are better at distinguishing children from this older age group.

The plots also show that the spread of the false non-match rate increased for each algorithm when the renewal period was extended from 5 to 10 years for each age group. For all algorithms, there was greater variation in the false non-match rate than false match rate data suggesting that algorithm performance over the six groups varied more based on mated pairs than non-mated pairs. Algorithms typically performed poorest on the 0–4 year age group with a renewal period of 10 years. This suggests that non-mated pairs were impacted more by the age of children in images. However, in many instances, the false non-match rate seemed to be poorer, with the 10 year renewal period of the 11–15 year age group producing higher false non-match rates than the 5–10 year age group with a 5 year renewal period. This suggests that mated pairs of teenagers with larger renewal periods were more difficult to be recognised than the mated pairs from the 5–10 year age group with shorter renewal periods.

8.4.3 *Algorithm Performance by Age Group and Renewal Period based on set False Match Rates*

Figure 50 displays bar charts to show how the false non-match rates varied when the false match rate was set at 0.001 for the three different age groups when the renewal period was 5 and 10 years.

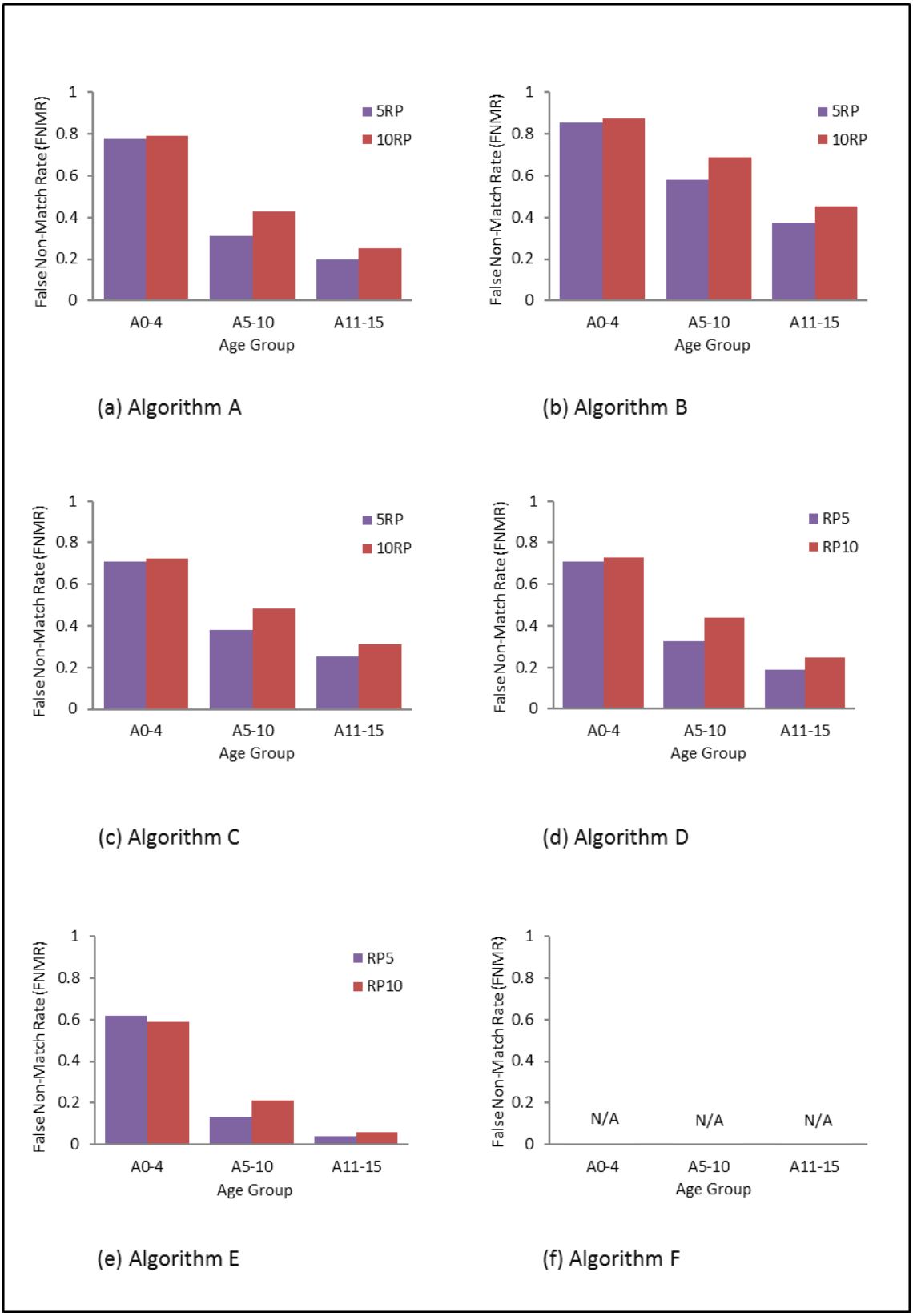


Figure 50. False non-match rate for the six groups based on a false match rate of 0.001 (A = age, RP = renewal period).

The bar charts show that the algorithms had consistently higher false non-match rates for the 10 year renewal period compared to the 5 year renewal period. This was consistent for all age groups and all algorithms, apart from Algorithm E. For this algorithm, the 5 year renewal period for the 0–4 year age group had a slightly higher false non-match rate of 0.618 compared to the 10 year renewal period of 0.590.

The charts also show that performance varied between algorithms and that some performed better at some age groups than others. For example, Algorithm A had a higher false non-match rate of 0.778 for the 0–4 year age group at a 5 year renewal period than Algorithm C with a false non-match rate of 0.710. However, Algorithm A had a lower false non-match rate (FNMR = 0.311) than Algorithm C (FNMR = 0.382) for the 5–10 year age group for the 5 year renewal period. Algorithm A also had a lower false non-match rate (FNMR = 0.197) than Algorithm C (FNMR = 0.252) on the 11–15 year age group for the 5 year renewal period. No false non-match rate data was available for Algorithm F (OpenBR) at a false match rate of 0.001 as performance was too poor.

8.4.4 *Practitioner Performance by Age Group and Renewal Period*

Figure 51 presents the descriptive statistics for overall facial comparison practitioner accuracy for the six different groups selected for this mock example.

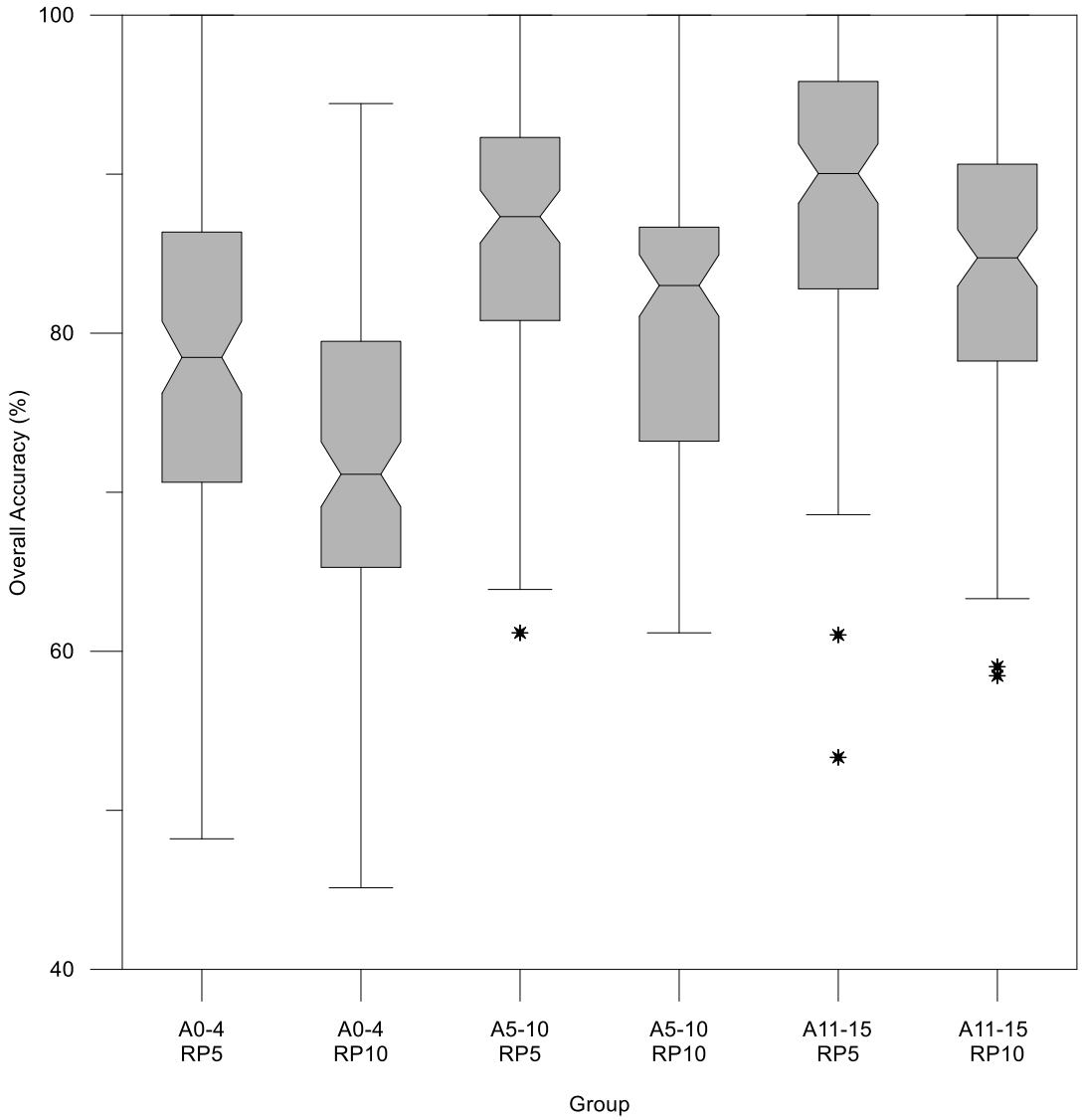


Figure 51. Overall practitioner accuracy for different groups (A= age, RP = renewal period).

As can be seen from the boxplots, accuracy was lower on images with a 10 year renewal period compared to those with a 5 year renewal period for the same age groups (0–4, 5–10, 11–15 years). A Wilcoxon Signed Rank test was conducted with a Bonferroni correction applied and showed that each of these was statistically significant. As with the previous Wilcoxon Signed Rank tests conducted in Study 2B, this study also reports the mean accuracy data along with the median data typically reported in non-parametric statistical tests for ease of comparison with past studies.

Overall accuracy for the 0–4 year age group dropped by 5.35% from a 5 year (*M* = 77.67%, *Mdn* = 78.48%) to a 10 year (*M* = 71.56%, *Mdn* = 71.13%) renewal period, *z* = -5.51, *p* < .001, *r* = -.36. Overall accuracy for the 5–10 year age group dropped 4.33% from a 5 year (*M* = 86.22%, *Mdn* = 87.33%) to a 10 year (*M* = 80.47%, *Mdn* = 83%) renewal period, *z* = -5.98, *p* < .001, *r* = -.39. Overall accuracy for the 11–15 year age group dropped 5.3% from a 5 year (*M* = 88.60%, *Mdn* = 90.04%) to a 10 year (*M* = 83.76%, *Mdn* = 84.74%) renewal period, *z* = -5.29, *p* < .001, *r* = -.34. These effect sizes are considered moderate.

8.4.5 *Practitioner Performance by Age Group and Renewal Period on Mated and Non-Mated Image Pairs*

Figure 52 displays the facial comparison practitioner accuracy results for the six different groups based on pair type (i.e., mated or non-mated).

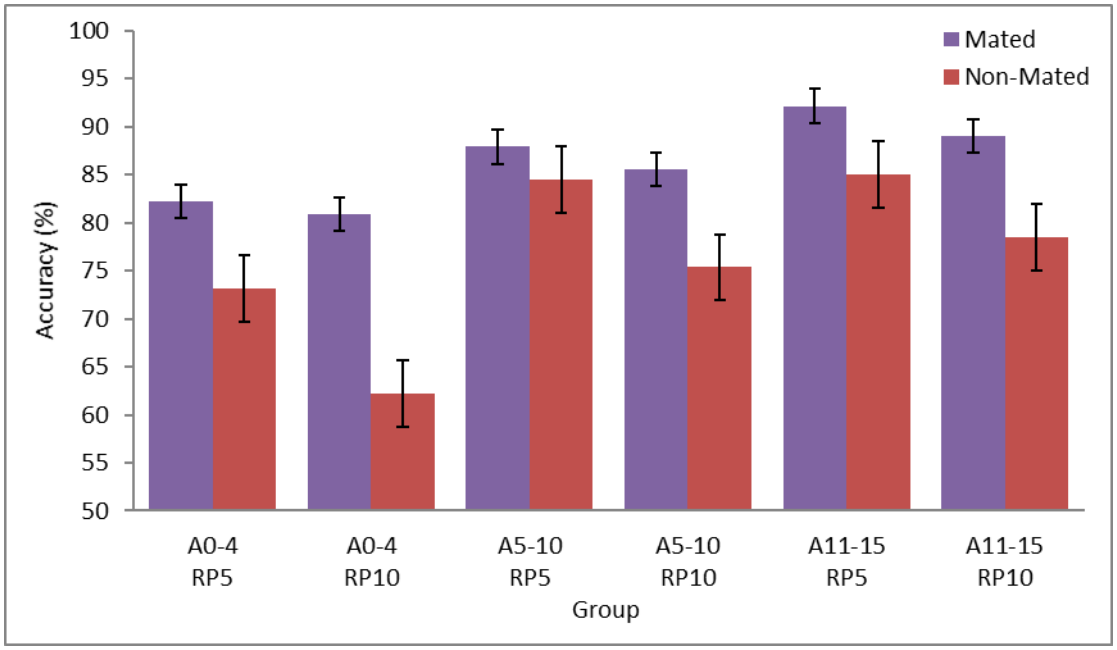


Figure 52. Practitioner accuracy (± 1 standard error) for different groups based on pair type (A = age, RP = renewal period).

The bar chart shows that facial comparison practitioners performed consistently better on mated images than on their non-mated counterparts. Mated accuracy stayed relatively consistent over each age group (0–4, 5–10, 11–15 years), regardless of renewal period being

5 or 10 years. No age group had a statistically significant difference in accuracy for mated pairs between the two renewal periods.

However, there was a larger difference between the 5 and 10 year renewal periods for each age group on non-mated image pairs. A Wilcoxon Signed Ranks test with a Bonferroni correction was conducted to determine if there was a significant difference between renewal periods of 5 and 10 years for all three age groups (0–4, 5–10, 11–15 years) based on pair type. Statistically significant differences over the two renewal periods were found for non-mated pairs for each age group. Accuracy on non-mated pairs for the 0–4 year age group dropped by 11.36% from a 5 year (*M* = 73.18%, *Mdn* = 75%) to 10 year (*M* = 62.21%, *Mdn* = 63.64%) renewal period, $z = -6.08$, $p < .001$, $r = -.39$. Non-mated accuracy for the 5–10 year age group dropped 8.50% from a 5 year (*M* = 84.54%, *Mdn* = 86.67%) to a 10 year (*M* = 75.36%, *Mdn* = 78.17%) renewal period, $z = -6.05$, $p < .001$, $r = -.39$. Non-mated accuracy for the 11–15 year age group dropped 7.93% from a 5 year (*M* = 85.03%, *Mdn* = 88.56%) to a 10 year (*M* = 78.48%, *Mdn* = 80.63%) renewal period, $z = -4.87$, $p < .001$, $r = -.31$. These effect sizes are considered moderate.

Table 11 provides the descriptive statistics and pairwise comparisons for the measures of discrimination and bias for each age group (0–4, 5–10, 11–15 years) over the two renewal periods (5 and 10 years).

Table 11. *Discrimination and Bias for Each Group*

	M (SD)	Median	Min – Max	Pairwise Comparisons
A' Discrimination				
Age 0–4				
Renewal Period 5 Years	.65 (.10)	.63	.50 – 1	z = -5.76, p < .001, r = -.37
Renewal Period 10 Years	.60 (.08)	.57	.49 – .87	
Age 5–10				
Renewal Period 5 Years	.75 (.11)	.75	.52 – 1	z = -5.87, p < .001, r = -.38
Renewal Period 10 Years	.68 (.11)	.66	.52 – 1	
Age 11–15				
Renewal Period 5 Years	.79 (.12)	.79	.50 – 1	z = -5.28, p < .001, r = -.34
Renewal period 10 Years	.72 (12)	.71	.51 – 1	
B" Bias				
Age 0–4				
Renewal Period 5 Years	-.15 (.54)	-.13	-1 – 1	z = -2.31, p = .021, r = -.15
Renewal Period 10 Years	-.27 (.49)	-.21	-1 – 1	
Age 5–10				
Renewal Period 5 Years	-.15 (.67)	-.15	-1 – 1	z = -1.78, p = .074, r = -.11
Renewal Period 10 Years	-.27 (.49)	-.21	-1 – 1	
Age 11–15				
Renewal Period 5 Years	-.26 (.77)	-.46	-1 – 1	z = -0.42, p = .676, r = -.03
Renewal Period 10 Years	-.31 (.65)	-.33	-1 – 1	

A significant difference was found in discrimination between the renewal periods of 5 and 10 years for each age group. This suggests that facial comparison practitioners were better at discriminating between mated (i.e., 'same') and non-mated (i.e., 'different') pairs when the renewal period was 5 years compared to when the renewal period was 10 years for each age group.

As the Bonferroni correction was applied, this required a p < .017. As such, there was no significant difference in bias between renewal periods of 5 and 10 years for any age group. This suggests that facial comparison practitioners were not biased to claim images with

renewal periods of 10 years were of the 'same' person more than they would with images of renewal periods of 5 years.

8.5 *Discussion*

The aim of this study was to show how the data generated in earlier chapters could be useful to agencies by providing empirical evidence for determining facial comparison performance for agency specific requirements (Requirement 3) identified in Study 1. Agency specific requirements cannot be discussed here due to privacy reasons; however this example provides a useful understanding of how valuable this data is and how it can be adapted to support agency needs. A discussion of the results for the algorithms and facial comparison practitioners is presented next.

8.5.1 *Algorithms*

The DETs showed that performance was similar on the 5 and 10 year renewal period for the 0–4 year age group for most algorithms. As performance for this age group was poor compared to the other two age groups, this performance may be explained due to children at this young age having less individuating facial features. Therefore, age variation does not impact on performance as drastically for this age group because performance is already poor and there are no individuating facial features available for comparison regardless of age variation. All algorithms showed performance improvements for the 5–10 and 11–15 year age groups on both the 5 and 10 year renewal periods compared to the 0–4 year age group. Performance for these two older age groups degraded when the renewal period was 10 years compared to 5 years. One explanation may be that the older the child, the more individuating facial features are present and therefore algorithm performance is better. However, larger age variations will still impact on performance due to the amount of facial change occurring over longer periods of time. One vendor mentioned during Study 1 that their algorithm relies on pattern and shape information. This supports the theory that facial changes are problematic for the algorithms.

The cumulative probability plots showed that algorithm performance was typically spread out for false match rate and false non-match rate data for the six groups under investigation (apart from the false match rate data for Algorithm A and D). This suggests that algorithm performance was impacted by both mated and non-mated pairs. The larger spread in the false

non-match rate data suggests that there was more variability in algorithm performance with mated pairs over the different groups than non-mated pairs. The plots show that the younger the age group, the poorer the performance on mated pairs. This finding suggests that the younger children are, the more the algorithms struggle to recognise children as themselves. The false match rate data also varied between the six groups (although Algorithm A and D results were less spread out suggesting less difference in performance over the six groups). The spread in the data and the poorer performance on the 0–4 year age group followed by the 5–10, then 11–15 year age groups suggests that the younger children are, the more difficult they are to distinguish from other children. The FRVT 2013 (Grother & Ngan, 2014) also found that younger children were both more difficult to recognise as themselves, but also more difficult to distinguish from others.

The false match rates were poorest with the 0–4 year age group with a 5 year then a 10 year renewal period, 5–10 year age group with a 5 then a 10 year renewal period, followed by 11–15 year age group with a 10 then a 5 year renewal period. This finding is perhaps due to younger children all looking alike, with minimal individuating features. Some facial growth patterns occur at around the same ages for most children (Ferrario, Sforza, Poggio, & Schmitz, 1998). Therefore, at a 10 year renewal period, there was more variance due to up to 10 years of facial growth occurring, causing children in image pairs to look more different than at the 5 year renewal period. As the age group advances, more individuating facial features form and so performance gradually improves. The false match rate was lowest for the 11–15 year age group at a 10 year renewal period. This could be due to the more defining facial features having developed around this age (Hunter et al., 2012). It could also be due to larger age variations allowing more time for typical facial growth patterns to progress.

The results for the false non-match rate when the false match rate was set at 0.001 were high for all algorithms, but particularly for the 0–4 year age group, both at a 5 and a 10 year renewal period. Given that security applications try to keep the false match rate as low as possible so that, for example, identity fraud occurs very infrequently, high levels of genuine people would be impacted as a result. For example, if a system is set so that false match rates occur on average 0.1% of the time (FMR = 0.001), Algorithm C would on average only allow 29% (FNMR = 0.710) of genuine children to successfully pass through an eGate system if they were between 0–4 years of age with a 5 year renewal passport.

As hypothesised, false non-match rates were higher for a renewal period of 10 years for each age group compared to the 5 year renewal period. However performance for the 0–4 year age group for the 5 and 10 year renewal period was similar, but very poor, regardless of renewal period. Performance was so poor again for OpenBR that there was no data available at the false match rates of 0.001. The best performing algorithm was Algorithm E. For the 11–15 year age group with a 5 year renewal period, the false non-match rate was 0.040 at a false match rate of 0.001. If the renewal period was extended to 10 years, the false non-match rate would be 0.059.

Although it would depend on what operational threshold an algorithm is set for, the results show that algorithms tested in this thesis (at a false match rate of 0.001) are not good enough to be used (for the purposes of this mock example) with images of children aged 0–4 years when there is a 5 or 10 year renewal period. As this mock example is for a national security application, error rates need to be low and will depend on the agency's willingness to accept any risk (including disruption) associated with the decisions. For example, depending on how many genuine travellers border agencies were willing to inconvenience if they failed to be accepted at an eGate, an agency may be willing to allow children aged 11–15 years to use an automated system if they were using one of the higher performing algorithms. Indeed, in Australia, children in this age group are already processing through SmartGate (DIBP, n.d.-a; n.d.-c). However, algorithm performance would degrade if the passport renewal period was extended to 10 years. Although, performance at a renewal period of 10 years is better than performance found for the two younger age groups (0–4, 5–10 years) with a 5 year renewal period. This then raises the question as to whether younger children require passport renewal periods shorter than 5 years or if the age of children using SmartGate when departing Australia should be raised until algorithm performance improves, if the algorithm used in SmartGate performs similarly to those evaluated here. This will depend on a number of factors that are outside the scope of this research.

It is important to note that this study was undertaken using controlled operational images typical of those found in a passport. However, images acquired at eGates such as SmartGate are not so controlled and can be impacted by lighting issues in the background (Wayman, 2016). In addition, this mock example was solely conducted to evaluate algorithm performance and an operational trial that incorporates all the additional variables that impact

on performance, such as human factors and lighting issues, would be the best approach to gain a true understanding of performance. Nevertheless, this mock example does highlight that even in controlled conditions, performance based on the age of the child in an image and the length of renewal period will indeed impact on algorithm performance. In an operational scenario, performance (with these particular state-of-the-art algorithms) can be expected to degrade further than that reported here.

The performance of every algorithm declined when the renewal period was extended from 5 to 10 years. Any proposal to extend the passport renewal period beyond 5 years without consideration of the impact on performance is not supported by this study. This study also highlights the importance of understanding how much state-of-the-art algorithms can differ in performance. It was shown that some algorithms were better than others for some groups in childhood but not others. Agencies that currently have algorithms implemented into their processes should keep in mind that the algorithm they have may not necessarily be the best performing algorithm for their needs.

Although this was a mock example, trials are underway in airports that allow travellers to use facial recognition technology from check-in to boarding without passports or other travel documents (Planet Biometrics, 2017). Implementation of such technology would require careful consideration, particularly in regards to the accuracy of algorithms used within such systems with young child travellers. This current study and Study 3A (Chapter 6) provide a good foundation to begin to understand state-of-the-art algorithm performance with children throughout childhood and with different renewal periods for this document-free application.

8.5.2 *Practitioners*

Results showed that overall accuracy for each age group was statistically significant over the two renewal periods of 5 and 10 years. As hypothesised, practitioners were significantly less accurate with images from a 10 year renewal period than images from the 5 year renewal period for all corresponding age groups (0–4, 5–10, 11–15 years). However, once the images were divided into pair type, it was evident that this significant difference was due to performance on non-mated pairs (10% difference) with performance on mated pairs staying relatively consistent over the two renewal periods (1% difference). Therefore, hypothetically, if passport renewal periods for children increased from 5 to 10 years, it may have minimal

impact on genuine customers. However, those trying to obtain a fraudulent child passport may be more successful if the validity period was extended from 5 to 10 years. Thus, in applications where a fraudulent passport of a child may be beneficial, such as in kidnapping or human trafficking cases where it may be necessary to transport victims across borders (Zhang, 2007), it is likely that it would be more difficult for passport processing staff and border staff to identify these cases.

It is also important to note that there was a considerable amount of variability in practitioner performance within each age group, but particularly in the 0–4 year age group at both a 5 and 10 year renewal period as shown in Figure 51. Practitioners that participated in this research do not currently conduct facial comparisons with images of children from this age group. This may explain some of the degradation in performance (i.e., lack of experience conducting facial comparisons with children at these ages), but it is more likely a reflection of the difficulty of conducting facial comparisons on children who lack distinguishing features. Training and testing with images of children from this age group may help to reduce the variability in practitioner performance to some extent. Alternative measures such as not relying on the face and conducting additional verification checks may also be considered as more appropriate (which is the current approach taken by the agency that participating practitioners were from).

This mock example highlights that rather than just presenting overall performance results, future research evaluating practitioner performance should also present mated and non-mated data separately. As there are different consequences for different agencies, depending on whether performance degrades on mated or non-mated pairs, it is critical to present data at this level so agencies are better informed and in turn, can better inform their business processes based on research findings.

8.5.3 *Summary*

This study was conducted to provide empirical data to feed into Requirement 3 (determining facial comparison performance for agency specific requirements). This study demonstrated the value and flexibility of the data collected during Study 3A and 3B to inform agency specific requirements. Although the example here was for a passport processing application, using good quality images, the data can be used by a range of agencies to demonstrate the upper bound levels of performance, and to understand how performance varies based on the

age-related variables under investigation. Such results can be used to inform agencies about whether changes in processes are practical from risk and performance perspectives, which algorithms may be most appropriate for their needs, and/or where to focus training for facial comparison practitioners.

Next, the discussion chapter will summarise the key findings from the six studies conducted in this thesis. Conclusions and recommendations are presented that aim to provide information of value to agencies, algorithm vendors, and researchers.

Chapter 9.

Summary, Recommendations, and Conclusion

9.1 *Introduction*

The overarching aim of this research was to determine the performance of algorithms and practitioners conducting facial comparisons with images of children. The focus was on age-related variables (i.e., age and age variation) that were expected to impact on performance given the amount of facial growth that occurs throughout childhood. To ensure the research was as operationally relevant as possible, Study 1 was designed to collect requirements directly from agencies. Four requirements were identified and examined in this thesis.

Requirement 1: Determining facial comparison performance with images of children and adults.

Requirement 2: Determining facial comparison performance with images of children at different ages and age variations.

Requirement 3: Determining facial comparison performance for agency specific requirements.

Requirement 4: Determining facial comparison performance with mated and non-mated images of children and adults (as examined as part of Requirements 1–3).

Requirement 5 (determining age estimation performance with images of children and adults) and Requirement 6 (determining the performance of age progressed images for facial comparison) were out of scope for this thesis, but will be discussed in Section 9.6 as part of future research recommendations.

Study 1 also enabled a better understanding of the practices and procedures followed when conducting facial comparisons within national security agencies in Australia and overseas so that the most appropriate paradigm could be investigated. It was determined that one-to-one facial comparisons would be the focus of this research as it was the most common undertaken by agencies. Furthermore, conducting one-to-one research provided the upper bound levels of performance as additional images that are displayed in a one-to-many paradigm were not present. Several million controlled operational images were provided for this research. Controlled images provided the upper bound levels of performance. Using controlled images also ensured age-related variables were examined as much as possible in isolation from other variables known to impact on performance. Based on the requirements collected from agencies and the research gaps identified in the literature, Study 2A, 2B, 3A, 3B, and 4 were developed.

State-of-the-art algorithms and facial comparison practitioners were evaluated under the same age-related conditions, but different approaches in methodology were necessary. This was to ensure that the methodology adopted for the algorithm studies followed best practices from previous algorithm evaluations (Grother et al., 2011; Yiu et al., 2015) and the facial comparison practitioner studies followed best practices from psychology (Burton, 2013; Calic, 2012; Heyer, 2013; White, Dunn, et al., 2015). This was whilst tailoring the methodologies to suit operationally relevant expectancies and using the largest number of samples to increase statistical significance of the results. Although OpenBR (Algorithm F) was tested to provide results from an algorithm that is publicly available, performance was often found to be too poor on the conditions evaluated within this thesis.

Typical performance measures from past algorithm and facial comparison practitioner studies were preserved. However, novel approaches were also incorporated into presenting the data in such a way that agencies could easily read and understand the results by using a simple one page figure to represent large amounts of data (i.e., a heat map data matrix). These heat map data matrices have been shown to several agencies and vendors and have been well received due to their ease of comprehension. They have also been presented effectively at conferences (Michalski et al., 2016) and recently by NIST in their latest FRVT reports (Grother et al., 2017). The matrices can be used as a reference table to guide decisions including where to direct additional mitigating strategies and training needs.

The methodology adopted for Study 3A and 3B was also carefully designed so that the results could be tailored specifically for individual agencies. Examples of how these studies could be tailored were provided in Chapter 8. For some agencies, the results were representative of the levels of performance of these algorithms and/or facial comparison practitioners in their agency under the conditions evaluated. For other agencies that work with uncontrolled images, the results provide insight into what the upper bound level of performance would likely be, making it beneficial to guide future decisions based on how age impacts performance throughout childhood and at different age variations. The data may also help agencies currently without facial recognition algorithms to determine if any of these algorithms will be of value for their needs. Algorithm vendors also benefit from the large-scale studies conducted with their algorithms in this thesis. The data provides vendors with insight into areas requiring additional improvement which can guide further development of their algorithms, with the potential to improve performance in future versions.

This final chapter presents a summary of the key findings for each of the requirements addressed and limitations of the research. Next, a series of recommendations for agencies conducting facial comparisons with algorithms and/or practitioners is provided. The recommendations also include suggestions for further research, as well as current research underway by the author to address additional knowledge gaps within this space. A short discussion is also presented to validate the contribution of the research to the field of facial comparison and to the national security agencies who currently make decisions predominantly based on anecdotal information. The chapter then ends with the overarching conclusion of the research.

9.2 *Key Findings: Facial Comparison Performance with Images of Children and Adults*

The impact of facial comparisons with images of children compared to images of adults was examined during Study 2A with algorithms and Study 2B with practitioners. Four key findings will be discussed.

Key Finding 1: Using the same operational threshold for images of children and adults is not optimal for algorithm performance.

Some agencies may be using the same operational threshold for children and adults. This may be due to function creep where a facial recognition system was implemented for use with adults but over time has also been required for use with children. It is also possible that an agency may have not considered using more than one threshold. It is important for agencies to understand that having one operational threshold for both adults and children can have an impact on performance levels with children, although the extent of this degradation varied between state-of-the-art algorithms. Table 12 presents the false non-match rate (and false match rate) performance of each algorithm with images of children at the false match rate of 0.001 based on images of adults and children respectively.

Table 12. *Algorithm Performance at a set False Match Rate of 0.001 based on Images of Adults or Children*

	Adult FMR of 0.001		Child FMR of 0.001
Algorithm	Child FMR	Child FNMR	Child FNMR
A	0.001	0.429	0.454[14]
B	0.005	0.574	0.728
C	0.003	0.424	0.485
D	0.006	0.340	0.447
E	0.002	0.235	0.260
F	N/A	N/A	N/A

When an overall threshold is used based on a threshold set at a false match rate of 0.001 with adult images, the false match rate for children was found to vary between 0.001 and 0.006 for the algorithms tested. For example, based on a false match rate of 0.001 with adult images, the best performing algorithm (Algorithm E) on average would falsely match non-mated pairs of children 0.2% of the time (FMR = 0.002) and correctly match mated pairs 76.5% of the time (FNMR = 0.235). If Algorithm E was based on a child's false match rate of 0.001, it would correctly match mated pairs 74% of the time (FNMR = 0.260). Therefore, using a threshold relevant for adults, although a 2.5% gain in true match rate performance with mated pairs is expected compared to a separate setting used for children, the number of non-mated pairs of children that would be incorrectly accepted by the system would be double. This would pose a problem in national security contexts since the aim is to keep the false match rate as low as possible to minimise the number of fraudulent users. Thus, for Algorithm D, if its threshold was set at an adult false match rate of 0.001, it would result in a false match rate of children that is six times higher, which may not be considered acceptable by the agencies.

[14] Although the Child FMR is also 0.001 at an Adult FMR of 0.001, the Child FNMR columns show different results. This is due to the rounding of the data.

Key Finding 2: There are large differences in performance between state-of-the-art algorithms and between practitioners.

The main aim of examining state-of-the-art algorithms was to determine the performance differences within each algorithm across age-related variables rather than between algorithms. However, it is important to note algorithm performance not only varied within algorithms, but also between algorithms based on age-related variables. This needs to be taken into consideration when purchasing a facial recognition system, as performance of algorithms, even from world leading vendors, can vary considerably. This can impact on the operational objectives of a national security agency.

Similarly, the interest of this research was to examine practitioners as a group rather than their individual performance. However, it is worth noting that individual differences between practitioners were extensive, even though practitioners were from the same agency, conducting the same tasks, and being exposed to the same training. These differences can be seen in the boxplots in Study 2B, Section 5.4.3. Other research has also shown that there are large individual differences between practitioners (White, Kemp, Jenkins, Matheson et al., 2014) and novices (Burton et al., 2010). White, Kemp, Jenkins, Matheson et al. (2014) found that these individual differences are not due employment duration in facial comparison roles.

Key Finding 3: Practitioners performed less accurately with images of children than images of adults.

The results of Study 2B showed that practitioners were less accurate, less confident, and took longer to make decisions with images of children than they did with images of adults. Past research has also found novices perform poorer with images of children than adults (Yadav et al., 2014). Given that this study incorporated images with up to a 10 year age variation, accuracy with images of children (*M* = 73.86%, *Mdn* = 74.75%) and adults (*M* = 92.05%, *Mdn* = 93%) was promising. This is particularly so when compared to previous studies using images of adults taken on the same day exhibiting around 80% accuracy (Megreya & Burton, 2007; White, Kemp, Jenkins, Matheson, et al., 2014; Zeng et al., 2012). However, studies are difficult to compare due to methodological differences, such as the experience of participants, image type, and whether or not participants were deadlined.

The advanced performance levels in the current research could reflect what practitioners are capable of in their operational settings given that they were presented with controlled coloured images and were not deadlined when making their decision. Previous research that did not deadline practitioners on one-to-many facial comparison tasks (Heyer, 2013) also produced better performance than those found by others that used a deadlined methodology (Megreya & Burton, 2006; White, Dunn, et al., 2015). Practitioners in the current research and in Heyer (2013) were instructed to work as quickly and accurately as possible, as would be expected in their operational environment. Hence, there was a trade-off between speed and accuracy which was up to the judgement of each practitioner. In these studies, accuracy seems to have been chosen over speed, whereas in deadlined studies they may not have this option. Despite these methodological differences, there is a consensus amongst research that facial comparisons with images of children are more difficult than with images of adults. This is likely to be due to the development of stable facial features by adulthood making comparisons easier.

Key Finding 4: Practitioners performed more accurately with mated image pairs of children and adults than non-mated image pairs.

The results presented in this thesis show that practitioners perform better with mated than non-mated image pairs of children and adults. The performance results for the Child and Adult groups based on pair type from Section 5.4.3 are summarised in Table 13.

Table 13. *Practitioners Performance for Child and Adult groups based on Mated and Non-Mated Image Pairs*

Performance Measure	Child Median (Mean)	Adult Median (Mean)
Mated Pairs		
Accuracy	83.67% (80.71%)	98% (95.99%)
Confidence	75.92% (73.17%)	89.80% (86.95%)
Response Time	15.92 sec (17.04 sec)	12.93 sec (13.06 sec)
Non-Mated Pairs		
Accuracy	68% (67.04%)	91.84% (88.11%)
Confidence	76.40% (74.26%)	87.20% (82.67%)
Response Time	16.94 sec (17.90 sec)	15.12 sec (15.14 sec)

Accuracy for both the mated Child and mated Adult groups was higher than their non-mated counterparts. Ceiling performance was obtained by practitioners comparing mated images of adults (*M* = 95.99%, *Mdn* = 98%). Using the mean for comparison purposes, mated accuracy with controlled images of children in this study (80.71%) was also high compared to past research by Ferguson (2015) with uncontrolled images of children (62.44%).

Although error rates were higher for non-mated pairs of children and adults than mated, the median error rate of non-mated pairs was approximately four times higher with images of children (32%) compared to images of adults (8.16%). This may be due to the less discriminating facial features in childhood compared to adulthood making children look more like each other.

Practitioners were also more accurate than confident on images of adults and mated pairs of children but were more confident than accurate on non-mated pairs of children. This indicates that practitioners were overconfident in their ability to make accurate decisions on non-mated images of children. Possible reasons for this are discussed as part of key finding 6 in the following section.

9.3 *Key Findings: Facial Comparison Performance with Images of Children at Different Ages and Age Variations*

A major finding from Study 3A and 3B was how much algorithm and practitioner performance varied depending on the age of the child in an image and the age variation between images. Given the extensive amount of data presented in Study 3A and 3B over 198 different conditions, the seven key findings outlined next will only broadly discuss the differences in performance across childhood and age variation.

Key Finding 1: Algorithm performance improves as the age of the child increases.

Algorithm performance in Study 3A was shown to progressively increase with advancing age. This is consistent with past research (Ferguson, 2015; Grother & Ngan, 2014; Mahalingam & Kambhamettu, 2012) and information provided by algorithm vendors during Study 1. It was also found that younger children (particularly those aged 0–2 years) were harder to recognise as themselves and to tell apart from others, as evidenced by the higher false non-match and false match rates in the younger years. Again, this is likely due to the less discriminating facial features available in younger children (Wilkinson, 2012). These findings support previous NIST FRVTs (Grother et al., 2017; Grother & Ngan, 2014).

Key Finding 2: Algorithm performance declines as the age variation increases.

Study 3A also found that algorithm performance declined as the age variation between images in a pair increased. This is consistent with past research (Lui et al., 2009; Mahalingam & Kambhamettu, 2012) and information provided by algorithm vendors in Study 1. Study 3A provided a unique opportunity to see the extent of this degradation at each year that the age variation increased up to 10 years and at each individual age in childhood (0–17 years). This is valuable data for any agency wanting to know at what point algorithm performance would be deemed inefficient with images of children at different ages, before age variation has too much impact on performance. Furthermore, algorithms found it difficult to tell different children apart if the age variation was shorter than 4 years (the false match rate was found to be higher in these conditions), suggesting that children of similar age are more similar in appearance than children of different ages. This finding is likely due to similar growth and development patterns occurring for all children at certain ages (Ricanek et al., 2013).

Key Finding 3: Threshold variation improves algorithm performance.

A suggestion by an algorithm vendor in Study 1 was to improve performance by using threshold variation. This means that the threshold is not set at one point but rather, fluctuates depending on the condition being examined, to gain improved performance. Similar to key finding 1 presented in Section 9.2, Study 3A found that threshold variation based on different ages and age variations also improved algorithm performance. This means that rather than using a false match rate based on images of adults, a false match rate set at 0.001 based on images of the same age and age variation being compared would improve performance further. For example, Table 14 compares the false non-match rates of Algorithm C with images of children at age 6 with a 4 year age variation, when the false match rate is set at 0.001 based on images of adults and images at the same age and age variation. This information was collated from the heat maps presented in Section 6.4.3.

Table 14. *Algorithm C's False Non-Match Rate Performance when the False Match Rate is set at 0.001 based on Different Image Groups*

FMR of 0.001 based on Images of……	FMR at Age 6 with a 4 Year Age Variation	FNMR
Adults	0.009	0.263
Children at age 6 with a 4 year age variation	0.001	0.389

As can be seen from the table, when Algorithm C is set based on an adult false match rate of 0.001, the false match rate for images of children at age 6 with a 4 year age variation at this threshold is nine times higher at 0.009. It should be noted that a threshold set for images of children at age 6 with a 4 year age variation instead would improve performance in contexts such as passport processing or border control as children would be considered to be low risk in terms of fraud detection, and thus there would be a reduction in the number of non-mated pairs that the practitioners would need to compare. However, this would impact on performance with mated pairs from this category, increasing the false non-match rate by 12.6% when compared to an overall threshold based on images of adults.

Key Finding 4: Practitioner accuracy increases as the age of the child increases.

Although it was expected that practitioners would be less accurate with younger children compared to older children, this difference was worse than expected. Table 15 presents the mean accuracy data based on the youngest age of a child in an image pair (i.e., all age variation data grouped by age). This data has been collated from Figure 41 of Study 3B.

Table 15. *Practitioner Accuracy (%) by Age (Years)*

Age	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Accuracy	64	71	76	80	79	79	82	83	84	83	84	84	85	85	88	87	88	89

The table shows that accuracy was much lower in the younger ages than the older ones. For example, the maximum difference in accuracy across the 4 youngest ages was 16 % (ranging from 64% at age 0 to 80% at age 3). However, the maximum difference across the 15 oldest ages was only 10% (ranging from 80% at age 3 to 89% at age 17). This performance degradation is likely due to younger children having less distinguishing facial features (Kozak et al., 2015; Ricanek et al., 2013). It could also be due to imaging issues where it is necessary to lay babies down when they cannot hold up their own head, or expression issues where it may be harder to capture a neutral expression of a baby than a person of any other age. However, for the practitioner studies, every effort was made to keep image pairs as standardised as possible. Hence, it is anticipated that the considerable performance differences are predominantly due to age-related changes in the face during these younger ages.

Key Finding 5: Practitioners find it difficult to tell younger children apart.

Practitioners performed poorly on non-mated pairs of younger children (i.e., ages 0-2 years). This suggests that practitioners found it more difficult to tell younger children apart. Table 16 presents mean practitioner accuracy across age based on pair type. This data has also been collated from Figure 41 of Study 3B.

Table 16. *Practitioner Accuracy (%) across Age (Years) based on Pair Type*

Age	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Mated	75	81	84	83	83	82	86	88	87	89	90	91	90	89	93	90	92	92
Non-Mated	53	61	67	76	76	77	78	79	81	78	78	77	79	81	83	83	85	87

The table shows that practitioners performed at chance level when presented with non-mated image pairs containing at least one baby (53%). Although both mated and non-mated performance improved with advancing age, the change across non-mated pairs (34%) doubled compared to mated pairs (17%). At age 0, there was a 22% difference in accuracy between mated and non-mated pairs. As age increased to 17 years, there was only a 5% difference in performance based on pair type. Again, these results are likely due to the lack of individuating features in early childhood to distinguish children from each other, but as children get older and develop more distinctive features, it becomes easier to recognise them as themselves and to tell them apart from others.

Key Finding 6: Accuracy declines and practitioners become less aware of their ability to make accurate decisions on non-mated pairs at longer age variations.

Statistical analyses in Study 3B revealed that practitioners did not significantly differ in accuracy on mated pairs across age variation. However, practitioner accuracy declined as age variation increased on non-mated pairs. Figure 53 collates mean accuracy and confidence data from Section 7.5.2 based on pair type across age variation.

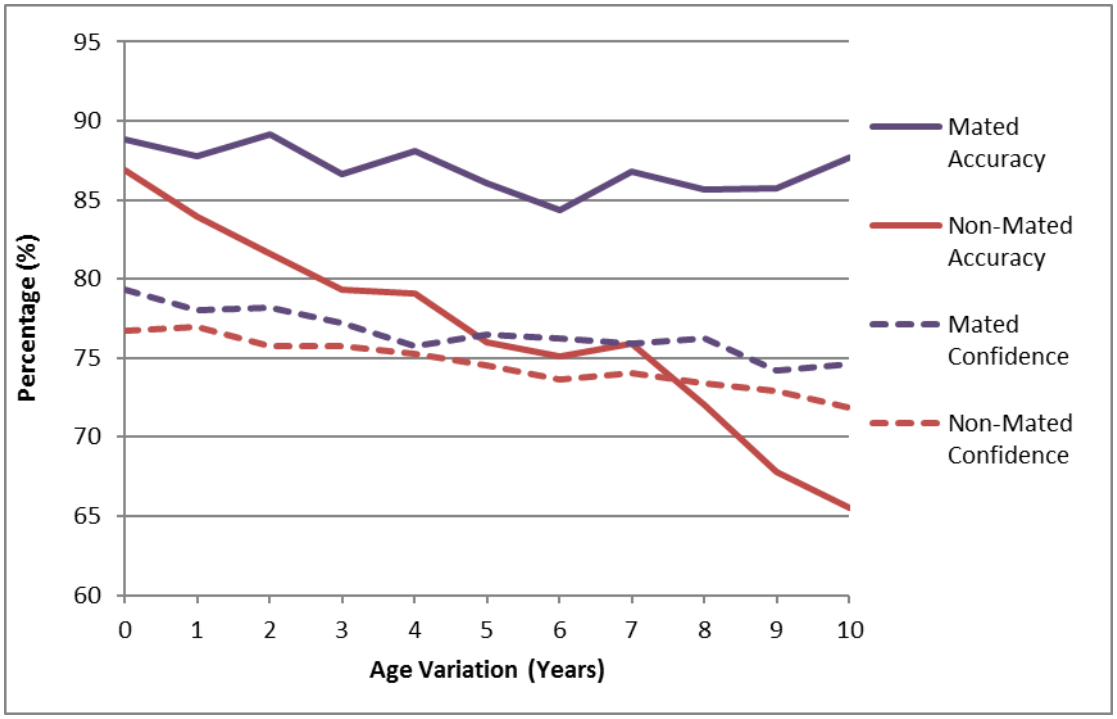


Figure 53. Accuracy and confidence based on pair type across age variation.

It is evident from this figure that practitioner accuracy remained relatively consistent across age variation on mated pairs. The confidence of practitioners also remained relatively consistent with mated and non-mated pairs across age variation with confidence gradually declining. However, practitioner accuracy dropped on non-mated pairs after a 7 year age variation, with little change in confidence. At a 10 year age variation, non-mated error rates (34%) were almost triple that of mated error rates (12%).

This decline in accuracy may be due to practitioners imagining what are 'possible and permissible' variations in appearance, as mentioned by Vernon (1952). Thus, as facial growth is extensive across childhood, variations in appearance of non-mated faces, particularly at longer age variations become 'possible and permissible' as mated faces. As a result, practitioners are no less confident in their decisions.

Another possible explanation, that may or may not be connected to the above, is that these practitioners are more commonly exposed to images of the 'same' person in their daily work and by default may have selected 'same' in the study. Non-mated pairs are likely to be easier when there are shorter age variations between images due to general knowledge about how

the face changes. Therefore, it may not seem as 'possible and permissible' that so much facial change occurs over short age variations. This leads to practitioners selecting 'same' unless they believe that so much facial change is not possible at shorter age variations. Thus, performance with non-mated pairs at shorter age variations was better than at longer age variations.

These results were in contrast to algorithm data showing that non-mated performance remained relatively consistent across age variation, but differed on mated pairs. This difference may be due to algorithms not having experience and knowledge about how the face changes over time, and practitioners placing emphasis on their general knowledge and bias on how the face changes and what is 'possible and permissible', which may have a negative impact on practitioner accuracy.

Key Finding 7: Practitioners make facial comparison decisions using different strategies.
Although all 120 practitioners that participated in Study 3B were from the same government agency and conduct facial comparisons as part of their daily role, these practitioners used different strategies and facial features to make their facial comparison decisions. For example, the most common feature compared was the ears and this was only mentioned by about half (52.50%) of the practitioners. The whole face was compared by 26 participants (21.67%), either prior to comparing facial features or in isolation from any other strategy. Another interesting finding was that 21 practitioners (17.5%) admitted to using their 'gut feeling' to make decisions with images of children, often when other strategies were not working. This was unexpected given that the majority (95%) of practitioners had participated in facial comparison training at some point during their career. This may be an indication that they had forgotten what they had been taught during this training or that the training was not relevant to facial comparisons with images of children.

9.4 *Key Findings: Facial Comparison Performance for Agency Specific Requirements*

A mock operational example was provided in Study 4 to demonstrate the value of the data collected in Study 3A and 3B. The example presented was to determine whether facial comparison performance with images of children from 0–4, 5–10, and 11–15 year age groups would differ if the passport renewal period was extended from 5 to 10 years. This was to

ascertain whether the current policy of a 5 year renewal period for children could be extended for any of these age groups without impacting on performance. This scenario is directly relevant to inform contexts including automated eGates in airports (algorithm), passport processing (practitioner), and manual border control (practitioner). Key findings from both the algorithm data and the practitioner data will be discussed.

Key Finding 1: Algorithm performance declines as the renewal period increases.

Algorithm performance for each age group declined when the passport renewal period was extended from 5 to 10 years. For the 0–4 year age group, algorithm performance was similar, but poor, for both the 5 and 10 year renewal periods. Poor performance in this age group is likely due to the lack of discriminating facial features making it difficult to recognise children as the same person and to tell them apart. As distinguishing facial features develop as children get older, algorithm performance improves, but the performance seems to still be impacted by larger age variations, as these facial features are still developing and stabilising. Vendors during Study 1 mentioned that pattern and shape information is used by algorithms which provides support to this explanation.

Algorithms were also impacted by both mated and non-mated image pairs, but more so by mated pairs across the different groups. The younger the age group, the poorer the performance, suggesting that younger children were more difficult to recognise as themselves. The results from this study suggest that agencies should not rely solely on an algorithm to make decisions with images of children aged 0–4 years, regardless of whether the renewal period is 5 or 10 years. The results for the 11–15 year age group may be considered feasible by border agencies to allow children in this age range to use eGates, if one of the higher performing algorithms was implemented. Indeed, children in this age range are already allowed to access SmartGate under some circumstances (DIBP, n.d.-a) and even younger in others (DIBP, n.d.-c). However, if the passport renewal period was to be extended to 10 years for children, algorithm performance would decline.

The poorer performance of the algorithms evaluated with the younger age groups (0–4, 5–10 years) also raised the question whether mitigating strategies need to be considered. For example, shorter passport renewal periods for younger children or raising the age that children can use SmartGate when departing Australia (if the algorithm used in SmartGate

performs similarly to those evaluated in this thesis). However, exploring this in more detail was outside the scope of this research.

Key Finding 2: Practitioner accuracy decreases when the renewal period increases (but only for non-mated pairs).

There was an approximately 5% difference in overall accuracy between a 5 year and 10 year renewal period for each of the 3 age groups (0–4, 5–10, 11–15 years). However, when divided into pair type, it was clear that for all age groups, performance remained relatively consistent with only a 1% difference on mated pairs across the two renewal periods. For the non-mated pairs, there was a higher difference in accuracy of practitioners between the two renewal periods (10% difference). This suggests that if passport renewal periods were to be extended for children's passports, it would impact considerably on the ability of practitioners to detect fraudulent passports. Therefore, victims of kidnapping or human trafficking, for example, may be harder to discover.

9.5 *Limitations of the Research*

The research conducted in this thesis accomplished its aim by providing empirical data to support operational requirements. However, there are several limitations that need to be considered when drawing conclusions. In regards to the algorithm studies, the state-of-the-art algorithms supplied for this research were the most up to date at the time this research commenced, hence they are now a few years old and a more recent version has been released by several of these vendors. However, the performance patterns within the data are likely to remain consistent and some agencies may still be using the same version of algorithms tested within this thesis (or older versions). Furthermore, the performance with images of children is also not likely to have improved drastically as the focus of most vendors has been on improving performance with images of adults, rather than children, as mentioned by vendors during Study 1. There are however aspirations to evaluate the most current algorithms on the same dataset to determine whether this theory holds true. As with all algorithm testing, it is also important to note that different datasets and/or different algorithms will provide different results to those presented within this thesis (Wayman et al., 2010).

One limitation of the practitioner studies was that they were conducted with an even number of mated and non-mated image pairs as discussed in Section 3.6.1. This is not typically encountered in operational settings, however, this methodology was chosen because realistic numbers of non-mated pairs encountered in operational settings are generally low and hence would not return enough data for statistical analyses. In addition, the number of non-mated pairs is typically unknown within a given operational setting and is also likely to differ considerably by agency. Regardless, the trends in data are likely to reflect what would occur in operational settings. For example, patterns in practitioner performance across childhood (e.g., accuracy increasing as the age of the child in an image increases) are not likely to change due to an equal presentation of mated and non-mated image pairs.

Another limitation of the research with practitioners is that they were forced to make a decision on each pair as either being the 'same' or 'different'. This forced choice is not representative of some operational settings where practitioners would have the option to submit images to a person with more expertise or ask other team members for help. This methodological decision is typically used in psychological research involving human participants to test one-to-one facial comparison performance, yet the reader is not informed that practitioners would likely have additional options in operational settings to escalate cases when they are unsure. This methodological decision was also raised as a concern by Dick (2015). However, the reason for such a methodological decision is that it enables an understanding of how practitioners would perform if they had to make that choice and ensures that enough data is collected for statistical purposes without practitioners being too cautious and opting out, rather than making decisions (which was shown in Heyer, 2013). This method also tests practitioner's own abilities on the variables under investigation to ensure they can be addressed by agencies, for example by using additional training. A balance is clearly necessary between ecologically valid studies and the examination of different phenomena (Heyer, 2013). This research has weighed up methodological decisions and justified them throughout this thesis. Furthermore, capturing the confidence of practitioners for every image pair in the studies allowed comparisons to be made between accuracy and confidence. From this data, it is clear that practitioners were generally aware of their own abilities in conducting facial comparisons albeit slightly more accurate than confident. It is possible that when practitioners reported low confidence, they would have sought help from their superior, other colleagues, or submitted the images to a more specialised area of their

agency. Essentially, this would raise the overall performance of the system (Dick, 2015). In areas where there was a considerable discrepancy between accuracy and confidence (such as non-mated images of children with large age variations where confidence stayed relatively stable but accuracy considerably declined), these could be investigated further to improve performance.

This research aimed to test facial comparison performance, not overall system performance in an agency. Depending on the operational context and what additional information is available, performance is expected to be much higher, as demonstrated by Dick (2015). In a passport processing context, the facial comparison task is just one of 200 checks performed (Gee, 2017). Thus, research presented in this thesis is based on facial comparison performance alone and claims should not be made as to how this agency and its practitioners perform as a whole when making identity decisions, as the task of conducting facial comparisons is just one part of a much larger process.

9.6 *Recommendations*

Recommendations for agencies conducting facial comparisons with images of children and future research directions are provided. The feasibility of incorporating these recommendations will vary by agency. A separate recommendation for algorithm vendors is also given.

Recommendation 1: Agencies should take careful consideration when purchasing a facial recognition system.

There are several factors that will determine the most appropriate facial recognition system for a national security agency to purchase. Algorithm performance, cost, mode of applications, after sales support, and computational expense are a few examples. With regards to algorithm performance, this varies considerably even between algorithms recognised as state-of-the-art. Therefore, implementing a better performing algorithm into an agency over another can reduce false non-match rates considerably at the predefined operational threshold. An agency considering purchasing a less accurate, but perhaps cheaper, algorithm/system should consider the trade-off between such a choice. For example, a less accurate algorithm could increase the risk of not detecting fraudulent users or require additional cost to employ more practitioners as a way to mitigate any concerns.

Recommendation 2: Agencies should consider using age information to improve algorithm performance.

Rather than perceiving age as a variable that impacts negatively on facial comparison performance, it could be used as a way to improve performance, particularly with images of children. This could be achieved by using biographical data that contains age information or age estimation algorithms.

Depending on the operational application, biographical information may be available that includes age. For example, in document processing applications, the date of birth of the person in an image and when the image was taken may be available, which can be used to calculate the age of the person in the image. This age information can then be used to calibrate the system. Therefore, threshold variation of the system is individually tailored rather than based on an overall threshold used by the system. This method incorporates age and age variations into the design of the system (Rebera & Mordini, 2013). Vendors surveyed during Study 1 mentioned dynamic selection and threshold variation as methods that could improve algorithm performance with images of children.

Another approach is to incorporate age estimation algorithms into processes to estimate the age of a person based on the facial features captured in an image. The output may be an absolute age or an age range. Age estimation could be used to filter a database thereby limiting the search only to those who belong to a specific age range (Costa-Abreu & Fairhurst, 2013; Yadav et al., 2014; Zhang, Lao, & Kurata, 2011). This has the potential to reduce the number of images an algorithm and/or a facial comparison practitioner may need to examine, thereby potentially reducing the number of errors made and increasing the speed of examination by focusing on the appropriate ages. For example, the author notes that during previous one-to-many testing conducted at DST Group, images of bald men returned candidate lists full of babies. Therefore, filtering the database by age would ensure only suitable matches are returned. This is a cost-effective and convenient approach that maximises the information extracted from a person with no additional interaction required (Costa-Abreu & Fairhurst, 2013). However, the age estimation approach relies on the accuracy of the age estimation algorithm. Age estimation was identified in Study 1 as one of the requirements that agencies were seeking empirical data for yet this was out of scope for this research (Requirement 5: determining age estimation performance with images of children and adults). NIST has

evaluated state-of-the-art age estimation algorithms as part of the FRVT 2013 (Ngan & Grother, 2014). Age estimation algorithms have also been tested in concurrent research by the author on the same images used in this thesis and these results are available to agencies upon request. Previous research (Ngan & Grother, 2014) and vendors with age estimation algorithms who were surveyed in Study 1 have found that age estimation algorithms perform more accurately with images of children than adults. This is for the same reason that facial recognition algorithms perform more poorly with images of children compared to adults. Namely, that as there is extensive facial growth occurring in childhood, it is easier for age estimation algorithms to estimate the age of a child in an image, as some growth and development patterns in childhood are consistent for all individuals (Ricanek et al., 2013). Therefore, age estimation is likely to be even more valuable with images of children than adults.

Recommendation 3: Facial comparison training should incorporate more information relevant to facial comparisons with images of children.

Study 3B captured information demonstrating that practitioners use different strategies when making facial comparison decisions (Section 7.4.3). Some practitioners look at the face as a whole or use their 'gut feeling', while others compare specific facial features. The approach taken by individual practitioners varied from one to another. This was an interesting finding given that these practitioners were from the same agency and the majority had been provided the same training. As mentioned as part of key findings in Section 9.3, this may be an indication that the training was conducted too long ago so practitioners do not remember the techniques taught. Alternatively, it could be that the techniques demonstrated during the course were not relevant or specific enough for facial comparisons with images of children. According to C. Thomas, former facial comparison training coordinator at the Department of Immigration and Border Protection, facial comparison training only typically involves a small portion of training with images of children. In addition, the focus of facial comparison training with images of children has been to explain which facial features used to conduct comparisons with images of adults are not as useful for comparisons with images of children (personal communication, 16 September, 2016).

Currently within Australia, there is no formally recognised facial comparison training course. Agencies do, however, conduct their own in-house training or invite staff from another agency

to provide training, such as the Department of Immigration and Border Protection. Progress has been made to establish a framework that guides facial recognition and comparison training needs for agencies. This is being conducted through the Facial Biometrics Centre of Expertise (FaBCoE). Based on the findings in this research, it is recommended that any formal training program developed incorporates more training and testing using images of children, and also provides a better understanding of which facial features practitioners should use for comparison purposes. As Study 3B suggests that non-mated image pairs of children are more difficult for practitioners than mated pairs, consideration should be given to providing more training with non-mated pairs that require practitioners to discriminate between children.

Given that not all agencies or practitioners within an agency may need to compare images of children, a separate module may be appropriate. As facial comparisons of children would be a more specialised module, it may be that practitioners need to complete certain modules prior including those involving adults. Once practitioners had successfully completed a module on children, they could then be allocated work that involves conducting comparisons with images of children as part of their regular work duties.

Recommendation 4: Agencies should assess facial comparison practitioner performance on a regular basis.

It is not enough for an agency to provide practitioners with brief facial comparison training at the beginning of their careers and hope that they retain this information for years to come. Furthermore, practitioners in most agencies are not provided with feedback on their facial comparison performance during their work duties due to the nature of the task. Therefore, there is currently no opportunity for practitioners to know if they are performing well on this task or not (unless they participate in research, which is not necessarily an accurate measure of their on-the-job performance). Practitioners should be tested on a regular basis to determine if their performance levels are at a standard deemed acceptable. If performance levels decline, further training or mentoring via a more experienced practitioner until performance improves may be considered. Feedback training has also been shown to be effective (Alenezi & Bindemann, 2013; Dowsett & Burton, 2014; White, Kemp, Jenkins, & Burton, 2014).

Testing practitioner performance may also demonstrate that some practitioners perform at acceptable levels with images of adults but not children (or vice versa). These practitioners could be provided with workloads that only involve facial comparisons with images of adults or provided with different tasking if they are poor at facial comparisons in general. This would improve the overall performance within the agency as outliers such as those seen in the boxplots in Section 5.4.3 (Study 2A, Chapter 5) could be recognised and addressed.

To ensure that practitioner performance stays at a peak level, a suggestion has been made by R. Heyer, Science Team Leader for Biometrics at DST Group to develop a central facial comparison practitioner hub in Australia to support the Australian government's National Facial Biometric Matching Capability (personal communication, 28 August, 2014). This hub would be staffed by practitioners from national security agencies and would provide them the opportunity to conduct facial comparisons on all types of images that national security agencies are exposed to, not just the images encountered in their own agency. The idea is that practitioners could be seconded into this role for a period of time after which they would return to their agency, more proficient on this task. This would help to develop a higher expertise level of practitioners and ultimately improve performance across national security agencies, while ensuring that those agencies with little or no facial comparison capability have access to expert decision making. Secondment into this hub could also be incorporated into formal training as a module to complete.

Recommendation 5: Agencies should collaborate with academia and industry as much as possible.

One of the most effective ways to gain the most realistic operational data that is of value to national security agencies is for academia, industry, and agencies to collaborate together on research programs. The research presented in this thesis is a prime example of what can be achieved with support from agencies (in the form of requirements, images, and participants) and industry (in the form of state-of-the-art algorithms, completed algorithm vendor surveys, and the development of an experimental application for practitioner testing).

Providing algorithm developers with access (under strict conditions) to databases for algorithm training and development would also result in more accurate algorithms being developed for agency specific purposes. Currently, a research program is underway at DST

Group in this area using uncontrolled images of children. The aim is to develop an algorithm specifically tailored for this type of image set.

For agencies that are considering the purchase of a facial recognition system, implementing other technologies or implementing different procedures into their operational environment, collaborating with academia and industry is a robust way to aid decision making to determine outcomes. Experienced researchers with knowledge on how to conduct scientific research to determine if new technologies or procedures are worth implementing can be invaluable to the agency. This could save time, money, and resources, as well as improve system performance.

Recommendation 6: Algorithm vendors should incorporate current findings into future developments of their facial recognition algorithms.

This research has provided algorithm vendors with a unique opportunity to have their own algorithms tested on extremely large datasets that they would not typically be able to access. This provides vendors with a better understanding as to the degradation of their algorithms with images of children compared to adults, as well as more granularity at individual ages throughout childhood and at age variations ranging up to 10 years.

During Study 1, vendors mentioned the value of being involved in this research. Vendors rely on training their algorithms on large databases of images to improve performance, but access to images of children is difficult. Although some vendors admitted their predominant focus is to improve algorithm performance with adults as this is where there is a bigger demand, vendors also admitted that customers had discussed using their algorithms with children. Given that many national security agencies use facial recognition algorithms provided by these vendors to help determine the identity of children, it is recommended that vendors incorporate this new knowledge into future developments of their algorithms. This could be in the form of a specific algorithm developed for children or more development with current versions of their algorithms that focuses on children.

Recommendation 7: Researchers should report performance data based on pair type.

A requirement identified in Study 1 was that agencies sought further information on performance based on pair type (Requirement 4). This requirement was incorporated into all three requirements investigated in this thesis due to its importance to agencies.

Understanding performance based on pair type is necessary for agencies due to different consequences for making wrong decisions, dependant on whether they are made on mated or non-mated pairs. These consequences vary based on the nature of the facial comparison task. Despite this, some studies only present overall performance data (e.g., Lanitis, 2008; Yadav et al., 2014; Zeng et al., 2012). Thus, future facial comparison research should provide data not just based on overall performance but based on pair type so that the results are more relevant to agencies.

In addition, when facial comparison studies do present results based on pair type, there is a discrepancy amongst studies as to whether performance is better on mated or non-mated pairs (Ferguson, 2015; Megreya & Burton, 2007; Megreya et al., 2010; White, Kemp, Jenkins, Matheson, et al., 2014). This discrepancy is likely due to methodological differences particularly when selecting non-mated pairs between studies. For example, Study 2B used a more ecologically valid method than those typically used, such as randomisation, which is shown to result in better non-mated performance due to less similarity in appearance between images, making them easier to detect (Calic, 2012). Furthermore, some research has been severely restricted in terms of the number of images available to select non-mated pairs. For example, the largest difference in performance between mated and non-mated pairs in a study was found in White, Kemp, Jenkins, Matheson, et al. (2014) with accuracy of participants (including practitioners) to be almost 20% higher with non-mated pairs. However, only 17 images of people from the same gender were available to select non-mated pairs from and the appearance of this group of people was considered diverse by the authors. Therefore, the pairs were easy to determine as different people which can skew performance results. This is especially problematic in situations where only overall performance data is provided as it cannot be determined how much of a role the method used to select non-mated pairs played in the overall results. Consequently, future research should present results based on pair type if not already doing so. This should include a thorough explanation on how images were selected for both pair types and an emphasis should be placed on the results based on pair type. This will help to explain any performance differences within and across studies and ensure the findings are of more relevance to agencies.

Recommendation 8: Further research should examine different paradigms, expertise levels, and image types.

The current research examined algorithm and practitioner performance using a one-to-one paradigm with controlled operational images. This was conducted as it provides the upper bound level of performance and was considered of most value to national security agencies. However, agencies also need to understand algorithm and practitioner performance with images of children on a one-to-many paradigm and with uncontrolled images and video, under similar conditions to those evaluated in this thesis. This would complement the current research.

It would also be valuable to evaluate practitioners that work at the examiner level on the same datasets to determine if their accuracy is higher than the reviewers tested in this thesis. This could be conducted using conditions similar to their operational environment where they have more tools, time, and resources to make decisions than practitioners at the reviewer level. Research testing examiner performance with access to their own tools is underway with uncontrolled images of adults through NIST (NIST, 2017).

The images presented in this thesis should also be tested under the same conditions with novices. This is to determine whether performance patterns stay consistent across different experience levels. For example, practitioners in Study 3A performed consistently on mated pairs across age variation, but performed more poorly on non-mated pairs as the age variation increased. It would be insightful to determine if novices performed in the same way so that more assumptions can be made about the difficulty of non-mated pairs of children. For example, to confirm that humans make decisions on non-mated pairs because they believe the changes in performance are 'possible and permissible' or whether this is a function of practitioners saying 'same' more in their daily role, which inadvertently decreased performance on non-mated pairs.

Recommendation 9: Further research should examine performance differences based on gender.

Although not considered a limitation of this research as it was not identified as a requirement during Study 1, it is still important to understand how gender impacts on facial comparison performance, especially with images of children. Separating the current data by gender could

determine if there are facial changes that occur at different ages for males compared to females that are considerable enough to impact on algorithm performance. Past research has shown that algorithms do show a gender bias and perform better with males than females (Klare et al., 2012). Further research with images of children across childhood based on gender would also be valuable and may help to improve algorithm performance.

It has been found that the gender of people in images of adults does not impact on the performance of novices (Zeng et al., 2012). In Ferguson's (2015) one-to-one study, a gender difference was found, with higher accuracy of participants (including practitioners) on female pairs (80.23%) than on male pairs (56.34%). However, Ferguson's one-to-many study found gender differences in the opposite direction with lower accuracy on female images (35.67%) than on male images (73.18%). Clearly, more research is required to investigate how performance differs based on gender across childhood. This knowledge may help to improve facial comparison practitioner training with images of children.

Recommendation 10: Further research should examine different technologies to improve performance with images of children.

Incorporating additional technologies may also improve facial comparison performance. For example, artificially ageing a face to compensate for the impact that age has on performance is used in some investigative applications when the age variation is large. This has traditionally been conducted by forensic artists (Frowd, Erickson, & Lampinen, 2014). However, this technique can now be automated and may be a feasible way to improve system performance in other national security contexts, particularly as technology advances. An age progression algorithm ages the image prior to converting it into a template for facial comparison and so it is independent of a facial recognition algorithm. Several researchers are currently working on age progression software (Kemelmacher-Shlizerman, Suwajanakorn, & Seitz, 2014; Lanitis, 2008). Collaborations between academia, industry, and agencies would help to test performance and deploy these developments into national security agencies (Partington, 2013).

Age progression was defined as Requirement 6 (determining the performance of age progressed images for facial comparison) in Study 1 as an area warranting further investigation, but was out of scope for this thesis. Part of the interest in this area involves

establishing whether algorithm and facial comparison practitioner performance would improve if artificially aged images produced by automated software were used. This would be used for comparison purposes instead of using an old outdated image. Accurate age progression software would be particularly valuable with images of younger children when a considerable amount of facial growth is occurring. Research in this space is currently in the initial scoping stages by the researcher.

Recommendation 11: Further research should examine different modalities and fusion to improve performance with images of children.

Multiple identifiers may be available in some circumstances for a child. For example, face, height, age, gender (i.e., soft biometrics) or perhaps face and fingerprint. Research that determines what modalities are most age invariant in childhood and fusing the most appropriate ones could improve performance with children. This may require different weightings for different modalities and perhaps may vary for different ages to gain optimal performance. Integrating faces and soft biometrics has been shown to improve algorithm performance with newborns than using the face alone (Tiwari, Singh, & Singh, 2012).

Fusion of the results of multiple facial recognition algorithms may also be effective in improving performance with images of children. Fusion of multiple images is another approach and has shown to improve algorithm performance (Robertson, Kramer, & Burton, 2015) and the performance of novices (White, Burton, Jenkins, & Kemp, 2014). Fusing the decisions of multiple practitioners has also been shown to be effective (Dowsett & Burton, 2015; White, Burton, Kemp, & Jenkins, 2013). However, these studies have been conducted with images of adults, thus future research should conduct similar studies with images of children to ensure fusion methods are effective with children rather than simply extrapolating results. For example, fusing images of a child with large age variations may actually hinder performance due to considerable variability in the face over time. Further research in these areas will help to determine whether any of these strategies should be incorporated into operational procedures to improve performance.

Recommendation 12: Further research should determine the best strategies and most suitable facial features for practitioners to use when conducting facial comparisons with images of children.

It is possible that certain facial features may be more appropriate for facial comparison at different ages than others. For example, Yadav et al. (2014) found that participants performed more accurately with younger children when just the binocular region (67.02%) was visible, rather than the whole face (60.41%). However, participants were most accurate on images of teenagers and young adults when the whole face was available for comparison (87%). It is recommended that similar research be conducted using practitioners and operational images. If these findings are replicated, it may be appropriate to incorporate this strategy into training involving facial comparisons of children.

Identifying what facial features can be used to discriminate between faces of different children and which features remain stable between images of the same child over age variation would be a valuable addition to the literature and likely to improve performance of practitioners. Eye tracking experiments may also help to determine what facial features practitioners are comparing and correlate this with accuracy to help determine the most effective facial features to compare for different ages in childhood. Furthermore, determining what facial features are stable over time will help to devise more robust facial recognition algorithms (Ramanathan & Chellappa, 2006; Ortege, Brodo, Bicego, & Tistarelli, 2009).

It is also important to consider when marks and freckles appear in childhood, as some form early in life. Therefore, understanding at what ages they form is also important to help explain any differences in multiple images of the same child. For example, if no freckles are present in one image but present in another, images may be the same child but the freckles may have formed in the time between when the images were taken.

Recommendation 13: A standards document should be developed to describe how the face changes throughout childhood.

For agencies to incorporate more knowledge of how the face changes in childhood and the most stable features, it is first necessary to know what they are. FISWG is currently writing a document to provide information on the stability of facial features of adults for facial comparison purposes (FISWG, 2016). A similar document would be advantageous for children

that collates empirical data to demonstrate the most/least stable facial features in childhood and at what ages these features become stable. This document would provide information about what are truly 'possible and permissible' facial changes throughout childhood rather than practitioners relying on their beliefs of what is 'possible and permissible'. The author is currently developing such a document. This document could be used to inform training of practitioners, algorithm development, and ensure practitioners working in this space are all conducting comparisons based on the same information and knowledge. This will also help to ensure that features can be reliably assessed between practitioners (Ferguson, 2015). This document would also be useful in circumstances when agencies just need to exclude a person rather than verify or identify them. For example, an enhanced ability to discriminate children from each other could help to exclude potential child victims from an investigative case. Additional research as discussed in Recommendation 12 could be incorporated into this document in subsequent versions and the information provided in this document could also help to shape future research.

9.7 *Contribution of the Research*

The research reported in this thesis has made a significant contribution to the field of facial comparison (both algorithm and practitioner based). It is the first one-to-one research to be driven by requirements collected from national security agencies in Australia and overseas that employ facial recognition systems and/or facial comparison practitioners with a focus on images of children. It also incorporated insights from state-of-the-art facial recognition vendors. In addition, it is the first research to use operational images and experimental methods to investigate the performance of both algorithms and practitioners with images of children over various age-related conditions.

The data collected in this research, particularly those displayed in the heat map data matrices, remains a rich source of information that can be further interrogated to answer agency specific questions, as well as more theoretical ones. Access to state-of-the-art algorithms, facial comparison practitioners, and a large database of controlled operational images makes this research ecologically valuable to a range of agencies that has not been possible in past research. This research has also provided valuable data for algorithm vendors to focus on any performance degradations identified in their algorithms based on the age-related variables examined in this thesis. This will help to reduce error rates and improve performance of their

algorithms in future versions. This is something that should be seen as valuable by industry due to the potential for this knowledge to reduce risk (Partington, 2013).

Access to operational images provided the largest examination of algorithm and practitioner performance with images of children to date. Evaluating performance at every age in childhood and age variations in yearly increments allowed intricate details to emerge in performance that has not been possible before. Study 3A, which was conducted with several million images of children and five state-of-the-art algorithms (and one open source algorithm where possible) was the largest algorithm study that has been conducted exclusively with images of children. Study 3B, which involved 120 facial comparison practitioners from a government agency, using 23,760 manually (and painstakingly) selected image pairs, was the single largest study of facial comparison practitioners that has been conducted to date.

During this research, the President's Council of Advisors on Science and Technology (PCAST) report was released criticising forensic comparison disciplines for not having enough statistical data to inform accuracy and error rates (PCAST, 2016). Facial image comparison is emerging as a forensic comparison discipline (Prince, 2013). Thus, the extensive amount of data presenting accuracy and error rates in this thesis by algorithms and practitioners are a significant contribution to this field. The knowledge gained from this research can also be incorporated into the ISO/IEC Report – Biometrics and Children (ISO/IEC 30110, 2015) that is currently severely lacking up to date information regarding face biometrics and children.

The findings from this research have been presented at a range of conferences in Australia and overseas (see Appendix Q). Interest both nationally and internationally has been exhibited through invitations to official meetings and international conferences. The Ross Vining Memorial Student Scholarship and Highly Commended Oral Presentation Award were also both received from the Australian and New Zealand Forensic Science Society in 2016 for this research.

This thesis has also identified areas requiring further research, some of which are currently being investigated by the author in collaboration with industry and national security agencies. The areas currently under investigation would not have been possible, nor identified as necessary, without the current research being undertaken.

9.8 *Conclusion*

Determining the identity of children is critical in a range of national security applications. In Australia, the *Migration Amendment (Strengthening Biometrics Integrity) Bill 2015* (Cth) was designed to allow collection of personal identifiers from children as young as 5 years old. The ISO/IEC Report – Biometrics and Children (ISO/IEC 30110, 2015) has also recently been developed to provide standards when collecting biometrics with children. These documents demonstrate that children are not exempt from procedures involving biometric collection and this is likely to increase in the future as technology advances and automation increases.

The face tends to be the most common biometric used by national security agencies to determine the identity of children. Despite this, very little research has been conducted to determine the performance of algorithms and/or practitioners employed to make facial comparison decisions with images of children. Given that the face significantly changes in childhood compared to adulthood, research examining performance with images of adults cannot simply be extrapolated to infer performance with images of children. Similarly, research that focuses on short age variations cannot be extrapolated to infer performance in operational settings when age variations can exceed 10 years, as highlighted in Study 1. The research reported in this thesis has aimed to address these omissions in the literature and provide empirical data that is of operational relevance to agencies, and that will be used as a foundation to continue further research in this space.

This research has highlighted the impact that age-related variables have on the performance of algorithms and practitioners when conducting facial comparisons with images of children. The findings of this research have generated a series of operationally relevant recommendations. The research has also provided a foundation on which further research in this space can be based. With collaborations between academia, industry, and agencies, there is scope to achieve high levels of accuracy when conducting facial comparisons with images of children. The progress experienced by algorithm vendors over the years through NIST FRVTs (Grother & Ngan, 2014; Grother et al., 2011; Phillips et al., 2007; Phillips et al., 2003) with images of adults, and the results demonstrated in this thesis with images of children are encouraging, and warrant further investigation. Technology is continually advancing, and Bills and Standards are being developed with little knowledge of algorithm and practitioner performance. Further research and development that focuses on images of children for

identity purposes is imperative, so that procedural decisions made in national security agencies are based on empirical data.

References

Abdi, H. (2009). Signal Detection Theory (SDT). In B. McGaw, P. L. Peterson & E. Baker (Eds.), *Encyclopedia of Education* (3rd ed.). New York: Elsevier.

Akhtar, Z., Rattani, A., Hadid, A., & Tistarelli, M. (2013). Face recognition under the ageing effect: A comparative analysis. *Image Analysis and Processing, 8157*, 309–318.

Albert, M., Ricanek, K., & Patterson, E. (2007). A review of the literature on the aging adult skull and face: Implications for forensic science research and applications. *Forensic Science International, 172*(1), 1–9.

Albert, M., Sethuram, A., & Ricanek, K. (2011). Implications of adult facial aging on biometrics. In M. Albert (Eds.), *Biometrics – unique and diverse applications in nature, science, and technology* (pp. 89–106). Vienna, Austria: InTech.

Anjos, A., Komulainen, J., Marcel, S., Hadid, A., & Pietikäinen, M. (2014). Face anti-spoofing: Visual approach. In S. Marcel, M. S. Nixon, S. Z. Li (Eds.), *Handbook of biometric anti-spoofing: Trusted biometrics under spoofing attacks.* London: Springer.

Anser Analytic Services (2011). *Technologies for identifying missing children* (Report No. 186277). West Virginia, USA: U.S. Department of Justice.

Attorney General's Department (2015). *Face matching services*. Retrieved from https://www.ag.gov.au/RightsAndProtections/IdentitySecurity/Documents/Fact-Sheet-National-Facial-Biometric-Matching-Capability.pdf

ACIC (2017). *Biometric Identification Services.* Retrieved from https://www.acic.gov.au/our-services/biometric-matching/biometric-identification-services

Australian Passport Office (n.d.). *Children's passports*. Retrieved from https://www.passports.gov.au/passportsexplained/childpassports/Pages/default.aspx

Babbie, E. (2010). *The practice of social research* (13th ed.). Belmont, CA: Thomson Wadsworth Publishing.

Bachiochi, P. D., & Weiner, S. P. (2002). Qualitative data collection and analysis. In S. G. Rogelberg (Eds.), *Handbook of research methods in industrial and organisational psychology* (pp. 161–183). Malden, MA: Blackwell.

Barrett, D. (2016, March). *Trafficking fears as children allowed to use electronic passport gates for the first time*. Retrieved from http://www.telegraph.co.uk/news/2016/03/28/trafficking-fear-as-children-allowed-to-use-electronic-passport/

BBC News. (2007). *10,000 passports go to fraudsters*. Retrieved from http://news.bbc.co.uk/2/hi/6470179.stm.

Best-Rowden, L., Hoole, Y., & Jain, A. (2016). Automatic face recognition of newborns, infants, and toddlers: A longitudinal evaluation. *Proceedings from the 15th International Conference of the Biometrics Special Interest Group (BIOSIG)*. Retrieved from https://pdfs.semanticscholar.org/571b/f99661a70e303f5803e2deef5c609a60ca25.pdf

Bharadwaj, S., Bhatt, H., Vatsa, M., & Singh, R. (2016). Domain specific learning for newborn face recognition. *IEEE Transactions on Information Forensics and Security, 99*, 1.

Bindemann, M., Avetisyan, M., & Blackwell, K. (2010). Finding needles in haystacks: Identity mismatch frequency and facial identity verification. *Journal of Experimental Psychology: Applied, 16*, 378–386.

Biometric Technology Today (2010, January). *Child identification program to use face biometrics*. Volume 2010(1), 12.

Bishara, S. E., Peterson, L. C., & Bishara, E. C. (1984). Changes in facial dimensions and relationships between the ages of 5 and 25 years. *American Journal of Orthodontics, 85*(3), 238-252.

Blackburn, D., Bone, J. M., & Phillips, P. J. (2001). *Facial Recognition Vendor Test (FRVT) 2000 - Evaluation Report*. Retrieved from https://www.nist.gov/sites/default/files/documents/2016/12/19/frvt_2000.pdf

Bromby, M. C. (2003). At face value? *New Law Journal Expert Witness Supplement, 28*, 302-303.

Bruce, V. (1994). Stability from variation: The case of face recognition. The M.D. Vernon memorial lecture. *The Quarterly Journal of Experimental Psychology, Section A, 47*(1), 5–28.

Bruce, V. (2012). Familiar face recognition. In C. Wilkinson & C. Rynn (Eds.), *Craniofacial identification,* (pp. 1–10). New York, NY: Cambridge University Press.

Bucks, R. S., Garner, M., Tarrant, L., Bradley, B. P., & Mogg, K. (2008). Interpretations of emotionally ambiguous faces in older adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 63*(6), 337–343.

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *The Quarterly Journal of Experimental Psychology, 66*(8), 1467–1485.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behaviour Research Methods, 42*, 286–291.

Butavicius, M., Parsons, K., McCormac, A., Foster, R., Whittenbury, A., & MacLeod, V. (2011). Assessment of the ThruVision T4000 Passive Terahertz Camera: A Human Factors Case Study. In L. Jain & E. Aidman (Eds.), *Innovations in Defence Support Systems* (Vol. 2, pp. 183–206).

Calic, D. (2012). *From the laboratory to the real world: evaluating the impact of impostors, expertise and individual differences on human face matching performance* (Doctoral dissertation). University of Adelaide, Australia.

Chellappa, R., Sinha, P., & Phillips, P. J. (2010). Face recognition by computers and humans. *IEEE Computer Society, 43*(2), 46–55.

Chen, B.-C., Chen, C.-S., & Hsu, W. H. (2015). Face recognition using cross-age reference coding with cross-age dataset. *IEEE Transaction on Multimedia, 17*(6), 804–815.

Chung, S., Christoudias, M., Darrell, T., Ziniel, S. I., & Kalish, L. A. (2012). A novel image-based tool to reunite children with their families after disasters. *Society for Academic Emergency Medicine, 19*, 1227–1234.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences* (2nd ed.). Hillsdale,New Jersey: Lawrence Erlbaum Associates.

Connor, S. (2014, August). *Passport officers no better than untrained amateurs at recognising faces, study finds*. Retrieved from http://www.independent.co.uk/travel/news-and-advice/passport-officers-no-better-than-untrained-amateurs-at-recognising-faces-study-finds-9676869.html

Creswell, J. W. (2014). *Research design* (3rd ed.). Thousand Oaks, CA: Sage Publications.

CrimTrac. (2016). Biometric Identification. Retrieved on 16 June 2016 from https://www.crimtrac.gov.au/biometric-identification.

De Silva, C., Roberts, P., & Dowden, J. (2014). NEC facial recognition whitepaper: It's all about the face - facial recognition. Retrieved from http://nec.com.au/campaign/neoface/NECA_Facial_Recognition_whitepaper_141113.pdf

Department of Foreign Affairs and Trade. (2016). *General photo guidelines.* Retrieved from https://www.passports.gov.au/passportsexplained/theapplicationprocess/passportphotographguidelines/Pages/default.aspx.

DIBP (2009). Facial comparison principles and application – participant guide. Commonwealth of Australia.

DIBP (n.d.-a). *Arrivals SmartGate*. Retrieved from https://www.border.gov.au/Trav/Ente/Goin/Arrival/Smartgateor-ePassport

DIBP (n.d.-b). *Biometrics collection*. Retrieved from https://www.border.gov.au/Trav/Visa/Biom

DIBP (n.d.-c). *Departures SmartGate*. Retrieved from https://www.border.gov.au/Trav/Ente/Goin/Departing/departureSmartGate

Dick, D. (2015, July). *An evolving natural experiment in the determination of human error in operational environments*. Paper presented at the International Conference on Evidence Law and Forensic Science, Adelaide, Australia.

Doty, L. A. (1996). *Statistical process control* (2nd ed.). New York, USA: Industrial Press Inc.

Du, J-X., Zhai, C-M., & Ye, Y-Q. (2013). Face aging simulation and recognition based on NMF algorithm with sparseness constraints. *Neurocomputing, 116*, 250–159.

Dunstone, T., & Yager, N. (2009). *Biometric system and data analysis: Design, evaluation, and data mining*. Eveleigh, New South Wales: Springer.

Enlow, D. (1990). *Facial growth* (3rd ed.). Philadelphia: PA; W.B. Saunders Company.

Erbilek, M., & Fairhurst, M. (2012). A methodological framework for investigating age factors on the performance of biometric systems. *Proceedings of the on Multimedia and Security,* 115–122.

Fairhurst, M. (2013). Ageing and biometrics: an introduction. In M. Fairhurst (Ed.), *Age factors in biometric processing* (pp. 3–16). London, UK: The Institution of Engineering and Technology.

Farkas, L. G., & Hreczko, T. A. (1994). Age-related changes in selected linear and angular measurements of the craniofacial complex in healthy North American Caucasians. In L. G. Farkas, *Anthropometry of the Head and Face* (2nd ed., pp. 89–102). New York: Raven Press.

Feik, S. A., & Glover, J. E. (1998). Growth of children's faces. In J. G. Clement & D. L. Ranson (Ed.), *Craniofacial Identification in Forensic Medicine* (pp. 203–224). New York: Oxford University Press.

Ferguson, E. L. (2015). *Facial identification of children: A test of automated facial recognition and manual facial comparison techniques on juvenile face images* (Doctoral dissertation). University of Dundee, United Kingdom.

Ferrario, V. F., Sforza, C., Poggio, C. E., & Schmitz, J. H. (1998). Facial volume changes during normal human growth and development. *The Human Anatomical Record, 250*, 480–487.

FISWG (2012). *Guidelines for facial comparison methods* (Version 1.0). Retrieved from https://fiswg.org/FISWG_GuidelinesforFacialComparisonMethods_v1.0_2012_02_02.pdf

FISWG (2016). *Physical stability of facial features of adults* (Version 1.0). Retrieved from https://fiswg.org/DRAFT_FISWG_Physical_Stability_of_Facial_Components_v1.0_20160202.pdf

Field, A. (2011). *Discovering statistics using IBM SPSS statistics: And sex and drugs and rock 'n' roll* (4th ed.). Los Angeles, USA: Sage.

Fladsrud, T. (2005). *Face recognition in a border control environment: Non-zero effort attacks' effect on false acceptance rate* (Master's thesis). Gjøvik University College, Norway.

Fletcher, K. I., Butavicius, M. A., & Lee, M. D. (2008). Attention to internal face features in unfamiliar face matching. *British Journal of Psychology, 99*, 379–394.

Fu, Y., Guo, G., & Huang, T. S. (2010). Age synthesis and estimation via faces: A survey. *Pattern Analysis and Machine Intelligence, 32*(11), 1955–1976.

Gagné, C. L., & Spalding T. L. (2014). Relation diversity and ease of processing for opaque and transparent English compounds. In F. Rainer, F. Gardani, H. C. Luschützky, & W. U. Dressler (Eds.), *Morphology and meaning: Selected papers from the 15th International Morphology Meeting* (pp. 153–162). Amsterdam: John Benjamins Publishing.

Gibson, L. (2010). *Forensic art essentials – A manual for law enforcement artists*. Burlington: Elsevier Science.

Goldman, J., & Maret, S. (2016). *Intelligence and information policy for national security: Key terms and concepts*. London, UK: Rowman & Littlefield.

Goyal, D., Nagar, S., & Kumar, B. (2014). An enhanced approach for face recognition of newborns using HMM and SVD coefficients. International Journal of Computer Applications, 88(14), 17–23.

Gray, D. E. (2009). *Doing research in the real world* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Graves, I., Butavicius, M., MacLeod, V., Heyer, R., Parsons, K., Kuester, N., . . . Johnson, R. (2011). The role of the facial comparison practitioner in image-based airport security technologies. In L. Jain & E. Aidman (Eds.), *Innovations in Defence Support Systems* (Vol. 2, pp. 147-181).

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Gross, R. (2005). Face databases. In S. Z. Li & A. K. Jain (Eds.), *Handbook of face recognition* (pp. 301–327). New York, NY: Springer.

Grother, P. J. (2004). *Face Recognition Vendor Test 2002 – Supplemental Report* (NIST Interagency Report 7083). Retrieved from http://www.face-rec.org/vendors/frvt2002_supplemental.pdf.

Grother, P., & Ngan, M. L. (2014). *Face Recognition Vendor Test (FRVT) – Performance of face identification algorithms* (NIST Interagency Report 8009). Retrieved from http://ws680.nist.gov/publication/get_pdf.cfm?pub_id=915761

Grother, P., & Ngan, M. (2015). *Child Exploitation Image Analysis (CHEX-IA) facial recognition evaluation: Concept, evaluation plan, and API*. Retrieved from http://www.nist.gov/itl/iad/ig/chexia-face.cfm.

Grother, P., Ngan, M., & Hanaoka, K. (2017, August). *Ongoing Face Recognition Vendor Test (FRVT) – Part 1: Verification* (NIST Interagency Report XXXX Draft). Retrieved from https://www.nist.gov/sites/default/files/documents/2017/06/20/frvt_report_2017_06_19.pdf

Grother, P. J., Quinn, G. W., & Phillips, P. J. (2011). *Multiple-Biometric Evaluation (MBE) 2010 – Report on the evaluation of 2D still-image face recognition algorithms* (NIST Interagency Report 7709). Retrieved from http://www.nist.gov/itl/iad/ig/mbe.cfm.

Guo, G. (2013). Age prediction in face images. In M. Fairhurst (Eds.), *Age factors in biometric processing* (pp. 231–251). London, UK: The Institution of Engineering and Technology.

Guo, G., Mu, G., & Ricanek, K. (2010). Cross-age face recognition on a very large database: The performance versus age intervals and improvement using soft biometric traits. *Proceedings of the 20th International Conference on Pattern Recognition*, 3392–3395.

Hancock, P. J. B. (2012). Unfamiliar face recognition. In C. Wilkinson & C. Rynn (Eds.), *Craniofacial identification* (pp. 11–23). New York, NY: Cambridge University Press.

Hancock, P. & McIntyre, A. H. (2012). *Training methods for facial image comparison of unknown faces: Review and recommendations*. Home Office. The Home Office Scientific Development Branch.

Hassaballah, M., & Aly, S. (2015). Face recognition: Challenges, achievements and future directions. *IET Computer Vision, 9*(4), 614–626.

Havard, C. (2007). *Eye movement strategies during face matching* (Doctoral dissertation). University of Glasgow, Scotland.

Heyer, R. (2013). *Understanding one-to-many unfamiliar face matching in the operational context: The impact of candidate list size, expertise, and decision aids on the performance of facial recognition system users* (Doctoral dissertation). University of Adelaide, Australia.

Heyer, R., MacLeod, V., Carter, L., Semmler, C., & Ma-Wyatt, A. (2017). *Profiling the facial comparison practitioner in Australia.* DST-GD-XXXX, DST Edinburgh, South Australia.

Hillstrom, A., Sauer, J., & Hope, L. (2011). *Training methods for facial image comparison: A literature review*. Retrieved from http://eprints.port.ac.uk/5730/1/Hillstrom_Sauer_Hope_HOSDB_report_2011.pdf.

Hole, G. & Bourne, V. (2010). *Face processing: Psychological, neuropsychological, and applied perspectives*. New York, NY: Oxford University Press.

Hole, M., McLindin, B., Hanton, K., Malec, C., Yiu, S. Y., & Hanly, G. (2015). *An overview of a DSTO developed human operator image comparison software tool – Comparer* (DSTO-GD-0855).

Hugenberg, K., Young, S. G., Bernstein, M. J., & Sacco, D. F. (2010). The categorisation-individuation model: An integrative account of the other-race recognition deficit. *Psychological Review, 117*(4), 1168–1187.

Hughes, F., Lichter, D., Oswald, R., & Whitfield, M. (2009). Face biometrics: A longitudinal study. *Proceedings of Student-Faculty Research Day*, C2.1–C2.8.

Hunter, D., Tidderman, B., & Perret, D. (2012). Facial ageing. In C. Wilkinson & C. Rynn (Eds.), *Craniofacial Identification* (pp. 57–75). Cambridge, New York: Cambridge University Press.

International Civil Aviation Organization. (2016). *ICAO MRTD photo guidelines*. Retrieved from http://www.icao.int/Security/mrtd/Downloads/Technical%20Reports/Annex_A-Photograph_Guidelines.pdf

Interpol (2015, October). *Interpol hosts first facial recognition symposium*. Retrieved from https://www.interpol.int/News-and-media/News/2015/N2015-156

ISO/IEC 19795-1:2006 (2006). *Information technology – Biometric performance testing and reporting – Part 1: Principles and framework* (Figure 1).

ISO/IEC 30110:2015 (2015). *Information Technology – Cross Jurisdictional and Societal aspects of Implementation of Biometric Technologies – Biometrics and Children*.

Jain, A. K., Arora, S. S., Best-Rowden, L., Cao, K., Sudhish, P. S., & Bhatnagar, A. (2015). *Biometrics for child vaccination and welfare: Persistence of fingerprint recognition for infants and toddlers* (Michigan State University Technical Report, MSU-CSE-15-7). Retrieved from https://pdfs.semanticscholar.org/05a9/542fbf047b770c930f2a14f9df610d9cf86c.pdf

Jain, A. K., Klare, B., & Park, U. (2012). Face matching and retrieval in forensics applications. *IEEE Multimedia in Forensics, Security, and Intelligence, 19*(1), 20–28.

Jain, A. K., Nandakumar, K., & Ross, A. (2016). 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters, 79*, 80–105.

Jain, A. K., Ross, A. A., & Nandakumar, K. (2011). *Introduction to biometrics*. New York: Springer.

Jain, A. K., Ross, A., & Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 14*(1), 1–29.

Jayasinghe, U., & Dharmaratne, A. (2009). Matching facial images using age related morphing changes. *World of Science, Engineering and Technology, 3*, 12–27.

Jenkins, R., & Burton, A. M. (2008). Limitations in facial identification: The evidence. *Justice of the Peace, 172*, 4–6.

Johnston, R. A., & Bindemann, M. (2013). Introduction to forensic face matching. *Applied Cognitive Psychology, 27*, 697–699.

Jordan, G. (2016, March). *Facial recognition coming to the forefront of biometric modalities*. SecureIDNews. Retrieved from https://www.secureidnews.com/news-item/facial-recognition-coming-to-the-forefront-of-biometric-modalities/

Joseph, J. (2017). Technoprison: Technology and prisons. In L. J. Moriarty (Ed.), *Criminal justice technology in the 21st century* (pp. 172–205). Illinois, USA: Charles C Thomas.

Kandhasamy, P. (2017). *Texture based hand vein pattern recognition*. Hamburg, Germany: Anchor Academic Publishing.

Khiyari, H. E., & Wechsler, H. (2016). Face recognition across time lapse using convolutional neural networks. *Journal of Information Security, 7*, 141–151.

Kenya Mission (2015). Kenyan passports requirements. Retrieved from https://www.kenyamission-un.ch/?Consular_Matters:Requirements_for_citizenship:Kenyan_passports_Requirements

Klare, B. F., Burge, M. J., Klontz, J. C., Vorder Bruegge, R. W., & Jain, A. K. (2012). *IEEE Transactions on Information Forensics and Security, 7*(6), 1789–1801.

Klontz, J. C., Klare, B. F., Klum, S., Jain, A. K., & Burge, M. J. (2013). *Open source biometric recognition. Proceedings of the IEEE Conference on Biometrics: Theory, Applications, and Systems.* Retrieved from

http://www.cse.msu.edu/biometrics/Publications/GeneralBiometrics/Klontzetal_OpenSourceBiometricRecognition_BTAS13.pdf

Kothari, C. R. (2004). *Research methodology: Methods and techniques* (2nd ed.). Daryaganj, ND: New Age International.

Kozak, F. K., Ospina, J. C., & Cardenas, M. F. (2015). Characteristics of normal and abnormal postnatal craniofacial growth and development. In M. M. Lesperance, & P.W. Flint (Eds.), *Cummings Pediatric Otolaryngology* (pp. 55–80). Philadelphia, PA: Elsevier.

Krishnaswamy, K. N., Sivakumar, A. I., & Mathirajan, M. (2006). *Management research methodology: Integration of principles, methods and techniques*. Panchsheel Park, ND: Pearson Education.

Kumar, R. (2011). *Research methodology: A step-by-step guide*. London: Sage Publications.

Lampinen, J. M., Erickson, W. B., Frowd, C. D., & Mahoney, G. (2017). Estimating the appearance of the missing: Forensic age progression in the search for missing persons. In S. J. Morewitz & C. S. Colls (Eds.), *Handbook of missing persons* (pp. 251–269). Cham, Switzerland: Springer International Publishing.

Lanitis, A. (2008). Evaluating the performance of face-aging algorithms. *Proceedings form the 8th IEEE International Conference on Automatic Face & Gesture Recognition,* 1–6.

Lanitis, A. (2009). Facial biometric templates and aging: Problems and challenges for artificial intelligence. *Proceedings of the Workshops of the 5th IFIP Conference on Artificial Intelligence Applications & Innovations*, 142–149.

Lanitis, A. (2010). A survey of the effects of aging on biometric identity verification. *International Journal of Biometrics, 2*(1), 34–52.

Lanitis, A., & Tsapatsoulis, N. (2011). Quantification evaluation of the effects of aging on biometric templates. *IET Computer Vision, 5*(6), 338–347.

Lanitis, A., Tsapatsoulis, N., & Maronidis, A. (2013). Review of ageing with respect to biometrics and diverse modalities. In M. Fairhurst (Ed.), *Age factors in biometric processing* (pp. 17–36). London, UK: The Institution of Engineering and Technology.

Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks, 8*(1), 98-113.

Lee, J. (2015, April). Biometric Institute warns government about collecting biometrics from children. Biometric update.com. Retrieved from http://www.biometricupdate.com/201504/biometrics-institute-warns-australian-government-about-collecting-biometrics-from-children

Leonard, K. R. (2016). Assessment of facial recognition system performance in realistic operation environments. In T. Bourlai (Eds.), *Face recognition across the imaging spectrum* (pp. 117–138). Switzerland: Springer International Publishing.

Li, Z., Gong, X., & Tao, D. (2016). Aging face recognition: A hierarchical learning model based on local patterns selection. *IEEE Transactions on Image Processing, 25*(5), 2146–2154.

Li, S. Z., & Jain, A. K. (2011). Introduction. In S. Z. Li & A. K. Jain (Eds.), *Handbook of face recognition* (pp. 1–15). London: Springer-Verlag.

Li, Z., Park, U., & Jain, A. K. (2011). A discriminative model for age invariant face recognition. *IEEE Transactions on Information Forensics and Security, 6*(3), 1028–1037.

Ling, H., Soatto, S., Ramanathan, N., & Jacobs, D. W. (2010). Face verification across age progression using discriminative methods. *IEEE Transactions on Information Forensics and Security, 5*(1), 82–91.

Lowie, W., Verspoor, M., & Seton, B. (2010). Conceptual representations in the multilingual mind. In M. Pütz & L. Sicola (Eds.), *Cognitive processing in second language acquisition: Inside the learner's mind* (pp. 135–148). Amsterdam: John Benjamins Publishing.

Lui, Y. M., Bolme, D., Draper, B. A., Beveridge, J. R., Givens, G., & Phillips, P. J. (2009). A meta-analysis of face recognition covariates. *Proceedings of the Third International Conference on Biometrics: Theory, Applications, and Systems*, 139–146.

MacLeod, V., & McLindin, B. (2011). Methodology for the evaluation of an international airport automated border control processing system. In L. C. Jain, E. V. Aidman & C. Abeynayake (Eds.), *Innovations in Defence Support Systems* (pp. 115–145). Berlin, Heidelberg: Springer.

Mahalingam, G., & Kambhamettu, C. (2012). Face verification of age separated images under the influence of internal and external factors. *Image and Vision Computing, 30*, 1052–1061.

Mann, M., & Smith, M. (2017). *Automated facial recognition technology: Recent developments and approaches to oversight*. Retrieved from http://unswlawjournal.unsw.edu.au/sites/default/files/04-mannsmith-advance-access-final.pdf

Martin, L. (2013). Biometrics. In J. R. Vacca (Eds.), *Computer and Information Security Handbook* (pp. 957-972). Massachusetts, USA: Elsevier.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The DET curve in assessment of task performance. *Proceedings of the Fifth European Conference on Speech Communication and Technology* (pp. 1895–1898).

Martin, J. L., Murphy, E., Crowe, J. A., & Norris, B. J. (2006). Capturing user requirements in medical device development: The role of ergonomics. *Physiological Measurement, 27*, R49–R62.

Matheny, A. P., & Dolan, A. B. (1975). Changes in eye colour during early childhood: Sex and genetic differences. *Annals of Human Biology, 2*(2), 191–196.

Maxwell, J. A. (2013). *Qualitative research design: An interactive approach* (3rd ed.). Thousand Oaks, CA: Sage Publications.

McNabb, D. E. (2010). Research methods for political science (2nd ed.). New York, NY: M.E. Sharpe Inc.

McNicol, D. (2005). *A primer for signal detection theory*. Mahwah, New Jersey: Erlbaum.

Meek, A. (2016). *Biopolitical media: Catastrophe, immunity and bare life*. New York, USA: Taylor & Francis.

Megreya, A. M., Bindemann, M., & Havard, C. (2011). Sex differences in unfamiliar face identification: Evidence from matching tasks. *Acta Psychologica, 137*, 83–89.

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition, 34*(4), 865–876.

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics, 69*(7), 1175–1184.

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied, 14*(4), 364–372.

Megreya, A. M., Sandford, A., & Burton, A. M. (2013). Matching face images taken on the same day or months apart: the limitations of photo ID. *Applied Cognitive Psychology 27*(6), 700–706.

Migration Amendment (Strengthening Biometrics Integrity) Bill 2015 (Cth). Retrieved from http://www.austlii.edu.au/cgi-bin/download.cgi/cgi-bin/download.cgi/download/au/legis/cth/bill_em/mabib2015576.pdf

Moss, P. (2015). *An Australian facial training standard*. Paper presented at the Unfamiliar Face Identification Group (UFIG) Meeting, 09–11 February, Sydney.

Moss, P. (2016). *Facial Identification Scientific Working Group (FISWG) and Facial Aptitude Competency and Error Testing (FACET) group update*. Paper presented at the Unfamiliar Face Identification Group (UFIG) Meeting, 8–10 February, Sydney.

NIST (2017). CHEXIA Face Recognition. Retrieved from https://www.nist.gov/programs-projects/chexia-face-recognition

National Policing Improvement Agency. (2009). *Facial Identification Guidance 2009*. Retrieved from http://library.college.police.uk/docs/acpo/facial-identification-guidance-2009.pdf

Newton, E. M. (2007). Strengths and weaknesses of biometrics. In L. J. Camp (Ed.), *Economics of identity theft: Avoidance, causes and possible cures* (pp. 109–124). Indiana, US: Springer.

Ngan, M., & Grother, P. (2014). *Face Recognition Vendor Test (FRVT): Performance of automated age estimation algorithms* (NIST Interagency Report 7995). Retrieved from http://www.nist.gov/customcf/get_pdf.cfm?pub_id=915238.

Ngan, M. L., & Grother, P. J. (2015). *Face Recognition Vendor Test (FRVT) – Performance of automated gender classification algorithms* (NIST Interagency Report 8052). Retrieved from http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8052.pdf

Ortega, M., Brodo, L., Bicego, M., & Tistarelli, M. (2009). Measuring changes in face appearance through aging. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Recognition Workshops*, 107–113.

Otto, C., Han, H., & Jain, A. (2012). How does aging affect facial components? *Proceedings for the European Conference on Computer Vision 7584*, 189–198.

O'Toole, A. J., Phillips, P. J., Jiang, F., Ayyad, J., Penard, N., & Abdi, H. (2007). Face recognition algorithms surpass humans matching faces over changes in illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(9), 1642–1646.

Ozbek, M., & Bindemann, M. (2011). Exploring the time course of face matching: Temporal constraints impair unfamiliar face identification under temporally unconstrained viewing. *Vision Research, 51*, 2145–2155.

Paller, A. S., & Mancini, A. J. (2016). *Hurwitz clinical pediatric dermatology: A textbook of skin disorders of childhood and adolescence* (5th ed.). Canada: Elsevier.

Panis, G., Lanitis, A., Tsapatsoulis, N., & Cootes, T. F. (2015). An overview of research on facial ageing using the FG-Net Ageing Database. *IET Biometrics, 5*(2), 37–46.

Papesh, M. H., & Goldinger, S. D. (2014). Infrequent identity mismatches are frequently undetected. *Attention, Perception, and Psychophysics, 76*(5), 1335–1349.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015, September). Deep Face Recognition. *British Machine Vision Conference, 1*(3), 6.

Partington, A. (2013). An industrial perspective on biometric age factors. In M. Fairhurst (Ed.), *Age factors in biometric processing* (pp. 37–62). London, UK: The Institution of Engineering and Technology.

Passport Canada. (2012). International comparison of passport-issuing authorities. Retrieved from http://www.cic.gc.ca/english/department/consultations/passport/pdf/2012-03-compare-eng.pdf

Patterson, E., Sethuram, A., Albert, M., Ricanek, K., & King, M. (2007). Aspects of age variation in facial morphology affecting biometrics. *Proceedings of the First IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 1-6.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Petrov, M. (2012). *Law enforcement applications of forensic face recognition*. Retrieved from http://www.planetbiometrics.com/creo_files/upload/article-files/whitepaper_facial_recognition_morphotrust.pdf

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Marques, J., . . . Worek, W. (2005). Overview of the face recognition grand challenge. *Computer Vision and Pattern Recognition, 1,* 947–954.

Phillips, P. J., Grother, P., Micheals R. J., Blackburn, D. M., Tabassi, E., & Bone, M. (2003). *Face Recognition Vendor Test 2002 – Evaluation Report* (NIST Interagency Report 6965). Retrieved from http://ws680.nist.gov/publication/get_pdf.cfm?pub_id=50767

Phillips, P. J., Jiang, F., Narvekar, A., Ayyad, J., & O'Toole, A. J. (2010). *An other-race effect for face recognition algorithms* (NIST Interagency Report 7666). Retrieved from http://bbs.utdallas.edu/facelab/publications/pdf/2011_ACM_Phillips.pdf.

Phillips, P. J., Moon, H., Rauss, P., & Rizvi, S. A. (1997). The FERET September 1996 database and evaluation procedure. *Audio and Video-Based Biometric Person Authentication 1206*, 395–402.

Phillips, P. J., Scruggs, W. T., OToole, A. J., Flynn, P. J., Bowyer, K. W., Schott, C. L., & Sharpe, M. (2007). *FRVT 2006 and ICE 2006 large-scale results* (NIST Interagency Report 7408). Gaithersburg, MD: NIST.

Planet Biometrics (2017, March). *Brisbane airport trial launches biometric trial*. Retrieved from http://www.planetbiometrics.com/article-details/i/5620/

Poh, N., Chan, C-H, Kittler, J., Fierrez, J., & Galbally, J. (2012). *BEAT – Biometrics Evaluation and Testing – D3.3: Description metrics for the evaluation of biometric performance*. Retrieved from https://www.beat-eu.org/project/deliverables-public/d3.3-description-of-metrics-for-the-evaluation-of-biometric-performance.

Prince, J. (2013). *To examine emerging police use of facial recognition systems and facial image comparison procedures.* Retrieved from http://www.churchilltrust.com.au/media/fellows/2012_Prince_Jason.pdf

Ramanathan, N., & Chellappa, R. (2006a). Face verification across age progression. *IEEE Transactions on Image Processing, 15*(11), 3349–3361.

Ramanathan, N., & Chellappa, R., (2006b). Modelling age progression in young faces. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1*, 387–394.

Ramanathan, N., Chellappa, R., & Biswas, S. (2009). Age progression in human faces: A survey. *Visual Languages and Computing, 15*, 3349–3361.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*(3), 510–532.

Rebera, A. P., & Mordini, E. (2013). Biometrics and ageing: social and ethical considerations. In M. Fairhurst (Ed.), *Age factors in biometric processing* (pp. 37–62). London, UK: The Institution of Engineering and Technology.

Rhodes, G. (2013). Face recognition. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 46–69). New York: Oxford University Press.

Ricanek, K., Bhardwaj, S., & Sodomsky, M. (2015). A review of face recognition against longitudinal child faces. *Proceedings of the 14th International Conference of the Biometrics Special Interest Group (BIOSIG 2015)*, 15-26.

Ricanek, K., & Boone, E. (2005). The effect of normal adult aging on standard PCA face recognition accuracy rates. *Proceedings of the International Joint Conference on Neural Networks, 4,* 2018–2022.

Ricanek, K., Mahalingam, G., Albert, A. M., & Vorder Bruegge, R. W. (2013). Human face ageing: A perspective analysis from anthropometry and biometrics. In M. Fairhurst (Ed.), *Age factors in biometric processing* (pp. 93–116). London, UK: The Institution of Engineering and Technology.

Ricanek, K., & Tesafaye, T. (2006). MORPH: a longitudinal image database of normal adult age-progression. *Proceedings of the IEEE Seventh International Conference of Automatic Face and Gesture Recognition*, 341–345.

Riopka, T., & Boult, T. (2003). *The eyes have it*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.5.9459&rep=rep1&type=pdf

Robertson, S., & Robertson, J. (2013). *Mastering the requirements process: Getting requirements right* (3 ed.). Westford, MA: Pearson Education, Inc.

Robinson, M. (2013, June). *Schoolgirl, 9, passes through Turkish customs with toy passport identifying her as a unicorn*. Retrieved from http://www.dailymail.co.uk/news/article-2340135/Schoolgirl-9-passes-Turkish-customs-toy-passport-identifying-UNICORN.html

Rose, J. A. & Lacher, D. C. (2017). *Managing public safety technology: Deploying systems in police, courts, corrections, and fire organisations*. New York, USA: Taylor & Francis.

Sadwick, R. (2017, January). *Stop human trafficking: The role of tech and public-private partnerships.* Retrieved from https://www.forbes.com/sites/rebeccasadwick/2017/01/10/stop-human-trafficking/#36bd43226294

Salkind, N. (2006). *Encyclopedia of measurement and statistics*. Thousand Oaks, California: SAGE Publications.

Schuckers, M. E. (2010). *Computational methods in biometric authentication: Statistical methods for biometric authentication*. London: Springer-Verlag.

Schwandt, T. A. (2007). *The Sage dictionary of qualitative inquiry*. Thousand Oaks, CA Sage Publications.

Semmler, C., Heyer, R. L., Ma-Wyatt, A., & MacLeod, V. (2013). Understanding expertise in unfamiliar face matching. *Australasian Experimental Psychology Conference*, Adelaide.

Sharma, M. (2008, April). *SmartGate passport check goes national*. Retrieved from http://www.theaustralian.com.au/australian-it-old/smartgate-passport-check-goes-national/news-story/1ffe941b4c52eb6387772bebef32f7b2

Sirovich, L. & Kirby, M. (1987). A low-dimensional procedure for the characterization of human faces. *Journal of Optical Society of America, 4*(3), 519-524.

Sforza, C., Grandi, G., Binelli, M., Dolci, C., De Menezes, M., & Ferrario, V. F. (2010). Age- and sex-related changes in three-dimensional lip morphology. *Forensic Science International, 200*, 182.e1–182.e7.

Sforza, C., Grandi, G., Binelli, M., Tommasi, D. G., Rosati, R., & Ferrario, V. F. (2009). Age- and sex-related changes in the normal human ear. *Forensic Science International, 187*, 110.e1–110.e7.

Sforza, C., Grandi, G., Catti, F., Tommasi, D. G., Ugolini, A., & Ferrario, V. F. (2009). Age- and sex-related changes in the soft tissues of the orbital region. *Forensic Science International, 185*, 115.e1–115.e8.

Sforza, C., Grandi, G., De Menzes, M., Tartaglia, G. M., & Ferrario, V. F. (2010). Age- and sex-related changes in the normal human nose. *Forensic Science International, 204*, 205.e1–205.e9.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology, 46*, 561–584.

Singh, R., Vatsa, M., Noore, A., & Singh, S. (2007). Age transformation for improving face recognition performance. In A. Gosh, R. De & S. Pal (Eds.), *Pattern recognition and machine intelligence* (pp. 576–583), Vol 4815. Berlin: Springer.

Sinha, P., Balas, B., Ostrovsky, Y., & Russell, R. (2006). Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE, 94*(11), 1948–1962.

Somanath, G., MV, R., & Kambhamettu, C. (2011). VADANA: A dense dataset for facial image analysis. *Proceedings from the IEEE International Conference on Computer Vision Workshops,* 2175–2182.

Spalding, P. M. (2004). Craniofacial growth and development: Current understanding and clinical considerations. In M. Miloro, P. Larsen, G. Ghali, & Teton Data Systems, *Peterson's principles of oral and maxillofacial surgery*. NC, USA: People's Medical Publishing House.

Spaun, N. A. (2009). Facial comparisons by subject matter experts: Their role in biometrics and their training. In M. Tistarelli & M. S. Nixon (Eds.), *Advances in Biometrics* (pp. 161–168). Berlin, Germany: Springer-Verlag.

Spaun, N. A. (2011). Face recognition in forensic science. In S. Z. Li & A. K. Jain (Eds.), *Handbook of face recognition* (pp. 655–670). London: Springer-Verlag.

Stanislaw, H. & Todorov, N. (1999). Calculation of signal detection theory measures. *Behaviour Research Methods, Instruments, & Computers, 31*(1), 137–149.

Stebbins, R. A. (2008). Exploratory data analysis. In L. M. Given (Ed.), *The Sage Encyclopedia of qualitative research methods* (pp. 326–328). Thousand Oaks, CA: Sage Publications.

Stephens, R. G., Semmler, C., & Sauer, J. D. (2017). The effect of the proportion of mismatching trials and task orientation on the confidence-accuracy relationship in unfamiliar face matching. *Journal of Experimental Psychology: Applied,* doi http://dx.doi.org/10.1037/xap0000130.

Suri, H. (2011). Purposeful sampling in qualitative research synthesis. *Qualitative Research Journal, 11*(2), 63–75.

Taylor, K. T. (2001). *Forensic Art and Illustration*. Boca Raton, FL: CRC Press.

The Guardian (2015, June). *French police detain six-year-old girl for three days over passport mistake.* Retrieved from https://www.theguardian.com/world/2015/jun/12/france-police-detain-girl-aged-six-three-days-passport-mistake

Tian, Y-L., Kanade, T., & Cohn, J. F. (2005). Facial expression analysis. In S. Z. Li & A. K. Jain (Eds.), *Handbook of face recognition*. New York, NY: Springer.

Tistarelli, M., Yadav, D., Vatsa, M., & Singh, R. (2013). Short- and long-time ageing effects in face recognition. In M. Fairhurst (Ed.), *Age factors in biometric processing* (pp. 253–275). London, UK: The Institution of Engineering and Technology.

Tiwari, S., Singh, A., & Singh, S. K. (2012). Intelligent method for face recognition of infant. *International Journal of Computer Applications, 52*(4), 46–50.

Towler, A., White, D., & Kemp, R. (2014). Evaluating training methods for facial image comparison: The face shape strategy does not work. *Perception, 43*, 214–218.

Turk, M. A. & Pentland, A. P. (1991). Face recognition using eigenfaces. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 586-591.

U.S. Department of Homeland Security (2015). *Biometric systems application note*. Retrieved from https://www.dhs.gov/sites/default/files/publications/Biometric-Sys-AppN_0615-508.pdf

U.S. Department of State (n.d.). *Passport and Visa fraud: A quick course*. Retrieved from https://www.state.gov/m/ds/investigat/c10714.htm

Vacca, J. R. (2007). *Biometric technologies and verification systems.* Massachusetts, USA: Elsevier.

Valladares, A. (2012). *The effect of ageing on facial recognition performance during passport image verification* (Master's thesis). University of Adelaide, Australia.

Verma, J. P. (2016). *Repeated measures design for empirical researchers*. Hoboken, New Jersey: John Wiley & Sons.

Vrankulj, A. (2013, March). *Walt Disney World introduces new RFID gate system*. Biometricupdate.com. Retrieved from http://www.biometricupdate.com/201303/walt-disney-world-introduces-biometric-verification-for-passholders

Walker, R. (2004). Applied qualitative research. In M. S. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *Encyclopedia of social science research methods* (pp. 20). Thousand Oaks, CA: Sage Publications.

Wayman, J. (2016). *Performance of SmartGate Australia*. International Biometric Performance Testing Conference, May 3–5 2016. Retrieved from http://www.nist.gov/itl/iad/ig/international-biometric-performance-testing-recording.cfm

Wayman, J., Possolo, A., & Mansfield, A. J. (2010). *Fundamental issues in biometric performance testing: A modern statistical and philosophical framework for uncertainty assessment*. Retrieved from https://www.nist.gov/sites/default/files/documents/2016/11/30/fundamentalissues_final.pdf

Wechsler, H. (2007). Reliable face recognition methods: System design, implementation and evaluation. Fairfax, VA: Springer Science.

Wei, X., & Li, C.-T. (2017). Face recognition technologies for evidential evaluation of video traces. In M. Tistarelli & C. Champod (Eds.), *Handbook of biometrics for forensic science* (pp. 177–193). Cham, Switzerland: Springer International Publishing.

Wen, D., Fang, C., Ding, X, & Zhang, T. (2010). Development of recognition engine for baby faces. *Proceedings of the International Conference on Pattern Recognition,* 3408–3410.

Wen, Y., Li, Z., & Qiao, Y. (2016). Latent factor guided convolutional neural networks for age-invariant face recognition. *Proceedings from the IEEE Conference on Computer Vision and Pattern Recognition*, 4893–4901.

Westbourne Group. (2017). *The Westbourne Group*. Retrieved from http://www.westbournegroup.com.au/

Weule, G. (2014, August). Passport officers poor at spotting fake ID photos. Retrieved from http://www.abc.net.au/science/articles/2014/08/19/4069251.htm

White, D**.**, Burton, A. M., Kemp, R. I., & Jenkins, R. (2013). Crowd effects in unfamiliar face matching. *Applied Cognitive Psychology, 27*(6), 769–777.

White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PLoS ONE 10*(10): e0139827. doi:10.1371/journal.pone.0139827.

White, D**.,** Kemp, R., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image comparison. *Psychonomic Bulletin & Review, 21* (1), 100–106.

White, D., Kemp, R., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PLOS ONE, 9* (8), 1–6. E103510.

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B: Biological Sciences, 282*(1814–1822).

Wicklin, R. (2010). *Statistical programming with SAS/IML software*. North Caroline, USA: SAS Institute Inc.

Wilkinson, C. (2012). Juvenile facial reconstruction. In C. Wilkinson & C. Rynn (Eds.) *Craniofacial identification* (pp. 254–260). Cambridge, New York: Cambridge University Press.

Wilkinson, C., & Evans, E. (2009). Are facial image analysis experts any better than the general public at identifying individuals from CCTV images? *Science and Justice 49*, 191–196.

Wilson, J. R. & Sharples, S. (2015). *Evaluation of human work* (4th ed.). Bosa Roca, United States: Taylor & Francis.

Yadav, D., Singh, R., Vatsa, M., & Noore, A. (2014). Recognising age-separated face images: Humans and Machines. *PLoS ONE, 9*(12), e112234.

Yiu, S. Y., McLindin, B., Malec, C., Bourn, S., & Hanton, K. (2015). *Multiple Facial Recognition Algorithm Test (FRAT) for the Department of Foreign Affairs and Trade (DFAT) – Testing methodology and metrics overview* (DSTO-GD-0852). Adelaide: Defence Science and Technology Group.

Yui, M. L., Bolme, D., Draper, B., Beveridge, J., Givens, G., & Phillips, P. J. (2009). A meta-analyses of face recognition covariates. *Proceedings of the IEEE Third International Conference on Biometrics: Theory, Applications, and Systems*, 1–8.

Zeng, J., Ling, H., Latecki, L. J., Fitzhugh, S., & Guo, G., (2012). Analysis of facial images across age progression by humans. *ISRN Machine Vision*, Article ID 505974.

Zhang, S. X. (2007). *Smuggling and human trafficking of human beings: All roads lead to America*. Westport, CT: Praeger Publishers.

Zhang, X., & Gao, Y. (2009). Face recognition across pose: A review. *Pattern Recognition, 4*(2), 2876–2896.

Appendices

Appendix A. Focus Group and Observations Information Sheet

INFORMATION SHEET

Requirements Collection

Brief Description of the Study

Face matching is a task that is necessary for many government agencies as part of processing, access control, or investigative applications. Many agencies have implemented face matching algorithms to assist human operators or replace some parts of the business process. Whilst the face matching performance of human operators and/or algorithms conducting face matching tasks with adult images is in many cases acceptable, it is currently not known how operators and algorithms perform when attempting to match faces of children.

This study is in support of a PhD research program investigating the performance of the human operators and commercial face matching algorithms when performing face matching tasks with a particular focus on children.

This study aims to understand the businesses processes and issues encountered by stakeholders and operators that manage and/or conduct facial matching tasks (human and/or operator) in real-world settings.

Your Part in the Study

You will be asked general questions regarding current business processes used in your agency with regards to face matching using human operators and/or face matching algorithms. If possible, you will also provide a demonstration of this business process. General questions will also be asked to identify what areas are believed to impact on human operators and/or face matching algorithm performance, particularly with children. This process is expected to take approximately one hour.

Risks of Participating

There are no risks to your health or wellbeing as a result of participating in this study. Any occupational health and safety issues will be identified on site and appropriate measures will be taken to control risks to participants. Participation is purely voluntary and you are free to withdraw at any time.

Statement of Privacy

All data collected during the experiment will be treated in the strictest confidence and stored on password protected computers. Furthermore, once the data is no longer required it will be destroyed. You will also have the opportunity to receive a summary of the research findings.

Investigator Contact Information:
Should you have any queries, complaints or concerns about the manner in which this project is conducted, please do not hesitate to contact the researchers in person, or you may prefer to contact the convener of the School of Psychology Human Ethics Subcommittee.

Dana Michalski
Psychology PhD Candidate
Ph. (08) 7389 7914
dana.michalski@dsto.defence.gov.au

Dr Paul Delfabbro
Convenor of the School of Psychology
Human Ethics Subcommittee
Ph. (08) 8303 5744
paul.delfabbro@adelaide.edu.au

Appendix B. Focus Group and Observations Consent Form

Australian Government
Department of Defence
Defence Science and
Technology Organisation

THE UNIVERSITY
OF ADELAIDE
AUSTRALIA

CONSENT FORM

Understanding the Business Processes of Face Matching in Operational Environments – Focus Groups / Demonstrations

1. I …..... hereby consent to participate in the above research project.

2. I have had the project, so far as it affects me, fully explained to my satisfaction by the researcher. My consent is given freely.

3. I am cooperating in this project on the condition that:
   - the information provided will be kept confidential
   - the research results will be made available to me at my request and any published reports of this study will preserve my anonymity.

4. I understand that:
   - I am free to withdraw from the project at any time with no detriment to my career or future health
   - I may not directly benefit from taking part in this research
   - A copy of this signed Consent Form and the Information Sheet will be provided to me if requested.

Participant's signature……Date……....

I certify that I have explained the study to the volunteer and consider that she/he understands what is involved and freely consents to participation.

Researcher's name…….....

Researcher's signature……..Date……

Appendix C. List of Example Focus Group Questions

1) For what purposes are facial comparisons conducted in this agency?
2) How are they conducted (i.e., automated, manually, or a combination)?
3) Can you explain the process of conducting facial comparisons in your agency?
4) What process is used (i.e., one-to-one or one-to-many)?
5) What sort of images are facial comparisons conducted on?
6) What quality are they?
7) What ages are the people in the images that facial comparisons are conducted on? (provide age range)
8) Why these particular ages?
9) What ages would you like to be able to conduct facial comparisons on? (just the ones mentioned or is there more?)
10) What age variations between images do you need to conduct facial comparisons on? (provide age variation range)
11) What ages are believed to be the most difficult to conduct facial comparisons on?
12) How long can staff spend manually comparing a pair of images?
13) Are staff ever deadlined?
14) Do you know how well staff perform when comparing images of children?
15) Do you know how well your facial recognition system performs with images of children?
16) What research do you think your agency would benefit from in regards to facial comparisons with images of children?
17) Are there any particular ages, age variations, or age groups that you would like practitioners and algorithms to be evaluated on?

Appendix D. Vendor Survey Questions

Facial Recognition Vendor Survey

This survey is in support of a PhD research program investigating the performance of automated facial recognition systems and the human operators performing face matching tasks with a particular focus on images of children's faces.

A key part of the early phases of this research is to gain an understanding of the current state of the art in facial recognition technology. Hence, this survey has been designed to gain a better understanding of your algorithms and what areas of DSTO's Collaborative Applied Research Program may be of most value to you. Participation is voluntary and information provided will remain anonymous.

It would be very useful if you could also supply any documents and/or research papers that you have access to involving the performance of your algorithms, especially those that include facial recognition of children.

When answering the questions, please identify if you are referring to one-to-one (1:1) verification performance or one-to-many (1:n) identification performance (or both).

Once completed, please return via post or email to:

Dana Michalski
Human Scientist / PhD Candidate
National Security and Intelligence, Surveillance and Reconnaissance Division
Defence Science and Technology Organisation
75 Labs
PO Box 1500
EDINBURGH
SOUTH AUSTRALIA 5111
dana.michalski@dsto.defence.gov.au

Please provide the following information:

Name:_____

Job Title/Level:_____

Company:_____

Area in Company:_____

Email:_____

1. Facial Recognition and Children

Many agencies are keen to implement automated facial recognition for children (defined here as those aged less than 18 years), but several have reported poor performance. We are interested in understanding in what ways images of children may impact on the performance of your algorithm(s).

1. Have customers mentioned their need or potential requirement to use automated facial recognition on children?

☐ Yes ☐ No

a) If so, in what contexts? (i.e. for what applications).

2. Are there differences in the performance of your facial recognition algorithm(s) when used to match images of children compared to when it is used to match images of adults?

☐ Yes ☐ No

a) If yes, please explain what these differences in performance are (both for 1:1 and 1:n).

b) If yes, do you know any ways to improve the performance of your algorithm(s) in this regard?

3. Has any research/testing been conducted to investigate the performance of <u>your</u> algorithm(s) when used with images of children?

☐ Yes ☐ No

a) If yes, please explain this research/testing including who conducted it (and where possible provide any findings, documents, and references). Was this 1:1, 1:n, or both?

4. Are you satisfied with the amount of research/testing that has been done in this regard?

☐ Yes ☐ No

a) **Please explain.**

5. **What ages less than 18 years (if any) does your algorithm(s) perform best on?**

a) **Please explain (if known) the reason for this better performance?**

6. **What ages less than 18 years (if any) does your algorithm(s) perform worst on?**

a) **Please explain (if known) the reason for this poorer performance?**

2. Age Separation (questions in this section refer to all ages, not just children)

Age separation refers to the period of chronological time that has lapsed between two or more images of the same person. We are interested in understanding in what ways age separation may impact on facial recognition algorithm performance.

7. **Are you aware of any limitations of your facial recognition algorithm(s) due to age separation (i.e. the number of years between when images of the same person were taken)?**

☐ Yes ☐ No

a) **If yes, please explain what these limitations are. Is this for 1:1 or 1:n?**

b) **If yes, do you know any way to improve the performance of your algorithm(s) in this regard?**

8. Has any research/testing been conducted to investigate how age separation (i.e. the number of years between when images of the same person were taken) affects the performance of your algorithm(s)?

☐ Yes ☐ No

a) If yes, please explain this research/testing including who conducted it (and where possible provide any findings, documents, and references).

9. Are you satisfied with the amount of research/testing that has been done in this regard?

☐ Yes ☐ No

a) Please explain.

3. Age Estimation (questions in this section refer to all ages)

We are interested in understanding the extent to which age estimation is a current customer requirement, as well as the availability of age estimation algorithms and their performance.

10. Have any customers mentioned their need or potential requirement to use age estimation on the faces within images?

☐ Yes ☐ No

a) If so, in what contexts? (i.e. for what applications?).

11. Do you have any algorithms that estimate the age of a person in an image?

☐ Yes ☐ No (go to question 16)

a) If yes, please explain if an exact age is provided or an age range (e.g. +/-5 years, variable, etc.).

12. Are there any minimum or maximum ages that the algorithm(s) can estimate?

☐ Yes ☐ No

a) If yes, please provide details.

13. **What research/testing has been conducted to evaluate the performance of your age estimation algorithm(s)? (Please attach any findings or references available).**

a) **Are you satisfied with the amount of research/testing that has been done in this regard?**

☐ Yes ☐ No

b) **Please explain.**

14. What ages does the algorithm(s) perform best on in regards to age estimation?

a) What is the reason (if known) for your algorithm's better performance on these ages?

15. What ages does the algorithm(s) perform worst on in regards to age estimation?

a) What is the reason (if known) for your algorithm(s) poorer performance on these ages?

4. Age Progression (questions in this section refer to all ages)

Age progression is the process of modifying a photograph of a person to represent the effect of ageing on their appearance. This process may be conducted manually by an artist using software tools such as Photoshop or by using automatic ageing software.

16. Have any of your algorithms been used on aged progressed images?

☐ Yes ☐ No

a) If yes, please explain. (e.g. how the imagery was manipulated, the ages of people in the original images, what ages they were aged progressed to, the results etc.) Please provide any findings or references available.

17. Do you have any age progression software?

☐ Yes ☐ No

a) If yes, please explain.

b) If yes, please provide details of the performance of this software.

5. Other

18. What information would you like to know regarding the performance of your algorithm(s) in the context of facial recognition on children specifically?

19. What other information would you like to know regarding the performance of your algorithm(s) in the context of facial recognition?

20. Please share any additional comments about any of the questions above or other thoughts.

Please attach any documents involving the performance of your algorithm(s), especially with regards to children

Thank you for taking the time to fill out this survey. Your input is greatly appreciated.

Appendix E. Failure-to-Enrol Rates

Table E1. Failure-to-Enrol Rates for each State-of-the-Art Algorithm by Different Groups (%)

Group	Algorithm A	Algorithm B	Algorithm C	Algorithm D	Algorithm E
Child	0.194	0.096	0.039	0.00012	0.177
Adult	0.008	0.103	0.015	0.00002	0.033
Age 0	1.498	0.590	0.246	0.001	1.165
Age 1	0.068	0.057	0.008	0	0.149
Age 2	0.041	0.079	0.016	0	0.095
Age 3	0.026	0.068	0.010	0	0.087
Age 4	0.009	0.042	0.008	0	0.063
Age 5	0.011	0.025	0.008	0	0.042
Age 6	0.005	0.016	0.008	0	0.035
Age 7	0.005	0.014	0.008	0	0.027
Age 8	0.004	0.015	0.009	0	0.028
Age 9	0.005	0.012	0.004	0	0.017
Age 10	0.003	0.010	0.006	0	0.016
Age 11	0.003	0.013	0.008	0	0.021
Age 12	0.005	0.016	0.011	0	0.016
Age 13	0.004	0.023	0.013	0	0.017
Age 14	0.006	0.026	0.016	0	0.021
Age 15	0.003	0.029	0.018	0	0.021
Age 16	0.003	0.022	0.010	0	0.019
Age 17	0.002	0.016	0.009	0	0.018

Child data is for all images of children (<18 years). Adult data is for all images containing a person above 18 years. The results show that images of children fail to enrol more than images of adults. Child images are then divided by age and shows that images of babies fail to enrol more than any other age.

Algorithm F (OpenBR) enrolled all images (i.e., the FTE is 0 for all rows).

Appendix F. Notched Boxplot Description

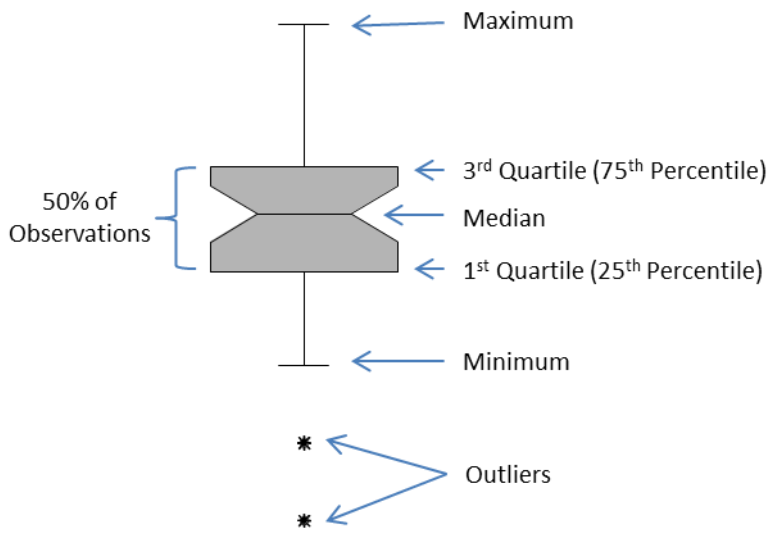


Figure F1. Diagram of a notched boxplot.

As can be seen, several components make up a notched boxplot:

Median - the middle quartile is represented by a horizontal line inside the box. Half of observations are greater than the median and half are below. Fifty percent of all observations are contained within the box;

1st quartile - the bottom line of the box, the 25th percentile i.e. 25% of the data has values below this line;

3rd quartile - the top line of the box, the 75th percentile i.e. 25% of the data has values above this line;

Minimum - at the end of the lower "whisker", the minimum data value;

Maximum - at the end of the upper "whisker", the maximum data value; and

Outliers - are any observations that differs a substantial amount from the rest of the data

Appendix G. Number of Mated Image Pairs used in Study 3A in each of the 198 Categories

Table G1. Number of mated image pairs used in each of the 198 categories. Approximately 30 times more non-mated pairs were used for each category (subject to variations based on failure-to-enrol rates by each algorithm).

		Age Variation (Years)										
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Youngest Age of Child in Image Pair (Years)	**0**	1845	3867	5659	7422	80555	83834	29250	14408	10551	21063	17024
	1	479	733	1185	2338	31699	42708	19556	10894	7382	11344	9911
	2	264	415	1366	1534	16369	21686	10502	6397	4352	5975	5087
	3	262	1012	1273	1141	15033	19429	9694	5731	3870	5501	4839
	4	360	713	674	1097	21072	21938	9590	5401	4076	6673	5299
	5	517	651	676	1338	37628	40762	13764	6847	5454	12567	9046
	6	420	550	582	1035	32350	38938	14687	8242	6003	11427	7927
	7	343	469	540	966	27054	34215	15359	8744	5937	8500	8068
	8	294	386	453	777	26117	32969	15026	8132	4835	9019	8103
	9	358	445	548	1006	31935	35136	14332	6746	6088	10461	8362
	10	464	489	581	1207	40236	40323	13168	8340	7365	12980	10004
	11	416	515	694	1572	39948	38611	17577	10722	8598	14162	9828
	12	396	494	629	1386	29764	37624	17247	9878	8045	10444	6006
	13	397	613	784	1508	34620	40392	17594	11402	7488	5725	2937
	14	527	776	944	3584	42574	44309	21764	12078	7462	4616	2937
	15	637	911	2002	5371	50158	52879	22469	12707	7853	5248	3385
	16	655	1469	3138	5830	50592	47476	20174	11460	7205	4698	3074
	17	1117	2522	3321	6227	41946	39711	16697	9168	5691	3672	2561

Appendix H. DETs for Algorithms A, B, C, D, and F displaying how Age Impacts on Performance for Age Variations 0–10 Years

See Section 3.5.1 for an explanation on how to interpret DETs.

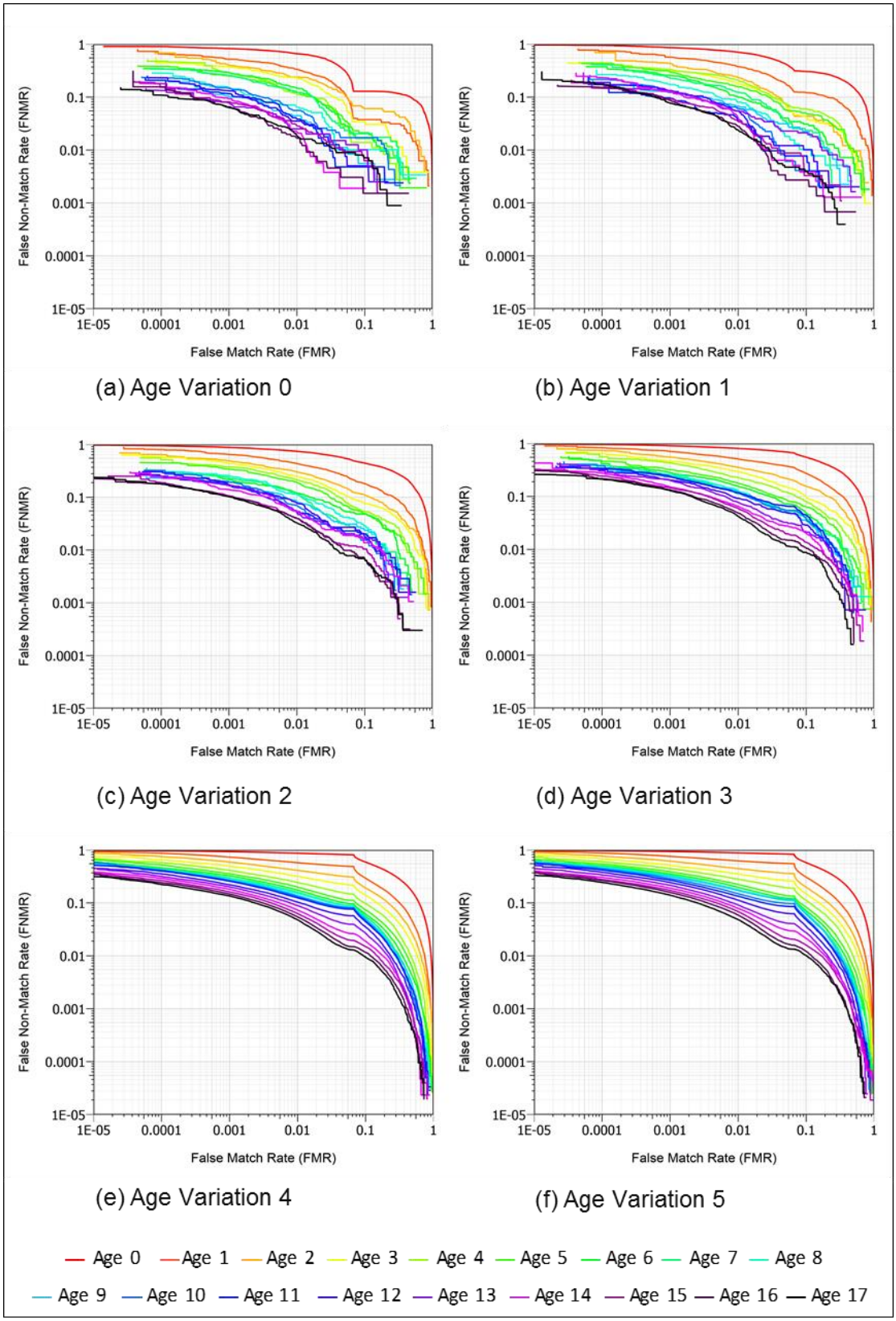


Figure H1. DETs for Algorithm A displaying how age impacts on performance for age variations spanning 0–5 years.

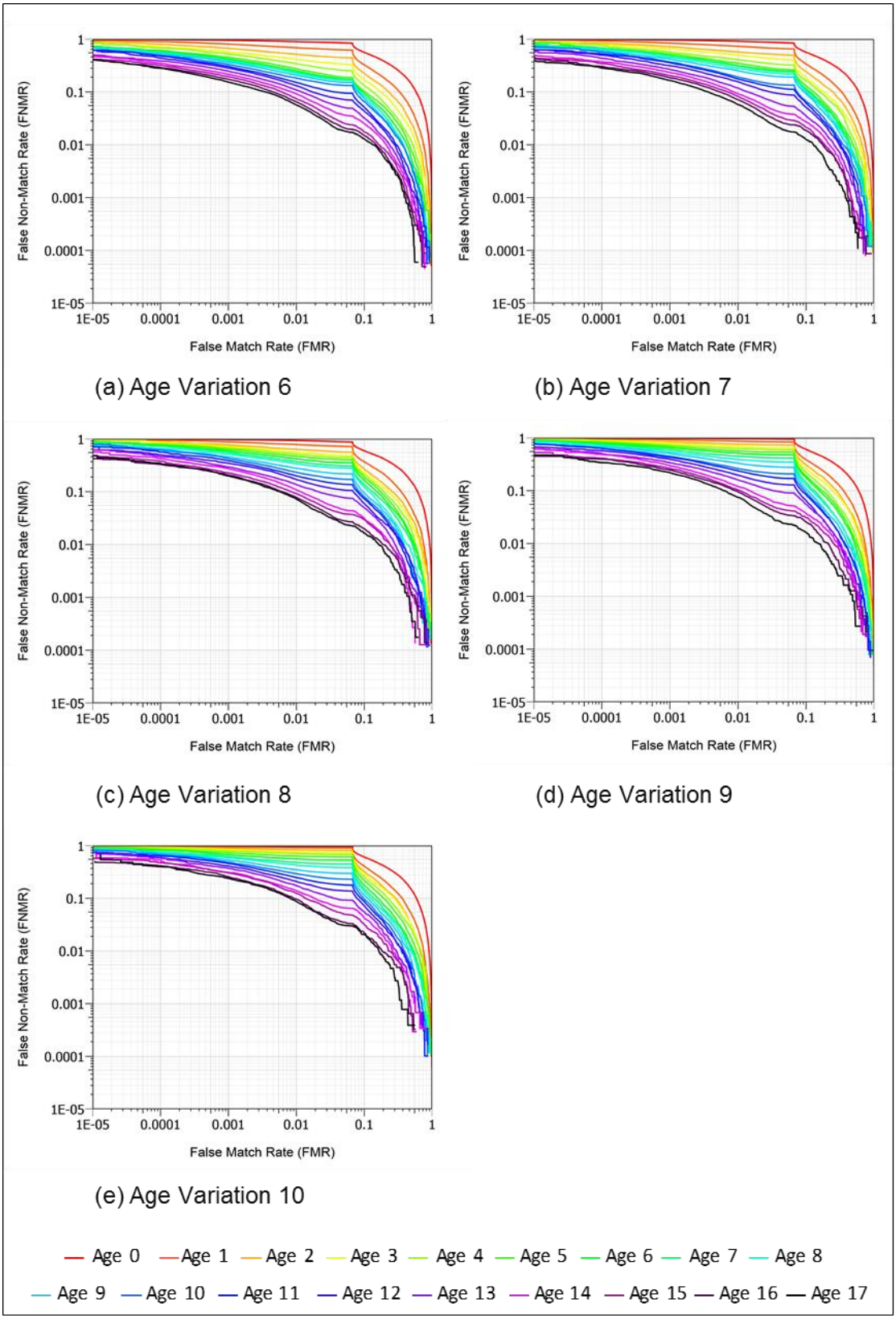


Figure H2. DETs for Algorithm A displaying how age impacts on performance for age variations spanning 6–10 years.

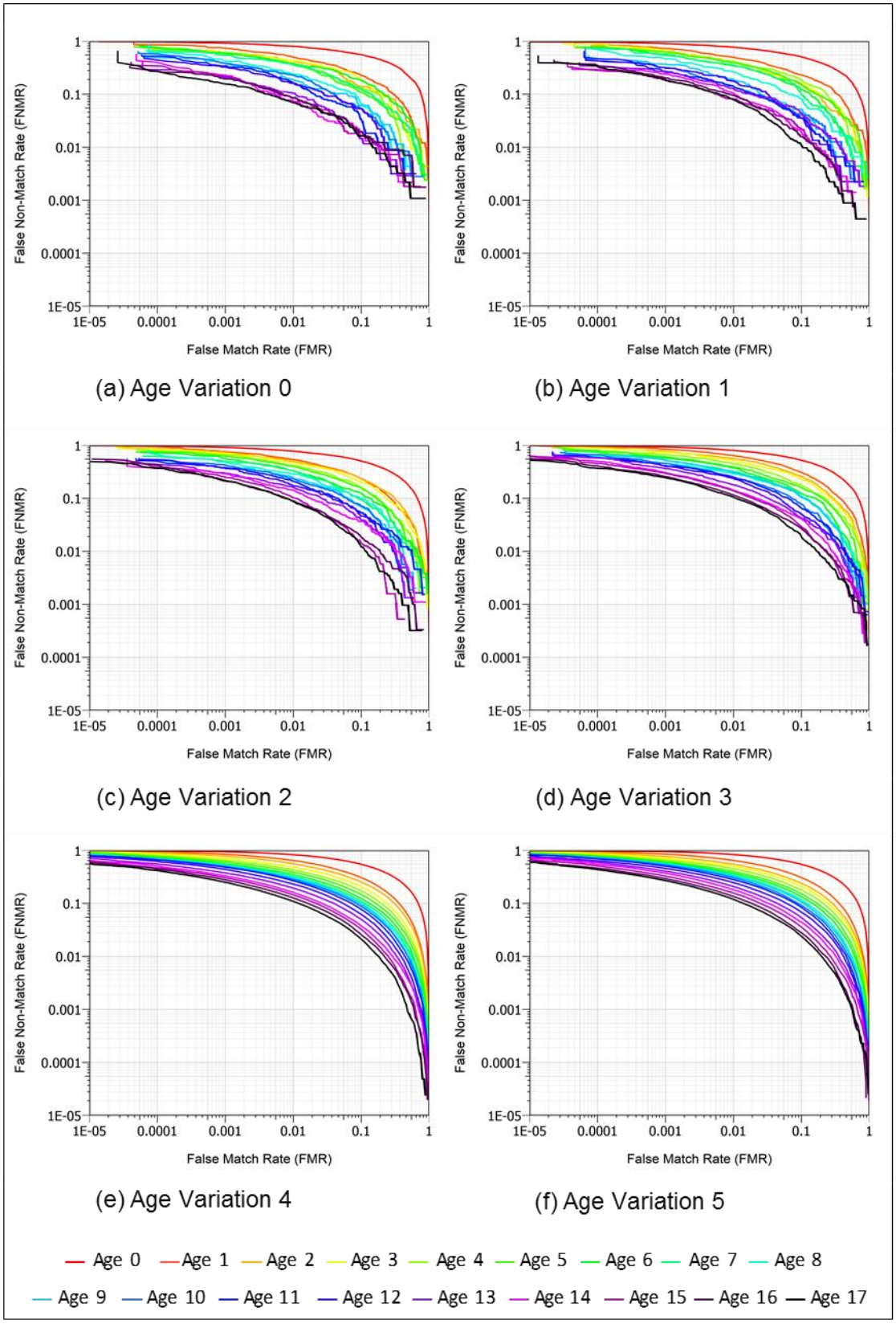


Figure H3. DETs for Algorithm B displaying how age impacts on performance for age variations spanning 0–5 years.

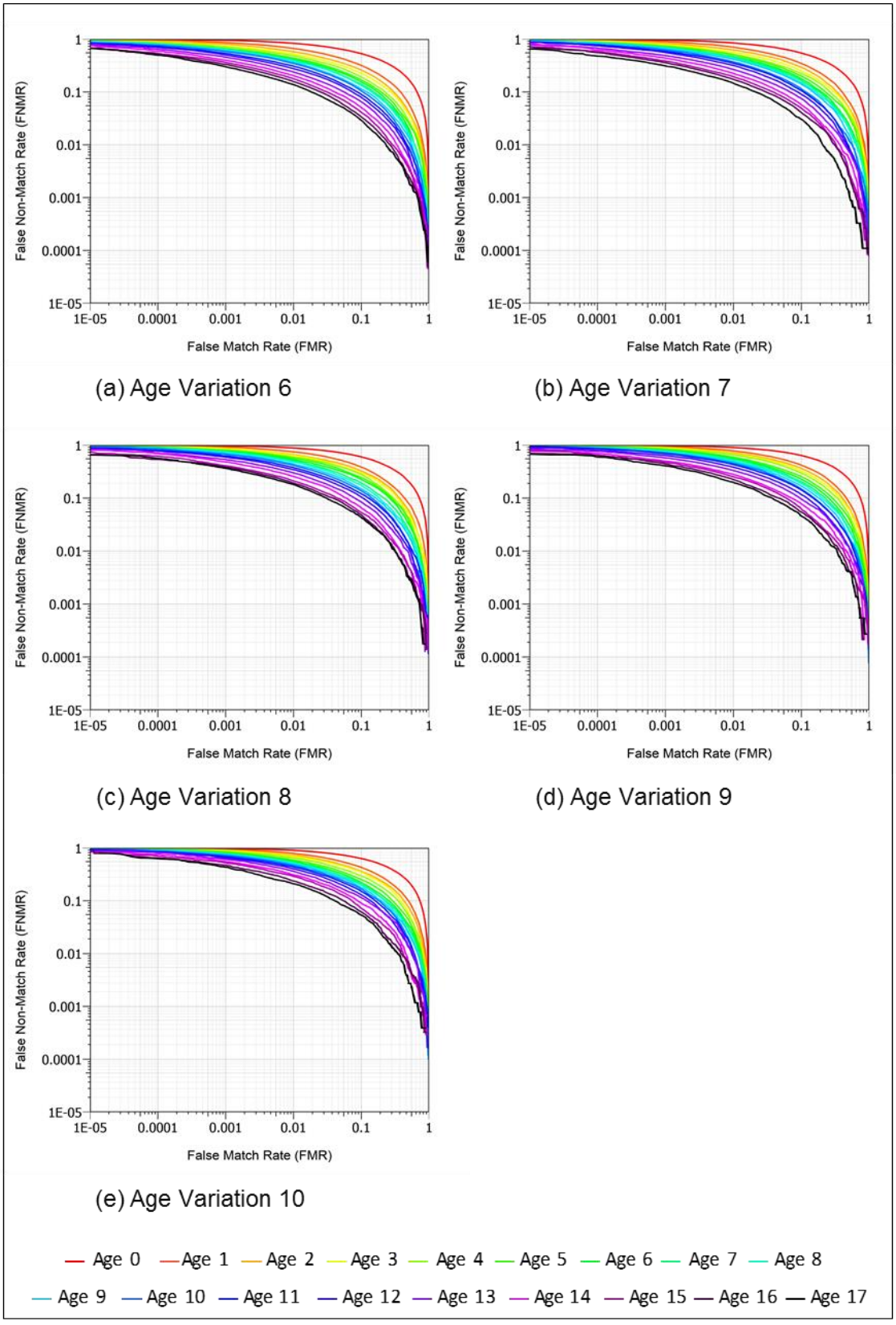


Figure H4. DETs for Algorithm B displaying how age impacts on performance for age variations spanning 6–10 years.

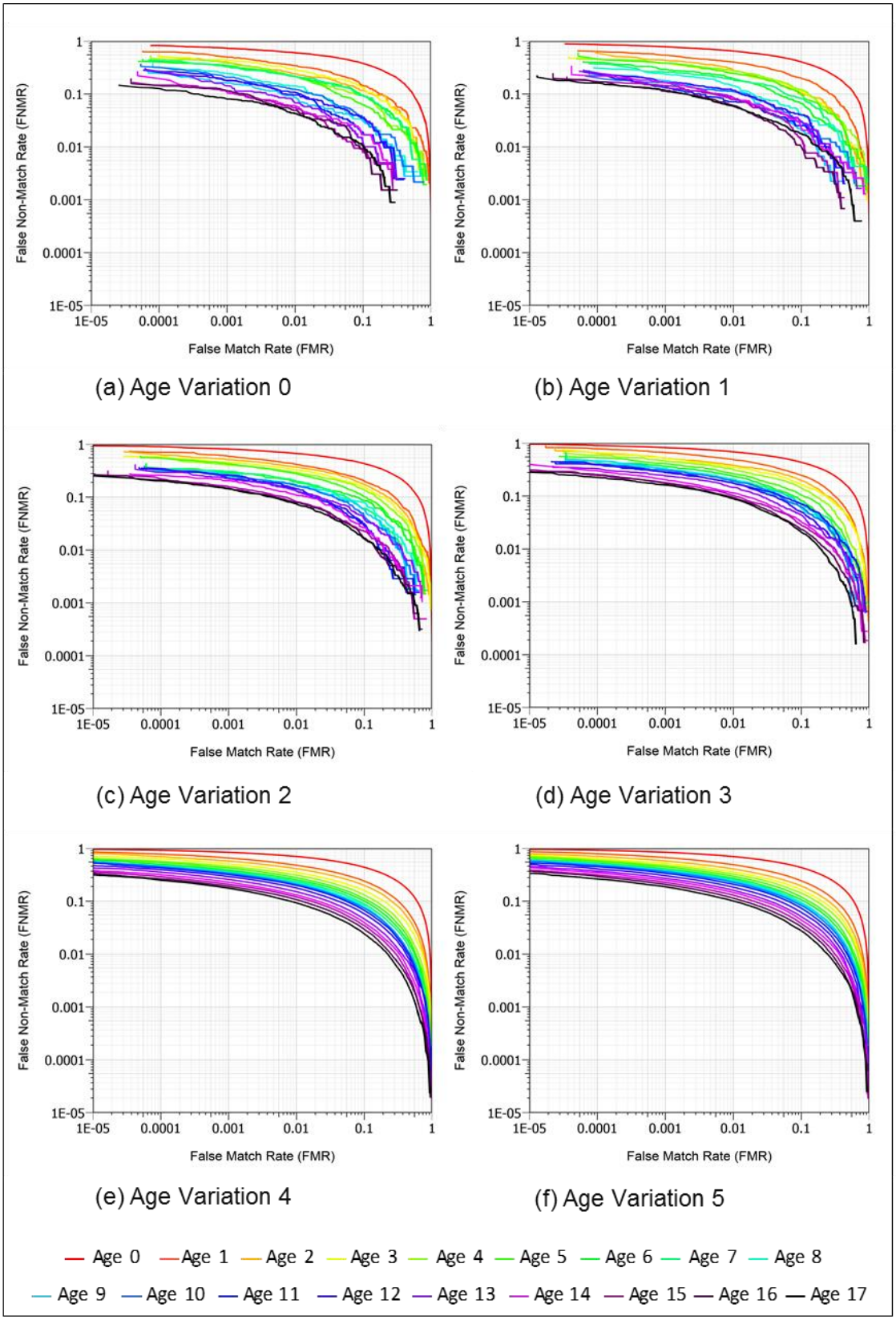


Figure H5. DETs for Algorithm C displaying how age impacts on performance for age variations spanning 0–5 years.

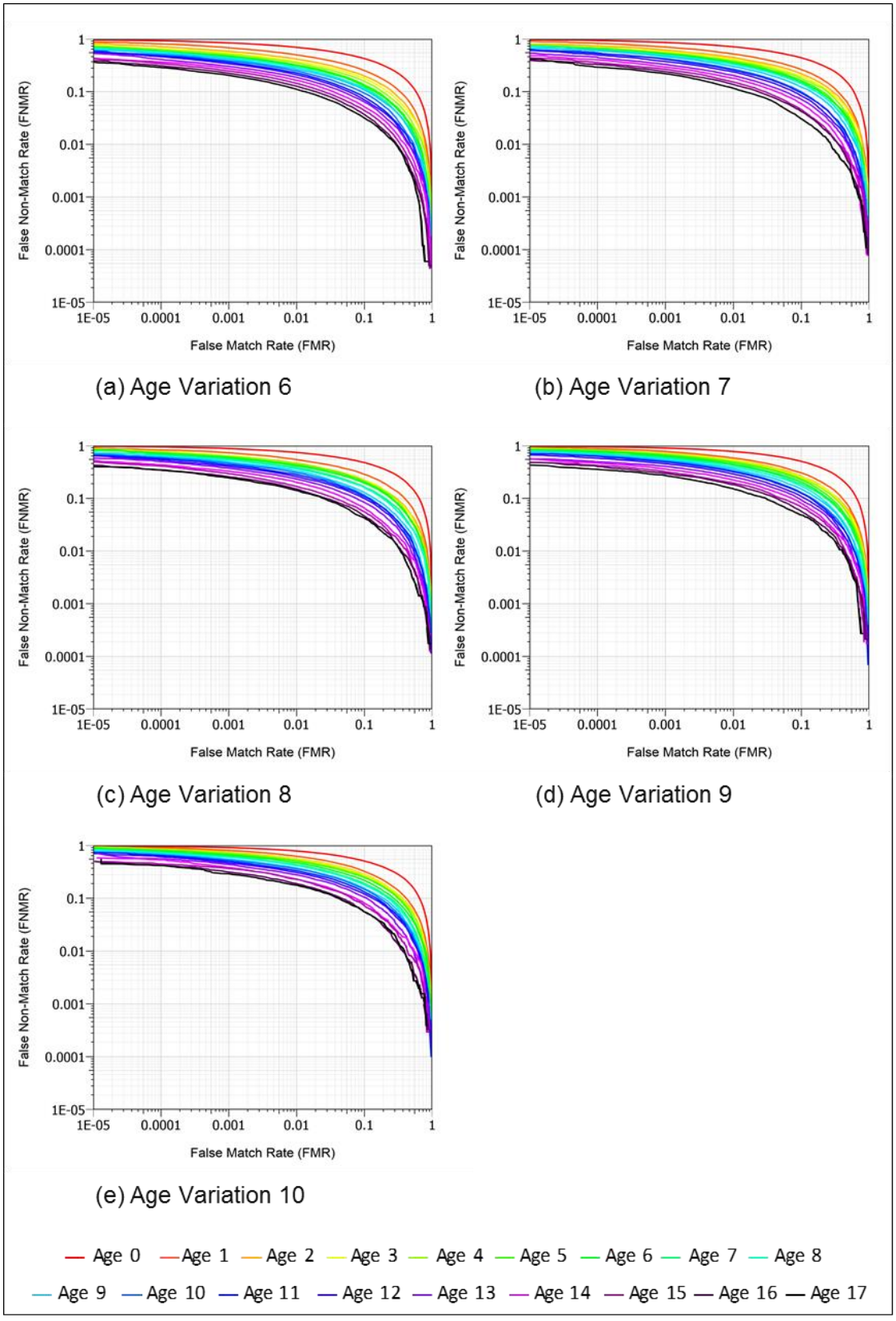


Figure H6. DETs for Algorithm C displaying how age impacts on performance for age variations spanning 6–10 years.

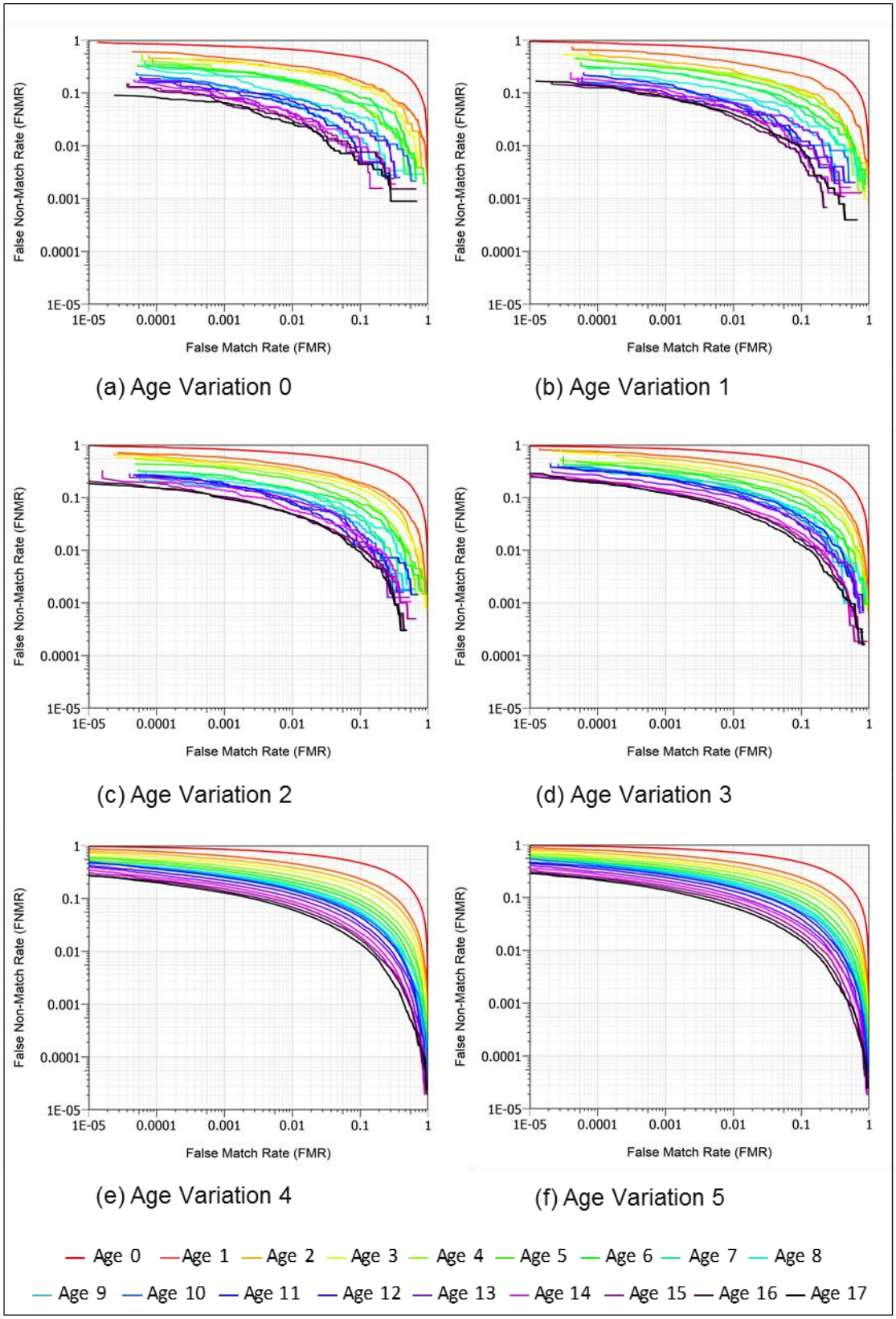


Figure H7. DETs for Algorithm D displaying how age impacts on performance for age variations spanning 0–5 years.

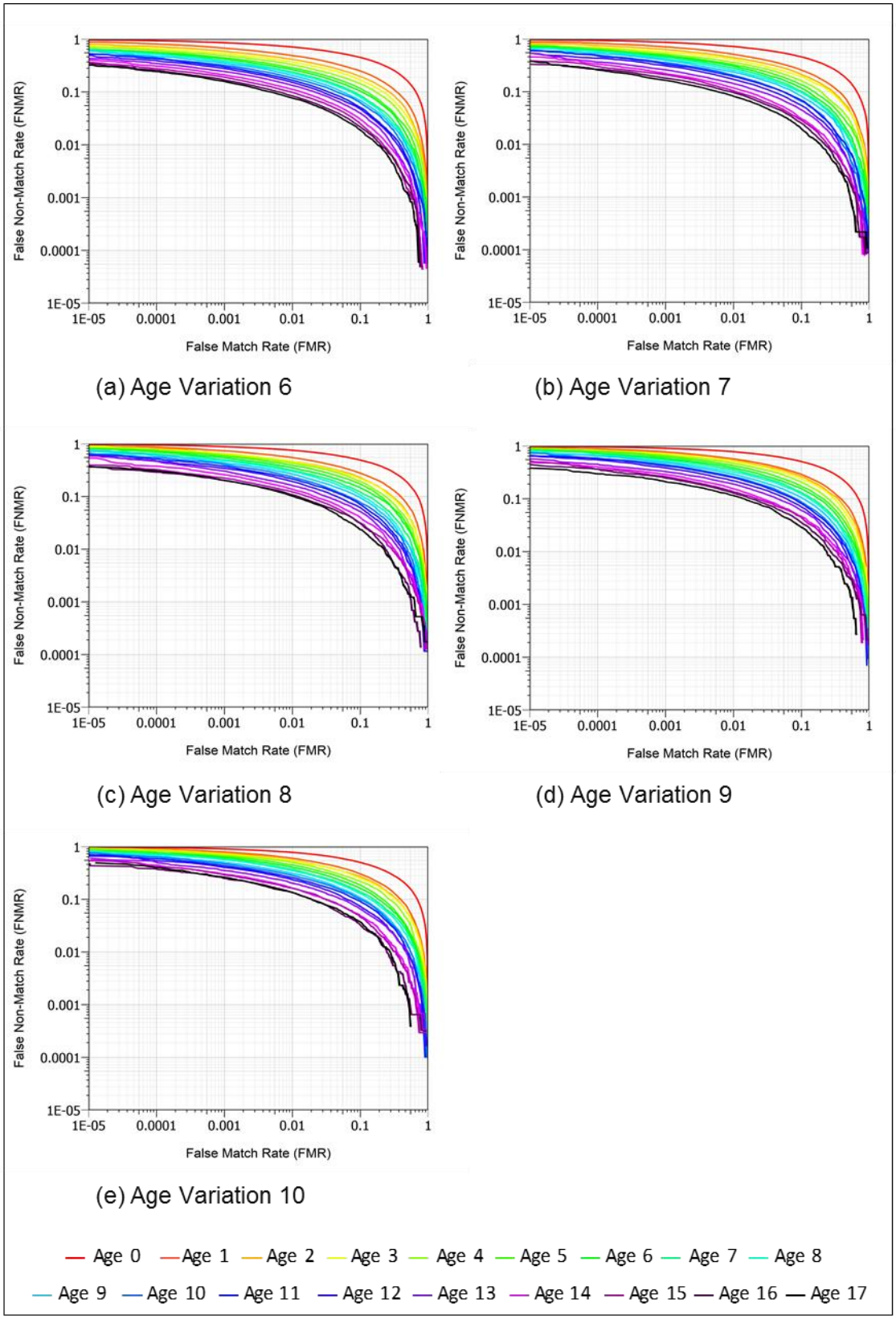


Figure H8. DETs for Algorithm D displaying how age impacts on performance for age variations spanning 6–10 years.

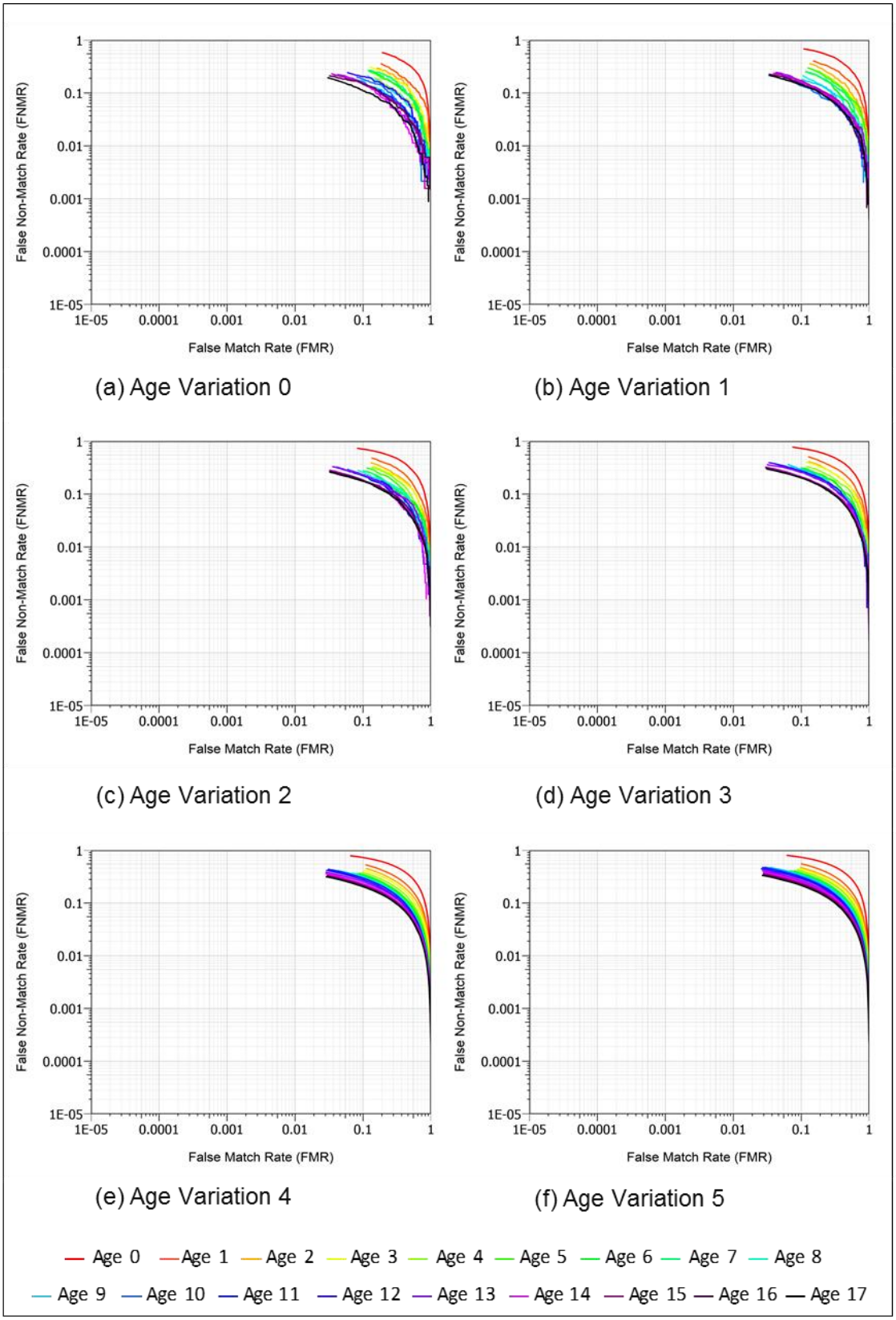


Figure H9. DETs for Algorithm F displaying how age impacts on performance for age variations spanning 0–5 years.

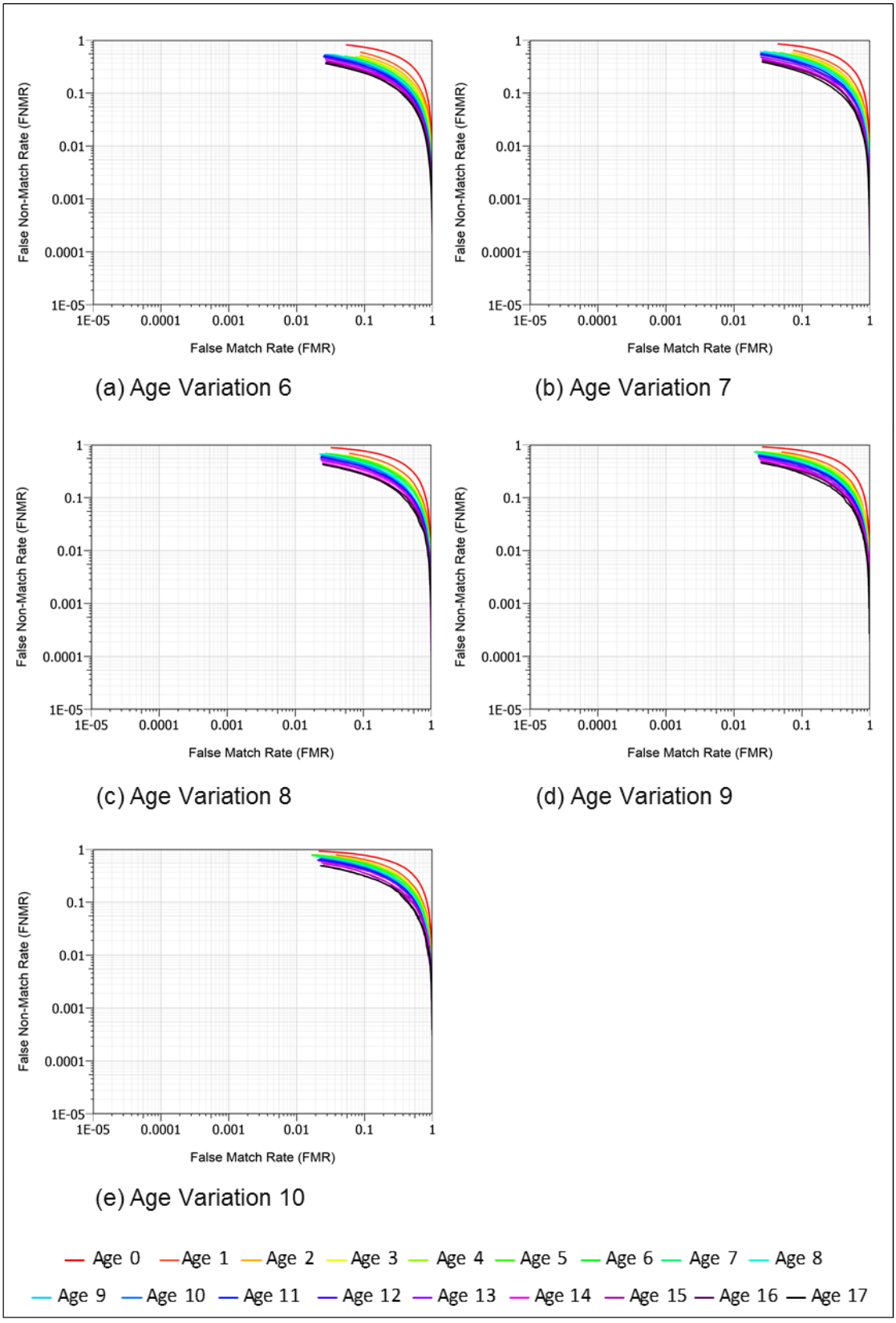


Figure H10. DETs for Algorithm F displaying how age impacts on performance for age variations spanning 6–10 years.

Appendix I. DETs for Algorithms A, B, C, D, and F displaying how Age Variation Impacts on Performance for Ages 0–17 Years

See Section 3.5.1 for an explanation on how to interpret DETs.

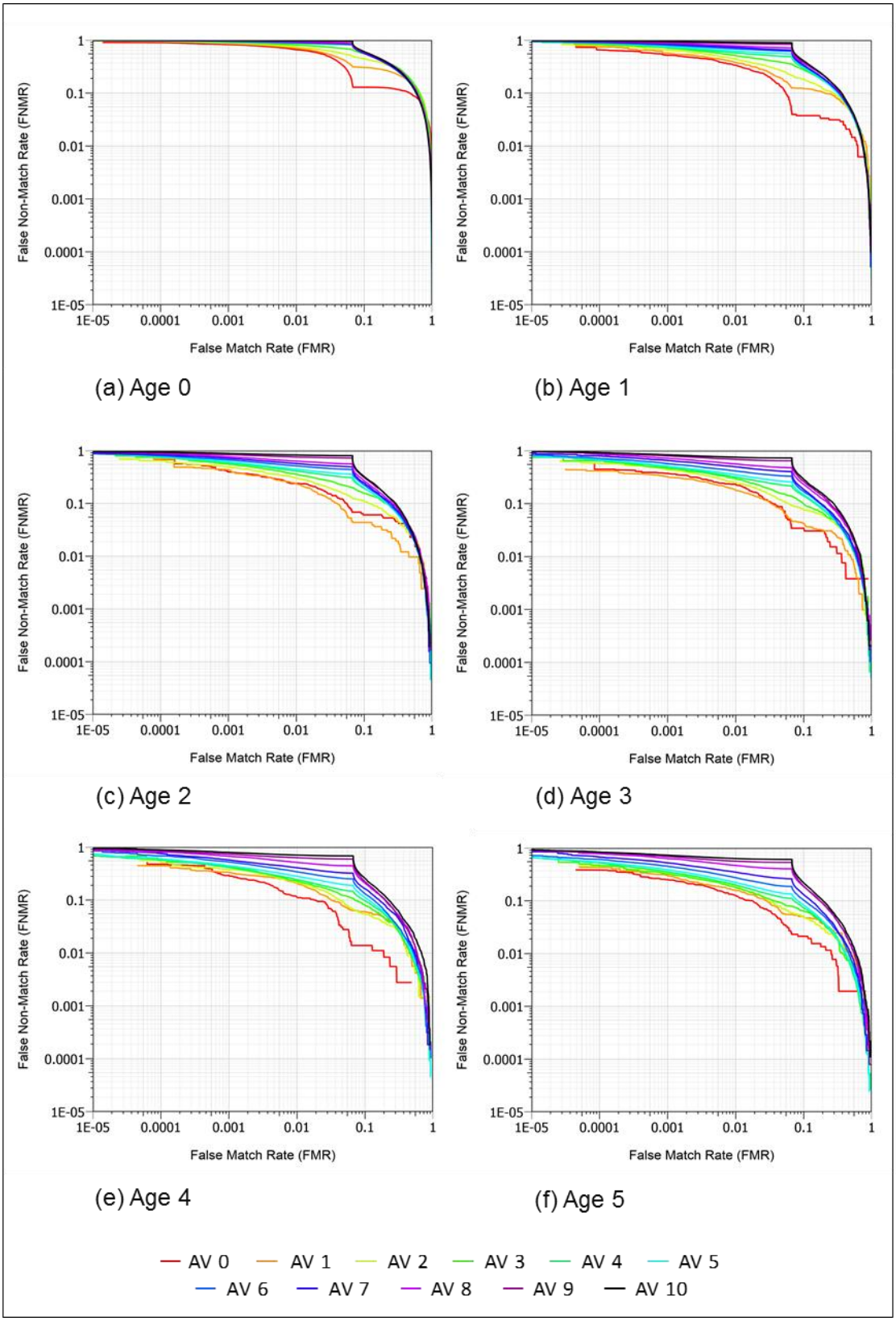


Figure I1. DETs for Algorithm A displaying how age variations impact on performance for ages 0–5 (AV = age variation in years).

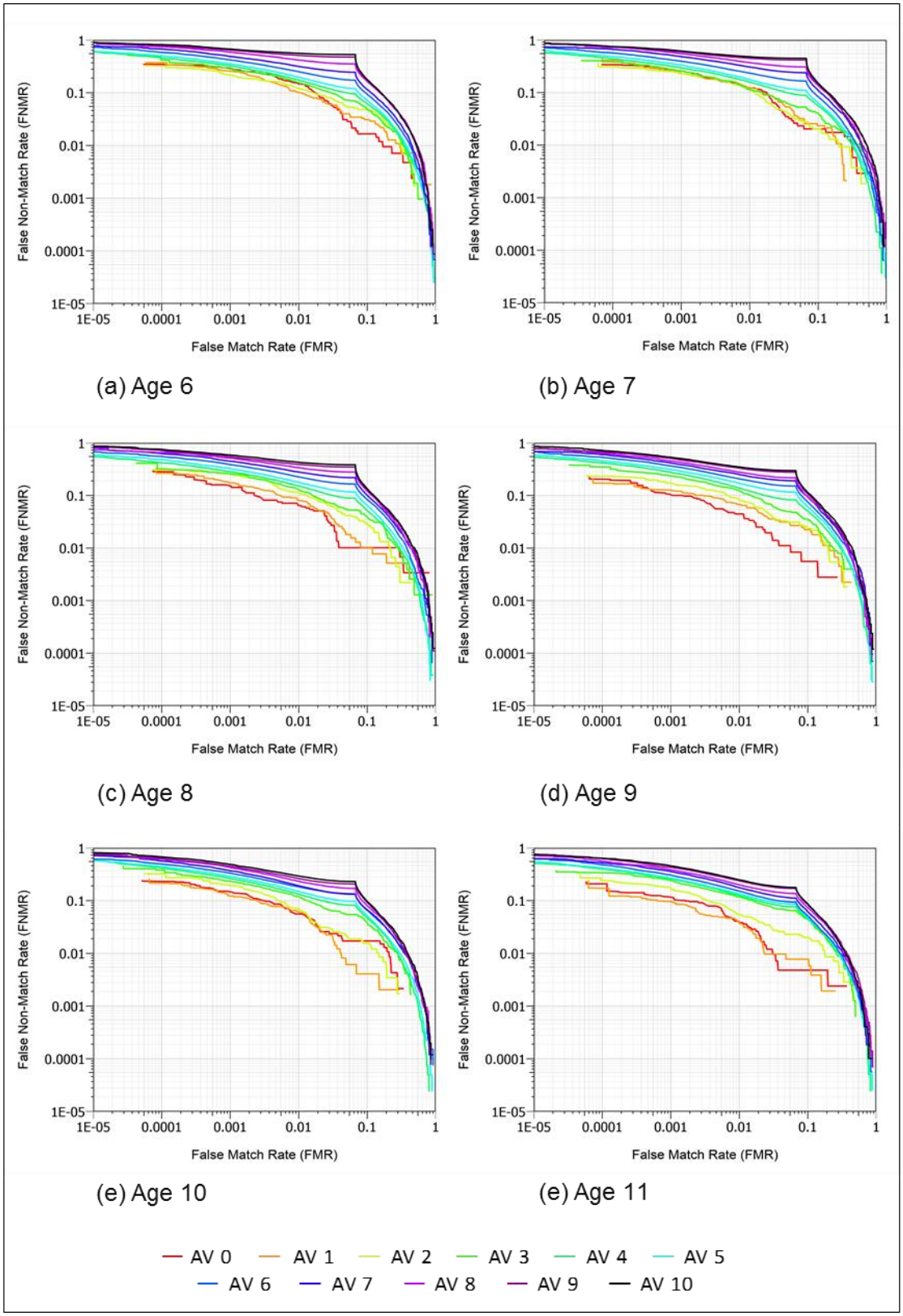


Figure I2. DETs for Algorithm A displaying how age variations impact on performance for ages 6–11 (AV = age variation in years).

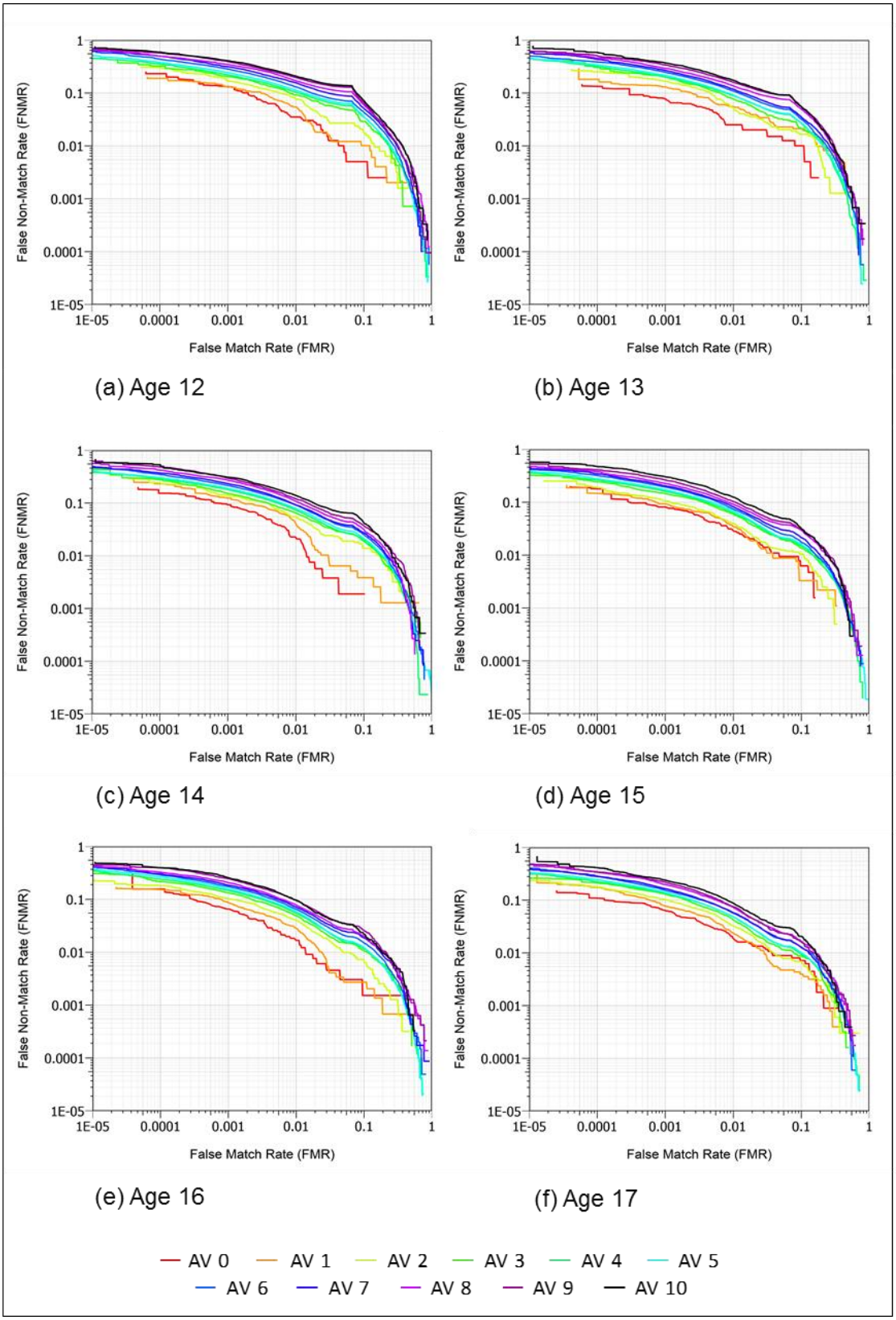


Figure I3. DETs for Algorithm A displaying how age variation impacts on performance for ages 12–17 (AV = age variation in years).

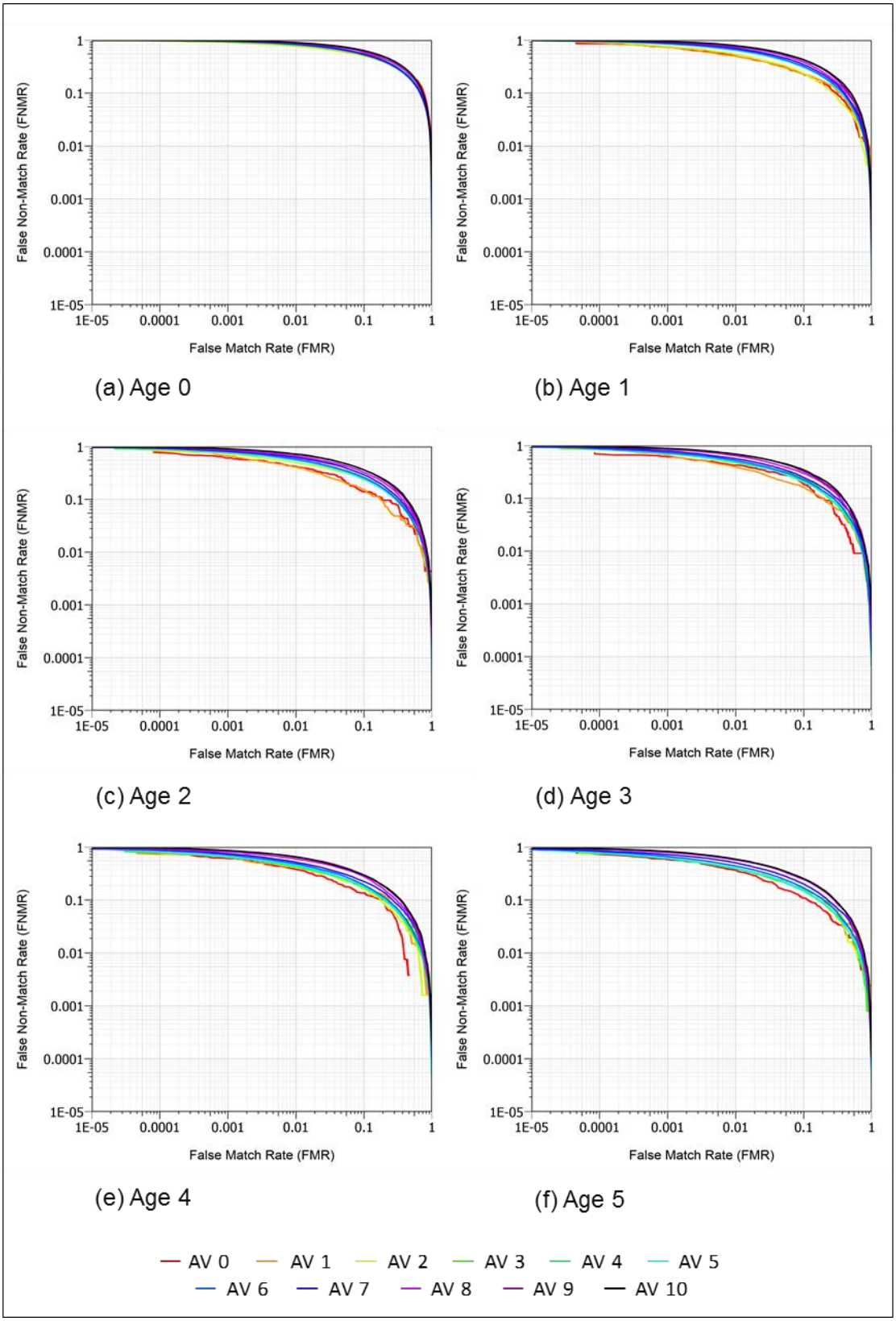


Figure I4. DETs for Algorithm B displaying how age variations impact on performance for ages 0–5 (AV = age variation in years).

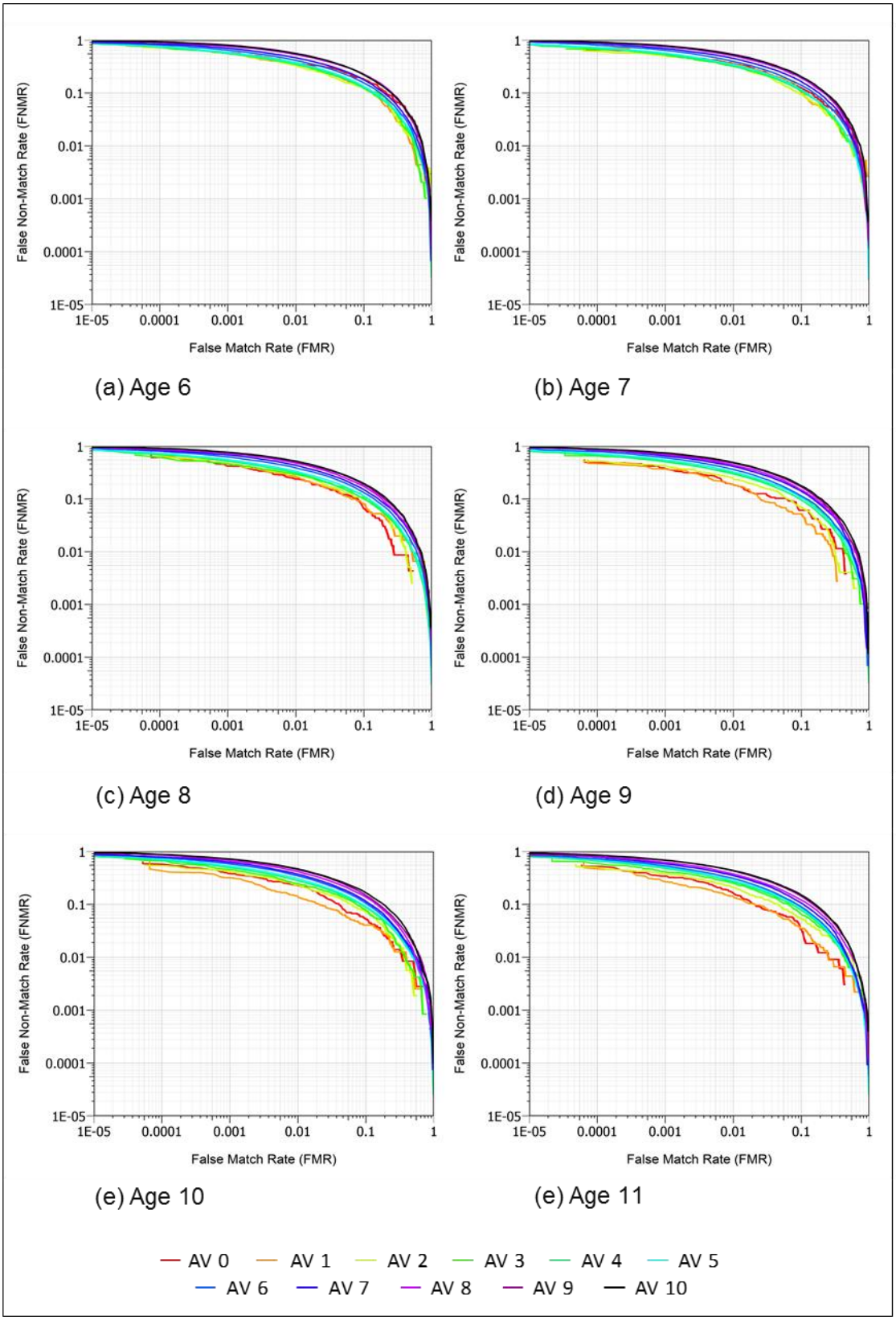


Figure I5. DETs for Algorithm B displaying how age variations impact on performance for ages 6–11 (AV = age variation in years).

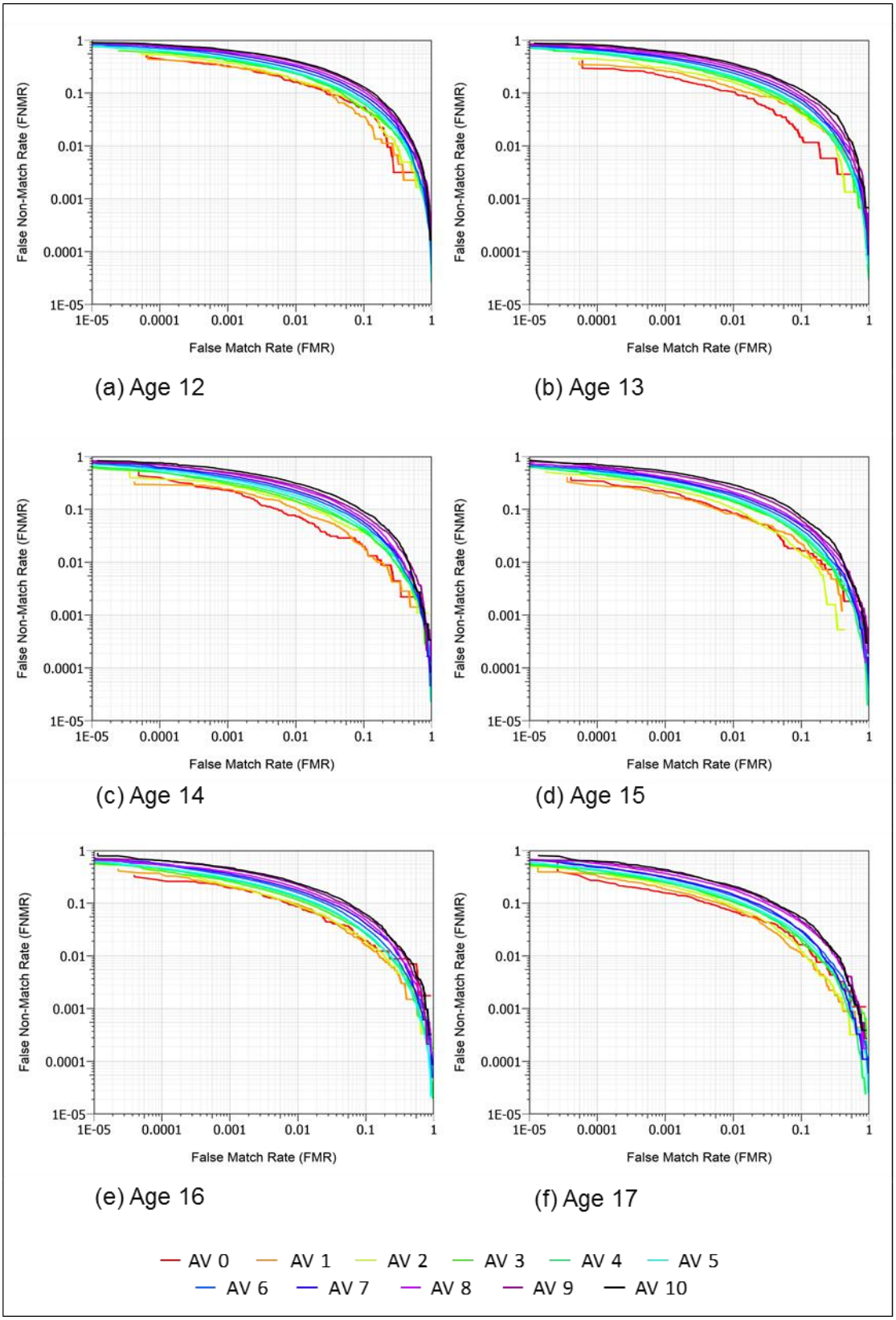


Figure I6. DETs for Algorithm B displaying how age variation impacts on performance for ages 12–17 (AV = age variation in years).

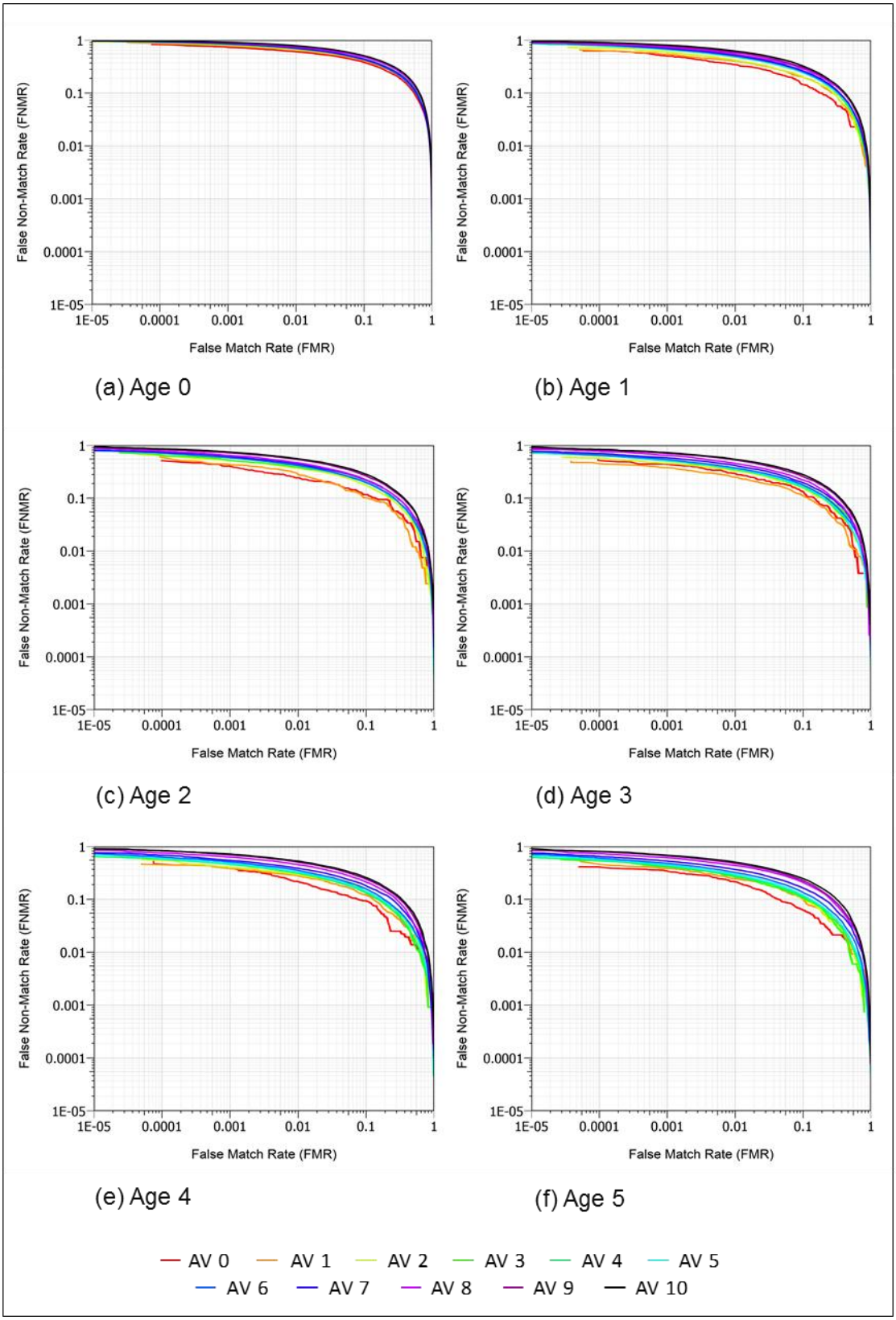


Figure I7. DETs for Algorithm C displaying how age variations impact on performance for ages 0–5 (AV = age variation in years).

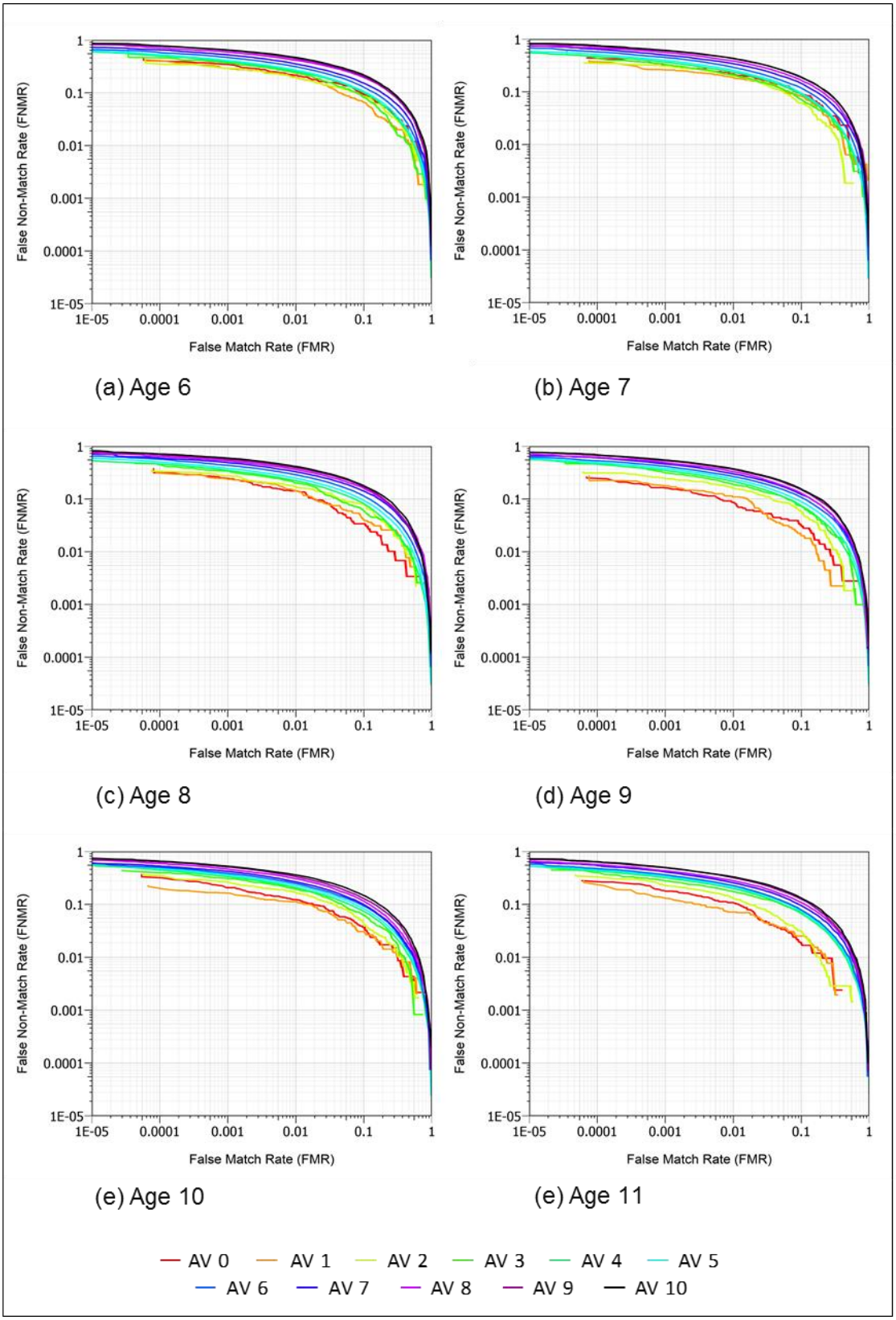


Figure I8. DETs for Algorithm C displaying how age variations impact on performance for ages 6–11 (AV = age variation in years).

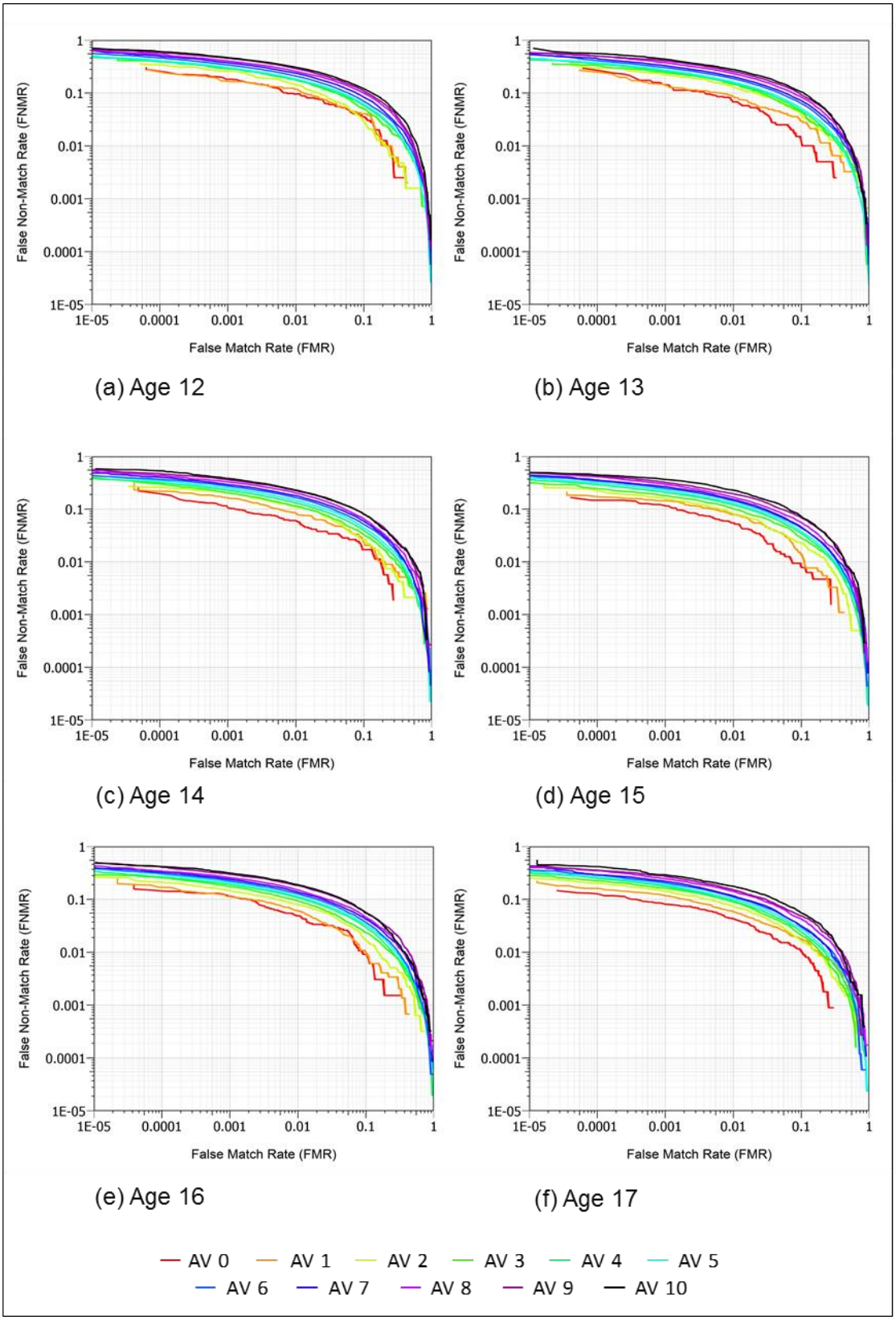


Figure I9. DETs for Algorithm C displaying how age variation impacts on performance for ages 12–17 (AV = age variation in years).

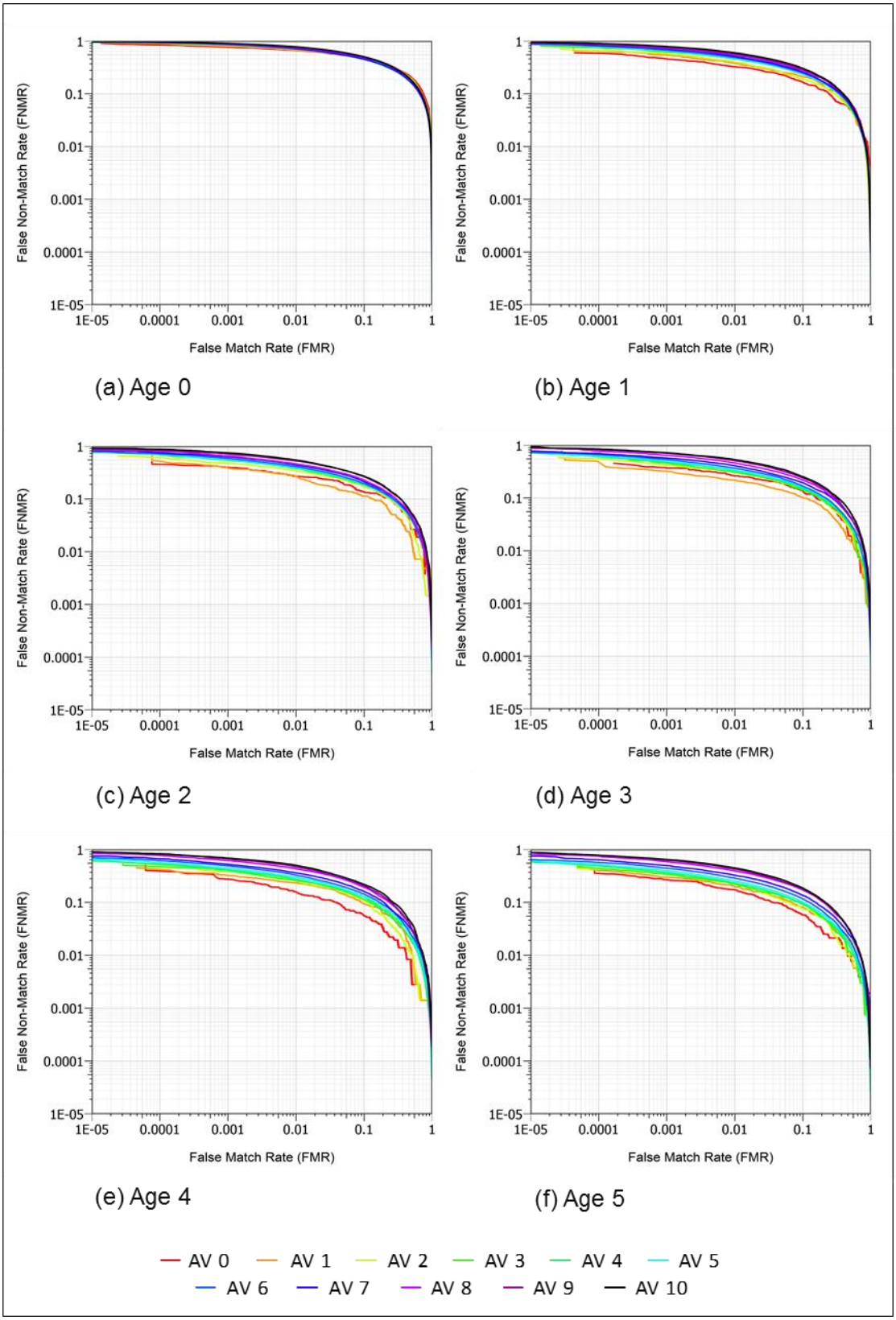


Figure I10. DETs for Algorithm D displaying how age variations impact on performance for ages 0–5 (AV = age variation in years).

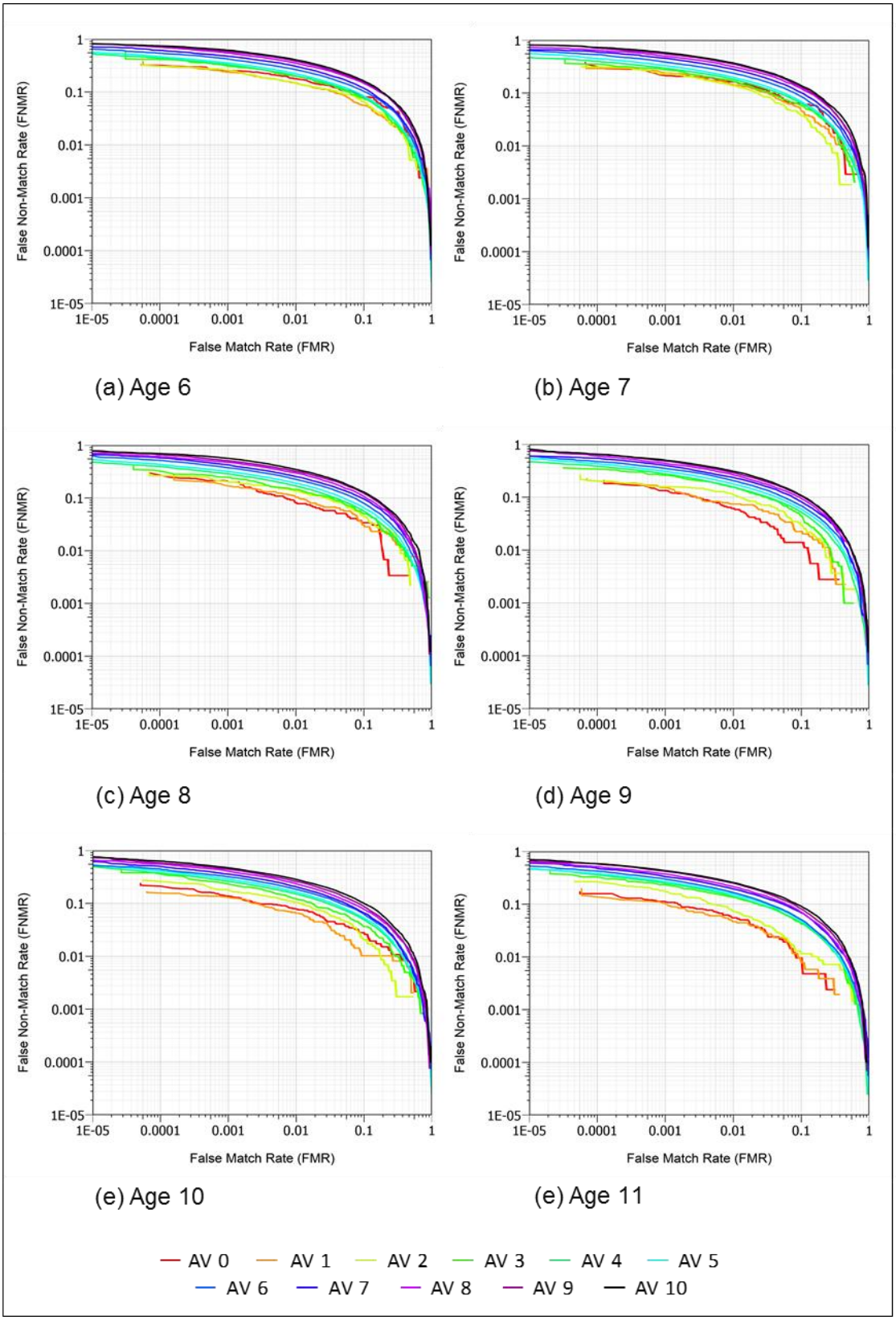


Figure I11. DETs for Algorithm D displaying how age variations impact on performance for ages 6–11 (AV = age variation in years).

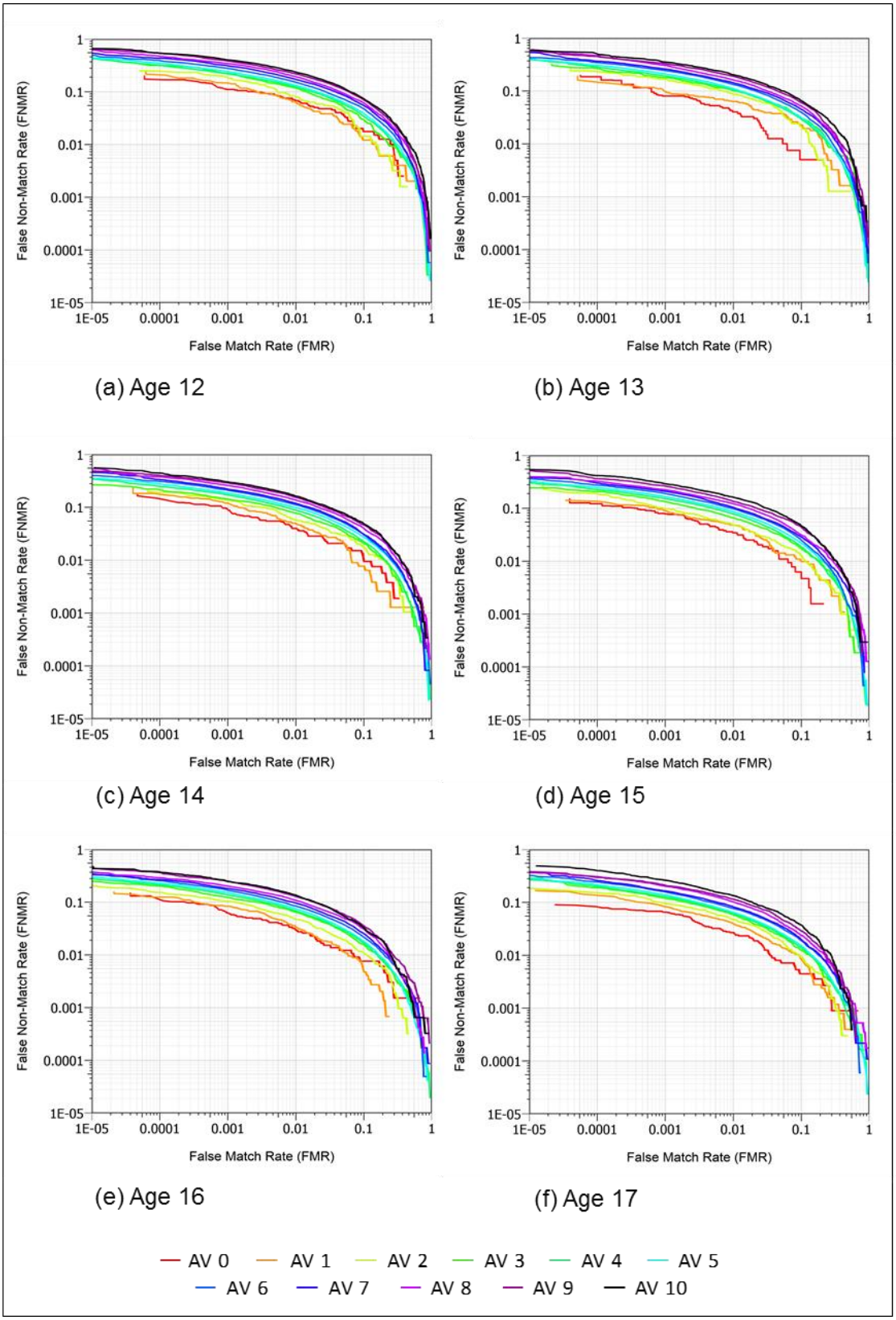


Figure I12. DETs for Algorithm D displaying how age variation impacts on performance for ages 12–17 (AV = age variation in years).

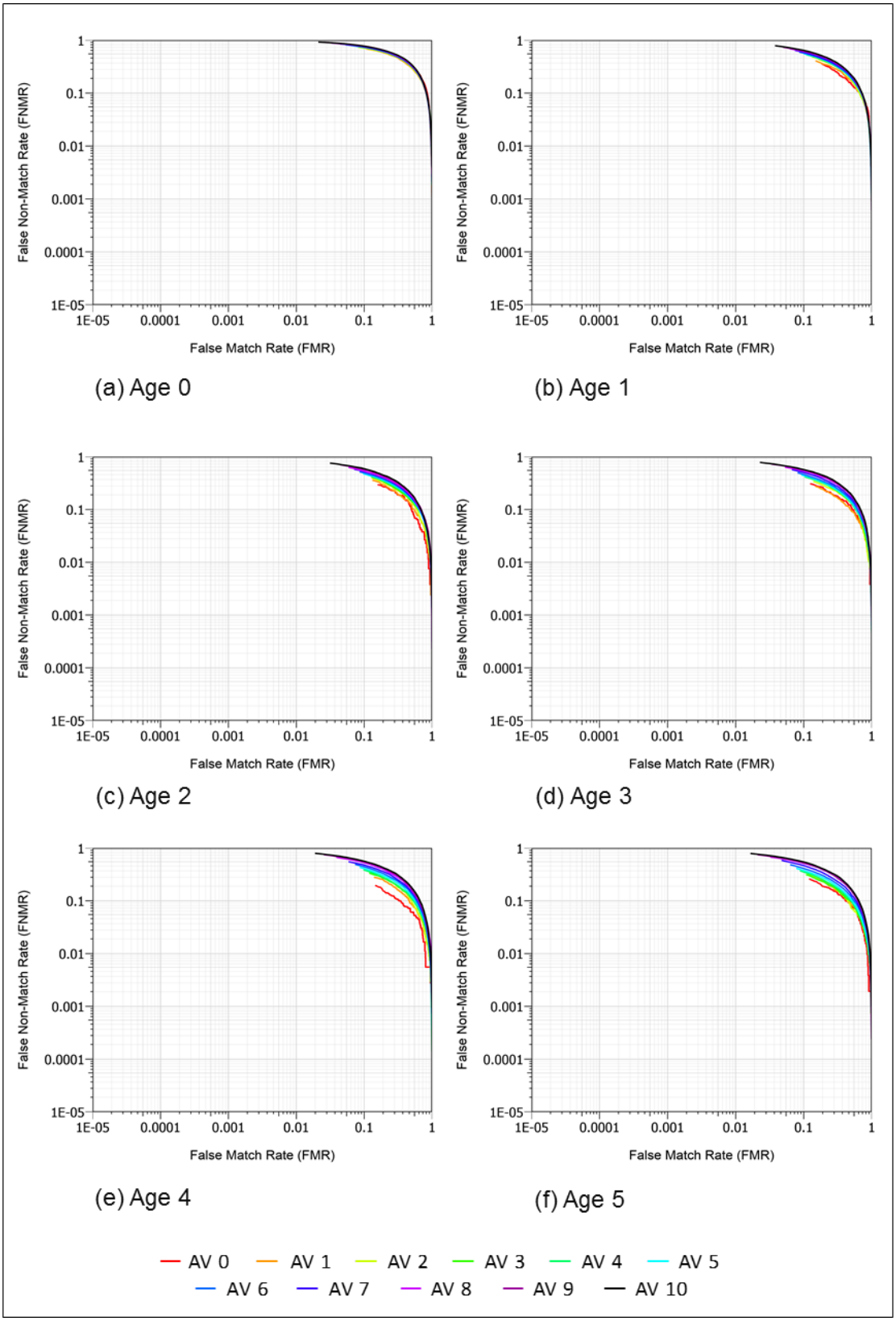


Figure I13. DETs for Algorithm F displaying how age variations impact on performance for ages 0–5 (AV = age variation in years).

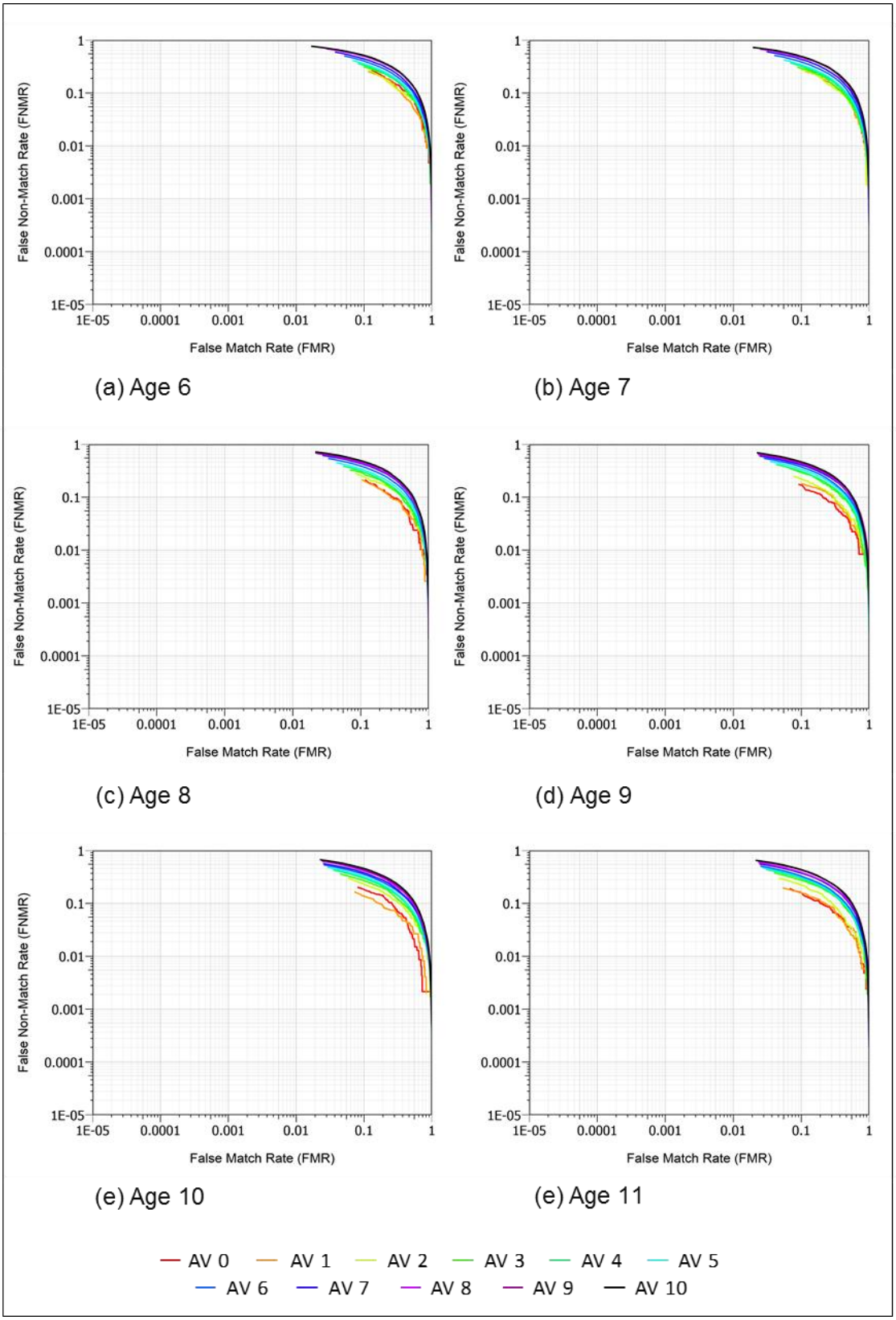


Figure I14. DETs for Algorithm F displaying how age variations impact on performance for ages 6–11 (AV = age variation in years).

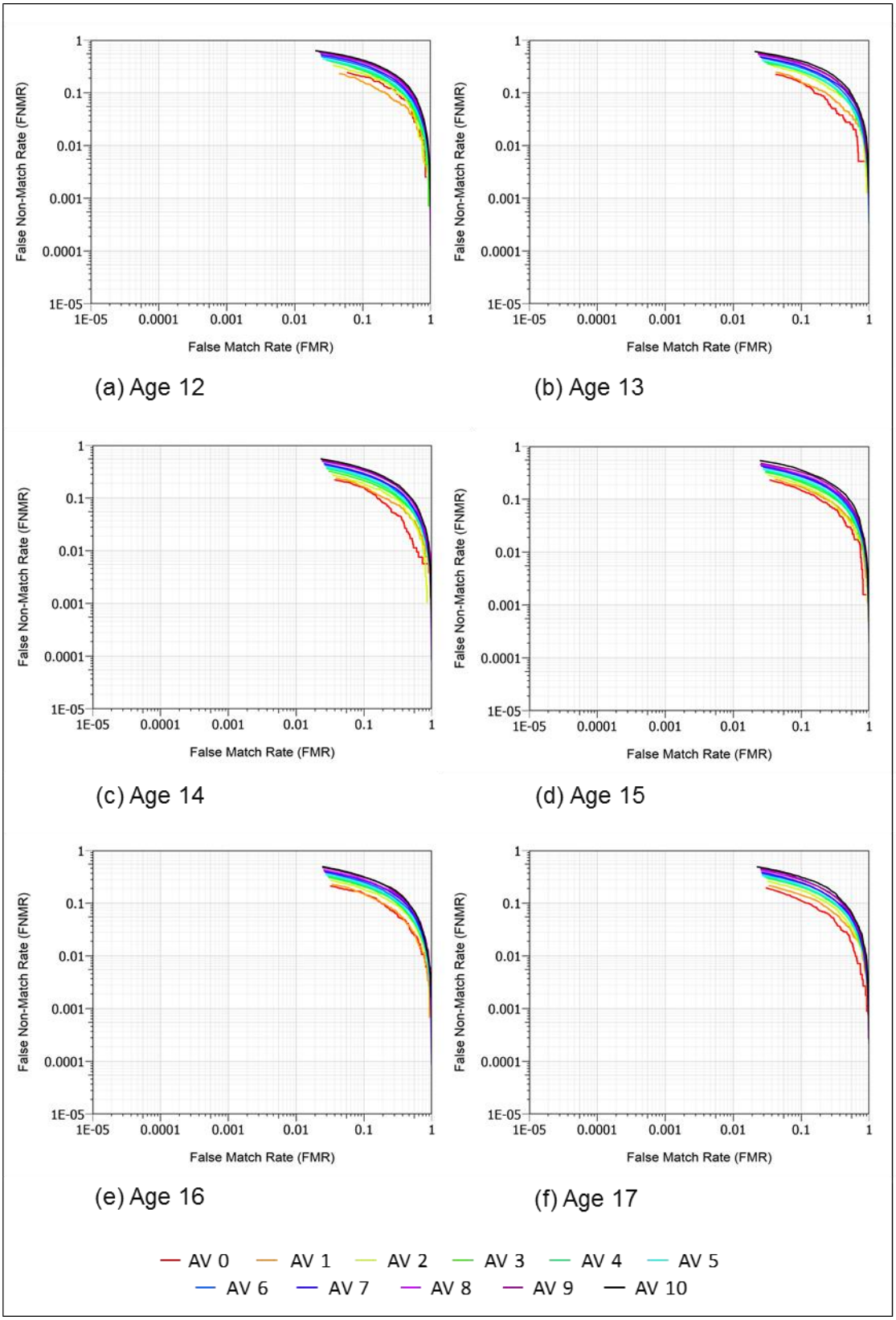


Figure I15. DETs for Algorithm F displaying how age variation impacts on performance for ages 12–17 (AV = age variation in years).

Appendix J. Cumulative Probability Plots for Algorithms A, B, C, D, and F displaying how Age Impacts on Performance for Age Variations 0–10 Years

See Section 3.5.2 for an explanation on how to interpret cumulative probability plots.

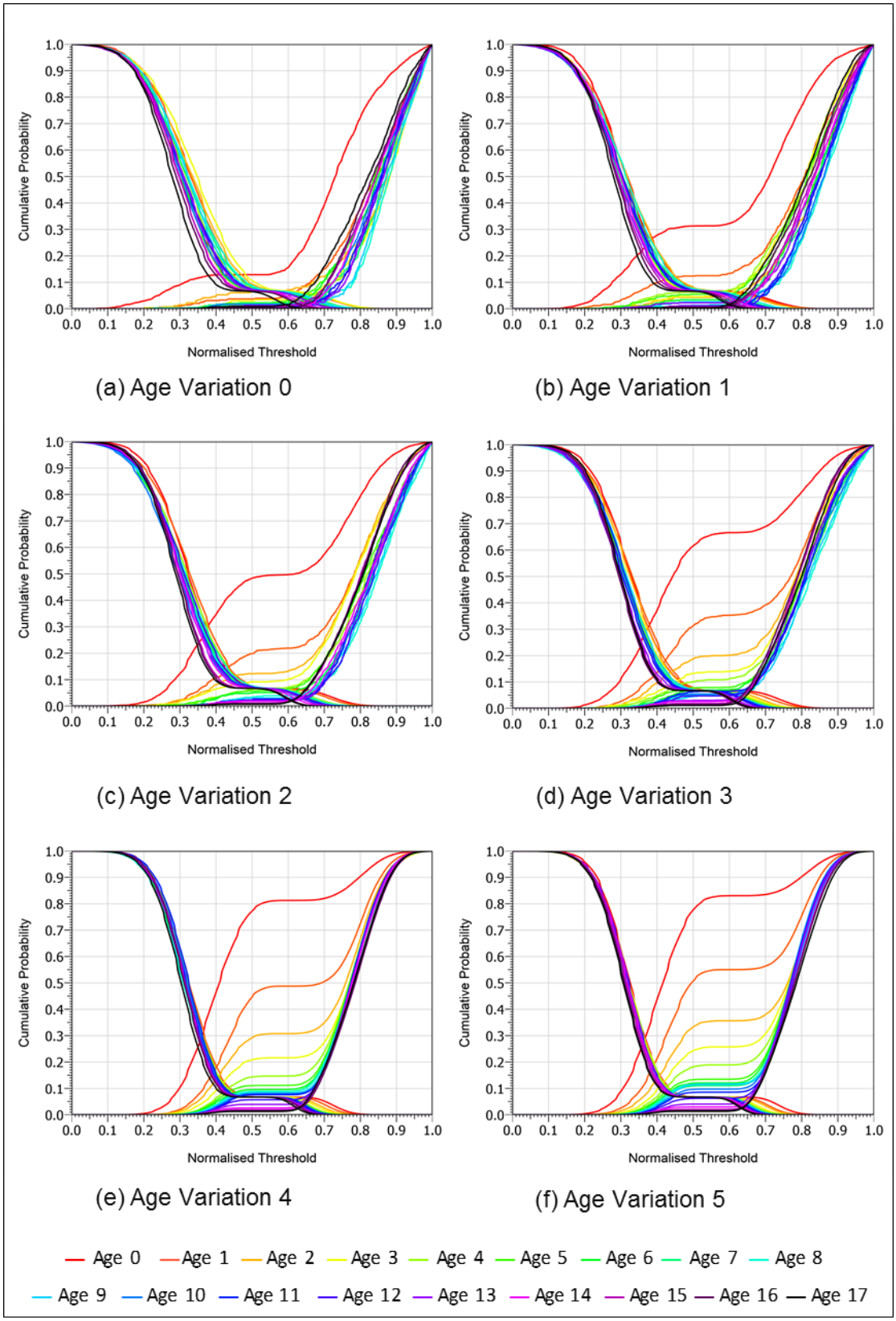


Figure J1. Cumulative probability plots for Algorithm A displaying how age impacts on performance for age variations spanning 0–5 years.

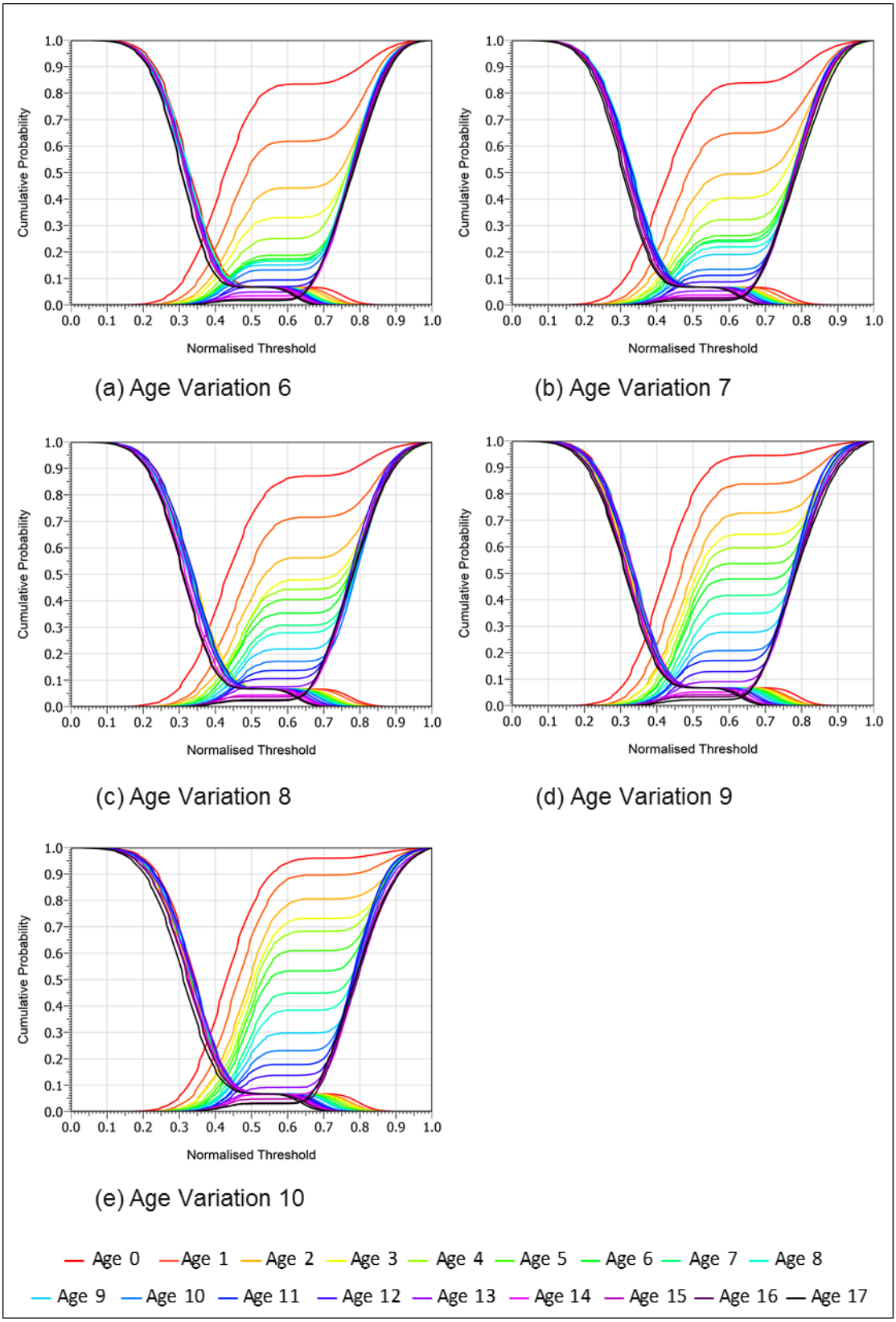


Figure J2. Cumulative probability plots for Algorithm A displaying how age impacts on performance for age variations spanning 6–10 years.

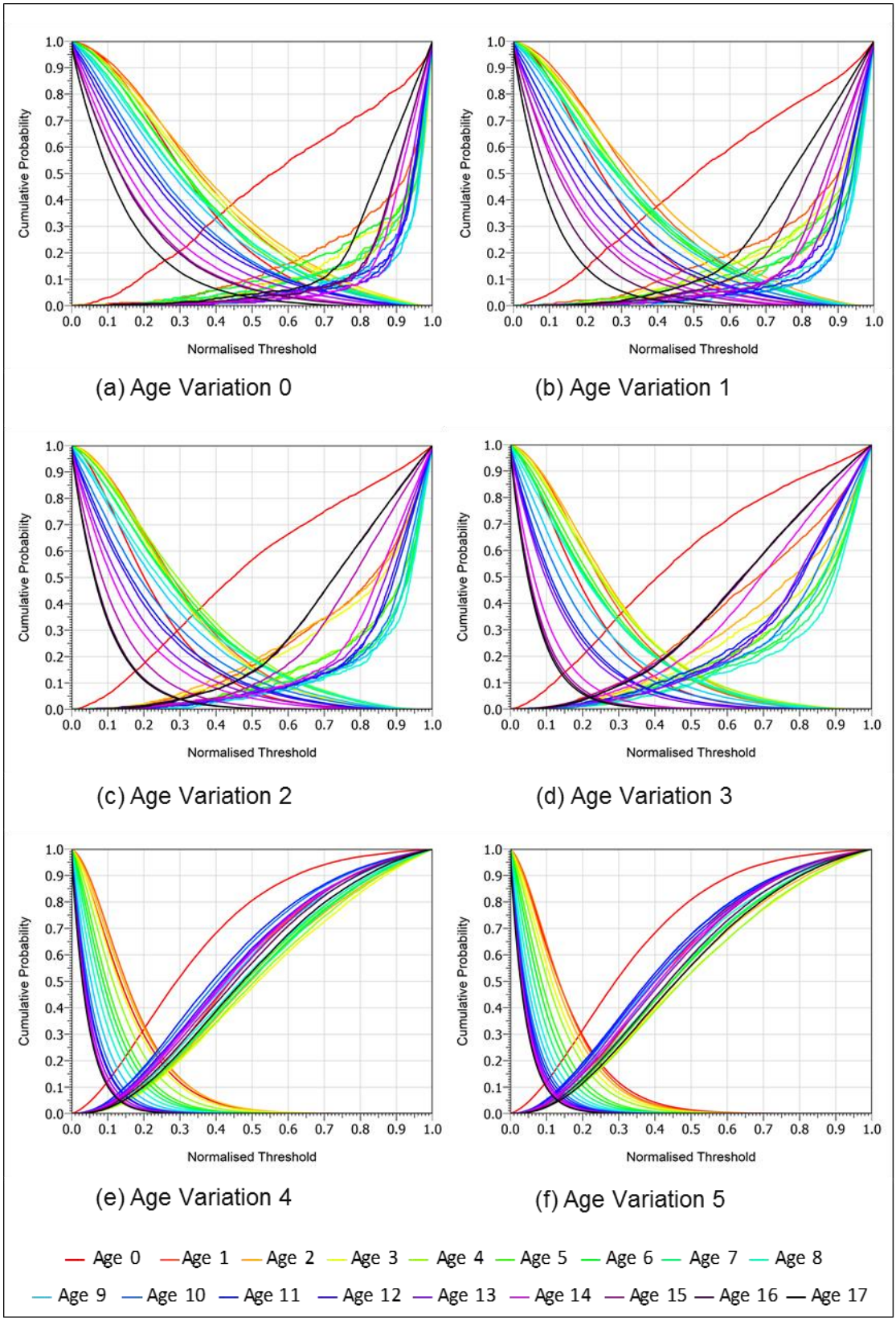


Figure J3. Cumulative probability plots for Algorithm B displaying how age impacts on performance for age variations spanning 0–5 years.

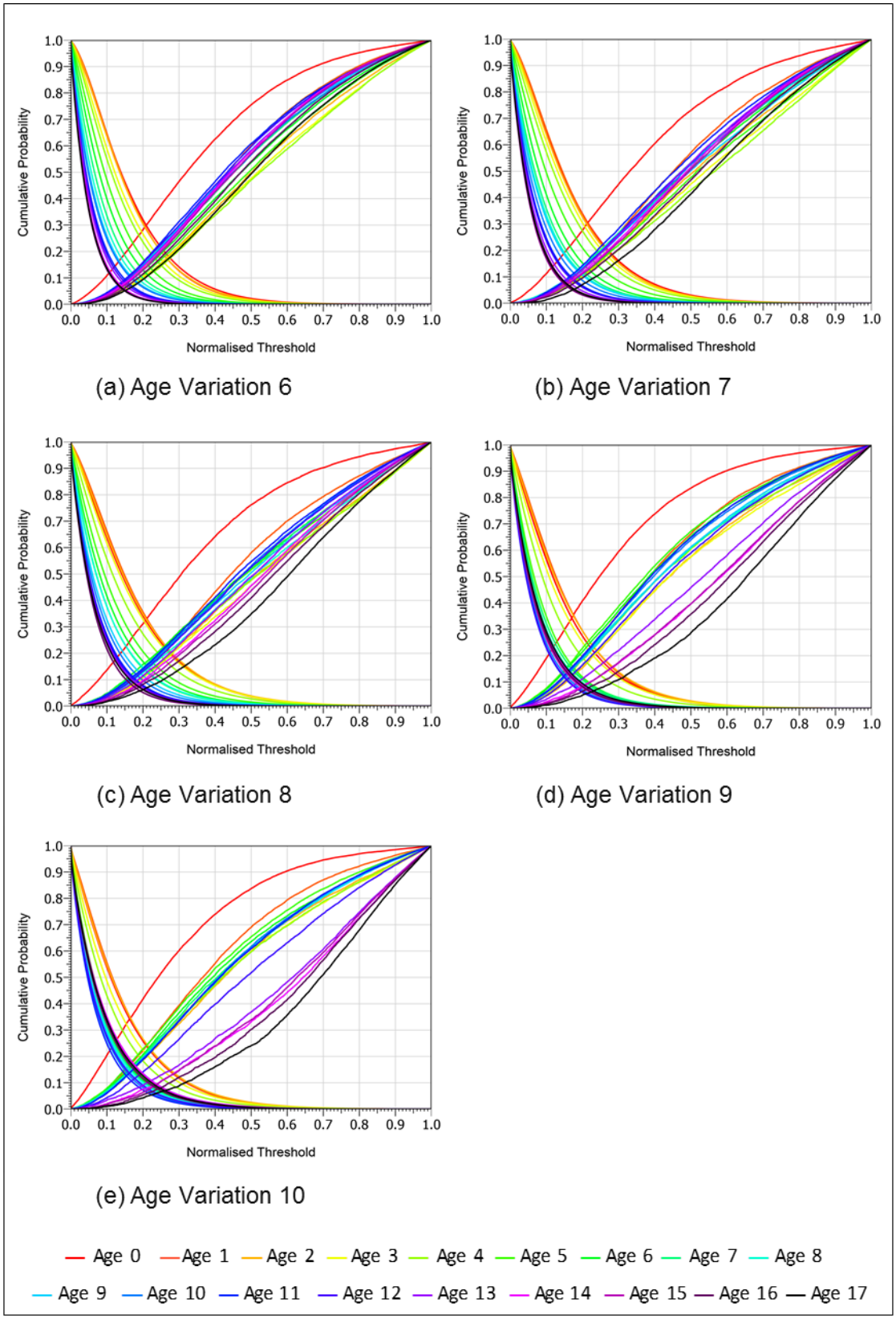


Figure J4. Cumulative probability plots for Algorithm B displaying how age impacts on performance for age variations spanning 6–10 years.

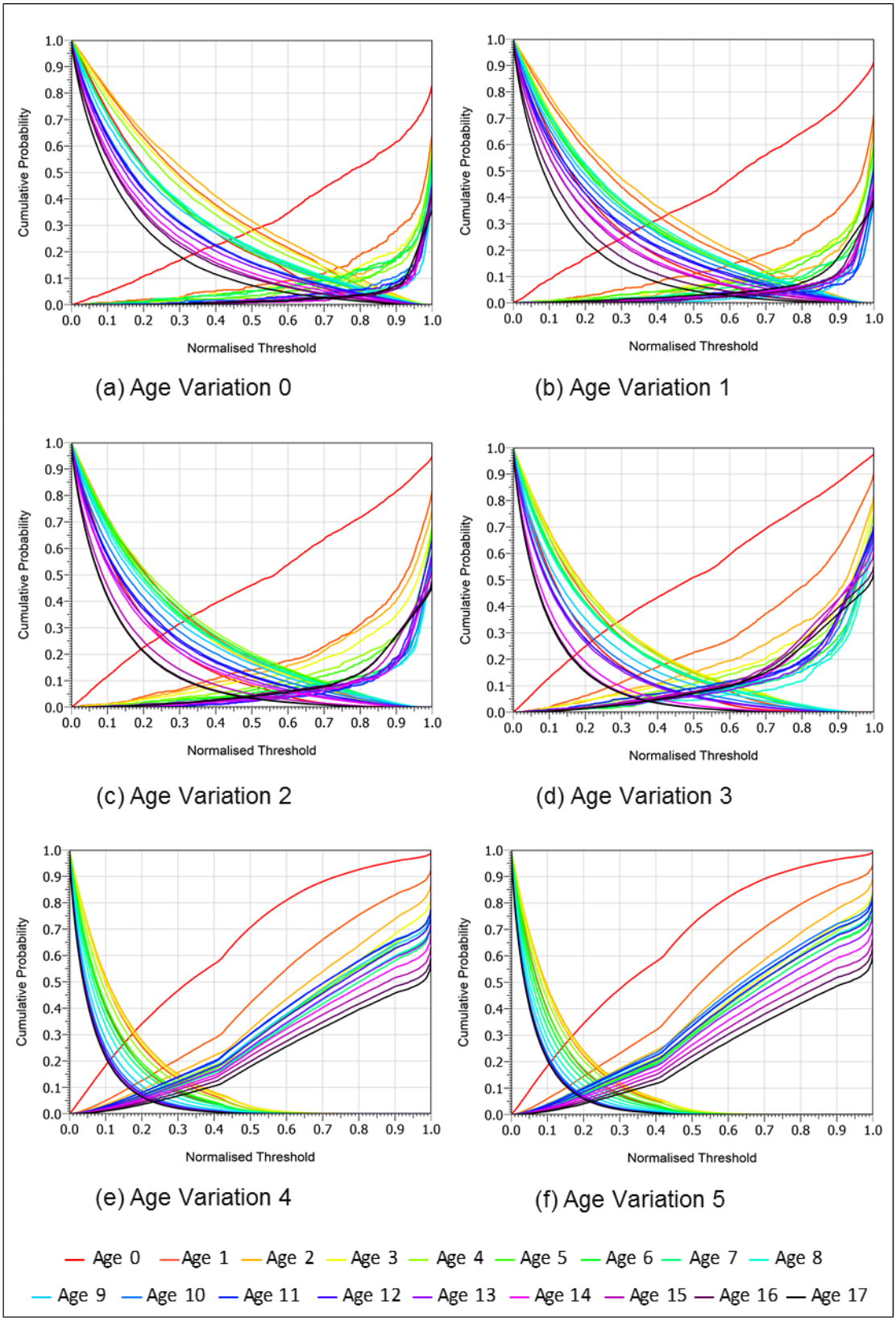


Figure J5. Cumulative probability plots for Algorithm C displaying how age impacts on performance for age variations spanning 0–5 years.

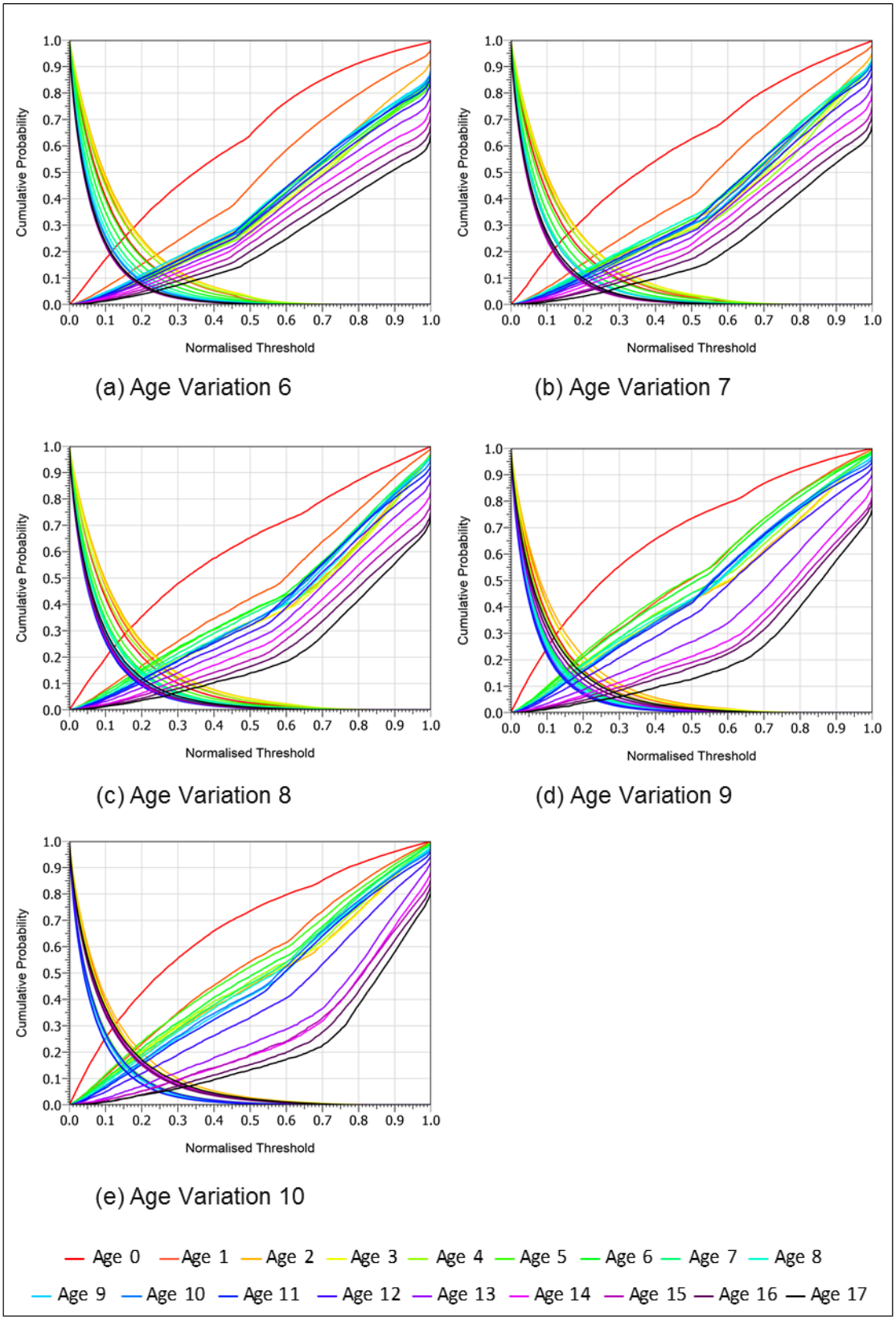


Figure J6. Cumulative probability plots for Algorithm C displaying how age impacts on performance for age variations spanning 6–10 years.

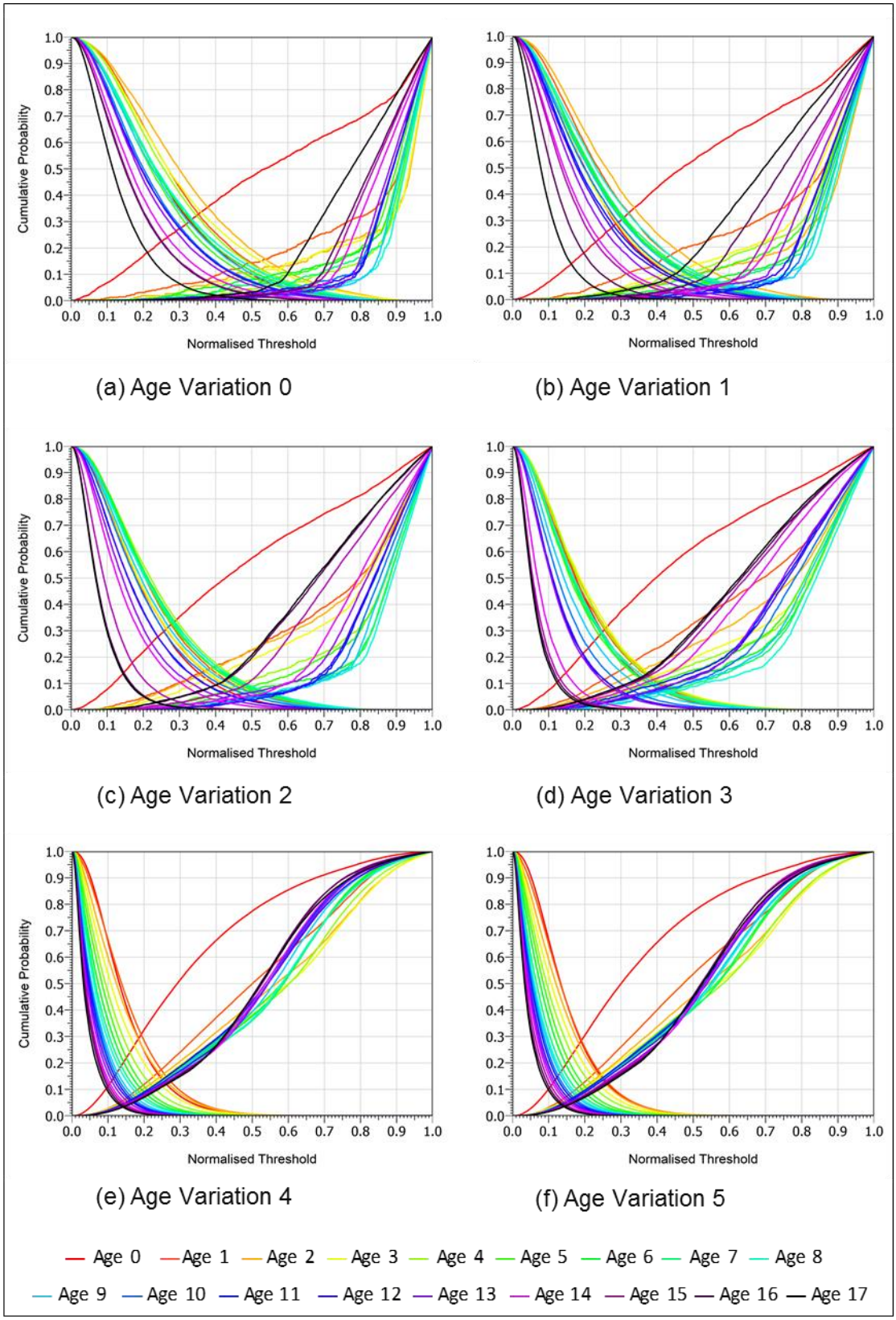


Figure J7. Cumulative probability plots for Algorithm D displaying how age impacts on performance for age variations spanning 0–5 years.

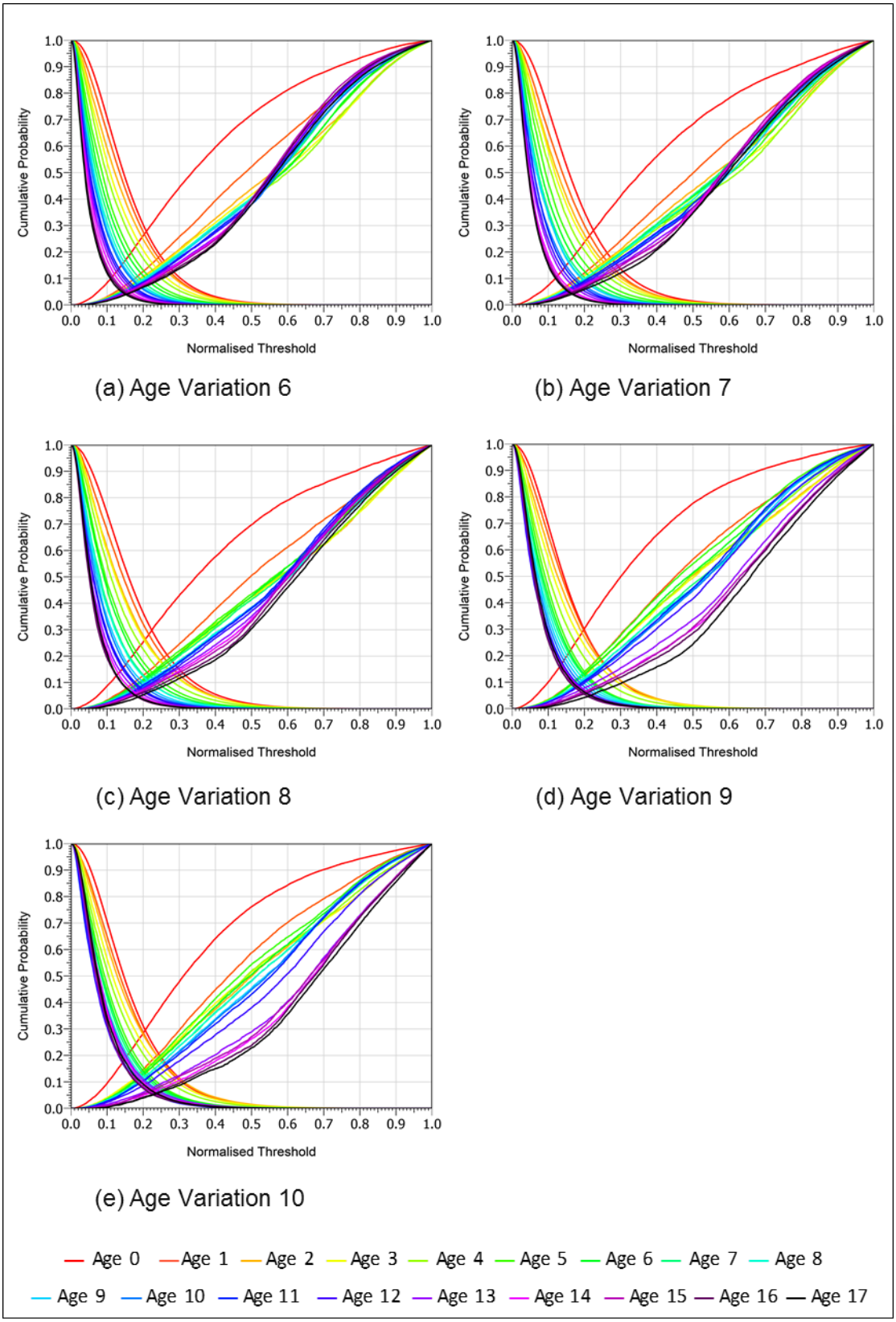


Figure J8. Cumulative probability plots for Algorithm D displaying how age impacts on performance for age variations spanning 6–10 years.

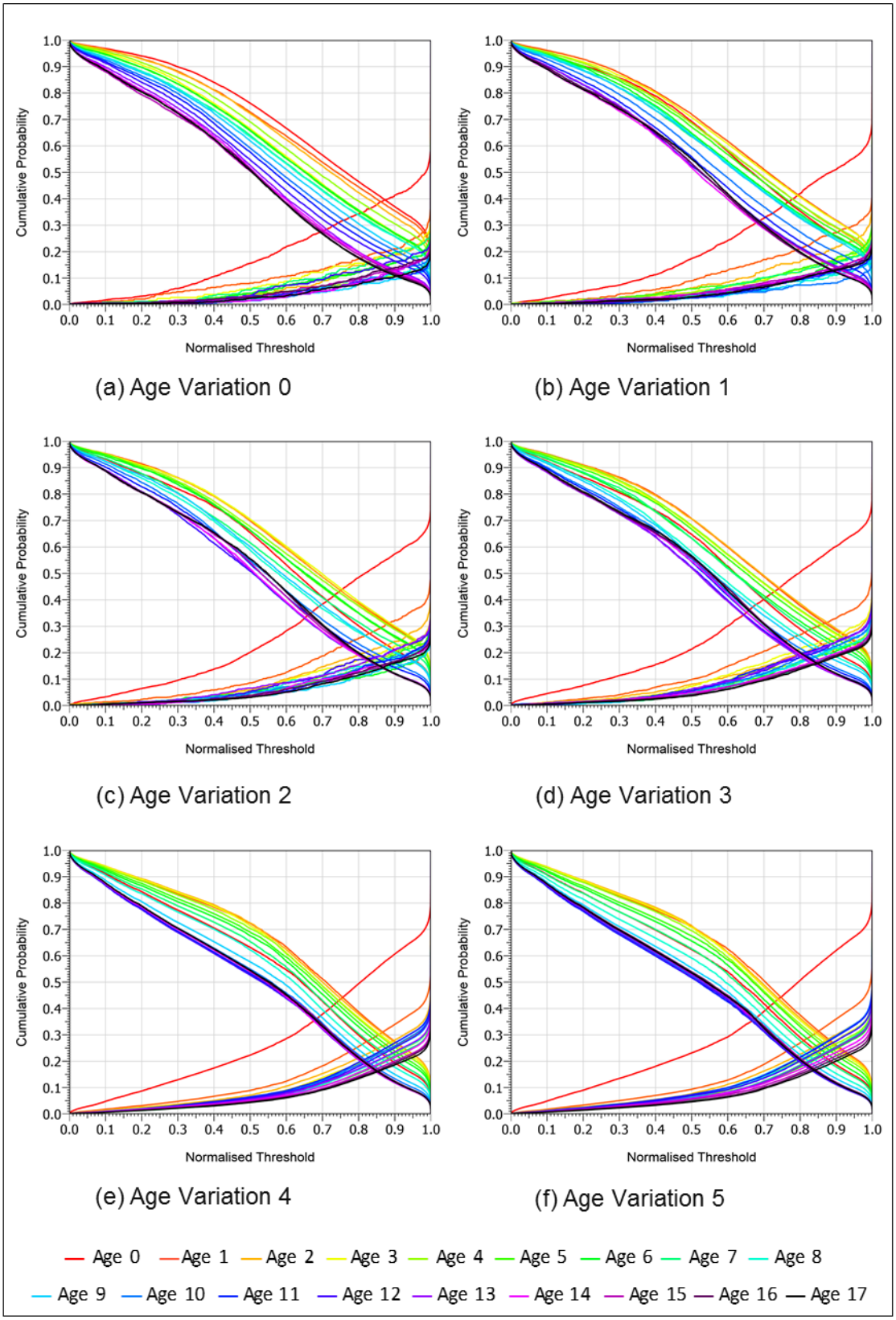


Figure J9. Cumulative probability plots for Algorithm F displaying how age impacts on performance for age variations spanning 0–5 years.

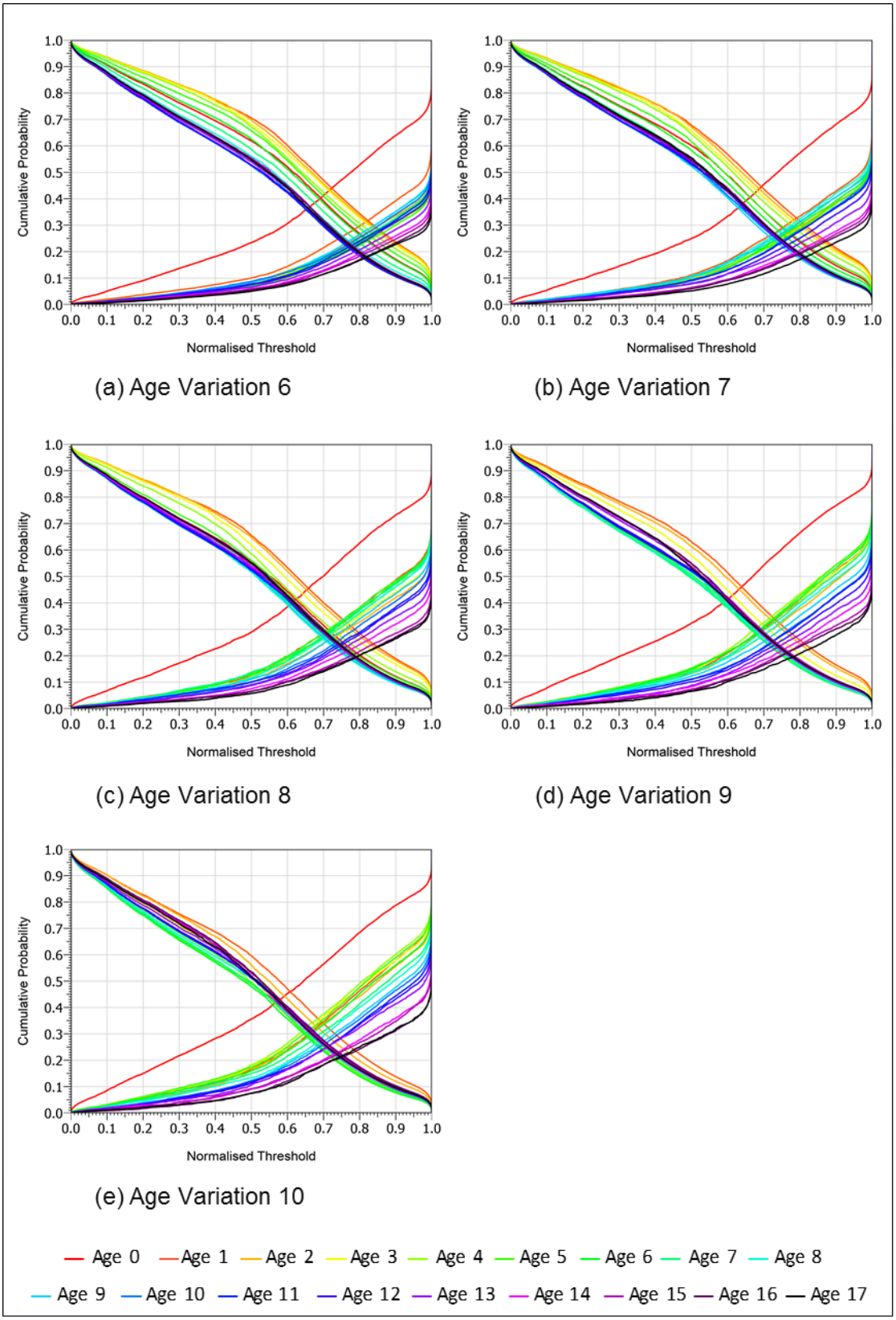


Figure J10. Cumulative probability plots for Algorithm F displaying how age impacts on performance for age variations spanning 6–10 years.

Appendix K. Cumulative Probability Plots for Algorithms A, B, C, D, and F displaying how Age Variation Impacts on Performance for Ages 0–10 Years

See Section 3.5.2 for an explanation on how to interpret cumulative probability plots.

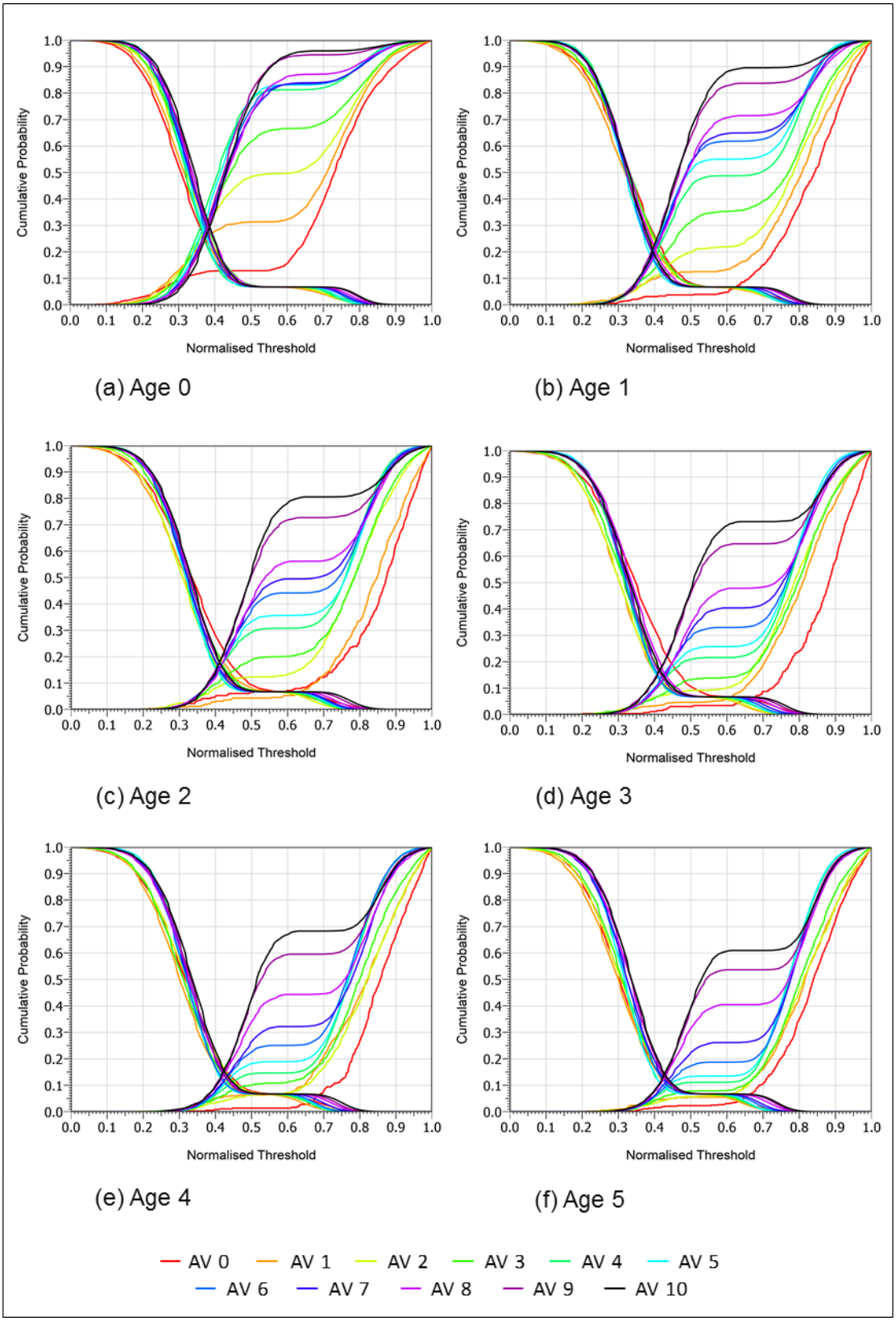


Figure K1. Cumulative probability plots for Algorithm A displaying how age variation impacts on performance for ages 0–5 (AV = age variation in years).

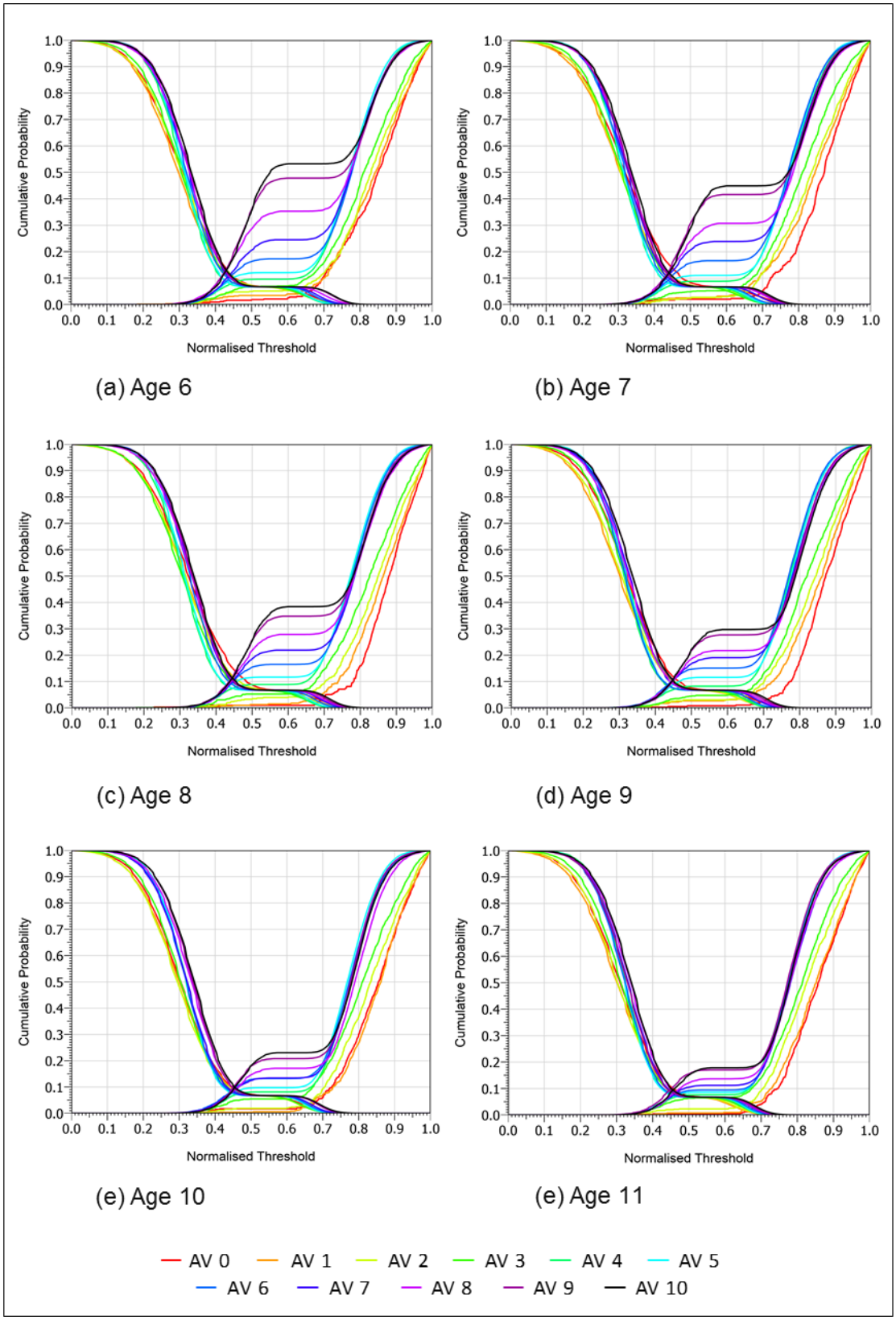


Figure K2. Cumulative probability plots for Algorithm A displaying how age variation impacts on performance for ages 6–11 (AV = age variation in years).

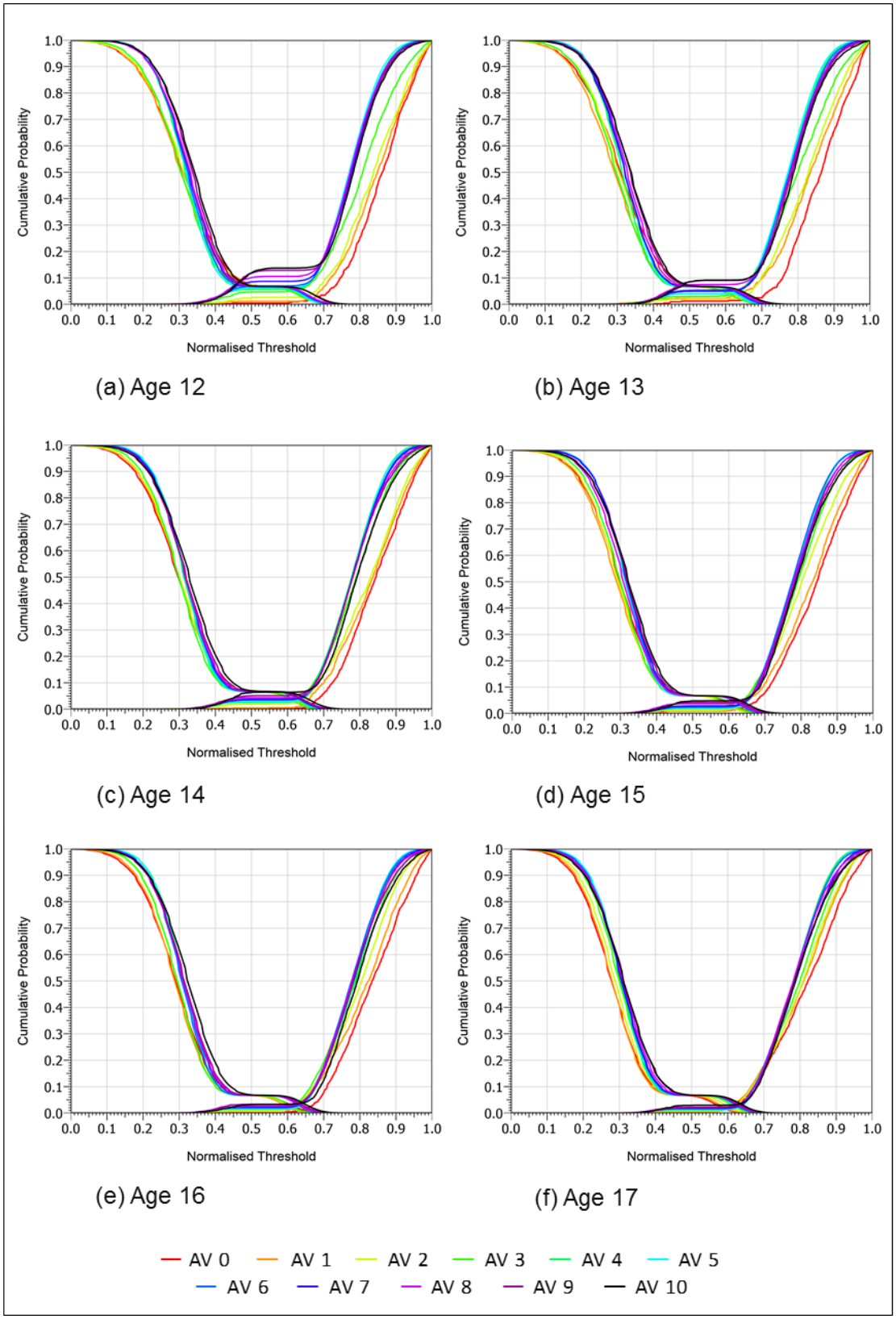


Figure K3. Cumulative probability plots for Algorithm A displaying how age variation impacts on performance for ages 12–17 (AV = age variation in years).

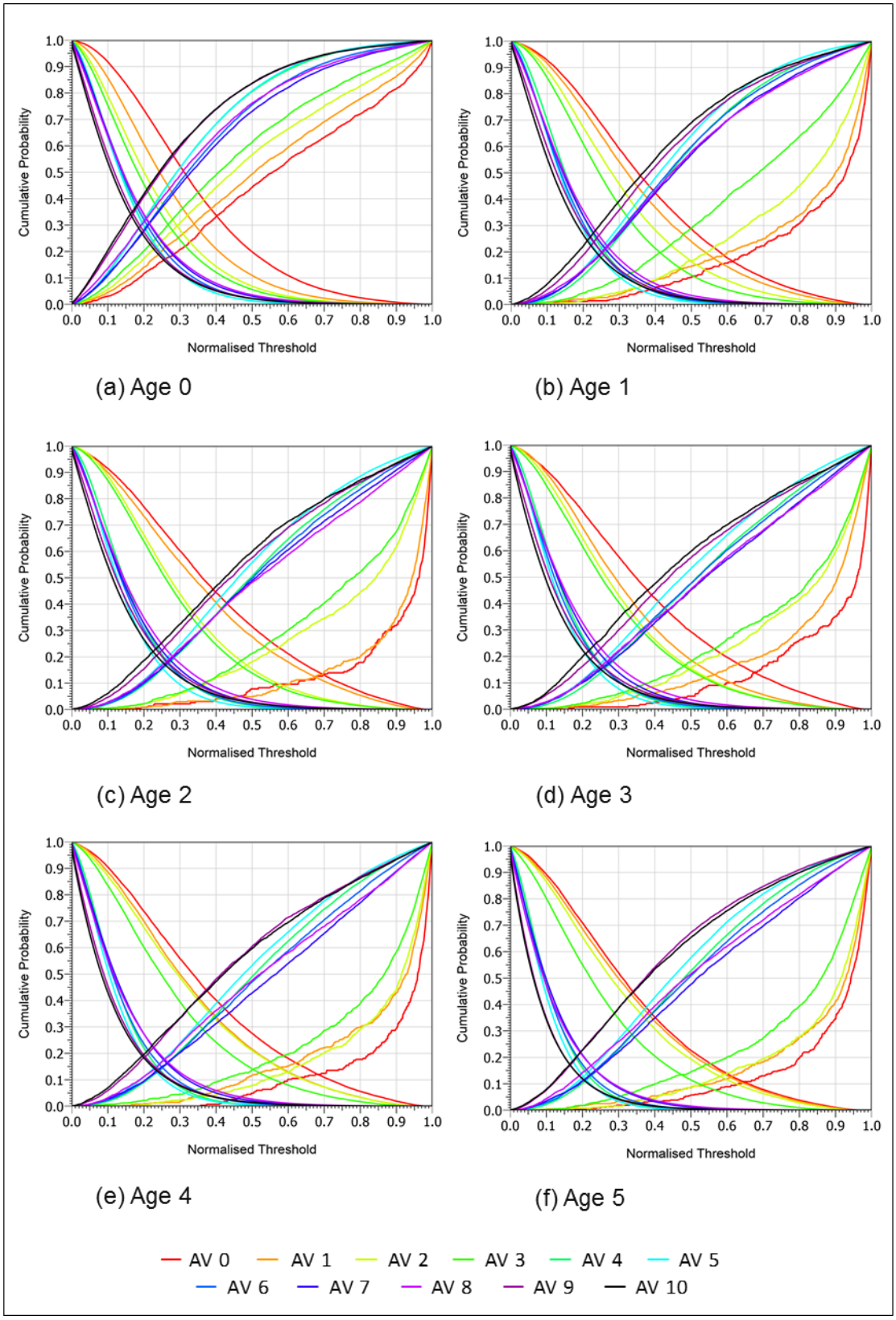


Figure K4. Cumulative probability plots for Algorithm B displaying how age variation impacts on performance for ages 0–5 (AV = age variation in years).

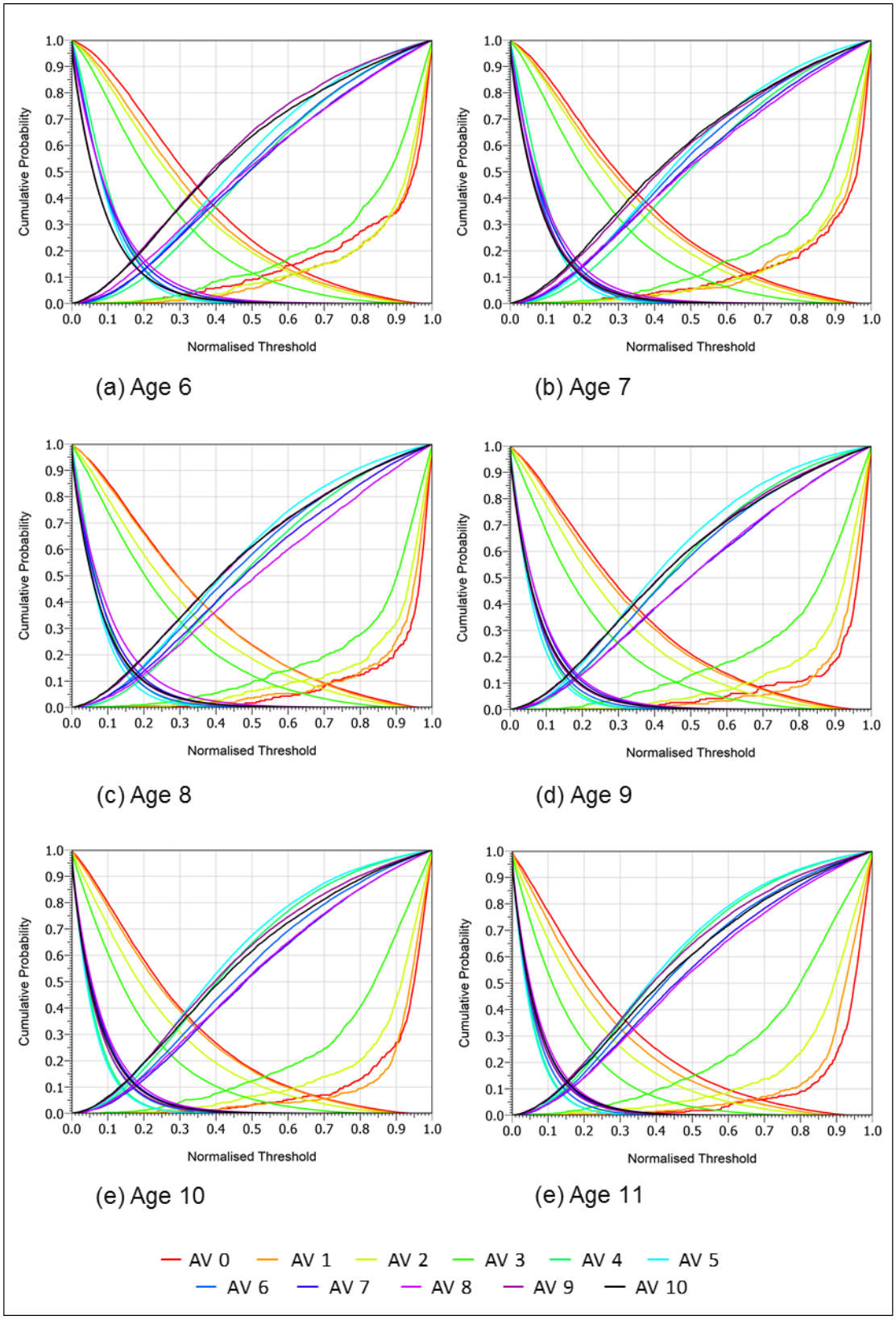


Figure K5. Cumulative probability plots for Algorithm B displaying how age variation impacts on performance for ages 6–11 (AV = age variation in years).

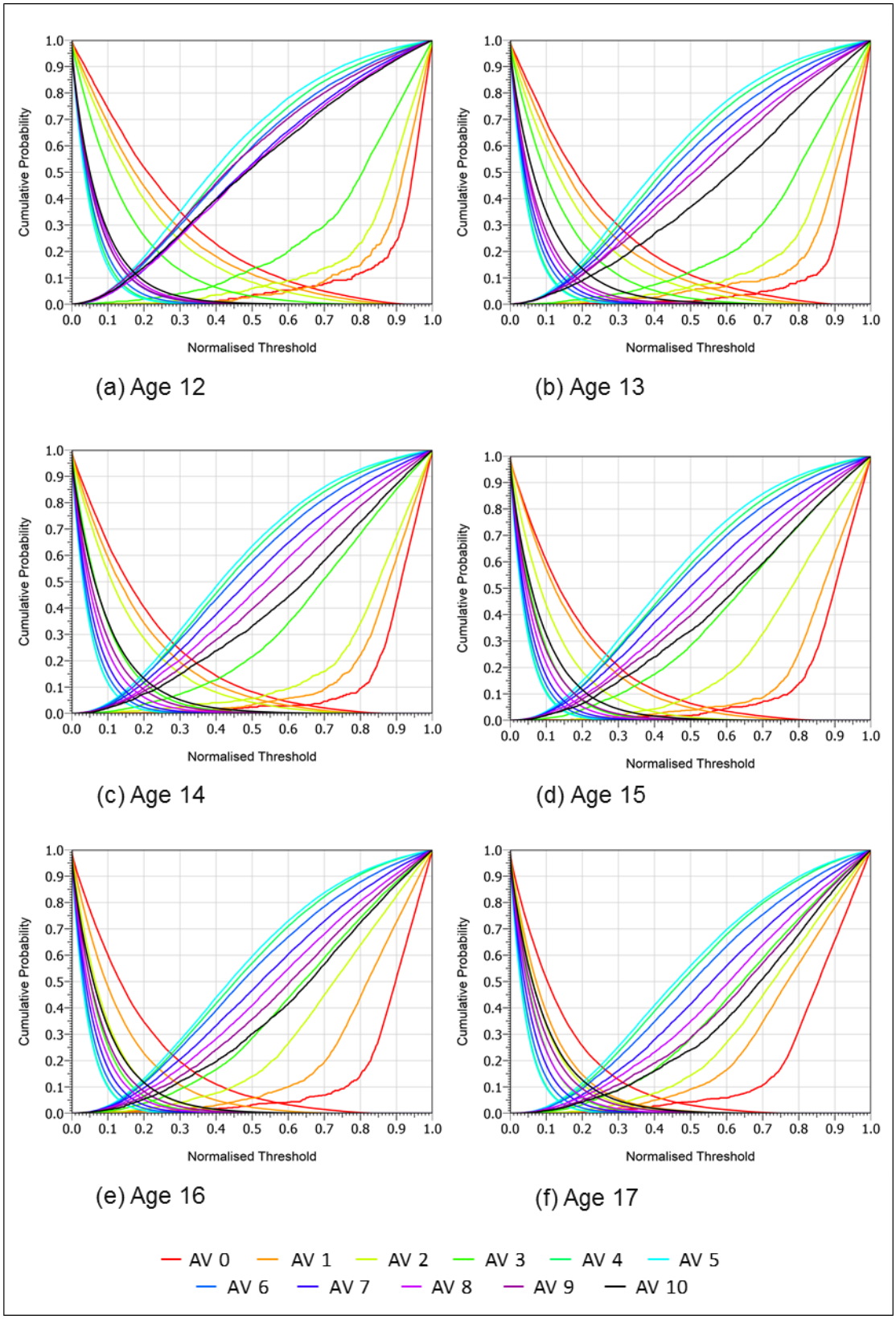


Figure K6. Cumulative probability plots for Algorithm B displaying how age variation impacts on performance for ages 12–17 (AV = age variation in years).

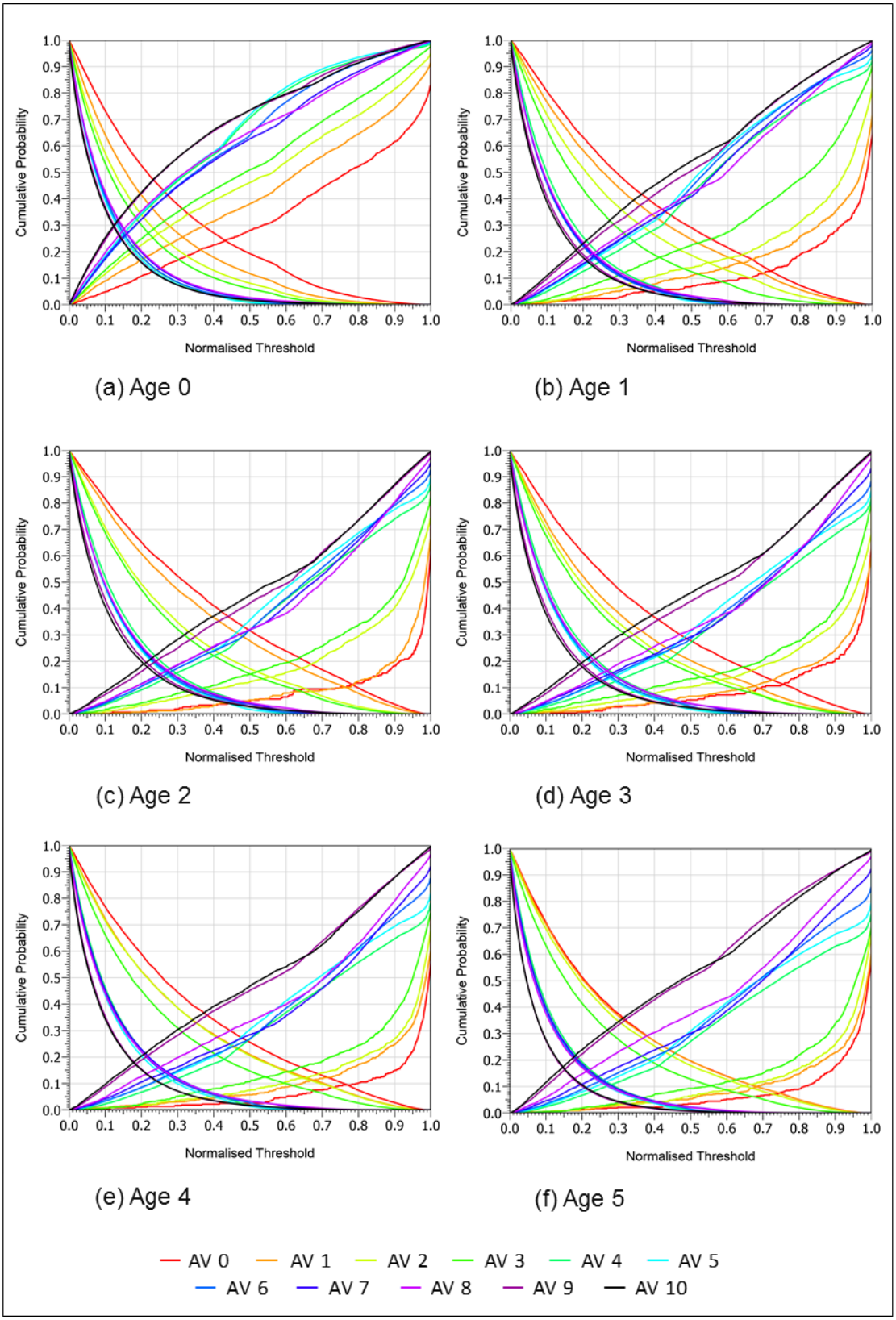


Figure K7. Cumulative probability plots for Algorithm C displaying how age variation impacts on performance for ages 0–5 (AV = age variation in years).

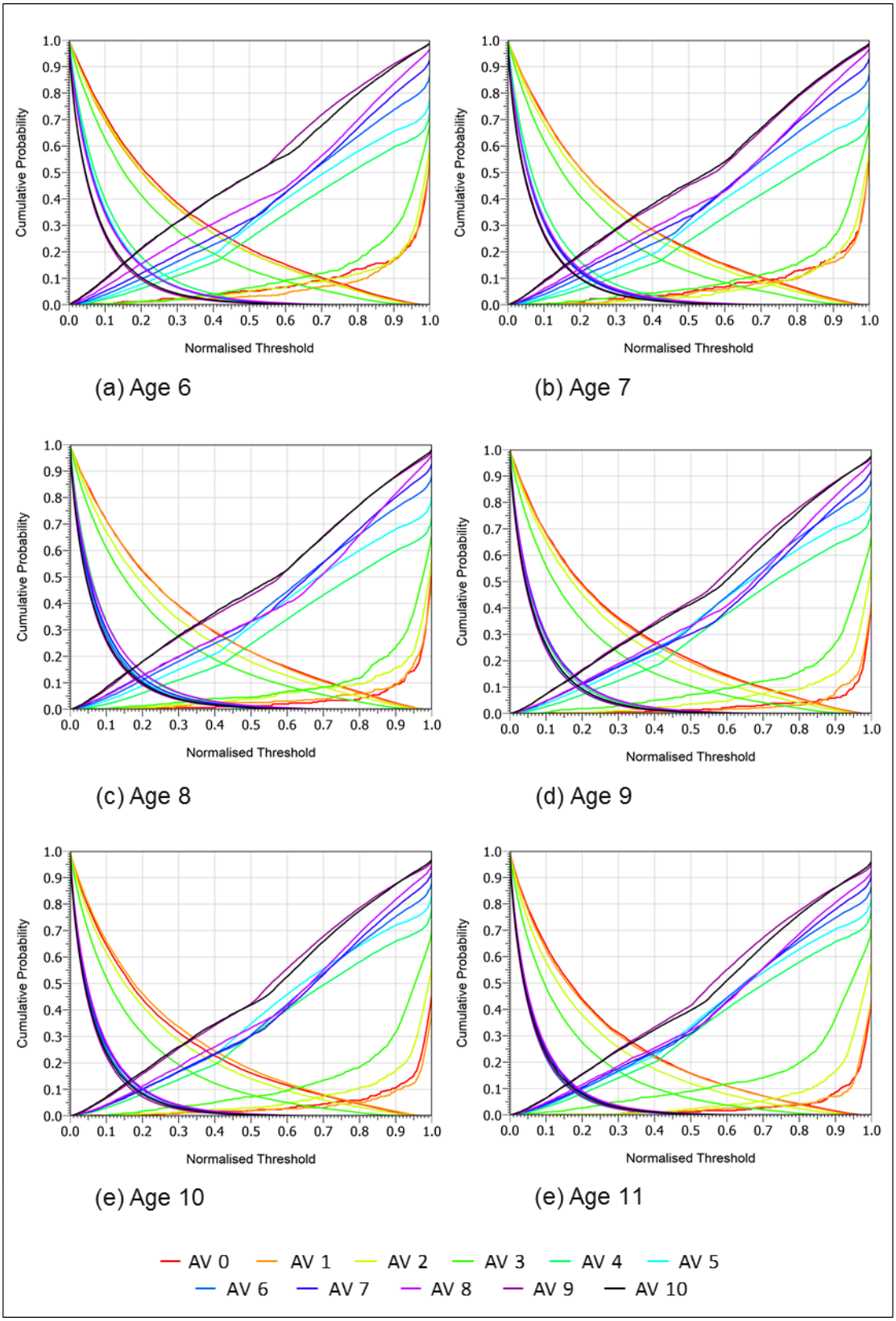


Figure K8. Cumulative probability plots for Algorithm C displaying how age variation impacts on performance for ages 6–11 (AV = age variation in years).

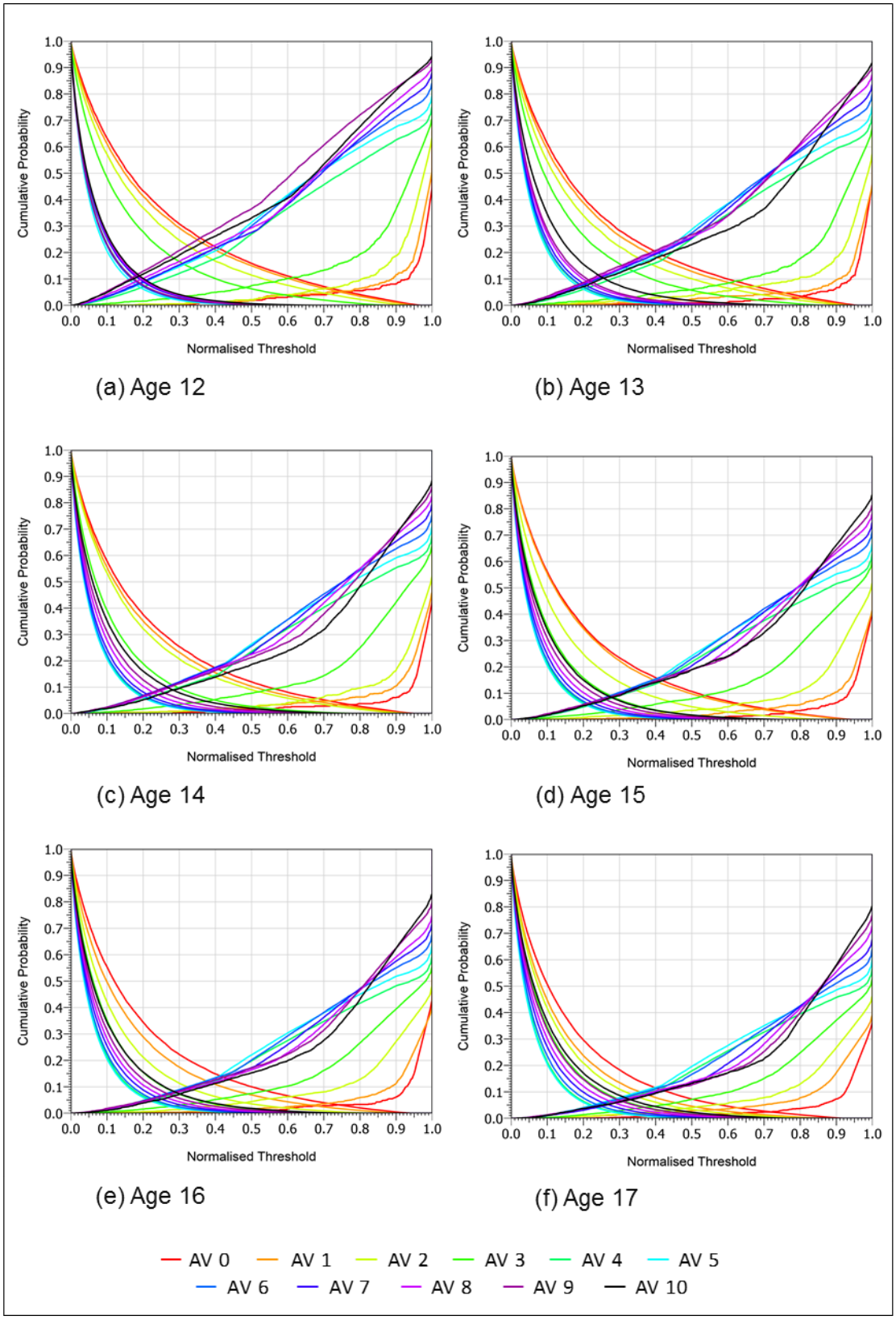


Figure K9. Cumulative probability plots for Algorithm C displaying how age variation impacts on performance for ages 12–17 (AV = age variation in years).

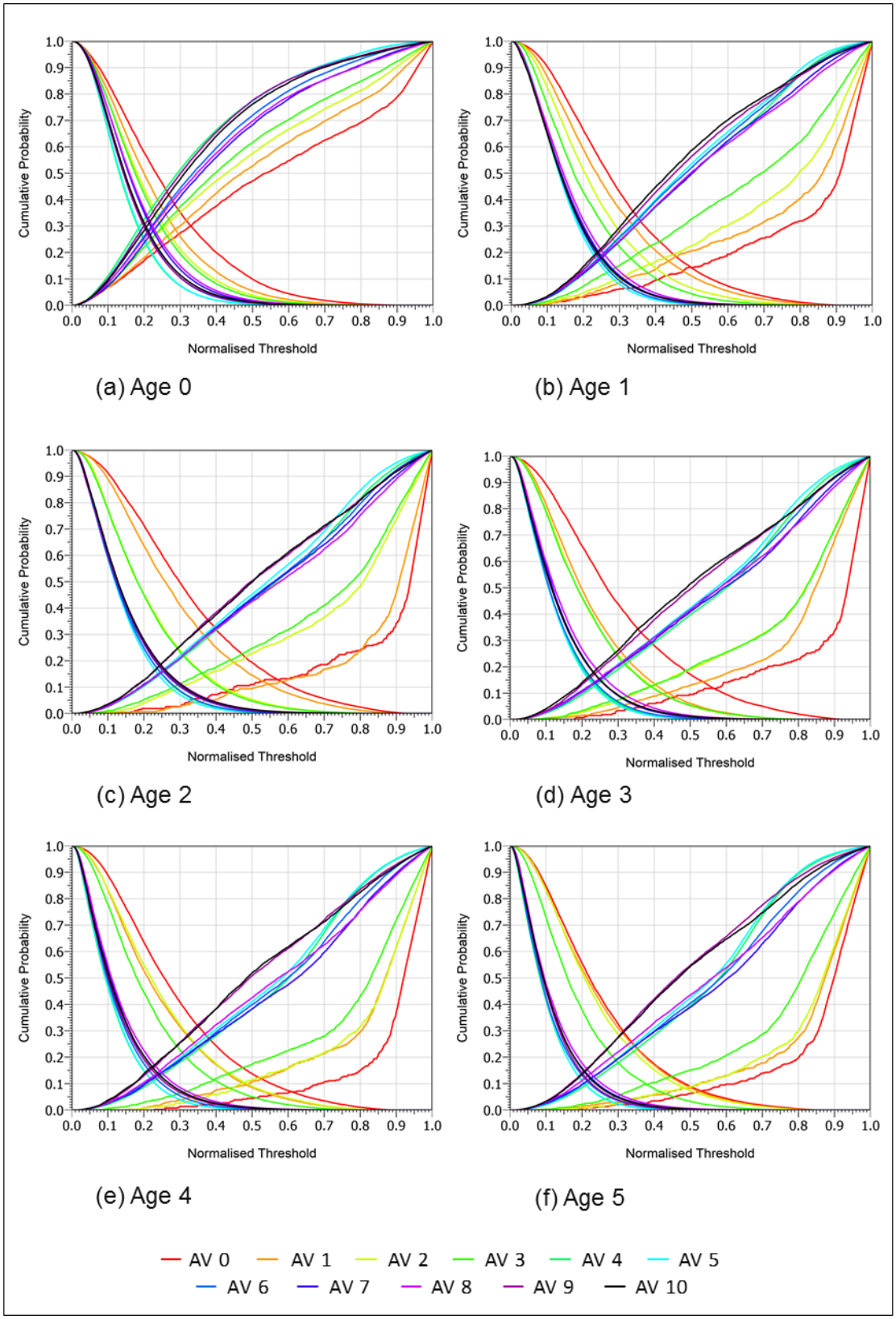


Figure K10. Cumulative probability plots for Algorithm D displaying how age variation impacts on performance for ages 0–5 (AV = age variation in years).

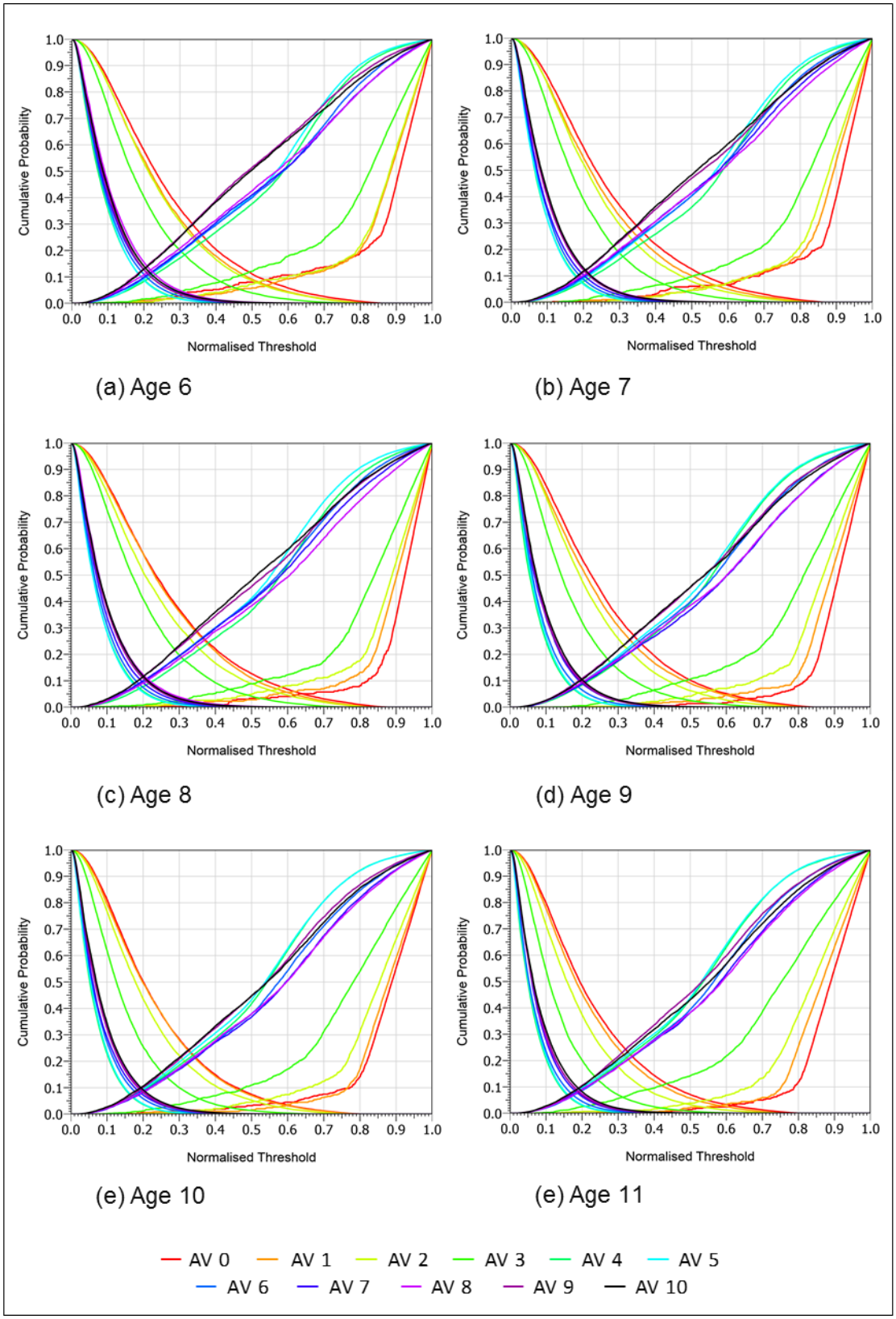


Figure K11. Cumulative probability plots for Algorithm D displaying how age variation impacts on performance for ages 6–11 (AV = age variation in years).

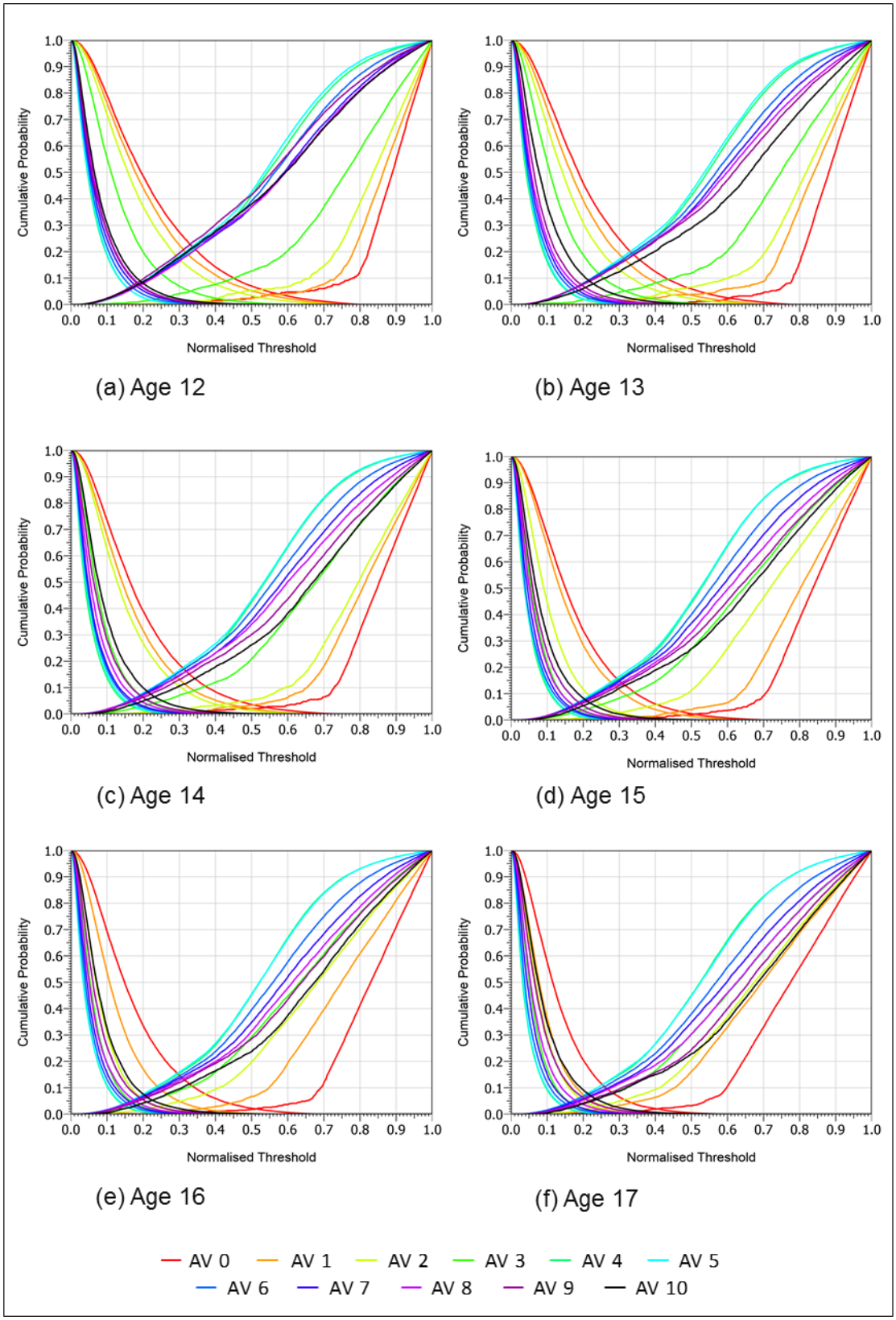


Figure K12. Cumulative probability plots for Algorithm D displaying how age variation impacts on performance for ages 12–17 (AV = age variation in years).

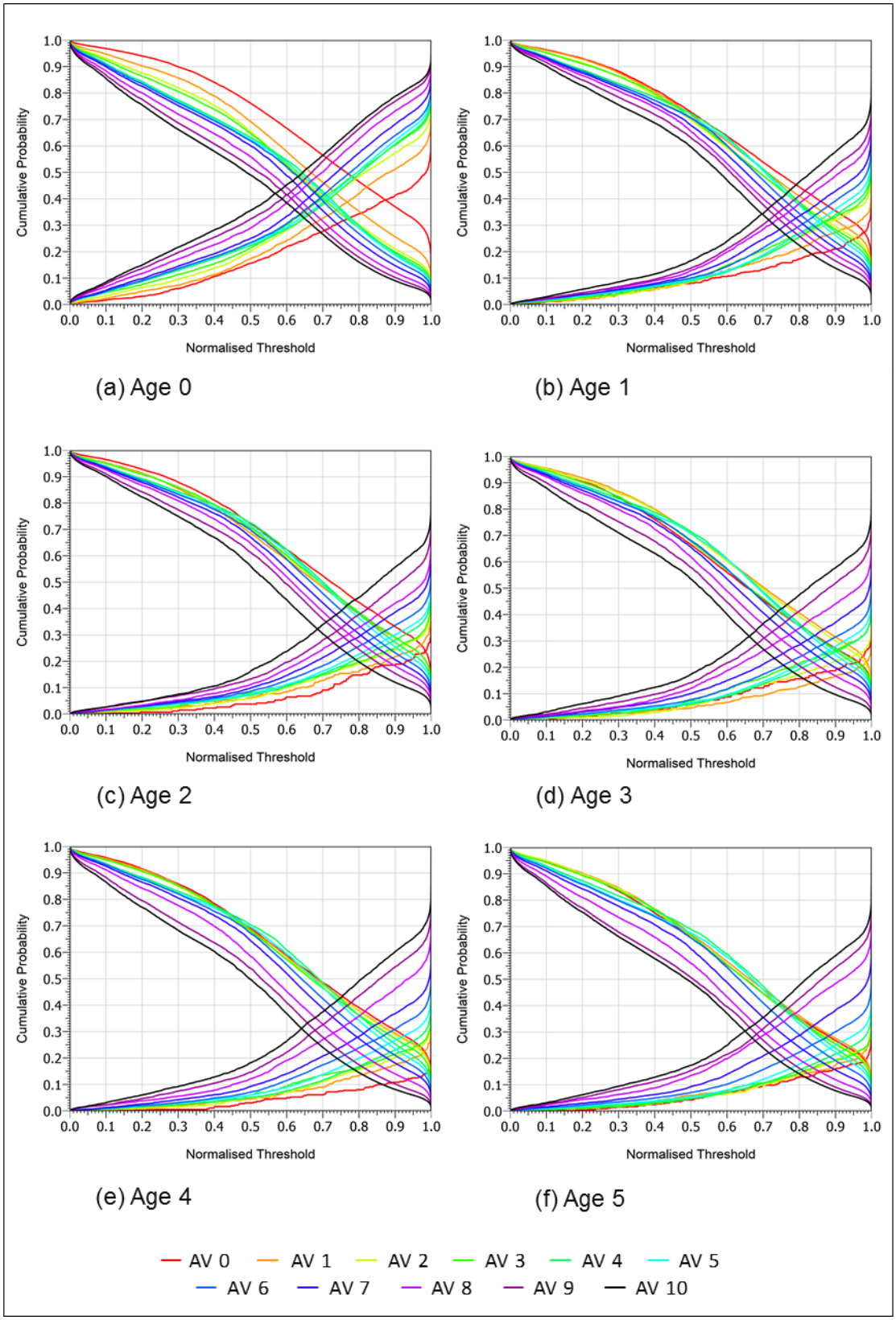


Figure K13. Cumulative probability plots for Algorithm F displaying how age variation impacts on performance for ages 0–5 (AV = age variation in years).

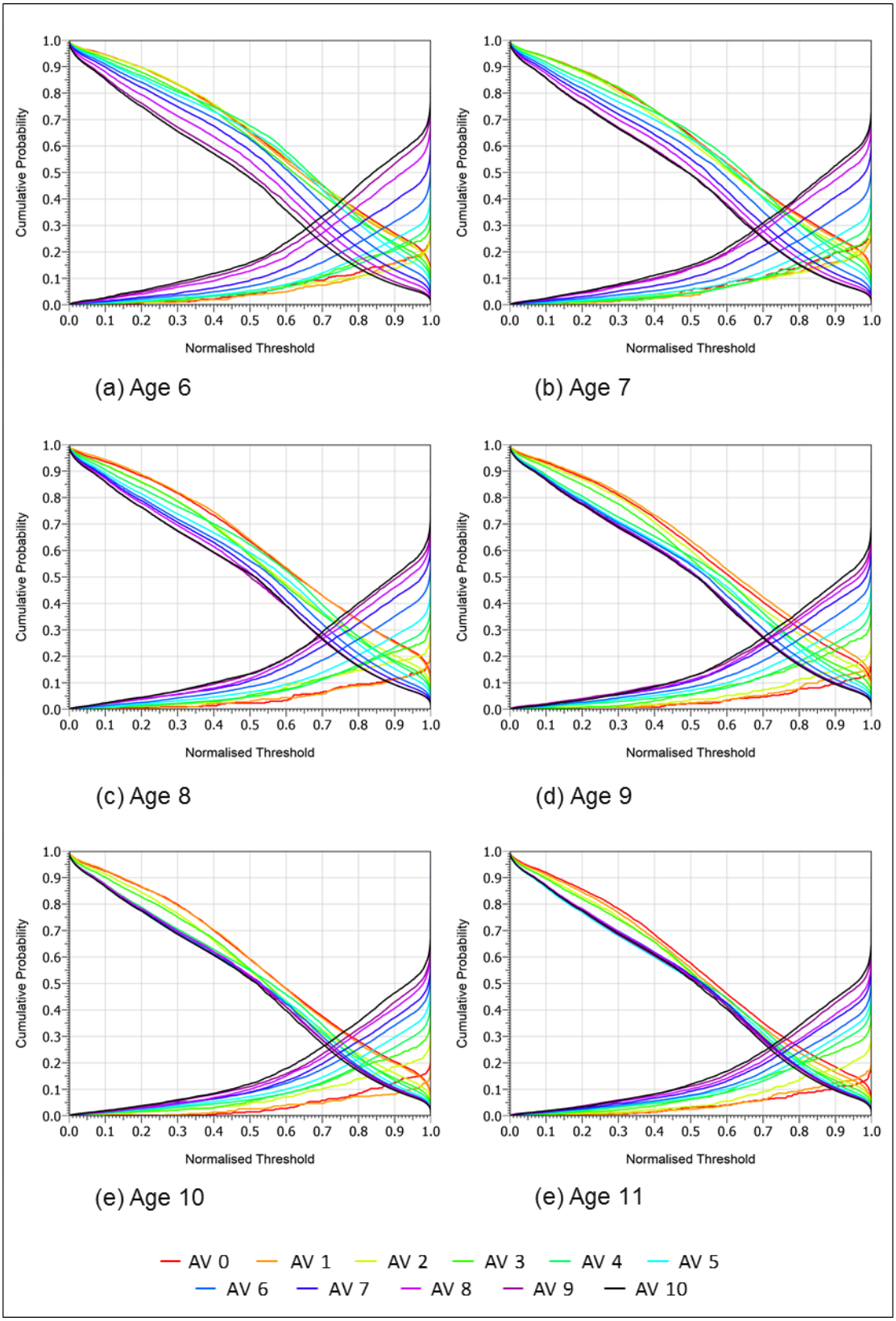


Figure K14. Cumulative probability plots for Algorithm F displaying how age variation impacts on performance for ages 6–11 (AV = age variation in years).

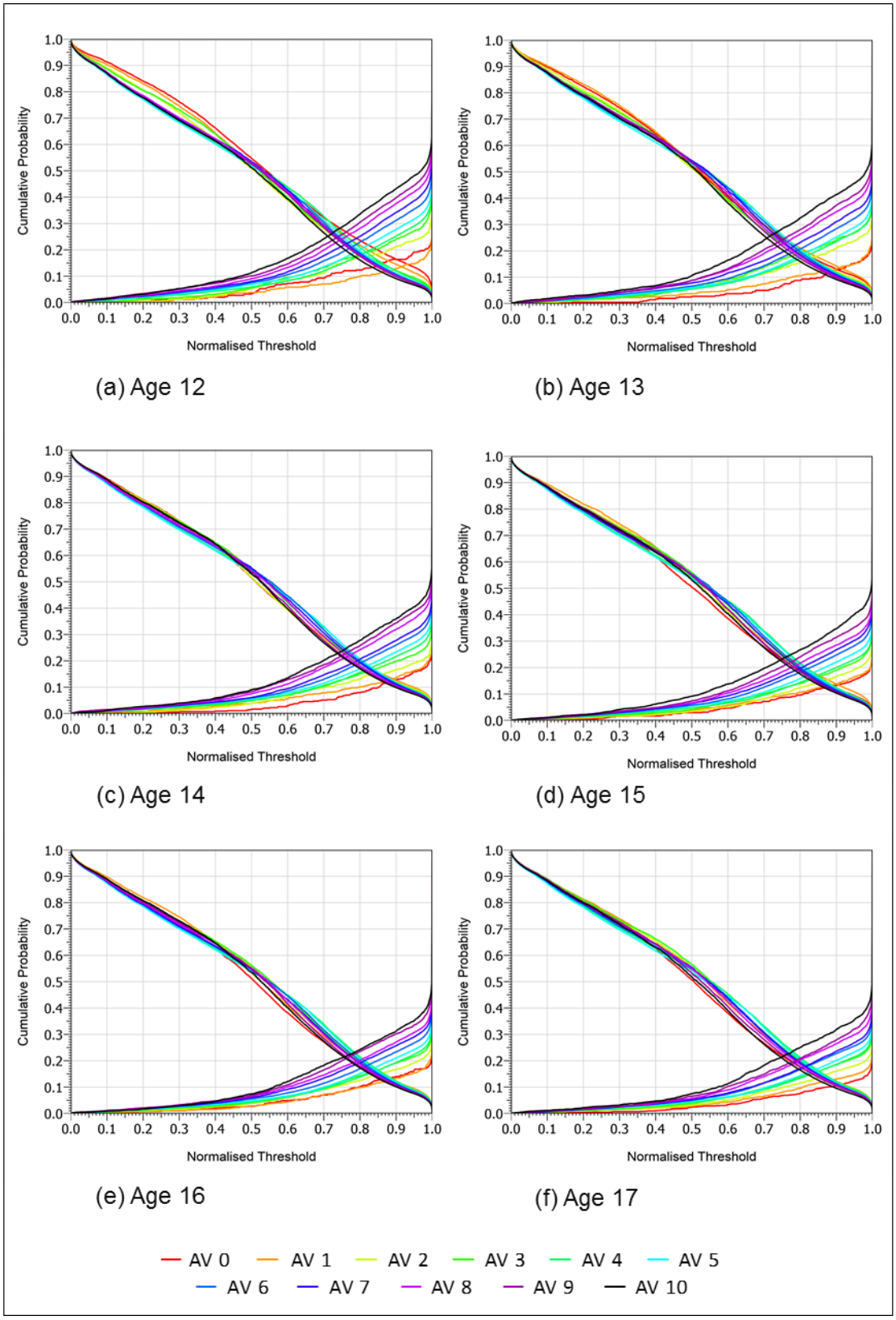


Figure K15. Cumulative probability plots for Algorithm F displaying how age variation impacts on performance for ages 12–17 (AV = age variation in years).

Appendix L. False Match Rate and False Non-Match Rate Data for Algorithms A, B, C, and D based on a Threshold Set at a False Match Rate of 0.001 with Images of Adults.

FMR

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	0.0511	0.0306	0.0174	0.0092	0.0040	0.0026	0.0012	0.0005	0.0002	0.0001	0.0000
1	0.0451	0.0369	0.0247	0.0158	0.0079	0.0049	0.0023	0.0013	0.0005	0.0002	0.0001
2	0.0349	0.0252	0.0188	0.0137	0.0078	0.0051	0.0029	0.0015	0.0009	0.0003	0.0002
3	0.0237	0.0209	0.0172	0.0111	0.0069	0.0049	0.0028	0.0017	0.0009	0.0004	0.0002
4	0.0193	0.0157	0.0125	0.0092	0.0056	0.0039	0.0025	0.0014	0.0007	0.0002	0.0001
5	0.0128	0.0128	0.0097	0.0073	0.0045	0.0032	0.0020	0.0010	0.0005	0.0002	0.0001
6	0.0122	0.0086	0.0069	0.0059	0.0037	0.0026	0.0014	0.0007	0.0004	0.0002	0.0001
7	0.0087	0.0089	0.0065	0.0045	0.0030	0.0019	0.0009	0.0005	0.0003	0.0002	0.0001
8	0.0075	0.0069	0.0050	0.0034	0.0020	0.0013	0.0007	0.0005	0.0003	0.0002	0.0002
9	0.0064	0.0059	0.0030	0.0027	0.0014	0.0009	0.0006	0.0004	0.0003	0.0003	0.0003
10	0.0039	0.0041	0.0030	0.0017	0.0010	0.0007	0.0005	0.0005	0.0004	0.0003	0.0004
11	0.0037	0.0026	0.0023	0.0014	0.0009	0.0006	0.0005	0.0005	0.0005	0.0004	0.0004
12	0.0032	0.0019	0.0017	0.0011	0.0008	0.0007	0.0007	0.0006	0.0005	0.0005	0.0004
13	0.0018	0.0022	0.0018	0.0012	0.0010	0.0009	0.0008	0.0007	0.0006	0.0006	0.0006
14	0.0021	0.0021	0.0012	0.0014	0.0012	0.0010	0.0009	0.0009	0.0008	0.0008	0.0005
15	0.0018	0.0018	0.0012	0.0014	0.0013	0.0011	0.0011	0.0009	0.0009	0.0009	0.0008
16	0.0015	0.0016	0.0016	0.0014	0.0014	0.0012	0.0011	0.0010	0.0010	0.0010	0.0009
17	0.0012	0.0018	0.0014	0.0017	0.0013	0.0012	0.0011	0.0010	0.0010	0.0010	0.0009

FNMR

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	0.317	0.538	0.697	0.816	0.899	0.920	0.946	0.970	0.985	0.993	0.996
1	0.136	0.217	0.343	0.472	0.593	0.665	0.761	0.830	0.904	0.950	0.974
2	0.141	0.140	0.240	0.329	0.431	0.490	0.596	0.694	0.774	0.885	0.935
3	0.141	0.134	0.209	0.287	0.345	0.402	0.500	0.618	0.731	0.844	0.907
4	0.097	0.164	0.199	0.235	0.285	0.346	0.449	0.541	0.699	0.821	0.880
5	0.109	0.146	0.194	0.200	0.253	0.303	0.399	0.527	0.689	0.798	0.850
6	0.143	0.109	0.139	0.190	0.236	0.287	0.395	0.527	0.668	0.763	0.809
7	0.131	0.130	0.141	0.170	0.232	0.285	0.403	0.537	0.642	0.717	0.749
8	0.071	0.093	0.146	0.193	0.243	0.306	0.416	0.519	0.611	0.662	0.707
9	0.050	0.085	0.135	0.185	0.254	0.320	0.404	0.498	0.548	0.606	0.640
10	0.095	0.080	0.139	0.204	0.263	0.316	0.382	0.437	0.502	0.549	0.581
11	0.079	0.056	0.128	0.213	0.266	0.298	0.348	0.408	0.452	0.511	0.539
12	0.071	0.109	0.151	0.203	0.232	0.264	0.313	0.354	0.400	0.454	0.489
13	0.063	0.103	0.142	0.188	0.200	0.223	0.264	0.294	0.331	0.379	0.412
14	0.068	0.097	0.136	0.141	0.172	0.191	0.219	0.244	0.288	0.318	0.352
15	0.071	0.077	0.096	0.134	0.150	0.166	0.191	0.207	0.233	0.275	0.316
16	0.053	0.069	0.093	0.121	0.132	0.150	0.168	0.188	0.209	0.247	0.268
17	0.060	0.066	0.090	0.108	0.124	0.132	0.151	0.166	0.195	0.222	0.252

Figure L1. Algorithm A.

FMR

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	0.2180	0.0883	0.0471	0.0300	0.0257	0.0210	0.0129	0.0079	0.0053	0.0025	0.0009
1	0.1656	0.0935	0.0622	0.0584	0.0391	0.0271	0.0159	0.0108	0.0070	0.0034	0.0013
2	0.0894	0.0683	0.0693	0.0544	0.0322	0.0207	0.0132	0.0087	0.0056	0.0024	0.0007
3	0.0673	0.0817	0.0672	0.0436	0.0260	0.0181	0.0119	0.0070	0.0041	0.0015	0.0003
4	0.0804	0.0659	0.0544	0.0389	0.0230	0.0152	0.0093	0.0053	0.0024	0.0009	0.0003
5	0.0616	0.0555	0.0389	0.0273	0.0179	0.0119	0.0065	0.0033	0.0013	0.0004	0.0001
6	0.0556	0.0432	0.0310	0.0206	0.0125	0.0080	0.0040	0.0016	0.0006	0.0003	0.0001
7	0.0427	0.0331	0.0219	0.0156	0.0079	0.0045	0.0021	0.0008	0.0005	0.0002	0.0001
8	0.0322	0.0220	0.0154	0.0083	0.0041	0.0022	0.0010	0.0005	0.0004	0.0002	0.0001
9	0.0207	0.0160	0.0088	0.0056	0.0019	0.0010	0.0005	0.0004	0.0002	0.0002	0.0001
10	0.0152	0.0087	0.0047	0.0017	0.0009	0.0005	0.0003	0.0003	0.0002	0.0002	0.0001
11	0.0094	0.0048	0.0019	0.0008	0.0005	0.0004	0.0002	0.0002	0.0002	0.0002	0.0001
12	0.0034	0.0014	0.0009	0.0005	0.0004	0.0003	0.0002	0.0002	0.0001	0.0001	0.0001
13	0.0010	0.0007	0.0004	0.0003	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001
14	0.0010	0.0002	0.0005	0.0003	0.0002	0.0003	0.0002	0.0002	0.0002	0.0001	0.0001
15	0.0006	0.0006	0.0004	0.0003	0.0003	0.0003	0.0002	0.0002	0.0002	0.0001	0.0001
16	0.0005	0.0004	0.0004	0.0002	0.0003	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001
17	0.0003	0.0005	0.0005	0.0003	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001

FNMR

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	0.400	0.520	0.617	0.692	0.744	0.769	0.812	0.865	0.920	0.958	0.979
1	0.140	0.217	0.296	0.396	0.455	0.517	0.606	0.693	0.783	0.869	0.932
2	0.125	0.169	0.272	0.334	0.409	0.455	0.544	0.640	0.727	0.854	0.925
3	0.218	0.155	0.232	0.325	0.371	0.426	0.498	0.601	0.740	0.856	0.925
4	0.103	0.188	0.210	0.274	0.328	0.396	0.489	0.588	0.742	0.869	0.928
5	0.118	0.183	0.223	0.261	0.317	0.375	0.480	0.625	0.776	0.878	0.925
6	0.190	0.165	0.194	0.258	0.315	0.385	0.514	0.665	0.800	0.882	0.917
7	0.166	0.186	0.211	0.258	0.334	0.418	0.571	0.706	0.807	0.874	0.904
8	0.116	0.148	0.221	0.288	0.385	0.479	0.623	0.736	0.810	0.866	0.905
9	0.092	0.126	0.237	0.347	0.448	0.542	0.657	0.743	0.799	0.856	0.886
10	0.149	0.121	0.274	0.396	0.496	0.579	0.673	0.729	0.793	0.839	0.875
11	0.135	0.165	0.327	0.431	0.532	0.585	0.651	0.717	0.767	0.827	0.859
12	0.190	0.275	0.361	0.467	0.516	0.564	0.631	0.691	0.746	0.797	0.846
13	0.184	0.250	0.341	0.440	0.478	0.519	0.583	0.638	0.695	0.757	0.810
14	0.197	0.260	0.325	0.382	0.432	0.472	0.526	0.579	0.647	0.696	0.738
15	0.215	0.200	0.303	0.360	0.391	0.427	0.482	0.528	0.572	0.639	0.709
16	0.206	0.229	0.275	0.336	0.358	0.393	0.450	0.488	0.530	0.613	0.661
17	0.170	0.202	0.261	0.313	0.330	0.364	0.405	0.440	0.496	0.567	0.613

Figure L2. Algorithm B.

FMR

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	0.0603	0.0304	0.0180	0.0104	0.0074	0.0059	0.0036	0.0022	0.0013	0.0007	0.0005
1	0.0687	0.0519	0.0297	0.0236	0.0156	0.0116	0.0074	0.0050	0.0032	0.0020	0.0013
2	0.0424	0.0321	0.0311	0.0264	0.0167	0.0121	0.0083	0.0054	0.0035	0.0022	0.0010
3	0.0298	0.0366	0.0328	0.0204	0.0142	0.0107	0.0073	0.0050	0.0029	0.0013	0.0006
4	0.0326	0.0313	0.0250	0.0168	0.0120	0.0089	0.0058	0.0039	0.0018	0.0008	0.0004
5	0.0233	0.0234	0.0194	0.0146	0.0104	0.0076	0.0047	0.0027	0.0013	0.0005	0.0003
6	0.0239	0.0185	0.0176	0.0106	0.0087	0.0058	0.0033	0.0019	0.0009	0.0004	0.0002
7	0.0182	0.0188	0.0146	0.0097	0.0067	0.0042	0.0021	0.0010	0.0006	0.0003	0.0002
8	0.0182	0.0139	0.0097	0.0070	0.0044	0.0027	0.0014	0.0007	0.0004	0.0003	0.0002
9	0.0120	0.0140	0.0068	0.0057	0.0027	0.0017	0.0009	0.0006	0.0004	0.0003	0.0002
10	0.0091	0.0073	0.0049	0.0029	0.0017	0.0011	0.0006	0.0004	0.0004	0.0003	0.0002
11	0.0064	0.0045	0.0037	0.0024	0.0012	0.0009	0.0006	0.0005	0.0004	0.0003	0.0003
12	0.0050	0.0031	0.0023	0.0016	0.0011	0.0007	0.0006	0.0005	0.0003	0.0003	0.0003
13	0.0016	0.0019	0.0013	0.0010	0.0008	0.0006	0.0005	0.0005	0.0004	0.0002	0.0003
14	0.0019	0.0023	0.0018	0.0009	0.0007	0.0006	0.0006	0.0005	0.0003	0.0004	0.0003
15	0.0014	0.0017	0.0010	0.0008	0.0007	0.0007	0.0005	0.0004	0.0005	0.0004	0.0002
16	0.0014	0.0009	0.0010	0.0009	0.0007	0.0006	0.0005	0.0004	0.0004	0.0003	0.0003
17	0.0010	0.0010	0.0010	0.0007	0.0007	0.0005	0.0005	0.0004	0.0004	0.0003	0.0003

FNMR

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	0.445	0.553	0.620	0.685	0.727	0.743	0.774	0.820	0.872	0.916	0.931
1	0.196	0.267	0.325	0.410	0.440	0.479	0.528	0.587	0.653	0.727	0.790
2	0.170	0.193	0.287	0.323	0.370	0.394	0.432	0.495	0.559	0.669	0.747
3	0.210	0.179	0.243	0.305	0.324	0.354	0.398	0.471	0.569	0.692	0.759
4	0.139	0.202	0.234	0.274	0.299	0.335	0.394	0.457	0.595	0.708	0.764
5	0.149	0.206	0.226	0.245	0.276	0.316	0.378	0.466	0.598	0.709	0.756
6	0.164	0.162	0.163	0.237	0.263	0.305	0.393	0.482	0.599	0.691	0.737
7	0.198	0.162	0.183	0.210	0.259	0.311	0.403	0.497	0.590	0.667	0.705
8	0.102	0.117	0.174	0.228	0.270	0.327	0.419	0.509	0.581	0.639	0.676
9	0.070	0.108	0.182	0.231	0.294	0.348	0.422	0.494	0.550	0.616	0.637
10	0.123	0.115	0.194	0.260	0.308	0.356	0.409	0.466	0.524	0.588	0.615
11	0.113	0.089	0.182	0.239	0.311	0.347	0.392	0.455	0.499	0.566	0.589
12	0.126	0.144	0.211	0.272	0.290	0.326	0.371	0.418	0.471	0.531	0.556
13	0.113	0.123	0.210	0.241	0.266	0.296	0.336	0.371	0.426	0.471	0.516
14	0.089	0.134	0.181	0.209	0.242	0.270	0.302	0.328	0.372	0.415	0.456
15	0.105	0.132	0.159	0.186	0.222	0.245	0.276	0.298	0.331	0.375	0.430
16	0.106	0.114	0.146	0.180	0.203	0.226	0.251	0.274	0.300	0.353	0.381
17	0.081	0.117	0.144	0.170	0.191	0.208	0.229	0.254	0.288	0.321	0.374

Figure L3. Algorithm C.

FMR

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	0.1154	0.0531	0.0319	0.0184	0.0109	0.0086	0.0051	0.0030	0.0014	0.0006	0.0004
1	0.1404	0.1026	0.0669	0.0507	0.0307	0.0224	0.0137	0.0091	0.0051	0.0023	0.0013
2	0.1050	0.0761	0.0658	0.0542	0.0336	0.0243	0.0158	0.0098	0.0053	0.0026	0.0013
3	0.0725	0.0851	0.0718	0.0463	0.0319	0.0235	0.0156	0.0098	0.0052	0.0019	0.0008
4	0.0824	0.0735	0.0626	0.0431	0.0281	0.0203	0.0128	0.0076	0.0033	0.0010	0.0005
5	0.0614	0.0553	0.0451	0.0391	0.0242	0.0172	0.0102	0.0051	0.0021	0.0007	0.0003
6	0.0608	0.0472	0.0428	0.0301	0.0194	0.0129	0.0067	0.0030	0.0013	0.0006	0.0004
7	0.0446	0.0433	0.0338	0.0238	0.0145	0.0087	0.0042	0.0019	0.0009	0.0006	0.0006
8	0.0403	0.0306	0.0238	0.0156	0.0091	0.0053	0.0026	0.0013	0.0008	0.0007	0.0007
9	0.0292	0.0283	0.0162	0.0130	0.0051	0.0032	0.0018	0.0011	0.0010	0.0008	0.0007
10	0.0202	0.0161	0.0091	0.0063	0.0032	0.0020	0.0014	0.0011	0.0010	0.0009	0.0007
11	0.0159	0.0106	0.0072	0.0041	0.0023	0.0017	0.0014	0.0013	0.0011	0.0008	0.0007
12	0.0086	0.0070	0.0040	0.0029	0.0019	0.0016	0.0014	0.0013	0.0010	0.0009	0.0007
13	0.0059	0.0047	0.0032	0.0021	0.0019	0.0018	0.0016	0.0013	0.0012	0.0009	0.0008
14	0.0036	0.0038	0.0024	0.0022	0.0020	0.0018	0.0016	0.0014	0.0013	0.0012	0.0010
15	0.0027	0.0028	0.0025	0.0021	0.0021	0.0018	0.0016	0.0016	0.0013	0.0013	0.0012
16	0.0023	0.0025	0.0023	0.0022	0.0021	0.0018	0.0016	0.0014	0.0013	0.0014	0.0011
17	0.0018	0.0025	0.0025	0.0023	0.0020	0.0018	0.0017	0.0015	0.0017	0.0012	0.0012

FNMR

Youngest Age of Child in Image Pair (Years) \ Age Variation (Years)	0	1	2	3	4	5	6	7	8	9	10
0	0.457	0.531	0.594	0.667	0.714	0.729	0.760	0.812	0.870	0.919	0.937
1	0.144	0.209	0.225	0.310	0.357	0.395	0.462	0.525	0.612	0.702	0.775
2	0.133	0.130	0.204	0.233	0.288	0.314	0.372	0.446	0.519	0.645	0.737
3	0.160	0.111	0.169	0.209	0.242	0.273	0.331	0.411	0.522	0.662	0.739
4	0.064	0.116	0.141	0.181	0.204	0.247	0.309	0.382	0.528	0.666	0.739
5	0.077	0.112	0.118	0.143	0.178	0.217	0.286	0.376	0.531	0.648	0.713
6	0.107	0.091	0.096	0.139	0.168	0.207	0.291	0.397	0.521	0.622	0.680
7	0.093	0.079	0.083	0.117	0.161	0.207	0.297	0.405	0.503	0.581	0.626
8	0.051	0.067	0.108	0.125	0.175	0.229	0.317	0.400	0.490	0.534	0.591
9	0.034	0.054	0.086	0.140	0.191	0.246	0.317	0.389	0.440	0.494	0.533
10	0.060	0.049	0.108	0.147	0.206	0.251	0.304	0.352	0.402	0.454	0.493
11	0.050	0.047	0.091	0.167	0.211	0.242	0.281	0.321	0.371	0.435	0.462
12	0.076	0.075	0.124	0.165	0.192	0.217	0.256	0.297	0.343	0.394	0.430
13	0.053	0.075	0.126	0.150	0.167	0.186	0.222	0.250	0.291	0.333	0.370
14	0.055	0.079	0.105	0.125	0.146	0.163	0.190	0.206	0.248	0.279	0.304
15	0.060	0.065	0.079	0.109	0.128	0.145	0.168	0.190	0.214	0.251	0.284
16	0.050	0.067	0.084	0.104	0.116	0.130	0.150	0.171	0.191	0.230	0.244
17	0.058	0.064	0.079	0.096	0.108	0.117	0.137	0.150	0.180	0.203	0.255

Figure L4. Algorithm D

Appendix M. False Non-Match Rate Data for every Age (0-17 Years) and Age Variation (0-10 Years) for Algorithms A, B, C, and D based on a False Match Rate of 0.001

Youngest Age of Child in Image Pair (Years)	Age Variation (Years)										
	0	1	2	3	4	5	6	7	8	9	10
0	0.828	0.863	0.884	0.911	0.933	0.941	0.949	0.960	0.971	0.982	0.986
1	0.520	0.570	0.650	0.711	0.737	0.762	0.805	0.841	0.880	0.919	0.944
2	0.399	0.431	0.494	0.557	0.591	0.618	0.669	0.719	0.764	0.841	0.887
3	0.374	0.321	0.435	0.465	0.495	0.514	0.571	0.656	0.721	0.798	0.837
4	0.286	0.331	0.377	0.433	0.412	0.443	0.524	0.573	0.673	0.738	0.794
5	0.250	0.298	0.339	0.320	0.354	0.389	0.452	0.523	0.633	0.691	0.733
6	0.293	0.258	0.216	0.282	0.326	0.351	0.421	0.488	0.592	0.647	0.675
7	0.257	0.230	0.231	0.260	0.295	0.333	0.397	0.482	0.551	0.596	0.619
8	0.153	0.174	0.249	0.257	0.286	0.326	0.386	0.451	0.515	0.543	0.588
9	0.106	0.124	0.170	0.234	0.275	0.313	0.366	0.420	0.459	0.508	0.533
10	0.147	0.123	0.193	0.234	0.266	0.291	0.333	0.369	0.427	0.449	0.490
11	0.115	0.095	0.166	0.240	0.257	0.269	0.307	0.359	0.386	0.430	0.459
12	0.131	0.134	0.170	0.205	0.220	0.242	0.283	0.313	0.347	0.393	0.411
13	0.076	0.131	0.170	0.202	0.200	0.217	0.245	0.268	0.294	0.340	0.372
14	0.093	0.125	0.141	0.153	0.182	0.191	0.216	0.235	0.270	0.296	0.311
15	0.082	0.093	0.106	0.147	0.160	0.173	0.194	0.205	0.228	0.267	0.305
16	0.066	0.086	0.104	0.133	0.148	0.157	0.173	0.184	0.206	0.247	0.257
17	0.063	0.077	0.103	0.128	0.136	0.141	0.153	0.166	0.199	0.222	0.241

Figure M1. Algorithm A.

Youngest Age of Child in Image Pair (Years)	Age Variation (Years)										
	0	1	2	3	4	5	6	7	8	9	10
0	0.920	0.903	0.914	0.931	0.946	0.951	0.952	0.960	0.974	0.977	0.978
1	0.741	0.729	0.756	0.844	0.845	0.853	0.861	0.896	0.925	0.934	0.941
2	0.607	0.663	0.727	0.768	0.779	0.769	0.801	0.829	0.867	0.903	0.910
3	0.612	0.665	0.704	0.742	0.709	0.706	0.734	0.790	0.850	0.883	0.885
4	0.608	0.631	0.660	0.669	0.652	0.659	0.712	0.752	0.814	0.860	0.866
5	0.582	0.622	0.626	0.611	0.607	0.616	0.665	0.723	0.791	0.819	0.831
6	0.592	0.566	0.558	0.577	0.557	0.587	0.648	0.705	0.769	0.808	0.797
7	0.542	0.543	0.505	0.532	0.531	0.566	0.644	0.693	0.758	0.786	0.776
8	0.422	0.467	0.475	0.461	0.522	0.560	0.631	0.687	0.737	0.758	0.772
9	0.383	0.368	0.429	0.516	0.508	0.551	0.607	0.657	0.700	0.734	0.753
10	0.385	0.317	0.426	0.454	0.496	0.529	0.582	0.625	0.669	0.713	0.724
11	0.343	0.267	0.378	0.414	0.480	0.504	0.536	0.587	0.621	0.690	0.685
12	0.326	0.309	0.364	0.399	0.438	0.457	0.501	0.558	0.593	0.638	0.650
13	0.204	0.261	0.301	0.362	0.375	0.406	0.454	0.489	0.533	0.575	0.609
14	0.230	0.241	0.292	0.305	0.330	0.371	0.400	0.438	0.481	0.500	0.555
15	0.217	0.187	0.252	0.298	0.305	0.327	0.363	0.385	0.408	0.472	0.521
16	0.196	0.201	0.212	0.258	0.278	0.294	0.332	0.359	0.385	0.441	0.470
17	0.158	0.183	0.216	0.246	0.251	0.270	0.301	0.312	0.366	0.406	0.435

Figure M2. Algorithm B.

Youngest Age of Child in Image Pair (Years)	Age Variation (Years)										
	0	1	2	3	4	5	6	7	8	9	10
0	0.733	0.785	0.813	0.828	0.844	0.846	0.845	0.860	0.886	0.904	0.911
1	0.501	0.542	0.591	0.668	0.658	0.670	0.682	0.705	0.736	0.780	0.806
2	0.405	0.428	0.512	0.518	0.577	0.577	0.585	0.618	0.644	0.718	0.747
3	0.439	0.378	0.457	0.523	0.507	0.518	0.541	0.582	0.649	0.704	0.730
4	0.386	0.386	0.402	0.467	0.456	0.478	0.511	0.546	0.631	0.693	0.722
5	0.346	0.382	0.426	0.410	0.425	0.446	0.478	0.532	0.612	0.666	0.683
6	0.345	0.293	0.292	0.367	0.389	0.410	0.467	0.526	0.593	0.640	0.650
7	0.327	0.260	0.307	0.330	0.366	0.396	0.449	0.498	0.559	0.604	0.611
8	0.245	0.236	0.276	0.338	0.349	0.382	0.436	0.489	0.531	0.565	0.599
9	0.162	0.178	0.248	0.311	0.348	0.377	0.418	0.462	0.491	0.534	0.549
10	0.208	0.164	0.267	0.308	0.335	0.361	0.388	0.422	0.464	0.510	0.527
11	0.178	0.130	0.228	0.279	0.320	0.339	0.364	0.411	0.432	0.493	0.502
12	0.182	0.166	0.259	0.285	0.292	0.304	0.339	0.374	0.405	0.456	0.467
13	0.144	0.139	0.221	0.241	0.258	0.271	0.303	0.328	0.376	0.408	0.438
14	0.104	0.164	0.204	0.204	0.228	0.251	0.274	0.293	0.328	0.363	0.382
15	0.115	0.145	0.156	0.179	0.208	0.227	0.247	0.267	0.295	0.320	0.366
16	0.115	0.112	0.146	0.173	0.191	0.208	0.223	0.240	0.258	0.298	0.313
17	0.081	0.117	0.144	0.162	0.177	0.188	0.204	0.220	0.245	0.273	0.292

Figure M3. Algorithm C.

Youngest Age of Child in Image Pair (Years)	Age Variation (Years)										
	0	1	2	3	4	5	6	7	8	9	10
0	0.768	0.802	0.819	0.848	0.855	0.853	0.856	0.868	0.885	0.901	0.904
1	0.459	0.539	0.577	0.620	0.642	0.652	0.675	0.711	0.739	0.771	0.791
2	0.394	0.386	0.477	0.539	0.552	0.557	0.589	0.632	0.655	0.712	0.752
3	0.370	0.321	0.420	0.456	0.480	0.495	0.535	0.574	0.648	0.708	0.721
4	0.272	0.324	0.384	0.403	0.410	0.439	0.490	0.525	0.618	0.667	0.687
5	0.267	0.298	0.355	0.338	0.365	0.394	0.442	0.497	0.590	0.629	0.648
6	0.255	0.238	0.259	0.306	0.329	0.353	0.422	0.475	0.546	0.585	0.617
7	0.213	0.237	0.230	0.275	0.294	0.329	0.385	0.455	0.495	0.548	0.580
8	0.207	0.168	0.203	0.247	0.288	0.321	0.374	0.427	0.477	0.510	0.566
9	0.131	0.151	0.159	0.259	0.275	0.309	0.350	0.397	0.435	0.482	0.502
10	0.136	0.129	0.179	0.232	0.262	0.287	0.319	0.358	0.405	0.443	0.470
11	0.111	0.103	0.174	0.234	0.250	0.267	0.299	0.334	0.374	0.420	0.433
12	0.111	0.146	0.181	0.223	0.219	0.242	0.278	0.312	0.342	0.385	0.402
13	0.081	0.093	0.166	0.178	0.192	0.210	0.244	0.260	0.304	0.328	0.351
14	0.091	0.117	0.137	0.142	0.172	0.185	0.211	0.220	0.259	0.290	0.302
15	0.078	0.090	0.098	0.133	0.153	0.166	0.191	0.210	0.224	0.264	0.290
16	0.060	0.083	0.104	0.123	0.138	0.152	0.168	0.181	0.202	0.247	0.250
17	0.065	0.083	0.095	0.119	0.128	0.138	0.158	0.165	0.202	0.210	0.264

Figure M4. Algorithm D.

Appendix N. Statistically Significant Differences in Practitioner Performance by Age on Mated Image Pairs

Table N. *Statistically Significant Differences in Practitioner Performance by Age (compared to other Ages in Childhood) on Mated Image Pairs (at p < .001)*

	Age (Years)																	
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Age (Years)	6	9	11	10	9	10	0	0	0	0	0	0	0	0	0	0	0	0
	7	10	14	11	10	11	14	14		1	1	1	1	1	1	1	1	1
	8	11		14	11	12				4	3	3	4	4	2	4	3	4
	9	12		16	12	13					4	4	5	5	3	5	4	5
	10	13			13	14					5	5			4		5	
	11	14			14	15									5			
	12	15			15	16									6			
	13	16			16	17									7			
	14	17			17										8			
	15																	
	16																	
	17																	

The data shows that practitioners perform significantly differently with mated image pairs containing a 0 year old as the youngest age compared to image pairs containing a 6 year old and above as the youngest age. Age 0 is the only age that is statistically significant from images containing an 8 year old. As a Friedman ANOVA test confirmed that there was no statistical significance amongst different age variations on mated image pairs, there is no table to present here.

Appendix O. Statistically Significant Differences in Practitioner Accuracy by Age on Non-Mated Image Pairs

Table O. *Statistically Significant Differences in Practitioner Performance by Age (compared to other Ages in Childhood) on Non-Mated Image Pairs (at p < .001)*

	Age (Years)																	
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Age (Years)	**2**	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3	4	6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	4	5	7	14	15		2	2	2	2	2	16	2	2	2	2	2	2
	5	6	8	15	16		17	17		17	17	17	17		3	3	3	3
	6	7	9	16	17											4	4	4
	7	8	10	17													11	6
	8	9	12															7
	9	10	13															9
	10	11	14															10
	11	12	15															11
	12	13	16															12
	13	14	17															
	14	15																
	15	16																
	16	17																
	17																	

The table shows that performance on non-mated pairs containing lower ages, particularly ages 0 and 1 were significantly different from other ages across childhood. It also shows that ages in the middle of childhood were only considerably different to ages at either end of childhood (i.e., ages 0–2 and 17).

Appendix P. Statistically Significant Differences in Practitioner Accuracy by Age Variation on Non-Mated Image Pairs

Table P. *Statistically Significant Differences in Practitioner Performance by Age Variation (compared to other Age Variations) on Non-Mated Image Pairs (at p < .001)*

	Age Variation (Years)										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Age Variation (Years)	**3**	5	5	0	0	0	0	0	0	0	0
	4	6	6	9	8	1	1	1	1	1	1
	5	7	8	10	9	2	2	10	2	2	2
	6	8	9		10	9	9		4	3	3
	7	9	10			10	10			4	4
	8	10								5	5
	9									6	6
	10										7

The table shows that generally, age variations that are further away from each other are more likely to result in statistically significant differences in practitioner accuracy. It also shows that the age variations on either end of the table (i.e., age 0 and 10) are statistically different from other age variations more than others.

Appendix Q. Conference Publications List

Michalski, D., (2017*). Facial comparisons with images of children: Current research and future research directions.* Invited paper at the Facial Identification Scientific Working Group, 23–27 October, Stafford, USA.

Michalski, D., McLindin, B., Heyer, R., & Semmler, C. (2017). *The impact of ageing on facial comparisons with images of children conducted by humans and automated systems.* Invited paper presented at the 12th International Conference of the Society for Applied Research in Memory and Cognition, 3–6 January, Sydney, Australia.

Michalski, D., McLindin, B., Heyer, R., & Semmler, C. (2016). *Facial comparisons with images of children: Human and algorithm performance.* Invited paper presented at Biometrics 2016, 18–10 October, London, UK.

Michalski, D., McLindin, B., Heyer, R., & Semmler, C. (2016). *The impact of extending the validity of children's passports from 5 to 10 years.* Poster presented at the Florey International Postgraduate Research Conference, 24 September, Adelaide, Australia.

Michalski, D., McLindin, B., Heyer, R., & Semmler, C. (2016). *Facial image comparisons of children: How does the age of children in images affect human and algorithm performance?* Paper presented at the Australian and New Zealand Forensic Society 23rd International Symposium on the Forensic Sciences, 18–23 September, Auckland, New Zealand.

Heyer, R., Michalski, D., Macleod, V., Lee, H., Kardos, M., & McLindin, B. (2016). *An update on Defence Science and Technology Group's unfamiliar facial identification and face recognition.* Paper presented at the Unfamiliar Facial Identification Group Conference, 8–10 February, Sydney, Australia.

Michalski, D., McLindin, B., Heyer, R., & Semmler, C. (2015). *Is facial image comparison as effective for verifying children as it is adults in security applications?* Poster presented at the Florey International Postgraduate Research Conference, 24 September 2015, Adelaide, Australia.

Michalski, D., McLindin, B., Heyer, R., & Semmler, C. (2014). *Understanding how age affects facial matching performance of humans and facial recognition systems: Requirements*

collection from whole of government and industry. Paper presented at the Defence Human Sciences Symposium, 19–21 November 2014, Adelaide, Australia.

Heyer, R., Michalski, D., Malec, C. & McLindin, B. (2014). *Vulnerabilities of Facial Recognition: The Impact of Glasses and Ageing on Facial Recognition Performance.* Paper presented at Biometrics 2014, 21–23 October, London.

Michalski, D., McLindin, B., Heyer, R., & Semmler, C. (2014). *Examining the performance of human operators and automated facial recognition systems for the identification of missing and exploited children.* Paper presented at the Australian and New Zealand Forensic Society 22nd International Symposium on the Forensic Sciences, 31 August - 4 September 2014, Adelaide, Australia.

Michalski, D., McLindin, B., Heyer, R., & Semmler, C. (2014). *Whole of government and industry requirements collection: Understanding how the ageing variable affects the facial comparison performance of humans and automated algorithms.* Poster presented at the Florey International Postgraduate Research Conference, 25 September 2014, Adelaide, Australia.

Michalski, D., McLindin, B., Heyer, R., & Semmler, C. (2014*). Empirically examining the impact of age related variables on facial recognition algorithms and human operators: A methodological primer.* Paper and poster presented at the Biometrics Institute Asia-Pacific Conference, 28–29 May 2014, Sydney, Australia.