Expert Reliability in Legal Proceedings: “Eeny, Meeny, Miny, Moe, With Which Expert Should We Go?”

Moa Lidén,⁎, Itiel E. Dror

⁎ UCL Department of Security and Crime Science, UK/Faculty of Law, Uppsala University, Sweden

A B S T R A C T

Between Expert Reliability refers to the extent to which different experts examining identical evidence make the same observations and reach the same conclusions. Some areas of expert decision making have been shown to entail questions with relatively low Between Expert Reliability, but the disagreement between experts is not always communicated to the legal actors forming decisions on the basis of the expert evidence. In this paper, we discuss the issues of Between Expert Reliability in legal proceedings, using forensic age estimations as a case study. Across national as well international jurisdictions, there is large variation in which experts are hired to conduct age estimations as well as the methods they use. Simultaneously, age estimations can be fully decisive for outcomes e.g. in asylum law and criminal law. Using datasets obtained from the Swedish legal context, we identify that radiologists and odontologists examining knees or teeth images to estimate age seem to disagree within their own disciplines (radiologist 1 v. radiologist 2 or odontologist 1 v. odontologist 2) as well as across different disciplines (radiologist v. odontologist) relatively often. This may have large implications e.g. in cases where only one expert from the respective field is involved. The paper discusses appropriate ways for legal actors to deal with the possibility of lacking Between Expert Reliability. This is indeed a challenging task provided that legal actors are legal experts but not necessarily scientific experts.

1. Introduction

Over the past centuries, the notion of who is to be considered and accepted as an expert has changed quite dramatically [1,2], and so have the ideas of if and how expertise should be brought into legal proceedings [3–5]. Today, experts are routinely used to answer questions that are outside the expertise of the legal decision makers, e.g. in civil, asylum, family, and criminal cases. With advances in science and technology, and growing specializations, we can expect that the Courts will increasingly rely on experts.

Contemporary research highlights many potential problems with experts in legal proceedings, even when it comes to forensic science experts. This includes difficulties with communicating forensic evidence in ways that are not only true to the science [6–8] but that can also be understood by the legal actors who are to form decisions on the basis of the evidence [9,10]. Also, in adversarial settings, perceptions that experts may be “hired guns”, that is, that they sell their professional expertise to the highest bidder and are prepared to say anything as long as they are paid to do so [11–13], or, that they just have allegiance to the side that hired them, make it difficult for legal decision makers to assess the true probative value of an expert statement [14].

Moreover, when experts present conflicting views on the same issue, this becomes a dilemma for legal actors who are often unable to properly evaluate which expert should be trusted [15,16]. If the experts cannot agree what is the right answer, then how can the legal actors be expected to? It can and has been argued that: “the system now in place is schizophrenic – the judge is at the same time incompetent to make expert generalizations, but is competent to choose between two conflicting expert generalizations. Asking a lawyer to adjudicate between two expert medical opinions is as ludicrous as asking a doctor to decide a dispute between two legal experts on a question of law [17].”

In this vacuum of knowledge there is a risk that the legal actors’ choices about which expert to trust, consciously or subconsciously, are based on arbitrary criteria, or other criteria which are not arbitrary but insufficient or even flawed, such as the experts’ formal credentials, training or experience with conducting similar assessments [17,18].

When disagreeing experts are pitted against one another in a Court, at least it is evident that there is a disagreement, and an expert opinion is not necessarily the absolute and only truth. This is important since there are situations in which the disagreement may not be evident. For instance, if there is only one expert who testifies in a case, it is unknown whether and to what extent the views of this expert are representative of the views of others in the same field. There can be many possible reasons as to why only one expert is involved, including that the defense does not have the resources to obtain their own expert, the state expert is considered sufficient by the Court, or because of legal actors’ insufficient understanding about scientifically contested questions etc. In retrospect, expert disagreement has been noted, for example in...
relation to the specificity of the so-called triad (bleeding in the eye retina, brain swelling and subdural hematoma) for the shaking of a baby (Shaken Baby Syndrome) [19]. Another example is expert disagreement on whether a child can repress into the unconscious mind memories of childhood sexual abuse and then recover those memories during therapy in adulthood [20]. Similarly, there may have been disagreement between experts that were involved during the investigative phase, e.g. two experts examining the same fingerprints in a forensic lab [21] or even DNA mixture interpretation [22,23]. However, such disagreements have often already been “solved” through discussion between the experts [24]. In some cases disagreements are ‘buried’ and not included in the forensic report, and sometimes they are included in the expert opinion presented at Court [25]. This more subtle form of expert disagreement adds an extra dimension, which further complicates legal decision makers’ task. Not only should they be able to choose between experts who disagree, but they should also be able to identify situations in which it is justified to question an expert opinion, since the expert (who may not have been disputed in Court) may actually not have given them a representative or accurate answer.

Some types of forensic assessments may involve opinions from several different kinds of experts who have diverse scientific backgrounds and who also use different methods to answer the same question. One such assessment, which has clear legal relevance across both national and international jurisdictions, is age estimation. Between jurisdictions, there is considerable variation in which type of experts are involved in age estimations (e.g. radiologists, odontologists, pediatricians, pathologists, psychologists), how many experts are involved, what type of assessments they conduct (e.g. wrist, teeth, knees, collarbones, pelvic bone, interviews) and whether and how any disagreement between the experts is integrated into the decision making process [26]. Such variations have two main consequences. Firstly, the implications of between expert disagreement varies between different jurisdictions, from never being communicated at all (e.g. because there is only one expert opinion), to being integrated into the decision making process, often to the advantage of the examined individual [26]. Secondly, it is possible that there is disagreement, not only among experts within the same discipline (e.g. radiology) but also among experts from different disciplines (e.g. radiology and odontology) who use different methods to estimate the age of the same individual. As such, the topic of so-called Between Expert Reliability [27–29] in age estimation is both complex and multifaceted. The specific purpose of this paper is to evaluate Between Expert Reliability. We use forensic age estimations, as an illustrative example to the broader issues of use of experts in legal proceedings. The paper examines these in the following three ways:

1. Reliability Between Radiologists (examining knees as well as knees/ wrists combined)
2. Reliability Between Odontologists (examining teeth)
3. Reliability Across Radiologists and Odontologists (examining knees and teeth)

Also, the purpose is to discuss how the lack of Between Expert Reliability is best taken into account by legal decision makers.

In the following, we first introduce the importance of age estimations in law. Then, we define Between Expert Reliability using the Hierarchy of Expert Performance (HEP) framework. Thereafter, we provide an overview of existing research relating to Between Expert Reliability in a range of forensic assessments. We then introduce the topic of Between Expert Reliability in forensic age estimation and present data relating specifically to radiologists’ assessments of whether knees and wrists have fully matured, and odontologists’ assessments of whether teeth have fully matured. These data (N = 10,871) from the Swedish Forensic Board of Medicine, entails age assessments of teeth and knees that have been conducted in real life asylum cases, as well as data from an evaluation conducted by the Swedish National Board of Health and Welfare (N = 968). Lastly, we discuss how legal decision makers best take into account issues of Between Expert Reliability in forensic age estimations specifically, as well as experts’ reliability in general.

2. Age estimations in law

Age estimations in the legal setting are of great importance, not the least in relation to asylum law, since age estimations in this context, for example, may be fully decisive for the odds of an individual being granted asylum [30–32]. Moreover, it can be decisive for whether and in what way the individual can be detained [33,34], as well as for access to fundamental rights and safeguards that children under 18 years of age are entitled to in line with the UN Convention on the Rights of the Child and other relevant International and European standards [35,36]. Age estimations are sometimes relevant in civil law, for instance in relation to questions of legal capacity [37] and legal guardianship [38].

Another very important context in which age estimations also have a crucial role, which is not so commonly discussed, is criminal law, both national and international criminal law. This is because many crimes require that an individual is legally classified as a child, such as child trafficking, child pornography, conscripting or enlisting child soldiers and rape or sexual exploitation of a child. Certainly, the age up to which an individual is considered a child varies across different jurisdictions. For instance, the range of different ages required for legally acceptable sexual consent is 14 to 18 years [39–41] and, for criminal responsibility, some US states do not legislate a minimum age at all [42,43], whereas other US states as well as other countries vary considerably, with for example 8 [44], 10 [45,46], 12 [47,48], 14 [49], 15 [49], 16 [50], or 18 [42,51,52] year limits [53–58]. Furthermore, for the same crime, sentence severity and length will vary depending on the age of the perpetrator, most commonly in the wide age range of 15–30 years, but here again, with much variation across jurisdictions [59–64].

Regardless of the variations in age limits, it is clear that age estimations are not only important but sometimes fully decisive for the outcome of a criminal case. For instance, in a Swedish case concerning several alleged sexual crimes against a child, both parties agreed that the sex had been consensual but the legal relevance of this consent was disputed [65]. With reference to the plaintiff’s civic registration age, the prosecutor claimed that the plaintiff was 14 years old at the time of the alleged acts, while the defense claimed she was at least 15 years old (the age of sexual consent in Sweden), based on the opinion of one expert who had examined an X-ray of the plaintiff’s wrist and indicated a two-year error margin [65]. The District Court convicted the defendant based on the plaintiff’s civic registration age, a conviction which was largely upheld by the Appellate Court [66], while The Supreme Court acquitted the defendant stating that there were reasonable doubts regarding the plaintiff’s age [65].

Similarly, in international jurisdictions, the crime of conscripting, enlisting or using child soldiers to actively participate in hostilities, make age estimations in relation to the 15 year threshold necessary [67,68]. If an alleged child soldier is classified as younger than 15 years, the individual is a victim, but, should the individual instead be classified as 15 years or older, the individual is, potentially, a perpetrator responsible both for his/her crimes as a soldier and for damages in relation to victims [69]. The difficulties with estimating age in this specific context (such as long time spans from the alleged crimes to investigations and proceedings) [70–72] can easily be understood through e.g. the International Criminal Court’s (ICC) cases The Prosecutor v. Bosco Ntaganda [70] and The Prosecutor v. Thomas Lubanga Dyilo [71].

Another situation in which age estimation has dramatic consequences is the determination of whether an individual found guilty of a crime is exempt from the death penalty or not [73].
Although age estimations are, undeniably, of clear and diverse legal relevance as well as personal relevance to those involved in the legal cases, it is also today rather commonly accepted that they are just estimations, not determinations of age. Indeed, “there is a broad consensus that physical and medical age assessment methods are not backed up by empirically sound medical science and that they cannot be assumed to result in a reliable determination of chronological age. Experts agree that physical and medical age assessment methods enable, at best, an educated guess.”[30] Hence, the unreliability of age assessments has been understood primarily as a result of the lack of empirically sound methods. This has been acknowledged by the experts themselves [74], international and national Courts [65,71] as well as the Council of Europe, Children’s Rights Division [30].

There are many possible explanations as to why these methods are not empirically sound, as well as attempts to improve the methods which includes, e.g., focusing on a wider range of study populations (in terms of ethnicity, age, socioeconomic factors etc. [75–79], improving imaging techniques [80–82], reconsidering which anatomical structures (wrist, knee, teeth, pelvic bone, collar bones etc.) to include [83,84] as well as the relationship and appropriate integration of estimations based on different anatomical structures [85–87].

While these methodological considerations are indeed important for the accuracy of age estimations, they all disregard one important source of error: The humans who apply the methods and whose judgments and interpretations underpin the estimations. Arguably, understanding the human error in forensic age estimations is more fundamental than understanding other sources of errors. Regardless of which methods, samples, etc. that are used, the humans are a constant element since they are the ones that interpret and draw conclusions from the material. As such, if different experts examining the same material disagree on what they see as well as what the implications are, it does not only matter what the material is, because the human element is also contributing to the error. Hence, in this research, we add to the already existing studies by focusing on a very different perspective, less on the methods employed for forensic age estimations and more on the humans who use the methods. This is applicable to many expert domains, where the methods are examined, without any consideration to the impact of the humans using them.

3. Human error, the Hierarchy of Expert Performance (HEP) and Between Expert Reliability

The topic of human error is multifaceted, including e.g. biases arising from contextual information [88,89], case hypotheses [90,91], emotional reactions [92] and time pressure [93]. Also, absent any biasing factors, it seems that experts can disagree both with the assessments of other experts as well as with their own previous assessments [12].

These two issues are both captured by the Hierarchy of Expert Performance (HEP) framework, see Fig. 1, in which the former type of issue is referred to as Biasability and the latter as Reliability [27]. Furthermore, the HEP framework makes two other fundamental distinctions, first n observations (Levels 1–4) vs. conclusions (Levels 5–8), and second, between vs. within experts (i.e., different experts conducting the same assessment vs. the same expert conducting the same assessment at different points in time). As such, the HEP framework describes in total eight different levels at which expert performance can be conceptualized, understood and evaluated.

Level 1 of the HEP concerns Within Expert Reliability in observations, which is the most basic level of expert performance and refers to whether the same expert, looking at the same evidence at different points in time, will observe the same data (e.g., note the same minutia in a fingerprint). Level 2 is about Between Expert Reliability in observations, i.e. whether different experts examining the same evidence will make the same observations. Level 3 concerns Within Expert Biasability in observations, i.e. whether the same expert observes the evidence differently when it is presented with irrelevant contextual information, whereas Level 4, Between Expert Biasability, concerns this problem as compared between experts. Levels 5–8 mirrors the same types of issues as those in Levels 1–4 but instead of relating to experts’ observations, here it relates to their conclusions. Hence, teasing apart two critical, but separate, aspects of decision making that are often lumped together: The perception, observation, of what the data is vs. the interpretation and conclusions reached based on the observed data. Level 5 refers to Within Expert Reliability in conclusions, i.e. whether the same expert conducting the exact same assessment at different points in time would reach the same conclusion. Similarly, Level 6 concerns Between Expert Reliability in conclusions, i.e., whether different experts conducting the exact same assessment reach the same conclusion. Level 7 Within Expert Biasability in conclusions and Level 8 Between Expert Biasability in conclusions concern whether the same or different experts would reach different conclusions as a result of contextual biasing information impacting the way conclusions are reached.

This paper focuses on Between Expert Reliability in experts’ conclusions (Level 6) [27]. It is unknown how the experts’ conclusions relate to their observations, but it can be noted that this relationship is not necessarily straightforward. Ideally, experts should make the same or similar observations and their observations should lead them to their conclusions. However, it can also be that, for example, two odontologists examining the same tooth X-ray observe different age-related morphological changes of the pulp cavity, not because the pulp cavity varies but because the experts’ perceptions of it vary. Based on their different observations the experts may reach different conclusions. Yet another possibility is that the experts prefer different conclusions, and this dictates their observations, often subconsciously, which is commonly referred to as confirmation bias [27,90]. However, as illustrated by Fig. 1, it is essential to tease apart issues of biasability and issues of reliability. The data in this study comes from experts who examined the same evidence within the same contextual information, and hence any differences in their conclusions are due to reliability.

To the extent that lacking Between Expert Reliability is communicated to legal decision makers the more specific explanation(s) for the disagreement will often be unknown. Notably, disagreement between experts is not necessarily proof that at least one of the experts must be incompetent, biased or unmotivated. Even among experts, some variation is expected due to the inherent task difficulty, i.e. because the task involves integration of multiple sources of data and/or complex and ambiguous data [12]. Disagreement that stems from inherent task difficulty is important disagreement that, ideally, should be integrated...
into the decision-making process. However, this type of disagreement is not necessarily easy to distinguish from other sources of disagreement such as individual evaluator differences, that is, patterns of stable individual differences among evaluators, as opposed to mere inaccuracy or random variation, which contributes to diverging opinions. For example, in a sample of 59 clinicians conducting a total of 4,498 evaluations of legal sanity, seven clinicians found 0% of the defendants insane while three clinicians found 50% of all defendants insane [94]. Similarly, some evaluators assign consistently higher or lower scores than others on the Psychopathy Checklist Revised (PCL-R), even when there are no obvious differences among examinees that might explain these scoring trends [28,95]. Stable patterns of differences suggest that evaluators may adopt different decision thresholds [96], e.g., what number and/or type of morphological changes odontologists want to see before they feel comfortable deciding that the examined individual is over the age of 18 years. These thresholds consistently shift experts’ opinions (or instrument scores) in a particular direction, especially when faced with ambiguous cases [97]. Differences in decision thresholds can also emerge from variations in personality and/or attitudes [98–101] or that the experts have different theoretical starting points regarding, for example, the relationship between certain types of changes and age.

At least in part, different examiners’ assessments can be calibrated through systematic training with pre-and post-assessments of Between Expert Reliability, think-out-loud protocols and/or group discussions etc. This also points to another possible explanation of disagreement; different and/or limited training and certification [12]. As a consequence, evaluations are sometimes performed by “occasional experts”, i.e., general clinicians/practitioners without specialized forensic training [12]. Indeed, training seems to produce more reliable evaluations, as exemplified by a 3 day training given in Hawaii as part of a certification process (including written test, submission of mock-report, peer review process etc.), which increased Between Expert Reliability for opinions on adjudicative competency (13% increase, p = .08), for legal sanity (17% increase, p = .04) and for violence risk assessments (29% increase, p = .001) [102]. Moreover, unstandardized methods are also a source of disagreement, as exemplified by a study with a sample of 434 forensic clinicians, in which the clinicians reported using no less than 286 different tools, many with unknown reliability or validity [103]. This is likely to be a major contributor to disagreement among forensic evaluators. Moving from unstandardized to standardized methods involves reaching consensus concerning the appropriate practice [103,104]. Also, structured tools with explicit scoring rules based on objective criteria yield higher field reliability than instruments involving more holistic or subjective judgments [105,106].

However, none of these can explain lacking Within Expert Reliability (level 5 in HEP, see Fig. 1), as the differences at this level are when the same examiner assessments are compared at different times. Hence, the training, threshold, personality, etc., are same, as we are not comparing across experts, but the same expert acts as their own control. However, studies that have such within-expert data are hard to conduct and rare.

4. Between Expert Reliability in forensic assessments

Today, lacking Between Expert Reliability has been noted in a range of forensic assessments, entailing both observations and conclusions. This includes e.g. fingerprinting evidence and DNA evidence, in relation to which experts reach a spectrum of different and conflicting conclusions when they examine the same evidence [21–23]. Furthermore, in Forensic Psychiatry and Psychology lacking Between Expert Reliability has been noted in e.g. the application of the Psychopathy Checklist-Revised (PCL-R) [28,107,108], which is widely used to support decision regarding offenders’ personal liberty. Although the more specific results vary across studies [109], among 280 trained raters (both researchers and clinicians) who watched six videotaped practice cases, values of ICC such as 0.75, 0.65 and 0.78 were noted [109,110]. Also, among 22 trained, but not yet experienced raters, applying the PCL-R in relation to four videotaped practice cases, 19% of the variance was attributable to scoring tendencies of individual raters [98]. Other examples are forensic evaluations of defendants referred for competence or sanity evaluations, which historically, in Hawaii, required three independent evaluations for all defendants [111]. Statistical analyses of these evaluations show that the three independent evaluators reached different conclusions regarding competence in 29% of the cases [112]. Regarding legal sanity, a review of 165 defendants revealed that three independent experts reached different conclusions in 45% of the cases [113]. Moreover, a recent meta-analysis concluded that for evaluations of adjudicative competency, one of the most common Forensic Psychology procedures, pairs of independent evaluators assessing the same defendant disagreed in approximately 15% – 30% of cases [114]. This corresponds to rater agreement coefficients (i.e. Cohens kappa) in the range of 0.30–0.65, which indicates fair to moderate agreement according to most kappa interpretation schemes [115]. In addition, evaluators tend to disagree in almost half (45%) of conditional release cases [116].

Lack of Between Expert Reliability has also been observed in several validation studies regarding different types of Forensic Medical assessments. For instance, among three observers with varying degrees of experience who were asked to examine cut wounds to determine which specific knife (out of five possible options) had been used to make the wounds, one observer (the least experienced) failed to identify the right knife in 26% of cases, whereas the other two observers had error rates of 8% [117]. This illustrates a relatively low level of agreement between experienced and inexperienced observers [117]. Some forensic labs, as well as units for Forensic Pathology, have established practices whereby they always refer certain types of assessment to a second expert from the same lab/unit and, should this expert disagree with the first expert, the opinion of a third expert will be decisive [118]. However, it is unknown to what extent these experts are truly independent of one another, which is relevant since knowledge of another’s assessment may introduce bias [119]. It is possible that due to e.g. collegiality or time pressure, these practices become more like consultations, or even verifications, rather than blind and independent second and third opinions [120].

5. Between Expert Reliability in forensic age estimations

Forensic age estimations (FAE) provides a good case study to the broader issue of Between Expert Reliability. In contrast to e.g., fingerprinting solely conducted by the same profession, using the same methods to answer the same question (e.g. whether two fingerprints match), forensic age assessment is broader as it can entail examiners with different professions, who use different methods to answer the same question, i.e. whether the examined individual is likely to have reached a certain age threshold. This occurs because FAEs are often composed of several different types of assessments by several different experts, such as radiologists, odontologists, geneticists and pathologists (and other types of age assessments by e.g. psychologists, pediatricians and legal officers in different capacities). Hence, examining FAE gives an insight to a multitude of expert opinions, not limited to cases of opinions composed by a single expert profession using a single methodology. Various forensic domains are different, and it can be discussed whether some are better than others in making age assessments. However, the issue of between expert reliability (or lack thereof) is irrespective of forensic discipline, it is a basic forensic decision making issue that is essential to examine regardless of forensic domain. Therefore, the generalizability of our discussion is broader. Similar to many domains, there is no internationally and generally accepted framework specifying best practices, save for recommendations to use multidisciplinary and holistic approaches [26]. Hence, the methods used, as well as the humans involved, are to a large extent determined
by national practice. This also means that the more specific Between Expert Reliability issues that may be of interest will vary from jurisdiction to jurisdiction.

The most commonly used medical methods (involving radiation) among European Union (EU) states, is the wrist X-ray which is used in 76.76% (23 out of 30) states, followed by the dental X-ray used in 63.33% (19 out of 30) of states and the collar bone X-ray used in 40% (12 out of 30) of the states [26]. Also, 10% (3 out of 30) states add the pelvic bone X-ray as an alternative method to be used occasionally in the process [26]. Other examples of methods used in single states are fourth rib analysis used in Portugal and MRI of the knee used in Sweden [26].

Most countries use two or more age indicators [26]. Dental X-rays, wrist X-rays and collar bone X-rays are all age indicators included in the recommendations by The Study Group on Forensic Age Diagnostics [121,122]. Apart from differences in the methods and experts involved, there are also variations in how the judgments of different experts are integrated and what level of agreement or disagreement is accepted for conclusions that an examined individual has or has not reached a certain age threshold [26].

In Sweden, for example, the procedure used for estimations relating to the 18 year age limit was created by the National Board of Forensic Medicine [85]. This procedure includes two biological age indicators, 1) Knees (magnetic resonance imaging of the distal femoral epiphysis) and 2) Teeth (X-ray of the third molars). In the years 2017–2018, 10,871 individuals seeking asylum were subjected to this procedure [26,26], since they claimed to be below 18 years but could not provide other evidence to support their claims. The procedure is as follows. For each individual applicant, two radiologists assess the knee and two odontologists assess the teeth [25]. Both for the radiologists and for the odontologists, there are three response options: 1) “Fully matured,” (for knees: stages 4 and 5 of Schmeling’s classification, for teeth; stage H, which is the final stage of the eight stages (A-H) included in the Demirjian categorization) [85,123,124], 2) “Not fully matured,” (for knees: stages 1–3 of Schmeling’s classification, for teeth; stages A-G Hmin the Demirjian categorization) or 3) “Inconclusive”. If one of the radiologists’ or one of the odontologists’ concludes that the knee or teeth are not fully matured, this conclusion takes precedence, meaning that the overall conclusion in relation to knees or teeth is that these have not fully matured. In other words, for a conclusion that knees, or teeth are fully matured, both experts have to agree. However, thereafter, the radiologists’ and odontologists’ conclusions are integrated by a forensic pathologist. In this integration, a decision rule is applied by the forensic pathologist, and what the more specific rule is depends on the sex of the examined individual, as research has identified sex differences [121,122].

Disagreement ranges from 77% (κ = 0.51 ± 0.04) to 98.30% agreement (κ = 0.68 ± 0.04), when different types of radiological techniques were used [127]. Also, disagreement has been noted among specialist neuroradiologists [128] as well as in a range of other medical domains, including for instance surgical opinions, biopsies and diagnostic anatomic pathology [129–140]. Radiology is a medical specialty and the need for continuous training to improve radiological skills has been acknowledged [141]. This is likely to be the case especially in relation to ambiguous X-rays. When it comes to interpretation of dental X-rays, the available between Expert Reliability data has been based on very small samples of X-rays (κ = 0.81) [142] or concern other types of assessments than age estimations (e.g. periodontal disease κ = 0.83) [143].

### 5.1. Radiologists and odontologists conclusion data

In this section we provide data relating to Between Expert Reliability in the conclusions of radiologists and odontologists. We first present the overall findings in Table 1. It summarizes four different types of Between Expert Reliability with various levels of disagreement. We then present the specific data relating to each type of Between Expert Reliability in Table 2 (Between Radiologist Reliability: knees and teeth) as well as in the following text (Between Radiologist and Odontologist Reliability: combined knees/wrists).

#### 5.1.1. Between Radiologist Reliability (knees) and between Odontologist Reliability (teeth)

Table 2 outlines the data (N = 10,871) for radiologists’ conclusions in relation to knees (MRI-images) as well as odontologists’ conclusions in relation to teeth (X-rays), which was obtained from asylum cases in 2017–2018, registered by The Swedish National Board of Forensic Medicine.

As outlined in Table 2, the overall percentage disagreement in radiologists’ conclusions in relation to knees was 8.05% (875 out of 10,871). Also, the overall percentage disagreement in odontologists’ conclusions in relation to teeth was 9.29% (1,010 out of 10,871).

Apart from the data presented in Table 2, there is also other data relevant for the question of Between Radiologist Reliability (knees). For instance, as a result of private initiatives, for 137 selected cases in which the knee was considered fully matured, external second opinions were obtained from German scientists [85,144]. These scientists came to the opposite conclusion, i.e., that the knee was not fully matured, in 55% (75 out of 137) of the cases, resulting in that the Swedish Board changed their conclusions. Although these figures cannot be generalized to all individuals who had their knees assessed as mature, they do point to the risk of a clear lack of Between Radiologist Reliability (knees) [85]. Moreover, The Swedish National Forensic Board of Medicine has also reviewed the radiologists’ assessments of knee X-rays (n = 219) by having their own expert radiologists (two radiologists

### Table 1

| Type of between Expert Reliability | Disagreement (%) |
|-----------------------------------|------------------|
| Between Radiologist Reliability: knees | 8.05–55.00       |
| Between Odontologist Reliability: teeth | 9.29            |
| Between Radiologist and Odontologist Reliability: knees and teeth | 11.98–42.66       |
| (opposite conclusions)           | 2.36–15.00       |
| Between Radiologist Reliability: combined knees/wrists | 2.10–2.40       |
| (19.70 only knee)                |                  |

Note. Table 1 summarizes several different data sets and the disagreement is therefore sometimes presented in ranges with the lowest and highest disagreement observed.
with expertise in child radiology), independently of one another, assess whether the knees were fully matured, not fully matured or inconclusive [145]. Of the knee X-rays used for these assessments, 210 were knee X-rays in relation to which the two radiologists had previously agreed that the knee was fully matured, whereas only 10 were knee X-rays about which the two radiologists had previously disagreed about whether the knee was fully matured. The overall level of agreement between the two external experts and the two experts from the National Forensic Board of Medicine was 89.50% (187 out of 209). However, this is overall relatively high percentage agreement is probably largely explained by the much higher base rate of X-rays (210 out of 220) in which the two expert radiologists had agreed. The overall level of agreement probably would have been lower, had more X-rays in relation to which the experts had disagreed been included.

5.1.2. Between Radiologist and Odontologist Reliability (knees and teeth)

The overall percentage disagreement when comparing the two radiologists’ conclusions to the two odontologists’ conclusions was higher than that observed when comparing radiologists to themselves and odontologists to themselves.

As outlined in Table 2, when both radiologists concluded that the knees were fully matured (n = 8432), the odontologists disagreed in 42.66% (3597 out of 8432) of the cases. In fact, the odontologists came to the opposite conclusion, i.e. they both agreed that the teeth were not fully matured in 15.00% (1264 out of 8432) of these cases.

Vice versa, when both odontologists concluded that the teeth were fully matured (n = 5493), the radiologists disagreed in 11.98% (658 out of 5494) of the cases. In 23.65% (130 out of 5493) of the cases, the radiologists came to the opposite conclusion as they both agreed that the teeth were not fully matured.

5.1.3. Between Radiologist Reliability (knees/wrists)

This data (N = 938) was obtained from an evaluation published by The Swedish National Board of Health and Welfare in 2018 [146]. In this study, two child radiologists assessed the level of maturity of knees, both including and excluding cartilage sequences, and wrists.

Among the two child radiologists, one with 2 years of experience and one with 30 years of experience within pediatric radiology [146], the overall level of disagreement, when combining both their knee (including cartilage sequence) and wrist assessments, was 2.70% [146]. Also when looking at their more specific assessments of different parts of the knee (the distal femur, proximal tibia), the overall level of disagreement ranged from 2.70% – 2.10% [146]. However, when two general radiologists, both with approximately 1.5 years of experience in general radiology, assessed the level of maturity of the knee (excluding cartilage sequence), the overall percentage disagreement in their conclusions was 19.70%, that is, nearly a fifth [146].

6. Discussion and conclusions

The Between Expert Reliability data presented in section 4 allows us to ascertain and examine the issue of expert decision making. It suggests that there may be quite high levels of disagreement both when comparing experts within the same discipline and when comparing experts from different disciplines who use different methods to estimate the age of the same individuals. The highest disagreement within the same discipline was noted among radiologists’ assessments of knees (55%) but this can be due to this study sample being relatively small and not randomly selected. Most notable is the disagreement between radiologists’ and odontologists’ (11.98–42.66%) assessments of the same individuals. There are five main implications of these findings.

Firstly, the lack of Between Radiologist Reliability as well as Between Odontologist Reliability suggests that in jurisdictions which rely on the judgment of just one expert, there is a risk of a lottery-like process as the expert may decide the outcome more than the evidence. Hence, unless an expert domain has been demonstrated to have high levels of Between Expert Reliability, one has to be very careful when an expert presents findings. Our data from forensic age assessment is similar to data found in other forensic domains, even DNA and fingerprinting.

From the point of view of legal decision makers, the most obvious sort of damage control is to appoint another expert. Today, the number of involved experts varies between different jurisdictions. Similar to in Sweden, also in the Netherlands two radiologists, separately, are required to agree, but this is in relation to another type of assessment than those discussed here, namely whether both clavicles are fused (collarbone X-ray), and this assessment is only conducted if one expert, previously has concluded that the wrist is completely fused [26]. In Finland, two experts are instead required to draw up a joint assessment in relation to a wrist X-ray, dental X-ray as well as a dental observation [26]. Guidelines also suggest that in cases of single examinations, two experts should agree on the results [26], but given the non-binding nature of these recommendations, as well as the lack of mapping of practices across the globe in this regard, it is unknown to what extent these recommendations are abided by. Thus, this is to be considered an encouragement to jurisdictions that rely on the judgment of just one expert to include at least one more, but also an encouragement to future research to map practices globally. When experts vary in their conclusions, using a single expert is very problematic. Even using two experts may be a problem, but less so than totally relying on a single expert.

Secondly, the lacking Between Radiologist and Odontologist Reliability implies that age estimations should not be made on the basis of experts from only one of these disciplines. As stated above, most European countries use two or more age indicators [26] but since it varies whether these indicators are forensic, what type of forensic assessments etc., this is in no way a guarantee that the issue has already been solved. From the perspective of legal decision makers, it is thus not only about appointing another expert, but also another type of expert. Furthermore, it is unknown which combination of assessments produce the most accurate results and, in this regard, more systematic empirical research is needed. Since dental, wrist and collarbone assessments are included in the recommendations by The Study Group on Forensic Age Diagnostics [121,122] and most commonly used [26], these are an appropriate first step. In the broader context of expert in
the Courtroom, the suggestion would be to try and solicit experts from various domains and perspectives, that can shed light on the decision at hand.

Thirdly, given that the two first conditions have been met, the next question for legal decision makers to answer is how to integrate any expert disagreement into the decision-making process. In line with many legal presumptions such as *in dubio pro reo* (“when in doubt for the accused”) and *in dubio mitius* (“more leniently in cases of doubt”), doubts stemming from expert disagreement should be to an individual’s advantage. However, in some cases it may be more obvious than others whose advantage should be prioritized and what this means more specifically. In asylum law, the so-called *presumption of minor age* provides fairly sound guidance. The European Court of Human Rights (ECHR) has, in for example *Yazgül Yılmaz v. Turkey* [147], stated that due to the scientific inaccuracy and unreliability of age assessments methods, age assessment results have to be presented with a margin of error. The Court also emphasized that, in light of the presumption of minor age and the best interest of the child, the margin of error should always be applied in favor of the person who has undergone age assessment. Furthermore, the person shall be treated as a child until any further evidence is provided to substantiate the age of the person. It can be argued that for the benefit of doubt to be applied appropriately and fully, not only should error rates in the methods themselves be to the individual’s advantage [26] but so should disagreement among the experts applying the methods. The Swedish requirement that both radiologists and both odontologists have to agree has been mentioned as an example of an “integrated protection mechanism and expression of the principle of benefit of the doubt” [26], but it should be emphasized that this practice only applies in relation to female individuals, given that the findings that female knees seem to mature earlier than male knees. It can be questioned whether this should be applied also in relation to males. According to the EASO’s evaluation, many states, but not all; like Finland and Romania [26] do apply the benefit of the doubt in the age assessment process since error rates are to the individuals’ advantage [26]. Also, some, but not all; like Bulgaria, Finland, Netherlands, Norway, Sweden and the UK, [26], consider inconclusive results to be in the individuals favor [26].

Thus, whether disagreement between experts (to the extent that different experts are included at all) is also to the advantage of the individual, is unknown. Also, when it comes to e.g. civil law, presuming that an individual is underaged may, arguably, be to the individual’s disadvantage if the individual in question wants to be able to work or make decisions without a legal guardian. Furthermore, in e.g. criminal cases a corresponding presumption of minor age is not necessarily consistent with the standard of proof beyond reasonable doubt and the rights of the accused. This is quite clearly the case when it comes to e.g. disputes about whether an individual has been old enough to provide legally acceptable consent to sex. Always presuming minor age in such a situation would also be a presumption, if not of the accused’s guilt, then at least to the accused’s clear disadvantage. However, it is equally clear that constantly resorting to the burden of proof in criminal cases may make criminal justice inefficient as there is likely to always be a fair amount of doubt (reasonable or not) in relation to the question of age. Hence, dealing with uncertainty appropriately is probably in the end about deciding what constitutes reasonable, as opposed to unreasonable, doubt regarding someone’s age. Given the inherently open nature of the BARD-standard and judges’ discretion to determine its meaning in a given case, there is often an unwillingness among legal actors to attempt to establish general rules that apply across criminal cases that involve age estimation.

In the interest of treating like cases alike, it may be appropriate for higher Courts to set precedence as to what evidence is required in this regard, and based on the data presented here, it seems reasonable to expect at the very least that two independent experts from different disciplines agree on the question of age for a finding to the defendant’s disadvantage. Another question for future research to answer is what experts and methods this would entail. Since prosecutors regularly have duties to take both incriminating and exonerating information into account, and also have the burden of proof in criminal cases, it seems reasonable that the expert evidence should be obtained by the prosecution, and, in fact, be incorporated already in the decision on whether to press charges as well as presented in Court. To assist legal actors in the difficult task of understanding and appropriately integrating expert disagreement, it is advisable for experts to report their findings in terms of age intervals with minimum and maximum ages rather than one single age [148] as this is likely to prevent communicational issues between the legal actors and the involved experts.

Fourthly, while the access to appellate procedures or other forms of judicial review vary between different jurisdictions as well as legal fields, much of the law applicable on asylum procedures in Europe [30,36,149–153] require that either the age assessment itself, or the legal decision formed on the basis of the age assessment (or both) should be subject to administrative or judicial appeal. Furthermore, these procedures should be child-sensitive, accessible for the child and his/her legal representative and the information on the possibility to appeal needs to be provided in a language that the child understands [154]. In 15 member states of the Council of Europe, the individual has a right to appeal the age assessment decision or to have it reviewed. In 13 countries, the person also has access to state funded legal counsel when seeking to appeal an age assessment decision or having it reviewed (out of 37 survey responses) [26]. Overall, when assessing whether individuals have sufficient legal remedies against decisions on age assessments, the EASO concluded that 8 states offer the possibility to challenge the age assessment decision separately, 9 states offer the possibility to challenge it as part of the international protection decision or simultaneously, while 2 states; Cyprus and Slovakia, do not offer legal remedies to the applicant against age assessments [26] and more information is needed from 11 states [26]. Given that age estimations are associated with several difficulties, lack of Between Expert Reliability being one of them, it can be questioned whether legal decision makers and/or policy makers should consider whether the legal remedies available against age assessments are sufficient also in other legal fields. Although the age evidence may be exactly the same, its legal significance may vary considerably across different legal fields (for instance, due to differences in standards of proof). It seems reasonable to further examine whether age assessment procedures, including legal remedies against the assessments, should be more streamlined.

Fifthly, for integration of disagreement to have any real effect on the quality of the decision-making process it is of course required that the disagreement has a sound basis. Ideally, disagreement should be based on the task difficulty, rather than undiscovered or unresolved differences in training, experience, theoretical starting points and decision thresholds among experts. Thus, if lacking Between Expert Reliability, which is traditionally considered a curse, is to be turned into an asset (providing multiple views on a difficult question) it has to be ensured that the experts disagree for the right reasons. In practice, this is probably difficult for legal decision makers to accomplish, or even contribute to, but it is instead a question for researchers and the experts themselves. Best practice guidelines and training designed to calibrate experts’ assessments should be useful in this regard. In addition, the potential of technical support, such as computer-assisted interpretation of X-ray or MRI images, as well as machine learning (AI), should be evaluated, as this may help reduce deviations between and within experts [26,155]. Although it is impossible to say exactly why the radiologists and odontologists (whose assessments form the basis of the data in this study) disagreed, there are indications that the disagreement is at least partially due to a lack of training, certification as well as *unstandardized methods*. This is because the implementation of the procedure used in Sweden today (see the full description in section 4) was not preceded by any validation of assessments, whereby apprentices make assessments under the supervision of experienced assessors.

M. Lidén and I.E. Dror

Science & Justice 61 (2021) 37–46
Acknowledgment

Declaration of Competing Interest

7. Ethical standards

7.1. Acknowledgment

This research was funded by Ragnar Söderberg’s Foundation and the Swedish Research Council.

References

[1] C.A.G. Jones, Expert Witnesses: Science, Medicine, and the Practice of Law, Oxford Socio-legal Studies (1994).
[2] B.M. Epp, Expert Witnesses: Law, 4th edition, 2017.
[3] C.M. Milroy, A Brief History of the Expert Witness, Acad Forensic Pathol, 2017.
[4] M. Lessof, Laws of men and laws of nature: the history of scientific expert testi-

ment in England and America, Clin. Med. (Northfield. Il) (2005), https://doi.org/10.1086/clinmedicine.5-5-526.
[5] C. Beech, The admissibility of expert opinions in Canada Courts, Law Bus. Rev. Am. 21 (2015) 361–368.
[6] C. Adam, Forensic Evidence in Court, 2016, https://doi.org/10.1002/ 9781119054443.
[7] W. O’Brien, Court scrutiny of expert evidence: recent decisions highlight the tensions, Int. J. Evol. Proof. 7 (2003) 172–184.
[8] B.D. Sales, D.W. Shuman, Experts in court: reconciling law, science, and profes-

sional knowledge, 2006, https://doi.org/10.1037/11155-000.
[9] R.C. Harris, Black Robes, white coats: The Puzzle of judicial policymaking and scientific evidence, 2008, https://doi.org/10.5860/choice.46-5887.
[10] C.A. Reid, L.M. Howes, Communicating forensic scientific expertise: An analysis of expert reports and corresponding testimony in Tasmanian courts, Sci. Justice (2020), https://doi.org/10.1016/j.scijus.2019.09.007.
[11] P.J. Candilis, R. Weinstock, R. Martinez, Forensic ethics and the expert witness, 2007, https://doi.org/10.1097/01.rfs.0000286498.93680.99.
[12] L.A. Guarnera, D.C. Murrie, M.T. Boccaccini. Why do forensic experts disagree? Sources of unreliability and bias in forensic psychology evaluations, Transl. Issues Psychol. Sci. (2017), https://doi.org/10.1177/1745240616650011.
[13] C. Kim, P.S. Koh, Court-appointed experts and accuracy in adversarial litigation, Int. J. Econ. Theory (2010), https://doi.org/10.1142/s1299519712500292.
[14] D.C. Murrie, M.T. Boccaccini, L.A. Guarnera, K.A. Rufino, Are forensic experts biased by the side that retained them? Psychol. Sci. (2013), https://doi.org/10.1177/0956797613481812.
[15] D. Sarewitz, The voice of science: Let’s agree to disagree, Nature, (2011), https://doi.org/10.1038/478007a.
[16] C. Taylor, P. Colman, Forensics: Experts disagree on statistics from DNA tawls, Nature (2010), https://doi.org/10.1038/461266c.
[17] M. Hor, When experts disagree, Singapore, J. Leg. Stud. (2000) 241–262.
[18] G. Edmond, Legal versus non-legal approaches to forensic science evidence, Int. J. Evol. Proof. (2016), https://doi.org/10.1177/1365712715513470.
[19] D. Tuerkheimer, Flawed convictions: “Shaken baby syndrome” and the inertia of injustice, 2014.
[20] C. Bannon, Recovered memories of childhood sexual abuse: should the courts get involved when mental health professionals disagree? Ariz. State Law J. 26 (1994) 855–895.
[21] I. Dör, R. Rosenthal, Meta-analytically quantifying the reliability and biasability of forensic experts, J. Forensic Sci. (2008), https://doi.org/10.1111/j.1556-4029.2008.00762.x.
[22] I.E. Dror, G. Hampikian, Subjectivity and bias in forensic DNA mixture interpre-

tation, Forensic Science, Sci. Justice, (2011), https://doi.org/10.1111/j.1189-9548.2011.00757.x.
[23] J.M. Butler, M.C. Kline, M.D. Coble, NIST interlaboratory studies involving DNA mixtures (MIX05 and MIX13): variation observed and lessons learned, Forensic Sci. Int. Genet. (2018), https://doi.org/10.1016/j.fsgen.2018.07.024.
[24] K.N. Ballantyne, G. Edmond, B. Found, Peer review in forensic science, Forensic Sci. Int. Genet. (2017), https://doi.org/10.1016/j.fsgen.2017.05.020.
[25] The Swedish National Board of Forensic Medicine, Medial Age Assessments, (n.d.).
[26] E.A.S.O. (EASO), Practical Guide on Age Assessment, 2018.
[27] L.E. Dör, A Hierarchy of Expert Performance, J. Appl. Res. Mem. Cogn. (2016), https://doi.org/10.1177/1747599116632579.
[28] L.E. Dör, D.C. Murrie, A hierarchy of expert performance applied to forensic psychological assessment, Psychol. Public Policy, Law (2018), https://doi.org/10.1037/law0000146.
[29] C.Y. Huang, R. Bull, Applying Hierarchy of Expert Performance (HEP) to Investigative Interview Evaluation - Strengths, Challenges and Future Directions, Psychiatry, Psychol. Law. (202AD), https://doi.org/10.1218719.2020.1770634.
[30] D. Wenke, Council of Europe, Children’s Rights Division, Age Assessments: Council of Europe Member States’ Policies, Procedures and Practices Respectful of Children’s Rights in the Context of Ignation, 2017.
[31] K. Schittenhelm, Implementing and rethinking the European Union’s Asylum Legislation: The Asylum Procedures Directive, Int. Migr. (2019), https://doi.

org/10.1111/mig.12533.
[32] T. Magner, Refugee, Asylum, and Related Legislation in the US Congress: 2013–2016, J. Migr. Hum. Secur. (2016), https://doi.org/10.14240/jmhs.v4i4.67.
[33] The UN Convention on the Rights of the Child, Article 37b, n.d.
[34] Committee on the Rights of the Child, General Comment No. 6, 2005.
[35] S.C. in E. Programme, Position Paper on Age Assessment in the Context of Separated Children in Europe, (n.d.).
[36] M. du Monde, Age Assessment for Unaccompanied Minors, When European Countries Deny Children their Childhood?, 2015.
[37] S. Ceseri, S. Scuro, Full Legal Capacity Acquired Before the Age of Majority, RSP 46 (2015) 297–304.
[38] A. Weatherburn, Y.E. Mellon, Child trafficking victims and legal guardians: Exploring the fulfilment of the EU trafficking directive in the context of the UK modern Slavery Act 2015 – Best practice or not fit for purpose? New J. Eur. Crim. Law (2019), https://doi.org/10.1177/2032284418836510.
[39] G. Zhu, S. van der Aa, Trends of age of consent legislation in Europe and China: towards a gender-neutral age of consent in China? Child Law, Crime, Policy Res. (2017), https://doi.org/10.1177/1749335217699293.
[40] H. Graupner, Sexual consent: the criminal law in Europe and outside of Europe, Arch. Sex. Behav. 29 (2000) 415–461.
[41] G. Zhu, S. van der Aa, A comparison of the gender-specificity of age of consent legislation in Europe and China towards a gender-neutral age of consent in China? R. J. Crim. Policy Res. (2017), https://doi.org/10.1007/s11419-017-9352-2.
[42] D. Cipriani, Children’s rights and the minimum age of criminal responsibility: A global perspective, 2009.
[43] K. Fitz-Gibbon, Protections for children before the law: An empirical analysis of the age of criminal responsibility, the abolition of dehi incapax and the merits of a developmental immunity defence in England and Wales, Criminol. Crim. Justice (2016), https://doi.org/10.1007/174898581632579.
[44] S. Government, No Title, (n.d.).
