Biometrische Stellungnahme zu den Referenzpublikationen von Kellinghaus et al. (2010a, 2010b)

Kellinghaus M, Schulz R, Vieth V, Schmidt S, Pfeiffer H, Schmeling A (2010a). Enhanced possibilities to make statements on the ossification status of the medial clavicular epiphysis using an amplified staging scheme in evaluating thin-slice CT scans. Int J Legal Med 124: 321–325.

Kellinghaus M, Schulz R, Vieth V, Schmidt S, Schmeling A (2010b). Forensic age estimation in living subjects based on the ossification status of the medial clavicular epiphysis as revealed by thin-slice multidetector computed tomography. Int J Legal Med 124(2):149–154.

Einleitung

Bei der multifaktoriellen, medizinischen Altersdiagnose in Österreich wird ein auf mehreren individuellen medizinischen Untersuchungen (körperlichen, zahnärztlichen, radiologischen und ggfs. computertomographischen Teiluntersuchungen) basierendes Modell angewandt. Insbesondere wird dabei auf die Empfehlungen zur Altersdiagnostik bei Jugendlichen und jungen Erwachsenen außerhalb des Strafverfahrens der Arbeitsgemeinschaft für Forensische Altersdiagnostik der Deutschen Gesellschaft für Rechtsmedizin (AGFAD) Rücksicht genommen. Da laut des UNHCR-Büros in Österreich in letzter Zeit von verschiedenen Seiten vermehrt Bedenken an der hinreichenden wissenschaftlichen Absicherung durch entsprechende Referenzstudien geäußert wurden, ist UNHCR mit dem Ersuchen an uns herangetreten, eine methodenkritische Stellungnahme zu diesen Bedenken bzw. zu diesen Referenzstudien zu erstellen. Bei den besagten Publikation handelt es sich um die oben angeführten Studien von Kellinghaus et al., die als Referenz im Rahmen von Untersuchungen des Schlüsselbeins herangezogen werden.

Die vorliegende Stellungnahme konzentriert sich vor allem auf die Aussagekraft der Publikationsresultate, auf deren Verallgemeinerbarkeit und insbesondere auf die Konsistenz der Publikationen mit allgemeinen wissenschaftlichen Standards, vornehmlich den AGFAD-Standards[1], insbesondere da in medizinischen Sachverständigengutachten direkt auf Referenzwerte aus diesen Publikationen Bezug genommen wird.

Vorausschickend ist anzumerken, dass sich die beiden Publikationen auf denselben Datensatz beziehen, wobei in der ersten Publikation die Gesamtstichprobe analysiert und in der zweiten Publikation ein gezielter Teilausschnitt davon neuerlich betrachtet wurde.

Diskussion der Publikationen

1. **Altersverteilung:**
   - **Kellinghaus et al. 2010a:** Die Altersverteilung sowohl in der Gesamtstichprobe als auch in den geschlechtlich segregierten Stichproben ist nicht gleichmäßig (siehe Abbildung 1 und Abbildung 2). Es zeigt sich ansatzweise eine dreigipfelige Verteilung, wobei insbesondere bei 15 und 16 Jahren ein Verteilungseinbruch zu verzeichnen ist. Ein Anpassungstest auf Gleichverteilung zeigt auch hoch signifikante statistische Abweichungen (K-S-Z=1.632, p=.010).

     Im Geschlechtervergleich wird insbesondere deutlich, dass sich die Altersverteilungen der weiblichen und männlichen PatientInnen unterscheiden, da es in der Gruppe der Männer <=16 deutlich weniger

[1] Lockemann U., Fuhrmann A., Püschel K., Schmeling A., Geserick G. (ohne Jahreszahl). Empfehlungen für die Altersdiagnostik bei Jugendlichen und jungen Erwachsenen außerhalb des Strafverfahrens. Arbeitsgemeinschaft für Forensische Altersdiagnostik der Deutschen Gesellschaft für Rechtsmedizin (AGFAD).

Probanden gibt als in der Gruppe der Frauen, während die Männer deutlich den Bereich >28 dominieren. Im Mittel sind die weiblichen PatientInnen demnach 21.9±7.2 Jahre, die männlichen 23.9±7.2 Jahre (t(500)=3.265, p=.001) und demnach signifikant unterschiedlich. Legt man einen Cut-Off für den in der Praxis relevanten Altersbereich (<18 und >=18), dann resultieren 66 (30.8%) weibliche PatientInnen und 61 (21.2%) männliche PatientInnen, was diesen Punkt weiter unterstreicht (p_{exakt}=.017).

Während die Frauen knapp noch als gleich verteilt gelten können (K-S-Z=1.280, p=.076), ist dies bei den Männern definitiv nicht der Fall (K-S-Z=1.982, p=.001).

Prinzipiell wäre ein Oversampling[2] im Bereich der 16-17-Jährigen gut argumentierbar, weil für die praktische Anwendung Fehlklassifikationen dieser Altersgruppe als volljährig das größte Risiko auf seiten der Betroffenen darstellen werden. Die Abweichungen gehen aber gerade nicht in diese Richtung – es gibt besonders wenige 16-Jährige (vgl. Abbildung 1).

Der AGFAD-Standard nach einer gleichmäßigen Altersverteilung *(vgl. AGFAD Richtlinien, S. 4, Punkt ‚gleichmäßige Altersverteilung')* wird hier sowohl in der Gesamtstichprobe als auch speziell in der Gruppe der Männer nicht erfüllt.

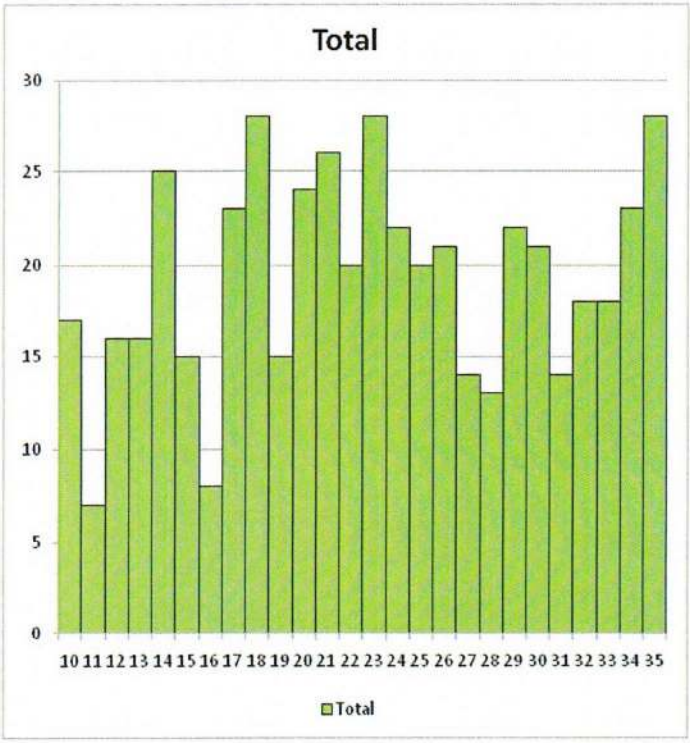

Abbildung 1: Altersverteilung in Gesamtstichprobe

[2] Gezielte/willentliche Abweichung von der Gleichverteilung, um für relevante Bevölkerungsgruppen größere Fallzahlen zu gewinnen.

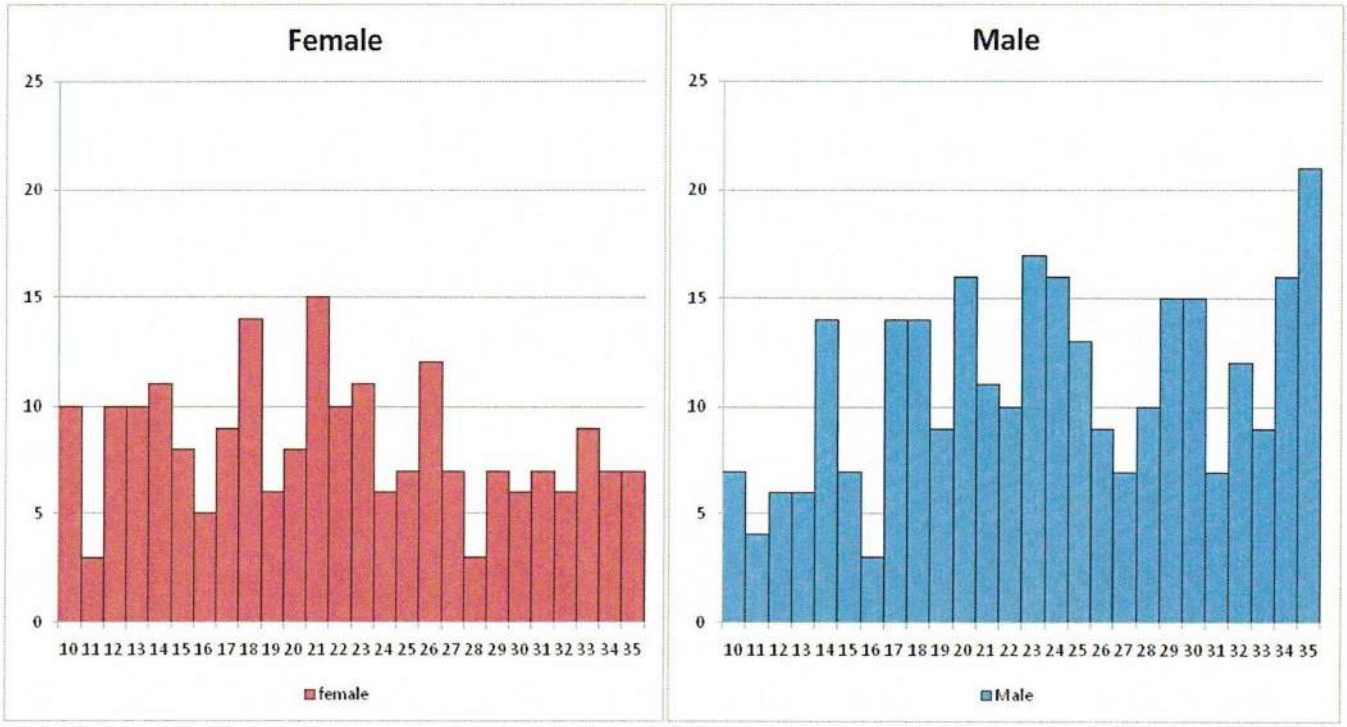

Abbildung 2: Altersverteilung nach Geschlecht in Kellinghaus et al. 2010a

- **Kellinghaus et al. 2010b:** Hier wurden PatientInnen der Stadien 2 und 3 reanalysiert. Die sich aus der Hauptstudie ergebenden Geschlechtsunterschiede sind demnach auch hier vorhanden; Männer sind im Mittel um ein Jahr älter (Frauen: 19.2±2.9; Männer: 20.2±2.8, t(183)=2.483, p=.014). Darüber hinaus zeigen die Verteilungen ganz deutlich (vgl. Abbildung 3), dass es bei den männlichen Patienten nur mehr sehr wenige Probanden (insgesamt 20) im sensiblen Altersbereich <18 gibt. Um eine Größenordnung der statistischen Unsicherheit in so kleinen Stichproben anzugeben: Selbst wenn die Wahrscheinlichkeit für eine Fehlzuordnung 10% betragen würde, bestünde immer noch eine Wahrscheinlichkeit von 12%, in einer derartigen Stichprobe keine einzige Fehlzuordnung zu beobachten. D.h. die statistische Information reicht nicht aus, um mit der gebotenen Sicherheit zu beurteilen, ob Fehlzuordnungen auszuschließen oder nur nicht sehr häufig sind.

  Wenngleich es sich wahrscheinlich um einen marginalen Irrtum handelt, soll doch angeführt werden, dass es in Tabelle 1 (Kellinghaus et al. 2010b, S. 322) gerade in der Gruppe der 16-jährigen Männer auch einen Fehler zu geben scheint, da dort vier 16-Jährige Männer aufscheinen, während es in der Gesamtgruppe (Kellinghaus et al. 2010a, Tabelle 1, S. 150) anfänglich nur drei 16-jährige Männer waren.

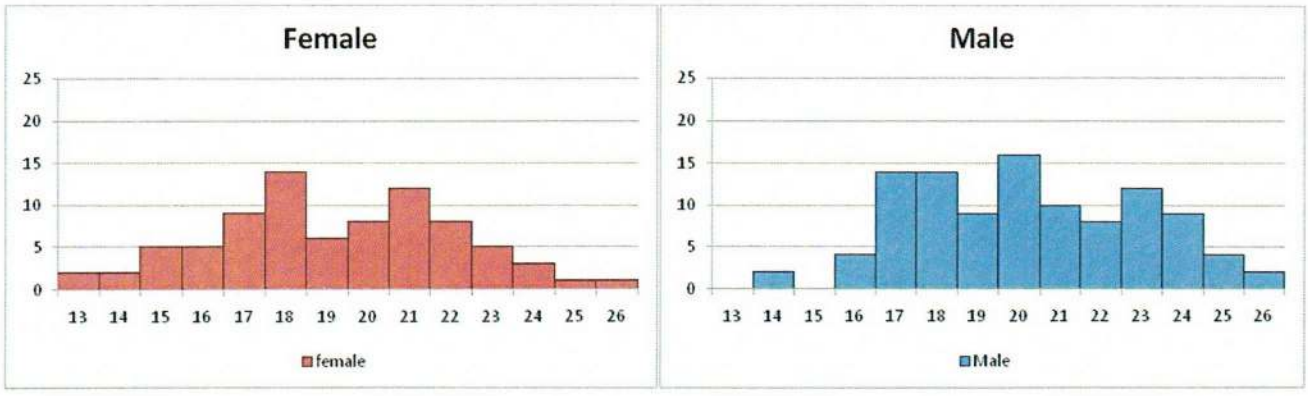

Abbildung 3: Altersverteilung in Kellinghaus et al. 2010b

Der AGFAD-Standard einer adäquaten Stichprobengröße unter Berücksichtigung der Zahl der erfassten Altersklassen und Bevölkerungsgruppen *(vgl. AGFAD Richtlinien, S. 4, Punkt ‚adäquate Stichprobengrößen')* wird hier speziell in der Gruppe der Männer nicht erfüllt.

2. Bestimmung der Stadien (Verblindung & Raterübereinstimmung):

- **Kellinghaus et al 2010a:** Positiv hervorzuheben ist, dass das Material hinsichtlich der Ossifikationsstadien verblindet[3] bewertet wurde (Kellinghaus et al. 2010a, S. 151, 2. Spalte Mitte). Allerdings wäre ein Verweis auf die Reliabilität[4] dieser Bewertungen in Form von Raterübereinstimmungsangaben (Kappa o.ä.) ebenfalls wichtig gewesen, da hier sehr wahrscheinlich Unschärfen in der Einstufung vorliegen.

- **Kellinghaus et al 2010b:** Im Artikel von 2010b fehlen hingegen Angaben zur Verblindung. Es gibt nur einen Verweis zu Schwierigkeiten in der Bestimmung des Stadiums und zum Einsatz eines ‚Measuring Tools'. Die Angabe, bei wievielen Fällen dieses eingesetzt werden musste, fehlt jedoch. Ein Kommentar zur Verblindung wäre insbesondere auch deshalb erforderlich gewesen, da bis auf eine alle AutorInnen der Studie 2010b auch an der Studie 2010a mitgewirkt hatten und dementsprechend mit dem Datenmaterial bereits befasst waren. Des Weiteren wären auch bei der Studie aus dem Jahr 2010b Angaben zur Reliabilität der Stadienbestimmungen (Raterübereinstimmungsangaben) wünschenswert gewesen.

Ein Verweis auf die Reliabilität dieser Bewertungen in Form von Raterübereinstimmungsangaben wäre im Sinne des AGFAD-Standards *(vgl. S. 4 Punkt 7: „genaue Beschreibung der Methodik")* zu fordern.

In 2010b unterbleibt die Schlüsselinformation in Bezug auf Verblindung, welche im Sinne des AGFAD-Standards *(vgl. S. 4 Punkt 7: „genaue Beschreibung der Methodik")* bzw. allgemeiner wissenschaftlicher Standards unverzichtbar ist. Ohne Verblindung wären die Ergebnisse hinfällig.

3. Fehlende/Unvollständige Fallzahlen in der Ergebnisdarstellung

- **Kellinghaus et al. 2010a:** Eine statistische Bewertung der Ergebnisse gemäß Tabelle 2, S. 152, wird sehr erschwert, da die Autoren ihre Gruppengrößen pro Stadium und Geschlecht nicht deutlich anführen, was sowohl nach wissenschaftlichen Standards als auch *gemäß AGFAD Richtlinien, S. 4 Punkt 9* gefordert wäre. Aus dem Diskussionsteil der Publikation lässt sich nur folgern, dass sich in Stadium 4 insgesamt 139 Fälle (88 Männer, 51 Frauen) und in Stadium 5 insgesamt 94 Fälle (55 Männer, 39 Frauen) befunden haben. Dies bedeutet, dass bei den Frauen für die Stadien 1-3 nur mehr 124 Fälle und für die Männer 145 Fälle verbleiben können. Gemäß Kellinghaus et al. 2010b wurde Stadium 2 und 3 reanalysiert, was einen ‚detektivischen' Rückschluss auf die Fallgrößen gestattet (Stadium 2: 53; Stadium 3: 152), aber nicht mehr nach dem Geschlecht getrennt. Offen bleibt nach wie vor: Wieviele Frauen gab es in Stadium 2, wieviele Männer gab es in Stadium 2, wieviele Frauen gab es in Stadium 3 und wieviele Männer gab es in Stadium 3?

  - Nicht verständlich ist außerdem, dass in Tabelle 2, entgegen der Altersverteilungstabelle (Tabelle 1, S. 150), das Alter nun auf zwei Dezimalen genau angegeben wurde, was ein Zusammenführen mit

[3] Den BeurteilerInnen der Stadien gemäß CT-Befund war das tatsächliche Alter nicht bekannt.

[4] Die Zuverlässigkeit dieser Stadienbewertung, d.h. speziell ob andere BeurteilerInnen gemäß CT-Befund zur selben Einstufung gelangen (i.e. Raterübereinstimmung) – üblicherweise wird diese durch den Kappa-Koeffizienten numerisch bewertet, der in akademischem Zusammenhang mindestens 0.7 betragen sollte.

Tabelle 1 erschwert. Beispielsweise ist nicht nachzuvollziehen, ob das Maximum aus Stadium 1, male, von Max=15.98 den 15-Jährigen oder bereits den 16-Jährigen aus Tabelle 1 zugerechnet wurde, etc.

- **Kellinghaus et al. 2010b:** Auch in Tabelle 2 (S. 323) werden erneut keine Fallzahlen angeführt. Dies wäre vor allem von Bedeutung gewesen, da es sich um deutlich reduzierte Gruppengrößen handelt. Wie zuvor schon erwähnt, umfasst Stadium 2 53 Fälle und Stadium 3 132 Fälle. Da nun beide Stadien auf 3 Substadien untergliedert und nach Geschlecht getrennt behandelt werden, entfallen im statistischen Schnitt lediglich 8.8 PatientInnen auf ein Segment in Stadium 2 und nur 22 PatientInnen auf ein Segment in Stadium 3. Betrachtet man in diesem Zusammenhang nun die statistischen Kenngrößen aus Tabelle 2 (vgl. Abbildung 5), wird insbesondere bei den Frauen in Stadium 2c deutlich, dass das Obere Quartil (UQ) und das Maximum nahezu zusammenfallen. Dies ist entweder ein Hinweis auf eine extrem rechtssteile Verteilung[5] in diesem Segment oder auf eine sehr kleine Gruppengröße. Dementsprechend sind die Vergleiche deutlich ‚underpowered'[6] und bringt etwa die statistische Überprüfung eines Geschlechtsunterschieds innerhalb der Stadien allein schon aufgrund der geringen Fallzahlen in den Segmenten keine inferenzstatistischen Überzufälligkeiten. Dies obwohl sich deskriptiv durchaus Unterschiede zwischen den Geschlechtern zeigen (siehe Abbildung 5).

  Bezüglich der Altersangaben in Tabelle 2 (S. 323) verwenden die AutorInnen nun eine Dezimale, was auch in dieser Studie ein Zusammenführen mit der Altersverteilungstabelle (Tabelle 1 S. 322) unnötig erschwert.

Die Gruppengrößen nach Stadien und Geschlecht sind sowohl in Kellinghaus et al. 2010a (teilweise) als auch insbesondere in Kellinghaus et al. 2010b (aus der Publikationen nicht entnehmbar) nicht dokumentiert, weshalb der AGFAD Standard (*S. 4, Punkt 9 „Angabe von Gruppengrößen, Mittelwert und einem Streuungsmaß für jedes untersuchte Merkmal")* nicht erfüllt ist. Die uneinheitliche Darstellung des Alters in Bezug auf Dezimalstellen erschwert die Nachvollziehbarkeit zusätzlich.

Der AGFAD Standard *„adäquate Stichprobengrößen unter Berücksichtigung der Zahl der erfassten Altersklassen und Bevölkerungsgruppen" (S. 4, Punkt 1)* ist bei Kellinghaus et al. 2010b schwerlich erfüllt, da die Fallzahl pro Substadium und Geschlecht (leider exakt nicht bekannt) anhand der verbleibenden Personen zu gering ist, um darzulegen, dass deren Eigenschaften jene einer wie auch immer definierten Zielpopulation verlässlich abbilden.

4. Stabilität von Minimum-Maximum-Angaben:

Minima und Maxima in Stichproben sind Extremwerte (oftmals sogar Ausreißer) und damit naturgemäß großen Schwankungen ausgesetzt. Nur mit Glück wird das wahre Minimum einer Referenzpopulation von z.B. 500 Personen genau in eine beobachtete Teilstichprobe von <100 Personen fallen. Die Beurteilung der Messunschärfe von Minima ist dementsprechend heikel und bedarf sorgfältiger statistischer Argumentation (etwa über Verteilungsannahmen). Die Publikationen bleiben bei der Angabe von Minima jedoch rein deskriptiv.

Bezüglich **Kellinghaus et al. 2010a, Tabelle 2, S. 152**: Betrachtet man den Range (Max-Min) bzw. den IQR (UQ-LQ) (vgl. Abbildung 4), so zeigt sich deutlich, dass ab Stadium 3 die Streuungen in beiden Geschlechtern zunehmen und in Stadium 4 am größten sind. Dass bei derart großer Variabilität überhaupt sichere untere Grenzen angegeben werden können, ist zwar nicht ausgeschlossen, bedarf wie gesagt aber tiefergreifender statistischer Auseinandersetzung.

[5] Eine Verteilung ist rechtssteil, wenn sich besonders viele Fälle nahe dem oberen Ende der Verteilung befinden.

[6] Statistische Macht (‚power') bezeichnet die Wahrscheinlichkeit, einen existenten Effekt in einer Stichprobe auch tatsächlich statistisch nachzuweisen.

Kellinghaus et al. 2010b: Vergleicht man das Minimum der Männer aus 2c (17.1) mit dem der Männer aus 3a (17.5), so zeigt sich nur eine Differenz von 0,4 Jahren. Beim Vergleich der Minima der Frauen aus 2b (15.4) mit denen aus 2c (15.6) sind es überhaupt nur mehr 0,2 Jahre. Dementsprechend scheint aufgrund der kleinen Stichprobe die Verteilung der Minima sehr zufällig zu sein. Das Ausreißermaximum in Stadium 3b männlich (vgl. Abbildung 5) ist mit 25.4 ein weiterer Indikator für die Instabilität von Maximum-Minimum-Betrachtungen.

In beiden Publikationen (2010a und 2010b) fehlen Angaben zur Schwankungsbreite von Minima und Maxima, welche von der AGFAD *(vgl. S. 4 Punkt 9: „Angaben von Gruppengröße, Mittelwert und einem Streuungsmaß für jedes untersuchte Merkmal")* gefordert werden. Minimum und Maximum sind deshalb als „Merkmale" aufzufassen, weil insbesondere das Minimum als Schätzer und Referenzgröße für die spätere Anwendung fungiert und damit über die Verwendung als deskriptive Stichprobencharakteristik weit hinausgeht. Es fehlen also Angaben zur statistischen Unsicherheit in Bezug auf eine der Schlüsselgrößen für die praktische diagnostische Umsetzung.

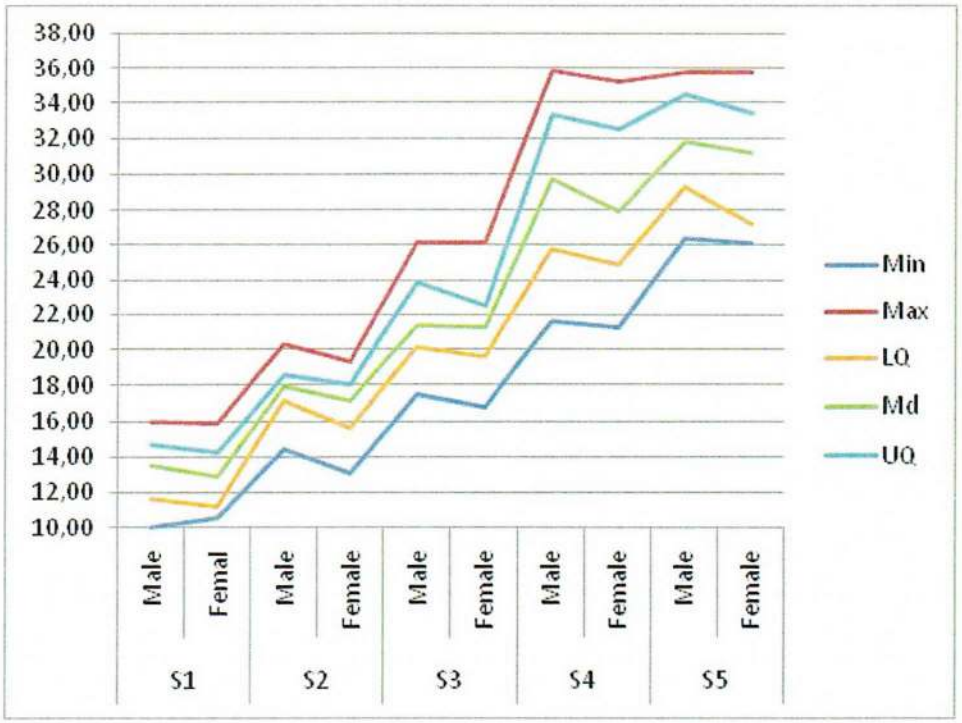

Abbildung 4: Altersgrenzen nach Geschlecht und Stadien (Kellinghaus et al. 2010a aus Tabelle 2, S. 152

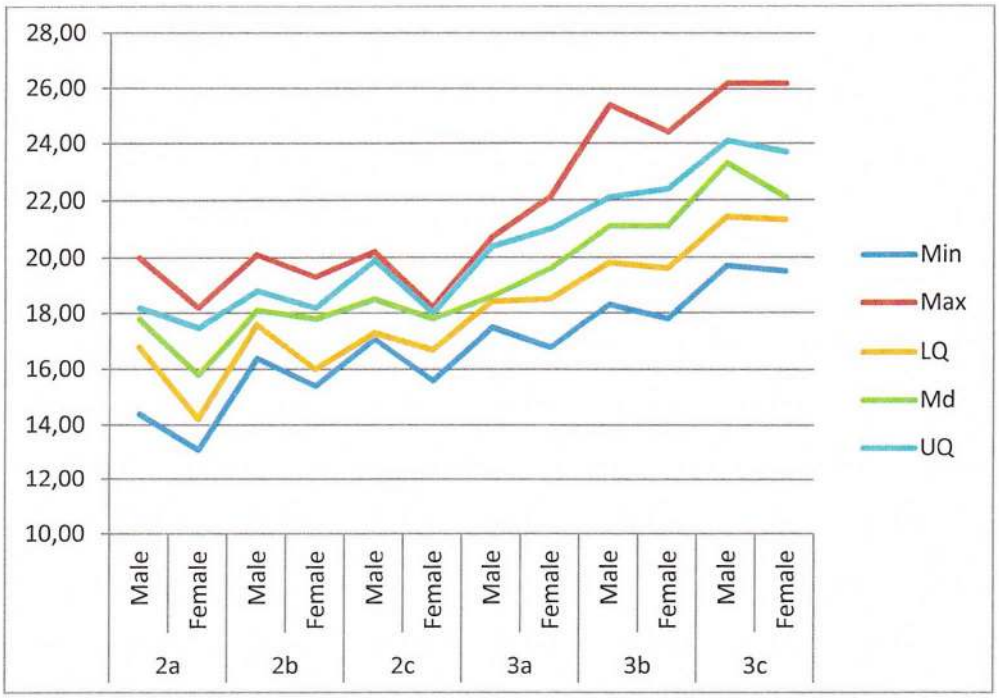

Abbildung 5: Altersgrenzen nach Geschlecht und Stadien (Kellinghaus et al. 2010b aus Tabelle 2, S. 323

5. **Angaben zur Referenzpopulation:** In Kellinghaus et al. 2010a und 2010b wird bezüglich der Referenzpopulation (gemeinsam für beide Publikationen) nur berichtet, dass es sich um eine Stichprobe von PatientInnen eines Unfallkrankenhauses handelt. Die Verteilung von Geschlecht, Alter und Gesundheitszustand ist bekannt (letztere jedoch nur punktuell), es fehlen aber Informationen bezüglich genetisch-geographischer Herkunft und sozioökonomischem Status, was laut AGFAD Richtlinien, S. 4 Punkt 8 „Angabe zur Referenzpopulation hinsichtlich genetisch-geographischer Herkunft, soziökonomischem Status, Gesundheitszustand" gefordert wird. Insbesondere handelt es sich um keine Zufallsstichprobe aus einem entsprechenden Stichprobenrahmen. Die Argumente, warum sich die Ergebnisse auch auf eine völlig anders geartete Zielpopulation übertragen lassen sollen, beschränken sich auf den Hinweis, dass keine ethnischen Unterschiede bekannt sind und dass die Knochenbildung benachteiligter Gruppen sich noch langsamer entwickeln würde. Folglich sollte die zu begutachtende Gruppe hinsichtlich der Diagnose eher im Vorteil sein. Ein empirischer Beleg für die Übertragbarkeit der Resultate fehlt jedoch zur Gänze. Ob die Population der KlientInnen eines Unfallkrankenhauses bezüglich Knochenbau, Knochenwachstum, Bruchwahrscheinlichkeit etc. mit der für den Diagnoseprozess relevanten Zielpopulation überhaupt vergleichbar ist, bleibt letztlich Spekulation. Insbesondere wird weder theoretisch noch empirisch darauf eingegangen, ob es zumindest in Einzelfällen zu einer Beschleunigung der Knochenbildung und damit zu einer nachteiligen Fehldiagnose für die zu diagnostizierende Person kommen kann.

In beiden Publikationen (2010a und 2010b) fehlen Angaben bezüglich genetisch-geographischer Herkunft und sozioökonomischem Status, was laut AGFAD Richtlinien, S. 4 Punkt 8 *„Angabe zur Referenzpopulation hinsichtlich genetisch-geographischer Herkunft, soziökonomischem Status, Gesundheitszustand"* gefordert wird. Ein empirischer Nachweis der Übertragbarkeit der Ergebnisse einer kaukasischen Stichprobe auf andere Populationen fehlt.

6. **Einfluss der Methodik:** In Kellinghaus et al 2010a wird argumentiert, dass die ‚CT slice thickness' einen Einfluss auf die Bestimmung des Ossifikationsstadiums hat (S. 150, 1. Spalte, oben). Das verwendete Material enthält aber wiederum vier verschiedene Dicken (0.6:3; 1.0:301; 1.25: 122 und 1.5:77). In den

vorliegenden Publikationen wird ein Einfluss des Materials aber nicht weiter behandelt, was laut AGFAD Richtlinie (S. 4, Punkt 7), *„genaue Beschreibung der Methodik"*, jedoch wünschenswert gewesen wäre.

Fazit

Der aus den Studien entnehmbare Beleg für eine Zulässigkeit der in den Referenzstudien vorgeschlagenen Methodik besteht im entscheidenden Kern darin, dass in der zugrunde liegenden Stichprobe kein Fall beobachtet wurde, in welcher eine unter 18-jährige Person gemäß ihres Ossifikationsstadiums als über 18 eingestuft worden wäre.

Im gegebenen Zusammenhang müssen die uns vorliegenden Artikel aus zweierlei Blickwinkeln diskutiert werden, nämlich zum Einen die Erfüllung wissenschaftlicher Standards aus akademischer Sicht, zum Anderen die Übertragbarkeit in die Praxis beim jetzigen Stand der Erkenntnis. Die in den Publikationen offen gelassenen Fragen könnten sich hier in der konkreten Anwendung für Betroffene drastisch auswirken, da in entsprechenden Altersgutachten direkt auf Referenzwerte aus diesen Publikationen Bezug genommen wird – insbesondere wenn es doch unter 18-Jährige in Stadium 3c geben sollte, diese in der Stichprobe nur nicht abgebildet wurden. Die in den Publikationen angeführten Minima finden sich nämlich tatsächlich 1:1 in dem von UNHCR vorgelegten anonymisierten Gutachten vom 09.08.2012 mit dem Verweis: „Ein Stadium 3b (Fusion v. Epi- & Metaphyse 1/3-2/3 der Wachstumsfuge) tritt dem aktuellen Forschungsstand folgend bei einer männlichen Klientel ab einem Mindestalter von 18.3 Jahren und im Median bei 21.1 Jahren auf, ein Stadium 3c (Fusion v. Epi- & Metaphyse >2/3 der Wachstumsfuge) ab einem Mindestalter von 19.7 Lebensjahren und im Median bei 23.3 Jahren."

Dementsprechend geht es bei der Bewertung der Texte auch darum, ob sie hinreichend wissenschaftliche Erkenntnisse liefern können, die in Hinblick auf die Zielpopulation verallgemeinerbar sind. Dies wird üblicherweise von einer einzelnen Studie (zur Erinnerung: es handelt sich hier um zwei Publikationen zu einer Studie) nicht verlangt und kann auch gar nicht verlangten werden. Normalerweise wird vor einer Akzeptanz wissenschaftlicher Ergebnisse insbesondere deren Replikation durch unabhängige WissenschaftlerInnen erwartet.

Zusammenfassend sind folgende Aspekte der Referenzstudien im Hinblick auf die AGFAD Standards und als Grundlage für Gutachten zur Altersdiagnose als problematisch zu klassifizieren:

- Die Stichprobe ist nicht repräsentativ für die Zielpopulation (und wegen des Samples aus einem Unfallkrankenhaus auch keine Zufallsauswahl aus einer „heimischen" Bevölkerung). Die Verallgemeinerbarkeit der Ergebnisse kann nur vermutet werden, ein empirischer Beleg dafür fehlt.
- Die Stichprobengrößen sind bei weitem zu klein, um die Neigung zu Fehldiagnosen für Subpopulationen ausschließen oder überhaupt beurteilen zu können. Selbst unter Annahme perfekter Repräsentativität für die Zielgruppe ergibt eine Näherungsrechnung der Gutachter, dass die Beobachtungen durchaus noch mit einer Fehlerrate von 5% im kritischen Alterssegment verträglich wären (Clopper-Pearson-Intervall für p einer Beobachtung von 0 aus 51 17-18-Jährigen). Die fehlenden Gruppengrößen nach Stadien und Geschlecht aus Kellinghaus et al. 2010a (teilweise) und in Kellinghaus et al. 2010b (vollständig) konnten auch durch eine Nachfrage per E-Mail an den korrespondierenden Autor nicht eruiert werden – das E-Mail von 02.08.2013 wurde bis dato nicht beantwortet.
- Die in klinischen Studien selbstverständliche Auskunft zur Verblindung der beteiligten RaterInnen fehlt in Kellinghaus et al 2010b.
- Derzeit gibt es außerdem noch keine Replikation der Studie in vergleichbarem Setting. Folglich beruht die Evidenz für das vorgeschlagene Verfahren auf einer einzigen Untersuchung. Replikationen von Forschungsergebnissen durch die *scientific community* sind aber wissenschaftliche Selbstverständlichkeit.

Dies unterstreichen die Ergebnisse einer australischen Untersuchung von Bassed et al. (2011)[7], welche an Verstorbenen durchgeführt wurde, die nämlich durchaus bei 17-jährigen Männern zu 6.1% Stadium 4 und zu 3% sogar Stadium 5 findet! Außerdem gibt diese Studie eine Inter-Raterreliabilität an, die mit 0.734 eher an der Untergrenze der Empfehlungen rangiert, was die Frage aufwirft, wie eindeutig diese Stadien überhaupt zuordenbar sind. Eine empirisch gesicherte Aufklärung dieser Diskrepanzen und unabhängige Replikation ist zu fordern, um wissenschaftlichen Standards zu genügen.

Empfehlungen

1. Replikationsstudie (am besten multizentrische Studie): Die Gültigkeit bzw. Verallgemeinerbarkeit der Resultate über das Verfahren aus den beiden Referenzpublikationen (ein Datensatz!) ist von unabhängiger Seite zu prüfen. Des Weiteren muss untersucht werden, ob sich das Verfahren auch in praxisnahem Umfeld mit geschulten RaterInnen bewährt. Bei der Überprüfung ist insbesondere darauf zu achten, dass die zur Beobachtung herangezogenen Fälle einer relevanten Zielpopulation möglichst repräsentativ und mit hinreichenden Fallzahlen nachgebildet sind (mindestens unter Berücksichtigung von Alter, Geschlecht, sozialer Schicht und Herkunftsregion).
2. Angaben zu Spezifität und Sensitivität der Stadien-/Altersbestimmungen: Die Dokumentation der vorliegenden Studien sollte um die relevanten Angaben zu den Fallzahlen ergänzt werden. Für eine vollständige Beurteilung der Eigenschaften des Verfahrens in beide Richtungen (also sowohl in Hinblick auf Fehldiagnosen als zu alt und zu jung[8]) sollten sogenannte ROC-Kurven angegeben werden, welche die Fehldiagnosen in beide Richtungen einander gegenüberstellen.
3. Zuverlässigere Angaben zu den Minima bzw. zu deren Schwankungsbreiten: Für diese Angaben sind weitere Daten erforderlich. Gäbe es mehr Studien dieser Art, könnten die Minima verlässlicher eingrenzt werden.

Wien, am 30.09.2013

Priv. Doz. MMag. Dr. Ivo Ponocny
Department for Applied Statistics and Economics
MODUL Private University Vienna

Mag.Dr. Elisabeth Ponocny-Seliger
Coaching, Empirische Sozialforschung &
Gender-Research

7 Bassed, R. B., Drummer, O. H., Briggs, C., & Valenzuela, A. (2011). Age estimation and the medial clavicular epiphysis: analysis of the age of majority in an Australian population using computed tomography. Forensic Science, Medicine, and Pathology, 7(2), 148-154.

8 Aus den Tabellen geht hervor, dass viele Personen über 18 noch nicht notwendigerweise in 3c sind. Bei zu niedriger Sensitivität (wenn also in vielen Fällen unter 18-Jährige als solche gar nicht erkannt werden) könnte sich aber die Frage nach Kosten-Nutzen-Abwägungen verschärft stellen.