Aus der Klinik und Poliklinik für Kieferorthopädie

der Universität Würzburg

Direktorin: Professor Dr. med. dent. Angelika Stellzig-Eisenhauer

„Künstliche Intelligenz zur vollständig automatisierten FRS-Auswertung: Bewertung der Auswertequalität verschiedener kommerzieller Anbieter im Vergleich zu einem menschlichen Goldstandard“

Inauguraldissertation

zur Erlangung der Doktorwürde der

Medizinischen Fakultät

der

Julius-Maximilians-Universität Würzburg

vorgelegt von

Lisa Marie Widmaier (geb. Wirth)

aus Neustadt bei Coburg

Würzburg, August 2023

Referent: Priv.-Doz. Dr. med. dent. Felix Kunz

Koreferent: Univ.-Prof. Dr. med. dent. Gabriel Krastl

Dekan: Univ.-Prof. Dr. med. Matthias Frosch

Tag der mündlichen Prüfung: 26.02.2024

Die Promovendin ist Zahnärztin.

Für meinen Ehemann und meine Eltern –
für alles, was sie für mich getan haben

Inhaltsverzeichnis

Diese wissenschaftliche Arbeit verzichtet aus Gründen der besseren Lesbarkeit bewusst auf die Verwendung genderspezifischer Sprache. Alle männlichen Formen beziehen sich gleichermaßen auf weiblich, divers, etc.

1 Einleitung

1.1 Fernröntgenseitenbild-Analyse (FRS-Analyse)

1.1.1 Historischer Überblick

„Könnte man doch das Gebiss im Schädel wie hinter Glas liegen sehen!“ [1]. Diesen durch den Kieferorthopäden A. M. Schwarz sehr treffend formulierten Wunsch teilten seinerzeit wahrscheinlich viele Kieferorthopäden, denn hierdurch hätten nicht nur klinisch sofort sichtbare Zahnfehlstellungen diagnostiziert, sondern zur besseren Beurteilung von Anomalien auch Informationen über den Einbau der Kiefer im Gesichtsschädel gewonnen werden können [2].

Den ersten großen Meilenstein auf dem Weg zur Erfüllung dieses Wunsches legte W. C. Röntgen 1895 in Würzburg mit der Entdeckung der Röntgenstrahlen [3]. 1922 fertigte A. J. Pacini die erste laterale Röntgenaufnahme eines menschlichen Schädels an [4]. Eine kieferorthopädische Auswertung und Interpretation der Röntgenbilder fehlten zu diesem Zeitpunkt jedoch noch.

Da die neuartige Röntgentechnik lange keinen Einzug in die Kieferorthopädie gehalten hatte, behalf man sich zunächst durch das 1922 von P. Simon entwickelte Gnathostatverfahren. Hierbei konnten anhand von wenigen anatomischen Punkten wie Porion und Orbitale und eigens angefertigten Gipsmodellen Aussagen über die schädelbezügliche Lage dentaler Areale getroffen werden [2, 5]. Dieses Verfahren diente als Grundlage für die spätere Entwicklung der Röntgenkephalometrie [2].

Im Jahre 1931 veröffentlichten schließlich der amerikanische Kieferorthopäde B. H. Broadbent und der deutsche Zahnarzt H. Hofrath unabhängig voneinander ihre Forschungen zur Fernröntgenseiten-Technik und definierten bereits einige charakteristische Landmarken, die zur Auswertung der Röntgenbilder verwendet werden konnten [6, 7]. Die gleichzeitige Darstellung des weichgewebigen Profils und der skelettalen Anteile des Gesichtsschädels revolutionierte die Kieferorthopädie maßgeblich.

Auf Grundlage der nun verfügbaren Fernröntgenseitenbilder (FRS) eröffnete sich eine neue Möglichkeit zur kieferorthopädischen Diagnostik und Behandlungsplanung: die Kephalometrie. Diese besteht sowohl aus dem Setzen röntgenologischer Punkte (skelettal, dental oder weichgewebig; anatomisch, röntgenologisch oder konstruiert), als

auch aus der Messung geometrischer Beziehungen zwischen diesen Punkten, beispielsweise Streckenlängen oder Winkel [8, 9]. In einem weiteren Schritt werden dann die patientenspezifischen Werte mit Durchschnittswerten verglichen, um Aussagen über die Art und das Ausmaß der Dysgnathie des Patienten treffen zu können. Generell ist anzumerken, dass viele kephalometrische Normwerte von Faktoren wie Alter, Geschlecht oder Ethnie des Patienten abhängig sind [2]. Im Jahre 1948 wurden durch W. Downs erstmals anhand von FRS-Auswertungen von 20 Kindern mit eugnather Okklusion Normwerte für die Kephalometrie beschrieben [10]. Die FRS-Analyse unterliegt seither einer stetigen Weiterentwicklung, sodass nach Nötzel et al. aktuell etwa 200 Messpunkte und circa 100 unterschiedliche FRS-Analysen verfügbar sind [11]. Zu nennen sind exemplarisch die Analysen nach C. C. Steiner, A. Björk, J. Jarabak oder T. Rakosi [12-15].

1.1.2 FRS-Analyse heute und Bedeutung in der Kieferorthopädie

Die FRS-Analyse ist bis heute eine der wichtigsten Komponenten der kieferorthopädischen Diagnostik und Behandlungsplanung [2]. Sie liefert wichtige Erkenntnisse über zugrunde liegende skelettale Gegebenheiten, wie beispielweise den Prognathiegrad des Ober- und Unterkiefers sowie deren Lagebeziehung zueinander, das Wachstumsmuster bzw. den Gesichtsschädelaufbau, die Achsenstellung und Position der Frontzähne und die Weichteilmorphologie des Patienten [9].

Die ursprünglich mittels analoger Röntgentechnik erstellten FRS wurden zunächst händisch durch den Kieferorthopäden ausgewertet. Hierzu wurde eine Acetat-Tracing-Folie über das analoge Röntgenbild geklebt und dieses auf einen Leuchtkasten gelegt. Mittels eines Bleistifts und Geodreiecks konnten dann die Punkte und Konturen anatomischer Strukturen durchgezeichnet und geometrisch vermessen werden. Dieses manuelle Vorgehen ist sehr zeitaufwendig und dauert, je nach Analyse und Erfahrung des Auswerters, durchschnittlich etwa 20 Minuten pro FRS-Auswertung [16].

Heute ist insbesondere bei digitaler Aufnahmetechnik lediglich das Setzen der Punkte selbst ein manueller Prozess, da die geometrische Vermessung der Winkel und Strecken durch spezialisierte Softwarelösungen vereinfacht werden kann [17].

1.1.3 Vollständig automatisierte FRS-Analyse

Das manuelle Setzen der Landmarken ist ein Prozess, dessen Qualität stark von der Erfahrung des Untersuchers abhängt und auch zwischen verschiedenen Untersuchern treten regelmäßig klinisch relevante Abweichungen auf [18-20]. Durch ungenau gesetzte

Punkte können Fehler in der geometrischen Auswertung und Vermessung sowie schlussendlich in der Interpretation der Auswertungen resultieren, sodass hierdurch auch Konsequenzen in der Behandlungsplanung entstehen können [8].

Vor diesem Hintergrund wurde in den letzten Jahrzehnten versucht, den noch manuellen Prozess der Landmarkendetektion ebenfalls zu automatisieren und somit zu objektivieren, um schlussendlich eine vollständig automatisierte Auswertung von FRS zu ermöglichen. Bereits 1986 untersuchten Lévy-Mandel et al. die Anwendung von Bildverarbeitungsprogrammen zur automatisierten Punktsetzung im FRS. Die FRS wurden mittels eines Medianfilters zur Verstärkung des Kontrasts und zur Reduktion des Rauschens vorbereitet und mit einem Kantendetektor vorverarbeitet, um anschließend die Landmarken durch einen Linienverfolgungsalgorithmus, der auf Basis menschlichen Wissens programmiert wurde, zu lokalisieren [21]. 1994 wandten Cardillo und Sid-Ahmed die Methode des Vorlagenabgleichs und morphologische Graustufenoperatoren an, um die Landmarken im FRS softwarebasiert platzieren zu lassen. Von den Autoren wurde beschrieben, dass 60-85% aller automatisiert gesetzten Landmarken in einem Radius von 2mm zur idealen Punktposition lagen [22]. Die Entwicklung von Bildverarbeitungsprogrammen zur automatisierten FRS-Analyse unterlag einer stetigen Weiterentwicklung [23-25], bis El-Feghi et al. 2004 erstmals den Bereich der künstlichen Intelligenz (KI) unter Anwendung des Machine Learnings für die FRS-Analyse beleuchteten [16]. 2017 beschrieben Arik et al. die erste Anwendung von Deep Convolutional Neural Networks, die bis heute zur automatisierten FRS-Analyse durch KI-Algorithmen verwendet werden [8].

1.2 Künstliche Intelligenz (KI) im Gesundheitswesen

1.2.1 Allgemeine Informationen

Unter dem Begriff „Künstliche Intelligenz“ versteht man ganz allgemein Computersysteme, die menschliche Entscheidungen nachahmen und dadurch Aufgaben erfüllen, die ursprünglich nur mit Hilfe menschlicher Intelligenz zu bewältigen waren [26]. Anwendungen von KI begegnen uns mittlerweile in vielen Bereichen des Alltags, beispielsweise in Sprachassistenten, bei Navigationssoftwares, Smart-Home-Steuerungen oder bei Programmen zum Übersetzen komplexer zusammenhängender Texte.

Die Begrifflichkeit der „Künstlichen Intelligenz“ geht auf J. McCarthy (erstmals gebraucht auf einer Konferenz in Hanover, USA im Jahre 1956) zurück [27], wobei A. Turing als Pionier der KI-Entstehung gilt [28]. Er brachte bereits 1947 in Manchester den Gedanken des selbstständigen Denkens von Maschinen ins Spiel und war durch die Entwicklung des Turing-Tests maßgeblich an der Entstehung von KI beteiligt [27]. Dieser Test prüft, ob die Intelligenz eines Computers mit der eines Menschen vergleichbar ist [29].

Grundlage der modernen KI-Entwicklung ist das sogenannte „Machine Learning“: anhand von Trainingsdaten können Computer „lernen“, das heißt, sie erkennen mit Hilfe von sich selbst anpassenden Algorithmen Gesetzmäßigkeiten und generieren auf dieser Basis mathematische Muster, anhand derer sie bestimmte Entscheidungen in äquivalenten Situationen vorhersagen können. Die Besonderheit ist dabei die Selbstständigkeit: der Computer erfasst allgemeine Muster und kann neue Eingabedaten mit seinem erlernten Wissen bearbeiten, ohne explizit auf diese Eingabedaten programmiert worden zu sein [30].

1.2.2 Künstliche neuronale Netzwerke

Die Architektur von KI-Algorithmen ist an den von Hubel und Wiesel im Jahre 1961 untersuchten Aufbau des primären visuellen Kortex von Wirbeltieren angelehnt [30, 31]. So sind die künstlichen Neuronen, ähnlich wie die Nervenzellen im menschlichen Gehirn, in komplexen Netzen mit einer variablen Anzahl an unterschiedlichen Schichten, sogenannten „Layers“, angeordnet [32]. Bei den künstlichen Netzen unterscheidet man die erste Schicht, den „Input Layer“, die letzte Schicht, den „Output Layer“, und die dazwischenliegenden versteckten oder verborgenen Schichten, die sogenannten „Hidden Layers“. Je mehr Layer ein solches System aufweist, desto komplexere Aufgaben können bewältigt werden und desto unempfindlicher wird die KI gegenüber irrelevanten Nebeninformationen [33]. Die Anzahl der versteckten Schichten wird auch als „Tiefe“ der KI beschrieben - ab einer gewissen Anzahl von Layers spricht man folglich von „Deep Neural Networks“ [30]. Grundsätzlich gibt es verschiedene Unterformen von Deep Neural Networks. Zu den Bekanntesten gehören die „Convolutional Neural Networks“ (CNN) und die „Recurrent Neural Networks“ (RNN). RNN werden hauptsächlich für Spracherkennung und -analyse genutzt, wohingegen CNN ihre Hauptanwendung im Bereich der Bilderkennung und Analyse von Bildinhalten finden [30].

1.2.3 Training einer KI

Bevor eine KI einsatzbereit ist, muss sie anhand einer Vielzahl von Beispielen trainiert werden. Dieses Lernen findet meist durch den Menschen überwacht statt, indem eine große Anzahl von Beispieldatensätzen mit einer entsprechenden Musterlösung zur Verfügung gestellt wird. Diese Form des Trainings wird auch als „Supervised Learning“ bezeichnet. Die KI versucht, ein Grundmuster in den Beispieldatensätzen zu erkennen, anhand dessen sich die vorgegebenen Musterlösungen aus den Eingabedaten herleiten lassen. Im Gegensatz dazu werden beim sogenannten „Unsupervised Learning“ Trainingsdatensätze ohne entsprechende Musterlösung zur Verfügung gestellt – die KI versucht dann, Grundmuster in den Beispieldaten zu erkennen, die sich vom allgemeinen Grundrauschen unterscheiden [34]. Insbesondere bei komplexen Fragestellungen ist meist ein überwachtes Lernen der KI erforderlich.

Im Folgenden wird der überwachte Lernvorgang am Beispiel einer KI zur Analyse von Bildinhalten beschrieben. Für das Training der KI wird eine Vielzahl möglichst heterogener Beispieldatensätze mit verschiedensten Bildinhalten verwendet. Die Gesamtheit aller Beispieldatensätze wird in Trainings- und Validierungsbilder unterteilt. Die Generierung großer Mengen an Beispielbilddaten ist schwierig und sehr zeitaufwendig. Um die Datenmenge weiter zu vergrößern, werden die Trainingsbilder oftmals gespiegelt, gedreht oder parallel verschoben, sodass hierdurch zusätzliche künstliche Trainingsdatensätze generiert werden. Dieser Vorgang wird als „Augmentation“ bezeichnet [32].

Zunächst werden die Trainingsbilder in den Input-Layer eingespeist und durchlaufen die verschiedenen Hidden Layers des Algorithmus. Innerhalb dieser passt sich die KI durch sogenannte variable Filter solange an, bis die im Output-Layer erzeugten Ausgaben mit den zu den Trainingsbildern gehörigen vorgegebenen Ergebnissen so gut wie möglich übereinstimmen und die KI die Trainingsbilder somit richtig auswerten kann [32]. Ab einer gewissen Anzahl von Layers spricht man vom „Deep Learning“ [30]. Zu Beginn des Trainings werden von der KI noch zufällige Bildmerkmale ausgegeben, jedoch adaptiert der Algorithmus sich zunehmend, sodass sich die Ergebnisse mit fortschreitendem Training verbessern. Allerdings besteht die Gefahr, dass der KI-Algorithmus irgendwann zur Verbesserung der Ergebnisse nicht mehr das allgemeine Muster erkennt, sondern beginnt, die Trainingsdaten auswendig zu lernen. Dieser Prozess wird als Overfitting bezeichnet [32].

Um dies zu vermeiden, werden die Validierungsbilder benötigt. Diese führen im Gegensatz zu den Trainingsbildern nicht zu einer weiteren Adaption des Algorithmus, sondern sind lediglich zum Vergleich mit den Trainingsbildern gedacht. Irgendwann wird die Genauigkeit der Auswertung innerhalb der Validierungsbilder stagnieren bzw. sich sogar verschlechtern, während die Genauigkeit der Auswertung innerhalb der Trainingsbilder (durch das Auswendiglernen) immer besser wird. Problematisch ist, dass die Ergebnisse der KI bei neuen Daten nun ungenauer und schlechter werden, da auch das Rauschen und die zufälligen Schwankungen innerhalb der Trainingsdaten erlernt wurden [33]. Die KI verliert somit durch das Auswendiglernen der Trainingsbilder bei Einsetzen des Overfittings die Möglichkeit zur Verallgemeinerung. Deshalb sollte das Training der KI zum Zeitpunkt des geringsten Fehlers innerhalb der Validierungsbilder beendet werden [32].

1.2.4 KI in Deutschland

Den hohen Stellenwert der KI-Entwicklung in Deutschland und deren Etablierung in der Medizin verdeutlicht die große Anzahl an Forschungsprojekten und Förderprogrammen vieler national und international tätiger Organisationen. 2006 wurde beispielsweise das Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS) ins Leben gerufen, das sich mit der Entwicklung von KI für unter anderem die Landwirtschaft, den Verkehr und auch die Medizin beschäftigt [35]. Ein aktuelles Forschungsprogramm des IAIS ist „LOTTE“ (Leitsystem zur Optimierung der Therapie traumatisierter Patient*innen bei der Erstbehandlung). Hierbei handelt es sich um eine KI, die die schnelle Entscheidungsfindung bei der Versorgung von Polytrauma-Patienten unterstützen soll. Insbesondere bei Erstaufnahme des schwerverletzten Patienten im Schockraum müssen schnell enorme Datenmengen, wie Unfallhergang, Vorerkrankungen, Dauermedikation, aktuelle Vitalparameter und vor allem Ergebnisse der aktuellen Bildgebungen analysiert und aufeinander abgestimmt werden. Auf dieser Grundlage müssen kurzfristig viele wichtige Entscheidung zur Priorisierung der verschiedenen Schritte im Rahmen der Notfallversorgung des Patienten getroffen werden, welche für den Patienten überlebenswichtig sind [36]. Laut Kleber et al. versterben in Deutschland aktuell noch 14% aller Polytrauma-Patienten im Krankenhaus [37]. Das Forschungsprojekt „LOTTE“ hat zum Ziel, den Menschen in der Entscheidungsfindung zu unterstützen und damit die Überlebenschancen des Patienten in der notfallmäßigen Frühversorgung zu verbessern [36].

Neben dem IAIS hat auch das Max-Planck-Institut für intelligente Systeme seinen Schwerpunkt in der Entwicklung und Anwendung künstlicher Intelligenz gesetzt. 2020 wurde unter anderem in Zusammenarbeit mit der Universität Tübingen, dem Deutschen Krebsforschungszentrum, dem European Molecular Biology Laboratory (EMBL) und der Universität Heidelberg eine neue Initiative zur Etablierung von KI und Robotik in der Medizin gestartet [38].

Darüber hinaus stellen Organisationen wie die Deutsche Forschungsgemeinschaft (DFG), das Bundesministerium für Bildung und Forschung (BMBF) und der KI Bundesverband finanzielle Fördermittel für Forschungsprojekte und Entwicklungen von KI sowie Möglichkeiten zum Austausch zwischen Forschung und Wirtschaft bzw. Politik zur Verfügung [39-41].

1.2.5 KI in der Humanmedizin

Auch in der Medizin ist ein rasanter Fortschritt in der Entwicklung von KI zu verzeichnen. Wichtige Anwendungsmöglichkeiten von KI in der Humanmedizin sind beispielsweise die Vermeidung unerwarteter Hypoxämien während Narkosen [42] oder die Selektion von überlebensfähigen Embryonen bei In-vitro-Fertilisationen [43].

Der überwiegende Anteil von KI in der Humanmedizin hat allerdings zum Ziel, Ärzte in der Auswertung von Bildgebungen im Rahmen der Röntgentechnik oder Magnetresonanztomographie (MRT) zu unterstützen [44]. Dieser wichtige Bereich der Medizin, in dem KI erfolgreich eingesetzt wird, wurde unter dem Namen „Radiomics“ bekannt. Bei dem Begriff Radiomics handelt sich um eine Wortneuschöpfung, die den Begriff „radiology“ und die Endung „-omics“ kombiniert [45]. Die Endung „-omics“ beschreibt dabei die Erhebung einer großen Menge von zunächst unspezifischen Parametern, die aus einer einheitlichen Datenquelle stammen [46]. Somit befasst sich das Gebiet Radiomics mit der Erkennung und klinischen Interpretation von morphologischen Auffälligkeiten in radiologischen Bilddateien [45].

Für die genaue Beschreibung und Charakterisierung von Krankheiten und der damit verbundenen Diagnosestellung, Prognose und Therapieplanung ist die Untersuchung der Läsion (beispielweise des Tumors) auf spezielle Biomarker erforderlich. Aus Röntgenbildern können moderne KI-Algorithmen riesige Mengen an Merkmalen extrahieren und dadurch die Datenbank von verfügbaren Biomarkern wesentlich vergrößern. Idealerweise kann durch zunehmende Etablierung von Radiomics in Zukunft möglicherweise auf weiterführende Untersuchungen wie Biopsien der Läsion verzichtet

werden [46]. Die Besonderheit von Radiomics ist nicht nur die Fähigkeit, extrem große Datenmengen zu verarbeiten, sondern auch die hohe Genauigkeit der Auswertungen: diskrete Strukturunterschiede im Röntgenbild sind für das menschliche Auge oft nicht sichtbar, jedoch in gewissen Situationen durch eine KI sicher diagnostizierbar [30]. Auf Grundlage der im Röntgenbild erkennbaren Strukturunterschiede sind moderne KI-Algorithmen in der Lage, Läsionen, wie beispielsweise Tumoren, zu identifizieren [30]. Dabei kann KI aber nicht nur die Gewebeveränderung erkennen, sondern auch Aussagen über den Subtyp des Tumors, die Prognose und Lebenserwartung, den Therapieerfolg und das Risiko eines Tumorrezidivs treffen [47-53]. Dies soll am folgenden Beispiel verdeutlicht werden.

2020 betrafen 11,7% aller Krebserkrankungen die weibliche Mamma. Damit stellt das Mamma-Karzinom die häufigste Krebsart dar [54]. In der Brustkrebsdiagnostik wird KI in den letzten Jahren vermehrt eingesetzt. Laut Becker et al. erreichen KI-Algorithmen bei der Diagnose von Mamma-Karzinomen im Rahmen der Mammographie bereits eine hohe Genauigkeit (ROCAUC = 0,82), sodass kein signifikanter Unterschied zu erfahrenen Radiologen (0,79 ≤ ROCAUC ≤ 0,87) besteht [55]. Bestimmte Bild- und damit Tumoreigenschaften wie beispielweise Grauwert, Konkavität oder Rundheit der Läsion korrelieren signifikant mit bestimmten molekularen Subtypen des Mamma-Karzinoms, sodass eine noch feinere Diagnostik des genauen Tumortyps möglich ist [56]. Allerdings wird KI nicht nur bei der Diagnosestellung, sondern auch in der anschließenden Therapieplanung eingesetzt. Aktuelle Forschungsprojekte beschäftigen sich mit der Entwicklung von KI, die die Radiatio automatisieren und bezüglich Körperform, des Schutzes angrenzender Gewebe und Reduktion der Strahlendosis individualisieren können [57]. KI-Algorithmen sind außerdem in der Lage, Aussagen über die Prognose und das Risiko der Mortalität zu treffen [58]. Anhand von Bildgebungen wie MRT und der damit verbundenen Bestimmung von Größe, Randmorphologie, Form und Anreicherungsmuster des Tumors können spezialisierte KI-Algorithmen das Risiko eines Tumorrezidivs erfolgsversprechend vorhersagen [53].

Neben Bildgebungen der Mammographie können auch Thorax-Röntgenaufnahmen durch KI auf vielfältige Art und Weise ausgewertet werden. In einer Studie von Cicero et al. wurde eine KI darauf trainiert, Kardiomegalien, Pleuraergüsse, Lungenödeme, Konsolidierungen und einen Pneumothorax zu erkennen. Die KI detektierte sogar kleinste Veränderungen, die klinisch leicht zu übersehen sind [59]. Thorax-Röntgenaufnahmen können auch zur Diagnose von Lungen-Karzinomen genutzt

werden. Laut Sung et al. galt Lungenkrebs 2020 mit 11,4% aller Karzinome als zweithäufigste Krebsart [54]. Entscheidend für das Überleben des Patienten ist eine möglichst frühe Diagnosestellung: bei frühzeitiger Entdeckung des Karzinoms erhöht sich die 5-Jahres-Überlebensrate um circa 50% [60]. Aus diesem Grund sollten Lungenknoten im Frühstadium genau untersucht und engmaschig überwacht werden. Die Computertomographie (CT) ist hierbei das bildgebende Mittel der Wahl. Anhand von Form, Struktur und vor allem Wachstumsrate der Knoten können KI-Algorithmen mit einer Genauigkeit von 84-90% zwischen benignen und malignen Knoten unterscheiden [60, 61]. Darüber hinaus können KI-basiert die Wahrscheinlichkeiten für eine Metastasenbildung, das Ansprechen auf eine Immuntherapie und sogar die Überlebenswahrscheinlichkeit des Patienten abgeschätzt werden [62-64].

Ein weiteres Einsatzgebiet von KI liegt in der Diagnostik von Alzheimer. Alzheimer gilt mit einer Prävalenz von etwa 5% in Europa als häufigste Form der Demenz und gewinnt mit einer zunehmend alternden Bevölkerung immer mehr an klinischer Bedeutung [65]. Mittels eines 18-FDG (Fluordesoxyglucose)-PET-CT des Gehirns können im Vergleich zu einer rein auf klinischen Befunden basierenden Diagnosestellung Alzheimer-Erkrankungen deutlich früher diagnostiziert werden, was einen entscheidenden zeitlichen Vorteil für die Therapie darstellt [66]. Das 18-FDG-PET-CT zeigt bei Alzheimer in bestimmten Bereichen des Gehirns (posteriorer Gyrus cinguli, parietotemporaler Kortex und Frontallappen) einen regionalen Hypometabolismus. Eine von Ding et al. untersuchte KI erreichte bezüglich der Früherkennung von Alzheimer eine Spezifität von 82% bei einer 100%igen Sensitivität und konnte die Diagnose sogar durchschnittlich 75,8 Monate vor der finalen Diagnose stellen [67]. Darüber hinaus sind KI in der Lage, den weiteren Verlauf der Erkrankung und sogar eine mögliche Entwicklung einer Demenz vorherzusagen [68].

1.2.6 KI in der Zahnmedizin

KI-Algorithmen finden mittlerweile auch in vielen Bereichen der Zahnmedizin Anwendung, dazu zählen allgemeine Zahnmedizin, Kariologie, Endodontie, Parodontologie, Kieferorthopädie und forensische Zahnmedizin [69].

Zhang et al. entwickelten eine KI, die in der Lage ist, das Ausmaß der fazialen Schwellung nach operativer Entfernung von impaktierten Weisheitszähnen vorherzusagen. Dazu wurden 15 Faktoren wie beispielsweise Alter und Geschlecht des Patienten, Anzahl der Wurzeln des zu extrahierenden Zahnes und chirurgisches

Vorgehen als Input festgelegt. Die KI konnte im Anschluss mit einer Genauigkeit von 98% eine leichte, mittlere oder starke Schwellung vorhersagen [70].

Vornehmlich befassen sich KI in der Zahnmedizin allerdings mit der Analyse von Röntgenbildern [71]. In den meisten europäischen Ländern entfällt der größte Anteil der erstellten Röntgenbilder auf die Zahnmedizin [69]. In einem Report des United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) aus dem Jahre 2010 werden im weltweiten Durchschnitt etwa 300 Röntgenaufnahmen pro 1000 Patienten pro Jahr dokumentiert [72]. Zu diesen zählen insbesondere Orthopantomogramme (OPGs), Bissflügel, Zahnfilme und FRS. Doch auch weitere bildgebende Verfahren, wie Fotos, 3D-Scans und Fluoreszenzbilder finden in der Zahnmedizin immer häufiger Anwendung und bieten damit viele Anwendungsmöglichkeiten für KI. Generell können KI-Algorithmen bei der Analyse von Bildmaterial verschiedene Aufgaben erfüllen: sie können charakteristische Strukturen wie Zähne oder Karies erkennen, Zähne segmentieren sowie klassifizieren [69].

In einer Studie von Lee et al. wurde eine KI (GoogLeNet Inception v3 CNN network) zur Detektion von Karies auf Zahnfilmen entwickelt. Sie zeigte eine Genauigkeit von 89% für Prämolaren und eine Genauigkeit von 88% für Molaren [73].

Nach Schwendicke et al. birgt der Einsatz von KI im Bereich der Parodontologie ebenfalls großes Potenzial: insbesondere eine Kombination von Röntgenbildern mit klinischen Daten wie Bleeding on Probing (BoP) oder Lockerungsgrad und anamnestischen Informationen wie Ernährung ist denkbar und kann neue Möglichkeiten zur individualisierten Therapieplanung und Prävention parodontaler Erkrankungen eröffnen [71]. Eine 2018 von Lee et al. entwickelte KI war bereits in der Lage, parodontal beeinträchtigte Zähne auf periapikalen Röntgenbildern zu erkennen und eine Prognose zum Erhalt der Zähne abzugeben. Für Prämolaren lag die Genauigkeit der Diagnose einer Parodontalerkrankung bei 81%, für Molaren bei 76,7%. Anhand einer Probe von 64 Prämolaren und 64 Molaren, die klinisch als stark parodontal geschädigt eingestuft wurden, lag die Genauigkeit der durch die KI vorhergesagten Extraktion bei 82,8%, für Molaren bei 73,4% im Vergleich zum menschlichen Goldstandard [74].

Auch im Bereich der forensischen Zahnmedizin finden KI-Algorithmen Anwendung [75]. Im Rahmen von Unfällen oder Verbrechen kann sich die Identifikation von Opfern schwierig gestalten, sodass auf zahnmedizinische Parameter zurückgegriffen werden muss. So eignet sich beispielsweise die Mandibula hervorragend für die Geschlechtsbestimmung eines Individuums. Patil et al. untersuchten 2020 eine

KI-basierte (Digimizer Image analysis software) Auswertung von OPGs hinsichtlich mandibulärer Parameter wie maximale Ramusbreite, kondyläre Höhe und Gonionwinkel. Alle untersuchten und durch die KI bestimmten Parameter zeigten im Medianwert statistisch signifikante Unterschiede zwischen Männern und Frauen. Gerade durch die Kombination mehrerer Faktoren ist die Geschlechtsbestimmung anhand von mandibulären Parametern somit sehr erfolgversprechend [76]. Andererseits bieten sich Röntgenaufnahmen von Weisheitszähnen zur Altersbestimmung an. De Tobel et al. entwickelten eine KI, die den Entwicklungsstand unterer Weisheitszähne anhand eines OPGs bestimmen und diesen mit dem Alter des Individuums korrelieren kann. Bei der Auswertung wurden zehn verschiedene Entwicklungsstadien unterschieden, eine durchschnittliche Abweichung der KI von 0,6 Entwicklungsstadien zum menschlichen Goldstandard wurde dokumentiert. Die Ergebnisse der KI waren somit mit den menschlichen Ergebnissen nahezu gleichzusetzen [77].

1.2.7 KI in der Kieferorthopädie

Auch im Fachgebiet der Kieferorthopädie gibt es inzwischen einige Ansätze, KI im klinischen Alltag zu integrieren. In der Literatur werden diesbezüglich verschiedene Anwendungsgebiete beschrieben.

1.2.7.1 KI zur Beurteilung des kieferorthopädischen Behandlungsbedarfs

Thanathornwong entwickelte 2018 eine KI, die anhand von Fotos und Kiefermodellen der Patienten verschiedene Parameter wie beispielsweise fehlende oder überzählige Zähne, die Bissstellung sowie den Overbite und Overjet analysieren kann und auf Grundlage dessen entscheidet, ob eine kieferorthopädische Behandlung erforderlich ist. Die untersuchten Parameter wurden in Anlehnung an allgemein gebräuchliche Indices zur Einstufung der kieferorthopädischen Behandlungsnotwendigkeit ausgewählt. Als Beispiel für solche Indices sind der Index of Orthodontic Treatment Need (IOTN) oder der Dental Aesthetic Index (DAI) zu nennen. Die Ergebnisse der KI wurden mit den Einschätzungen zweier Kieferorthopäden mit mehrjähriger Berufserfahrung verglichen und es konnte eine hohe Übereinstimmung gezeigt werden (Kappa-Wert im Vergleich zu Kieferorthopäden A: 1; Kappa-Wert im Vergleich zu Kieferorthopäden B: 0,894) [78].

1.2.7.2 KI in der KFO-Diagnostik

Das aktuell wohl am weitesten entwickelte Einsatzgebiet von KI in der kieferorthopädischen Diagnostik ist die FRS-Analyse. Wie bereits beschrieben, gehört

die Erstellung und Analyse eines FRS zur kieferorthopädischen Routinediagnostik [2]. Untersucher- und erfahrungsabhängig können allerdings signifikante Unterschiede bei der Lokalisation der Punkte auftreten, welche weitereichende klinische Konsequenzen haben können [8, 18-20]. Durch den Einsatz von KI wird der manuelle Prozess der Landmarkenidentifikation automatisiert und somit objektiviert.

Die Mehrheit der Studien zum Thema KI-gestützte FRS-Analyse untersucht die Genauigkeit von KI auf Grundlage der metrischen Abweichungen zwischen den automatisiert gesetzten Landmarken und einem menschlichen Goldstandard. In einer Meta-Analyse zum Thema Deep Learning zur automatisierten FRS-Analyse aus dem Jahre 2021 beschreiben Schwendicke et al., dass KI bereits eine hohe Genauigkeit beim Setzen der Landmarken erreichen. Konkret war der überwiegende Anteil der untersuchten KI in der Lage, die Punkte in einem Radius von bis zu 2mm zur idealen Punktposition zu setzen [79]. In der Literatur gelten Abweichungen von maximal 2mm als klinisch ausreichend genau, da diese meist innerhalb der ersten Standardabweichung liegen [8, 80-82].

Allerdings ist neben dem Betrag der Abweichung auch die Richtung der Abweichung von der idealen Punktposition entscheidend. Bei einer Winkelmessung würde eine Abweichung der Punktlokalisation entlang eines Winkelschenkels beispielsweise nicht zu einer Veränderung des gemessenen Parameters führen, selbst wenn eine Abweichung von mehr als 2mm zur idealen Punktposition vorläge [32]. Daher ist nach Santoro et al. nicht die Positionierung der einzelnen Landmarken, sondern die Genauigkeit der auf diesen Landmarken basierenden kieferorthopädischen Parameter für die spätere Therapieplanung entscheidend. Somit sollte auch die FRS-Analyse als Ganzes, nicht nur die Lokalisation der einzelnen Punkte, untersucht werden [83].

In einer Studie von Kunz et al. wurde eine KI anhand von 1792 FRS darauf trainiert, 18 Landmarken automatisch zu erkennen und auf dieser Basis eine vollständige FRS-Analyse zu generieren. Als Goldstandard wurde der Medianwert zwölf erfahrener Untersucher für alle Parameter definiert. Im Vergleich zu diesem Goldstandard wurden mit Ausnahme des SN_MeGo keine statistisch signifikanten Unterschiede festgestellt und die mittleren Differenzen zwischen der KI und menschlichem Goldstandard waren mit Abweichungen von maximal 0,37° bei Winkelmessungen und maximal 0,20mm bei metrischen Messungen nur sehr gering. Diese Abweichungen wurden von den Autoren als klinisch nicht relevant eingestuft [32].

Die Analyse des Restwachstums eines Patienten ist ein weiterer essenzieller Bestandteil der kieferorthopädischen Behandlungsplanung. Bei der Therapie einer skelettalen Klasse II steht die sagittale Wachstumsförderung des Unterkiefers im Vordergrund, welche allerdings nur bei ausreichendem Wachstumspotential der Mandibula erfolgversprechend ist [84]. Um beispielsweise zu beurteilen, ob ab einem bestimmten Alter des Patienten noch mit funktionskieferorthopädischen Geräten behandelt werden kann oder bereits auf eine festsitzende Klasse-II-Mechanik zurückgegriffen werden sollte, ist die Einschätzung des noch zu erwartenden Wachstumspotenzials erforderlich. Da die individuelle Wachstumsdynamik in der Adoleszentenphase großen Schwankungen unterliegt, ist hierfür die alleinige Berücksichtigung des chronologischen Alters nicht ausreichend [85, 86]. Vielmehr sollte zusätzlich das skelettale Alter des Patienten bedacht werden [87-89]. Hierfür eignen sich sowohl die radiologische Analyse der Handwurzelknochen als auch die Wirbelkörperanalyse unter Verwendung der CVM-Methode (Cervical Vertrebral Maturation) [84, 90, 91]. Zwischen diesen beiden Methoden lässt sich eine signifikante Korrelation feststellen [87, 92]. Der große Vorteil der CVM-Methode ist, dass zur Bestimmung des Wachstumsstadiums, anders als bei der Handröntgenaufnahme, keine zusätzlichen Röntgenbilder angefertigt werden müssen, sondern die Diagnostik anhand des FRS erfolgt, welches im Rahmen der kieferorthopädischen Routinediagnostik meist ohnehin erstellt wird [93]. Generell werden für die Wirbelkörperanalyse die Halswirbelkörper C2-C4 herangezogen und auf Grundlage ihrer Morphologie sechs Wachstumsstadien unterschieden. Die Wachstumsstadien CS1 und CS2 gelten als präpuberal, die Stadien CS3 und CS4 als circumpuberal und die Stadien CS5 sowie CS6 werden als postpuberal eingestuft [93].

Amasya et al. untersuchten 2020 die Analysequalität einer künstlichen Intelligenz zur Wirbelkörperanalyse anhand von 647 FRS von Patienten im Alter von zehn bis 30 Jahren. Die Ergebnisse der KI wurden mit den Entscheidungen von drei Radiologen und einem Kieferorthopäden mit jeweils mehrjähriger Berufserfahrung verglichen und zeigten für den Vergleich der KI mit jeder menschlichen Auswertung einen gewichteten Kappa-Koeffizienten von minimal 0,8 bis maximal 0,91. Insgesamt wurde eine Übereinstimmung von KI und Mensch von 58,3% erzielt [94]. Von deutlich höheren Übereinstimmungen zwischen KI und Mensch berichtete die Arbeitsgruppe um Seo et al. 2021. Anhand von 600 FRS von Patienten im Alter von sechs bis 19 Jahren wurden sechs Softwarelösungen zur automatisierten CVM-Analyse miteinander verglichen. Die Entscheidung der KIs wurde der Entscheidung eines Radiologen mit über zehnjähriger

Berufserfahrung gegenübergestellt und für alle sechs Anbieter wurden Richtigkeiten von über 90% dokumentiert [95].

1.2.7.3 KI zur therapeutischen Entscheidungsfindung

Im Rahmen der Therapieplanung muss ein Kieferorthopäde auf Grundlage der diagnostischen Unterlagen häufig komplexe und irreversible Entscheidungen treffen. Dazu zählen beispielsweise die Fragestellungen, ob im Rahmen der kieferorthopädischen Behandlung Zähne extrahiert werden müssen oder ob ein Patient aufgrund der Ausprägung der Kieferfehlstellung kombiniert kieferorthopädisch-kieferchirurgisch behandelt werden sollte [96, 97].

Extraktionsentscheidungen gestalten sich häufig schwierig, da eine Vielzahl an funktionellen, klinischen, radiologischen und soziokulturellen Faktoren in die Entscheidung einfließt [98-100]. Besonders bei Grenzfällen ist die Entscheidungsfindung nicht einfach und von der Ausbildung und der klinischen Erfahrung des Kieferorthopäden abhängig [101, 102]. Somit ist es nicht verwunderlich, dass unterschiedliche Behandler, insbesondere bei Grenzfällen, unterschiedliche Entscheidung treffen [103, 104]. Vor diesem Hintergrund ist es sinnvoll, dass KI den Behandler in der Entscheidungsfindung unterstützen.

Xie et al. entwickelten eine KI, die darauf trainiert wurde, bei elf- bis fünfzehnjährigen Patienten die Entscheidung für oder gegen eine kieferorthopädisch indizierte Extraktionstherapie zu treffen. Hierzu analysiert die KI zwei nicht quantifizierbare Werte (Heredität und exponierte, protrudierte Frontzähne bei inkompetentem Lippenschluss), fünf Werte aus der Modellanalyse (Engstand jeweils im Ober- und Unterkiefer, Overbite, Overjet und Ausprägung der Spee-Kurve), sowie 18 Werte aus der FRS-Analyse (beispielsweise WITS-Wert und Frontzahninklinationen). Die für die Extraktionsentscheidung relevantesten Parameter waren die Achsenstellung der unteren Frontzähne (gemessen zum Mandibularplanum) sowie der Lippenschluss (exponierte Frontzähne durch inkompetenten Lippenschluss als Kriterium für Extraktion). Die Übereinstimmung der KI mit der Entscheidung von Kieferorthopäden als Goldstandard wurde anhand von 20 Patienten analysiert und lag bei 80% [105].

Jung et al. beschrieben im Jahr 2016 eine KI, welche neben der Indikationsstellung für bzw. gegen eine kieferorthopädisch indizierte Extraktionstherapie zusätzlich das Extraktionsmuster, beispielsweise eine Kombination aus Extraktionen erster und zweiter Prämolaren in den verschiedenen Quadranten, vorhersagen kann. Die KI verwendet

zwölf Werte aus der FRS-Analyse, die weitestgehend mit den Werten der oben beschriebenen KI von Xie et al. übereinstimmen. Zusätzlich werden sechs Indices, wie der „Maxillary Arch Length Discrepancy Index“, der „Mandibular Arch Length Discrepancy Index“ oder der „Protrusion Index“ herangezogen. Die Besonderheit dieser KI ist, dass sie mehrere Outputs aufweist: sie unterscheidet zunächst zwischen Ex- und Non-Ex-Therapie, gibt aber zusätzlich an, welches Extraktionsmuster (Kombinationen aus unterschiedlichen Extraktionen im Ober- und Unterkiefer) sinnvoll ist und nennt das voraussichtliche Ausmaß der Frontzahnretraktion. Die generelle Extraktionsentscheidung stimmte zu 93% mit dem in der Studie definierten menschlichen Goldstandard überein und die Entscheidung bezüglich des Extraktionsmusters zeigte eine Übereinstimmung von 89% mit dem Goldstandard [101].

Eine von Li et al. 2019 veröffentlichte KI kann neben der Extraktionsentscheidung und der Auswahl des Extraktionsmusters zusätzlich die erforderliche Verankerung der 6-Jahr-Molaren festlegen. Die Übereinstimmung zwischen KI und dem Menschen lag für die Extraktionsentscheidung selbst bei 94%, für das Extraktionsmuster bei 84,2% und für die Verankerung bei 92,8% [96].

Eine weitere grundsätzliche Therapieentscheidung in der Kieferorthopädie ist, ob zur Einstellung einer stabilen Okklusion eine kieferverlagernde Dysgnathie-Operation erforderlich ist oder nicht. Diese Entscheidung ist von essenzieller Bedeutung in der Therapieplanung. Nach Abschluss des Wachstums können Kieferfehllagen zwar bis zu einem gewissen Ausmaß im Sinne einer Camouflage-Therapie ausschließlich kieferorthopädisch kompensiert werden, jedoch ist eine kausale Therapie im Erwachsenenalter nur im Rahmen eines kombiniert kieferorthopädisch-kieferchirurgischen Vorgehens möglich [84]. Insbesondere bei Grenzfällen ist die Entscheidung, ob ein kombiniert kieferorthopädisch-kieferchirurgisches Vorgehen notwendig ist, von großer Bedeutung, da die Behandlungspläne der beiden verschiedenen Vorgehensweisen maßgeblich voneinander abweichen. Soll ein kombiniertes Vorgehen vermieden werden, ist eine Kompensationsbehandlung notwendig, die sich von der Dekompensationsbehandlung als Vorbereitung auf die Umstellungsosteotomie hinsichtlich Verankerung und geplanter Zahnbewegung grundlegend unterscheidet. Die Entscheidung für oder gegen ein kombiniertes Vorgehen muss daher zu einem Zeitpunkt getroffen werden, bevor irreversible Maßnahmen wie Extraktionen bleibender Zähne durchgeführt werden [106]. Aufgrund unterschiedlicher Behandlungsphilosophien wird die Entscheidung für bzw. gegen ein kombiniertes

Vorgehen von verschiedenen Behandlern oftmals abweichend beurteilt. Diese Tatsache macht es insbesondere für weniger erfahrene Kieferorthopäden schwierig, diese komplexe Entscheidung zu treffen [97]. Hier können KI hilfreich sein.

Eine Arbeitsgruppe um Choi et al. stellte 2019 eine KI vor, die den Menschen bei der Entscheidung OP vs. Non-OP und zusätzlich Extraktion vs. Non-Extraktion unterstützen kann. Die KI verwendet die gleichen zwölf FRS-Parameter und sechs Indices wie die oben beschriebene KI von Jung et al. zur Extraktionsentscheidung, wurde jedoch darauf trainiert, drei verschiedene Output-Klassifikationen auszugeben: die Entscheidung OP vs. Non-OP, die Art der OP (Klasse-II- oder Klasse-III-OP) und die Entscheidung Extraktion vs. Non-Extraktion. Die Ergebnisse der KI wurden mit den Entscheidungen eines Kieferorthopäden mit über zehnjähriger Berufserfahrung verglichen. Dabei lag die Übereinstimmung innerhalb des Testdatensatzes von 112 Patientenfällen zwischen der KI und dem Menschen für die OP-Entscheidung bei 96%, für die OP-Art bei 100% und für die Extraktionsentscheidung bei 100% für Klasse-II-Patienten bzw. 86% für Klasse-III-Patienten. Die Erfolgsrate für die Gesamtdiagnose (OP-Art und Extraktionsentscheidung) wurde innerhalb des Testdatensatzes mit 90% angegeben [106].

1.2.7.4 KI zur Prognose des Therapieergebnisses

Häufig ändert sich durch eine kombiniert kieferorthopädisch-kieferchirurgische Behandlung das äußere Erscheinungsbild des Patienten maßgeblich [107]. Die Verbesserung der fazialen Ästhetik ist in vielen Fällen der ausschlaggebende Faktor für den Patienten, sich für eine kombinierte Therapie zu entscheiden [108, 109]. Auch im Rahmen der Aufklärung vor Beginn einer solchen Behandlung können KI-Algorithmen zur Simulation des fazialen Erscheinungsbildes nach Behandlungsende zum Einsatz kommen. Ter Horst et al. stellten 2021 eine KI vor, die mit 3D-Fotografien und digitalen Volumentomographien (DVTs) von Patienten vor und nach Unterkiefervorverlagerung (bilaterale sagittale Spaltosteotomie) im Rahmen einer Klasse-II-OP trainiert wurde und die Veränderungen im Weichgewebe anhand einer Foto-Simulation vorhersagen kann. Zu 64,3% werden Simulationen mit hoher Genauigkeit erreicht (Fehler von ≤ 1mm) und zu 92,2% werden Simulationen mit mittlerer Genauigkeit (Fehler von ≤ 2mm) erreicht. Für die untere Gesichtsregion war der mittlere absolute Fehler der KI sogar signifikant geringer als der des bisher zur Vorhersage der Weichteilveränderungen gebräuchlichen Massentensormodells (MTM) [110].

1.2.8 Kommerzialisierung von KI in der Zahnmedizin und Kieferorthopädie

KI sind mittlerweile auch in beträchtlicher Anzahl für den Praxisalltag verfügbar. Zu nennen ist beispielsweise die seit März 2021 erhältliche Software dentalXrai Pro (dentalXrai GmbH, Berlin, Deutschland). Der KI-Algorithmus kann anhand von Bissflügeln und OPGs Zähne erkennen und Restaurationen, Karies oder apikale Läsionen identifizieren. Diese werden auf dem Röntgenbild farbig markiert und können automatisch in die Patientenkartei übertragen werden [71, 111].

Der überwiegende Anteil der kommerziellen KI-Anbieter entfällt jedoch auf das Themengebiet der automatisierten FRS-Analyse. Hier sind DentaliQ.ortho der CellmatiQ GmbH (Hamburg, Deutschland), WebCeph der AssembleCircle Corp (Seongnam-si, Korea), AudaxCeph von Audax d.o.o. (Ljubljana, Slowenien) und CephX von Orca Dental AI (Herzliya, Isreal) anzuführen, die online weltweit für Kieferorthopäden zur Verfügung stehen.

Damit sich Kliniker im Alltag auf die KI-basierten Software-Lösungen verlassen können, müssen diese hohe Anforderungen hinsichtlich der Analysequalität erfüllen. Allerdings ist bei vielen kommerziellen Anbietern von KI-Lösungen unklar, auf welcher Datengrundlage die KI trainiert wurde und vielfach fehlt es an vergleichbaren wissenschaftlichen Untersuchungen bezüglich der zu erwartenden Auswertgenauigkeit [69].

1.3 Ziel der Arbeit

Die Entwicklung von künstlicher Intelligenz schreitet in allen medizinischen Bereichen zügig voran und kann Ärzte sowohl in der Diagnostik als auch in der therapeutischen Entscheidungsfindung zunehmend unterstützen. Auch in der Kieferorthopädie erlangen KI-Algorithmen einen immer höheren Stellenwert.

Bis dato wurde die meiste Forschungsarbeit im Bereich der automatisierten FRS-Analyse geleistet, sodass es in den letzten Jahren zu einem exponentiellen Anstieg an wissenschaftlichen Publikationen zu diesem Thema gekommen ist. Nach Schwendicke et al. weisen viele aktuelle Studien zum Thema KI in der FRS-Analyse allerdings noch erhebliche Limitationen auf und lassen daher keine eindeutigen Schlussfolgerungen bezüglich der klinischen Anwendbarkeit der KI zu. Hinzu kommt, dass einige KI bereits kommerziell angeboten werden, wenngleich die wissenschaftliche Datengrundlage noch unklar ist [112]. Schwendicke et al. betonen deshalb in ihrer Meta-

Analyse, dass in zukünftigen Studien die Robustheit, Generalisierbarkeit und vor allem die klinische Anwendbarkeit von KI genauer untersucht werden sollte [79].

Nach Durchsicht der aktuellen Fachliteratur gibt es bisher keine einheitlichen wissenschaftlichen Untersuchungen, die Aufschlüsse über die Genauigkeit und damit über die klinische Anwendbarkeit der aktuell auf dem Markt verfügbaren kommerziellen KI-Anbieter für automatisierte FRS-Analysen geben. Die wenigen bisher vorhandenen Studien zur Genauigkeit der Auswertungen einzelner KI-Anbieter sind häufig nicht miteinander vergleichbar, da zum einen der Goldstandard nicht einheitlich definiert und zum anderen unterschiedliche kieferorthopädische Parameter analysiert wurden. Insbesondere für klinisch tätige Kieferorthopäden ist jedoch eine gute wissenschaftliche Grundlage von essenzieller Bedeutung, um solche Softwarelösungen zuverlässig im Praxisalltag integrieren zu können.

Ziel der vorliegenden Arbeit war es daher, die Auswertegenauigkeit der aktuell auf dem Markt verfügbaren kommerziellen KI-Anbieter für automatisierte FRS-Analysen zu untersuchen und mit einem durch menschliche Experten definierten Goldstandard zu vergleichen. Daraus leiteten sich folgende Nullhypothesen ab, die auf Grundlage der vorliegenden Arbeit akzeptiert oder verworfen werden sollten:

- Es besteht kein statistisch signifikanter Unterschied zwischen der automatisierten FRS-Auswertung durch DentaliQ.ortho und dem menschlichen Goldstandard.
- Es besteht kein statistisch signifikanter Unterschied zwischen der automatisierten FRS-Auswertung durch WebCeph und dem menschlichen Goldstandard.
- Es besteht kein statistisch signifikanter Unterschied zwischen der automatisierten FRS-Auswertung durch AudaxCeph und dem menschlichen Goldstandard.
- Es besteht kein statistisch signifikanter Unterschied zwischen der automatisierten FRS-Auswertung durch CephX und dem menschlichen Goldstandard.

2 Material und Methoden

2.1 Studiendesign

Die vorliegende Studie wurde in Übereinstimmung mit der Deklaration von Helsinki durchgeführt.

Die für die Studie verwendeten FRS wurden im Rahmen der kieferorthopädischen Routinediagnostik erstellt und lagen zum Zeitpunkt der Studienplanung bereits vor. Zudem handelt es sich bei einem FRS um ein Bildgebungsverfahren, welches nicht zur Identifikation einer Person geeignet ist. Um einen adäquaten Datenschutz zu gewährleisten, wurden alle patientenbezogenen Metadaten, wie Name, Geschlecht oder Alter der Patienten vor der Verarbeitung der Datensätze gelöscht. Die FRS wurden für die weitere Verarbeitung mit einer zufällig generierten 36-stelligen Kombination aus Zahlen und Buchstaben anonymisiert. Aus diesen Gründen war kein Ethikvotum zur Durchführung der Studie erforderlich.

2.2 Patienten und FRS

Die in der Studie verwendeten FRS stammen aus einer privaten kieferorthopädischen Praxis und wurden mit Hilfe eines digitalen Orthophos-XG-5-Gerätes von Sirona im Rahmen der klinischen Routinediagnostik zur kieferorthopädischen Therapieplanung aufgenommen (Sirona Dental Systems GmbH, Bensheim, Deutschland). Aus einem Pool von 3000 FRS wurden zufällig 50 FRS für die vorliegende Studie ausgewählt. Die FRS wiesen eine große Heterogenität bezüglich der Gebissphase und der konservierenden bzw. prothetischen Vorbehandlungen der Patienten sowie der bei den Patienten zum Zeitpunkt der Bilderstellung eingesetzten kieferorthopädischen Apparaturen auf.

2.3 FRS-Analyse

Die für die vorliegende Studie verwendete kephalometrische Analyse basiert auf 15 gebräuchlichen Landmarken und neun gängigen Parametern (acht anguläre Parameter und ein Streckenverhältnis). Auf metrische Parameter wurde bewusst verzichtet, da zum Zeitpunkt der Datenerhebung nicht alle verfügbaren kommerziellen KI-Anbieter die metrische Referenzskala selbstständig erkennen konnten

beziehungsweise nur vereinzelt unterschiedliche metrische Parameter in den Analysen der kommerziellen KI-Anbietern enthalten waren.

Die Zusammenstellung der Werte für die kephalometrische Analyse dieser Studie orientiert sich an der Publikation von Kunz et al. [32]. Abbildung 1 veranschaulicht die Position der Landmarken im FRS anhand einer schematischen Zeichnung (eigene Abbildung) und Tabelle 1 beschreibt die Definition der einzelnen Landmarken. Tabelle 2 zeigt die auf diesen Landmarken basierenden Parameter. Die Abbildungen 2-4 visualisieren diese anhand der drei Kategorien skelettal sagittale, skelettal vertikale und dentale Analyse im FRS (eigene Abbildungen).

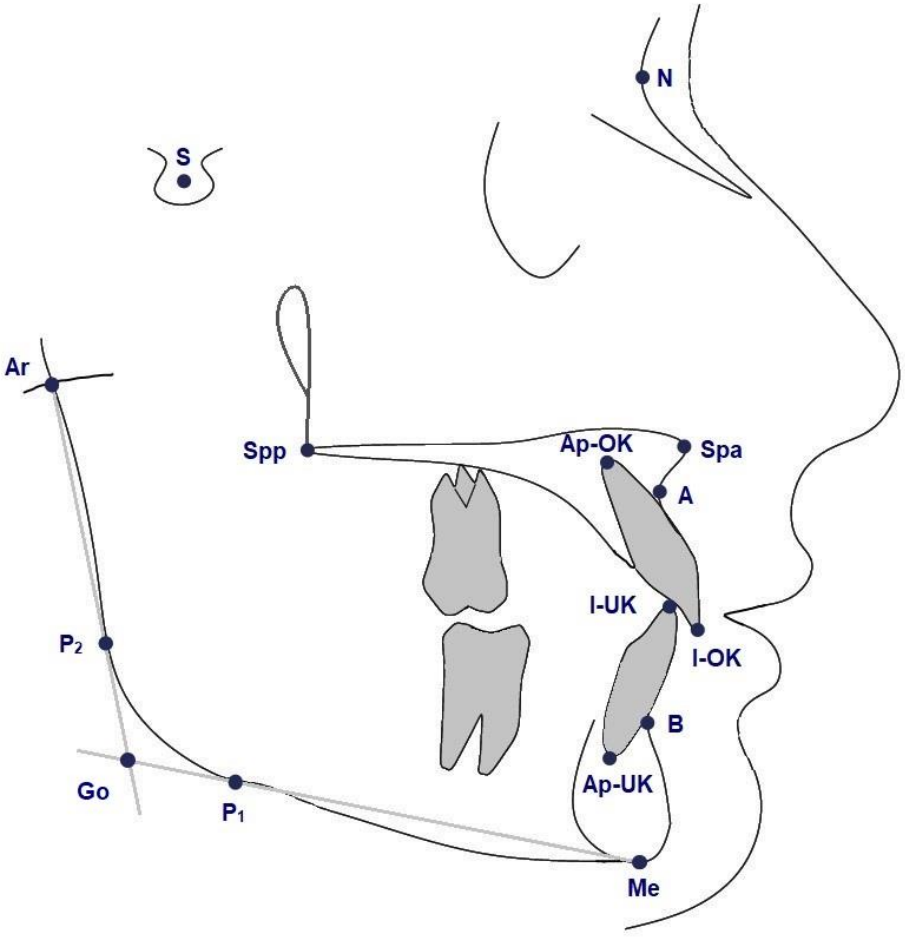


Abbildung 1: Position der für die Studie verwendeten Landmarken im FRS (eigene Abbildung).

Tabelle 1: Definition der Landmarken für die kephalometrische Analyse.

Landmarke	Abkürzung	Definition
Sella	S	Mittelpunkt der Fossa hypophysialis
Nasion	N	Anteriorster Punkt der Sutura nasofrontalis in der Median-Sagittal-Ebene
A-Punkt	A	Posteriorster Punkt der äußeren Kurvatur des Processus alveolaris maxillae in der Median-Sagittal-Ebene
B-Punkt	B	Posteriorster Punkt der äußeren Kurvatur des Processus alveolaris mandibulae in der Median-Sagittal-Ebene
Spina nasalis anterior	Spa	Anteriorster Punkt der knöchernen Spina nasalis anterior in der Median-Sagittal-Ebene (= anteriore Begrenzung der Maxilla)
Spina nasalis posterior	Spp	Radiologischer Schnittpunkt der vorderen Wand der Fossa pterygopalatina mit dem Nasenboden (= posteriore Begrenzung der Maxilla)
Articulare	Ar	Radiologischer Schnittpunkt zwischen dem hinteren Rand des Ramus ascendens mandibulae und dem äußeren Rand der Schädelbasis
P_1-Punkt	P_1	Kaudalster Punkt an der äußeren Kurvatur des Corpus mandibulae im Bereich der Protuberantia masseterica
P_2-Punkt	P_2	Dorsalster Punkt des Ramus ascendens mandibulae im Bereich des Kieferwinkels
Menton	Me	Kaudalster Punkt an der äußeren Kontur der Symphysis mandibulae
Gonion	Go	Schnittpunkt des Mandibularplanums (Me-P_1) und der Tangente am Ramus ascendens mandibulae (Ar-P_2) *(konstruierte Landmarke)*
Incision superior	I-OK	Spitze der Inzisalkante des am weitesten anterior gelegenen oberen mittleren Schneidezahnes
Incision inferior	I-UK	Spitze der Inzisalkante des am weitesten anterior gelegenen unteren mittleren Schneidezahnes
Apicale superior	Ap-OK	Wurzelspitze des am weitesten anterior gelegenen oberen mittleren Schneidezahnes in der Längsachse der Zahnwurzel
Apicale inferior	Ap-UK	Wurzelspitze des am weitesten anterior gelegenen unteren mittleren Schneidezahnes in der Längsachse der Zahnwurzel

Tabelle 2: Definition der Parameter für die kephalometrische Analyse.

	Parameter	Einheit	Definition und Interpretation
Skelettal sagittale Analyse	SNA	°	Winkel zwischen Sella, Nasion und A-Punkt → *sagittale Position der Maxilla*
	SNB	°	Winkel zwischen Sella, Nasion und B-Punkt → *sagittale Position der Mandibula*
	ANB	°	Winkel zwischen A-Punkt, Nasion und B-Punkt → *skelettale Klasse*
Skelettal vertikale Analyse	SN_SpP	°	Winkel zwischen der vorderen Schädelbasis (= Sella-Nasion-Linie) und dem Oberkieferplanum (= Spa-Spp-Linie) → *Neigung des Oberkiefers*
	SN_MeGo	°	Winkel zwischen der vorderen Schädelbasis (= Sella-Nasion-Linie) und dem Unterkieferplanum (= Menton-Gonion-Linie) → *Neigung des Unterkiefers*
	SpP_MeGo	°	Winkel zwischen dem Oberkieferplanum (= Spa-Spp-Linie) und dem Unterkieferplanum (= Menton-Gonion-Linie) → *skelettal offener / skelettal tiefer Biss*
	Gesichtshöhen-verhältnis	%	Verhältnis der posterioren (= Sella-Gonion-Linie) und der anterioren (= Nasion-Menton-Linie) Gesichtshöhe zueinander → *Wachstumsmuster*
Dentale Analyse	O1_SN	°	Winkel zwischen dem oberen mittleren Schneidezahn (= Linie zwischen Incision superior und Apicale superior) und der anterioren Schädelbasis (= Sella-Nasion-Linie) → *Inklination des oberen mittleren Frontzahnes*
	U1_MeGo	°	Winkel zwischen dem unteren mittleren Schneidezahn (= Linie zwischen Incision inferior und Apicale inferior) und dem Unterkieferplanum (= Menton-Gonion-Linie) → *Inklination des unteren mittleren Frontzahnes*

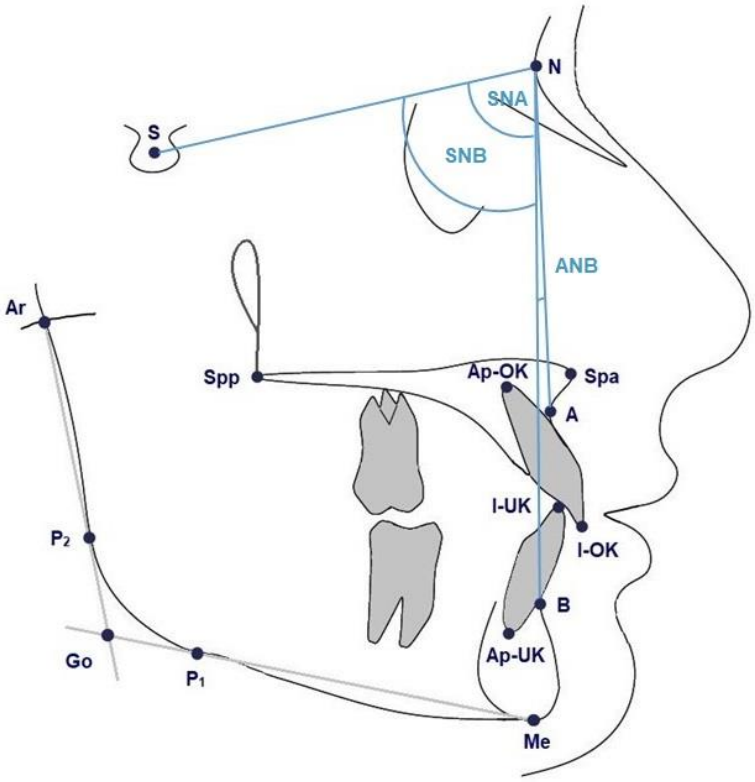


Abbildung 2: Skelettal sagittale Analyse (eigene Abbildung).

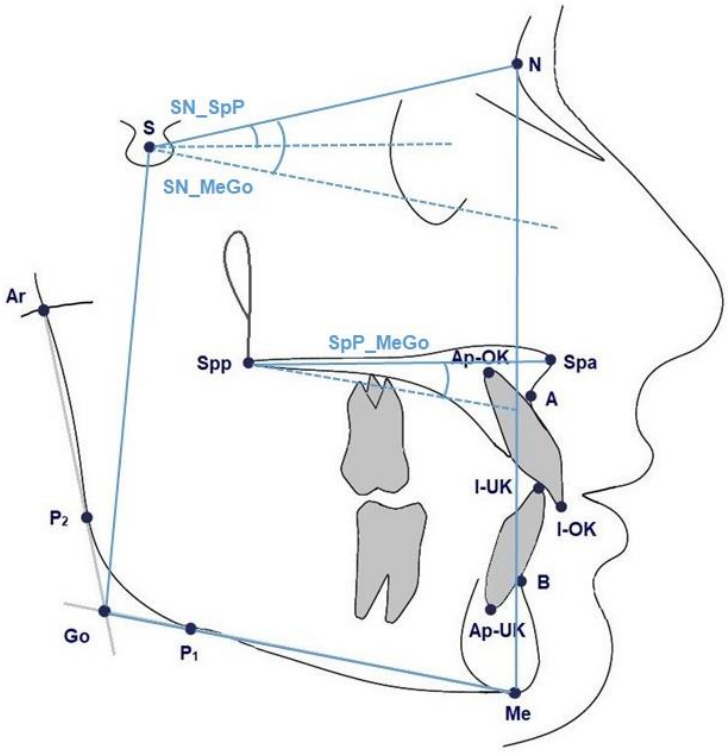


Abbildung 3: Skelettal vertikale Analyse (eigene Abbildung).

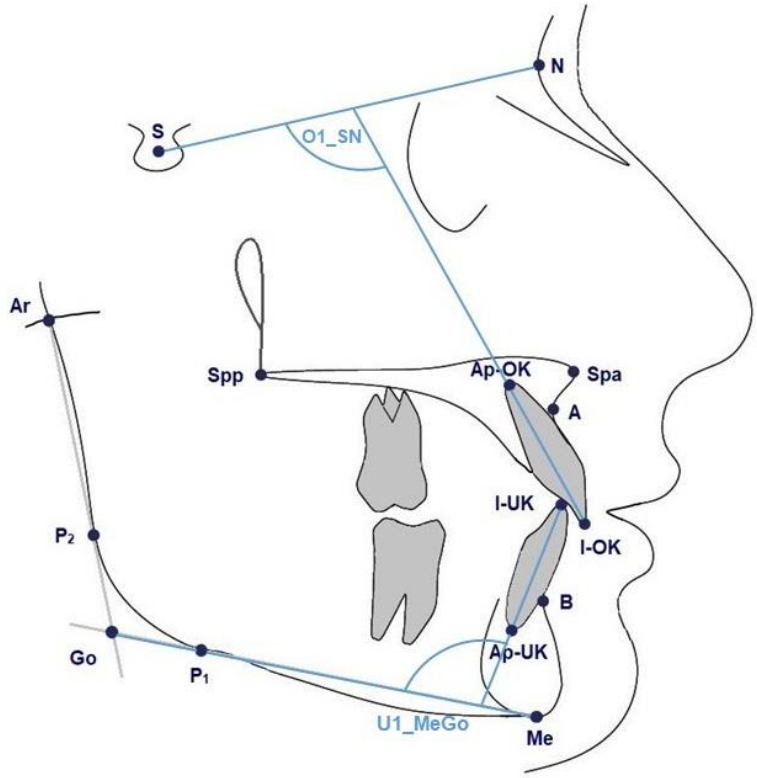


Abbildung 4: Dentale Analyse (eigene Abbildung).

2.4 Definition des menschlichen Goldstandards

Zur Festlegung des Goldstandards wurden alle 50 FRS durch zwölf erfahrene Untersucher der Poliklinik für Kieferorthopädie des Universitätsklinikums Würzburg digital ausgewertet. Bei den Untersuchern handelte es sich jeweils hälftig um Fachzahnärzte für Kieferorthopädie und hälftig um Zahnärzte in der fachzahnärztlichen Weiterbildung. Nach Auswertung aller FRS wurde für jeden der neun Parameter auf jedem FRS der Medianwert aller zwölf Untersucher als menschlicher Goldstandard definiert. Dieses Verfahren zur Definition des menschlichen Goldstandards wurde bereits in einer früheren Untersuchung beschrieben [32].

Zudem wurden von den 50 FRS zufällig 20 FRS ausgewählt und durch jeden Untersucher mit einem zeitlichen Abstand von mehreren Wochen erneut ausgewertet, um die Intraraterreliabilität der Untersucher zu analysieren.

2.5 Auswahl der kommerziellen KI-Anbieter

Die Auswahl der untersuchten kommerziellen KI-Anbieter erfolgte im März 2021. Ziel war es, alle auf dem Markt verfügbaren KI, die auf Basis eines deep-learning-Algorithmus eine vollständig automatisierte FRS-Analyse anbieten, in die Studie einzuschließen. Ein zusätzliches Einschlusskriterium war, dass alle in den Tabellen 1 (sh. Seite 21) und 2 (sh. Seite 22) aufgeführten Punkte und Parameter in den FRS-Analysen der KI enthalten waren.

Folgende kommerzielle KI-Anbieter erfüllten diese Einschlusskriterien und wurden folglich in die Studie aufgenommen:

- DentaliQ.ortho der CellmatiQ GmbH (Hamburg, Deutschland),
- WebCeph der AssembleCircle Corp (Seongnam-si, Korea),
- AudaxCeph von Audax d.o.o. (Ljubljana, Slowenien) und
- CephX von Orca Dental AI (Herzliya, Isreal).

Eine weitere Software in App-Form, CephNinja von Cyncronus LLC (Bothell, USA), wurde nicht in die vorliegende Studie einbezogen, da durch die KI einige Punkte und Parameter unserer FRS-Analyse entweder nicht oder nur mit extremen Abweichungen generiert wurden. Auf Nachfrage bei den Entwicklern erhielten wir die Information, dass es sich bei der Version von CephNinja, die zum Zeitpunkt der Datenerhebung für die vorliegende Studie genutzt wurde, um die erste Version der KI handele und weitere Verbesserungen notwendig seien.

Im Folgenden werden die vier untersuchten kommerziellen KI-Anbieter zur besseren Übersicht visualisiert.

2.5.1 DentaliQ.ortho

Die Software DentaliQ.ortho der CellmatiQ GmbH aus Deutschland bietet als Analyse eine Zusammenstellung diverser skelettaler und dentaler Parameter an. Eine Auswahl zwischen verschiedenen Analysen ist nicht möglich. Die Auswertung der FRS erfolgt online, wobei die Software nach den ersten zehn Auswertungen kostenpflichtig ist. Abbildung 5 zeigt die automatisierte FRS-Analyse anhand eines Beispiels, Abbildung 6 zeigt die zugehörige Auswertung des FRS.

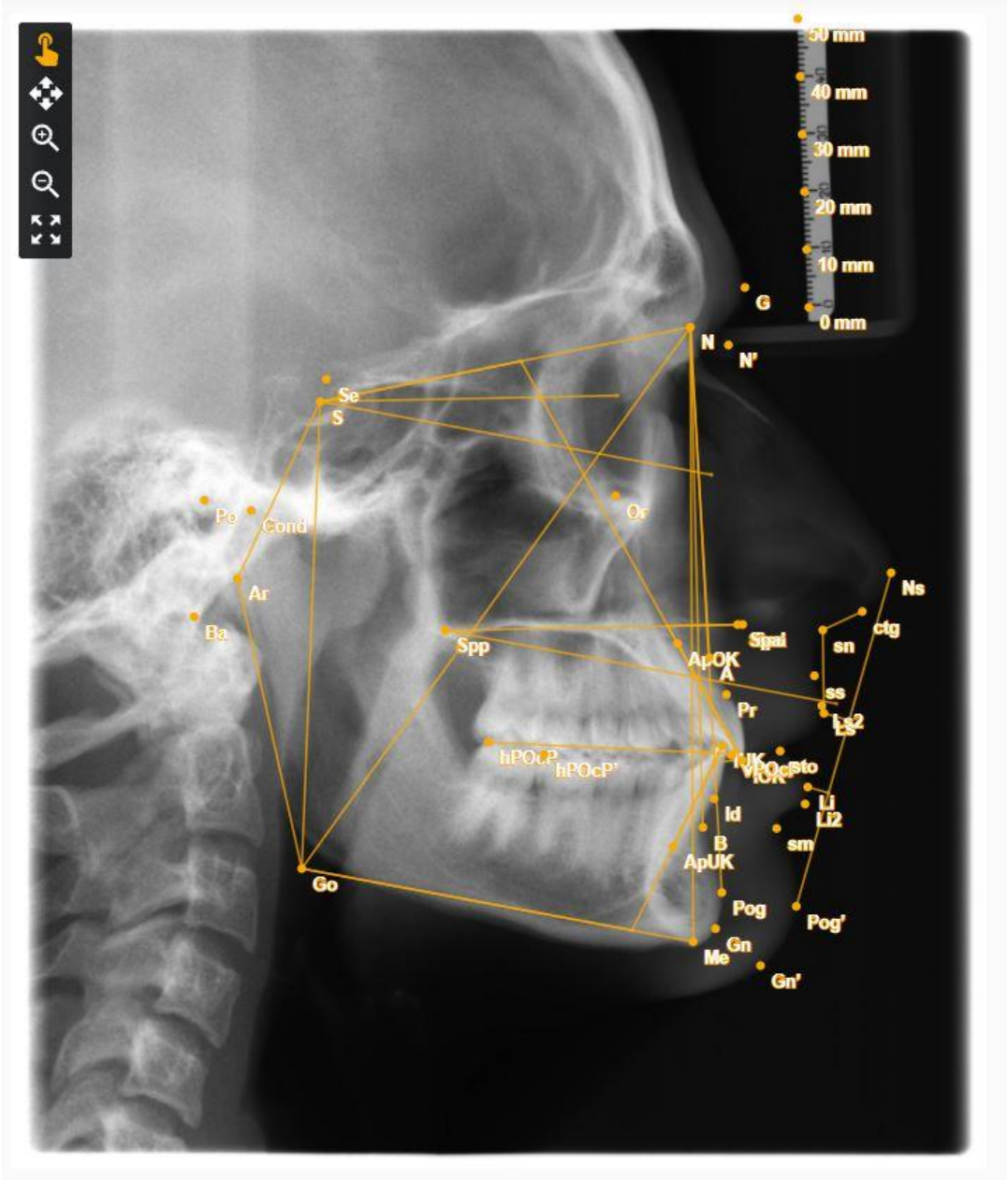


Abbildung 5: FRS-Analyse mit DentaliQ.ortho, https://ortho.dentaliq.ai/ceph/analysis.

Gewähltes ethnisches Profil: Kaukasisch

Analyse	Norm/Toleranz	Messwert	Abweichung	Befund
Einbau der Kieferbasen				
SNA	81° ± 3.5°	81.93°	+0.93°	Oberkiefer orthognath
SNB	78° ± 3°	80.07°	+2.07°	Unterkiefer orthognath
SNPog	78.5° ± 3°	81.82°	+3.32°	* Stark ausgeprägtes Kinn
Skelettale Klasse				
ANB	2° ± 2°	1.86°	-0.14°	Skelettale Klasse I
iANB/ANB	2° ± 1°	1.86°	-0.14°	Skelettale Klasse I
WITS	0mm ± 2mm	-0.35mm	-0.35mm	Skelettale Klasse I
Wachstumsmuster				
NSAr	123.5° ± 5°	126.42°	+2.92°	Durchschnittliche Lage der Kiefergelenk-Grube
SArGo	142° ± 6°	142.44°	+0.44°	Unterkiefer orthognath
ArGoMe	128.5° ± 6°	113.1°	-15.4°	* Horizontales Wachstumsmuster = anteriore Rotation des UK
Go1	55° ± 4°	48.2°	-6.8°	* Verkleinert
Go2	72.5° ± 4.5°	64.9°	-7.6°	* Verkleinert
Summen_W	394° ± 5°	381.96°	-12.04°	* Horizontales Wachstumsmuster
SN_SpP	7° ± 3°	10.21°	+3.21°	* Retroinklination des Oberkiefers
SN_MeGo	33.5° ± 5°	21.95°	-11.55°	* Horizontales Wachstumsmuster
SpP_MeGo	26.5° ± 5°	11.74°	-14.76°	* Skelettal tiefer Biss
GHV	65% ± 4%	76.03%	+11.03%	* Horizontales Wachstumsmuster
Dentale Analyse				
1_SN	103° ± 6.5°	107.66°	+4.66°	OK-Front achsengerecht
1_SpP	70° ± 6°	62.13°	-7.87°	* OK-Front labial
1_MeGo	93° ± 6°	105.54°	+12.54°	* UK-Front labial
1OK_1UK	130° ± 9°	124.84°	-5.16°	Stabil
NPog_IOK	6.5mm ± 3mm	5.67mm	-0.83mm	OK-Front in Orthoposition
NPog_IUK	3mm ± 2.5mm	1.84mm	-1.16mm	UK-Front in Orthoposition
Weichteilanalyse				
OL	-2mm ± 2mm	-5.3mm	-3.3mm	* Retrusive Oberlippe
UL	-0.5mm ± 2mm	-4.37mm	-3.87mm	* Retrusive Unterlippe
NLW	102° ± 7°	114.28°	+12.28°	* Vergrößert

Abbildung 6: FRS-Auswertung mit DentaliQ.ortho, https://ortho.dentaliq.ai/ceph/analysis.

2.5.2 WebCeph

Bei der Software WebCeph der AssembleCircle Corp aus Korea erfolgt die FRS-Analyse ebenfalls online. Die Basisversion der Software ist kostenfrei. Der Untersucher kann zwischen elf verschiedenen Analysen wählen. Für die vorliegende Studie wurden Werte aus der WebCeph-, der Eastman- und der Jarabak-Analyse verwendet. Abbildung 7 zeigt beispielhaft die Auswertung eines FRS mittels WebCeph-Analyse.

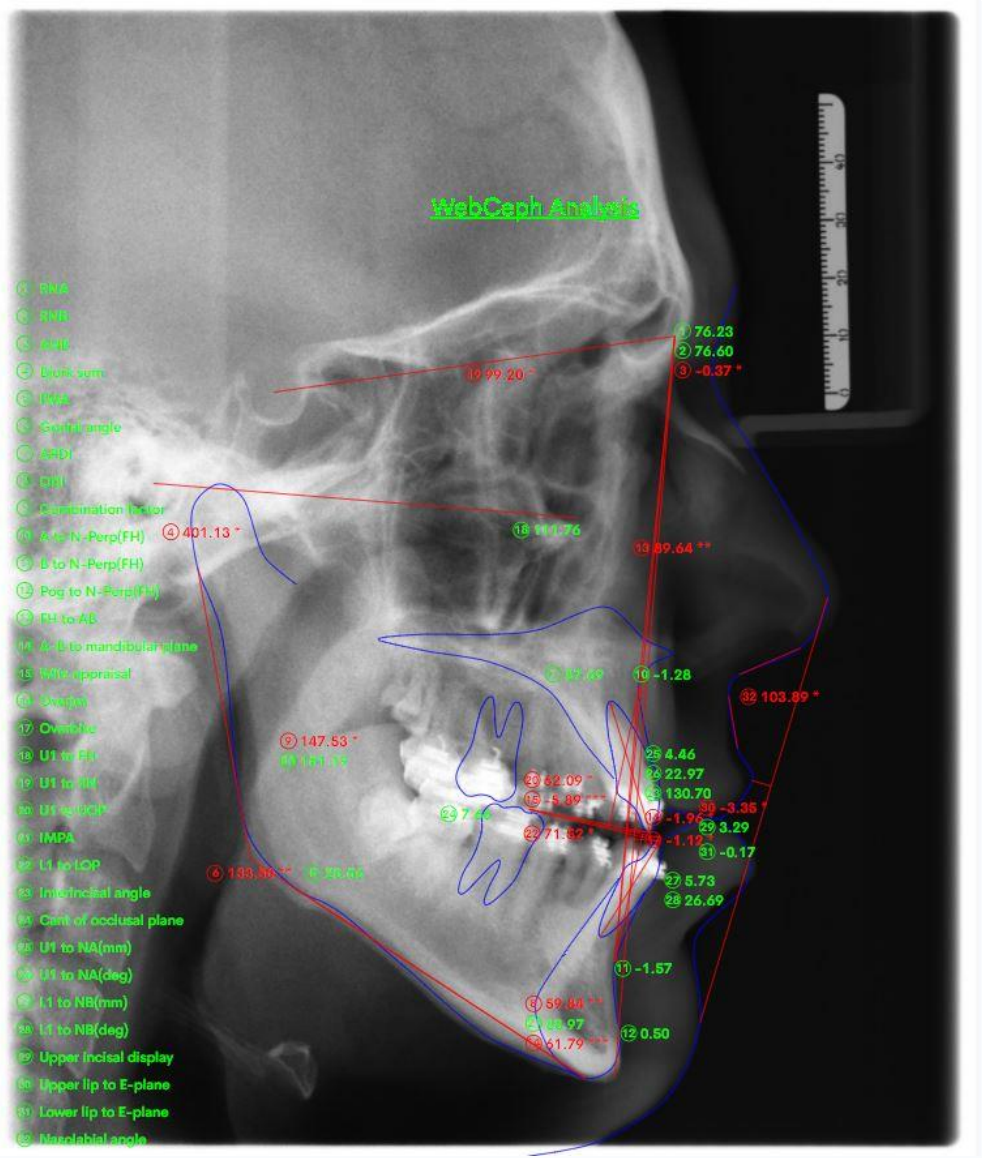


Abbildung 7: FRS-Analyse mit WebCeph, WebCeph-Analyse, https://webceph.com/de/records/3hw4awqAR44t/2021-01-08/analysis/.

2.5.3 AudaxCeph

Bei AudaxCeph von Audax d.o.o. aus Slowenien erfolgt die FRS-Analyse nach Download einer Software. Nach einem 30-tägigen Testzeitraum ist die Nutzung der Software kostenpflichtig. Zum Zeitpunkt der Datenerhebung wurden insgesamt 21 verschiedene Standard-Analysen angeboten, weitere Analysen standen zum zusätzlichen Download bereit. Für die Studie wurden Werte aus den Analysen

„UniLjubljana“ und „ABO American Board“ verwendet. In Abbildung 8 ist die Analyse „UniLjubljana“ zu erkennen, Abbildung 9 zeigt die zugehörige Auswertung des FRS.

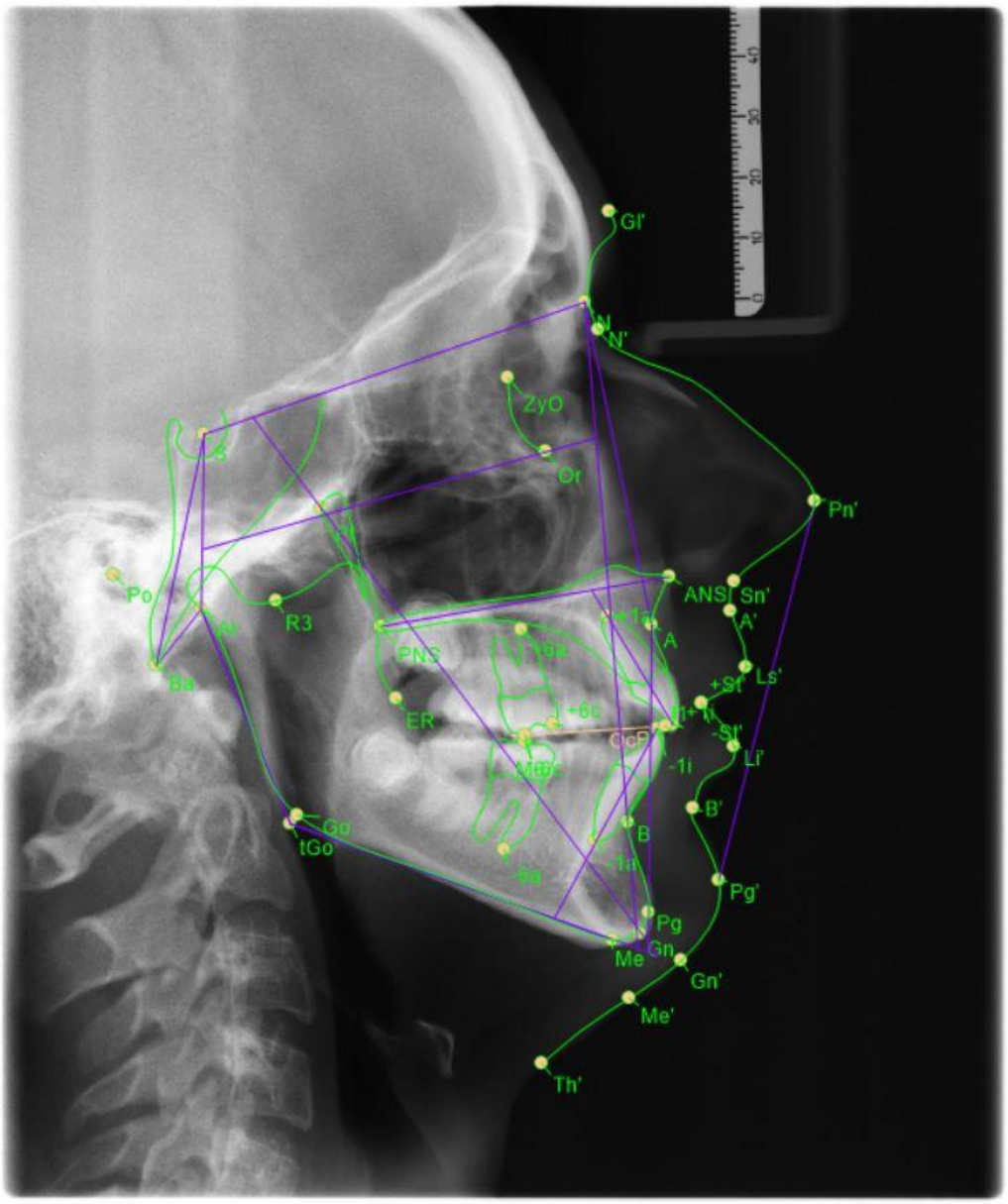
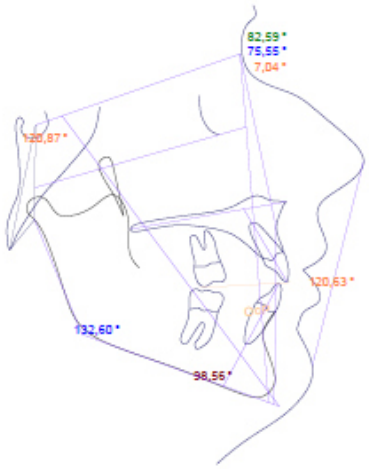


Abbildung 8: FRS-Analyse mit AudaxCeph, Analyse „UniLjubljana“, aus Programm AudaxCeph Empower.



MEASUREMENT		NORMAL VALUE	VALUE	DIFFERENCE	BIAS
SAGITAL RELATIONS					
SNA	°	82±2	83	1	|
SNB	°	80±3	76	-4	•|
ANB	°	2±2	7	5	|••
Wits	mm	1±1	6	5	|•••
SN-Ba	°	130±4	121	-9	••|
SNPg	°	81±3	77	-4	•|
+1/APg	mm	2±2	3	1	|
VERTICAL RELATIONS					
SN/FH	°	6±2	3	-3	•|
NL/NSL	°	9±3	9	0	|
ML-NSL	°	32±4	39	7	|•
NL/ML'	°	24±4	30	6	|•
Facial axis	°	90±4	87	-3	|
PFH/AFH	%	64±4	62	-2	|
Bjork	°	396±4	399	3	|
Gonial angle	°	120±7	133	13	|•
DENTAL ANALYSIS					
+1/NL	°	109±6	111	2	|
+1/NA	°	22±2	19	-3	•|
+1i/NA	mm	4±1	0	-4	•••|
-1/ML'	°	90±1	99	9	|•••
-1/NB	°	25±2	33	8	|•••
-1i/NB	mm	4±1	6	2	|••
-1/APg	°	22±1	28	6	|•••
-1i/APg	mm	1±1	1	0	|
+1/-1	°	135±5	121	-14	••|
SOFT TISSUE					
Ls'/E-line	mm	-4±2	-4	0	|
Li'/E-line	mm	-2±2	-3	-1	|

Abbildung 9: FRS-Auswertung mit AudaxCeph, Analyse „UniLjubljana“, aus Programm AudaxCeph Empower.

2.5.4 CephX

CephX von Orca Dental AI aus Israel stellt ebenfalls eine Online-Plattform zur FRS-Auswertung zur Verfügung. Nach einer siebentägigen Testphase ist die Nutzung der Software kostenpflichtig. Zum Zeitpunkt der Datenerhebung wurden 58 unterschiedliche Analysen angeboten. Da drei der untersuchten Punkte bzw. Parameter in der Mehrzahl der verfügbaren Analysen anders als in unserer Studie definiert wurden (Gonion, Mandibularplanum und Gesichtshöhenverhältnis), wurde nach Rücksprache mit einer Mitarbeiterin der Orca Dental AI eine individualisierte FRS-Analyse für diese Studie erstellt. Abbildung 10 zeigt beispielhaft das durch die KI automatisierte Setzen der Punkte. Die gewünschte Analyse wird in einem zweiten Schritt ausgewählt. Abbildung 11 zeigt die eigens für diese Studie entwickelte FRS-Analyse.

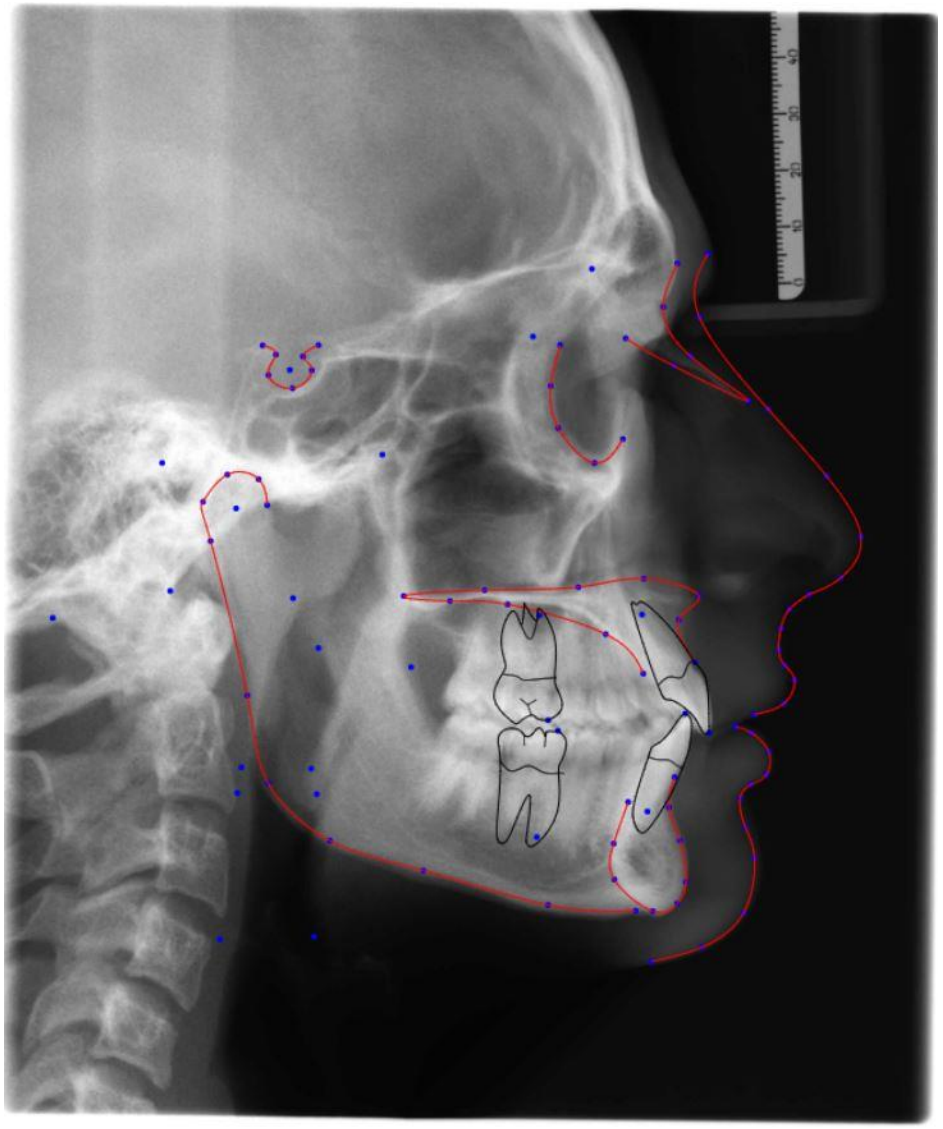


Abbildung 10: FRS-Analyse mit CephX, https://cloud.cephx.com/cephx/cephx.jsp.

Descriptor	Meas.	Type	Mean	Sd	Patient	Graph	Comment
SNA		Deg	82.0	2.0	83.25	-(*)+	
SNB		Deg	80.0	2.0	81.06	-(*)+	
ANB		Deg	2.0	2.0	2.19	-(*)+	
Wits Appraisal		mm	0.0	2.0	3.55	-(|*)+	Class II Skeletal problem
SN-SpP = SN to maxillary plane		Deg	7.0	3.0	9.46	-(*)+	
SN-MeGo (constructed Gonion)		Deg	33.5	5.0	27.99	-(*|)+	
SpP-MeGo (with constructed gonion)		Deg	26.5	5.0	18.52	-(*|)+	
Facial proportion (Jarabak) (with constructed gonion)		%	65.0	4.0	68.5	-(*)+	
U1 to SN		Deg	103.0	4.0	109.91	-(|*)+	
L1_MeGo (with constructed gonion)		Deg	93.0	6.0	92.75	-(*)+	
U1 TO NPg		mm	6.5	3.0	5.16	-(*)+	
L1 to NPg		mm	3.0	2.5	1.08	-(*)+	

Abbildung 11: FRS-Auswertung mit CephX, https://cloud.cephx.com/cephx/cephx.jsp.

2.6 Datenerhebung

Die Datenerhebung erfolgte im März 2021 durch eine erfahrene Mitarbeiterin der Poliklinik für Kieferorthopädie des Universitätsklinikums Würzburg. Eine erneute Überprüfung der Werte fand im Juli 2021 statt, um eine mögliche Veränderung der Messwerte durch zwischenzeitliche Updates der KI auszuschließen. Keiner der untersuchten kommerziellen KI-Anbieter zeigte dabei im Vergleich zur ersten Auswertung Abweichungen.

Zur Auswertung der FRS wurde der Datensatz aus 50 FRS in alle zu untersuchenden KI-Algorithmen importiert. Nach der automatisierten Auswertung wurden die von den kommerziellen KI-Anbietern ermittelten Werte für die neun untersuchten kephalometrischen Parameter exportiert. Aufgrund von ausgeprägten Artefakten wurden nachträglich drei FRS von der Untersuchung ausgeschlossen, sodass die statistische Auswertung anhand von 47 FRS erfolgte.

2.7 Statistische Auswertung

Die statistische Auswertung erfolgte durch Herrn Priv.-Doz. Dr. Felix Kunz in enger Koordination mit dem Diplom-Mathematiker und professionellen Statistiker Herrn Florian Zeman, stellvertretender Leiter des Zentrums für Klinische Studien des Universitätsklinikums Regensburg, unter Verwendung der Softwares IBM® SPSS® Statistics 25.0 für Windows (IBM, Ehningen, Deutschland) und R (Version 4.1.0).

Zur Überprüfung der Reliabilität der Analysen der zwölf Untersucher, die zur Definition des menschlichen Goldstandards verwendet wurden, wurden Intraklassen-Korrelations-Koeffizienten (*ICC*) verwendet. Mit Hilfe der *ICC* wurden die Intraraterreliabilität für jeden Untersucher und jeden Parameter sowie die Interraterreliabilität für jeden Parameter bestimmt.

Für die statistische Auswertung der Daten wurden zunächst ANOVA mit Messwiederholung durchgeführt, um die vier kommerziellen Anbieter und den Goldstandard miteinander zu vergleichen. In diesem Zusammenhang wurden die Mittelwerte (*M*), die Standardabweichungen der Mittelwerte (*SD*) und die *p*-Werte nach Greenhouse-Geisser angegeben. Angesichts der Anzahl der ausgewerteten FRS war eine Normalverteilung der Werte anzunehmen.

Im nächsten Schritt wurde im Rahmen des Post-hoc-Tests mit paarweisen Vergleichen untersucht, ob zwischen menschlichem Goldstandard und der Vorhersage der verschiedenen kommerziellen KI-Anbieter signifikante Unterschiede bestehen. Dazu wurden die mittlere Differenz (kommerzieller KI-Anbieter - Mensch), das 95%-Konfidenzintervall der mittleren Differenz mit Unter- und Obergrenze sowie der *p*-Wert dokumentiert. Um die Alpha-Fehler-Kumulierung auszugleichen, wurde die Bonferroni-Korrektur verwendet.

Im Anschluss wurden für jeden Parameter Bland-Altman-Plots erstellt. Auf der x-Achse wurde der menschliche Goldstandard, auf der y-Achse jeweils die Differenz zwischen der Vorhersage der kommerziellen KI-Anbieter und dem menschlichen Goldstandard aufgetragen [113]. Dabei dienten die Bland-Altman-Plots vor allem zur Beurteilung der Genauigkeit der KI-basierten FRS-Analysen, wobei sich die Genauigkeit eines bestimmten Verfahrens aus der Richtigkeit und der Präzision zusammensetzt. Die Richtigkeit eines Messverfahrens beschreibt, wie nah das arithmetische Mittel einer großen Anzahl an Messwerten an dem eigentlich richtigen Referenzwert liegt. Dabei müssen die einzelnen Messwerte nicht unbedingt eng beieinanderliegen. Dies wird gesondert durch die Präzision beschrieben, welche den Grad der Übereinstimmung zwischen den einzelnen Messwerten angibt. Um eine hohe Präzision zu erreichen, müssen die Messwerte nur eng beieinander lokalisiert sein, dies kann aber auch an einem anderen Punkt als dem eigentlich richtigen Referenzwert erfolgen [114]. Ziel eines Messverfahrens ist immer eine hohe Genauigkeit, also ein hohes Maß an Richtigkeit und Präzision. Dies ist in Abbildung 12 veranschaulicht.

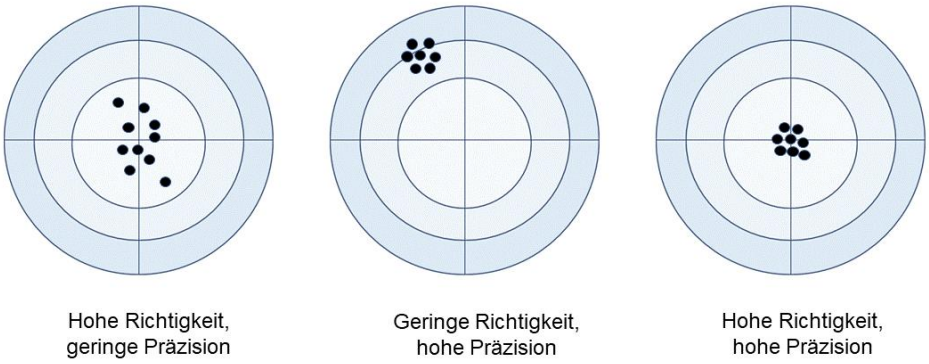


Abbildung 12: Graphische Darstellung von Richtigkeit und Präzision (eigene Abbildung).

Die mittleren Differenzen zwischen den Analysen wurden als Maßstab für die durchschnittliche Richtigkeit der kommerziellen KI-Anbieter verwendet. Von einer hohen durchschnittlichen Richtigkeit wurde bei mittleren Differenzen zum menschlichen Goldstandard von < 0,5° bzw. % ausgegangen, von einer moderaten durchschnittlichen Richtigkeit bei mittleren Differenzen zwischen 0,5° und 1° bzw. % und von einer niedrigen durchschnittlichen Richtigkeit bei mittleren Differenzen von > 1° bzw. %. Die 95% Limits of Agreement (LoA, mittlere Differenz ± 1,96 * Standardabweichung der mittleren Differenzen) dienten zur Beurteilung der Präzision der Analysen der kommerziellen KI-Anbieter. Die Präzision wurde bei einer Standardabweichung der Differenzen von < 1,5° bzw. % als hoch, bei einer Standardabweichung zwischen 1,5° und 2,5° bzw. % als moderat und bei einer Standardabweichung von > 2,5° bzw. % als niedrig bewertet.

Außerdem wurden Regressionsanalysen zur Beurteilung des proportionalen Fehlers durchgeführt. Unter einem proportionalen Fehler versteht man eine systematische Abweichung vom Standardwert, das heißt, die von den kommerziellen KI-Anbietern vorhergesagten Parameter sind im Durchschnitt bei kleinen Werten kleiner bzw. bei größeren Werten größer als der menschliche Goldstandard oder genau umgekehrt. Im Rahmen der Regressionsanalyse wurde die Differenz zwischen den Werten der kommerziellen KI-Anbieter und dem menschlichen Goldstandard als Kriterium (unabhängige Variable) und der menschliche Goldstandard als Prädiktor (abhängige Variable) definiert. Die aus den Regressionsanalysen errechneten Regressionsgeraden wurden zusätzlich in die entsprechenden Bland-Altman-Plots eingefügt, um mögliche proportionale Fehler visuell darzustellen.

Für alle aufgeführten statistischen Analysen wurde das Signifikanzniveau auf 5% festgelegt.

3 Ergebnisse

3.1 Reliabilität des Goldstandards

Für alle Untersucher und Parameter konnte eine hohe Intraraterreliabilität nachgewiesen werden (alle *ICC* > .800 mit $p < .001$). Auch die Interraterreliabilität war für alle untersuchten Parameter hoch (alle *ICC* > .900 mit $p < .001$).

3.2 Ergebnisse der ANOVA mit Messwiederholung und der paarweisen Vergleiche

Tabelle 3 stellt die Ergebnisse der ANOVA mit Messwiederholung dar, Tabelle 4 veranschaulicht die Ergebnisse der paarweisen Vergleiche mittels Post-hoc-Test.

Anhand der ANOVA mit Messwiederholung war festzustellen, dass für alle neun untersuchten Parameter statistisch signifikante Unterschiede zwischen den fünf Auswertungen vorlagen.

Die paarweisen Vergleiche zum Goldstandard zeigen, dass es zwischen den vier kommerziellen Anbietern große Unterschiede hinsichtlich der durchschnittlichen Richtigkeit der Analysen gab.

Tabelle 3: Vergleich der Vorhersagen der KI-basierten Auswertungen der kommerziellen Anbieter und des menschlichen Goldstandards. Deskriptive Statistik mit Mittelwert (*M*) und Standardabweichung (*SD*). ANOVA mit Messwiederholung mit *p*-Wert (Greenhouse-Geisser).

	Parameter	Einheit	Auswertung	*M*	*SD*	*p*
Skelettal sagittale Analyse	SNA	°	Menschlicher Goldstandard	81,23	2,61	.000**
			DentaliQ.ortho	81,12	2,54	
			WebCeph	81,29	2,81	
			AudaxCeph	82,59	2,70	
			CephX	81,58	2,58	
	SNB	°	Menschlicher Goldstandard	78,56	2,86	.000**
			DentaliQ.ortho	78,60	3,02	
			WebCeph	78,36	3,09	
			AudaxCeph	79,67	2,87	
			CephX	78,64	2,94	
	ANB	°	Menschlicher Goldstandard	2,65	2,26	.002**
			DentaliQ.ortho	2,52	2,07	
			WebCeph	2,93	1,82	
			AudaxCeph	2,94	2,10	
			CephX	2,95	2,10	
Skelettal vertikale Analyse	SN_SpP	°	Menschlicher Goldstandard	7,33	2,78	.000**
			DentaliQ.ortho	7,39	2,70	
			WebCeph	8,02	1,86	
			AudaxCeph	5,67	2,79	
			CephX	7,76	2,54	
	SN_MeGo	°	Menschlicher Goldstandard	30,62	7,07	.000**
			DentaliQ.ortho	30,60	7,20	
			WebCeph	31,68	6,45	
			AudaxCeph	29,97	7,02	
			CephX	35,15	6,27	
	SpP_MeGo	°	Menschlicher Goldstandard	23,38	6,34	.000**
			DentaliQ.ortho	23,21	6,36	
			WebCeph	23,66	5,46	
			AudaxCeph	24,30	6,12	
			CephX	27,38	4,92	
	Gesichts-höhen-verhältnis	%	Menschlicher Goldstandard	68,15	6,29	.000**
			DentaliQ.ortho	68,21	6,35	
			WebCeph	66,94	5,43	
			AudaxCeph	69,65	6,10	
			CephX	63,08	4,94	
Dentale Analyse	O1_SN	°	Menschlicher Goldstandard	103,96	6,74	.013*
			DentaliQ.ortho	103,87	6,04	
			WebCeph	103,04	4,80	
			AudaxCeph	103,64	5,08	
			CephX	104,96	5,89	
	U1_MeGo	°	Menschlicher Goldstandard	94,15	7,34	.000**
			DentaliQ.ortho	93,82	6,55	
			WebCeph	93,81	4,65	
			AudaxCeph	94,65	6,04	
			CephX	88,19	5,52	

***Signifikanz für $p < .05$ / **Signifikanz für $p < .01$**

Tabelle 4: Vergleich zwischen Vorhersage der kommerziellen KI-Anbieter und dem menschlichen Goldstandard. Post-hoc-Analyse der ANOVA mit Messwiederholung mit mittlerer Differenz, 95%-Konfidenzintervall und *p*-Wert (*p*).

	Parameter	Einheit	Menschlicher Goldstandard vs. kommerzielle KI-Anbieter		*Mittlere Differenz*	*95% Konfidenzintervall*		*p*
						Untergrenze	*Obergrenze*	
Skelettal sagittale Analyse	SNA	°	Menschlicher Goldstandard vs.	DentaliQ.ortho	-0,11	-0,56	0,33	1.000
				WebCeph	0,06	-0,91	1,02	1.000
				AudaxCeph	1,36	0,81	1,91	.000**
				CephX	0,35	-0,19	0,89	.602
	SNB	°	Menschlicher Goldstandard vs.	DentaliQ.ortho	0,03	-0,38	0,44	1.000
				WebCeph	-0,21	-0,97	0,56	1.000
				AudaxCeph	1,10	0,62	1,59	.000**
				CephX	0,08	-0,38	0,54	1.000
	ANB	°	Menschlicher Goldstandard vs.	DentaliQ.ortho	-0,13	-0,42	0,17	1.000
				WebCeph	0,28	-0,20	0,76	.874
				AudaxCeph	0,29	0,05	0,53	.008**
				CephX	0,30	0,01	0,60	.041*
Skelettal vertikale Analyse	SN_SpP	°	Menschlicher Goldstandard vs.	DentaliQ.ortho	0,06	-0,48	0,60	1.000
				WebCeph	0,69	-0,26	1,64	.365
				AudaxCeph	-1,66	-2,28	-1,04	.000**
				CephX	0,43	-0,40	1,27	1.000
	SN_MeGo	°	Menschlicher Goldstandard vs.	DentaliQ.ortho	-0,02	-0,48	0,44	1.000
				WebCeph	1,06	-0,09	2,20	.090
				AudaxCeph	-0,66	-1,29	-0,02	.038*
				CephX	4,53	3,82	5,23	.000**
	SpP_MeGo	°	Menschlicher Goldstandard vs.	DentaliQ.ortho	-0,16	-0,70	0,37	1.000
				WebCeph	0,28	-0,86	1,42	1.000
				AudaxCeph	0,92	0,45	1,39	.000**
				CephX	4,01	3,17	4,85	.000**
	Gesichtshöhenverhältnis	%	Menschlicher Goldstandard vs.	DentaliQ.ortho	0,06	-0,33	0,45	1.000
				WebCeph	-1,20	-2,41	0,00	.050
				AudaxCeph	1,50	0,83	2,17	.000**
				CephX	-5,07	-5,89	-4,25	.000**
Dentale Analyse	O1_SN	°	Menschlicher Goldstandard vs.	DentaliQ.ortho	-0,09	-1,06	0,87	1.000
				WebCeph	-0,93	-2,96	1,11	1.000
				AudaxCeph	-0,32	-1,65	1,00	1.000
				CephX	1,00	-0,18	2,17	.163
	U1_MeGo	°	Menschlicher Goldstandard vs.	DentaliQ.ortho	-0,33	-1,35	0,70	1.000
				WebCeph	-0,33	-2,18	1,51	1.000
				AudaxCeph	0,51	-0,67	1,69	1.000
				CephX	-5,96	-7,45	-4,47	.000**

***Signifikanz für *p* < .05 / **Signifikanz für *p* < .01**

3.2.1 DentaliQ.ortho vs. menschlicher Goldstandard

Die von DentaliQ.ortho vorhergesagten Ergebnisse waren insgesamt sehr ähnlich zum menschlichen Goldstandard. Für keinen der insgesamt neun gemessenen Parameter unterschieden sich die Mittelwerte statistisch signifikant vom menschlichen Goldstandard - der *p*-Wert lag für alle Parameter bei $p = 1.000$. Die größte Übereinstimmung zwischen menschlichem Goldstandard und DentaliQ.ortho wurde für die Unterkieferneigung SN_MeGo mit einer mittleren Differenz von $\Delta = 0{,}02°$, die größte Abweichung wurde für die Inklination des unteren mittleren Frontzahnes U1_MeGo mit einer mittleren Differenz von $\Delta = 0{,}33°$ ermittelt.

3.2.2 WebCeph vs. menschlicher Goldstandard

Auch im Vergleich zwischen WebCeph und menschlichem Goldstandard waren größtenteils sehr ähnliche Ergebnisse zu verzeichnen. Für keinen Parameter wurden statistisch signifikante Unterschiede zum menschlichen Goldstandard festgestellt. Die höchste durchschnittliche Richtigkeit wurde für den SNA-Winkel mit einer mittleren Differenz von $\Delta = 0{,}06°$, die niedrigste durchschnittliche Richtigkeit für das Gesichtshöhenverhältnis mit einer mittleren Differenz von $\Delta = 1{,}2\%$ zum menschlichen Goldstandard beobachtet.

3.2.3 AudaxCeph vs. menschlicher Goldstandard

In der Gegenüberstellung zwischen AudaxCeph und menschlichem Goldstandard waren deutlichere Abweichungen zu verzeichnen. Für alle skelettal sagittalen und alle skelettal vertikalen Parameter wurden statistisch signifikante Unterschiede zum menschlichen Goldstandard ermittelt – die *p*-Werte lagen für diese Parameter bei $p < .05$. Lediglich die dentale Analyse (O1_SN und U1_MeGo) zeigte mit $p = 1.000$ keine statistisch signifikanten Unterschiede zum menschlichen Goldstandard. Die kleinste mittlere Differenz wurde mit $\Delta = 0{,}29°$ für den ANB-Winkel, die größte mittlere Differenz wurde mit $\Delta = 1{,}66°$ für den Parameter SN_SpP gemessen.

3.2.4 CephX vs. menschlicher Goldstandard

Im paarweisen Vergleich zwischen CephX und menschlichem Goldstandard zeigten fünf der untersuchten neun Parameter statistisch signifikante Abweichungen zum menschlichen Goldstandard. Dabei handelte es sich um ANB, SN_MeGo, SpP_MeGo, das Gesichtshöhenverhältnis und U1_MeGo. Für den ANB-Winkel lag der *p*-Wert bei $p = .041$, für SN_MeGo, SpP_MeGo, das Gesichtshöhenverhältnis und U1_MeGo lag

der *p*-Wert bei $p = .000$. Zudem wiesen diese vier Parameter durchgehend große absolute Abweichungen von > 4° bzw. > 5% zum menschlichen Goldstandard auf. Als am fehleranfälligsten kann somit die skelettal vertikale Analyse beschrieben werden. Die kleinste mittlere Differenz wurde mit $\Delta = 0{,}08°$ für SNB, die größte mittlere Differenz wurde mit $\Delta = 5{,}96°$ für U1_MeGo ermittelt.

3.3 Ergebnisse der Bland-Altman-Plots

Bei der folgenden Auswertung wurden die durchschnittliche Richtigkeit, die Präzision und der proportionale Fehler der untersuchten kommerziellen KI-Anbieter für jeden einzelnen Parameter bewertet. Da im Falle einer niedrigen durchschnittlichen Richtigkeit selbst bei einer hohen Präzision nicht von einer guten Genauigkeit einer Analyse ausgegangen werden kann, wurde keine dezidierte Beurteilung der Präzision bei Vorliegen einer niedrigen durchschnittlichen Richtigkeit vorgenommen.

3.3.1 Skelettal sagittale Analyse

Die Ergebnisse der Bland-Altman-Plots für die skelettal sagittale Analyse sind in Abbildung 13 bis Abbildung 15 dargestellt.

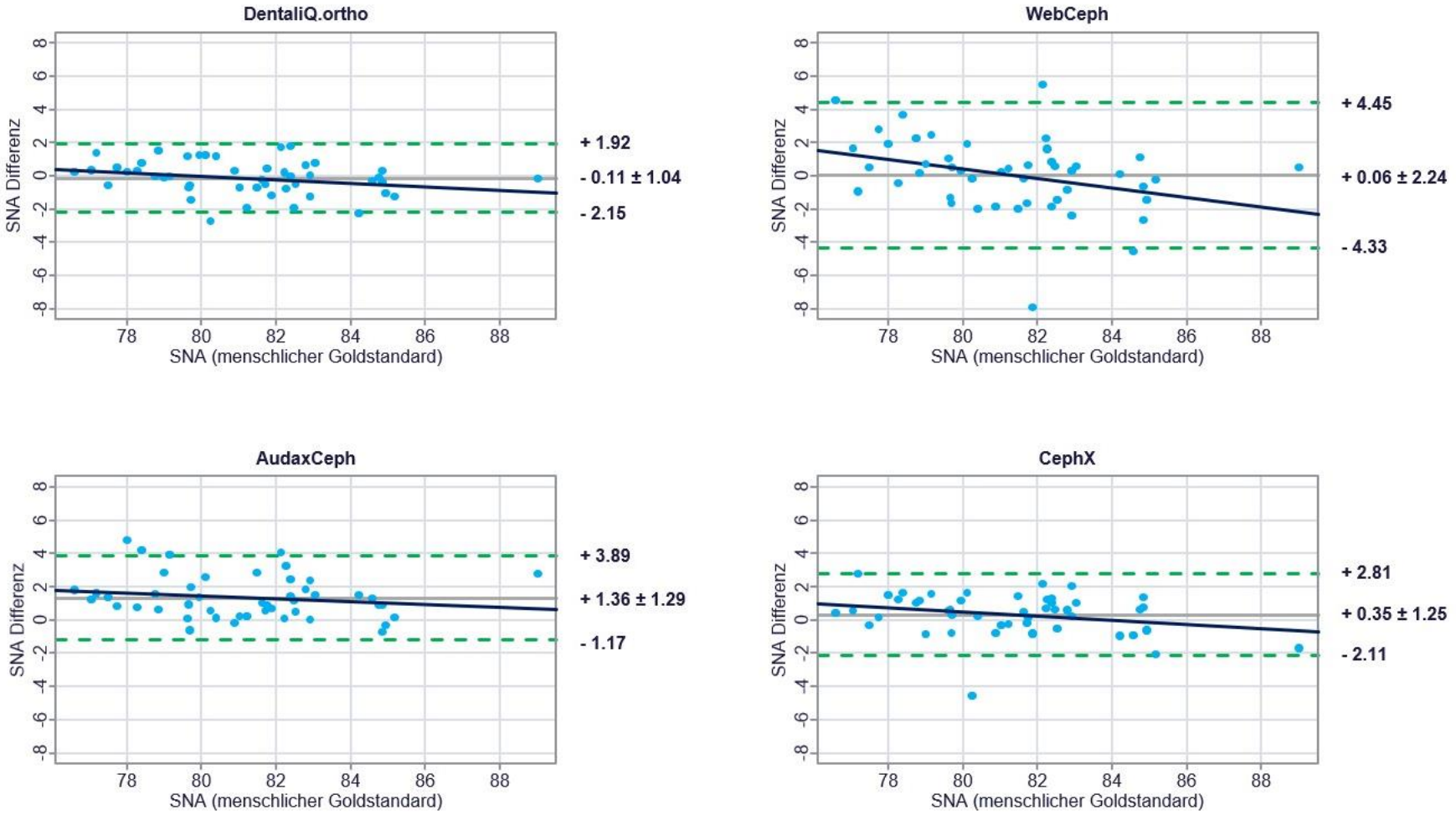


Abbildung 13: Bland-Altman-Plots für SNA.

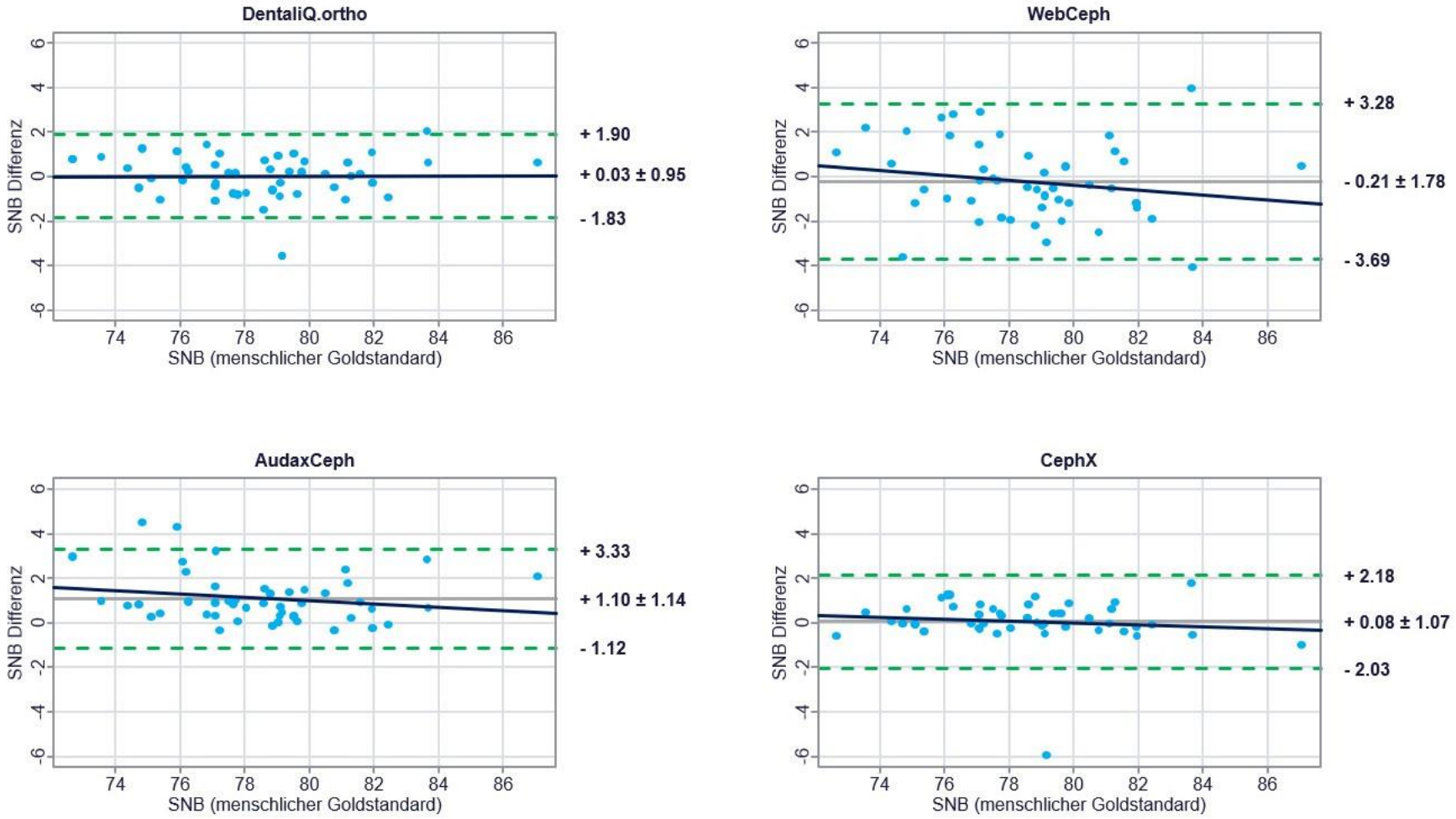


Abbildung 14: Bland-Altman-Plots für SNB.

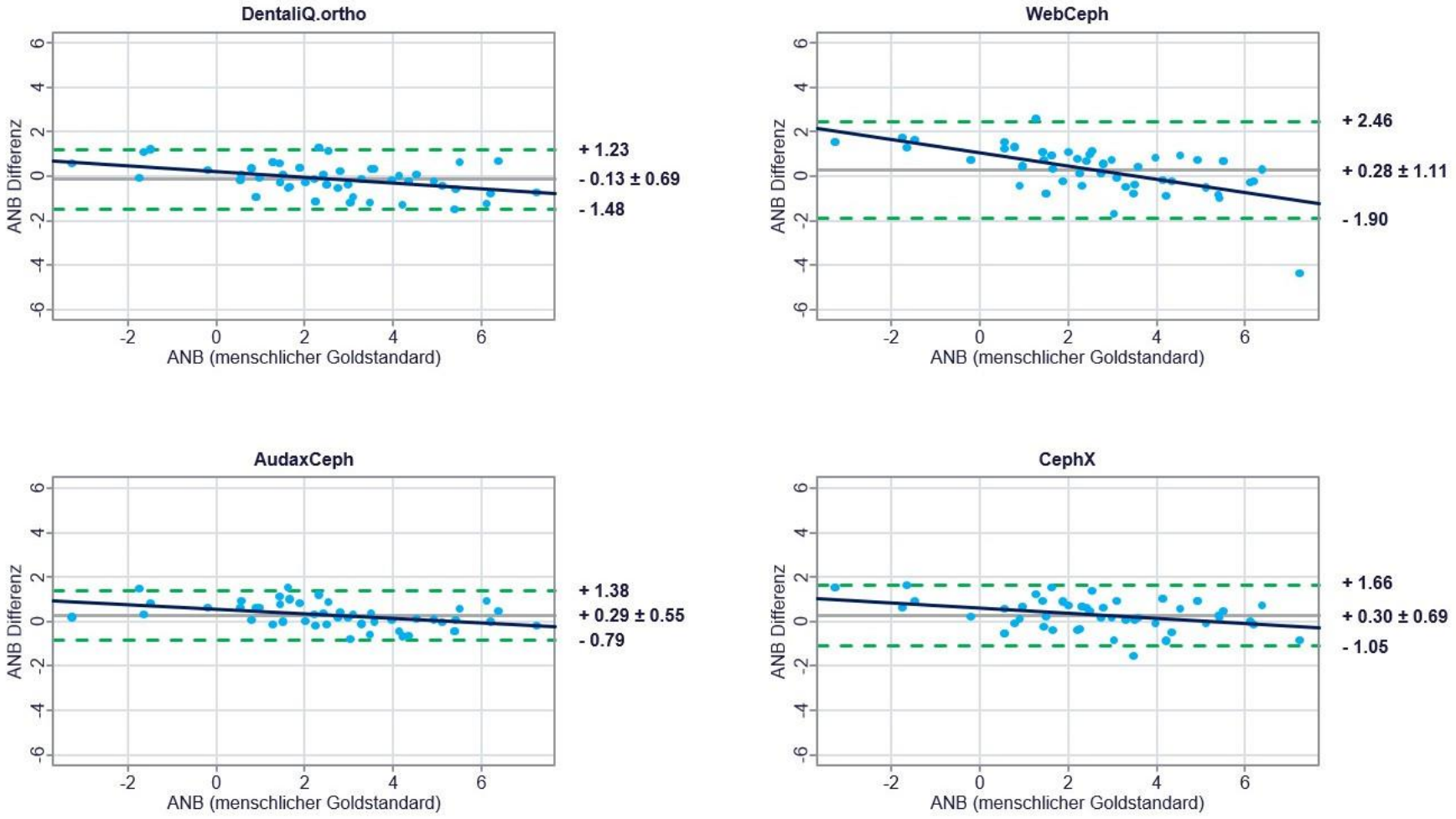


Abbildung 15: Bland-Altman-Plots für ANB.

3.3.1.1 SNA

Für den SNA-Winkel zeigte DentaliQ.ortho mit einer mittleren Differenz von Δ = 0,11° zum menschlichen Goldstandard eine hohe durchschnittliche Richtigkeit. Auch die Präzision war als hoch zu beschreiben, die LoA lagen bei -2,15° und +1,92°. Das Risiko eines proportionalen Fehlers war als gering einzustufen.

Der Anbieter Web-Ceph erreichte mit einer mittleren Differenz von Δ = 0,06° zum Goldstandard die im Vergleich zu den anderen kommerziellen KI-Anbietern höchste durchschnittliche Richtigkeit, allerdings war die Präzision lediglich als moderat zu beschreiben, die LoA lagen hier bei -4,33° und +4,45°. Zudem war der proportionale Fehler der Analyse von WebCeph vergleichsweise am höchsten.

Die durchschnittliche Richtigkeit der Analyse von AudaxCeph war mit einer mittleren Differenz von Δ = 1,36° zum menschlichen Goldstandard, also mit einer Abweichung von mehr als einem Grad im Vergleich zu den anderen drei Anbietern, als niedrig zu bewerten.

CephX zeigte mit einer mittleren Differenz von Δ = 0,35° zum Goldstandard analog zu den Anbietern DentaliQ.ortho und WebCeph eine hohe durchschnittliche Richtigkeit. Die Präzision der Analyse konnte mit LoA bei -2,11° und +2,81° ebenfalls als hoch beschrieben werden. Das Risiko eines proportionalen Fehlers war gering.

3.3.1.2 SNB

Ähnlich dem SNA-Winkel zeigte DentaliQ.ortho mit der vergleichsweise geringsten mittleren Differenz von Δ = 0,03° zum menschlichen Goldstandard auch für den SNB-Winkel eine hohe durchschnittliche Richtigkeit. Die Präzision war im Vergleich zu den anderen drei untersuchten kommerziellen KI-Anbietern am höchsten, die LoA lagen bei -1,83° und +1,90°. Zudem war von nahezu keinem proportionalen Fehler auszugehen.

Die durchschnittliche Richtigkeit der Analyse des Anbieters WebCeph war mit einer mittleren Differenz von Δ = 0,21° zum Goldstandard ebenfalls als hoch zu bewerten. Die Präzision hingegen kann mit LoA bei -3,69° und +3,28° als moderat beschrieben werden. Analog zum SNA-Winkel war der proportionale Fehler der Analyse von WebCeph vergleichsweise am höchsten.

Der Anbieter AudaxCeph erreichte wie auch beim Parameter SNA mit einer mittleren Differenz von $\Delta = 1{,}10°$ zum menschlichen Goldstandard lediglich eine niedrige durchschnittliche Richtigkeit.

Die durchschnittliche Richtigkeit von CephX konnte als hoch beschrieben werden, es wurde eine mittlere Differenz von $\Delta = 0.08°$ zum Goldstandard dokumentiert. Auch die Präzision war hoch, die LoA lagen bei -2,03° und +2,18°. Zudem zeigt der entsprechende Bland-Altman-Plot einen geringen proportionalen Fehler.

3.3.1.3 ANB

Für den ANB-Winkel zeigten alle Anbieter sehr geringe mittlere Differenzen zum menschlichen Goldstandard, die kleinste mittlere Differenz lag für DentaliQ.ortho bei $\Delta = 0{,}13°$, die größte mittlere Differenz für CephX bei $\Delta = 0{,}30°$. Bei allen kommerziellen KI-Anbietern konnte somit von einer hohen durchschnittlichen Richtigkeit ausgegangen werden.

Auch die Präzision war für den ANB-Winkel durchweg hoch, die höchste Präzision konnte für AudaxCeph mit LoA von -0,79° und +1,38° ermittelt werden.

Der Anbieter WebCeph zeigte, vergleichbar mit der Analyse des SNA- und SNB-Winkels, den höchsten proportionalen Fehler.

3.3.2 Skelettal vertikale Analyse

Die Ergebnisse der Bland-Altman-Plots für die skelettal vertikale Analyse sind in Abbildung 16 bis Abbildung 19 dargestellt.

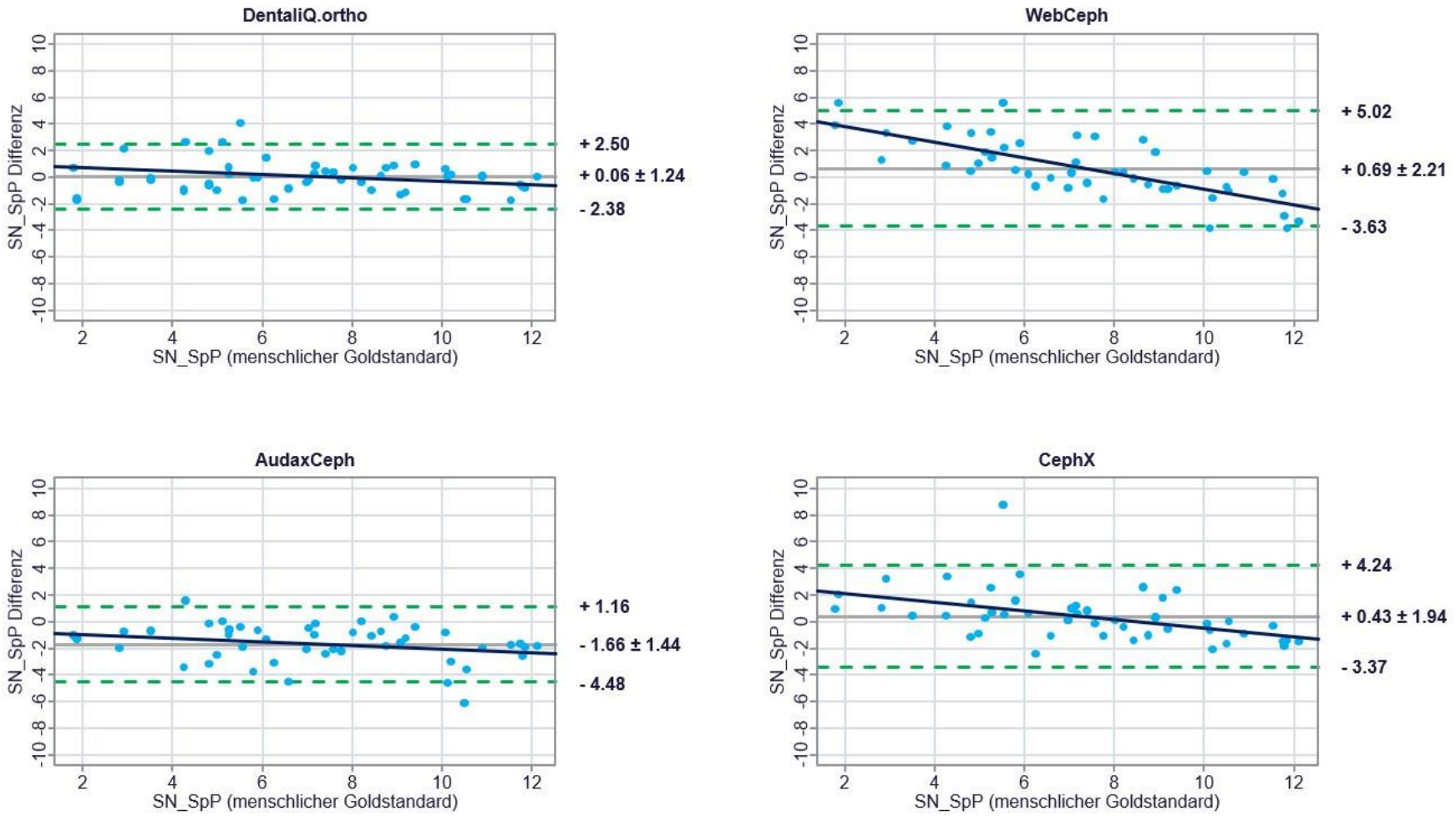


Abbildung 16: Bland-Altman-Plots für SN_SpP.

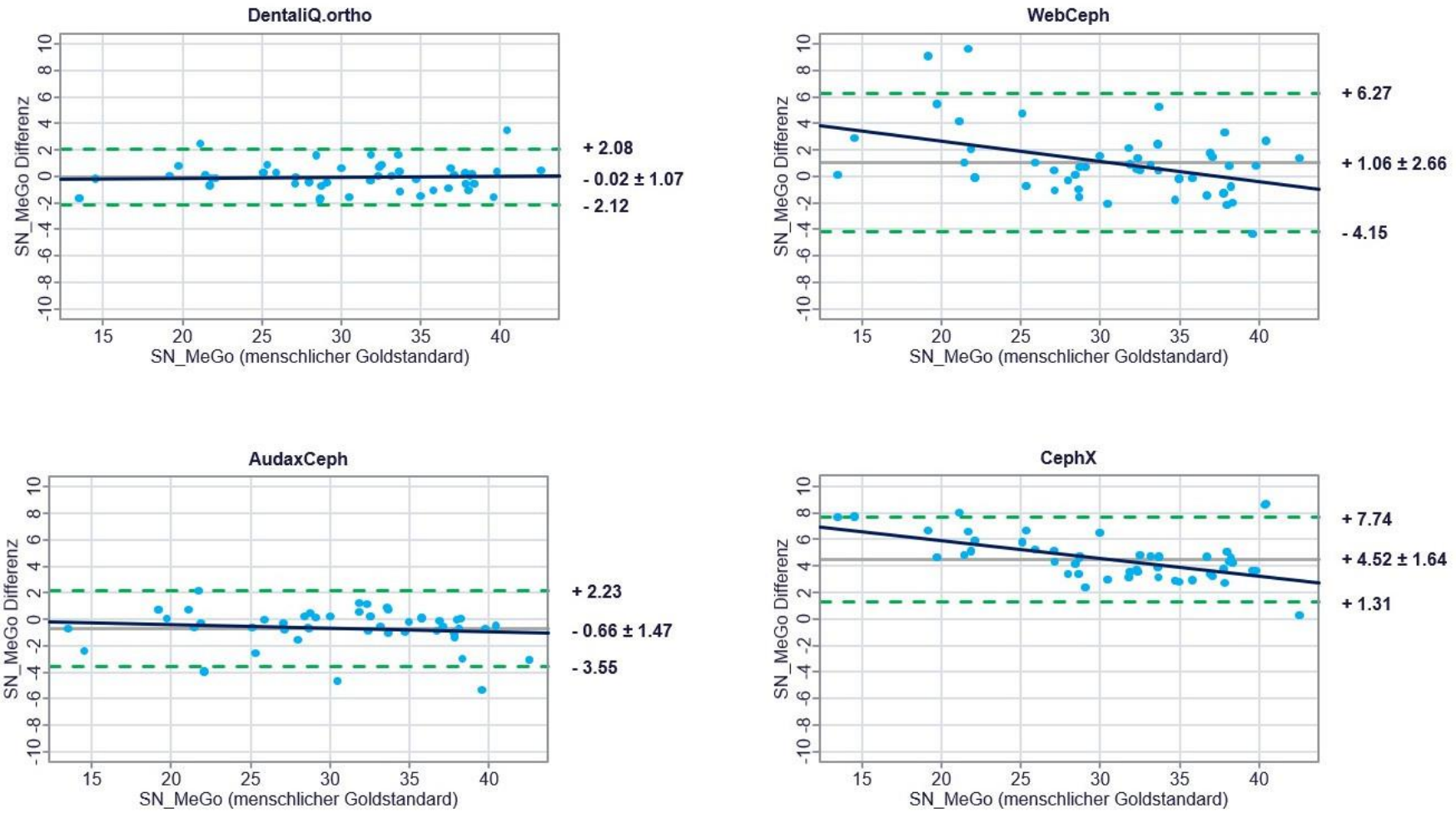


Abbildung 17: Bland-Altman-Plots für SN_MeGo.

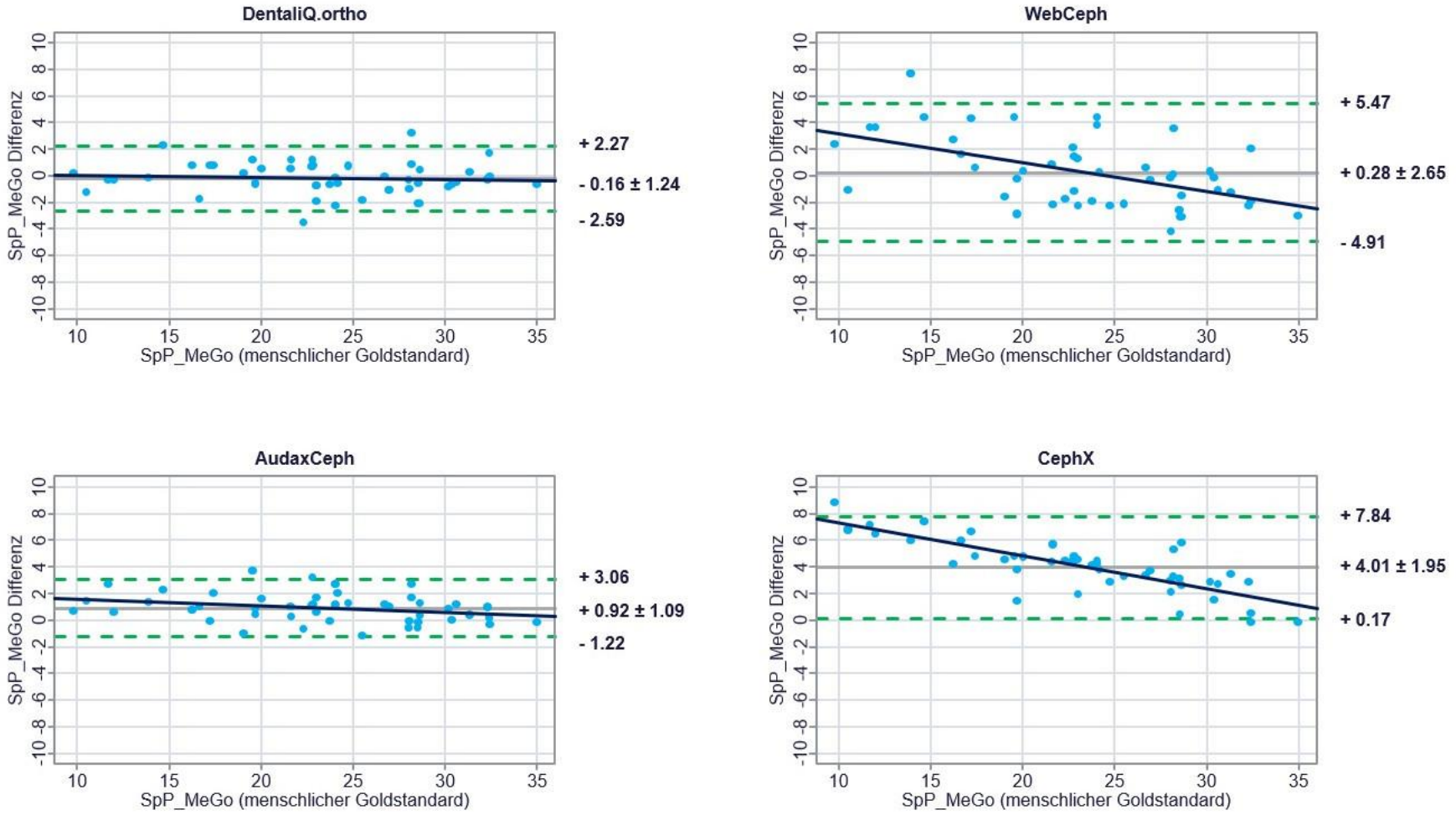


Abbildung 18: Bland-Altman-Plots für SpP_MeGo.

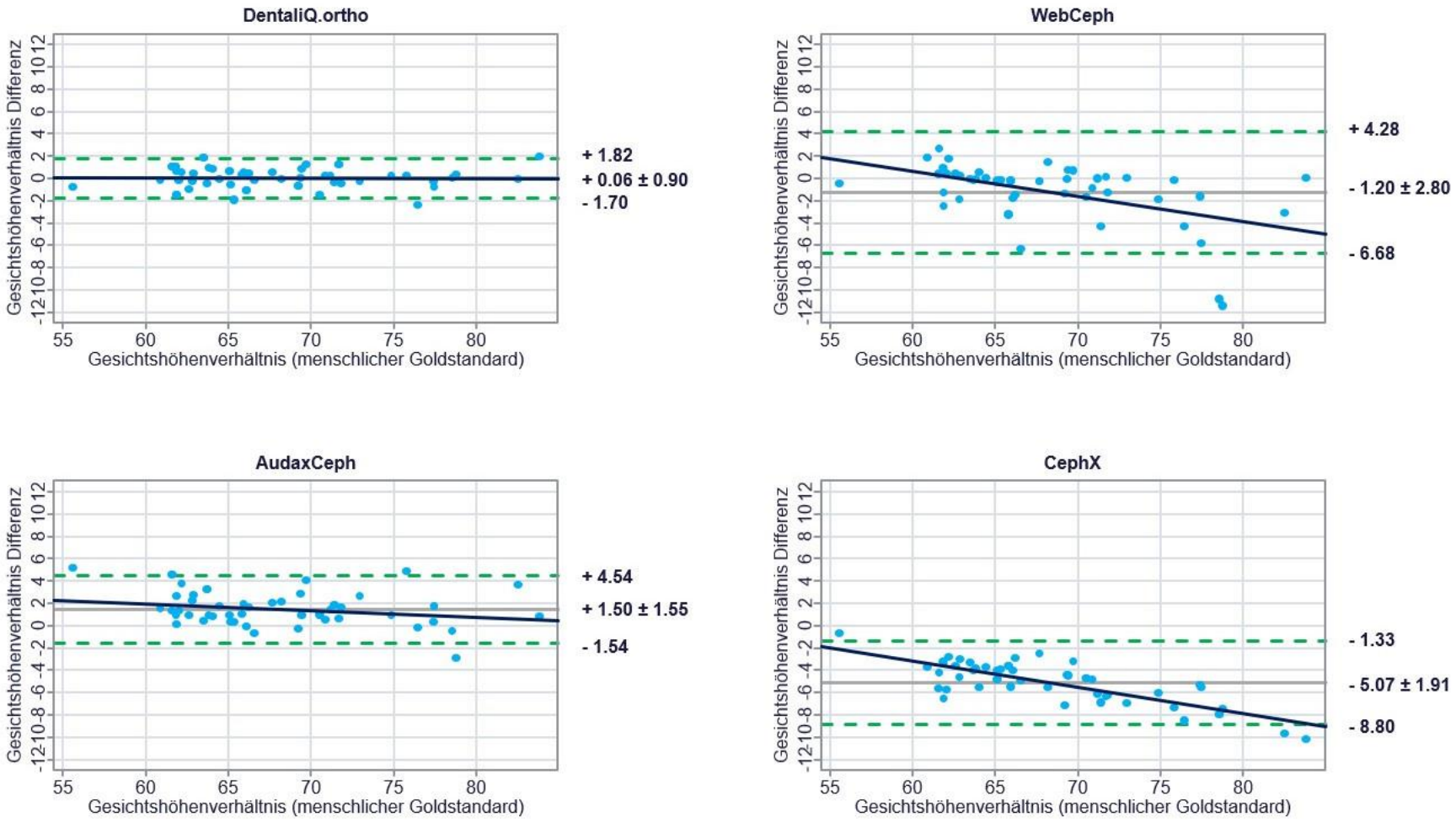


Abbildung 19: Bland-Altman-Plots für Gesichtshöhenverhältnis.

3.3.2.1 SN_SpP

Bezüglich des Winkels SN_SpP zeigte DentaliQ.ortho eine hohe durchschnittliche Richtigkeit mit einer mittleren Differenz zum menschlichen Goldstandard von Δ = 0,06°. Die Präzision war ebenfalls als hoch zu beschreiben, die LoA lagen bei -2,38° und +2,50°. Zudem war das Risiko eines proportionalen Fehlers als gering einzustufen.

Der Anbieter WebCeph erzielte mit einer mittleren Differenz von Δ = 0,69° zum Goldstandard eine moderate durchschnittliche Richtigkeit. Auch die Präzision lag im moderaten Bereich. Außerdem zeigte der Bland-Altman-Plot ein deutlich erhöhtes Risiko eines proportionalen Fehlers.

Übereinstimmend mit dem SNA- und SNB-Winkel erreichte die Analyse von AudaxCeph für den Winkel SN_SpP eine niedrige durchschnittliche Richtigkeit. Es wurde eine mittlere Differenz von Δ = 1,66° zum menschlichen Goldstandard dokumentiert.

Die durchschnittliche Richtigkeit der Analyse von CephX war als hoch zu bewerten, die mittlere Differenz zum Goldstandard lag bei Δ = 0,43°. Mit LoA bei -3,37° und +4,24° lag die Präzision im mittleren Bereich. Im Vergleich zu DentaliQ.ortho und AudaxCeph wurde ein erhöhtes Risiko eines proportionalen Fehlers dokumentiert.

3.3.2.2 SN_MeGo

Für den Winkel SN_MeGo zeigte DentaliQ.ortho eine hohe durchschnittliche Richtigkeit mit einer mittleren Differenz zum menschlichen Goldstandard von Δ = 0,02°. Die Präzision war ebenfalls als hoch zu bewerten, die LoA lagen bei -2,12° und +2,08°. Es war nahezu kein proportionaler Fehler vorhanden.

Der Anbieter WebCeph erzielte mit einer mittleren Differenz von Δ = 1,06° zum Goldstandard lediglich eine niedrige durchschnittliche Richtigkeit.

Die durchschnittliche Richtigkeit von AudaxCeph war mit einer mittleren Differenz von Δ = 0.66° zum menschlichen Goldstandard als moderat zu beschreiben. Die LoA lagen bei -3,55° und +2,23°, sodass von einer hohen Präzision ausgegangen werden konnte. Auch der proportionale Fehler war als gering einzustufen.

Für den Winkel SN_MeGo erzielte der Anbieter CephX lediglich eine niedrige durchschnittliche Richtigkeit, es wurde eine mittlere Differenz von Δ = 4,52° zum Goldstandard dokumentiert.

3.3.2.3 SpP_MeGo

Der Anbieter DentaliQ.ortho erzielte für den Winkel SpP_MeGo mit einer mittleren Differenz von Δ = 0,16° zum menschlichen Goldstandard die höchste durchschnittliche Richtigkeit. Auch die Präzision war als hoch zu bewerten, die LoA lagen bei -2,59° und +2,27°. Das Risiko eines proportionalen Fehlers war sehr gering.

Die durchschnittliche Richtigkeit der Analyse von WebCeph kann mit einer mittleren Differenz von Δ = 0.28° zum Goldstandard ebenfalls als hoch beschrieben werden. Allerdings war die Präzision als niedrig einzustufen und das Risiko eines proportionalen Fehlers deutlich erhöht.

Die Analyse von AudaxCeph erreichte mit einer mittleren Differenz von Δ = 0,92° zum menschlichen Goldstandard eine moderate durchschnittliche Richtigkeit. Die Präzision war als hoch, das Risiko eines proportionalen Fehlers als gering zu beschreiben.

Analog zum Winkel SN_MeGo war die durchschnittliche Richtigkeit der Analyse von CephX als niedrig zu bewerten, da eine große mittlere Differenz von Δ = 4,01° zum menschlichen Goldstandard dokumentiert wurde.

3.3.2.4 Gesichtshöhenverhältnis

Bei der Bestimmung des Gesichtshöhenverhältnisses erzielte DentaliQ.ortho mit einer mittleren Differenz von Δ = 0,06% zum menschlichen Goldstandard eine hohe durchschnittliche Richtigkeit. Die Präzision war mit LoA bei -1,70% und +1,82% ebenfalls als hoch zu bewerten. Der zugehörige Bland-Altman-Plot zeigt nahezu keinen proportionalen Fehler.

Die Analysen von WebCeph, AudaxCeph und CephX waren hinsichtlich ihrer durchschnittlichen Richtigkeit als niedrig zu bewerten, es wurden mittlere Differenzen von Δ = 1,20%, Δ = 1,50% und Δ = 5,07% zum menschlichen Goldstandard dokumentiert.

3.3.3 Dentale Analyse

Die Ergebnisse der Bland-Altman-Plots für die dentale Analyse sind in Abbildung 20 und Abbildung 21 dargestellt.

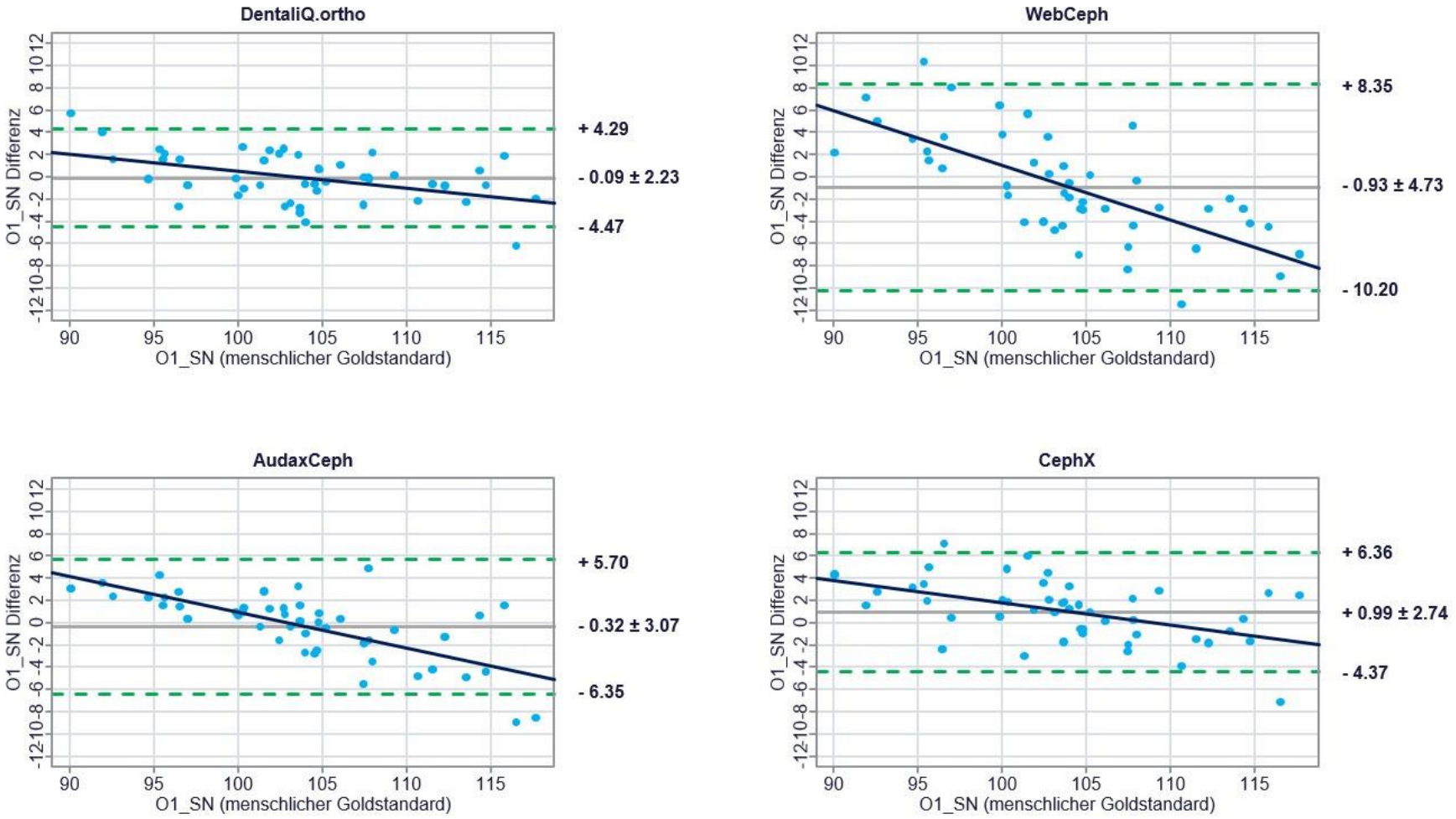


Abbildung 20: Bland-Altman-Plots für O1_SN.

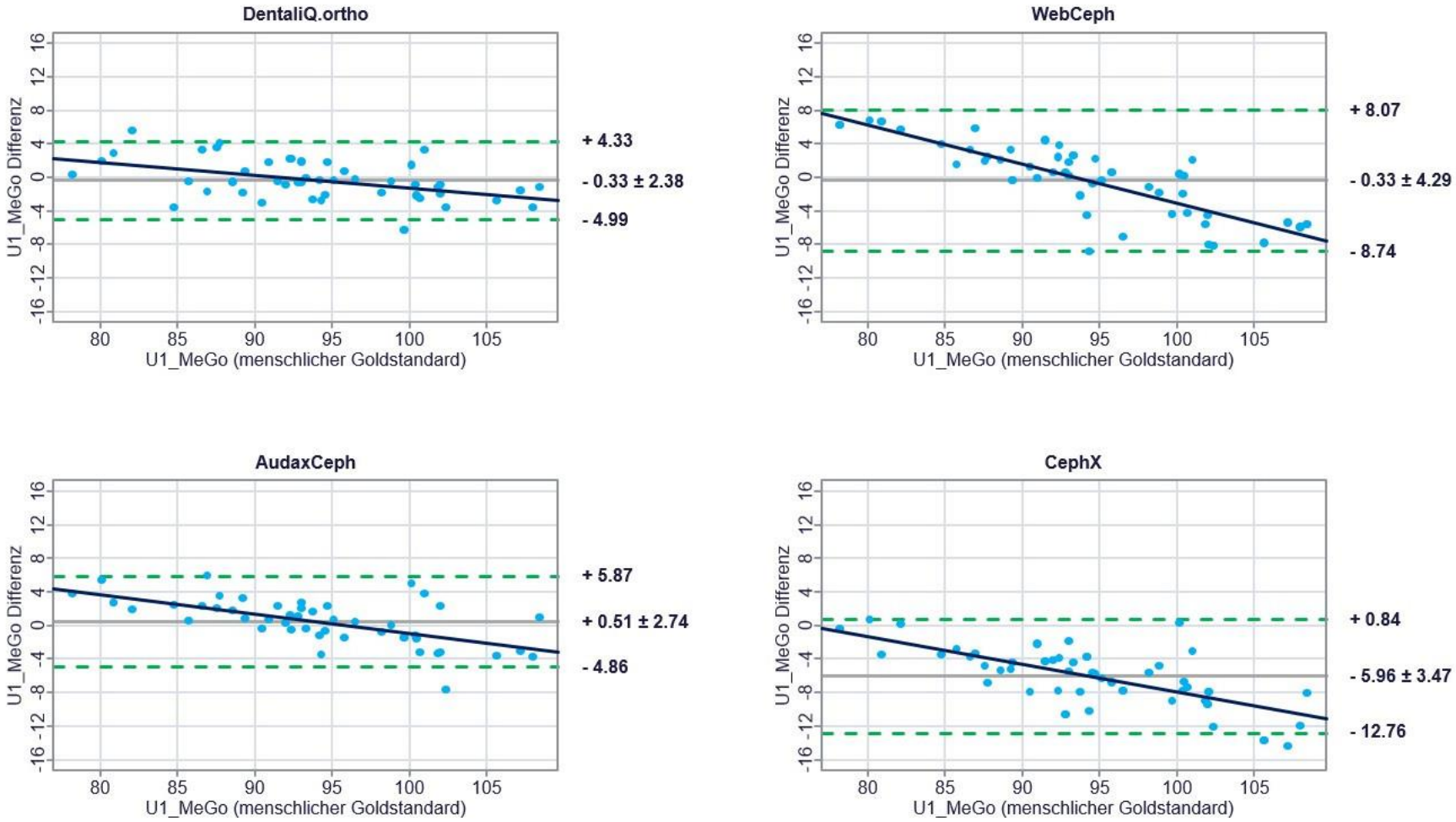


Abbildung 21: Bland-Altman-Plots für U1_MeGo.

3.3.3.1 O1_SN

Für den Winkel O1_SN erreichte DentaliQ.ortho mit einer mittleren Differenz von $\Delta = 0{,}09°$ zum menschlichen Goldstandard eine hohe durchschnittliche Richtigkeit. Die Präzision war mit LoA bei -4,47° und +4,29° als moderat zu beschrieben. Der entsprechende Bland-Altman-Plot zeigte einen moderat ausgeprägten proportionalen Fehler.

Die durchschnittliche Richtigkeit der Analyse von WebCeph war mit einer mittleren Differenz von $\Delta = 0{,}93°$ zum menschlichen Goldstandard als moderat einzustufen. Die Präzision war mit LoA bei -10,20° und +8,35° allerdings sehr niedrig und das Risiko eines proportionalen Fehlers sehr hoch.

Die Analyse von AudaxCeph erzielte zwar eine hohe durchschnittliche Richtigkeit, allerdings war die Präzision mit LoA bei -6,35° und +5,70° analog zu WebCeph als niedrig zu bewerten. Auch für diesen Anbieter wurde ein hohes Risiko eines proportionalen Fehlers dokumentiert.

Die durchschnittliche Richtigkeit des Anbieters CephX konnte mit einer mittleren Differenz von $\Delta = 0{,}99°$ zum menschlichen Goldstandard als moderat beschrieben werden. Die Präzision war, vergleichbar mit WebCeph und AudaxCeph, erneut niedrig. Der Bland-Altman-Plot zeigt auch für CephX einen erhöhten proportionalen Fehler.

3.3.3.2 U1_MeGo

Bezüglich des Winkels U1_MeGo konnte eine hohe durchschnittliche Richtigkeit für DentaliQ.ortho mit einer mittleren Differenz zum menschlichen Goldstandard von $\Delta = 0{,}33°$ ermittelt werden. Die Präzision war mit LoA bei -4,99° und +4,33° als moderat zu beschreiben. Das Risiko eines proportionalen Fehlers war als moderat einzustufen.

Der Anbieter WebCeph erreichte ebenfalls eine hohe durchschnittliche Richtigkeit. Allerdings war die Präzision mit LoA bei -8,74° und +8,07° sehr niedrig und es wurde ein hohes Risiko eines proportionalen Fehlers dokumentiert.

AudaxCeph erzielte eine moderate durchschnittliche Richtigkeit. Analog zu WebCeph war die Präzision der Analyse allerdings niedrig. Auch hier zeigte der Bland-Altman-Plot ein erhöhtes Risiko eines proportionalen Fehlers.

Für den Anbieter CephX wurde mit einer mittleren Differenz von $\Delta = 5{,}96°$ zum menschlichen Goldstandard eine niedrige durchschnittliche Richtigkeit dokumentiert.

4 Diskussion

Innerhalb der letzten Jahre schritt die Entwicklung von KI im Bereich der Kieferorthopädie rasch voran, wobei sich die meisten Publikationen in der Literatur zum Thema KI-gestützte FRS-Analyse finden lassen [32, 79, 82]. Für Kieferorthopäden ist dies von großem Interesse, da die FRS-Analyse zur kieferorthopädischen Routinediagnostik zählt und somit einen nicht unerheblichen Teil des klinischen Alltags eines Kieferorthopäden ausmacht [2, 93]. Ursprünglich war die manuelle Auswertung von FRS ein sehr zeitintensiver und aufwendiger Prozess [16]. KI-Algorithmen hingegen sind in der Lage, FRS innerhalb von Sekunden auszuwerten. Allerdings sollte dabei beachtet werden, dass bei der Nutzung von KI fehlerhafte FRS-Analysen strengstens zu vermeiden sind, da diese in weiterer Folge möglicherweise zu Fehlern in der kieferorthopädischen Behandlungsplanung führen können [115].

In den bisherigen Studien wurde vornehmlich die metrische Genauigkeit der Punktplatzierung durch KI untersucht [79, 81, 116-118]. Vergleiche der Genauigkeiten der kephalometrischen Parameter an sich und damit der gesamten FRS-Analyse sind in der aktuell verfügbaren Literatur deutlich seltener [115, 119-121]. Zudem sind bereits einige KI zur FRS-Analyse kommerziell verfügbar, ohne dass ausreichend fundierte wissenschaftliche Daten zur klinischen Anwendbarkeit zugrunde liegen [79].

Daher war das Ziel der vorliegenden Arbeit, die aktuell vorhandenen kommerziellen KI-Anbieter anhand von kieferorthopädischen Parametern bezüglich der Genauigkeit der automatisierten FRS-Analysen zu bewerten, indem die Ergebnisse der automatischen Auswertungen mit einem qualitativ hochwertigen menschlichen Goldstandard verglichen wurden. In diesem Zuge sollten die eingangs beschriebenen Nullhypothesen, dass keine statistisch signifikanten Unterschiede zwischen der FRS-Auswertung der jeweiligen kommerziellen KI-Anbieter und dem menschlichen Goldstandard bestehen, entweder akzeptiert oder verworfen werden.

Die Ergebnisse der vorliegenden Arbeit zeigen, dass deutliche Unterschiede bezüglich der Analysequalität der verschiedenen kommerziellen Anbieter für KI-unterstützte FRS-Auswertungen bestehen.

4.1 Diskussion der Methodik

4.1.1 Auswahl der Patienten und FRS

Um die Patienten keiner zusätzlichen Röntgenstrahlung auszusetzen, wurde für die vorliegende Arbeit auf Röntgenbilder aus einem bereits vorhandenen Datenpool zurückgegriffen. Dabei handelte es sich um eine Gesamtheit von 3000 FRS, die alle aus einer privaten kieferorthopädischen Praxis stammen und im Rahmen der kieferorthopädische Routinediagnostik mit dem gleichen Gerät (Orthophos-XG-5, Sirona Dental Systems GmbH, Bensheim, Deutschland) aufgenommen wurden. Das Orthophos-XG-Gerät gilt als gebräuchlich und wird in der Zahnmedizin häufig eingesetzt [122-124].

Aus den insgesamt 3000 FRS wurden zufällig 50 FRS für die Studie ausgewählt. Aufgrund von Artefakten und zu starken Überlagerungen mussten drei FRS ausgeschlossen werden, sodass schlussendlich 47 FRS Grundlage für die statistische Analyse waren. Dabei orientierte sich die Anzahl der analysierten FRS an der Studie von Kunz et al. [32]. Durch die zufällige Auswahl der FRS wurde keine Selektion der Patienten bezüglich Faktoren wie Alter, Geschlecht, skelettaler Klasse oder Wachstumsmuster vorgenommen. Die FRS zeigten Patienten in verschiedenen Gebissphasen, mit und ohne festsitzende Apparaturen und mit und ohne konservierende und prothetische Versorgungen. Auf diese Art und Weise sollte ein möglichst heterogener Patientenpool mit großer Variabilität repräsentiert werden [125].

4.1.2 Beurteilung der FRS-Analyse

Im Rahmen der vorliegenden Studie wurden für die FRS-Analyse in Anlehnung an die Analyse nach Rakosi 15 gebräuchliche Punkte und neun gängige Parameter (acht anguläre Parameter, ein Streckenverhältnis) verwendet [12]. Die Parameter konnten in drei Kategorien, nämlich in die skelettal sagittale, die skelettal vertikale und die dentale Analyse, gruppiert werden. Damit kann die in der vorliegenden Studie verwendete FRS-Analyse als sehr umfassend beschrieben werden. Auf metrische Parameter wurde bewusst verzichtet, da der Anbieter AudaxCeph zum Zeitpunkt der Datenerhebung nicht in der Lage war, die Metrik anhand der im FRS abgebildeten Metallskalierung selbst festzulegen. Ein händisches Setzen der Punkte zur Festlegung der Metrik hätte einerseits zu Ungenauigkeiten geführt, andererseits sollte in der vorliegenden Studie die vollständig automatisierte FRS-Analyse ohne Eingreifen des Menschen im Mittelpunkt stehen. Zudem waren die in den Analysen angebotenen metrischen Parameter bei

DentaliQ.ortho, WebCeph und CephX nicht einheitlich. Die Parameter O1_NPog und U1_NPog waren zwar in der Analyse von DentaliQ.ortho enthalten, nicht aber bei CephX. Der Anbieter WebCeph verzichtete auf den Parameter U1_NPog, sodass schlussendlich keine Vergleichbarkeit der vier kommerziellen KI-Anbieter hinsichtlich metrischer Parameter gegeben war.

Um systematische Fehler zu vermeiden, wurde bei der kephalometrischen Analyse darauf geachtet, dass eine einheitliche Definition aller Punkte, Strecken und Winkel bei allen kommerziellen KI-Anbietern und dem menschlichen Goldstandard besteht. Dies ist insbesondere vor dem Hintergrund, dass für einzelne Punkte und Parameter abweichende Definitionen in der Literatur beschrieben werden, von großer Wichtigkeit.

Beispielsweise kann der Punkt Gonion unterschiedlich definiert werden. Eine Möglichkeit ist, wie in der vorliegenden Arbeit, Gonion als Schnittpunkt zwischen Mandibularplanum (Me-P_1) und der Tangente am Ramus ascendens mandibulae (Ar-P_2) zu definieren. Dabei handelt es sich um eine konstruierte Landmarke, so wie von Ricketts definiert [126]. Eine andere Möglichkeit ist, den Punkt Gonion über den Schnittpunkt der Winkelhalbierenden des Winkels zwischen Mandibularplanum und Tangente an den Ramus ascendens mandibulae mit dem äußeren Rand der Mandibula zu konstruieren. Diese Möglichkeit wird vom American Board of Orthodontics favorisiert [127].

Auch das Mandibularplanum kann abweichend definiert werden. Man unterscheidet die Definition des Unterkieferplanums als Menton-Gonion-Linie beispielsweise nach Sassouni, wie bei der vorliegenden Arbeit verwendet, von der Definition des Unterkieferplanums als Gnathion-Gonion-Linie nach Ricketts [126, 128].

Das Gesichtshöhenverhältnis nach Jarabak, wie in der vorliegenden Studie verwendet, wird als Verhältnis der posterioren (= Sella-Gonion-Linie) und der anterioren (= Nasion-Menton-Linie) Gesichtshöhe zueinander definiert [15].

Bei der Software CephX war im Hinblick auf Gonion zunächst nicht ersichtlich, wie dieser Punkt in den verschiedenen Analysen genau definiert wurde. Zudem wurden in den standardmäßig angebotenen Analysen das Mandibularplanum und das Gesichtshöhenverhältnis nicht übereinstimmend mit unserer Studie definiert. Konkret wurde das Mandibularplanum anders als in der vorliegenden Arbeit häufig als Gnathion-Gonion-Linie beschrieben. Das Gesichtshöhenverhältnis wurde in der Mehrzahl der Standardanalysen zudem nicht über die Nasion-Menton-, sondern über die Nasion-Gnathion-Linie definiert, sodass dementsprechend zunächst von abweichenden

Ergebnissen ausgegangen werden musste. Aufgrund dieser Diskrepanzen wurde von der Firma Orca Dental AI eine für diese Studie individualisierte FRS-Analyse erstellt. Auf Nachfrage wurde uns zweimal bestätigt, dass in der individualisierten Analyse die Definitionen aller verwendeten Parameter, insbesondere auch des Gonions, des Mandibularplanums und des Gesichtshöhenverhältnisses, mit der in der Studie verwendeten FRS-Analyse übereinstimmen.

4.1.3 Beurteilung des menschlichen Goldstandards

Zur Beurteilung der Analysequalität der verschiedenen Anbieter wurden die Ergebnisse der KI-basierten FRS-Auswertung mit den durch den Menschen gemessenen Werten verglichen. Diese werden in der Literatur bei der Untersuchung von KI-basierter Röntgenbildauswertung gegenwärtig als Goldstandard verwendet [32, 118, 129]. Allerdings unterliegt auch die Auswertung durch menschliche Experten intra- und interindividuellen Schwankungen, sodass einzelne Ausreißer oder Fehlplatzierungen von Punkten sich nicht gänzlich ausschließen lassen [19, 130, 131].

Um dennoch ein hohes Qualitätsniveau für den Goldstandard zu erzielen, wurden alle 50 für die Studie ausgewählten FRS von jeweils zwölf erfahrenen Untersuchern der Poliklinik für Kieferorthopädie des Universitätsklinikums Würzburg ausgewertet. Die Messergebnisse von mehr als zwei Untersuchern für die Festlegung des menschlichen Goldstandards zu verwenden, beschreiben auch Kamoen et al. in ihrer Studie als sinnvoll [19]. Der Medianwert aller zwölf Untersucher wurde für den jeweiligen kephalometrischen Parameter als Goldstandard definiert. Der Vorteil bei der Verwendung des Medianwertes für den Goldstandard liegt darin, dass Ausreißer unter den menschlichen Analysen nach oben oder unten sicher ausgeschlossen werden können [32].

In einer Vielzahl von Publikationen zur Beurteilung der Auswertequalität von KI wurde die menschliche Auswertung zwar auch als Goldstandard festgelegt, allerdings wurde die Auswertung häufig nur von einer einzelnen Person durchgeführt [115, 121, 132-134]. Dadurch können systematische Fehler und ein personenbezogener Bias nicht ausgeschlossen werden.

Zudem wurde die Analyse in der vorliegenden Studie zeitlich versetzt, das heißt in zwei Etappen, durchgeführt, um tagesformabhängige Fehler zu vermeiden. Prasad et al. berichten von deutlichen Abweichungen der kephalometrischen Messungen bei einem

Untersucher zu verschiedenen Zeitpunkten [135]. Diese möglichen systematischen Fehler sollten in der vorliegenden Arbeit ausgeschlossen werden.

Um die Intraraterreliabilität zu bestimmen, wurden jeweils 20 FRS durch einen Untersucher doppelt analysiert. Diese war für alle Untersucher und Parameter sehr hoch (alle ICC > .800 mit $p < .001$). Auch die Interraterreliabilität erreichte für alle Parameter sehr hohe Werte (alle ICC > .900 mit $p < .001$). Nach Koo et al. sind *ICC* von über 75% Indikatoren für eine gute Verlässlichkeit, *ICC* von über 90% für eine exzellente Verlässlichkeit [136].

4.1.4 Bewertung der Statistik

Zur Beurteilung der Genauigkeit einer Auswertung im Vergleich zu einem menschlichen Goldstandard ist es erforderlich, die Richtigkeit und die Präzision zu bewerten [114].

Die Parameter Richtigkeit und Präzision lassen sich aus Bland-Altman-Plots sehr gut herauslesen, daher wurden diese für die vorliegende Arbeit verwendet [137, 138]. Zusätzlich lassen sich anhand dieser Darstellung proportionale Fehler über die Regressionsgerade gut visualisieren [137].

Auch wenn statistisch signifikante Unterschiede zwischen den Ergebnissen eines kommerziellen KI-Anbieters und dem menschlichen Goldstandard zu verzeichnen sind, müssen diese nicht zwangsläufig klinisch relevant sein. In der Literatur werden Abweichungen von 1mm als kaum klinisch relevant, sogar Abweichungen von 2mm oder 2° als klinisch akzeptabel beschrieben, da diese meist innerhalb der ersten Standardabweichung liegen [80, 139].

Genauere Einteilungen bezüglich durchschnittlicher Richtigkeit und Präzision fehlen in der Literatur allerdings bisher. Daher wurden aus Gründen der besseren Kategorisierbarkeit durch einen klinisch erfahrenen Experten auf Grundlage der Bland-Altman-Plots zusätzlich Einteilungen in hoch, moderat und niedrig festgelegt.

4.2 Diskussion der Ergebnisse

4.2.1 DentaliQ.ortho

In einer Studie von Kunz et al. zur Bewertung der Auswertequalität von DentaliQ.ortho im Vergleich zu einem menschlichen Goldstandard wurde lediglich für den Parameter SN_MeGo ein statistisch signifikanter Unterschied ermittelt. Zudem waren die mittleren

Differenzen zum menschlichen Goldstandard sehr gering, sodass im Rahmen dieser Publikation von keinen klinisch relevanten Abweichungen ausgegangen wurde [32].

Diese Ergebnisse stimmen fast vollständig mit der vorliegenden Studie überein, da für alle untersuchten Parameter kein statistisch signifikanter Unterschied der Auswertung von DentaliQ.ortho zum menschlichen Goldstandard ermittelt werden konnte und durchweg hohe Genauigkeiten erzielt wurden.

Moreno und Gebeile-Chauty veröffentlichten 2022 eine Untersuchung zur metrischen Genauigkeit der Punktplatzierung durch DentaliQ.ortho im Vergleich zu WebCeph und zum menschlichen Goldstandard. Dabei lag die Erfolgsrate, die mit einer maximalen Abweichung von 2mm vom menschlichen Goldstandard bestimmt wurde, für DentaliQ.ortho bei 66,5%, für WebCeph bei 57,2%. DentaliQ.ortho war WebCeph damit überlegen, wenngleich der Unterschied nicht klinisch signifikant war. Die höchsten Richtigkeiten konnten unter anderem für die Inzisalkanten ermittelt werden [118].

Auch in der vorliegenden Arbeit war die Genauigkeit der von DentaliQ.ortho ermittelten FRS-Parameter der Genauigkeit von WebCeph überlegen. Da das Ziel der vorliegenden Arbeit allerdings nicht die Untersuchung der metrischen Abweichungen zum menschlichen Goldstandard war, ist ein genauerer Vergleich zur Studie von Moreno und Gebeile-Chauty nicht möglich.

4.2.2 WebCeph

Mahto et al. untersuchten 2022 in ihrer Publikation die Genauigkeit der KI-basierten FRS-Analyse durch WebCeph im Vergleich zur menschlichen manuellen FRS-Auswertung. Dazu wurden zufällig 30 prätherapeutische FRS ausgewählt und zur Festlegung des menschlichen Goldstandards durch einen einzelnen erfahrenen Kieferorthopäden manuell durchgezeichnet. Die FRS-Analyse umfasste zwölf Parameter, von denen vier Parameter (SNA, SNB, ANB und U1_MeGo) übereinstimmend in unserer Arbeit verwendet wurden. Im Anschluss wurde die Übereinstimmung der von WebCeph ermittelten Ergebnisse mit dem menschlichen Goldstandard anhand von *ICC*-Werten überprüft. *ICC* unter 0,75 wurden als schlechte bis moderate Übereinstimmung, *ICC* zwischen 0,75 und 0,9 wurden als gute Übereinstimmung und *ICC* über 0,9 wurden als exzellente Übereinstimmung definiert. Alle untersuchten Parameter erreichten einen *ICC* über 0,75, sieben Parameter erreichten einen *ICC* über 0,9. Für die Parameter SNA und SNB waren in der Publikation von Mahto et al. *ICC* von 0,879 bzw. 0,899 beschrieben, daher konnte für diese beiden

Parameter von einer guten Übereinstimmung mit dem menschlichen Goldstandard ausgegangen werden. Für die Parameter ANB und U1_MeGo waren ICC von 0,908 bzw. 0,915 beschrieben, was für eine exzellente Übereinstimmung mit dem menschlichen Goldstandard spricht [115]. Auch in der vorliegenden Arbeit waren für alle untersuchten Parameter, damit auch SNA, SNB, ANB und U1_MeGo keine statistisch signifikanten Unterschiede zum menschlichen Goldstandard zu verzeichnen. Allerdings sind für die Genauigkeit der KI-basierten Analyse zudem das Maß der Richtigkeit und Präzision entscheidend. Diese wurden in der Studie von Mahto et al. nicht untersucht. Hier fällt anhand der vorliegenden Arbeit auf, dass die Präzision von WebCeph bei allen vier genannten Parametern im Vergleich zu den anderen untersuchten kommerziellen KI-Anbietern niedriger war und zudem recht große proportionale Fehler vorhanden waren. Außerdem ist es sinnvoll, anders als in der Studie von Mahto et al., den menschlichen Goldstandard zur Vermeidung eines personenbezogenen Bias nicht nur anhand der Analyse eines einzelnen Untersuchers festzulegen. Dies empfehlen auch Kamoen et al. in ihrer Studie [19]. Mahto et al. beschrieben die von WebCeph ermittelten kephalometrischen Messwerte abschließend als recht genau, trafen aber gleichzeitig die Aussage, dass es zwingend erforderlich sei, dass die automatisierte FRS-Analyse zur Vermeidung von Fehlern bei der Diagnosestellung und Therapieplanung durch einen erfahrenen Kieferorthopäden kontrolliert wird [115].

Auch Yassir et al. verglichen 2022 die Genauigkeit der KI-basierten FRS-Analyse durch WebCeph mit der menschlichen Auswertung unter Zuhilfenahme der Software AutoCAD. Anhand von 50 prätherapeutischen FRS wurden elf Parameter ermittelt, davon fünf übereinstimmend mit der vorliegenden Arbeit (SNA, SNB, ANB, SN_MeGo und U1_MeGo). In der Publikation von Yassir et al. wurde der menschliche Goldstandard ebenfalls nur durch einen einzelnen Untersucher, der selbst kein Kieferorthopäde war, sondern von einem Kieferorthopäden trainiert wurde, festgelegt. Analog zur vorliegenden Arbeit wurden Bland-Altman-Plots zum Vergleich zwischen WebCeph und menschlichem Goldstandard erstellt, die einen Vergleich mit unseren Ergebnissen ermöglichen. Grundsätzlich zeigten von den fünf genannten Parametern bis auf SNA alle Parameter statistisch signifikante Unterschiede zum menschlichen Goldstandard. Interessant waren hierbei die Bland-Altman-Plots: für SNA, SNB, ANB und SN_MeGo waren die LoA und damit die Präzision vergleichbar mit unserer Arbeit, lediglich für U1_MeGo lagen die LoA deutlich weiter auseinander, sodass eine deutlich schlechtere Präzision im Vergleich zu den Ergebnissen der vorliegenden Arbeit nachgewiesen wurde. Die Parameter SNB und insbesondere U1_MeGo wurden von den Autoren als

nicht mehr klinisch akzeptabel eingestuft. Zudem errechnete WebCeph für die Frontzahninklinationen deutlich zu niedrige Werte. Yassir et al. schlussfolgerten, dass unterschiedliche Probleme wie unzureichende Punktesetzung und Widersprüchlichkeit der Messungen mit der Nutzung von WebCeph einhergehen und empfahlen daher, WebCeph nur mit großer Sorgfalt und Überprüfung durch einen Kliniker zu verwenden [121].

4.2.3 AudaxCeph

AudaxCeph zeigte in der vorliegenden Studie für sieben von neun untersuchten Parametern statistisch signifikante, für sechs Parameter sogar statistisch hochsignifikante Unterschiede zum menschlichen Goldstandard. Lediglich O1_SN und U1_MeGo wiesen keine statistisch signifikanten Unterschiede zum menschlichen Goldstandard auf. Leider liegen bisher keine vergleichbaren Studien zur Untersuchung der Genauigkeit der von AudaxCeph bestimmten FRS-Paramater vor.

Ristau et al. veröffentlichten 2022 eine Studie, die die metrische Genauigkeit der Punktplatzierung durch AudaxCeph im Vergleich zum menschlichen Goldstandard untersucht. In insgesamt 60 FRS wurden 13 Punkte gesetzt - sowohl von zwei Kieferorthopäden als auch von der KI-basierten Software AudaxCeph. Die Genauigkeit der Punktplatzierung durch AudaxCeph wurde anhand eines Koordinatensystems mit der menschlichen Punktsetzung verglichen. Hierbei waren lediglich für zwei Punkte, Porion und Apicale inferior, statistisch signifikante Unterschiede zum menschlichen Goldstandard zu verzeichnen [140].

Ein genauer Vergleich mit der vorliegenden Studie ist schwierig, da in unserer Studie nicht die metrische Platzierung der Punkte, sondern wie von Santoro et al. gefordert, die Gesamtheit der FRS-Analyse mit Berechnung der kephalometrischen Winkel und Strecken im Vordergrund steht [83]. Die Vergleichbarkeit ist auch vor dem Hintergrund der Tatsache, dass Fehler in der Punktesetzung sich hinsichtlich der Ergebnisse der therapierelevanten FRS-Parameter entweder gegenseitig abschwächen oder verstärken können, nur stark eingeschränkt gegeben [83, 141]. Beispielsweise würde eine fehlerhaft Punktplatzierung der KI auf einem Winkelschenkel nicht zu einer Veränderung des gemessenen Winkels führen [32].

4.2.4 CephX

In ihrer Studie aus dem Jahr 2020 untersuchten Meric und Naoumova die Genauigkeit der KI-basierten FRS-Analyse durch CephX im Vergleich zur manuellen menschlichen

Analyse anhand von 40 FRS und zwölf FRS-Parametern. Zudem wurden die App CephNinja und die Software Dolphin mit der manuellen menschlichen Auswertung verglichen. CephX schnitt dabei im Vergleich zu CephNinja und Dolphin am schlechtesten ab und wies für die Parameter SN_MeGo, I-NA (mm) und I-NB (°) statistisch signifikante Unterschiede zum menschlichen Goldstandard auf [142]. Vergleichbar mit der vorliegenden Studie ist nur der Parameter SN_MeGo. Der menschlich bestimmte Mittelwert in der Studie von Meric et al. lag bei 33,8 ± 6,9°, der durch CephX ermittelte Mittelwert bei 40,3 ± 6,5°. Damit lag ein hochsignifikanter Unterschied mit $p < .01$ zwischen den beiden Auswertungen vor [142]. In der vorliegenden Studie lag der menschlich bestimmte Mittelwert für SN_MeGo bei 30,62 ± 7,07°, der durch CephX ermittelte Mittelwert bei 35,15 ± 6,27°. Damit lag ebenfalls ein hochsignifikanter Unterschied mit $p < .01$ vor. In der Studie von Meric et al. wurde anschließend eine manuelle Korrektur der durch CephX automatisiert gesetzten Punkte durchgeführt. Durch diese Korrektur war für SN_MeGo kein statistisch signifikanter Unterschied zum menschlichen Goldstandard mehr zu verzeichnen [142]. Dies ist für die vorliegende Studie aber von untergeordneter Bedeutung, da hier nur die Genauigkeit der vollautomatisierten FRS-Analyse untersucht wurde. Insgesamt schlussfolgerten Meric et al., dass die vollautomatisierte, KI-basierte FRS-Analyse durch CephX noch genauer werden muss [142].

Wie in Kapitel 4.1.2 beschrieben, traten wir aufgrund von Zweifeln bezüglich der Übereinstimmung der Definition des Punktes Gonion, des Mandibularplanums und des Gesichtshöhenverhältnisses mit der in unserer Studie verwendeten Definition in Kontakt mit einer Mitarbeiterin der Orca Dental AI. Diese generierte eine eigens für diese Studie individualisierte FRS-Analyse und bestätigte uns wiederholt die Verwendung der von uns genutzten Definition der oben genannten Parameter. Dennoch lassen die Ergebnisse den Rückschluss zu, dass insbesondere Gonion möglicherweise trotz mehrmaliger Zusicherung der Orca Dental AI doch anders als in unserer Studie definiert wurde. Alle Parameter, die den Punkt Gonion enthalten, nämlich SN_MeGo, SpP_MeGo, das Gesichtshöhenverhältnis und U1_MeGo, wiesen statistisch hochsignifikante Unterschiede mit $p = .000$ zum menschlichen Goldstandard auf. Mit mittleren Differenzen von nahezu 6° zum menschlichen Goldstandard am Beispiel von U1_MeGo konnte die durchschnittliche Richtigkeit für diese vier Parameter als gering und damit klinisch inakzeptabel beschrieben werden.

4.2.5 Anbieterübergreifende Ergebnisse

Aufgrund der Schwierigkeit der präzisen Identifikation der Punkte gilt insbesondere die Messung der Frontzahninklination in der Literatur als fehleranfällig, sowohl in der manuellen als auch in der digitalen Analyse. Dies könnte auf die Tatsache zurückzuführen sein, dass sich viele Strukturen in den Bereichen eines FRS, die zur Platzierung der Landmarken für die dentale Analyse benötigt werden, röntgenologisch überlagern. Zudem zeigen diese Landmarken, auch wenn sie von menschlichen Experten gesetzt werden, erhöhte Abweichungen [142-145].

Da alle KI-Algorithmen auf Grundlage von menschlichen Daten trainiert werden, ist es nicht verwunderlich, dass auch die automatisierte FRS-Analyse im Hinblick auf die Frontzahninklination fehleranfällig ist. Die Genauigkeit von insbesondere DentaliQ.ortho, aber auch von AudaxCeph für die dentale Analyse und CephX für die Inklination des oberen mittleren Frontzahnes kann noch als klinisch akzeptabel gewertet werden. Hingegen können die Ergebnisse von CephX für die Inklination des unteren mittleren Frontzahnes und von WebCeph nicht mehr als klinisch akzeptabel beschrieben werden.

4.3 Beantwortung der Hypothesen

Die eingangs formulierten Nullhypothesen lassen sich anhand der in dieser Arbeit erhobenen Daten folgendermaßen bewerten:

- *Es besteht kein statistisch signifikanter Unterschied zwischen der automatisierten FRS-Auswertung durch DentaliQ.ortho und dem menschlichen Goldstandard.*

  Die Nullhypothese kann auf Grundlage der ANOVA mit Messwiederholung akzeptiert werden. Allerdings muss dabei berücksichtigt werden, dass die Präzision der Analyse von DentaliQ.ortho für zwei Parameter (O1_SN und U1-MeGo) und der zugehörige proportionale Fehler im moderaten Bereich lagen.

- *Es besteht kein statistisch signifikanter Unterschied zwischen der automatisierten FRS-Auswertung durch WebCeph und dem menschlichen Goldstandard.*

  Die Nullhypothese kann auf Grundlage der ANOVA mit Messwiederholung ebenfalls akzeptiert werden. Jedoch war die Präzision für alle Parameter

vergleichsweise am niedrigsten und der proportionale Fehler bei nahezu allen Parametern im Vergleich am größten.

- *Es besteht kein statistisch signifikanter Unterschied zwischen der automatisierten FRS-Auswertung durch AudaxCeph und dem menschlichen Goldstandard.*
  Die Nullhypothese muss für sieben von neun untersuchten Parametern verworfen werden und kann somit nur für zwei Parameter (O1_SN und U1_MeGo) akzeptiert werden. Für diese beiden Parameter war die Präzision allerdings niedrig und der proportionale Fehler hoch.

- *Es besteht kein statistisch signifikanter Unterschied zwischen der automatisierten FRS-Auswertung durch CephX und dem menschlichen Goldstandard.*
  Die Nullhypothese muss für fünf von neun Parametern ebenfalls verworfen werden und kann somit nur für vier Parameter (SNA, SNB, SN_SpP und O1_SN) akzeptiert werden. Jedoch war die Präzision für den Parameter SN_SpP moderat, für den Parameter O1_SN niedrig. Zudem war der proportionale Fehler für diese beiden Parameter vergleichsweise erhöht.

4.4 Schlussfolgerung

Bei der vorliegenden Arbeit handelt es sich um die erste Untersuchung, die mehrere kommerzielle KI-Anbieter zur vollständig automatisierten FRS-Analyse hinsichtlich ihrer Auswertequalität im Vergleich zu einem hochwertigen menschlichen Goldstandard untersucht und die Anbieter damit untereinander vergleichbar macht. Diese Arbeit kann somit eine wissenschaftlich fundierte Grundlage für die Entscheidung für oder gegen eine Implementierung von KI zur automatisierten FRS-Auswertung in den kieferorthopädischen Praxisalltag liefern.

Die Ergebnisse lassen den Rückschluss zu, dass der Anbieter DentaliQ.ortho unter allen untersuchten kommerziellen KI-Anbietern die höchste Auswertequalität erreicht. Die Präzision und das Risiko eines proportionalen Fehlers für die dentale Analyse lagen allerdings im moderaten Bereich. WebCeph weist trotz statistisch nicht signifikanter Unterschiede der Mittelwerte der untersuchten Parameter zum menschlichen Goldstandard durchweg eine vergleichsweise geringe Präzision und einen hohen

proportionalen Fehler auf. AudaxCeph zeigt bei sechs von neun Parametern hochsignifikante Unterschiede zum menschlichen Goldstandard, CephX bei vier von neun Parametern. Daher sind diese drei kommerziellen KI-Anbieter DentaliQ.ortho zum aktuellen Zeitpunkt noch unterlegen.

Die Entwicklung von KI schreitet zum aktuellen Zeitpunkt rasch voran, sodass zum einen stetig neue KI-Anwendungen auf dem Markt erscheinen und zum anderen bestehende Softwarelösungen weiter präzisiert werden. Die Daten für die vorliegende Arbeit wurden bereits 2021 erhoben, sodass etwaige Verbesserungen oder Updates an den untersuchten Softwarelösungen nach diesem Zeitpunkt nicht in die Ergebnisse dieser Arbeit einfließen. Daher ist es seitens wissenschaftlicher Auswertungen auch zukünftig erforderlich, die Analysequalität von KI-Lösungen wiederkehrend kritisch zu beurteilen – gleichzeitig ist es seitens der Anbieter kommerzieller Produkte ebenfalls erforderlich, Veränderungen an den Algorithmen eindeutig zu kennzeichnen und die Datengrundlage der Algorithmen detailliert und transparent darzustellen.

Insgesamt stellen KI-Algorithmen zur automatisierten FRS-Auswertung eine vielversprechende Unterstützung im klinischen Alltag dar und durch ihren Gebrauch lassen sich sowohl tagesformabhängige menschliche Fehler vermeiden als auch Zeit einsparen. Allerdings bestehen mitunter noch deutliche Abweichungen vom menschlichen Goldstandard und ausgeprägte Qualitätsunterschiede zwischen den kommerziellen KI-Anbietern. Daher sollten KI zum aktuellen Zeitpunkt nur als Ergänzung und unter Aufsicht erfahrener Kliniker angewandt werden.

5 Zusammenfassung

KI-Algorithmen erlangen in vielen Bereichen des Lebens wie auch in der Medizin und Zahnmedizin immer größere Bedeutung. Auch die in der Kieferorthopädie verwendete KI-basierte FRS-Analyse rückt zunehmend in den Fokus und kann den Kieferorthopäden im klinischen Alltag unterstützen. Dafür müssen KI ein hohes Maß an Genauigkeit erreichen. Bereits jetzt werden einige dieser Services kommerziell angeboten, wobei eine adäquate Datengrundlage allerdings häufig fehlt. Ziel der vorliegenden Studie war es daher, die aktuell verfügbaren kommerziellen KI-Anbieter zur FRS-Analyse bezüglich ihrer Analysequalität mit einem menschlichen Goldstandard zu vergleichen.

Auf 50 FRS wurden durch zwölf erfahrene Untersucher 15 Landmarken identifiziert, auf deren Basis neun relevante Parameter vermessen wurden. Der Medianwert dieser zwölf Auswertungen wurde für jeden Parameter bei jedem FRS als Goldstandard definiert und als Referenz für die Vergleiche mit vier verschiedenen kommerziellen Anbietern für KI-basierte FRS-Analysen (DentaliQ.ortho, WebCeph, AudaxCeph, CephX) festgelegt. Die statistische Auswertung erfolgte mittels ANOVA mit Messwiederholung, paarweiser Vergleiche mittels Post-hoc-Test und Bland-Altman-Plots.

DentaliQ.ortho zeigte für alle neun untersuchten Parameter keinen statistisch signifikanten Unterschied zum menschlichen Goldstandard und es konnte insgesamt von einer hohen Genauigkeit der Auswertungen ausgegangen werden. Auch für WebCeph war kein statistisch signifikanter Unterschied zum menschlichen Goldstandard zu verzeichnen. Allerdings war die Präzision im Vergleich zu den anderen Anbietern für alle Parameter am geringsten und der proportionale Fehler bei nahezu allen Parametern am höchsten. AudaxCeph wies für sieben Parameter statistisch signifikante, für sechs davon sogar statistisch hochsignifikante Unterschiede zum menschlichen Goldstandard auf. Für CephX wurden für fünf Parameter statistisch signifikante, davon für vier Parameter statistisch hochsignifikante Unterschiede zum menschlichen Goldstandard ermittelt. Insbesondere für die dentale Analyse war für alle untersuchten kommerziellen KI-Anbieter eine vergleichsweise niedrigere Genauigkeit zu verzeichnen.

Die Ergebnisse zeigen, dass noch deutliche Qualitätsunterschiede zwischen den kommerziellen KI-Anbietern für die vollständig automatisierte FRS-Analyse bestehen. Vor dem Hintergrund der Zeitersparnis und Qualitätssicherung sind KI zwar vielversprechend, sollten aber zum aktuellen Zeitpunkt nur unter Aufsicht durch menschliche Experten zum Einsatz kommen.

6 Literaturverzeichnis

1. Schwarz, A.M., *Die Röntgenostatik: die kieferorthopädische Diagnose am Fern-Röntgenbild*. 1958: Urban & Schwarzenberg.

2. Schopf, P., *Curriculum Kieferorthopädie Band I*. Vol. 4. 2008, Berlin: Quintessenz Verlags-GmbH.

3. Röntgen, W.C., *Über eine neue Art von Strahlen*. 1895, Würzburg: Sonderdruck aus den Sitzungsberichten der Würzburger Physik-medic Gesellschaft: Verlag der Stahel'schen Königlichen Hof- und Universitätsbuchhandlung.

4. Pacini, A., *Roentgen ray anthropometry of the skull.* J Radiol, 1922. 3(8): p. 322-31.

5. Simon, P.W., *Grundzüge einer systematischen Diagnostik der Gebiss-Anomalien: nebst Darbietung einer neuen Einteilung auf Grund der gnathostatischen Untersuchungsmethoden: ein Handbuch für Forschung und Praxis*. 1922, Berlin: H. Meusser.

6. Broadbent, B.H., *A NEW X-RAY TECHNIQUE and ITS APPLICATION TO ORTHODONTIA.* The Angle Orthodontist, 1931. 1(2): p. 45-66 DOI: 10.1043/0003-3219(1931)001<0045:Anxtai>2.0.Co;2.

7. Hofrath, H., *Die Bedeutung der Röntgenfern- und Abstandsaufnahme für die Diagnostik der Kieferanomalien.* Fortschritte der Orthodontik in Theorie und Praxis, 1931. 1(2): p. 232-258 DOI: 10.1007/BF02002578.

8. Arık, S., Ibragimov, B., and Xing, L., *Fully automated quantitative cephalometry using convolutional neural networks.* J Med Imaging (Bellingham), 2017. 4(1): p. 014501 DOI: 10.1117/1.Jmi.4.1.014501.

9. Kahl-Nieke, B., *Einführung in die Kieferorthopädie*. Vol. 2. 2001, München, Jena: Urban & Fischer Verlag.

10. Downs, W.B., *Variations in facial relationships; their significance in treatment and prognosis.* Am J Orthod, 1948. 34(10): p. 812-40 DOI: 10.1016/0002-9416(48)90015-3.

11. Nötzel, F., Schultz, C., and Hartung, M., *Fernröntgenseitenbild-Analyse*. 2007, Köln: Deutscher Zahnärzte Verlag.

12. Rakosi, T., *Atlas und Anleitung zur praktischen Fernröntgenanalyse*. Vol. 2. 1988, München-Wien: Hanser.

13. Steiner, C.C., *Cephalometrics In Clinical Practice.* The Angle Orthodontist, 1959. 29(1): p. 8-29 DOI: 10.1043/0003-3219(1959)029<0008:Cicp>2.0.Co;2.

14. Björk, A. and Lundström, A., *Introduction to orthodontics*. 1960, New York: McGraw-Hill.

15. Jarabak, J.R. and Fizzell, J.A., *Technique and Treatment with Light-wire Edgewise Appliances*. 1972: C. V. Mosby Company.

16. El-Feghi, I., Sid-Ahmed, M., and Ahmadi, M., *Automatic localization of craniofacial landmarks for assisted cephalometry.* Pattern Recognition, 2004. 37: p. 609-621.

17. Nishimoto, S., et al., *Personal Computer-Based Cephalometric Landmark Detection With Deep Learning, Using Cephalograms on the Internet.* J Craniofac Surg, 2019. 30(1): p. 91-95 DOI: 10.1097/scs.0000000000004901.

18. Haynes, S. and Chau, M.N., *Inter- and intra-observer identification of landmarks used in the Delaire analysis.* Eur J Orthod, 1993. 15(1): p. 79-84 DOI: 10.1093/ejo/15.1.79.

19. Kamoen, A., Dermaut, L., and Verbeeck, R., *The clinical significance of error measurement in the interpretation of treatment results.* Eur J Orthod, 2001. 23(5): p. 569-78 DOI: 10.1093/ejo/23.5.569.

20. Kim, H., et al., *Web-based fully automated cephalometric analysis by deep learning.* Comput Methods Programs Biomed, 2020. 194: p. 105513 DOI: 10.1016/j.cmpb.2020.105513.

21. Lévy-Mandel, A.D., Venetsanopoulos, A.N., and Tsotsos, J.K., *Knowledge-based landmarking of cephalograms.* Comput Biomed Res, 1986. 19(3): p. 282-309 DOI: 10.1016/0010-4809(86)90023-6.

22. Cardillo, J. and Sid-Ahmed, M.A., *An image processing system for locating craniofacial landmarks.* IEEE Trans Med Imaging, 1994. 13(2): p. 275-89 DOI: 10.1109/42.293920.

23. Grau, V., et al., *Automatic localization of cephalometric Landmarks.* J Biomed Inform, 2001. 34(3): p. 146-56 DOI: 10.1006/jbin.2001.1014.

24. Davis, D.N. and Taylor, C.J., *A blackboard architecture for automating cephalometric analysis.* Med Inform (Lond), 1991. 16(2): p. 137-49 DOI: 10.3109/14639239109012123.

25. Yue, W., et al., *Automated 2-D cephalometric analysis on X-ray images by a model-based approach.* IEEE Trans Biomed Eng, 2006. 53(8): p. 1615-23 DOI: 10.1109/tbme.2006.876638.

26. Barbour, A.B., et al., *Artificial Intelligence in Health Care: Insights From an Educational Forum.* J Med Educ Curric Dev, 2019. 6: p. 2382120519889348 DOI: 10.1177/2382120519889348.

27. Konrad, E., *Zur Geschichte der Künstlichen Intelligenz in der Bundesrepublik Deutschland*, in *Sozialgeschichte der Informatik: Kulturelle Praktiken und Orientierungen*, D. Siefkes, et al., Editors. 1998, Deutscher Universitätsverlag: Wiesbaden. p. 287-296.

28. Weidlich, V. and Weidlich, G.A., *Artificial Intelligence in Medicine and Radiation Oncology.* Cureus, 2018. 10(4): p. e2475 DOI: 10.7759/cureus.2475.

29. Copeland, B.J., *The Turing Test*.* Minds and Machines, 2000. 10(4): p. 519-539 DOI: 10.1023/A:1011285919106.

30. Yasaka, K., et al., *Deep learning with convolutional neural network in radiology.* Jpn J Radiol, 2018. 36(4): p. 257-272 DOI: 10.1007/s11604-018-0726-3.

31. Hubel, D.H. and Wiesel, T.N., *Receptive fields, binocular interaction and functional architecture in the cat's visual cortex.* J Physiol, 1962. 160(1): p. 106-54 DOI: 10.1113/jphysiol.1962.sp006837.

32. Kunz, F., et al., *Artificial intelligence in orthodontics : Evaluation of a fully automated cephalometric analysis using a customized convolutional neural network.* J Orofac Orthop, 2020. 81(1): p. 52-68 DOI: 10.1007/s00056-019-00203-8.

33. LeCun, Y., Bengio, Y., and Hinton, G., *Deep learning.* Nature, 2015. 521(7553): p. 436-44 DOI: 10.1038/nature14539.

34. Kleesiek, J., et al., *Wie funktioniert maschinelles Lernen?* Der Radiologe, 2020. 60(1): p. 24-31 DOI: 10.1007/s00117-019-00616-x.

35. Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS. *Aktuelle Forschungsprojekte*. 2021 02.05.2021]; Available from: https://www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz.html.

36. Antweiler, D., et al. *Künstliche Intelligenz im Krankenhaus. Potenziale und Herausforderungen - eine Fallstudie im Bereich der Notfallversorgung*. 2020 02.05.2021]; Available from: https://newsletter.fraunhofer.de/public/a_14338_S4Jdz/file/data/1965_Fraunhofer_IAIS_Whitepaper_LOTTE_web.pdf.

37. Kleber, C., et al., *Rettungszeit und Überleben von Schwerverletzten in Deutschland.* Der Unfallchirurg, 2012. 115: p. 345-350.

38. Rijkhoek, K.G., Callaghan, V., and Fleiter, D. *Neue Initiative zu KI in der Medizin gestartet*. 2020 02.05.2021]; Available from: https://www.is.mpg.de/de/news/new-ai-in-medicine-initiative-launched.

39. Deutsche Forschungsgemeinschaft. *Forschungsgruppen und Kolleg-Forschungsgruppen im Bereich „Künstliche Intelligenz“*. 2020 02.05.2021]; Available from: https://www.dfg.de/foerderung/info_wissenschaft/2020/info_wissenschaft_20_08/.

40. Bundesministerium für Bildung und Forschung. *Künstliche Intelligenz*. 2020 02.05.2021]; Available from: https://www.bmbf.de/de/kuenstliche-intelligenz-5965.html.

41. KI Bundesverband. *Unsere Mission*. 2020 02.05.2021]; Available from: https://ki-verband.de/mission/.

42. Lundberg, S.M., et al., *Explainable machine-learning predictions for the prevention of hypoxaemia during surgery.* Nat Biomed Eng, 2018. 2(10): p. 749-760 DOI: 10.1038/s41551-018-0304-0.

43. Wang, R., et al., *Artificial intelligence in reproductive medicine.* Reproduction, 2019. 158(4): p. R139-r154 DOI: 10.1530/rep-18-0523.

44. Topol, E.J., *High-performance medicine: the convergence of human and artificial intelligence.* Nat Med, 2019. 25(1): p. 44-56 DOI: 10.1038/s41591-018-0300-7.

45. Murray, J.M., et al., *Wie funktioniert Radiomics?* Der Radiologe, 2020. 60(1): p. 32-41 DOI: 10.1007/s00117-019-00617-w.

46. Scheckenbach, K., *Radiomics: Big Data Instead of Biopsies in the Future?* Laryngorhinootologie, 2018. 97(S 01): p. S114-s141 DOI: 10.1055/s-0043-121964.

47. Skogen, K., et al., *Measurements of heterogeneity in gliomas on computed tomography relationship to tumour grade.* J Neurooncol, 2013. 111(2): p. 213-9 DOI: 10.1007/s11060-012-1010-5.

48. Yasaka, K., et al., *Quantitative computed tomography texture analysis for estimating histological subtypes of thymic epithelial tumors.* Eur J Radiol, 2017. 92: p. 84-92 DOI: 10.1016/j.ejrad.2017.04.017.

49. Lubner, M.G., et al., *CT Textural Analysis of Large Primary Renal Cell Carcinomas: Pretreatment Tumor Heterogeneity Correlates With Histologic Findings and Clinical Outcomes.* AJR Am J Roentgenol, 2016. 207(1): p. 96-105 DOI: 10.2214/ajr.15.15451.

50. Kickingereder, P., et al., *Radiomic Profiling of Glioblastoma: Identifying an Imaging Predictor of Patient Survival with Improved Performance over Established Clinical and Radiologic Risk Models.* Radiology, 2016. 280(3): p. 880-9 DOI: 10.1148/radiol.2016160845.

51. Huang, Y., et al., *Radiomics Signature: A Potential Biomarker for the Prediction of Disease-Free Survival in Early-Stage (I or II) Non-Small Cell Lung Cancer.* Radiology, 2016. 281(3): p. 947-957 DOI: 10.1148/radiol.2016152234.

52. Goh, V., et al., *Assessment of response to tyrosine kinase inhibitors in metastatic renal cell cancer: CT texture as a predictive biomarker.* Radiology, 2011. 261(1): p. 165-71 DOI: 10.1148/radiol.11110264.

53. Li, H., et al., *MR Imaging Radiomics Signatures for Predicting the Risk of Breast Cancer Recurrence as Given by Research Versions of MammaPrint, Oncotype DX, and PAM50 Gene Assays.* Radiology, 2016. 281(2): p. 382-391 DOI: 10.1148/radiol.2016152110.

54. Sung, H., et al., *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.* CA Cancer J Clin, 2021. 71(3): p. 209-249 DOI: 10.3322/caac.21660.

55. Becker, A.S., et al., *Deep Learning in Mammography: Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer.* Invest Radiol, 2017. 52(7): p. 434-440 DOI: 10.1097/rli.0000000000000358.

56. Ma, W., et al., *Breast Cancer Molecular Subtype Prediction by Mammographic Radiomic Features.* Acad Radiol, 2019. 26(2): p. 196-201 DOI: 10.1016/j.acra.2018.01.023.

57. Poortmans, P.M.P., et al., *Winter is over: The use of Artificial Intelligence to individualise radiation therapy for breast cancer.* Breast, 2020. 49: p. 194-200 DOI: 10.1016/j.breast.2019.11.011.

58. Tran, W.T., et al., *Personalized Breast Cancer Treatments Using Artificial Intelligence in Radiomics and Pathomics.* J Med Imaging Radiat Sci, 2019. 50(4 Suppl 2): p. S32-s41 DOI: 10.1016/j.jmir.2019.07.010.

59. Cicero, M., et al., *Training and Validating a Deep Convolutional Neural Network for Computer-Aided Detection and Classification of Abnormalities on Frontal Chest Radiographs.* Invest Radiol, 2017. 52(5): p. 281-287 DOI: 10.1097/rli.0000000000000341.

60. Song, Q., et al., *Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images.* J Healthc Eng, 2017. 2017: p. 8314740 DOI: 10.1155/2017/8314740.

61. Nibali, A., He, Z., and Wollersheim, D., *Pulmonary nodule classification with deep residual networks.* Int J Comput Assist Radiol Surg, 2017. 12(10): p. 1799-1808 DOI: 10.1007/s11548-017-1605-6.

62. Huynh, E., et al., *CT-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer.* Radiother Oncol, 2016. 120(2): p. 258-66 DOI: 10.1016/j.radonc.2016.05.024.

63. Jiang, M., et al., *Assessing PD-L1 Expression Level by Radiomic Features From PET/CT in Nonsmall Cell Lung Cancer Patients: An Initial Result.* Acad Radiol, 2020. 27(2): p. 171-179 DOI: 10.1016/j.acra.2019.04.016.

64. Aerts, H.J., et al., *Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach.* Nat Commun, 2014. 5: p. 4006 DOI: 10.1038/ncomms5006.

65. Niu, H., et al., *Prevalence and incidence of Alzheimer's disease in Europe: A meta-analysis.* Neurologia, 2017. 32(8): p. 523-532 DOI: 10.1016/j.nrl.2016.02.016.

66. Scheltens, P., et al., *Alzheimer's disease.* Lancet, 2016. 388(10043): p. 505-17 DOI: 10.1016/s0140-6736(15)01124-1.

67. Ding, Y., et al., *A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using (18)F-FDG PET of the Brain.* Radiology, 2019. 290(2): p. 456-464 DOI: 10.1148/radiol.2018180958.

68. Liu, X., et al., *Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease.* Transl Res, 2018. 194: p. 56-67 DOI: 10.1016/j.trsl.2018.01.001.

69. Schwendicke, F., et al., *Convolutional neural networks for dental image diagnostics: A scoping review.* J Dent, 2019. 91: p. 103226 DOI: 10.1016/j.jdent.2019.103226.

70. Zhang, W., et al., *Predicting postoperative facial swelling following impacted mandibular third molars extraction by using artificial neural networks evaluation.* Sci Rep, 2018. 8(1): p. 12281 DOI: 10.1038/s41598-018-29934-1.

71. Schwendicke, F., Dommisch, H., and Krois, J., *Künstliche Intelligenz in der Bildanalytik Chancen und Herausforderungen für die Parodontologie.* Parodontologie, 2020. 31(4): p. 417-423.

72. United Nations Scientific Committee on the Effects of Atomic Radiation, *Sources and Effects of Ionizing Radiation, United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) 2008 Report to the General Assembly with Scientific Annexes*. 2010: United Nations.

73. Lee, J.H., et al., *Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm.* J Dent, 2018. 77: p. 106-111 DOI: 10.1016/j.jdent.2018.07.015.

74. Lee, J.H., et al., *Diagnosis and prediction of periodontally compromised teeth using a deep learning-based convolutional neural network algorithm.* J Periodontal Implant Sci, 2018. 48(2): p. 114-123 DOI: 10.5051/jpis.2018.48.2.114.

75. Khanagar, S.B., et al., *Developments, application, and performance of artificial intelligence in dentistry - A systematic review.* J Dent Sci, 2021. 16(1): p. 508-522 DOI: 10.1016/j.jds.2020.06.019.

76. Patil, V., et al., *Artificial neural network for gender determination using mandibular morphometric parameters: A comparative retrospective study.* Cogent Engineering, 2020. 7(1): p. 1723783 DOI: 10.1080/23311916.2020.1723783.

77. De Tobel, J., et al., *An automated technique to stage lower third molar development on panoramic radiographs for age estimation: a pilot study.* The Journal of forensic odonto-stomatology, 2017. 35(2): p. 42-54.

78. Thanathornwong, B., *Bayesian-Based Decision Support System for Assessing the Needs for Orthodontic Treatment.* Healthc Inform Res, 2018. 24(1): p. 22-28 DOI: 10.4258/hir.2018.24.1.22.

79. Schwendicke, F., et al., *Deep learning for cephalometric landmark detection: systematic review and meta-analysis.* Clin Oral Investig, 2021. 25(7): p. 4299-4309 DOI: 10.1007/s00784-021-03990-w.

80. Chen, Y.J., et al., *The effects of differences in landmark identification on the cephalometric measurements in traditional versus digitized cephalometry.* Angle

Orthod, 2004. 74(2): p. 155-61 DOI: 10.1043/0003-3219(2004)074<0155:Teodil>2.0.Co;2.

81. Hwang, H.W., et al., *Automated identification of cephalometric landmarks: Part 2- Might it be better than human?* Angle Orthod, 2020. 90(1): p. 69-76 DOI: 10.2319/022019-129.1.

82. Lee, J.H., et al., *Automated cephalometric landmark detection with confidence regions using Bayesian convolutional neural networks.* BMC Oral Health, 2020. 20(1): p. 270 DOI: 10.1186/s12903-020-01256-7.

83. Santoro, M., Jarjoura, K., and Cangialosi, T.J., *Accuracy of digital and analogue cephalometric measurements assessed with the sandwich technique.* Am J Orthod Dentofacial Orthop, 2006. 129(3): p. 345-51 DOI: 10.1016/j.ajodo.2005.12.010.

84. Proffit, W., et al., *Contemporary Orthodontics*. Sixth Edition ed. 2019: Elsevier.

85. Khanagar, S.B., et al., *Scope and performance of artificial intelligence technology in orthodontic diagnosis, treatment planning, and clinical decision-making - A systematic review.* J Dent Sci, 2021. 16(1): p. 482-492 DOI: 10.1016/j.jds.2020.05.022.

86. Kim, D.W., et al., *Prediction of hand-wrist maturation stages based on cervical vertebrae images using artificial intelligence.* Orthod Craniofac Res, 2021. 24 Suppl 2: p. 68-75 DOI: 10.1111/ocr.12514.

87. Alkhal, H.A., Wong, R.W., and Rabie, A.B., *Correlation between chronological age, cervical vertebral maturation and Fishman's skeletal maturity indicators in southern Chinese.* Angle Orthod, 2008. 78(4): p. 591-6 DOI: 10.2319/0003-3219(2008)078[0591:Cbcacv]2.0.Co;2.

88. Demirjian, A., et al., *Interrelationships among measures of somatic, skeletal, dental, and sexual maturity.* Am J Orthod, 1985. 88(5): p. 433-8 DOI: 10.1016/0002-9416(85)90070-3.

89. Fishman, L.S., *Chronological versus skeletal age, an evaluation of craniofacial growth.* Angle Orthod, 1979. 49(3): p. 181-9 DOI: 10.1043/0003-3219(1979)049<0181:Cvsaae>2.0.Co;2.

90. Baccetti, T., Franchi, L., and McNamara, J.A., Jr., *An improved version of the cervical vertebral maturation (CVM) method for the assessment of mandibular growth.* Angle Orthod, 2002. 72(4): p. 316-23 DOI: 10.1043/0003-3219(2002)072<0316:Aivotc>2.0.Co;2.

91. Fishman, L.S., *Radiographic evaluation of skeletal maturation. A clinically oriented method based on hand-wrist films.* Angle Orthod, 1982. 52(2): p. 88-112 DOI: 10.1043/0003-3219(1982)052<0088:Reosm>2.0.Co;2.

92. Mito, T., Sato, K., and Mitani, H., *Cervical vertebral bone age in girls.* Am J Orthod Dentofacial Orthop, 2002. 122(4): p. 380-5 DOI: 10.1067/mod.2002.126896.

93. McNamara, J.A., Jr. and Franchi, L., *The cervical vertebral maturation method: A user's guide.* Angle Orthod, 2018. 88(2): p. 133-143 DOI: 10.2319/111517-787.1.

94. Amasya, H., et al., *Validation of cervical vertebral maturation stages: Artificial intelligence vs human observer visual analysis.* Am J Orthod Dentofacial Orthop, 2020. 158(6): p. e173-e179 DOI: 10.1016/j.ajodo.2020.08.014.

95. Seo, H., et al., *Comparison of Deep Learning Models for Cervical Vertebral Maturation Stage Classification on Lateral Cephalometric Radiographs.* J Clin Med, 2021. 10(16) DOI: 10.3390/jcm10163591.

96. Li, P., et al., *Orthodontic Treatment Planning based on Artificial Neural Networks.* Sci Rep, 2019. 9(1): p. 2037 DOI: 10.1038/s41598-018-38439-w.

97. Kim, Y.H., et al., *Influence of the Depth of the Convolutional Neural Networks on an Artificial Intelligence Model for Diagnosis of Orthognathic Surgery.* J Pers Med, 2021. 11(5) DOI: 10.3390/jpm11050356.

98. Burrow, S.J., *To extract or not to extract: a diagnostic decision, not a marketing decision.* Am J Orthod Dentofacial Orthop, 2008. 133(3): p. 341-2 DOI: 10.1016/j.ajodo.2007.11.016.

99. Suhail, Y., et al., *Machine Learning for the Diagnosis of Orthodontic Extractions: A Computational Analysis Using Ensemble Learning.* Bioengineering (Basel), 2020. 7(2) DOI: 10.3390/bioengineering7020055.

100. Real, A.D., et al., *Use of automated artificial intelligence to predict the need for orthodontic extractions.* Korean J Orthod, 2022. 52(2): p. 102-111 DOI: 10.4041/kjod.2022.52.2.102.

101. Jung, S.K. and Kim, T.W., *New approach for the diagnosis of extractions with neural network machine learning.* Am J Orthod Dentofacial Orthop, 2016. 149(1): p. 127-33 DOI: 10.1016/j.ajodo.2015.07.030.

102. Etemad, L., et al., *Machine learning from clinical data sets of a contemporary decision for orthodontic tooth extraction.* Orthod Craniofac Res, 2021. 24 Suppl 2: p. 193-200 DOI: 10.1111/ocr.12502.

103. Baumrind, S., et al., *The decision to extract: Part 1--Interclinician agreement.* Am J Orthod Dentofacial Orthop, 1996. 109(3): p. 297-309 DOI: 10.1016/s0889-5406(96)70153-1.

104. Luke, L.S., Atchison, K.A., and White, S.C., *Consistency of patient classification in orthodontic diagnosis and treatment planning.* Angle Orthod, 1998. 68(6): p. 513-20 DOI: 10.1043/0003-3219(1998)068<0513:Copcio>2.3.Co;2.

105. Xie, X., Wang, L., and Wang, A., *Artificial neural network modeling for deciding if extractions are necessary prior to orthodontic treatment.* Angle Orthod, 2010. 80(2): p. 262-6 DOI: 10.2319/111608-588.1.

106. Choi, H.I., et al., *Artificial Intelligent Model With Neural Network Machine Learning for the Diagnosis of Orthognathic Surgery.* J Craniofac Surg, 2019. 30(7): p. 1986-1989 DOI: 10.1097/scs.0000000000005650.

107. Patcas, R., et al., *Applying artificial intelligence to assess the impact of orthognathic treatment on facial attractiveness and estimated age.* Int J Oral Maxillofac Surg, 2019. 48(1): p. 77-83 DOI: 10.1016/j.ijom.2018.07.010.

108. Oland, J., et al., *Motives for surgical-orthodontic treatment and effect of treatment on psychosocial well-being and satisfaction: a prospective study of 118 patients.* J Oral Maxillofac Surg, 2011. 69(1): p. 104-13 DOI: 10.1016/j.joms.2010.06.203.

109. Patcas, R., et al., *Motivation for orthognathic treatment and anticipated satisfaction levels-a two-centre cross-national audit.* J Craniomaxillofac Surg, 2017. 45(6): p. 1004-1009 DOI: 10.1016/j.jcms.2017.03.012.

110. Ter Horst, R., et al., *Three-dimensional virtual planning in mandibular advancement surgery: Soft tissue prediction based on deep learning.* J Craniomaxillofac Surg, 2021 DOI: 10.1016/j.jcms.2021.04.001.

111. dentalXrai GmbH. *Innovative KI-Technologie für die Zahnmedizin*. 2021 09.05.2021]; Available from: https://www.dentalxr.ai/technologie/.

112. Schwendicke, F., et al., *Artificial intelligence in dental research: Checklist for authors, reviewers, readers.* J Dent, 2021. 107: p. 103610 DOI: 10.1016/j.jdent.2021.103610.

113. Krouwer, J.S., *Why Bland-Altman plots should use X, not (Y+X)/2 when X is a reference method.* Stat Med, 2008. 27(5): p. 778-80 DOI: 10.1002/sim.3086.

114. Menditto, A., Patriarca, M., and Magnusson, B., *Understanding the meaning of accuracy, trueness and precision.* Accreditation and Quality Assurance, 2007. 12(1): p. 45-47 DOI: 10.1007/s00769-006-0191-z.

115. Mahto, R.K., et al., *Evaluation of fully automated cephalometric measurements obtained from web-based artificial intelligence driven platform.* BMC Oral Health, 2022. 22(1): p. 132 DOI: 10.1186/s12903-022-02170-w.

116. Le, V.N.T., et al., *Effectiveness of Human-Artificial Intelligence Collaboration in Cephalometric Landmark Detection.* J Pers Med, 2022. 12(3) DOI: 10.3390/jpm12030387.

117. Bulatova, G., et al., *Assessment of automatic cephalometric landmark identification using artificial intelligence.* Orthod Craniofac Res, 2021. 24 Suppl 2: p. 37-42 DOI: 10.1111/ocr.12542.

118. Moreno, M. and Gebeile-Chauty, S., *[Comparative study of two software for the detection of cephalometric landmarks by artificial intelligence].* Orthod Fr, 2022. 93(1): p. 41-61 DOI: 10.1684/orthodfr.2022.73.

119. Kılınç, D.D., et al., *Evaluation and comparison of smartphone application tracing, web based artificial intelligence tracing and conventional hand tracing methods.* J Stomatol Oral Maxillofac Surg, 2022 DOI: 10.1016/j.jormas.2022.07.017.

120. Alqahtani, H., *Evaluation of an online website-based platform for cephalometric analysis.* J Stomatol Oral Maxillofac Surg, 2020. 121(1): p. 53-57 DOI: 10.1016/j.jormas.2019.04.017.

121. Yassir, Y.A., Salman, A.R., and Nabbat, S.A., *The accuracy and reliability of WebCeph for cephalometric analysis.* J Taibah Univ Med Sci, 2022. 17(1): p. 57-66 DOI: 10.1016/j.jtumed.2021.08.010.

122. Chinem, L.A., et al., *Digital orthodontic radiographic set versus cone-beam computed tomography: an evaluation of the effective dose.* Dental Press J Orthod, 2016. 21(4): p. 66-72 DOI: 10.1590/2177-6709.21.4.066-072.oar.

123. Ludlow, J.B., Davies-Ludlow, L.E., and White, S.C., *Patient risk related to common dental radiographic examinations: the impact of 2007 International Commission on Radiological Protection recommendations regarding dose calculation.* J Am Dent Assoc, 2008. 139(9): p. 1237-43 DOI: 10.14219/jada.archive.2008.0339.

124. Vasil'ev, Y., Paulsen, F., and Dydykin, S., *Anatomical and radiological features of the bone organization of the anterior part of the mandible.* Ann Anat, 2020. 231: p. 151512 DOI: 10.1016/j.aanat.2020.151512.

125. Pracharktam, N., et al., *Cephalometric assessment in obstructive sleep apnea.* American Journal of Orthodontics and Dentofacial Orthopedics, 1996. 109(4): p. 410-419 DOI: https://doi.org/10.1016/S0889-5406(96)70123-3.

126. Ricketts, R.M., *Bioprogressive Therapie. 2. Aufl.* 1988, Heidelberg: Hüthig.

127. American Board of Orthodontics. *Cephalometric Tracings. Construction of the Mandibular Plane*. 2022 [cited 2022 14.08.2022]; Available from: http://www.americanboardortho.com/media/4omnm4kg/construction-of-the-mandibular-plane.pdf.

128. Sassouni, V., *A roentgenographic cephalometric analysis of cephalo-facio-dental relationships.* American Journal of Orthodontics, 1955. 41(10): p. 735-764 DOI: https://doi.org/10.1016/0002-9416(55)90171-8.

129. Zhou, J., et al., *Development of an Artificial Intelligence System for the Automatic Evaluation of Cervical Vertebral Maturation Status.* Diagnostics (Basel), 2021. 11(12) DOI: 10.3390/diagnostics11122200.

130. Gonçalves, F.A., et al., *Comparison of cephalometric measurements from three radiological clinics.* Braz Oral Res, 2006. 20(2): p. 162-6 DOI: 10.1590/s1806-83242006000200013.

131. Wang, C.W., et al., *Evaluation and Comparison of Anatomical Landmark Detection Methods for Cephalometric X-Ray Images: A Grand Challenge.* IEEE Trans Med Imaging, 2015. 34(9): p. 1890-900 DOI: 10.1109/tmi.2015.2412951.

132. Liu, J.K., Chen, Y.T., and Cheng, K.S., *Accuracy of computerized automatic identification of cephalometric landmarks.* Am J Orthod Dentofacial Orthop, 2000. 118(5): p. 535-40 DOI: 10.1067/mod.2000.110168.

133. Park, J.H., et al., *Automated identification of cephalometric landmarks: Part 1-Comparisons between the latest deep-learning methods YOLOV3 and SSD.* Angle Orthod, 2019. 89(6): p. 903-909 DOI: 10.2319/022019-127.1.

134. Hwang, H.W., et al., *Evaluation of automated cephalometric analysis based on the latest deep learning method.* Angle Orthod, 2021. 91(3): p. 329-335 DOI: 10.2319/021220-100.1.

135. Prasad, S., Denotti, G., and Farella, M., *Effect of prior knowledge about treatment on cephalometric measurements.* J Orthod, 2022: p. 14653125221094333 DOI: 10.1177/14653125221094333.

136. Koo, T.K. and Li, M.Y., *A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research.* J Chiropr Med, 2016. 15(2): p. 155-63 DOI: 10.1016/j.jcm.2016.02.012.

137. Giavarina, D., *Understanding bland altman analysis.* Biochemia medica, 2015. 25(2): p. 141-151.

138. Gerke, O., *Reporting standards for a Bland–Altman agreement analysis: A review of methodological reviews.* Diagnostics, 2020. 10(5): p. 334.

139. Schulze, R.K., Gloede, M.B., and Doll, G.M., *Landmark identification on direct digital versus film-based cephalometric radiographs: a human skull study.* Am J Orthod Dentofacial Orthop, 2002. 122(6): p. 635-42 DOI: 10.1067/mod.2002.129191.

140. Ristau, B., et al., *Comparison of AudaxCeph®'s fully automated cephalometric tracing technology to a semi-automated approach by human examiners.* Int Orthod, 2022: p. 100691 DOI: 10.1016/j.ortho.2022.100691.

141. Ongkosuwito, E.M., et al., *The reproducibility of cephalometric measurements: a comparison of analogue and digital methods.* Eur J Orthod, 2002. 24(6): p. 655-65 DOI: 10.1093/ejo/24.6.655.

142. Meriç, P. and Naoumova, J., *Web-based Fully Automated Cephalometric Analysis: Comparisons between App-aided, Computerized, and Manual Tracings.* Turk J Orthod, 2020. 33(3): p. 142-149 DOI: 10.5152/TurkJOrthod.2020.20062.

143. Chan, C.K., et al., *Effects of cephalometric landmark validity on incisor angulation.* Am J Orthod Dentofacial Orthop, 1994. 106(5): p. 487-95 DOI: 10.1016/s0889-5406(94)70071-0.

144. Paixão, M., et al., *Comparative study between manual and digital cephalometric tracing using Dolphin Imaging software with lateral radiographs.* Dental Press Journal of Orthodontics, 2010. 15: p. 123-130 DOI: 10.1590/S2176-94512010000600016.

145. Baumrind, S. and Frantz, R.C., *The reliability of head film measurements. 2. Conventional angular and linear measures.* Am J Orthod, 1971. 60(5): p. 505-17 DOI: 10.1016/0002-9416(71)90116-3.

Appendix

I Abkürzungsverzeichnis

18-FDG	18-Fluordesoxyglucose
ANOVA	Analysis of Variance
BMBF	Bundesministerium für Bildung und Forschung
BoP	Bleeding on Probing
CNN	Convolutional Neural Network
CT	Computertomographie
CVM	Cervical Vertebral Maturation
DAI	Dental Aesthetic Index
DFG	Deutsche Forschungsgemeinschaft
DGKFO	Deutsche Gesellschaft für Kieferorthopädie e.V.
DVT	Digitale Volumentomographie
EMBL	European Molecular Biology Laboratory
FRS	Fernröntgenseitenbild
IAIS	Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme
ICC	Intraklassen-Korrelations-Koeffizient
IOTN	Index of Orthodontic Treatment Need
KI	Künstliche Intelligenz
LOA	Limits of Agreement
LOTTE	Leitsystem zur Optimierung der Therapie traumatisierter Patient*innen bei der Erstbehandlung
M	Mittelwert
MRT	Magnetresonanztomographie
MTM	Massentensormodell
OP	Operation
OPG	Orthopantomogramm
PET	Positronenemissionstomographie
RNN	Recurrent Neural Networks
SD	Standardabweichung
UNSCEAR	United Nations Scientific Committee on the Effects of Atomic Radiation

II Abbildungsverzeichnis

III Tabellenverzeichnis

IV Danksagung

Zunächst möchte ich mich bei Frau Professorin Dr. Angelika Stellzig-Eisenhauer für die Überlassung des Dissertationsthemas und die fortlaufende wissenschaftliche Unterstützung bedanken. Ebenso gilt ihr mein Dank für die umfassende kieferorthopädische Ausbildung, die stetige Förderung und den fortwährenden Rückhalt.

Ein besonderer Dank geht an Herrn Priv.-Doz. Dr. Felix Kunz für die hervorragende Betreuung und Zusammenarbeit, sowohl auf fachlicher als auch auf menschlicher Ebene. Ich konnte in wissenschaftlicher und klinischer Hinsicht immer auf seine Zeit, Unterstützung und in jeglicher Richtung kompetenten Ratschläge vertrauen und so viel von ihm lernen.

Außerdem möchte ich mich bei Herrn Professor Dr. Gabriel Krastl für die Übernahme des Koreferats bedanken.

Darüber hinaus danke ich Herrn Florian Zeman für die statistische Unterstützung.

Von ganzem Herzen danke ich meiner Familie. Meinem Ehemann Dr. Benjamin Widmaier, der mich mit seinem Optimismus immer wieder ermutigt und mehr als nur einmal seine eigenen Interessen hintenangestellt hat, um mir zur Seite zu stehen. Meinen Eltern Verena und Ralf Wirth für die liebevolle Unterstützung in jeglicher Hinsicht und dafür, dass sie mir diesen Lebensweg ermöglicht haben.

V Lebenslauf

VI Eigene Veröffentlichungen

1. Kunz, F., Stellzig-Eisenhauer, A., Widmaier, L. M., Zeman, F., Boldt, J., *Assessment of the quality of different commercial providers using artificial intelligence for automated cephalometric analysis compared to human orthodontic experts.*

   Das Paper wurde im Juni 2023 durch das Journal of Orofacial Orthopedics / Fortschritte der Kieferorthopädie zur Publikation angenommen.

2. Kunz, F., Widmaier, L., Boldt, J., Keß, S., Zeman, F., Stellzig-Eisenhauer, A., *Untersuchung der Auswertequalität kommerzieller Anbieter für KI-basierte FRS-Analysen im Vergleich zu einem Experten-Goldstandard.*

   Dieser wissenschaftliche Beitrag wurde im Rahmen der 95. Jahrestagung der DGKFO (Deutsche Gesellschaft für Kieferorthopädie e.V.) im September 2023 als Posterbeitrag angenommen.