Aus der Klinik und Poliklinik für Radiologie

Klinikum der Ludwig-Maximilians-Universität München

Direktor: Prof. Dr. med. Jens Ricke

Die Klinische Evaluation und Optimierung von Algorithmen der Künstlichen Intelligenz in der Diagnostischen Radiologie

Kumulative Habilitationsschrift

zur Erlangung der Venia Legendi

im Fach „Experimentelle Radiologie“

vorgelegt von

Dr. med. Johannes Rückel, B.Sc.

aus Bremen

2022

Inhaltsverzeichnis

1. Einleitung

Der Einsatz künstlicher Intelligenz [*artificial intelligence*; AI] zur automatisierten Mustererkennung in radiologischen Bilddaten verbreitete sich ab den 1980/90er Jahren in Form von „CAD-Systemen“ (*computer-aided diagnosis;* CAD) (1), dabei vorzugsweise bei höhergradig standardisierten Untersuchungen wie beispielsweise Mammographien (2) oder der Detektion pulmonaler Rundherde in thorakalen Computertomographien (3). So wurden CAD-Systeme beispielsweise in den USA im Jahr 2008 bereits in ca. 70% aller Mammographie-Studienauswertungen genutzt (2), wobei der diagnostische Mehrwert jedoch selbst bei dieser hochstandardisierten Aufnahmetechnik umstritten blieb (1): Verschiedene Studien belegten bestenfalls keinen eindeutigen Mehrwert (4), eine teils reduzierte diagnostische Genauigkeit von Radiologen durch Zuhilfenahme einer CAD-Assistenz (5), häufigere Verlaufskontrollen sowie bioptische Probenentnahmen (6) oder auch einen erhöhten Zeitaufwand der Befundung bei CAD-assistierter Bildinterpretation (7).

Während der konkrete klinische Mehrwert von CAD-Systemen in vielen klinischen Anwendungen umstritten bleiben, nahm das grundsätzliche Interesse an einer (semi-)automatisierten oder Software-assistierten radiologischen Bildinterpretation stetig zu. Beispielsweise Bezug nehmend auf Zahlen aus England stieg hier im Zeitraum 2012-2015 die Anzahl ausgebildeter Radiologen um 5%, dem gegenüber nahm jedoch die Anzahl durchgeführter CT- / MRT-Untersuchungen disproportional um 29% / 26% zu [The Royal College of Radiologists 2016] (8). Notwendigerweise verkürzte Interpretationszeiten pro Bild sind mit höheren Fehlerquoten assoziiert (9); allgemein wird der Anteil fehlerbehafteter Befunde auf ca. 3-5% geschätzt (10). Der Einsatz künstlicher Intelligenz zur assistierten Bildinterpretation verspricht diese Lücke zu füllen - zwischen einerseits einem notwendigerweise effizienterem klinisch-radiologischem Workflow bei steigenden Untersuchungszahlen und andererseits den unverändert hohen Qualitätsansprüchen an die radiologische Bildinterpretation.

Maschinelles Lernen basiert auf neuronalen Netzwerken, bestehend aus einer Eingangsschicht (*„input layer“)*, einer oder mehrerer Zwischenschichten (*„hidden layer“)*

sowie einer Ausgangsschicht („output *layer")*. Jeder dieser Schichten wiederum besteht aus sog. künstlichen Neuronen, die mit denselben der vor- und nachgeschalteten Schichten über Aktivierungsfunktionen verknüpft sind. Die über die Jahre deutliche Zunahme der verfügbaren Rechenleistung ermöglichte zuletzt das Einführen und Trainieren zunehmend komplexer neuronaler Netzwerke; die Komplexität hierbei charakterisiert durch die hohe Anzahl hintereinander geschalteter Zwischenschichten. Die so erreichte komplexe Informationsweitergabe durch das neuronale Netzwerk prägte den Begriff „*deep learning*". Hintergrund des maschinellen „Lernens" ist dabei, dass neuronale Netzwerke basierend auf bekannten Trainingsdaten die eigenen, intrinsischen Aktivierungsfunktionen zwischen den neuronalen Schichten ausgerichtet am vorliegenden Referenzstandard der Trainingsdaten selbst anpassen. Auf diese Weise können neuronale Netzwerke beispielsweise auf die Klassifikation von radiologischen Bilddaten entsprechend definierten Merkmalen (z. B. abgebildeten Pathologien) trainiert werden.

Entgegen der eingangs erwähnten etablierten CAD-Systeme, versprechen moderne *Deep-Learning*-Algorithmen eine deutlich höhere Leistungsfähigkeit. Eine Vielzahl hochrangig publizierter Studien konnte inzwischen klinische Anwendungen demonstrieren, in denen die Genauigkeit einer maschinellen Bildklassifikation durch entsprechend vortrainierte neuronale Netze jener einer visuellen Analyse durch medizinische Experten gleichkommt: So beispielsweise hinsichtlich der Detektion diabetischer Netzhauveränderungen oder Papillenschwellungen in Fundusfotografien (11,12), hinsichtlich der Klassifikation von Hauttumoren (13), der Detektion kritischer Befunde in Computertomographien (CT) des Schädels (14) oder hinsichtlich der Brustkrebserkennung in Screening-Mammographien (15,16).

Während leistungsstarke *Deep-Learning*-Algorithmen zur Mustererkennung in radiologischen Bilddaten vielversprechend sowohl hinsichtlich *Workflow*-Optimierung (z. B. Arbeitslistentriagierung zur priorisieren Befundung auffälliger Untersuchungen) als auch diagnostischer Qualität (z. B. Reduktion übersehener Befunde oder Reduktion der je nach radiologischer Untersuchung teils beträchtlichen *Inter-Reader*-Variabilität) sind, bleiben Limitationen und rechtliche Hürden zu bedenken: So bleibt die Nachvollziehbarkeit, anhand welcher Bildmerkmale eine algorithmische Entscheidungsfindung zu Stande kommt,

beispielsweise häufig eingeschränkt. Auch die Kenntnis über die Algorithmus-*Performance* in verschiedenen relevanten Patienten-Subgruppen bleibt essenziell, um ggf. bestimmte Bilddaten einer automatisierten Befundung vorenthalten zu können (z. B. im Falle von im Bild enthaltenen *Confoundern*, welche die Algorithmen mutmaßlich deutlich beeinflussen können). Robuste Algorithmen müssen weiterhin geräte- und institutsunabhängig eine leistungsstarke Bildklassifikation anbieten können – diese ist jedoch maßgeblich von den verwendeten Trainingsdaten und den hierbei verwendeten Referenz-*Labels* abhängig. Vor dem Hintergrund dieser Limitationen ist vor einem Einsatz in der klinischen Routine eine hinreichende externe Validierung diagnostischer Algorithmen notwendig.

Ziel dieser Habilitationsschrift ist es, im Bereich der thorakalen Bildgebung moderne *Deep-Learning*-Algorithmen der künstlichen Intelligenz klinisch zu evaluieren. Einerseits geht es dabei um die Identifikation klinischer Anwendungsgebiete, innerhalb derer eine Softwareassistenz einen klinischen Mehrwert in der diagnostischen Bildgebung verspricht. Andererseits geht es gleichermaßen um die Identifikation algorithmischer Schwachstellen sowie um deren entwicklungstechnische Adressierung, um den Weg letzten Endes diagnostisch robuster *Deep-Learning*-Algorithmen in die klinische Routine der diagnostischen Radiologie voranzutreiben.

Fokussierend auf die konventionelle Projektionsradiographie als auch auf die Computertomographie des Thorax werden im Folgenden die einschlägigen Originalarbeiten als kumulative Habilitationsschrift inhaltlich zusammengefasst sowie in den wissenschaftlichen Kontext eingeordnet.

2. Künstliche Intelligenz im konventionellen Röntgen-Thorax

2.1 Detektion von Pneumothoraces: Einliegende Thorax-Drainagen als *Confounder*

Impact of Confounding Thoracic Tubes and Pleural Dehiscence Extent on Artificial Intelligence Pneumothorax Detection in Chest Radiographs

Rueckel J, Trappmann L, Schachtner B, Wesp P, Hoppe BF, Fink N, Ricke J, Dinkel J, Ingrisch M, Sabel BO.

Invest Radiol. 2020 Dec; 55(12):792–8.

DOI: 10.1097/RLI.0000000000000707

Diverse *Deep-Learning*-Algorithmen zur automatisierten Röntgen-Thorax-Analyse wurden auf Basis öffentlich verfügbarer Datensätze mit jeweils mehreren Hunderttausenden Bilddaten entwickelt und validiert, so beispielsweise auf Grundlage der öffentlich verfügbaren Datensätze *ChestX-ray 14* oder *PLCO* (Prostate, Lung, Colorectal and Ovarian cancer screening trial) (17,18). Diese öffentlich verfügbaren Bilddatensätze enthalten Referenz-*Labels* für verschiedene Pathologien (nicht jedoch für miterfasstes Fremdmaterial). Die Qualität und Einheitlichkeit der Referenz-*Labels* (meist automatisiert extrahiert aus zugehörigen radiologischen Befunden, z. B. durch *natural language processing* [NLP]) wird als deutliche Limitation dieser Datensätze angesehen. Auch wird in diesen Datensätzen nicht zwischen posterior-anterioren (stehender Patient) und anterior-posterioren (liegender Patient) Projektionsradiographien unterschieden. Trotzdem erreichen auf diesen Datensätzen trainierte und validierte Algorithmen eine teils vielversprechende Leistungsfähigkeit hinsichtlich der Bildklassifikation entsprechend verschiedener Pathologien, so z. B. der ebenfalls öffentlich verfügbare Algorithmus „CheXNet“ (19,20). Hierbei wird jedoch wie meist üblich auf einem vom Training ausgeschlossenen Anteil der öffentlichen Bilddatensätze validiert – unter Verwendung gleichermaßen zustande gekommener Referenz-*Labels* sowie ohne die Möglichkeit relevanter Subgruppenanalysen.

Die o. g. Originalarbeit etabliert eine externe Validierungskohorte bestehend aus 1.652 Pneumothorax-positiven Liegend-Röntgen-Thoraces (*supine chest X-ray*, SCXR) sowie 4.782

Pneumothorax-negativen SCXRs. Alle Projektionsradiographien wurden seitengetrennt hinsichtlich miterfasster einliegender Thoraxdrainagen annotiert; die vorliegenden Pneumothoraces wurden weiterhin hinsichtlich ihrer Größe (entsprechend der maximalen pleuralen Dehiszenz) in drei Subgruppen (<1cm, 1-2cm, >2cm) unterteilt. Erwartungsgemäß war in dem so etablierten Validierungsdatensatz der Anteil einliegender Thoraxdrainagen in den Pneumothorax-positiven SCXRs (73,3% für einseitige Pneumothoraces) deutlich höher als der in den Pneumothorax-negativen SCXRs (13,1 %) – vergleichbare Angaben bezüglich der verwendeten öffentlichen Trainingsdaten sind mangels entsprechenden Fremdmaterial-assoziierten Annotationen nicht verfügbar. Basierend dem etablierten Datensatz (siehe Tabelle 1) erfolgte die externe Validierung inkl. Subgruppenanalysen für zwei verschieden Algorithmen, die auf eingangs erwähnten öffentlichen Datensätzen trainiert worden sind.

Verglichen mit der entsprechenden Originalarbeit Rajpurkar et al. erreichte der Algorithmus „CheXNet“ hinsichtlich der Detektion von Pneumothoraces eine Subgruppen-übergreifend reduzierte Performance auf dem etablierten Validierungsdatensatz mit einer *area under receiver operating characteristics* (AUROC) von 0.765 (0.750 – 0.779), verglichen mit einer AUROC von 0,8887 entsprechend Rajpurkar et al.. Weiterhin zeigte sich eine deutliche Abhängigkeit der erreichbaren AUROCs von der Größe der zu detektierenden Pneumothoraces (siehe Abbildung 1): Während Pneumothoraces <1cm mit einer AUROC von 0.722 (0.698 – 0.745) detektiert wurden, erreichte „CheXNet“ für die Detektion größerer Pneumothoraces >2cm eine AUROC von 0.818 (0.796 – 0.840). Weiterhin wurden Thoraxdraingen als massiver *Confounder* identifiziert (siehe Abbildung 1): So konnte „CheXNet“ solche Bilder mit Pneumothoraces und einliegender Thoraxdrainage gegenüber Kontrollfällen ohne Pneumothorax sowie ohne einliegende Thoraxdrainage mit einer AUROC von bis zu 0.875 (0.854 – 0.897) identifizieren. Gegensätzlich konnten jedoch selbst große Pneumothoraces > 2cm ohne einliegende Thoraxdrainagen nicht gegenüber Pneumothorax-negativen Kontrollbildern mit einliegenden Thoraxdrainagen differenziert werden (AUROC 0.550 [0.495 – 0.605]). Dieselben Effekte sowohl hinsichtlich der Größe vorhandener Pneumothoraces also auch hinsichtlich einliegender Thoraxdrainagen konnten sowohl für „CheXNet“ (trainiert auf dem Datensatz ChestX-ray 14) also auch für einen zweiten auf öffentlichen Datensätzen trainierten Algorithmus (trainiert auf den Datensätzen ChestX-ray 14 und PLCO) demonstriert werden.

	TT		
	Yes, n (%)	**No, n (%)**	**Sum, n (%)**
Unilateral PTX (n = 1476)			
A Dehiscence <1 cm	446 (72.4)	170 (27.6)	616 (41.7)
B Dehiscence 1–2 cm	341 (76.1)	107 (23.9)	448 (30.5)
C Dehiscence >2 cm	295 (71.6)	117 (28.4)	412 (27.8)
Sum, n (%)	1082 (73.3)	394 (26.7)	
Bilateral PTX (n = 176)			
A Max. dehiscence <1 cm	36 (78.2)	10 (21.8)	46 (26.1)
B Max. dehiscence 1–2 cm	62 (96.9)	2 (3.1)	64 (36.4)
C Max. dehiscence >2 cm	56 (84.8)	10 (15.2)	66 (37.5)
Sum, n (%)	154 (87.5)	22 (12.5)	
Control cases (n = 4782)			
PTX-negative	627 (13.1)	4155 (86.9)	4782

PTX-positive cases are radiologically annotated for PTX size, PTX location (unilateral vs bilateral), and inserted TTs. PTX-negative control cases are radiologically annotated for inserted TTs.

Abbreviations: PTX, pneumothorax; TT, thoracic tube.

***Tabelle 1**: Validierungskohorte mit Annotationen entsprechend der Größe von einseitigen / beidseitigen Pneumothoraces sowie mitabgebildeten Thoraxdrainagen in den zugrundeliegenden Röntgen-Thorax-Aufnahmen. Aus Rueckel et al. Investigative Radiology 2020. Abkürzungen: Pneumothorax – PTX; TT – Thoraxdrainage / thoracic tube.*

Damit konnte in o. g. Originalarbeit gezeigt werden, dass unter Verwendung verbreiteter öffentlicher Datensätze vermeintliche auf die Detektion von Pneumothoraces trainierte Algorithmen zugleich akzidentiell auf die Detektion von Thoraxdrainagen trainiert werden. Dies unterstreicht den unverzichtbaren Wert hochqualitativ annotierter Trainingsdaten sowie die Gefahr, die von möglicherweise auch unbekannten *confoundern* in verwendeten Trainingsdatensätzen für *Deep-Learning*-Algorithmen ausgeht – insbesondere bei Verwendung von Algorithmusarchitekturen, die keine örtliche Zuordnung im Bild zu jenen für die Algorithmus-Entscheidung relevanten Bildbereich erlauben.

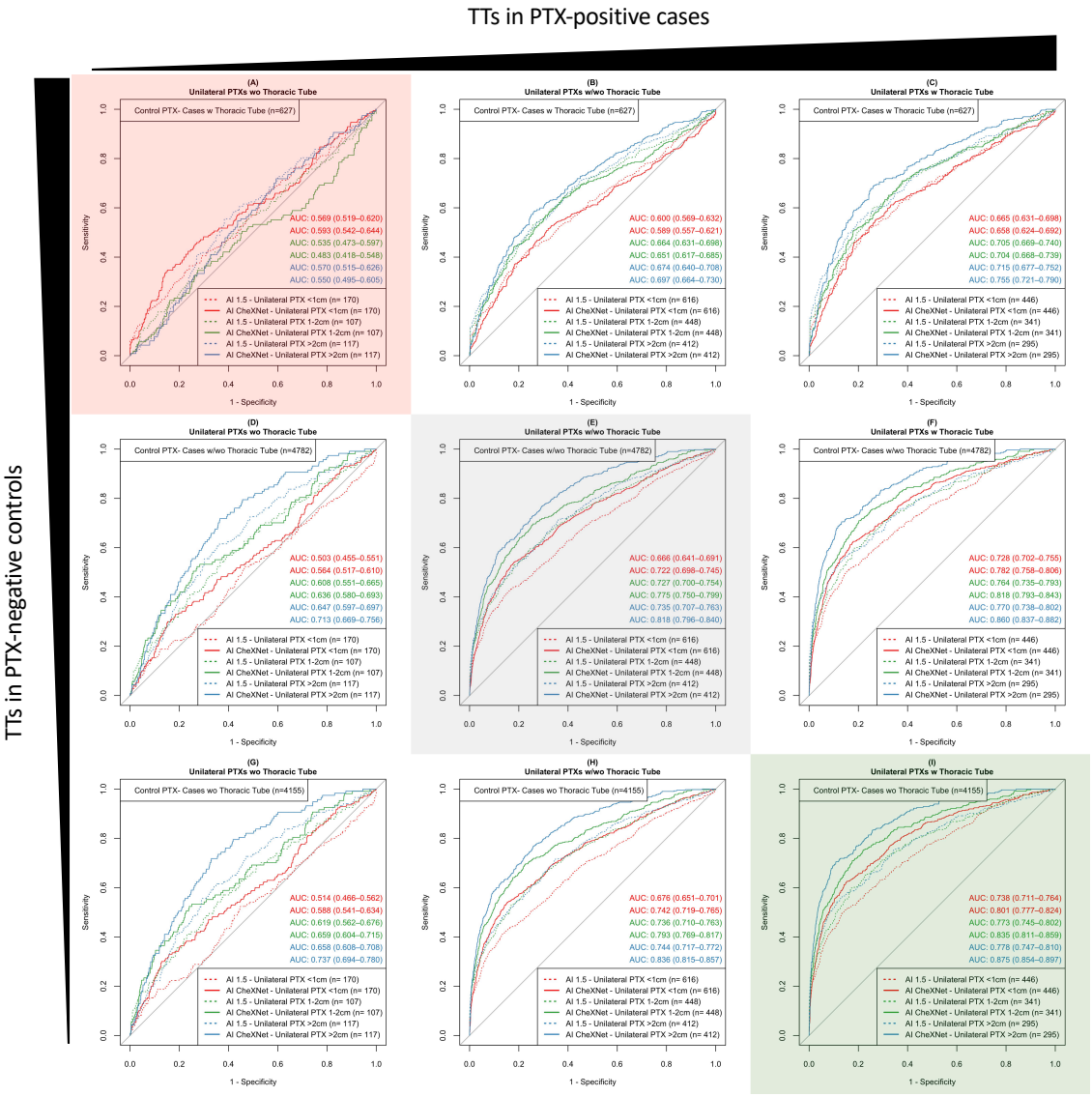

Abbildung 1: *Leistungsfähigkeit zweier Algorithmen (AI_CheXNet, AI_1.5) hinsichtlich der Pneumothorax-Detektion, quantifiziert mittels area under receiver operating characteristics (AUC, AUROC) in Abhängigkeit einliegender Thoraxdrainagen in Pneumothorax-positiven Fallbildern und/oder Pneumothorax-negativen Kontrollbildern: Die erreichbaren AUROCs korrelieren positiv mit dem Anteil einliegender Thoraxdrainagen in Pneumothoax-positiven sowie negativ mit dem Anteil einliegender Thoraxdrainagen in Pneumothorax-negativen Röntgenbildern. Damit konnte demonstriert werden, dass beide Algorithmen zumindest teilweise für die Detektion einliegender Thoraxdrainagen (neben vorhandenen Pneumothoraces) trainiert worden sind – ein mutmaßlich akzidenteller Effekt durch mangelnde Berücksichtigung dieses Confounders in den verwendeten Trainingsdaten. Aus Rueckel et al. Investigative Radiology 2020. Abkürzungen: Area under receiver operating characteristics – AUC; Pneumothorax – PTX.*

2.2 Detektion von Pneumothoraces: Algorithmus-Optimierung

Pneumothorax detection in chest radiographs: optimizing artificial intelligence system for accuracy and confounding bias reduction using in-image annotations in algorithm training

Rueckel J, Huemmer C, Fieselmann A, Ghesu FC, Mansoor A, Schachtner B, Wesp P, Trappmann L, Munawwar B, Ricke J, Ingrisch M, Sabel BO.

Eur Radiol. 2021 Mar 27.

DOI: 10.1007/s00330-021-07833-w

Die in vorherigem Abschnitt ausgeführten Limitationen der automatisierten Pneumothorax-Detektion in Abhängigkeit der Größe der pleuralen Dehiszenz sowie in Abhängigkeit einliegender Thoraxdrainagen als relevanter *confounder* wurden in dieser Originalarbeit durch algorithmische Verbesserungen begegnet. Die Hälfte der Bilder jeder Subgruppe (Subgruppen entsprechend der Größe der Pneumothoraces sowie einliegender Thoraxdrainagen) der zuvor (siehe 2.1, Tabelle 1) etablierten SCXR-Kohorte wurde dabei dem Algorithmus-Training zugeführt; die andere Hälfte der Bilder verblieb zur Algorithmus-Validierung. Basierend auf dieser verbliebenen Validierungskohorte wurden in der Studie vier verschiedene Algorithmen verglichen: Einerseits erneut der öffentlich verfügbare Algorithmus „CheXNet“ als etablierter Benchmarking-Algorithmus. Andererseits drei schrittweise verbesserte Algorithmen [Prototypen, *Siemens Healthineers*], deren zugrundeliegende Trainingsdaten schrittweise weg von öffentlich verfügbaren Bilddaten / Referenz-Labels hin zu händisch annotierten Bilddaten verschiedener Institutionen adjustiert wurden. Außerdem wurde im letzten Schritt eine Algorithmus-Architektur etabliert, die im Algorithmus-Training eine Berücksichtigung Pathologie-assoziierter Pixelannotationen im Bild erlaubt. So wurden im zuletzt etablierten Algorithmus ausschließlich solche Pneumothorax-positiven SCXRs verwendet, deren pleurale Dehiszenzlinie über annotierte Pixelkoordinaten definiert wurde (siehe Abbildung 2). Dieselbe Algorithmus--Architektur erlaubt ebenso eine örtliche Zuordnung später detektierter Pathologien im Bild.

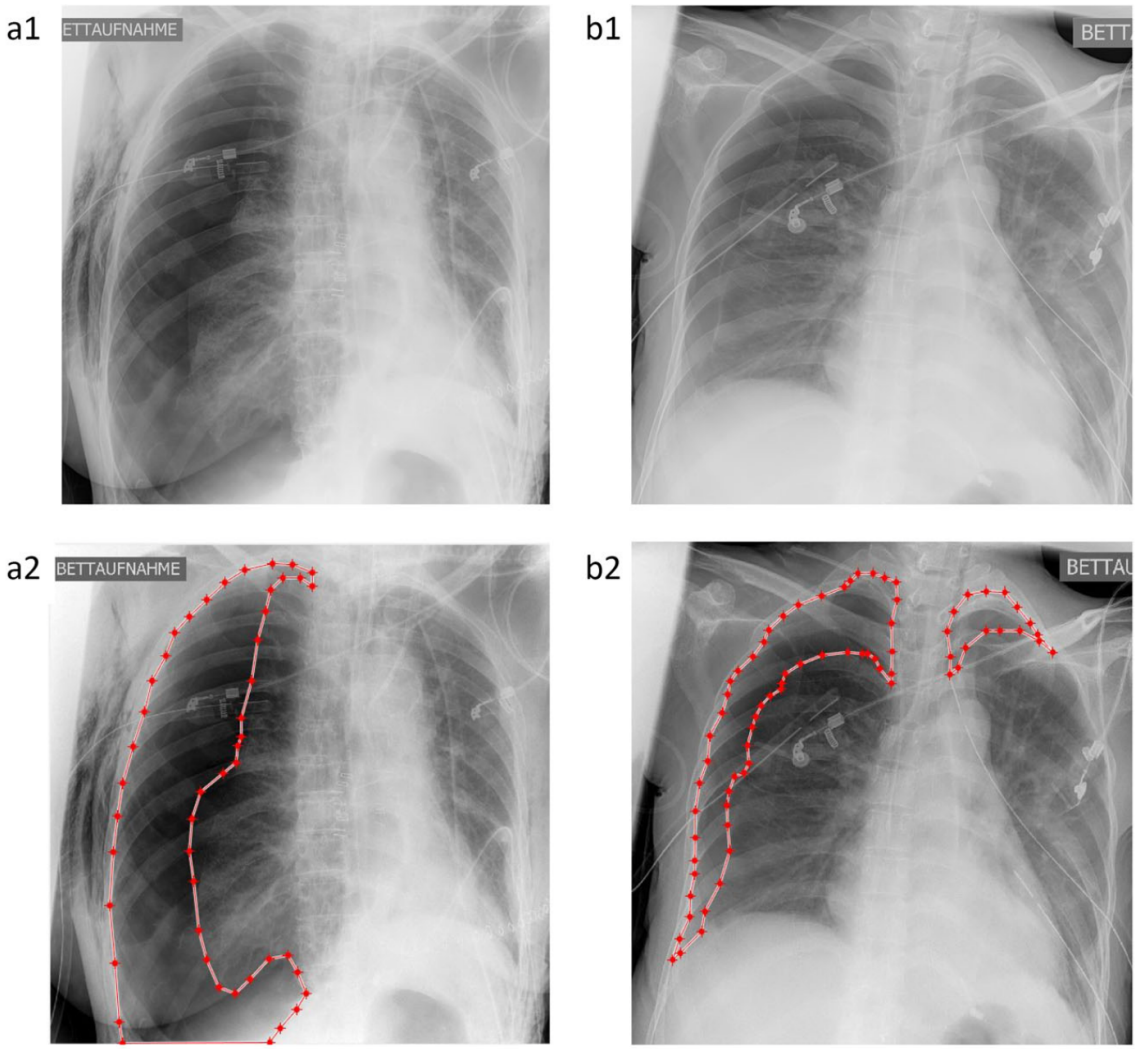

Abbildung 2: *Annotierte Pixelkoordinaten der Thoraxwand (Pleura parietalis) sowie der pleuralen Dehiszenzlinie (Pleura visceralis); das über die Pixelkoordinaten definierte Polygon wurde im Algorithmus-Training berücksichtigt. Aus Rueckel et al. European Radiology 2021.*

Es konnte im Vergleich aller Algorithmen gezeigt werden, dass die Einbeziehung von Pixelannotationen der pleuralen Dehiszenzlinie im Algorithmus-Training sowohl die generelle Leistungsfähigkeit der Pneumothorax-Detektion signifikant steigert als auch den *confounding bias* durch einliegende Thoraxdrainagen partiell unterdrückt (siehe Abbildung 3). Insbesondere der Vergleich der beiden zuletzt etablierten Algorithmen („Algorithmus 1“ und „Algorithmus 2“) mit gleicher Netzwerkarchitektur jedoch unterschiedlicher Zusammensetzung der verwendeten Trainingsdaten offenbarte diesen Effekt deutlich: Obwohl die Gesamtzahl der dem Training zugeführten Bilddaten für „Algorithmus 2“ deutlich geringer war (n=75.067 vs n=112.120), führten die zusätzlichen konsequenten Pixelannotionen der pleuralen Dehiszenzlinien im Algorithmus-Training zu einer deutlich gesteigerten resultierenden Algorithmus-Performance (AUROCs 0,877 [0.861 – 0.893] vs

0,726 [0.703 – 0.748]). Auch der *confounding bias* durch einliegende Thoraxdrainagen ließ sich zumindest reduzieren: So war „Algorithmus 1“ weiterhin nicht in der Lage Pneumothoraces ohne einliegende Thoraxdrainagen gegenüber Pneumothorax-negativen SCXRs mit einliegenden Thoraxdrainagen zu detektieren (AUROCs 0.491-0.587), während „Algorithmus 2“ in diesem Setting zumindest Pnumothoraces mit einer pleuralen Dehiszenz > 2cm mit einer AUROC von 0.921 [0.871 – 0.971] detektieren konnte (siehe Abbildung 3a).

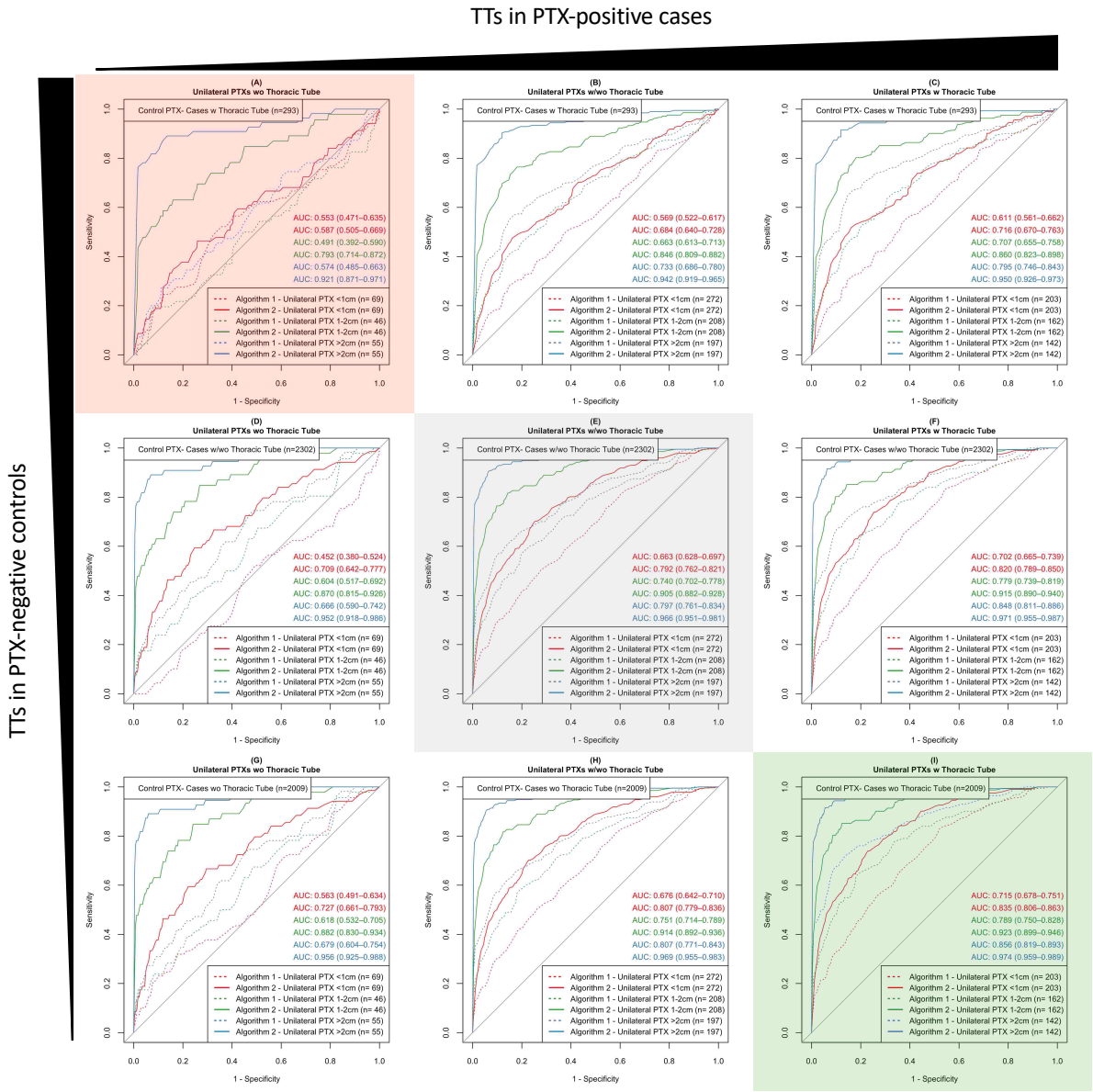

Abbildung 3: *Auswertung äquivalent zu Abbildung 2. Die Berücksichtigung von Pneumothorax-Annotationen im Training von Algorithmus 2 (durchgezogene Linien) verbessert gegenüber Algorithmus 1 (keine Pixelannotationen im Training berücksichtigt, gestrichelte Linien) deutlich die Leistungsfähigkeit der Pneumothorax-Detektion und reduziert deren Abhängigkeit von einliegenden Thoraxdrainaen. Aus Rueckel et al. European Radiology 2021. Abkürzungen: Area under receiver operating characteristics – AUC; Pneumothorax – PTX.*

Letztlich konnte durch die Studie insbesondere gezeigt werden, dass sich Effekte von *confoundern* in den Trainingsdaten durch direkte Pathologie-assoziierte Pixelannotationen in den Trainingsdaten supprimieren lassen. Dies ist insbesondere relevant, da Thoraxdrainangen als *confounder* der Pneumothorax-Detektion nur aufgrund der klinischen Offensichtlichkeit bekannt geworden sind (siehe Abschnitt 2.1), jedoch generell auch von unbekannten *confoundern* in Trainingsdaten ausgegangen werden muss – und sich auch deren Effekte mutmaßlich supprimieren lassen, wenn nicht der (unbekannte) *confounder* selbst, sondern die zu detektierende Pathologie für das Algorithmus-Training annotiert wird.

2.3 Detektion von Pneumonien: Algorithmus-Evaluation mit CT als Referenzstandard

---

Artificial Intelligence Algorithm Detecting Lung Infection in Supine Chest Radiographs of Critically Ill Patients With a Diagnostic Accuracy Similar to Board-Certified Radiologists

Rueckel J, Kunz WG, Hoppe BF, Patzig M, Notohamiprodjo M, Meinel FG, Cyran CC, Ingrisch M, Ricke J, Sabel BO.

Crit Care Med. 2020 May 20

DOI: 10.1097/CCM.0000000000004397

Die Detektion von Pleuraergüssen sowie insbesondere Pneumonie-suspekten Konsolidierungen und Belüftungsstörungen in Röntgen-Thoraces unterliegt vor allem bei anterior-posterioren Liegenaufnahmen einem relevanten Interpretationsspielraum und ist entsprechend durch eine hohe Inter-*Reader*-Variabilität und letztlich reduzierte diagnostische Genauigkeit limitiert (21). Für einige AI-Algorithmen konnten hinsichtlich der Detektion von Pneumonien oder Belüftungsstörungen bereits mit Radiologen vergleichbare diagnostische Genauigkeiten demonstriert werden, dabei jedoch weiterhin basierend auf öffentlichen Datensätzen mit den hier verfügbaren Referenz-*Labels* sowie ohne dezidierte Unterscheidung zwischen Stehend- und Liegendaufnahmen (19,22). In o. g. Originalarbeit wurde ein ebenfalls auf öffentlichen Datensätzen trainierter Algorithmus (Datensätze: ChestX-ray 14 & PLCO, insgesamt 297.541 Bilddatensätze) extern validiert [Prototyp, *Siemens Healthineers*] und mit der Performance radiologischer Fachärzte auf einem instituts-internen Validierungsdatensatz verglichen; dabei neu im wissenschaftlichen Kontext unter Verwendung von Computertomographien als Referenzstandard sowie mit einem Fokus auf Liegenaufnahmen (mehrheitlich akquiriert auf der Intensivstation).

Retrospektiv eingeschlossen wurden 166 Patienten, die innerhalb von 90 Minuten einer Röntgen-Thorax-Liegendaufnahme sowie einer Computertomographie (CT) ohne zwischenzeitliche Intervention unterzogen wurden. Radiologische Fachärzte beurteilten die SCXRs hinsichtlich Pleuraergüssen und Pneumonie-suspekten Konsolidierungen auf einer dreistufigen Lickert-Skala. Die ärztliche Performance wurde mit der algorithmischen

Genauigkeit verglichen, dabei basierend auf der ärztlicherseits unabhängig erfolgten Evaluation der zugehörigen CTs als Referenzstandard (siehe Abbildung 4).

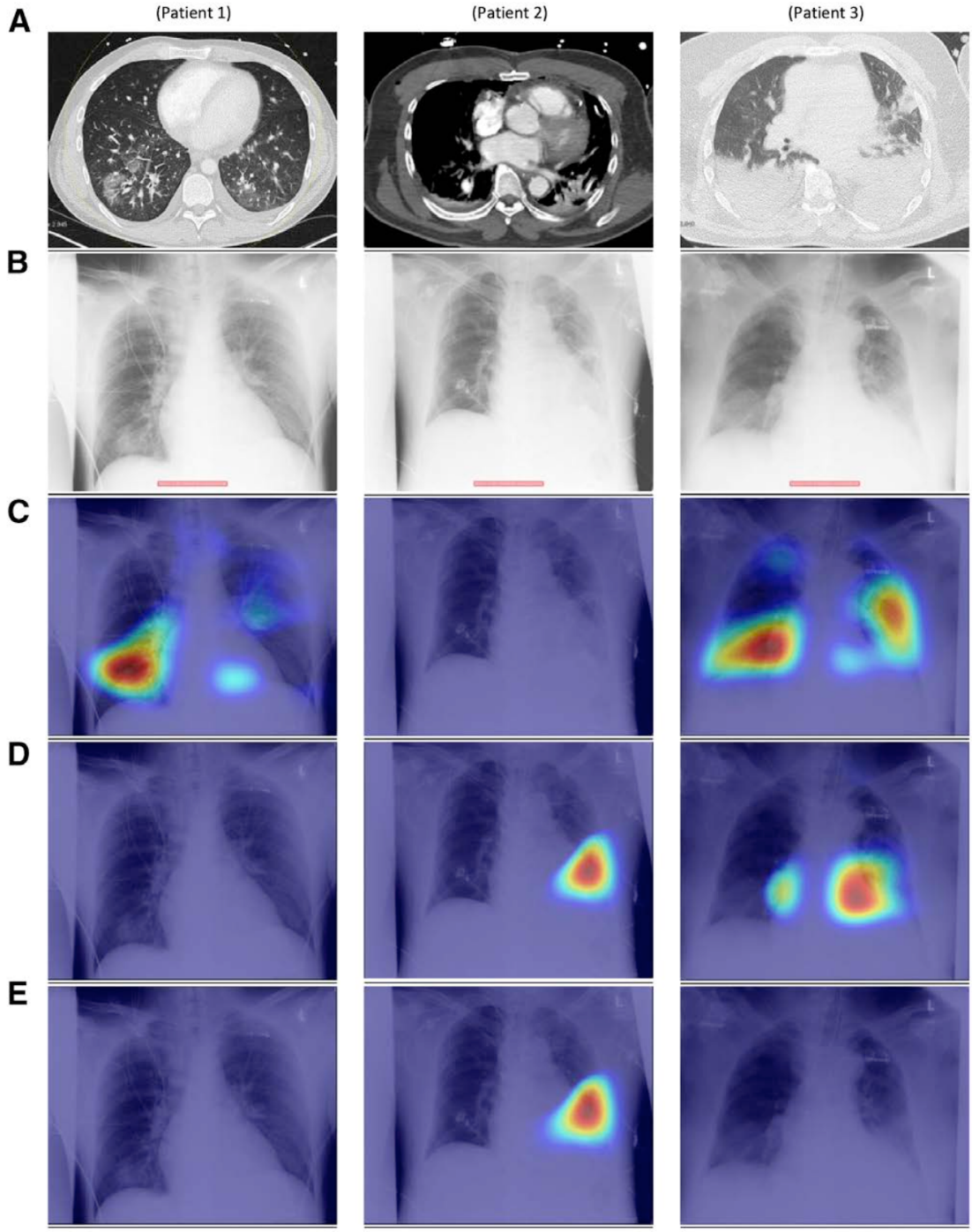

***Abbildung 4**: Algorithmische Analyse von Röntgen-Thoraces (B) mit vorliegenden Computertomographien des Thorax (A) als Referenzstandard; Heat-Map-basierte Darstellung jener Bildbereiche mit erhöhter CNN-Aktivität hinsichtlich der Bildklassifikation bzgl. Pneumonie-suspekten Konsolidierungen (C), Pleuraergüssen (D) und Atelektasen (E). Bildbeispiele: (Patient 1) - Bilaterale Pneumonie, CT-morphologische Milchglastrübungen; (Patient 2) - Diskrete Pleuraergüsse mit angrenzenden Belüftungsstörungen; (Patient 3) – Bilaterale Pleuraergüsse mit Belüftungsstörungen und linksseitig führend Pneumonie-suspekten Veränderungen. Aus Rueckel et al. Critical Care Medicine 2020. Abkürzungen: Convolutional Neural Network - CNN.*

Demonstriert werden konnte zunächst im Vergleich zur CT die deutlich eingeschränkte diagnostische Güte der SCXR-Befundung: Für die Pneumonie-Detektion erreichten radiologische Fachärzte eine maximale Genauigkeit von 73% (Sensitivität 69%, Spezifität 74%), für Pleuraergüsse eine maximale Genauigkeit von 68% (Sensitivität 67%, Spezifität 69%). Sowohl gemessen an den AUROCs als auch gemessen an den diagnostischen Metriken definierter ROC-Operationspunkte unterschied sich die ärztliche jedoch nicht signifikant von der algorithmischen Performance: AUROCs 0,779 (0,723 – 0,836) vs 0,737 (0,659 – 0,815) für die Pneumonie-Detektion sowie AUROCs 0,698 (0,646 – 0,749) vs 0,740 (0,662 – 0,817) für die Detektion von Pleuraergüssen. Die vom Algorithmus erreichten AUROCs waren dabei erwartungsgemäß geringer bei CT-basiertem Referenzstandard verglichen mit einem Referenzstandard basierend auf der visuell-radiologischen SCXR-Interpretation. Gemessen am zuletzt genannten Referenzstandard lässt ich die Performance des in vorliegender Studie charakterisierten Algorithmus mit vergleichbar validierten Algorithmen im wissenschaftlichen Kontext vergleichen: Dabei übertraf die Performance des in dieser Originalarbeit validierten Algorithmus hinsichtlich der Pneumonie-Detektion drei von vier gleichermaßen auf öffentlichen Datensätzen trainierten und anderweitig publizierten Algorithmen – obwohl gegensätzlich zu den Vergleichsstudien (17,19,22,23) die Validierung o. g. Studie auf die mutmaßlich schwieriger interpretierbaren SCXRs (Liegendaufnahmen) limitiert war. Gemessen an entsprechend *Youden*-Statistik [maximale Summe aus Sensitivität und Spezifität, (24)] optimierten algorithmischen Operationspunkten wurden vier Patienten mit CT-morphologisch bestätigter Pneumonie vom Algorithmus, nicht jedoch entsprechend der visuell-radiologischen SCXR-Interpretation identifiziert. Gleichermaßen identifizierte der Algorithmus im Gegensatz zu den radiologischen Fachärzten sechs zusätzliche Patienten mit CT-morphologisch bestätigten Pleuraergüssen. Selbst verglichen mit radiologischen Fachärzten – und mutmaßlich deutlich ausgeprägter verglichen mit nicht-radiologisch spezialisierten Klinikern sowie unter Verwendung weiter optimierter Algorithmen – kann daher ein Mehrwert durch eine z. B. AI-assistierte SCXR-Interpretation unterstellt werden.

2.4 Künstliche Intelligenz zur Röntgen-Thorax-Interpretation in der Notaufnahme

---

Artificial Intelligence in Chest Radiography Reporting Accuracy - Clinical Value in the Emergency Unit Setting without 24/7 Radiology Coverage

Rudolph J, Huemmer C, Ghesu FC, Mansoor A, Preuhs A, Fieselmann A, Fink N, Dinkel J, Koliogiannis V, Schwarze V, Goller S, Fischer M, Jörgens M, Khaled NB, Vishwanath RS, Balachandran A, Ingrisch M, Ricke J, Sabel BO, **Rueckel J.**

Invest Radiol. 2022 Feb; 57 (2) – (ePub ahead of print)

DOI: 10.1097/CCM.0000000000004397

Mit der Befundung von Röntgen-Thorax-Aufnahmen sind nicht ausschließlich Radiologen, sondern z. B. in kleineren Versorgungsstrukturen ohne durchgehende Betreuung durch eine radiologische Fachabteilung ebenso klinisch tätige Ärzte ohne dezidierte radiologische Weiterbildung konfrontiert. Diagnostisch leistungsstarke AI-Algorithmen zur Detektion der wichtigsten Pathologien können hier als Entscheidungsunterstützung für klinische Kollegen einen relevanten Mehrwert erbringen. In o. g. Originalarbeit wurde ein AI-System zur Detektion von Pneumonie-suspekten Konsolidierungen, Pleuraergüssen, Pneumothoraces sowie pulmonalen Rundherden auf Basis nicht-öffentlich verfügbarer, radiologisch annotierter Bilddatensätze trainiert. Dieses AI-System wurde in o. g. Originalarbeit unter Verwendung von 563 Röntgen-Thoraces (AP-Projektion, Untersuchungen in der Notaufnahme) validiert. Alle Röntgen-Thoraces des Validierungsdatensatzes wurden von 9 verschiedenen Ärzten (3 radiologische Fachärzte, 3 radiologische Assistenzärzte, 3 nicht-radiologische Assistenzärzte mit Notaufnahmen-Erfahrung) unter Verwendung Lickert-skalierter, Pathologie-spezifischer Konfidenz-Scores unabhängig retrospektiv ausgewertet (siehe Abbildung 5). Neu im wissenschaftlichen Kontext ist hierbei einerseits die statistische Berücksichtigung der generellen Unsicherheit der Röntgen-Thorax-Befundung unter Verwendung verschiedener Konfidenzniveaus, andererseits der Vergleich eines AI-Systems auch mit Nicht-Radiologen zur Quantifizierung eines Mehrwertes als Entscheidungsunterstützung für klinisch tätige Ärzte.

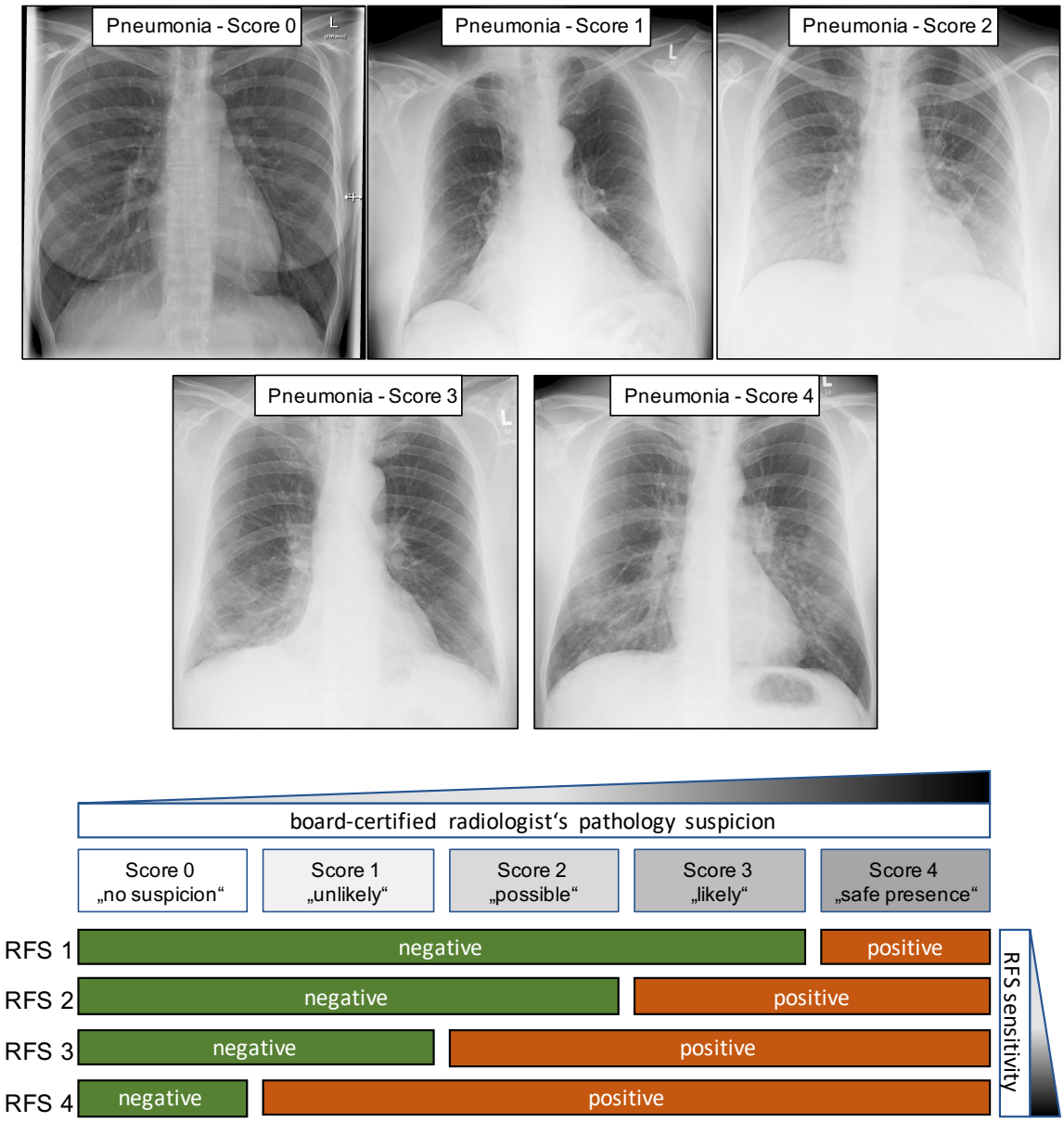

Abbildung 5 *- Darstellung der Lickert-skalierten Konfidenz-Scores für vorliegende Pathologien (hier exemplarische Bildbeispiele mit aufsteigendem Verdacht auf eine vorliegende Pneumonie). Die fünfstufigen Konfidenz-Scores der radiologischen Fachärzte wurden mittels Score-Pooling in vier verschiedene binäre Referenzstandards (RFS) unterschiedlicher Sensitivität konvertiert, letztlich verwendete Referenzstandards gebildet als Konsensus aller drei involvierten radiologischen Fachärzte. Aus Rudolph et al. Investigative Radiology 2022. Abkürzungen: Referenzstandard - RFS.*

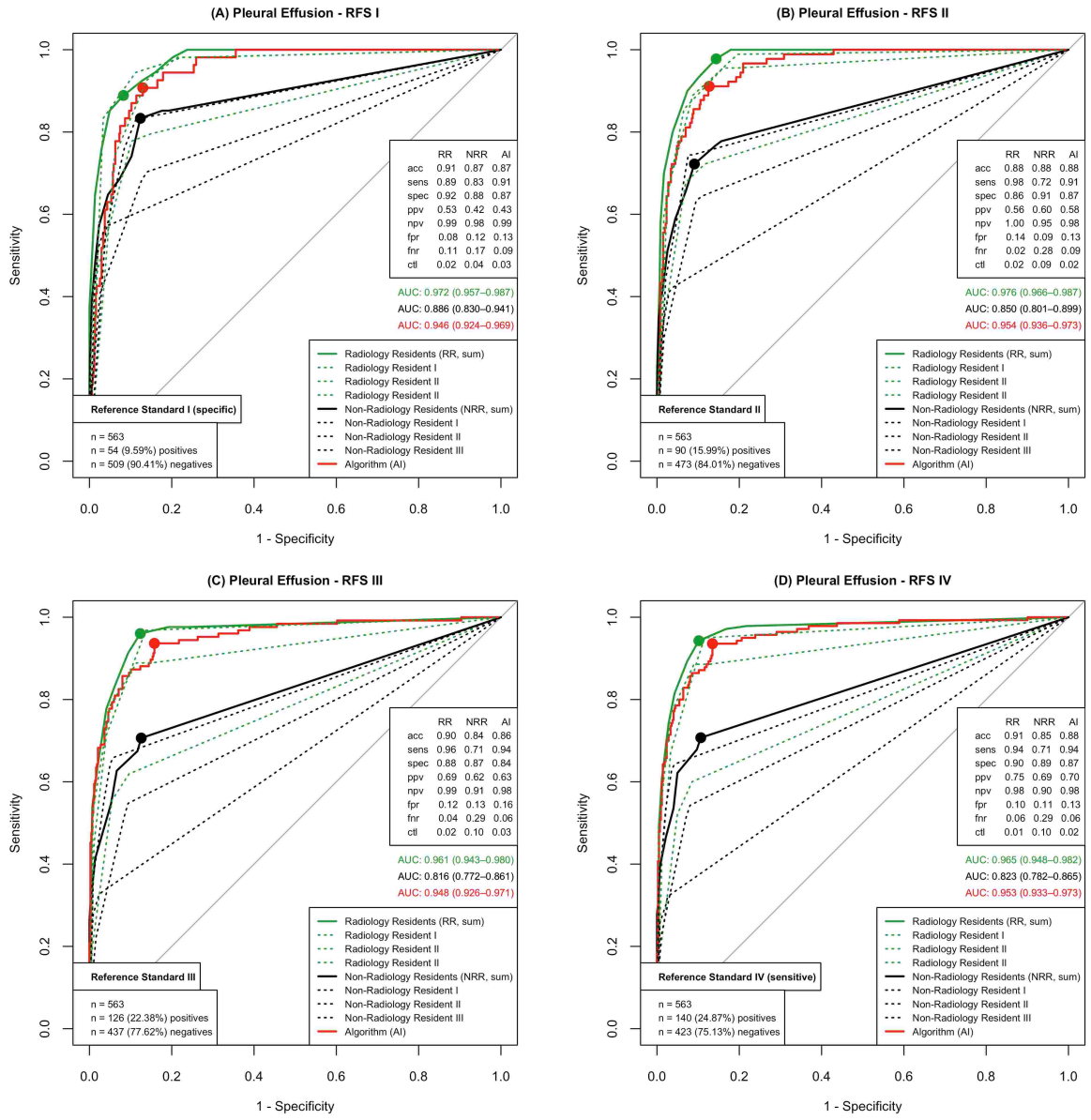

Abbildung 6: *Exemplarische Auswertung hinsichtlich der Detektion von Pleuraergüssen, dabei gemessen an aufsteigend sensitiven Referenzstandards (A-D) sowie quantifiziert mittels receiver operating characteristics: Die diagnostische Genauigkeit von Algorithmus (rot) und radiologischen Assistenzärzten (grün) bleibt unabhängig von der Sensitivität des fachärztlichen Referenzstandards (A-D) vergleichbar hoch. Die diagnostische Genauigkeit nicht-radiologischer Assistenzärzte (schwarz) ist signifikant niedriger und nimmt mit zunehmend sensitivem Referenzstandard weiter ab (A-D). Operationspunktoptimierung entsprechend maximaler Summe aus Sensitivität und Spezifität, zugehörige Metriken rechtsseitig tabellarisch dargestellt. Aus Rudolph et al. Investigative Radiology 2022 (entspricht Figures 4-B1/2/3/4, hier andere Darstellung zwecks Übersichtlichkeit). Abkürzungen: Referenzstandard - RFS; Radiology Resident (radiologischer Asssistenzart) – RR; Non-Radiology Resident (nicht-radiologischer Assistenzarzt) – NRR; Artificial Intelligence Algorithm – AI; area under receiver operating characteristics – AUC; Accuracy – acc; Sensitivität – sens; Spezifität - spec; Positiver Vorhersagewert – ppv; Negativer Vorhersagewert – npv; Falsch-Positiv-Rate – fpr; Falsch-Negativ-Rate – fnr.*

Es konnte auf dem Validierungsdatensatz gemessen am sensitivsten Facharztstandard demonstriert werden, dass das AI-System für die Detektion von Pleuraergüssen (AUC 0.953 [0.933-0.973]), Pneumothoraces (AUC 0.940 [0.893-0.988]) und pulmonalen Rundherden (AUC 0.883 [0.830-0.936]) vergleichbar mit erfahrenen radiologischen Assistenzärzten und signifikant besser verglichen mit Notaufnahme-erfahrenen Nicht-Radiologen abschneidet. Dabei schneidet hier das AI-System gegenüber den Nicht-Radiologen umso besser ab, je sensitiver der binäre Facharzt-Referenzstandard aus den mehrstufigen Lickert-basierten Konfidenz-Scores gepoolt wird (exemplarisch dargestellt für die Auswertung hinsichtlich Pleuraergüssen, siehe Abbildung 6): Es wird daher statistisch ersichtlich, dass Nicht-Radiologen von dem AI-System insbesondere bei der Detektion weniger eindeutiger / diskreter Bildbefunde profitieren würden. Hinsichtlich der Detektion von Pneumonie-suspekten Konsolidierungen (AUC 0.847 [0.808-0.887]) schneidet das AI-System schlechter als radiologische Assistenzärzte, vergleichbar mit dem statistischen Konsensus dreier nicht-radiologischer Assistenzärzte sowie besser als jeder individuelle nicht-radiologische Assistenzarzt ab.

Damit konnte in dieser Originalarbeit ein AI-System für die Detektion von Basis-Pathologien in Röntgen-Thoraces entwickelt werden, das in einem klinisch repräsentativen Validierungsszenario einen Mehrwert als Entscheidungsunterstützung mindestens für nicht-radiologisch spezialisierte Ärzte aufweist, die beispielsweise in kleineren Krankenhäusern ohne 24h-Betreuung durch eine radiologische Fachabteilung auch selbstständig Röntgen-Thoraces interpretieren müssen.

3. Künstliche Intelligenz in der thorakalen Computertomographie

3.1 KI-assistierte orthograde Vermessung der thorakalen Aorta

---

Artificial intelligence assistance improves reporting efficiency of thoracic aortic aneurysm CT follow-up

Rueckel J, Reidler P, Fink N, Sperl J, Geyer T, Fabritius MP, Ricke J, Ingrisch M, Sabel BO.

Eur J Radiol. 2021 Jan; 134:109424

DOI: 10.1016/j.ejrad.2020.109424

Neben der qualitativen Pathologie-Detektion in radiologischen Bilddatensätzen, verspricht eine AI-basierte Softwareassistenz ebenfalls einen relevanten Mehrwert bei der quantitativen Vermessung Pathologie-assoziierter anatomischer Strukturen. Deren visuell-manuelle Vermessung ist einerseits mit einem teils beträchtlichen Zeitaufwand sowie andererseits mit teils relevanter Inter-*Reader*-Variabilität vergesellschaftet. Ein hierbei relevantes klinisches Beispiel ist die Leitlinien-gerechte orthograde Vermessung der thorakalen Aorta, z. B. zur Verlaufskontrolle von Aneurysmen: Einerseits ist die messtechnische Verlaufskontrolle essenziell, um das Risiko einer interventionellen / operativen Therapie mit dem einer Ruptur abzuwägen; letztere ist mit einer Mortalität von 97-100% vergesellschaftet und das Risiko einer Ruptur steht im Zusammenhang mit der zeitlichen Größenzunahme (25,26). Andererseits ist die notwendige Leitlinien-gerechte, orthograde, vollständige und notwendigerweise exakte Vermessung der thorakalen Aorta an neun verschiedenen, anatomisch definierten Messpositionen zeitaufwändig und unterliegt – auch mitbedingt durch notwendige multiplanare Rekonstruktionen – einer erheblichen Fehleranfälligkeit (27–32).

In o. g. Originalarbeit wird ein Algorithmus [Prototyp, *Siemens Healthineers*] klinisch validiert, der automatisiert orthograde multiplanare Rekonstruktionen (MPR) an entsprechend den Leitlinien der *American Heart Association* (33) anatomisch definierten Positionen erstellt und hier den maximalen kontrastierten Innendurchmesser der Aorta vermisst. Eingeschlossen wurden 18 Patienten mit einer aneurysmatischen Erweiterung mindestens der Aorta ascendens, die in zwei EKG-getriggerten CT-Angiographien (*Baseline* & *Follow-Up*)

verlaufskontrolliert wurde. Im Rahmen des retrospektiven Studien-*Readings* wurden in jeder CT die Aortendurchmesser an neun anatomischen Positionen durch drei unabhängige radiologische Ärzte vermessen, dabei in einem ersten *Reading* ohne sowie mit zeitlichem Versatz (*washout-period*) in einem zweiten *Reading* mit Bereitstellung graphisch aufbereiteter algorithmischer Ergebnisse. Als Studienendpunkte wurden definiert: Die Abweichung der algorithmischen von den radiologisch-manuellen Vermessungen inkl. dem Vergleich mit den ursprünglichen radiologischen Befunden, der Einfluss einer Algorithmus-Assistenz auf die letztlich dokumentierten Aortendurchmesser (absolut sowie die Veränderung über die Zeit i. S. e. Aneurysma-Verlaufskontrolle), der Einfluss einer Algorithmus-Assistenz auf die Inter-*Reader*-Variabilität sowie der Zeitaufwand für eine Leitlinien-gerechte Vermessung mit / ohne algorithmische Unterstützung.

Im Abgleich mit den initialen radiologischen Befunden offenbarte die zeitaufwändige Leitlinien-gerechte Aortenvermessung eine zusätzliche dilatative Affektion der Aortenwurzel und/oder des Aortenbogens für 80% jener Aneurysmen, die entsprechend des initialen radiologischen Befundes ausschließlich der Aorta ascendens zugeschrieben wurden. Der notwendige Zeitaufwand für die umfangreiche Leitlinien-gerechte Verlaufsbeurteilung von Aneurysmen unter Einbeziehung von zwei CT-Untersuchungen und aller Messpositionen konnte durch eine algorithmische Assistenz von 13:01 Minuten auf 04:46 Minuten (gemittelt über alle radiologischen *Reader* und eingeschlossenen Patienten) gesenkt werden. Außerdem konnte durch eine AI-Assistenz ebenfalls die Inter-*Reader*-Variabilität von durchschnittlich 1,16mm auf 0,42mm gesenkt werden. Eine signifikante Reduktion der Inter-*Reader*-Variabilität wurde dabei insbesondere für die anatomisch-distalen Messpositionen erreicht mit hier einem deutlich geringeren Anteil algorithmisch vorgeschlagener Messungen, die einer manuelle Korrektur bedurften. Gemittelt über alle radiologischen *Reader* wurden insgesamt 33,6% der algorithmischen Vermessungen korrigiert – entweder durch Korrektur in den algorithmisch generierten MPRs oder durch deren manuelle Neuausrichtung. Überproportional häufig korrigiert wurden dabei Vermessungen der unmittelbar suprakardialen Aorta, mutmaßlich bedingt durch hier auftretende Pulsationsartefakte und entsprechend algorithmische Ungenauigkeiten. Letztlich wich die konventionelle von der algorithmisch-assistierten Verlaufsbeurteilung (gemessen an der Durchmesseränderung von *Baseline*- zu *Follow-Up*-Untersuchung) gemittelt über alle Messpositionen um 0,75mm ab –

diese Abweichung überstieg jedoch nicht signifikant die zugehörige Inter-*Reader*-Variabilität. Zusammenfassend konnte damit das hohe Potenzial einer algorithmisch assistieren Aortenvermessung sowohl hinsichtlich einer signifikanten Zeitersparnis als auch hinsichtlich einer reduzierten Inter-*Reader*-Variabilität demonstriert werden. Anhaltspunkte für eine algorithmische Weiterentwicklung ergeben sich aus den Studienergebnisse insbesondere für eine Verlaufskontrolle basierend auf einer automatisierten sektoriellen Volumetrie anstatt basierend auf Einzeldurchmessern – eine so quantifizierbare Verlaufskontrolle von Aneurysmen ist mutmaßlich weniger abhängig von Pulsationsartefakten oder einer durch Plaquebildungen möglicherweise im gehobenen Patientenalter asymmetrischen Windkesselfunktion der v. a. suprakardialen Aorta.

3.2 Reduktion von *Missed Findings* in der CT-Polytrauma-Spirale durch KI-Assistenz

---

Reduction of missed thoracic findings in emergency whole-body computed tomography using artificial intelligence assistance

Rueckel J, Sperl J, Kaestle S, Hoppe BF, Fink N, Rudolph J, Schwarze V, Geyer T, Strobl FF, Ricke, J, Ingrisch M, Sabel BO.

Quant Imaging Med Surg. 2021; No 6 (June 2021), Vol 11

DOI: 10.21037/qims-20-1037

Schwere Traumata bleiben eine der häufigsten Todesursachen bei Patienten eines Alters jünger als 45 Jahre (34,35). Die Zeit bis zur diagnostischen Computertomographie ist ein etablierter klinischer Qualitätsindikator für die Versorgung polytraumatisierter Patienten – eine Zeitspanne von 30-60 Minuten inklusive der radiologischen Bildinterpretation wird empfohlen (36). Die Kombination aus Patienten mit mutmaßlich erstmaliger umfangreicher (Ganzkörper-)Bildgebung und der zeitkritischen radiologischen Befundung erhöht das Risiko, dass bisher unbekannte Nebenbefunde übersehen werden. Relevante Diskrepanzen zwischen initialen radiologischen Berichten und im Verlauf zusätzlich erstellter Begutachtungen durch erfahrene Experten wurde durch Studien quantifiziert (37,38). Eine AI-basierte Bildanalyse parallel zur initialen radiologischen Bildinterpretation in der Notfallsituation verspricht eine zeiteffiziente Reduktion übersehener (Neben-)Befunde. Die inhaltliche Breite möglicher CT-Befunde bleibt jedoch eine Herausforderung vor dem Hintergrund von *Deep-Learning*-Algorithmen, die üblicherweise entsprechend der verwendeten Trainingsdaten für Detektion sehr eng definierter Bildmerkmale ausgelegt sind.

In o. g. Originalarbeit wird der klinische Mehrwert einer Softwareplattform (*AI-Rad Companion Chest CT, Siemens Healthineers*) validiert, in die verschiedene Einzelalgorithmen eingebunden sind. Dabei wurden in eingangs erwähnten Polytrauma-Bildgebungen Algorithmen für die Detektion folgender thorakaler (Neben-)befunde validiert: Automatische Aortenvermessung (siehe auch Abschnitt 3.1), Detektion von Sinterungsfrakturen der thorakalen Wirbelkörper, Detektion von Lungenrundherden, die automatische Volumetrie der

Herzgröße (u. a. relativ zum Lungenvolumen) sowie die automatische Detektion kalzifizierter Koronarplaques. Retrospektiv eingeschlossen wurden 105 konsekutiv in der Notaufnahme vorstellige Patienten, die eine Ganzkörperbildgebung („Schockraumspirale“, analysiert wurden die thorakalen Abschnitte der in arterieller KM-Phase akquirierten Bildgebung) im Zeitraum von Januar 2019 bis November 2019 erhalten haben. Die algorithmisch detektierten Befunde wurden graphisch aufgearbeitet, radiologisch auf Plausibilität geprüft und mit den initialen radiologischen Berichten verglichen.

Unter den 105 algorithmisch assistiert ausgewerteten CT-Bildgebungen zeigte sich eine relevante Anzahl von Nebenbefunden, die in den initialen radiologischen Berichten nicht erwähnt wurden: Bis zu 25 Patienten (23,8%) mit einer Kardiomegalie oder als grenzwertig groß klassifizierter Herzgröße, 17 Patienten (16,2%) mit kalzifizierten Koronarplaques, 34 Patienten (32,4%) mit einem den Populations-Mittelwert um mindestens zwei Standardabweichungen übersteigendem Durchmesser der thorakalen Aorta, zwei Patienten (1,9%) mit Lungenrundherden >6mm sowie zwölf Patienten (11,4%) mit Sinterungsfrakturen der Brustwirbelsäule (darunter zwei Frakturen bildmorphologisch mutmaßlich akuter Genese). Dies unterstreicht in dem untersuchten klinischen *Setting* das hohe Potenzial, die relevante Anzahl initial nicht detektierter (oder zumindest im radiologischen Bericht nicht erwähnter) (Neben-)befunde durch eine Software-Assistenz relevant zu reduzieren und greift dabei auf einen Ansatz zurück, verschieden Einzelalgorithmen Plattform-basiert im Sinne der Erweiterung des klinischen Fokus zu bündeln.

Nichts desto trotz zeigt auch diese Studie die dringende Notwendigkeit, die Mehrzahl der analysierten Einzelalgorithmen gerade in Bezug auf die erreichte Spezifität sowie hinsichtlich der Robustheit gegenüber im Notfall auch suboptimaler Bildqualität zu steigern: Die Quantifzierung der Aortendurchmesser sowie die qualitative Detektion von Koronarplaques wurde deutlich limitiert durch Pulsationsartefakte mangels EKG-getriggerter Bildakquisition, die Detektion von Koronarplaques zusätzlich durch intravasales Kontrastmittel und entsprechend demgegenüber schwieriger zu detektierender / volumetrierender kalzifizierter Plaques. Die Detektion von pulmonalen Rundherden und mutmaßlicher Wirbelkörpersinterungen offenbarte eine hohe Anzahl falsch-positiver Befunde, die im Anschluss zeitaufwändig wieder verworfen werden müssten: Von 64 AI-detektierten

Sinterungsfrakturen wurden 24 mangels Plausibilität wieder als „falsch positiv“ verworfen, während 13 Sinterungsfrakturen ausschließlich vom Algorithmus detektiert und radiologisch als plausible Detektion validiert wurden. Hinsichtlich 81 ausschließlich vom Algorithmus detektierten Lungenläsionen wurden 17 (21%) der Detektionen als nicht plausibel verworfen, 49 (73%) der Detektionen als nicht malignomsuspekt klassifiziert (Trauma-assoziiert, perifissurale Lymphknoten, Granulome, narbig / postentzündlich) und nur drei der Läsionen (3,7%) als ggf. kontrollbedürftig eingeordnet. Letztlich stellt sich damit trotz dem demonstrierten Potenzial, die Anzahl verpasster Befunde mittels KI-Assistenz zu reduzieren, die Frage nach der klinischen Effektivität und einem möglichen Einfluss auf das klinische *decision making*: Ein hoher Anteil auch teils falsch-positiver Detektionen durch eingeschränkt spezifische Algorithmen bedarf zum einen ein zeitaufwändiges Verwerfen nicht plausibler algorithmisch detektierter Befunde. Zum anderen stellt sich die Frage, ob eine ggf. vorhandene Hemmschwelle, algorithmisch detektierte Befunde zu negieren, zu einem relevanten *confirmation bias* und letztlich zu ggf. einer disproportional erhöhten Anzahl indizierter Verlaufsuntersuchungen oder anderweitiger Abklärungen führt. Diesbezüglich sind weitere Studien notwendig.

4. Verzeichnis der Originalarbeiten

1. **Rueckel J**, Trappmann L, Schachtner B, Wesp P, Hoppe BF, Fink N, Ricke J, Dinkel J, Ingrisch M, Sabel BO. Impact of Confounding Thoracic Tubes and Pleural Dehiscence Extent on Artificial Intelligence Pneumothorax Detection in Chest Radiographs. Invest Radiol. 2020 Dec; 55(12):792–8. DOI: 10.1097/RLI.0000000000000707

2. **Rueckel J**, Huemmer C, Fieselmann A, Ghesu FC, Mansoor A, Schachtner B, Wesp P, Trappmann L, Munawwar B, Ricke J, Ingrisch M, Sabel BO. Pneumothorax detection in chest radiographs: optimizing artificial intelligence system for accuracy and confounding bias reduction using in-image annotations in algorithm training. Eur Radiol. 2021 Mar 27. DOI: 10.1007/s00330-021-07833-w

3. **Rueckel J**, Kunz WG, Hoppe BF, Patzig M, Notohamiprodjo M, Meinel FG, Cyran CC, Ingrisch M, Ricke J, Sabel BO. Artificial Intelligence Algorithm Detecting Lung Infection in Supine Chest Radiographs of Critically Ill Patients With a Diagnostic Accuracy Similar to Board-Certified Radiologists. Crit Care Med. 2020 May 20. DOI: 10.1097/CCM.0000000000004397

4. **Rueckel J**, Reidler P, Fink N, Sperl J, Geyer T, Fabritius MP, Ricke J, Ingrisch M, Sabel BO. Artificial intelligence assistance improves reporting efficiency of thoracic aortic aneurysm CT follow-up. Eur J Radiol. 2021 Jan;134:109424. DOI: 10.1016/j.ejrad.2020.109424

5. **Rueckel J**, Sperl J, Kaestle S, Hoppe BF, Fink N, Rudolph J, Schwarze V, Geyer T, Strobl FF, Ricke, J, Ingrisch M, Sabel BO. Reduction of missed thoracic findings in emergency whole-body computed tomography using artificial intelligence assistance. Quant Imaging Med Surg. 2021. DOI: 10.21037/qims-20-1037

6. Rudolph J, Huemmer C, Ghesu FC, Mansoor A, Preuhs A, Fieselmann A, Fink N, Dinkel J, Koliogiannis V, Schwarze V, Goller S, Fischer M, Jörgens M, Khaled NB, Vishwanath RS, Balachandran A, Ingrisch M, Ricke J, Sabel BO, **Rueckel J**. Artificial Intelligence in Chest Radiography Reporting Accuracy - Clinical Value in the Emergency Unit Setting without 24/7 Radiology Coverage. Invest Radiol. In Press.

5. Literatur

1. Oakden-Rayner L. The Rebirth of CAD: How Is Modern AI Different from the CAD We Know? Radiology: Artificial Intelligence. 2019 May;1(3):e180089.
2. Rao VM, Levin DC, Parker L, Cavanaugh B, Frangos AJ, Sunshine JH. How widely is computer-aided detection used in screening and diagnostic mammography? J Am Coll Radiol. 2010 Oct;7(10):802–5.
3. Li Q. Recent progress in computer-aided diagnosis of lung nodules on thin-section CT. Comput Med Imaging Graph. 2007 Jul;31(4–5):248–57.
4. Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson ANA, Miglioretti DL, et al. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med. 2015 Nov;175(11):1828–37.
5. Fenton JJ, Taplin SH, Carney PA, Abraham L, Sickles EA, D'Orsi C, et al. Influence of Computer-Aided Detection on Performance of Screening Mammography. New England Journal of Medicine. 2007 Apr 5;356(14):1399–409.
6. Gilbert FJ, Astley SM, Gillan MGC, Agbaje OF, Wallis MG, James J, et al. Single reading with computer-aided detection for screening mammography. N Engl J Med. 2008 Oct 16;359(16):1675–84.
7. Tchou PM, Haygood TM, Atkinson EN, Stephens TW, Davis PL, Arribas EM, et al. Interpretation time of computer-aided detection at screening mammography. Radiology. 2010 Oct;257(1):40–6.
8. Clinical radiology UK workforce census 2016 report. Clinical radiology. 2016;54.
9. Sokolovskaya E, Shinde T, Ruchman RB, Kwak AJ, Lu S, Shariff YK, et al. The Effect of Faster Reporting Speed for Imaging Studies on the Number of Misses and Interpretation Errors: A Pilot Study. J Am Coll Radiol. 2015 Jul;12(7):683–8.
10. Brady AP. Error and discrepancy in radiology: inevitable or avoidable? Insights Imaging. 2016 Dec 7;8(1):171–82.
11. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016 13;316(22):2402–10.
12. Milea D, Najjar RP, Jiang Z, Ting D, Vasseneix C, Xu X, et al. Artificial Intelligence to Detect Papilledema from Ocular Fundus Photographs. New England Journal of Medicine. 2020 Apr 30;382(18):1687–95.
13. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 02;542(7639):115–8.
14. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet. 2018 Dec 1;392(10162):2388–96.
15. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020 Jan;577(7788):89–94.
16. Lotter W, Diab AR, Haslam B, Kim JG, Grisot G, Wu E, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. Nature Medicine. 2021 Feb;27(2):244–9.
17. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of

Common Thorax Diseases. arXiv:170502315 [cs] [Internet]. 2017 May 5 [cited 2018 Oct 29]; Available from: http://arxiv.org/abs/1705.02315
18. Gohagan JK, Prorok PC, Hayes RB, Kramer BS, Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial Project Team. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. Control Clin Trials. 2000 Dec;21(6 Suppl):251S-272S.
19. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv:171105225 [cs, stat] [Internet]. 2017 Nov 14 [cited 2018 Oct 29]; Available from: http://arxiv.org/abs/1711.05225
20. Weng X. CheXNet for Classification and Localization of Thoracic Diseases [Internet]. 2020 [cited 2020 May 1]. Available from: https://github.com/arnoweng/CheXNet
21. Kunz WG, Patzig M, Crispin A, Stahl R, Reiser MF, Notohamiprodjo M. The Value of Supine Chest X-Ray in the Diagnosis of Pneumonia in the Basal Lung Zones. Acad Radiol. 2018;25(10):1252–6.
22. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med. 2018;15(11):e1002686.
23. Yao L, Poblenz E, Dagunts D, Covington B, Bernard D, Lyman K. Learning to diagnose from scratch by exploiting dependencies among labels. arXiv:171010501 [cs] [Internet]. 2017 Oct 28 [cited 2019 Sep 2]; Available from: http://arxiv.org/abs/1710.10501
24. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32–5.
25. Johansson G, Markström U, Swedenborg J. Ruptured thoracic aortic aneurysms: a study of incidence and mortality rates. J Vasc Surg. 1995 Jun;21(6):985–8.
26. Yiu RS, Cheng SWK. Natural history and risk factors for rupture of thoracic aortic arch aneurysms. J Vasc Surg. 2016;63(5):1189–94.
27. Erbel R, Aboyans V, Boileau C, Bossone E, Bartolomeo RD, Eggebrecht H, et al. 2014 ESC Guidelines on the diagnosis and treatment of aortic diseases: Document covering acute and chronic aortic diseases of the thoracic and abdominal aorta of the adult. The Task Force for the Diagnosis and Treatment of Aortic Diseases of the European Society of Cardiology (ESC). Eur Heart J. 2014 Nov 1;35(41):2873–926.
28. Mora CE, Marcus CD, Barbe CM, Ecarnot FB, Long AL. Maximum Diameter of Native Abdominal Aortic Aneurysm Measured by Angio-Computed Tomography. Aorta (Stamford). 2015 Apr 1;3(2):47–55.
29. Banno H, Kobeiter H, Brossier J, Marzelle J, Presles E, Becquemin J-P. Inter-observer Variability in Sizing Fenestrated and/or Branched Aortic Stent-grafts. European Journal of Vascular and Endovascular Surgery. 2014 Jan 1;47(1):45–52.
30. Oshin OA, England A, McWilliams RG, Brennan JA, Fisher RK, Vallabhaneni SR. Intra- and interobserver variability of target vessel measurement for fenestrated endovascular aneurysm repair. J Endovasc Ther. 2010 Jun;17(3):402–7.
31. Lu T-LC, Rizzo E, Marques-Vidal PM, Segesser LK von, Dehmeshki J, Qanadli SD. Variability of ascending aorta diameter measurements as assessed with electrocardiography-gated multidetector computerized tomography and computer assisted diagnosis software. Interact Cardiovasc Thorac Surg. 2010 Feb;10(2):217–21.
32. Quint LE, Liu PS, Booher AM, Watcharotone K, Myles JD. Proximal Thoracic Aortic Diameter Measurements at CT: Repeatability and Reproducibility According to Measurement Method. Int J Cardiovasc Imaging. 2013 Feb;29(2):479–88.
33. Hiratzka LF, Bakris GL, Beckman JA, Bersin RM, Carr VF, Casey DE, et al. 2010

ACCF/AHA/AATS/ACR/ASA/SCA/SCAI/SIR/STS/SVM guidelines for the diagnosis and management of patients with Thoracic Aortic Disease: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines, American Association for Thoracic Surgery, American College of Radiology, American Stroke Association, Society of Cardiovascular Anesthesiologists, Society for Cardiovascular Angiography and Interventions, Society of Interventional Radiology, Society of Thoracic Surgeons, and Society for Vascular Medicine. Circulation. 2010 Apr 6;121(13):e266-369.

34. Trauma Facts - The American Association for the Surgery of Trauma [Internet]. 2020 [cited 2020 Mar 23]. Available from: http://www.aast.org/trauma-facts
35. Motta-Ramírez GA. [The radiologist physician in major trauma evaluation]. Gac Med Mex. 2016 Aug;152(4):534–46.
36. McKechnie PS, Kerslake DA, Parks RW. Time to CT and Surgery for HPB Trauma in Scotland Prior to the Introduction of Major Trauma Centres. World J Surg. 2017;41(7):1796–800.
37. Guven R, Akca AH, Caltili C, Sasmaz MI, Kaykisiz EK, Baran S, et al. Comparing the interpretation of emergency department computed tomography between emergency physicians and attending radiologists: A multicenter study. Niger J Clin Pract. 2018 Oct;21(10):1323–9.
38. Howlett DC, Drinkwater K, Frost C, Higginson A, Ball C, Maskell G. The accuracy of interpretation of emergency abdominal CT in adult patients who present with non-traumatic abdominal pain: results of a UK national audit. Clin Radiol. 2017 Jan;72(1):41–51.

.