Aus dem Institut für Diagnostische und Interventionelle Radiologie
der Universität zu Köln
Direktor: Universitätsprofessor Dr. med. David Maintz

GPT-4 Analyse von MRT-Berichten bei Verdacht auf Myokarditis: Eine multizentrische Studie

Inaugural-Dissertation zur Erlangung der Doktorwürde
der Medizinischen Fakultät
der Universität zu Köln

vorgelegt von
Kenan Kaya
aus Köln

promoviert am 23.12.2024

Gedruckt mit Genehmigung der Medizinischen Fakultät der Universität zu Köln

2025

Dekanin/Dekan: Universitätsprofessor Dr. med. G. R. Fink

1. Gutachter: Privatdozent Dr. med. L. U. Pennig
2. Gutachter: Privatdozent Dr. med. univ. T. Hickethier

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Dissertationsschrift ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Bei der Auswahl des Themas und Auswertung des Materials sowie bei der Herstellung des Manuskriptes habe ich Unterstützungsleistungen von folgenden Personen erhalten:

Herrn PD Dr. med. Lenhard Urs Pennig und Herrn Dr. med. Jonathan Kottlors

Weitere Personen waren an der Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe einer Promotionsberaterin/eines Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertationsschrift stehen.

Die Dissertationsschrift wurde von mir bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Für die vorliegende kumulative Dissertation habe ich das Paper selbstständig verfasst. Außerdem habe ich selbstständig die Daten in das Large Language Model mittels eines Prompts eingegeben und die statistischen Ergebnisse in Kooperation mit Dr. Astha Jaiswal mit RStudio ausgewertet. Die zugrundeliegenden MRT-Berichte wurden in Zusammenarbeit mit Herrn PD Dr. med. Lenhard Urs Pennig und Dr. med. Jonathan Kottlors selbstständig sortiert.

Erklärung zur guten wissenschaftlichen Praxis:

Ich erkläre hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten (Amtliche Mitteilung der Universität zu Köln AM 132/2020) der Universität zu Köln gelesen habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen.

Köln, 30.09.2024

Unterschrift: .....

Danksagung

Ich möchte mich bei allen Personen bedanken, die mich bei dieser Promotion begleitet haben. Mein besonderer Dank gilt PD Dr. med. Lenhard Pennig und Dr. med. Jonathan Kottlors für die Überlassung des Themas und enge Betreuung während der Durchführung des Projektes. Außerdem möchte ich mich bei meiner Familie und meiner Freundin Matina Tzianopoulou für Ihre Unterstützung bedanken. Insbesondere Dir Matina bin ich sehr dankbar, dass Du jederzeit für mich da bist, nie an mir zweifelst und immer hinter mir stehst. Ich liebe Dich von tiefstem Herzen.

Widmung

Mit aufrichtiger Liebe und tiefstem Respekt widme ich diese Dissertation meiner Mutter Frau Dr. med. Ergül Kaya. Mama, Du bist die größte Unterstützung und Inspiration in meinem Leben. Auch wenn diese Arbeit nur ein kleiner Anteil im Vergleich zu deiner bedingungslosen Hingabe ist, möge sie dennoch meine unermessliche Dankbarkeit für Dich zum Ausdruck bringen. Alles, was ich bin und alles, was ich werde, habe ich allein Dir zu verdanken.

In Liebe,
Kenan

INHALTSVERZEICHNIS

Abkürzungsverzeichnis

2D – Zweidimensional
ACS – Akutes Koronarsyndrom
BSG – Blutsenkungsgeschwindigkeit
CK – Creatinkinase
CK-MB – Creatinkinase-MB
CMR – Kardiale Magnetresonanztomographie
CRP – C-reaktives Protein
CT – Computertomographie
DCM – Dilatative Kardiomyopathie
EZV – Extrazelluläres Volumen
EGE – Early Gadolinium Enhancement
EGEr – Early Gadolinium Enhancement Ratio
ESC – European Society of Cardiology
Gd – Gadolinium
GPT-4 – Generative Pre-trained Transformer 4
Hs-TnI – Hochsensitives Troponin I
KI – Künstliche Intelligenz
KM – Kontrastmittel
LGE – Late Gadolinium Enhancement
LLC – Lake Louise Criteria (dt: Lake-Louise-Kriterien)
LLM – Large Language Model
LVEF – Linksventrikuläre Ejektionsfraktion
NLP – Natural Language Processing
MIP – Maximum-Intensity-Projektionen
MOLLI – Modified Look-Locker Imaging
MPR – Multiplanare Rekonstruktionen
MRT – Magnetresonanztomographie
NSTEMI – Nicht-ST-Hebungsinfarkt
PACS – Picture Archiving and Communication System
PLMs – Pre-trained language models
PSIR – Phase-Sensitive Inversion-Recovery
RIS – Radiologieinformationssystem

RLHF – Reinforcement Learning With Human Feedback

ROI – Region of interest

SAx – Kurze Achse (engl: Short axis)

ShMOLLI – Shortened MOLLI

SI – Signalintensität

SNR – Signal-Rausch-Verhältnis (engl: Signal-Noise-Ratio)

TE – Echozeit

TI – Inversionszeit

TR – Repetitionszeit

1. Zusammenfassung

Das Ziel dieser Arbeit war es, den Nutzen eines Large Language Models (LLMs) namens Generative Pre-trained Transformer 4 (GPT-4) hinsichtlich der text-basierten Diagnose einer Myokarditis basierend auf dem radiologischen Befundbericht einer Herz-MRT, der Patientenanamnese und Laborwerten zu untersuchen. Hierzu wurden Befundberichte von 396 Patienten von acht deutschen Universitätsklinika verwendet, welche eine Herz-MRT bei Verdacht auf Myokarditis erhielten. Anhand der Anweisung: „Bitte entscheiden Sie auf der Grundlage des radiologischen Berichts, der bereitgestellten Patienteninformationen und der klinischen Parameter, ob eine Myokarditis vorliegt oder nicht. Bitte antworten Sie entweder mit ja oder nein" sollten sowohl GPT-4 als auch drei Radiologen mit jeweils einem (R1), zwei (R2) und vier Jahren (R3) Erfahrung in der kardiovaskulären Bildgebung anhand der gegebenen Daten entscheiden, ob eine Myokarditis vorliegt oder nicht. Die abschließende Beurteilung des Befundes, ob eine Myokarditis vorliegt oder nicht, wurde den Radiologen und GPT-4 nicht vorgelegt. Deren Beurteilung wurde gegenüber der Consensus-Bewertung der vorliegen Daten zweier Fachärzte mit acht und zehn Jahren Erfahrung in der kardiovaskulären Bildgebung, welche als Referenzstandard dienten, verglichen. Sensitivität, Spezifität und Genauigkeit wurden berechnet.

GPT-4 erzielte eine Genauigkeit von 83%, eine Sensitivität von 90% und eine Spezifität von 78%, was mit den Ergebnissen des Radiologen mit einem Jahr Erfahrung (R1: 86%, 90%, 84%, p=.14) vergleichbar war, jedoch geringer als bei erfahreneren Ärzten ausfiel (R2: 89%, 86%, 91%, p=.007 und R3: 91%, 85%, 96%, p<.001). Sowohl GPT-4 als auch die Radiologen zeigten eine verbesserte diagnostische Leistung, wenn Berichte die Ergebnisse des T1- und T2-Mappings enthielten, wobei dies bei den Radiologen mit einer und vier Jahren Erfahrung statistisch signifikant war (p=.004 bzw. p=.02).

Zusammenfassend lässt sich sagen, dass GPT-4 eine hohe Genauigkeit bei der Diagnose von Myokarditis anhand von MRT-Berichten in einem umfangreichen, multizentrischen Datensatz erreichte und somit als unterstützendes diagnostisches Werkzeug, besonders für weniger erfahrene Ärzte, eingesetzt werden könnte. Weitere Untersuchungen sind erforderlich, um das gesamte Potenzial zu erfassen und die übrigen Aspekte der Einbindung großer Sprachmodelle in die medizinische Entscheidungsfindung genauer zu untersuchen.

2. Einleitung

Die häufigste Todesursache weltweit sind Erkrankungen des Herzkreislaufsystems[1]. Als bedeutende Ursache für kardiale Morbidität und Mortalität macht die Myokarditis bei Patienten unter 40 Jahren 20-40 % der Fälle des plötzlichen Herztods aus[2,3]. Durchschnittlich erkranken weltweit jährlich ca. 23/100.000 der Gesamtpopulation an einer Myokarditis[4], wobei vor allem junge, männliche Patienten betroffen sind[5]. Im Rahmen einer viralen Infektion wird bei 1-5 % der Patienten eine myokardiale Beteiligung angenommen[6]. Obwohl eine frühzeitige und genaue Diagnose der Myokarditis unerlässlich ist, um das Risiko eines Fortschreitens der Erkrankung zu minimieren, bleibt die korrekte Diagnose aufgrund der vielfältigen klinischen Erscheinungsbilder und Laborbefunde der Myokarditis weiterhin eine Herausforderung[7].

In diesem Kontext hat sich die Magnetresonanztomographie (MRT) unter Verwendung verschiedener Sequenztechniken für Morphologie und Gewebecharakterisierung als Goldstandard in der nicht-invasiven Diagnostik bei Patienten mit Verdacht auf eine Myokarditis durchgesetzt[8]. Die Diagnose einer Myokarditis mittels kardialer Magnetresonanztomographie (CMR) setzt ein hohes Maß an radiologischem Fachwissen voraus sowie die Fähigkeit, verschiedene Bildmerkmale in unterschiedlichen Sequenzen zu interpretieren[9]. Experten für kardiovaskuläre Bildgebung verfügen über die notwendige Expertise und Erfahrung, um die Diagnose einer Myokarditis in der CMR verlässlich zu stellen. Im Gegensatz dazu zeigen unerfahrene Radiologen eine deutlich geringere Genauigkeit bei der Interpretation dieser anspruchsvollen Befunde. Dies führt zu einer höheren Wahrscheinlichkeit von Fehldiagnosen, da sie die subtilen Anzeichen der Myokarditis übersehen oder falsch deuten[10].

Zahlreiche Studien haben die Möglichkeit und das Potenzial des Einsatzes von künstlicher Intelligenz (KI) in der medizinischen Entscheidungsfindung gezeigt, exemplarisch in der bild-basierten Detektion von Melanomen[11]. Auch im Bereich der Radiologie ist der Einsatz von KI von großem Interesse in Forschung und klinischer Routine[12], primär um visuelle Daten zu analysieren und zu interpretieren[13]. So konnte in einer kürzlich publizierten Studie von Saha et al. ein künstliches neuronales Netzwerk Prostatakarzinome in der MRT besser als Radiologen detektieren[13]. In den letzten Jahren weitet sich der Nutzen der KI in der Radiologie vermehrt auch auf die Analyse von Textdaten, z.B. CT- oder MRT-Anforderungen[14] oder radiologische Befundberichte[15], welche das Fundament für die Dokumentation und Kommunikation in der Radiologie darstellen, aus. Hier eröffnen sogenannte Large Language Models (LLMs) neue Möglichkeiten für die Verarbeitung von textbasierten medizinischen Informationen[14]. In diesem Kontext stellt das „Generative Pre-trained Transformer fourth-generation model“ (GPT-4) ein besonders leistungsfähiges Modell dar. GPT-4 ist ein künstliches neuronales Netzwerk, das in der Lage ist, logische und semantisch genaue Antworten auf textbasierte Eingaben und Fragen zu generieren[16].

GPT-4 wurde mithilfe einer großen Sammlung von Textdaten aus dem Internet trainiert und für eine Vielzahl sprachbezogener Aufgaben wie Textvervollständigung, Übersetzung und die Beantwortung von Fragen optimiert. Der Nutzen von GPT-4 in der Radiologie konnte schon für die Optimierung von Arbeitsprozessen[14] und zur klinischen Entscheidungsfindung auf Grundlage von radiologischen Befundberichten[15] in diversen Publikationen gezeigt werden. Dank seiner Fähigkeit, textbasierte Informationen zu analysieren und zu integrieren, könnte GPT-4 auch einen Nutzen hinsichtlich der textbasierten Detektion von Myokarditis basierend auf der Interpretation verschiedener Bildmerkmale in unterschiedlichen Sequenzen, klinischer Daten und Laborparameter zeigen. Das Ziel dieser Arbeit war es, den Nutzen von GPT-4 hinsichtlich der text-basierten Diagnose einer Myokarditis basierend auf dem radiologischen Befundbericht einer CMR, der Patientenanamnese und Laborwerten zu untersuchen.

Die Struktur dieser Arbeit ist wie folgt: Zuerst werden die Grundlagen der MRT, insbesondere der CMR, erläutert. Danach wird das Themengebiet der Myokarditis hinsichtlich ihrer klinischen Präsentation und Diagnostik näher dargelegt. Darauffolgend wird es eine Einführung in das Gebiert der künstlichen Intelligenz geben und die verschiedene Einsatzmöglichkeiten dieser in der Radiologie mit besonderem Fokus auf LLMs beleuchtet. Zum Abschluss werden nach der Publikation die Ergebnisse und Limitationen der aktuellen Arbeit diskutiert sowie ein Ausblick auf in Zukunft kommende Weiterentwicklungen und Anwendungsmöglichkeiten der künstlichen Intelligenz in der Radiologie gegeben.

2.1. Grundlagen der MRT

Die Magnetresonanztomografie (MRT) beruht auf der Erfassung von Interaktionen zwischen rotierenden Atomkernen und einem externen Magnetfeld. Diese rotierenden Atomkerne erzeugen durch ihren eigenen Drehimpuls, der auch als „Kernspin“ bekannt ist, kleine Magnetfelder. Das am häufigsten vorkommende chemische Element im menschlichen Körper ist hierbei der Wasserstoff. Aufgrund seiner einzigartigen Struktur bestehend aus einem einzelnen Proton (H+), ist das Wasserstoffatom besonders geeignet für die MRT, da es einen sehr effizienten und geeigneten Kernspin bietet. Unter normalen Bedingungen sind die Wasserstoffkerne zufällig und ungeordnet ausgerichtet. Wird jedoch ein starkes äußeres Magnetfeld angelegt, wie es bei der MRT der Fall ist, richten sich die Kernspins entlang der Magnetfeldlinien aus. Dies geschieht ähnlich wie bei einer Kompassnadel, die sich nach Norden und Süden bzw. von Kopf bis Fuß ausrichtet. Während dieses Prozesses präzedieren die Kernspins mit einer bestimmten Frequenz, die als „Larmorfrequenz“ bezeichnet wird. Diese Frequenz beschreibt die Rotationsgeschwindigkeit der Protonen in einem bestimmten Magnetfeld und variiert je nach chemischer Zusammensetzung des Moleküls. Ein zusätzliches Magnetfeld, der sogenannte „Anregungspuls“, wird im Radiofrequenzbereich angewendet und hat eine Frequenz, die mit den Kernspins resoniert. Dieses Magnetfeld ändert die Ausrichtung der Kernspins um einen Winkel von 90 Grad, auch „Flipwinkel“ genannt, relativ zum bestehenden Magnetfeld. Nachdem dieses zweite Magnetfeld abgeschaltet wird, kehren die Kernspins in ihre ursprüngliche Ausrichtung entlang des Magnetfeldes zurück, ein Vorgang der als „Relaxation“ bezeichnet wird. Während dieser Rückkehr in den Ausgangszustand erzeugen die Wasserstoffprotonen eine transversale magnetische Komponente. Diese Komponente kann von einer Empfangsspule erfasst werden, die am Körper des Patienten platziert ist[17,18].

Die Resonanzfrequenz wird maßgeblich durch die chemische Zusammensetzung der Umgebung beeinflusst. Das empfangene Signal enthält zwar wertvolle Informationen über die chemische Zusammensetzung und damit über die unterschiedlichen Gewebetypen im untersuchten Objekt, jedoch ist dieses Signal ohne zusätzliche Informationen nur als Gesamtsumme aller angeregten Kernspins empfangbar. Dies bedeutet, dass es nicht möglich ist, die genauen Positionen der verschiedenen Gewebetypen zu bestimmen, ohne eine räumliche Zuordnung vorzunehmen. Um dieses Problem zu lösen und eine präzise räumliche Auflösung zu ermöglichen, werden zusätzliche Magnetfeldgradienten in allen drei Raumrichtungen (x, y und z) eingesetzt. Diese Gradientenfelder erzeugen ein dreidimensionales System, in dem das Magnetfeld ortsabhängig variiert. Durch die Einführung dieser Gradienten ändern sich die Frequenzen der Spins entlang der Magnetfeldgradienten in Abhängigkeit von ihrer Position im Raum.

Dadurch können gezielt Spins in bestimmten Bereichen des untersuchten Objektes angeregt werden. Mit Hilfe eines sogenannten Anregungspulses in Kombination mit einem Magnetfeldgradienten können nur die Spins angeregt werden, welche die spezifische Resonanzbedingung erfüllen und beispielsweise zu einer bestimmten Schicht gehören. Diese selektive Anregung ermöglicht es, gezielt Spins in einer bestimmten Schicht des Objekts anzuregen. Das empfangene Signal stellt daher eine Mischung aus den Signalen aller Spins in dieser spezifischen Schicht dar. Um innerhalb dieser ausgewählten Schicht das Signal weiter in einzelne Voxel (dreidimensionale Bildpunkte) zu unterteilen und letztendlich ein detailliertes Bild zu erzeugen, werden zusätzliche, temporäre Magnetfeldgradienten angewendet. Diese Gradienten führen zu einer Phasen- und Frequenzverschiebung der Spins, die es ermöglicht, den räumlichen Ursprung der Spins zu bestimmten (Phasen- und Frequenzkodierung). Bei jeder durchgeführten Messung hat der Spin eine spezifische Frequenz und Phase. Gleichzeitig wird ein Signal aufgezeichnet, das die aufsummierten Signale aller Spins in dieser Schicht darstellt. Durch die Kodierung der Phase und Frequenz kann die genaue Position jedes einzelnen Spins bestimmt werden, was durch die Anwendung der Fourier-Transformation erreicht wird. Die notwendigen Einzelmessungen werden in Abhängigkeit von der gewählten Größe der Akquisitionsmatrix durchgeführt. Häufig werden 256 Messungen pro Schicht durchgeführt, um eine hohe räumliche Auflösung zu gewährleisten. Diese Einzelmessungen werden im sogenannten k-Raum gespeichert, welcher der gewählten Matrixgröße entspricht. Die Datenpunkte in der Mitte des k-Raums beeinflussen hauptsächlich den Bildkontrast, während die Datenpunkte der Peripherie die Konturen des Bildes definieren. Durch die Anwendung der Fourier-Transformation können die im k-Raum gesammelten Informationen in den Bildraum übertragen werden. Dadurch entsteht ein detailliertes anatomisches Bild, das die genaue räumliche Verteilung und Beschaffenheit der verschiedenen Gewebetypen darstellt[17,19].

Sequenztechniken

Die wichtigsten Arten von Bildgebungssequenzen sind die Gradienten-Echo- und Spin-Echo-Sequenzen. Bei der Gradienten-Echo-Sequenz wird das Signal durch einen einzigen Radiofrequenzpuls und einen abwechselnd in positiver und negativer Richtung wirkenden Magnetfeldgradienten erzeugt[20]. Im Gegensatz dazu wird das Signal bei der Spin-Echo-Sequenz durch die Anwendung von zwei aufeinanderfolgenden Radiofrequenzpulsen erzeugt, die zunächst einen 90°- und 180°-Puls umfassen[21]. Hierbei erwähnenswert sind die T1- und T2-Gewichtung.

T1-Gewichtung

Die T1-Gewichtung wird durch das Einstellen bestimmter Parameter erzielt. Dafür ist eine kurze Repetitionszeit (TR) notwendig, welche die Zeit zwischen den aufeinanderfolgenden Anregungen der Spins definiert. Zusätzlich wird eine kurze Echozeit (TE) eingestellt, welche die Zeitspanne zwischen der Anregung des Spins und der Messung des Signals repräsentiert. Diese Kombination führt dazu, dass die longitudinale Relaxationszeit, auch als T1-Zeit bekannt, den Bildkontrast beeinflusst. Die T1-Relaxationszeit variiert je nach Gewebeart. Gewebe mit einer kurzen T1-Relaxationszeit, wie Blut und Fett, können sich schneller erholen und bereits bei der nächsten Anregung ein Signal erzeugen. Diese Gewebe erscheinen daher auf den Bildern hyperintens, also heller. Im Gegensatz dazu benötigen Gewebe mit einer langen T1-Relaxationszeit, wie der Liquor cerebrospinalis, mehr Zeit zur Erholung und erzeugen daher kein oder nur ein schwaches Signal. Diese Gewebe erscheinen hypointens, also dunkler. Der Grad der T1-Gewichtung kann weiter gesteigert werden, indem die TR noch kürzer gewählt und der Winkel des Anregungspulses erhöht wird. Ein steilerer Anregungspulswinkel für dazu, dass die Unterschiede in der T1-Relaxationszeit noch deutlicher hervortreten, was den Kontrast im Bild weiter verstärkt[22].

T2-Gewichtung

Die T2-Gewichtung wird durch die Wahl einer langen TR und einer langen TE erreicht. Diese Parameterkombination sorgt dafür, dass die transversale Relaxationszeit, auch als T2-Zeit bekannt, den Kontrast des Bildes dominiert, während der Einfluss der longitudinalen Relaxationszeit (T1-Zeit) minimiert wird. In der T2-Gewichtung variieren die Relaxationszeiten ebenfalls je nach Gewebeart erheblich. Gewebe mit einer kurzen T2-Relaxationszeit, wie beispielsweise Knochen, erholen sich schnell und geben nur ein schwaches Signal ab. Diese Gewebe erscheinen auf den Bildern hypointens, also dunkler. Im Gegensatz dazu haben Gewebe mit einer langen T2-Relaxationszeit, wie der Liquor cerebrospinalis, eine längere Erholungsphase und geben ein stärkeres Signal ab. Diese Gewebe erscheinen hyperintens, also heller, auf den T2-gewichteten Bildern[22].

2.1.1. Bestandteile und Sequenztechniken der kardialen MRT

Die kardiale Magnetresonanztomographie (CMR) ist neben der transthorakalen (TTE) und transösophagealen Echokardiographie (TEE)[23], der Computertomographie (CT)[24], nuklearmedizinischen Verfahren[25] und der invasiven Koronarangiografie[26] eine der zentralen Methoden in der bildgebenden Diagnostik kardialer Erkrankungen[27]. Eine besondere Stärke der CMR liegt darin, dass sie sowohl die morphologische als auch die funktionelle Beurteilung des Herzens und eine Gewebecharakterisierung ohne die Verwendung ionisierender Strahlung ermöglicht. Diese umfassende Diagnosefähigkeit erklärt die vielfältigen klinischen Anwendungen der CMR im Rahmen chronischer und akuter Pathologien. Sie wird zur Untersuchung angeborener Herzfehler ebenso eingesetzt wie zur Diagnose von ischämischen und nicht-ischämischen Kardiomyopathien sowie kardialer Tumore[27]. Auch bei der Beurteilung der Ausprägung von Klappenerkrankungen, primär als Domäne der Echokardiografie angesehen, nimmt sie eine immer wichtigere Rolle ein[28]. Weitere Anwendungsgebiete umfassen die Abklärung kardialer Speichererkrankungen wie die Amyloidose[29] sowie die Diagnose und Verlaufsbeurteilung einer Myokarditis[30].

Herz- und Atembewegungen mit entsprechenden Bewegungsartefakten stellen besondere Herausforderungen an die CMR dar. Daher werden die Sequenzen der CMR vorwiegend unter Verwendung von EKG-Triggerung während eines endexpiratorischen Atomstopps akquiriert[31]. Die Synchronisation der Akquisition auf den Herzzyklus erfolgt anhand des aufgenommenen EKG-Signal des Patienten: Die R-Welle des EKGs wird erkannt und löst einen Synchronisationsimpuls für die Signalakquisition aus. Dadurch kann das Herz entweder zu mehreren Zeitpunkten im Verlauf des Herzzyklus (Cine-Bildgebung) oder zu einem einzelnen Zeitpunkt (Standbildgebung) dargestellt werden[31].

Im Folgenden wird auf die wichtigsten Sequenztypen der CMR zur Diagnose einer Myokarditis eingegangen. Hierbei wird zwischen solchen unterschieden, welche die Analyse von Morphologie und Funktion ermöglichen, sowie solchen, welche das myokardiale Gewebe charakterisieren.

Sequenz für Morphologie und Funktion

a) Balanced Steady-State Free Precession “Cine”-Sequenz

Eine der Standardsequenzen in der CMR ist die während eines Atemstops akquirierte EKG-getriggerte zweidimensionale (2D) Balanced Steady-State Free Precession (bSSFP)-Sequenz, eine Gradientenechosequenz mit einem Flipwinkel von weniger als 90°. Dies ermöglicht eine Verkürzung der TR ohne erheblichen Signalverlust und generiert Bilder mit

einem T1-/T2-Mischkontrast. In der bSSFP-Sequenz erscheinen Flüssigkeiten und Fett deutlich heller als anderes Gewebe, wobei insbesondere das fließende Blut aufgrund des sogenannten „inflow enhancements“ ein starkes Signal aufweist. Dies führt zu einem hervorragenden Kontrast zwischen dem signalreichen Blut und dem signalarmen Myokard, was in der kardialen Bildgebung von Vorteil ist[32].
Da die zeitliche Auflösung der standardmäßig verwendeten bSSFP-Technik zu gering ist, um einen Herzzyklus vollständig zu erfassen, werden aus mehreren Herzzyklen Aufnahmen der Herzaktion zu verschiedenen Zeitpunkten akquiriert und anschließend zu einer Cine-Schleife zusammengesetzt. Die Cine-Sequenz liefert wertvolle Informationen über Morphologie und Funktion des Herzens, erlaubt die Detektion von Wandbewegungsstörungen und ermöglicht in der kurzen Achse die Quantifizierung der linksventrikulären Funktion, wofür sie den nicht-invasiven Goldstandard darstellt[33,34]. Zudem ist neben der Bestimmung der Masse des linken Ventrikels (LV) auch die Beurteilung von Klappenpathologien und etwaiger Obstruktionen des links- oder rechtsventrikulären Ausflusstrakts möglich[35].

Sequenzen der Gewebecharakterisierung

Die klassischen Sequenzen, namentlich die T2-gewichtete Sequenz zur Ödembildgebung und die T1-gewichtete Sequenz nach intravenöser Gabe von Gadolinium (Gd)-haltigen Kontrastmittel (Late Gadolinium Enhancement, LGE) sowie die neueren Mapping-Techniken (T1- und T2-Mapping) ermöglichen die nicht-invasive myokardiale Gewebecharakterisierung in der CMR[31]. Zudem wird aus historischen Gründen hinsichtlich der Entwicklung der CMR zur Diagnose der Myokarditis im Rahmen dieser Dissertation auch auf das Early Gadolinium Enhancement (EGE) eingegangen. Diese Sequenz findet jedoch im heutigen klinischen Alltag in der Regel keine Anwendung mehr[30].

a) <u>T2-gewichtete fettsupprimierte Sequenzen</u>

Die während eines Atemstops akquirierte EKG-getriggerte 2D T2-gewichtete Short-Tau-Inversion-Recovery (STIR)-Sequenz nutzt eine Kombination aus nicht-selektiven und selektiven 180° Inversionsvorbereitungspulsen, gefolgt von einer langen TI. Dadurch wird die Magnetisierung des Blutes unterdrückt, indem alle Spins außerhalb des Bildbereichs invertiert werden. Gleichzeitig bleiben die Gewebe innerhalb des stationären Bildbereichs, einschließlich des Myokards unbeeinflusst, da die beiden Pulse dort keinen Nettoeffekt auf die Spins haben. Die Sequenz wird primär in der kurzen Achse akquiriert und dient zur Erkennung eines Myokardödems bei akuten kardialen Erkrankungen wie einer Myokarditis oder einem Myokardinfarkt[36].

b) Early Gadolinium Enhancement

Das EGE basiert auf der Akquisition einer EKG-getriggerten 2D T1-gewichteten Turbo-Spin-Echo (TSE) Sequenz vor und innerhalb von 3 Minuten nach der intravenösen Verabreichung des Kontrastmittels, welche häufig unter freier Atmung in der kurzen Achse des linken Ventrikels oder im Vierkammerblick akquiriert wird. Das absolute myokardiale Enhancement wird als prozentuale Zunahme der Signalintensität des Myokards vor und nach der Kontrastmittelgabe in Relation zur Zunahme der Intensität des Skelettmuskels (üblicherweise Musculus pectoralis major) berechnet, basierend auf der untenstehenden Gleichung[37,38]:

$$Absolutes\ Enhancement\ (\%) = 1 - \frac{\text{Signalintensität post Gd} - \text{Singalintensität prä Gd}}{\text{Singalintensität prä Gd}}$$

Ein erhöhtes EGE wird durch die Early Gadolinium Enhancement Ratio (EGEr) bestimmt, welche durch eine Signalintensitäts-Ratio zwischen dem Myokard und dem Skelettmuskel von ≥ 4,0 oder durch ein absolutes Enhancement des Myokards von ≥ 45 % definiert wird. Die Gleichungen hierzu lauten[37,38]:

$$Early\ gadolinium\ enhancement\ ratio\ (EGE) = \frac{\text{Enhancement (Myokard)}}{\text{Enhancement (Skelettmuskel)}}$$

$$Enhancement = \frac{\text{Signalintensität post Gd} - \text{Singalintensität prä Gd}}{\text{Singalintensität prä Gd}}$$

Der Zweck der EGEr besteht darin, ein insgesamt erhöhtes Verteilungsvolumen von Gadolinium im intravaskulären Raum und interstitiellen Raum während der frühen Auswaschphase zu erkennen. Das EGE ermöglicht die Darstellung der vermehrten Durchblutung und des Kapillarlecksyndroms bei akuten kardialen Pathologien wie bei einer Myokarditis[37].

c) Late Gadolinium Enhancement

Für die LGE-Bildgebung wird üblicherweise eine EKG-getriggerte 2D/3D Inversion-Recovery T1-gewichtete Gradienten-Echo-Sequenz verwendet, welche 10-15 Minuten nach Kontrastmittelgabe entlang der unterschiedlichen Herzachsen während eines Atemstops akquiriert wird. Hierbei ist die manuelle Einstellung einer mittels einer Look-Locker-Sequenz

zu bestimmenden Inversionszeit (TI) notwendig, um das Signal des gesunden Myokards zu unterdrücken. Da sich das extrazelluläre Gadolinium im fibrotisch oder narbig verändertem Myokard akkumuliert, stellen sich solche Veränderungen in den LGE-Sequenzen hyperintens dar[39,40].

Die LGE-Bildgebung ermöglicht die nicht-invasive Vitalitätsdiagnostik des Myokards und stellt den nicht-invasiven Goldstandard für dessen Gewebecharakterisierung dar[31]. Zudem ermöglicht das differente Verteilungsmuster im LGE eine Unterscheidung zwischen ischämischen und nicht-ischämischen Kardiomyopathien[41]. Des Weiteren beinhaltet die Ausdehnung des LGE einen hohen prognostischen Wert hinsichtlich des Auftretens von schweren kardiovaskulären Ereignissen (MACE) bei diversen Kardiomyopathien[42,43].

d) Myokardiales Mapping

Die primär in der kurzen Achse unter EKG-Triggerung während eines Atemstopps akquirierten 2D „Mapping-Sequenzen“ ermöglichen die Detektion akuter und chronischer Myokardveränderungen anhand der Bestimmung von T1- und T2-Relaxationszeiten. Sie ergänzen die oben genannten LGE- und T2-gewichtete Sequenzen. Das T1-Mapping quantifiziert myokardiale Gewebeveränderungen wie Fibrose oder die Akkumulation von Amyloid während das T2-Mapping eine quantitative Bestimmung des im Rahmen akuter Pathologien auftretenden Ödems ermöglicht. Nach Kontrastmittelgabe kann das T1-Mapping unter Berücksichtigung des vorzugsweise tagesaktuellen Hämatokritwertes zudem zur Quantifizierung der Extrazellulärvolumen (EZV)-Fraktion genutzt werden[31,44].

T1-Mapping und EZV-Fraktion

Von den verfügbaren Techniken hat sich die „Modified Look-Locker Inversion-Recovery“ (MoLLI) als Standardverfahren zur Erstellung von T1-Karten etabliert[45,46]. Hierbei wird nach jeder Inversion die longitudinale Relaxationskurve bei unterschiedlichen Inversionszeiten abgetastet, wodurch Einzelbilder mit verschiedenen Inversionszeiten entstehen. Basierend auf diesen Einzelbildern wird mithilfe von Curve-Fitting die T1-Zeiten approximiert und in Form einer Karte (engl: Map) dargestellt[46]. Die T1-Zeiten sind abhängig von verwendetem MRT-System und Feldstärke sowie von der jeweiligen myokardialen Gewebezusammensetzung: So sind die T1-Zeiten bei einem erhöhten Anteil an freiem Wasser (zum einen bei einem myokardialen Ödem, z.B. im Rahmen einer akuten Myokarditis/eines akuten Infarkts, zum anderen bei einer Fibrose, z.B. im Rahmen einer hypertrophen Kardiomyopathie) verlängert sie, während myokardiale Lipid- oder Eisenablagerungen sie verkürzen[46].

Die T1-Zeiten können sowohl nativ als auch 10 Minuten nach Kontrastmittelgabe akquiriert werden. Zwar haben die T1-Zeiten nach Kontrastmittelgabe alleinstehend eine geringere klinische Relevanz, jedoch ist unter Berücksichtigung der Zeiten vor und nach Kontrastmittelgabe die Berechnung der EZV-Fraktion möglich. Diese beschreibt das Verhältnis zwischen zellulärem und extrazellulärem Anteil des Myokards. Die EZV-Fraktion ermöglicht eine Quantifizierung von akuten (wie bei einer Myokarditis) und chronischen (wie bei einer Amyloidose) myokardialen Texturstörungen[46].

T2-Mapping

Das T2-Mapping basiert ebenfalls auf einem pixelweisen Curve-Fitting über Einzelbilder, welche aber im Gegensatz zum T1-Mapping mit unterschiedlichen Echo- anstelle von Inversionszeiten akquiriert werden. Für das T2-Mapping wird vorwiegend eine T2-gewichtete Gradienten-Spin-Echo Sequenz (T2 GraSE) genutzt[47]. Für jede Schicht werden mehrere Bilder bei unterschiedlichen Echozeiten aufgenommen, um die T2-Abfallkurve zu rekonstruieren und eine pixelgenaue Anpassung zur Erstellung der Karten zu ermöglichen[46].

Hinsichtlich der klinischen Anwendung haben zahlreiche Studien die diagnostischen Vorzüge der T1- und T2-Mapping-Techniken für ischämische und nicht-ischämische Kardiomyopathien gegenüber den Standardsequenzen der Gewebecharakterisierung gezeigt, da sie im Vergleich zu diesen kein gesundes Referenzgewebe zur primär subjektiven Bestimmung bzw. Detektion pathologischer myokardialer Veränderungen erfordern, sondern diese objektiv quantitativ abbilden[9,48,49].

2.2. Myokarditis

Pathophysiologie

Die Pathogenese der Myokarditis ist weiterhin noch nicht vollständig geklärt und weiterhin Gegenstand der Forschung[5]. Auch wenn diverse Ätiologien bekannt sind, weisen alle Formen am Ende die gleichen myokardialen Gewebsprozesse auf: Zelluläre Infiltration, Hyperämie, Ödem, Nekrose und zum Ende eine mögliche Fibrose[37]. Die Entzündungsreaktion kann lokal begrenzt oder diffus im Myokard verteilt sein[50], jedoch ist anzumerken, dass ein zunächst lokaler Prozess auch im Verlauf disseminiert das gesamte Myokard befallen kann[51]. Der Myokarditisverlauf lässt sich in drei verschiedene Phasen einteilen. Während der akuten viralen Phase (innerhalb der ersten drei Tage), dringt das Virus in die Kardiomyozyten mittels

spezifischer Rezeptoren ein und schädigt die Zellen direkt [2,52]. Hierauf folgt die subakute Phase (etwa 10 Tage) mit einer autoimmunen Reaktion, in welcher das körpereigene Immunsystem eine direkte Zerstörung der befallenen Kardiomyozyten bewirkt[2]. Zum Schluss erreicht die Myokarditis die chronische Phase, die normalerweise bis zu 90 Tage andauert, in einigen Fällen jedoch deutlich länger sein kann[52]. In dieser Phase erfolgt die Ersetzung des geschädigten bzw. nekrotischen Gewebe durch Narbenbildung[53]. Wenn währenddessen durch die entzündliche Reaktion eine Entfernung des Erregers erreicht wird, kann sich der Herzmuskel vollständig regenerieren und der Patient gilt als genesen[54] und es resultiert eine Wiederherstellung bzw. Verbesserung der in der akuten Phase bisweilen eingeschränkten linksventrikulären Ejektionsfraktion (LVEF)[55]. In einigen Fällen kann jedoch eine dilatative Kardiomyopathie (DCM) entstehen, nachdem sich Autoimmunzellen sowie Kreuzantikörpern gegen kardiale Proteine wie Myosin oder Muskarinrezeptoren gebildet haben. Dieser Mechanismus führt zu einer fortschreitenden Schädigung der Herzmuskelzellen und zu strukturellen Umbauprozessen, die letztendlich zu kardialen Funktionsstörungen während der Systole und Diastole münden und hieraus eine Herzinsuffizienz resultieren kann[56]. Genetische Varianten, die mit einer DCM oder arrhythmischen Kardiomyopathie in Verbindung stehen, werden bei 8-16 % der Patienten mit einer Myokarditis festgestellt[57,58].

Klinische Präsentation

Die meisten Patienten mit einer Myokarditis zeigen keine Symptome. Wenn Symptome auftreten, können sie stark variieren und reichen von milden Beschwerden bis hin zu akut lebensbedrohlichen Zuständen[56]. In groß angelegten Patientenkollektiven zeigte sich, dass die akute Myokarditis häufig bei einem mittleren Alter zwischen 30 und 45 Jahren auftrat und etwa 60-80 % der Betroffenen Männer waren[59]. Patienten mit Myokarditis präsentieren sich zu 82-85 % der Patienten mit Brustschmerzen, 19-49 % mit Atemnot, 58-65 % mit Fieber, sowie 5-7 % mit Synkopen[59]. Zudem trat bei 3-9 % der Patienten mit akuter Myokarditis ein kardiogener Schock auf[59]. Palpitationen des Herzens sind als mögliches klinisches Symptom ebenfalls zu finden[60]. Als zusätzliche, sowie eher untypische Symptome können gastrointestinale und grippeähnliche Beschwerden bereits vorher auftreten[61]. Die Myokarditis stellt somit eine zu beachtende Differenzialdiagnose des thorakalen Brustschmerzes dar, auch im Krankenhaussetting. So konnte in eine Studie von Prepoudis et al. in einem Patientenkollektiv von 2533 Patienten, die sich in einer Notaufnahme mit Brustschmerzen vorstellten, für die Myokarditis eine Inzidenz von 1,1 %, für Perikarditis von 1,9 %, und für einen Nicht-ST-Hebungsinfarkt (NSTEMI) von 21,6 % beobachtet werden[62]. Zu beachten ist, dass der größte Anteil der Patienten Brustschmerzen nicht-kardialer Genese aufwies (76,2 %)[62].

Diagnostik

Der erste Schritt in der diagnostischen Abklärung einer Myokarditis umfasst neben der körperlichen Untersuchung und Anamnese die Bestimmung von Laborwerten bzw. Biomarkern. Unspezifische Parameter, die bei allgemeinen Entzündungsprozessen im Körper erhöht sind, wie die Blutsenkungsgeschwindigkeit (BSG), das C-reaktive Protein (CRP), oder die Anzahl an Leukozyten, gelten als Indikatoren[63]. Herzspezifische Biomarker, die häufig bei einer Myokarditis erhöht sind, umfassen hochsensitive kardiale Troponine (Hs-cTn) und die Creatinkinaseform MB (CK-MB). Diese Marker weisen eine hohe Spezifität und Sensitivität für myokardiale Schädigungen auf[64]. Es ist jedoch zu berücksichtigen, dass normale Werte dieser Parameter eine Myokarditis nicht ausschließen.
Hinsichtlich der apparativen Diagnostik sind im EKG bei 62-96 % der Patienten mit akuter Myokarditis abnormale Befunde anzutreffen[59]. So werden ST-Streckenhebungen bei etwa 58-70 % der Patienten mit akuter Myokarditis beobachtet[59]. T-Inversionen stellen eine weitere mögliche EKG-Veränderung dar. So traten beispielsweise bei 9% der Patienten mit Myokarditis in einer Studie von Di Bella et. während der EKG-Ableitung im Rahmen der CMR diese Veränderungen auf, welche bei 57% der Patienten nach 48h noch persistierten[65].
In der TTE zeigen sich bei einer Myokarditis häufig eine verdickte Myokardwand und eine gesteigerte Echogenität des Myokards[59]. Weitere häufig beobachtete Veränderungen sind segmentale Wandbewegungsstörungen wie die Hypokinesie, welche typischerweise die Hinter- und Lateralwand des linken Ventrikels betrifft, diastolische Funktionsstörungen und ein Perikarderguss. In der Anfangsphase einer akuten Myokarditis sind die LV-Dimensionen in der Regel normwertig und die LVEF erhalten, jedoch kann die Herzfunktion in den ersten Tagen rasch abnehmen und zu einer kardiopulmonalen Verschlechterung mit Intensivpflichtigkeit führen[59].
Aufgrund der oben genannten ST-Streckenhebungen stellt das akute Koronarsyndrom (ACS) eine wichtige Differenzialdiagnose dar, welche ausgeschlossen werden sollte. Daher wird bei etwa 46-95 % der erwachsenen Patienten mit akuter Myokarditis eine invasive Koronarangiographie oder eine Herz-CT zum Ausschluss eines ACS durchgeführt, insbesondere bei Patienten mit kardiovaskulären Risikofaktoren, Brustschmerzen, ST- und T-Veränderungen oder segmentalen Wandbewegungsstörungen im Echokardiogramm[59].

Gemäß den European Society of Cardiology (ESC)-Leitlinien von 2013 wird die Endomyokardbiopsie (EMB) aus folgenden Gründen als der Goldstandard für die Diagnose einer Myokarditis gesehen[7]:

Direkte Gewebeuntersuchung: Die EMB ermöglich die direkte Untersuchung des Myokardgewebes, was das Vorhandensein einer Entzündung bestätigen und die zugrundeliegende Ursache der Myokarditis, wie virale Infektionen oder Autoimmunprozesse, identifizieren kann. Dieser direkte Ansatz liefert eindeutige Beweise, die nicht-invasive Tests nicht bieten können[7,59].

Charakterisierung der Entzündung: Die Biopsie kann die Art der vorliegenden Entzündung charakterisieren (z. B. Riesenzellmyokarditis, eosinophile Myokarditis, Sarkoidose), was entscheidend für die Bestimmung der richtigen Behandlung und Prognose ist[7]. Sie ist zudem die einzig anerkannte Methode, um zwischen einer autoimmunen und viralen Myokarditis zu differenzieren[66].

Erhöhte diagnostische Genauigkeit: Die EMB wird durch molekulare Analysen, wie den Nachweis viraler Genome mittels PCR-Analyse, ergänzt, was zusätzlichen diagnostischen Wert bieten und systemische Infektionen ausschließen kann[7].

Dennoch hat die EMB auch nicht zu vernachlässigende Nachteile. Aufgrund der Invasivität des Eingriffs besteht ein Risiko für Komplikationen wie Infektionen, Perforationen der Herz- und Gefäßwände, Embolien, Blutungen und Arrhythmien[50,67]. Die Komplikationsrate reicht von 0-0,8% für erfahrene bis zu 6% für unerfahrene Untersucher[7,66,68].

Für die finale Diagnose einer Myokarditis muss ein positiver diagnostischer Test (pathologische EKG-Veränderungen, erhöhte kardiale Biomarker, positiver Befund in der EMB, positiver Befund in der CMR wie unten weiter ausgeführt, positiver Befund in der TTE) sowie ein klinisches Symptom (Brustschmerzen, Herzinsuffizienz, Extrasystolen/Palpitationen, unerklärbarer kardiogener Schock) vorliegen. Bei asymptomatischen Patienten hingegen sind zwei positive diagnostische Tests nötig[7].

Aufgrund ihrer fehlenden Invasivität wird die CMR in der Regel der EMB vorgezogen[59]. Auf die diagnostischen Kriterien einer Myokarditis in der CMR und ihre historische Entwicklung wird im folgenden Kapitel näher eingegangen.

2.2.1. Originale Lake-Louise-Kriterien

Die „International Consensus Group on Cardiovascular Magnetic Resonance in Myocarditis" wurden im Jahr 2006 gegründet. Ihr Ziel war es einheitliche CMR-basierte Kriterien für die

Diagnose einer akuten Myokarditis zu entwickeln und dadurch die Standardisierung und Verbesserung der nicht-invasiven Diagnostik voranzutreiben. Das Ergebnis dieser Bemühungen war die Veröffentlichung der sogenannten „Lake-Louise-Kriterien“ im Jahr 2009[37].

Die originalen Kriterien fokussierten sich hierbei auf drei wesentliche Merkmale der Myokardentzündung, von denen für die positive Diagnose einer Myokarditis zwei von drei Kriterien vorliegen mussten:

a) Ödem in der T2-gewichteten Sequenz

Myokardiales Ödem, also eine verlängerte T2-Relaxationszeit des Myokards bzw. eine erhöhte Signalintensität in T2-gewichteten Aufnahmen bedingt durch erhöhte Durchlässigkeit der Zellmembranen mit vermehrtem Natrium-Einstrom und Nettoausstrom von Wasser und daraus resultierender transmembraner Leckage größerer Moleküle, sowie möglichem Funktionsverlust der Zellen. Neben der visuellen Analyse ist auch eine quantitative Bestimmung der Signalintensität des gesamten Myokards in Relation zur Thoraxmuskulatur mit Bestimmung des T2-Signalintensitätsverhältnis möglich[37], welches durch ein Signalintensitätsverhältnis zwischen Myokard und der Skelettmuskulatur von ≥ 2,0 definiert wird[37].

b) Hyperämie im EGE

Die regionale Vasodilatation (Hyperämie) aufgrund einer Myokarditis führt zu einer gesteigerten Aufnahme von Kontrastmittel im intravaskulären bzw. interstitiellem Raum, welches auf Basis der mittels der EGE-Technik bestimmbaren erhöhten EGE Ratio (EGEr) zu detektieren ist[37].

c) Nekrose und/oder Fibrose im LGE

In der akuten Phase der Myokarditis dringt Gadolinium zusammen mit Wasser im Rahmen der Nekrose durch die geschädigten Zellmembranen in den Kardiomyozyten ein, was das Verteilungsvolumen von Gadolinium vergrößert und die durch Myokarditis befallenen Anteile des Myokards im LGE als hyperintense Läsionen in vorzugsweise subepikardialer Lokalisation sichtbar macht. Sobald die entzündlichen und nekrotischen Veränderungen abklingen, wird in der chronischen Phase das zuvor funktionsfähige Gewebe durch ein Netz aus Fibrozyten ersetzt. Dadurch vergrößert sich erneut das Gadolinium-Verteilungsvolumen im extrazellulären Raum während der Auswaschphase und kann

mittels LGE dargestellt werden. Da das LGE sowohl in der akuten als auch in der chronischen Phase der Myokarditis sichtbar sein kann, ist es allein nicht ausreichend, um zwischen den beiden Phasen unterscheiden zu können[37].

Nach Veröffentlichung zeigten diverse Studien Nachteile der T2-gewichteten Sequenz und vor allem der EGE-Technik. So ist die Qualität der T2-gewichteten Aufnahmen häufig durch Signalinhomogenitäten beeinträchtigt, welche die Detektion eines myokardiales Ödem, vor allem wenn dieses diffus ausgeprägt ist, erschweren kann[48,69,70]. Beide sind anfällig für Artefakte, insbesondere durch Atembewegungen und Arrhythmien[69]. Die EGE hängt zudem von der korrekten Wahl der Schnittebene und der Auswahl der richtigen Segmente ab[70]. Diese technischen Nachteile spiegeln sich auch in der diagnostischen Evaluation der Lake-Louise Kriterien: So konnten beispielsweise Luetkens et al. demonstrieren, dass T1- und T2-Mapping gemeinsam mit LGE-Bildgebung eine signifikant höhere diagnostische Leistungsfähigkeit bei Patienten mit akuter Myokarditis als die Lake-Louise Kriterien alleine aufwiesen.

2.2.2. Revidierte Lake-Louise-Kriterien von 2018

Als Reaktion auf die oben genannten Erkenntnisse wurden die ursprünglichen Lake-Louise-Kriterien im Jahr 2018 umfassend überarbeitet indem das EGE als diagnostisches Kriterium nicht mehr berücksichtigt wird und die Mapping-Techniken hinzugefügt wurden[8]. Nach den 2018 aktualisierten Lake-Louise-Kriterien stützt sich die Diagnose einer Myokarditis mittels CMR auf die Kombination von jeweils mindestens einem T1-basierten Kriterium (verlängerte T1-Relaxationszeiten, erhöhte EZV-Fraktion oder positives entzündliches LGE) und einem T2-basierten Kriterium (verlängerte T2-Relaxationszeiten, sichtbares Myokardödem oder ein erhöhtes T2-Signalintensitätsverhältnis)[8]. Abbildung 1 zeigt den Vergleich zwischen den originalen und revidierten Lake-Louise-Kriterien von 2018.

In einer Studie von Luetkens et al., welche die ursprünglichen Lake-Louise-Kriterien mit den 2018 aktualisierten Kriterien verglich, wurde deutlich, dass die revidierten Kriterien im Vergleich zu den ursprünglichen Kriterien die Diagnose einer akuten Myokarditis signifikant verbessert. So zeigten die revidierten Kriterien bei einer klinisch diagnostizierten Myokarditis eine Sensitivität von 87,5% und Spezifität von 96,2% bei 40 Patienten, während die originalen Kriterien eine Sensitivität von 72,5% und Spezifität von 96,2% aufwiesen. Es ist bemerkenswert, dass kein Patient, der nach den ursprünglichen Kriterien diagnostiziert wurde, in den aktualisierten Kriterien übersehen wurde[30].
Die aktualisierten Kriterien von 2018 verbessern die diagnostische Leistung der CMR im Vergleich zu den ursprünglichen Kriterien erheblich, und die Integration der neuen

parametrischen Bildgebungstechniken in routinemäßige Diagnoseprotokolle stellt heute an den meisten Kliniken und Praxen den Alltag dar.

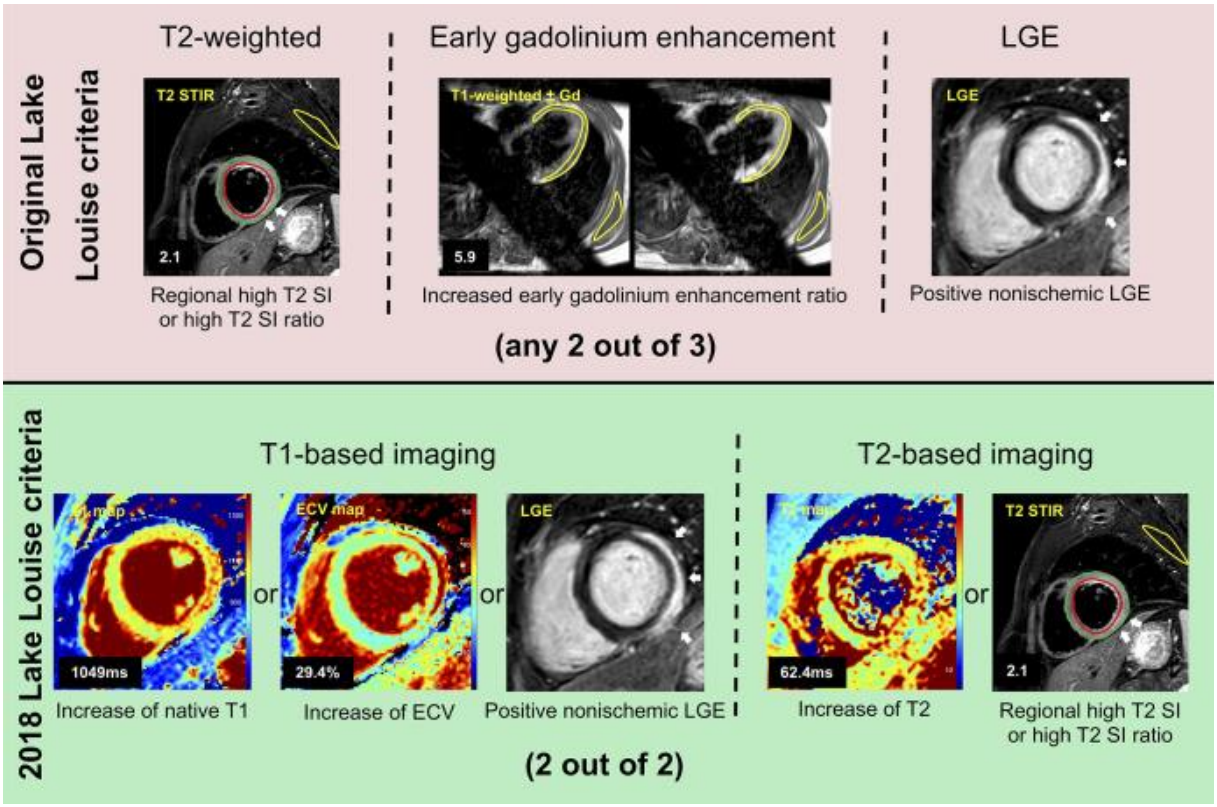

Abbildung 1: Vergleich der originalen Lake-Louise-Kriterien mit denen der 2018 revidierten und aktuell gültigen Version. LGE: Late Gadolinium Enhancement, Gd: Gadolinium, ECV: Extrazelluläres Volumen. STIR: Short Tau Inversion Recovery, SI: Signalintensität.
Abbildung aus Luetkens JA, Faron A, Isaak A, *et al.* Comparison of Original and 2018 Lake Louise Criteria for Diagnosis of Acute Myocarditis: Results of a Validation Cohort. *Radiol Cardiothorac Imaging* 2019; **1**: e190010.

2.3. Künstliche Intelligenz

Der Begriff “Künstliche Intelligenz” (KI) oder intelligente Maschinen umfasst unter anderem komplexe Algorithmen, die in der Lage sind, Aufgaben autonom auszuführen und sich an unbekannte Situationen anzupassen. Die beiden wichtigsten Konzepte in diesem Bereich sind Machine Learning und Deep Learning.

2.3.1. Machine Learning und Deep Learning

Machine Learning setzt Algorithmen ein, um Daten zu analysieren, daraus zu lernen und basierend auf diesen Erkenntnissen Entscheidungen zu treffen. Ziel ist es, Muster in den Daten zu erkennen, die zur Vorhersage oder Entscheidungsfindung genutzt werden können. Eine spezielle Form des Machine Learning ist das sogenannte Deep Learning, das auf komplexen neuronalen Netzwerken basiert. Deep Learning hebt sich vom traditionellen Machine Learning durch seine Fähigkeit ab, Informationen in mehreren Schichten zu verarbeiten und besonders effektiv bei der Analyse und Interpretation von Bildern, wie beispielsweise radiologischen MRT- oder CT-Aufnahmen, zu arbeiten. Im Gegensatz zum klassischen Machine Learning, bei dem relevante Bildmerkmale manuell extrahiert und dem Algorithmus zugeführt werden müssen, erlernen Deep Learning Methoden diese Merkmale direkt aus den gegebenen Bilddaten. Die künstlichen neuronalen Netzwerke, die im Deep Learning zum Einsatz kommen, sind von der Struktur des menschlichen Gehirns inspiriert, insbesondere vom Hirnkortex, um menschliche Lernprozesse zu imitieren. Diese Netzwerke, auch als Deep Learning Modelle bezeichnet, sind in der Lage, eigenständig zu lernen und „intelligente“ Entscheidungen zu treffen, indem sie komplexe Muster in den Daten identifizieren. Ein solches neuronales Netzwerk besteht aus mehreren Schichten, die jeweils eine spezifische Rolle bei der Verarbeitung der Bilddaten spielen. Die erste Schicht, die sogenannte Eingabeschicht (engl: input layer), empfängt die Rohdaten, wie beispielsweise die einzelnen Voxel eines Bildes. In den nachfolgenden tieferen Schichten, die man auch als verborgene Schicht (engl: hidden layers) kennt, werden die Daten durch verschiedene Abstraktionsebenen hindurch verarbeitet, wobei jede Schicht zunehmend komplexere Merkmale aus den Bildinformationen extrahiert. Die letzte Schicht, die Ausgabeschicht (engl: output layer), liefert schließlich das Endergebnis, wie die Klassifizierung oder Segmentierung eines Bildes. Dank dieser mehrstufigen Verarbeitungsstruktur sind Deep Learning Modelle in der Lage, hochkomplexe Muster in Bilddaten zu erkennen und darauf basierende Entscheidungen zu treffen[71–74].
Abbildung 2 zeigt die Struktur eines neuronalen Netzwerks.

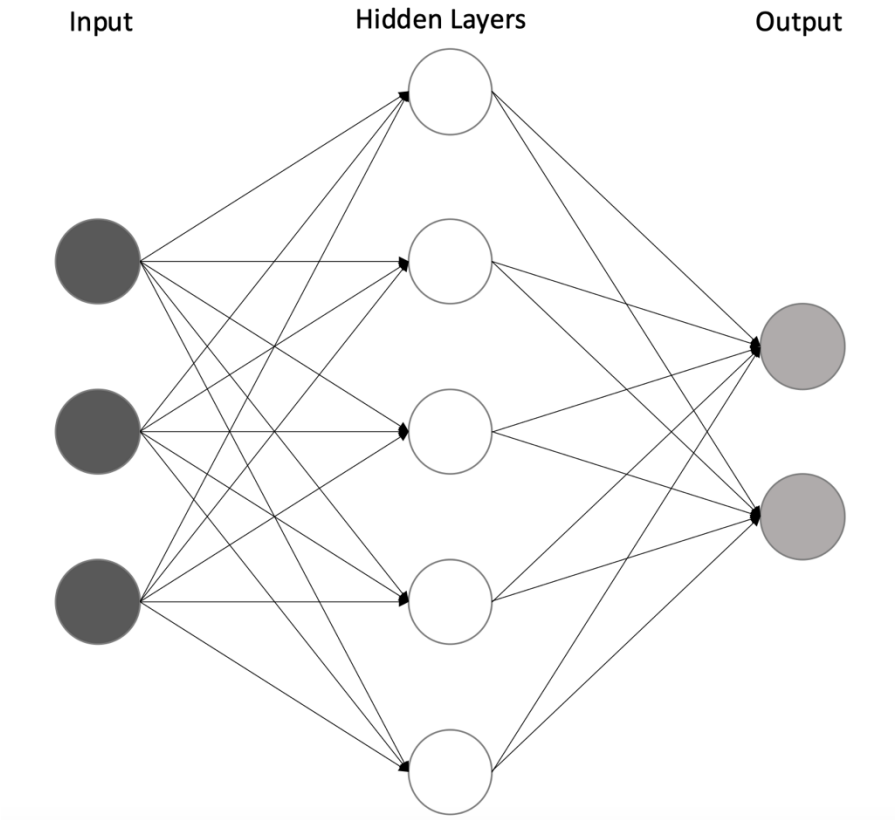

Abbildung 2: Struktur eines neuronalen Netzwerks

2.3.2. Large Language Models (LLMs)

Die KI bietet potenzielle Anwendungen im nahezu gesamten radiologischen Workflow, darunter die Verbesserung der Bildqualität (etwa durch Verkürzung der Bildaufnahmezeit und/oder Reduzierung der Strahlendosis), die Nachbearbeitung von Bildern (wie Bildannotation und Bildsegmentierung) sowie die Interpretation von Bildern, beispielsweise zur Vorhersage von Diagnosen[75]. Mit der Weiterentwicklung der natürlichen Sprachverarbeitung (Natural Language Processing; NLP) und insbesondere großer Sprachmodelle (Large Language Models; LLMs) wird klar, dass KI-Anwendungen in der Radiologie weit über bildbezogene Aufgaben hinausgehen. LLMs könnten einen erheblichen Einfluss auf die Radiologie haben, da Radiologen vor allem textbasierte Berichte verfassen, in denen sie ihre Interpretation diagnostischer Bilder und deren klinische Bedeutung darlegen[76]. Die Entwicklung der Sprachmodellierung lässt sich technisch in drei Stufen unterteilen: statistische Sprachmodelle, neuronale Sprachmodelle und vortrainierte Sprachmodelle (Pre-trained language models; PLMs)[76].

Abbildung 3 zeigt einen Überblick der Entwicklung der jeweiligen Sprachmodelle.

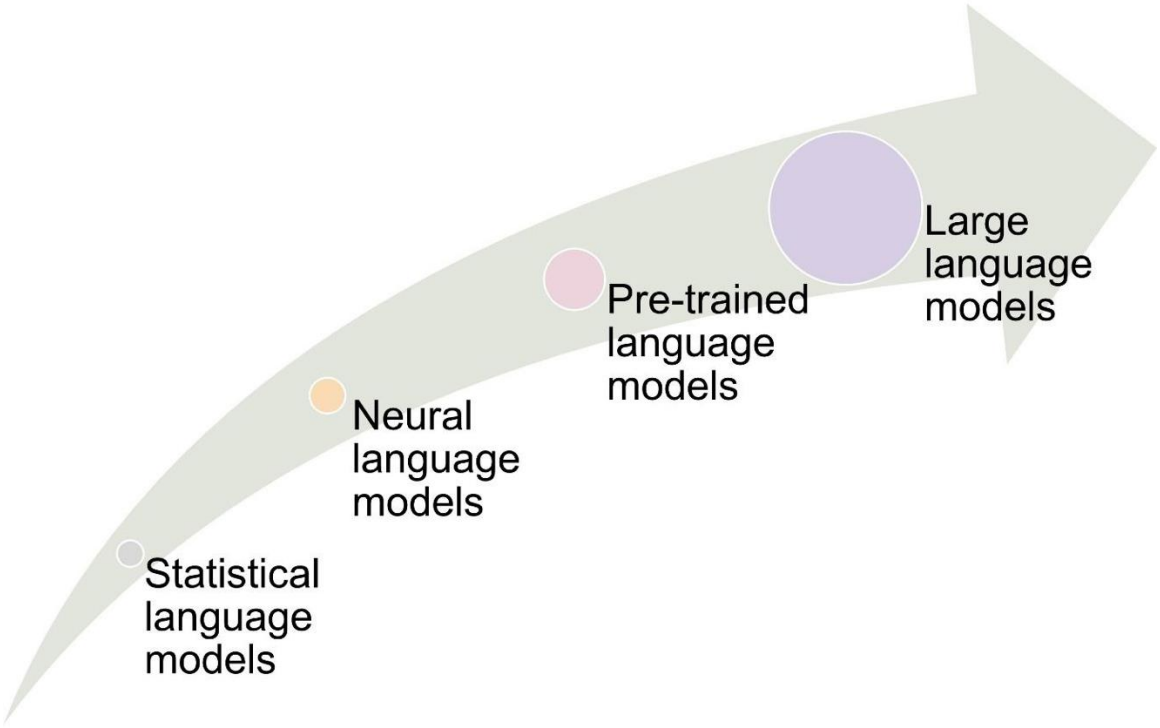

Abbildung 3: Entwicklungen der verschiedenen Sprachmodelle.
Abbildung aus Akinci D'Antonoli T, Stanzione A, Bluethgen C, *et al.* Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagnostic and Interventional Radiology* 2024; **30**: 80–90.

PLMs werden einmalig mit unüberwachten Lernmethoden trainiert, das heißt, sie lernen Muster aus unbeschrifteten Daten anhand einer riesigen Menge an Textdaten. Diese Modelle können dann für eine Vielzahl von Aufgaben genutzt werden, ohne dass ein erneutes Training von Grund auf erforderlich ist. Dank ihrer Zero-Shot- und Few-Shot-Lernfähigkeiten sind PLMs in der Lage, zu generalisieren und sich mit wenig oder gar keinem zusätzlichen Training an neue Aufgaben und Daten anzupassen[76]. Diese groß angelegten PLMs zeigen im Vergleich zu kleineren Modellen überraschende Verhaltensunterschiede und entwickeln neue Fähigkeiten, die es ihnen ermöglichen, komplexe Aufgaben zu bewältigen, wie etwa kontextbasiertes Lernen, das Befolgen von Anweisungen und schrittweises logisches Denken[77,78]. PLMs sind in der Lage, durch kontextbasiertes Lernen die gewünschten Ergebnisse zu erzielen, ohne dass zusätzliches Training oder Gradient-Anpassungen nötig sind. Sie können Ergebnisse für neue Aufgaben allein anhand von Anweisungen liefern, ohne dass explizite Beispiele gegeben werden müssen. Daher hat die Forschungsgemeinschaft den Begriff LLMs für diese riesigen PLMs eingeführt, die bis zu Hunderte von Milliarden Parametern umfassen können[79,80].

LLMs basieren typischerweise auf einer Transformer-Architektur, die aus rechnerischer Sicht sehr gut parallelisierbar ist[81]. Transformer bestehen im Wesentlichen aus Encodern und Decodern, die jeweils über einen speziellen Aufmerksamkeitsmechanismus verfügen[81]. Dieser Mechanismus verwendet eine Punktproduktoperation, um Ähnlichkeitswerte zu berechnen, wodurch das Modell in der Lage ist, bestimmten Eingaben mehr Aufmerksamkeit zu schenken als anderen, unabhängig von deren Position in der Eingabesequenz. Dadurch kann das Modell den Kontext eines Wortes besser erfassen. Im Gegensatz zu rekurrenten neuronalen Netzwerken ermöglicht der Aufmerksamkeitsmechanismus dem Modell außerdem, den gesamten Satz oder sogar den gesamten Absatz auf einmal zu betrachten, anstatt Wort für Wort vorzugehen. Abbildung 4 zeigt die Struktur, auf denen LLMs basieren[76].

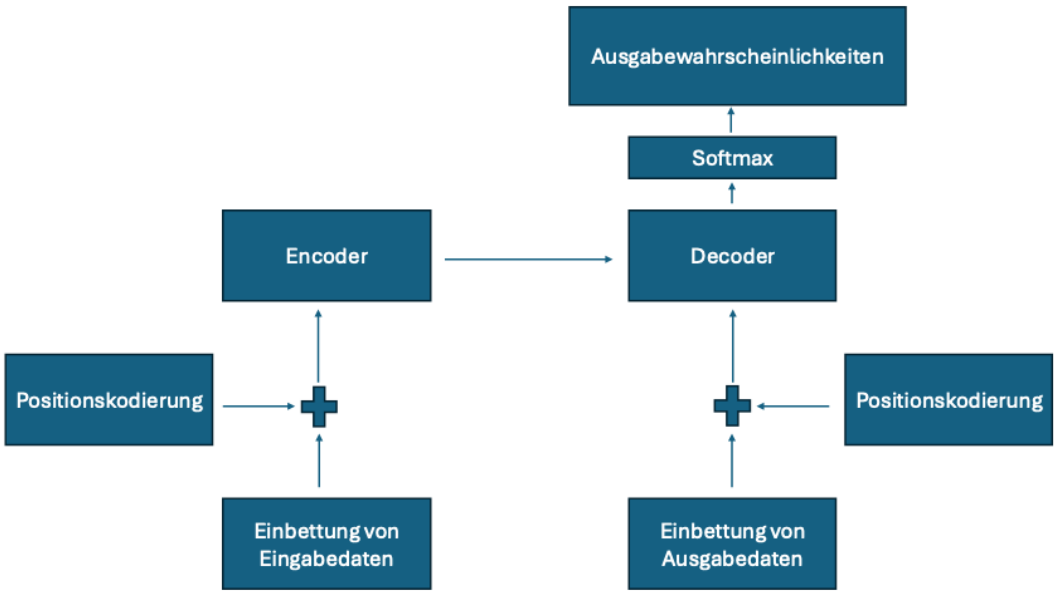

Abbildung 4: Schematische Struktur von LLMs.
Transformer bestehen aus mehreren Schichten von Encodern und Decodern. Mittels der Einbettung von Ein- und Ausgabedaten erzeugt die Softmax Wahrscheinlichkeitsvorhersagen und somit Ausgabewahrscheinlichkeiten. Abbildung modifiziert aus Akinci D'Antonoli T, Stanzione A, Bluethgen C, *et al.* Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagnostic and Interventional Radiology* 2024; **30**: 80–90.

Ein Prompt ist im Kontext von LLMs eine Eingabe, die dazu dient, die Ausgabe des Modells zu steuern. Diese Prompts bestehen häufig aus Sequenzen in natürlicher Sprache, können jedoch auch andere Formen von strukturierten Informationen umfassen. Sowohl die Syntax des Prompts (wie Struktur, Länge, Reihenfolge) als auch dessen semantischer Inhalt (wie Wortwahl und Tonfall) beeinflussen die Ergebnisse der LLMs maßgeblich[82].

Zum Zeitpunkt der Veröffentlichung der Publikation, auf der die vorliegende Dissertation basiert, veröffentlicht worden ist, sind die neuesten von OpenAI (San Francisco, Kalifornien, USA) veröffentlichten LLMs GPT-3.5, GPT-4 und ChatGPT. Diese Tools basieren alle auf der Transformer-Architektur, wie das Akronym GPT erkennen lässt. Angesichts der bisherigen Entwicklungen im Bereich von LLMs stellen ChatGPT und GPT-4 zwei herausragende Errungenschaften dar, die den Standard für die Fähigkeiten bestehender KI-Systeme erheblich erhöht haben[83].

Das GPT-3.5-Modell ist eine optimierte Version von GPT-3 und wurde als ein "Completion"-Modell trainiert, was bedeutet, dass es passende Wörter generieren kann, die den Eingabewörtern folgen. Im Gegensatz dazu ist GPT-4 ein völlig neues, großes multimodales Modell, das mithilfe von Reinforcement Learning mit menschlichem Feedback (Reinforcement Learning With Human Feedback; RLHF) weiterentwickelt wurde, um besser den menschlichen Erwartungen zu entsprechen. Die Erweiterung von Texteingaben auf multimodale Signale wird als ein bedeutender Fortschritt betrachtet. Insgesamt übertrifft GPT-4 im Vergleich zu GPT-3.5 in der Fähigkeit, komplexe Aufgaben zu lösen, was sich in einer deutlich verbesserten Leistung bei verschiedenen Bewertungstests zeigt[84].

ChatGPT, entwickelt auf Basis von GPT-3.5 und GPT-4, wurde gezielt darauf ausgelegt, Konversationsantworten zu generieren und als Modell im Dialogstil optimiert. Durch den Einsatz von RLHF wurde es weiter verfeinert[85]. Hierbei wurden die Ausgaben des Modells von Menschen bewertet, und ein Belohnungssystem half dabei, das Modell besser an menschliche Erwartungen anzupassen. Diese Anpassung ist möglicherweise ein Schlüsselfaktor für den Erfolg und hat seit der Einführung von ChatGPT großes Interesse in der KI-Community geweckt, da es enormes Potenzial für die menschliche Kommunikation zeigt. Die Implementierung von ChatGPT in dialogbasierten Interaktionen eröffnet vielfältige Möglichkeiten für die Mensch-Computer-Interaktion. Dank seiner Fähigkeit, Kontext zu verstehen, sinnvolle Antworten zu generieren und den Gesprächsfluss aufrechtzuerhalten, ist es ein wertvolles Werkzeug in zahlreichen Bereichen, wie beispielsweise im Kundensupport, bei Brainstorming-Sitzungen, der Inhaltserstellung und der Nachhilfe[86].

2.3.3. Verwendung von LLMs in der Radiologie

Die Verwendung von LLMs in der Radiologie bietet das Potenzial, fast jeden Schritt im radiologischen Arbeitsablauf positiv zu beeinflussen - von der Entscheidung, ob und wie eine Bildgebung durchgeführt werden soll, bis hin zur Unterstützung von Ärzten und Patienten bei der Interpretation von Ergebnissen. Trotz dieser vielversprechenden Perspektive erfordert jeder Anwendungsfall vor dem Einsatz eine gründliche Validierung. Die Anforderungen an eine akzeptable Leistung hängen von verschiedenen Faktoren ab, darunter die spezifische Art der Anwendung, ihr beabsichtigter Nutzen und das Risiko von Fehlern. Auch wenn in naher Zukunft für die meisten radiologischen Anwendungen weiterhin eine menschliche Überwachung erforderlich sein wird, sind LLMs gut geeignet, um die Effizienz und Qualität in der Radiologie zu steigern[16]. Die wichtigsten klinischen Einsatzmöglichkeiten werden im Folgenden beschrieben.

Vor der Bildgebung: Entscheidungsunterstützung, klinische Anamnese und Protokollierung

Aktuelle Richt- und Leitlinien aus verschiedensten Gesellschaften unterstützen Ärzte dabei zu entscheiden, wann eine Bildgebung sinnvoll ist und welche Modalitäten dafür am besten geeignet sind. LLMs können dabei helfen, die klinische Entscheidungsfindung zu optimieren[87,88]. Ein Beispiel dafür ist ein LLM, welche die Bildgebungsempfehlungen des American College of Radiology anwendete und dabei bessere Ergebnisse als Radiologen erzielte - und das zu geringeren Kosten[89]. Die Auswahl der korrekten Bildgebungsmodalität und Befundqualität ist abhängig von einer suffizienten klinischen Anamnese und Fragestellung im Rahmen der Anforderungen im Radiologieinformationssystem (RIS)[90,91]. LLMs können Radiologen bei der Auswahl der passenden Bildgebungsprotokolle unterstützen, eine wichtige aber oft monotone und zeitintensive Aufgabe[16]. So wurden In der Vergangenheit NLP-Werkzeuge für die automatische Protokollierung in bestimmten Bereichen entwickelt[92]. LLMs bieten jedoch die Möglichkeit flexibler und genauer zu arbeiten[16]. So konnte eine Studie von Gertz et al. zeigen, dass GPT-4 mit Zero-Shot-Prompting, also ohne vorherige exemplarische Fälle, eine hohe Sensitivität (84%) bei der korrekten Auswahl der passenden Bildgebungsmodalität, Kontrastmittelgabe und -phase basierend auf klinische Anamnese und Fragestellung im RIS im Vergleich zur Protokollfestlegung eines Radiologen mit 17 Jahren Berufserfahrung erzielte[14].

Nach der Bildgebung: Befunderstellungstools und Befundextraktion

Der Einsatz von LLMs kann den Arbeitsprozess bei der Erstellung von Befunden effizienter gestalten[16]. Sie können beispielsweise automatisch Befundeindrücke generieren, indem sie wichtige Befunde zusammenfassen[93], teils sogar aus mehreren Befunden[15], unstrukturierte bzw. freitextbasierte Befunde in strukturierte Befunde umwandeln[94] und hierbei relevante Differentialdiagnosen vorschlagen[95]. Wenn diese Werkzeuge präzise genug sind, könnten sie Radiologen dabei unterstützen, Befunde schneller und mit weniger Aufwand zu erstellen, während gleichzeitig Fehler durch das Auslassen wichtiger Informationen verringert werden[16]. Das Fein-Tuning mit Daten aus radiologischen Befunden trägt zur Verbesserung der Leistung automatisierter Berichtseindrücke bei[96]. Einige dieser Technologien werden bereits in kommerziellen Diktierlösungen wie PowerScribe Smart Impression von Nuance Communications (Burlington, Massachusetts, USA) und Omni Impressions von Rad AI (San Francisco, Kalifornien, USA) verwendet[16]. LLMs wie GPT-4 zeigen auch eine hohe Leistungsfähigkeit bei der Extraktion und Klassifizierung von Informationen aus Radiologiebefunden[97]. Durch ihre starke Fähigkeit, mit wenigen Beispielen zu lernen, könnten LLMs von Klinikern flexibel in elektronischen Krankenakten genutzt werden. So könnten beispielsweise Benachrichtigungen eingerichtet werden, die auf wichtige, für das jeweilige Fachgebiet relevante Befunde hinweisen, wie etwa Thrombosen oder anderweitige bedeutsame Zufallsbefunde[16].

Verbesserung der Patientenkommunikation: Beantwortung von Fragen und Erläutern von Befunden

Eine niedrige Patientencompliance wird mit einer geringeren Teilnahme bei Bildgebungsscreenings und schlechteren gesundheitlichen Ergebnissen in Zusammenhang gebracht[98]. Oft haben Patienten Fragen zu ihren Bildgebungen, jedoch fehlt ihnen häufig der rechtzeitige Zugang zu einem Arzt. In diesem Kontext bieten LLMs großes Potenzial, allgemeine medizinische Fragen zu beantworten und können dies in verschiedenen Sprachen tun[16]. Zum Beispiel lieferten sie Antworten, die von medizinischen Fachkräften als einfühlsamer und vorzuziehen gegenüber den Antworten von Ärzten in Online-Foren bewertet wurden[99]. Des Weiteren zeigte eine Studie von Salam et. al, dass komplexe CMR-Befunde mittels GPT-4 einfach und verständlich für den medizinischen Laien erklärt werden können[100], was ebenfalls die Patientenkommunikation und das Wohlfühlsein verbessern kann. Der Einsatz von LLMs in diesem Bereich weist Potenzial auf[101], doch ihre Genauigkeit und Sicherheit der Anwendung müssen weiter untersucht werden[16].

LLM-basierte Tools, die speziell darauf ausgelegt sind, die Patientencompliance zu fördern, sollten in überwachten klinischen Umgebungen getestet werden, etwa zur Unterstützung von Ärzten bei der präziseren und effizienteren Formulierung von Antworten. Da das Verhalten von LLMs durch Nutzereingaben beeinflusst werden kann, einschließlich unangemessener Anfragen, ist es notwendig patientenorientierte Tools mit eingeschränktem Funktionsumfang und geeigneten Schutzvorkehrungen zu entwickeln, um die Sicherheit zu gewährleisten[16].

Insgesamt ist es wichtig zu betonen, dass LLMs in der Radiologie zwar ein nützliches Werkzeug darstellen können, aber die Expertise von Radiologen ergänzen und nicht ersetzen sollten. Ein besonderes Problem bei ChatGPT ist seine Neigung, auch bei falschen Antworten mit großer Zuversicht aufzutreten. Diese Eigenschaft könnte bei klinischer Anwendung potenziell negative Auswirkungen haben[102].

2.4. Fragestellung

Die Diagnose einer Myokarditis erfordert eine umfassende Analyse multimodaler Daten, zu denen in der Regel die CMR, klinische Symptome und Blutwerte gehören. Eine präzise Interpretation und Integration der MRT-Befunde setzt dabei ein tiefgehendes radiologisches Fachwissen und Erfahrung voraus. GPT-4, ein LLM, könnte hier eine Unterstützung hinsichtlich der text-basierten Diagnose der Daten bieten. Das Ziel dieser Studie war es, die Fähigkeit von GPT-4 hinsichtlich der Diagnose einer Myokarditis basierend auf dem radiologischen Befundbericht der CMR, der klinischen Symptome und Blutwerte, falls vorhanden, zu untersuchen und gegenüber drei Radiologen mit einem, zwei und vier Jahren Berufserfahrung verglichen. Die Leistung von GPT-4 und der menschlichen Leser wurde mit dem Referenzstandard von zwei Experten in kardiovaskulärer Bildgebung mit jeweils acht und zehn Jahren Erfahrung verglichen. Weder die Radiologen noch GPT-4 hatten Zugang zu den Bilddaten oder der finalen Beurteilung des Befundes.

3. Publikation

Journal of Cardiovascular Magnetic Resonance 26 (2024) 101068

Contents lists available at ScienceDirect

Journal of Cardiovascular Magnetic Resonance

journal homepage: www.sciencedirect.com/journal/jocmr

Original Research

Generative Pre-trained Transformer 4 analysis of cardiovascular magnetic resonance reports in suspected myocarditis: A multicenter study

Kenan Kaya[a,*,1], Carsten Gietzen[a,1], Robert Hahnfeldt[a], Maher Zoubi[b], Tilman Emrich[e,k,l], Moritz C. Halfmann[e], Malte Maria Sieren[f,g], Yannic Elser[f], Patrick Krumm[c], Jan M. Brendel[c], Konstantin Nikolaou[c], Nina Haag[d], Jan Borggrefe[d], Ricarda von Krüchten[h], Katharina Müller-Peltzer[h], Constantin Ehrengut[i], Timm Denecke[i], Andreas Hagendorff[j], Lukas Goertz[a], Roman J. Gertz[a], Alexander Christian Bunck[a], David Maintz[a], Thorsten Persigehl[a], Simon Lennartz[a], Julian A. Luetkens[b], Astha Jaiswal[a], Andra Iza Iuga[a], Lenhard Pennig[a,2], Jonathan Kottlors[a,2]

[a] Institute for Diagnostic and Interventional Radiology, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany
[b] Institute for Diagnostic and Interventional Radiology, Faculty of Medicine and University Hospital Bonn, University of Bonn, Bonn, Germany
[c] Department of Radiology, Diagnostic and Interventional Radiology, University of Tübingen, Tübingen, Germany
[d] Institute for Radiology, Neuroradiology and Nuclear Medicine Johannes Wesling University Hospital/Mühlenkreiskliniken, Bochum/Minden, Germany
[e] Department of Diagnostic and Interventional Radiology, University Medical Center of the Johannes-Gutenberg-University, Mainz, Germany
[f] Department of Radiology and Nuclear Medicine, UKSH, Campus Lübeck, Lübeck, Germany
[g] Institute of Interventional Radiology, UKSH, Campus Lübeck, Lübeck, Germany
[h] Department of Diagnostic and Interventional Radiology, Medical Center, Faculty of Medicine, University of Freiburg, Freiburg, Germany
[i] Department of Diagnostic and Interventional Radiology, University of Leipzig, Leipzig, Germany
[j] Department of Cardiology, University of Leipzig, Leipzig, Germany
[k] Division of Cardiovascular Imaging, Department of Radiology and Radiological Science, Medical University of South Carolina, Charleston, South Carolina, USA
[l] German Centre for Cardiovascular Research, Partner Site Rhine-Main, Mainz, Germany

ARTICLE INFO

Keywords:
Cardiovascular magnetic resonance
Generative Pre-trained Transformer 4
Artificial intelligence
Large language models
Myocarditis

ABSTRACT

Background: Diagnosing myocarditis relies on multimodal data, including cardiovascular magnetic resonance (CMR), clinical symptoms, and blood values. The correct interpretation and integration of CMR findings require radiological expertise and knowledge. We aimed to investigate the performance of Generative Pre-trained Transformer 4 (GPT-4), a large language model, for report-based medical decision-making in the context of cardiac MRI for suspected myocarditis.

Methods: This retrospective study includes CMR reports from 396 patients with suspected myocarditis and eight centers, respectively. CMR reports and patient data including blood values, age, and further clinical information were provided to GPT-4 and radiologists with 1 (resident 1), 2 (resident 2), and 4 years (resident 3) of experience in CMR and knowledge of the 2018 Lake Louise Criteria. The final impression of the report regarding the radiological assessment of whether myocarditis is present or not was not provided. The performance of Generative pre-trained transformer 4 (GPT-4) and the human readers were compared to a consensus reading (two board-certified radiologists with 8 and 10 years of experience in CMR). Sensitivity, specificity, and accuracy were calculated.

Results: GPT-4 yielded an accuracy of 83%, sensitivity of 90%, and specificity of 78%, which was comparable to the physician with 1 year of experience (R1: 86%, 90%, 84%, p = 0.14) and lower than that of more experienced

Abbreviations: AI, artificial intelligence; CK, creatine kinase; CK-MB, creatine kinase-MB; CRP, C-reactive protein; GPT-4, Generative Pre-trained Transformer 4; Hs-cTn, high-sensitive cardiac troponin; LGE, late gadolinium enhancement; LLC, Lake Louise Criteria; LLM, large language model; LVEF, left ventricular ejection fraction; CMR, cardiovascular magnetic resonance; ROC, receiver operating characteristic; LV, left ventricular; EDV, end-diastolic volume; ESV, end-systolic volume; EF, ejection fraction; CO, cardiac output; CI, cardiac index; HR, heart rate; CT, computed tomography; LV EDD, left ventricular end-diastolic dimension; BSA, body surface area; ED, end-diastole

* Corresponding author. Institute of Diagnostic and Interventional Radiology, University Hospital of Cologne, Kerpener Straße 62, 50937 Cologne, Germany.
E-mail address: kenan.kaya@uk-koeln.de (K. Kaya).

[1] These authors contributed equally as first authors.
[2] These authors contributed equally as senior authors.

https://doi.org/10.1016/j.jocmr.2024.101068
Received 18 April 2024; Received in revised form 4 July 2024; Accepted 24 July 2024

physicians (R2: 89%, 86%, 91%, p = 0.007 and R3: 91%, 85%, 96%, p < 0.001). GPT-4 and human readers showed a higher diagnostic performance when results from T1- and T2-mapping sequences were part of the reports, for residents 1 and 3 with statistical significance (p = 0.004 and p = 0.02, respectively).
Conclusion: GPT-4 yielded good accuracy for diagnosing myocarditis based on CMR reports in a large dataset from multiple centers and therefore holds the potential to serve as a diagnostic decision-supporting tool in this capacity, particularly for less experienced physicians. Further studies are required to explore the full potential and elucidate educational aspects of the integration of large language models in medical decision-making.

1. Introduction

Myocarditis represents an important cause of cardiac morbidity and mortality, leading to up to 20–40% of sudden cardiac deaths in patients younger than 40 years [1,2]. Although an early and accurate diagnosis of myocarditis is mandatory to reduce the risk of progression, the correct diagnosis still poses a challenge in modern cardiology because of the variety of clinical representations and laboratory findings of myocarditis [3].

In this context, cardiac cardiovascular magnetic resonance (CMR) has evolved as a reliable non-invasive diagnostic tool in patients with suspected myocarditis [4]. The diagnosis of myocarditis using CMR requires a high level of radiological expertise and the ability to interpret various image characteristics in different sequences [5]. In 2009, the Lake Louise Criteria (LLC) were introduced for the diagnosis of myocarditis and were supplemented by quantitative mapping techniques in 2018 [6]. According to the revised LLC, diagnosis of myocarditis can be made when two main criteria are met: at least one T1-based criterion (increased myocardial T1 relaxation time, increased extracellular volume fraction, or positive late gadolinium enhancement [LGE]) and at least one T2-based criterion (increased myocardial T2 relaxation time or visual myocardial edema/increased T2 signal intensity ratio) [6]. While proficient cardiovascular imaging experts can make precise diagnoses of myocarditis, inexperienced radiologists exhibit a much lower level of accuracy in interpreting these distinct findings, leading to a higher likelihood of incorrect diagnoses [7].

Several studies have highlighted the feasibility and potential of utilizing artificial intelligence (AI) in medical decision-making, particularly in radiology [8,9]. These studies predominantly concentrate on AI-based processing of visual information [10–13]. However, textual information is the cornerstone for documentation and communication in radiology [14,15]. Recent advances in large language models (LLM) have opened new opportunities for processing such text-based medical information [16–19]. One LLM that has shown remarkable capabilities is the Generative Pre-trained Transformer (GPT-4), developed by OpenAI (San Francisco, California, USA) [20,21]. GPT-4 is a fourth-generation deep learning model able to generate logical and semantically accurate responses to text-based input information and questions [22]. GPT-4 has been trained using a large collection of text data extracted from the World Wide Web and has been optimized for various language-related tasks, such as text completion, translation, and question answering. Experimental studies indicated that the predecessor model GPT-3 showed promising results in medical question-answering tasks, achieving passing scores in medical licensing examinations [23]. The use of LLMs, such as GPT-4, for immediate clinical decision-making based on radiology report texts, could provide several benefits, such as improved diagnostic accuracy and reduced variability in decision-making processes. With the ability to analyze and integrate text-based information, such models could aid in the interpretation of various image characteristics in different sequences as well as clinical information and laboratory results to identify cases of myocarditis.

The aim of this study was, therefore, to investigate the performance of GPT-4 for diagnosing myocarditis using different styles of CMR reports as well as clinical information and blood values from various study centers, and to compare its performance to radiologists with different levels of experience in cardiovascular imaging.

2. Materials and methods

2.1. Ethics

This retrospective study received ethical approval (23-1061-retro) and informed consent was waived due to the retrospective design of the investigation. Beyond the patient's aggregated age and sex, no personal information about the patient was transmitted to the GPT-4 model, especially no patient-identifying information was provided to the AI.

2.2. Data acquisition

Radiology departments of eight tertiary care medical centers were advised to each retrospectively screen their database and randomly select a total of 50 CMR reports of patients who were referred for suspected myocarditis. MRI examinations were performed according to respective in-house protocols for myocarditis. Furthermore, the patient's age, gender, clinical symptoms of the patients, and a board-certified radiology report with a final diagnosis of the examination needed to be available. Centers were advised to provide the patient's age, gender, and clinical symptoms. Additionally, laboratory results were provided by the centers, if available. Laboratory results included C-reactive protein (CRP), creatine kinase (CK), creatine kinase-MB (CK-MB), and high-sensitive cardiac troponin (Hs-cTn).

The following data were retrieved from the reports as baseline characteristics for the cohort: left ventricular ejection fraction (LVEF), LGE pattern (subepicardial, mid-myocardial, subendocardial, transmural, and absent LGE), mapping characteristics, additional image findings, and final diagnosis of cardiomyopathies.

2.3. Data preparation

The final impression of the report was extracted from the texts. Furthermore, reports were not included if significant artifacts or poor image quality was reported hindering the ability to make a certain diagnosis. After assessment regarding inclusion and exclusion criteria by the leading center (1), reports were excluded, if provided report data were insufficient or afflicted with errors such as missing text information.

The radiology report in [blinded for submission] language as well as clinical symptoms, laboratory values (if available), and aggregated patient age and gender were compiled into a text dataset in one Word document (Microsoft Office, Redmond, Washington) per patient.

Furthermore, subgroups were established based on the availability of (a) T1- and T2-mapping sequences, (b) laboratory values, and (c) structured reports (Fig. 1). Laboratory results were defined as available if all of the following were available: CRP, CK, CK-MB, and Hs-cTn (as shown in Graphical Abstract).

2.4. GPT-4

GPT-4 was accessed via OpenAIs (San Francisco, California, USA) web interface platform ChatGPT (https://chat.openai.com/) within the timeframe between March and July 2023 [21]. All text datasets were copied separately to the platform using one chat per text dataset. GPT-4 was prompted with evaluating each dataset using zero-shot prompting (Fig. 2).

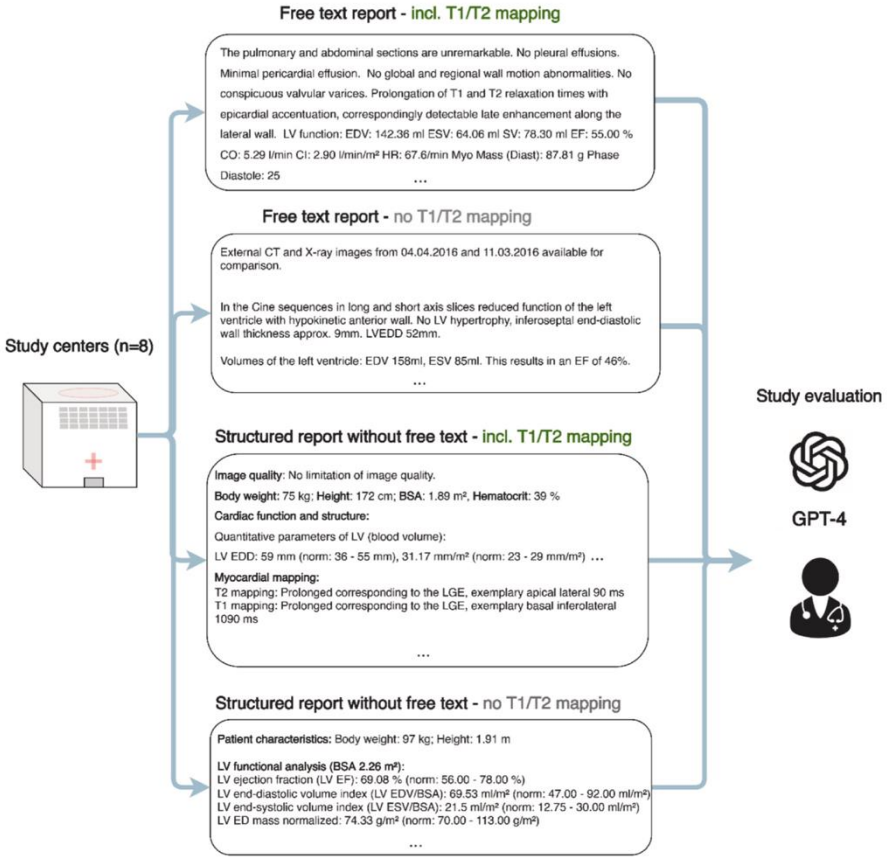

Fig. 1. Exemplary styles of reports being included in this study: free text with T1/T2-mapping, free text without T1/T2-mapping, structured report with T1- and T2-mapping, and structured report without T1/T2-mapping. *LV* left ventricular, *EDV* end-diastolic volume, *ESV* end-systolic volume, *EF* ejection fraction, *CO* cardiac output, *CI* cardiac index, *HR* heart rate, *CT* computed tomography, *LV EDD* left ventricular end-diastolic diameter, *BSA* body surface area, *ED* end-diastole, *LGE* late gadolinium enhancement, *GPT-4* Generative Pre-trained Transformer 4.

2.5. Human reader

To provide a comparison of the performance of GPT-4 to human readers, the datasets were reviewed by three radiology residents with 1 (R.H.; resident 1), 2 (K.K.; resident 2), and 4 years (C.G.; resident 3) of experience in cardiovascular MRI. Knowledge of the 2018 LLC was a necessary precondition to serve as a human reader [6]. The evaluations were conducted independently and without a dedicated time limit (Fig. 2).

2.6. Prompting

The prompt for the human reader and GPT-4 was as follows:
"Please decide on the presence or absence of myocarditis based on the radiological report, provided patient information and clinical parameters. Please respond with either 'yes' or 'no'."

2.7. Reference standard

The reference standard was established by the diagnosis of myocarditis based on the assessment of two board-certified radiologists with 8 (L.P.) and 10 (A.I.) years of experience in CMR who reviewed the above-mentioned data and performed a consensus reading (Fig. 2). Consensus refers to a general agreement among the members of a particular group, each of whom has some level of autonomy in making decisions [24]. All human readers strictly adhered to the 2018 Lake Louise diagnostic criteria for myocarditis [6].

2.8. Statistical analysis

Statistical data analysis was performed using R version 4.4.1 (San Francisco, California, USA). The accuracy, precision, recall (sensitivity), F1 score, and specificity of the performance of GPT-4 and the human readers were calculated by comparing their evaluation to the reference standard and assessed using contingency tables. Dichotomous performance data were compared using McNemar's test or Pearson's chi-squared test. A p-value < 0.05 was considered statistically significant. Figures were plotted using the ggplot2 package (Hadley Wickham, New Zealand). Continuous variables were reported as mean and standard deviation. Demographic characteristics were compared using the chi-squared test for categorical variables and the Mann-Whitney U test for continuous variables.

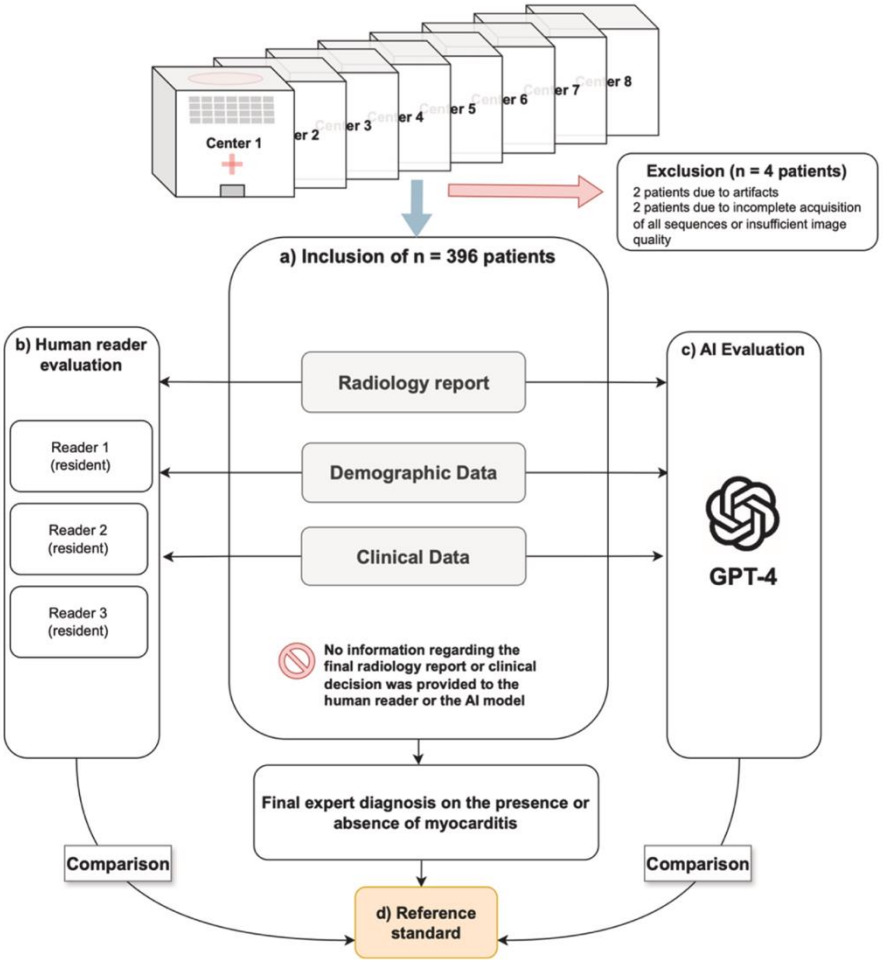

Fig. 2. Workflow of the study design. The reference standard was established by the diagnosis of myocarditis based on the assessment of two board-certified radiologists with 8 and 10 years of experience in cardiovascular imaging, respectively. *GPT-4* Generative Pre-trained Transformer 4.

3. Results

3.1. Baseline characteristics

Of the available 400 reports, 4 patients were excluded due to artifacts or poor image quality and incomplete acquisition of all sequences (Fig. 2). Consequently, 396 patients were included for further analysis. Based on the final assessment in the reports, myocarditis was the most frequent diagnosis with 163 of 396 patients (41.2%), followed by ischemic cardiomyopathy with 23 of 396 patients (5.8%). Table 1 provides detailed results for the final diagnosis of CMR studies from the impressions of the individual reports by the respective centers. Regarding LGE, a subepicardial pattern was the most prevalent localization in 143 of 396 patients (36.1%), followed by mid-myocardial lesions. Table 2 lists the LGE pattern based on the reports. A mean LVEF of 54.9 ± 12.1% was observed.

3.2. Assessment by the expert reader

According to the assessment by the expert readers (consensus), 171 of 396 patients (43.2%) showed myocarditis (group 1), whereas 225 of 396 patients (56.8%) did not (group 2). Patients in group 1 were significantly younger (38.6 ± 17.7 vs 44.4 ± 17.6 years; $p = 0.001$) and predominantly male (76.0% vs 56.9%; $p < 0.001$). Table 3 lists the demographic values for each center.

Table 1
Final diagnosis of CMR based on the final diagnosis of the reports from the respective centers.

Final diagnosis	n = 396 patients
Non-ischemic cardiomyopathies	238/396 (60.1%)
Chemotherapy-induced toxicity	1/396 (0.3%)
Dilated cardiomyopathy	24/396 (6.1%)
Hypertrophic cardiomyopathy	12/396 (3.0%)
Non-compaction cardiomyopathy	1/396 (0.3%)
Myocarditis	163/396 (41.2%)
Pericarditis	23/396 (5.8%)
Sarcoidosis	3/396 (0.8%)
Takotsubo cardiomyopathy	7/396 (1.8%)
Ischemic cardiomyopathy	23/396 (5.8%)
Valvulopathy	6/396 (1.5%)
Uncertain findings	6/396 (1.5%)
No finding	143/396 (36.1%)

CMR cardiovascular magnetic resonance.
Table 1: Data are numbers (%) of 396 cases with their final diagnosis.

Table 2
LGE pattern based on the reports.

LGE localization	n = 396 patients
Subendocardial	24/396 (6.1%)
Mid-myocardial	78/396 (19.7%)
Subepicardial	143/396 (36.1%)
Transmural	35/396 (8.8%)
Absence of LGE	172/396 (43.3%)

LGE late gadolinium enhancement.
Table 2: Data are numbers (%) of 396 cases with their LGE localization in the text reports.

3.3. Performance of GPT-4

Compared to the expert reading, GPT-4 had an accuracy of 83%, specificity of 78%, and sensitivity of 90%. Table 4 provides detailed results for the performance of GPT-4.

The re-test-evaluation within the timeframe between July and August 2023 showed a 100% concordance between the results of the first and second GPT-4 evaluations.

3.4. Performance of the radiologists

Compared to the expert reading, resident 1 showed an accuracy of 86%, resident 2 of 89%, and resident 3 of 91%. The performance of resident 1 was comparable to GPT-4 ($p = 0.14$) whereas the more experienced readers showed superior results ($p = 0.007$ and $p < 0.001$, respectively). Table 4 gives detailed results for the performance of the radiologists. The experienced radiology residents (residents 2 and 3) showed no significant difference in accuracy ($p = 0.22$). Fig. 3 depicts confusion matrices for the performance of GPT-4 and radiologists compared to the reference standard.

Fig. 4 provides two examples of text-based analysis created by GPT-4 and the final assessment of GPT-4 and the human readers.

3.5. Subgroup analysis

3.5.1. Distribution of the subgroups

T1- and T2-mapping was available from 250 of 396 patients (63.1%), laboratory values from 166 of 396 patients (41.9%), and structured reports from 246 of 396 patients (62.1%). The distribution of subgroups is presented in Table 5, as assessed by the expert reader.

3.5.1.1. Subgroup laboratory values. For the subgroup according to laboratory values available vs unavailable, Table 6 gives detailed results. For GPT-4 (accuracy 85% vs 81%), resident 1 (accuracy 84% vs 89%), resident 2 (accuracy 87% vs 91%), and resident 3 (accuracy 89% vs 94%), no statistically significant differences in performance were observed between the subgroups with or without available laboratory values ($p = 0.44$, $p = 0.22$, $p = 0.34$, and $p = 0.13$, respectively).

3.5.1.2. Subgroup T1- and T2-mapping sequences. GPT-4 (accuracy 79% vs 86%) as well as all residents, resident 1 (accuracy 79% vs 90%), resident 2 (accuracy 85% vs 91%), and resident 3 (accuracy 86% vs 94%), had an improved performance when mapping sequences were part of the reports. Residents 1 and 3 showed significant differences in their diagnostic performance regarding the availability of mapping sequences ($p = 0.004$ and $p = 0.02$, respectively). Table 7 summarizes the results of the mapping subgroup analysis.

Table 3
Demographic characteristics of patients with myocarditis and without myocarditis.

		Myocarditis			No myocarditis		
	N	N	Age	Sex	N	Age	Sex
Center 1	50	27	34.1 ± 14.6 (18, 62)	5 F, 22 M	23	42.9 ± 19.0 (18, 73)	12 F, 11 M
Center 2	50	25	42.5 ± 21.6 (16, 83)	8 F, 17 M	25	52 ± 21.3 (9, 85)	14 F, 11 M
Center 3	50	18	51.7 ± 16.7 (19, 75)	6 F, 12 M	32	52 ± 18.2 (26, 86)	7 F, 25 M
Center 4	50	19	42.4 ± 15.2 (24, 71)	4 F, 15 M	31	38.9 ± 12.9 (18, 62)	14 F, 17 M
Center 5	50	21	35 ± 15.8 (18, 77)	7 F, 14 M	29	37.7 ± 11.0 (20, 60)	10 F, 19 M
Center 6	48	16	29.8 ± 16.3 (3, 61)	4 F, 12 M	32	42.3 ± 20.5 (11, 84)	12 F, 20 M
Center 7	48	24	36.8 ± 18.4 (15, 77)	2 F, 22 M	24	42.3 ± 15.7 (16, 77)	13 F, 11 M
Center 8	50	21	37.8 ± 15.5 (20, 80)	5 F, 16 M	29	47.2 ± 16.4 (19, 83)	15 F, 14 M
All centers	396	171	38.6 ± 17.7 (3, 83)	41 F, 130 M	225	44.4 ± 17.6 (9, 86)	97 F, 128 M

N number (ages are reported as means ± standard), *F* female, *M* male.
Values in parentheses are ranges.
Table 3: Data of the respective centers with their means ± standard deviation of age.

Table 4
Performance of GPT-4 and radiology residents with 1 (resident 1), 2 (resident 2), and 4 years (resident 3) of experience compared to the reference standard.

	Accuracy	Precision	Recall (sensitivity)	F1 score	Specificity
GPT-4	0.83 (330/396)	0.76 (154/203)	0.90 (154/171)	0.82 (308/374)	0.78 (176/225)
Resident 1	0.86 (342/396) (p = 0.14)	0.81 (154/191)	0.90 (154/171)	0.85 (308/362)	0.84 (188/225)
Resident 2	0.89 (352/396) **(p = 0.007)**	0.88 (147/167)	0.86 (147/171)	0.87 (294/338)	0.91 (205/225)
Resident 3	0.91 (361/396) **(p < 0.001)**	0.94 (146/156)	0.85 (146/171)	0.89 (292/327)	0.96 (215/225)

GPT-4 Generative Pre-trained Transformer 4, *F1 score* measurement of predictive performance.
The difference in accuracy between GPT-4 and radiology residents is shown as p values, bold indicates statistical significance.

3.5.1.3. Subgroup structured report. For GPT-4 (accuracy 81% vs. 85%), Resident 1 (accuracy 86% vs. 87%), Resident 2 (accuracy 93% vs. 87%), and Resident 3 (accuracy 91% vs. 91%), no significant differences were observed between structured and free-text radiological reports (p = 0.49, p = 0.99, p = 0.09, and p > 0.99, respectively). Please refer to Table 8 for detailed results.

4. Discussion

In this study, AI-assisted diagnosis of myocarditis solely based on CMR reports, laboratory results, and clinical information using GPT-4 was compared to the assessment of radiology residents with different levels of experience. Of note, neither the human readers nor GPT-4 had access to any imaging data. Using a consensus reading of two CMR experts as the reference standard, GPT-4 achieved a sufficient diagnostic performance, which was comparable to a first-year resident. While the availability of laboratory values showed a lower accuracy for GPT-4 based diagnosis, structured reports and available mapping sequences improved its diagnostic performance, albeit without yielding statistical significance.

The integration of diverse data sources has become increasingly important in the medical field. In this context, AI is playing a crucial role in supporting decision-making by assisting in the analysis of complex medical data and improving treatment planning [25–27]. Recently, GPT-4 has been widely recognized for its exceptional proficiency in assessing textual information, representing a major stride forward in natural language processing technology [21]. In a recently published study, GPT-4 succeeded in presenting complex medical findings in a simplified and understandable way for laypersons [28]. In other previous studies, GPT-4 has already shown its potential for applications in radiology. In this context, GPT-4 was able to provide assistance in the radiological workflow by enabling automated determination of radiologic study and protocol based on request forms [15], standardizing radiology reports [29], detecting errors in radiology reports [30], and transforming of free-text reports into structured reporting [31]. Furthermore, GPT-4 is capable of giving diagnostic

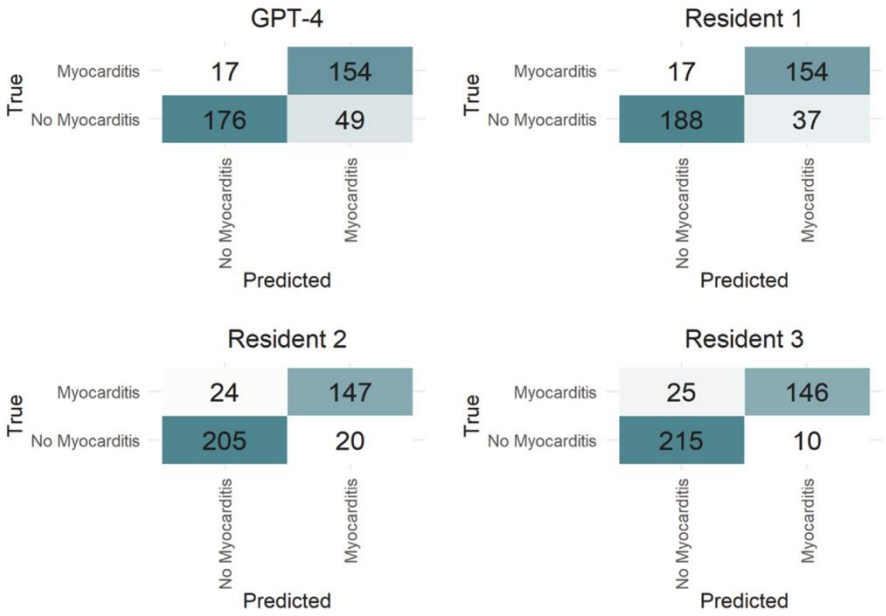

Fig. 3. Confusion matrices for performance of GPT-4, residents 1, 2, and 3 compared to the reference standard. GPT-4: Generative Pre-trained Transformer 4.

Fig. 4. Proofreading examples by GPT-4 based on the given clinical data, laboratory values, and the radiology report compared to the assessment of the human readers. *CRP* C-reactive protein, *CK* creatine kinase, *CK-MB* creatine kinase-MB, *Hs-Trop* high sensitive troponin. *IVSD* interventricular septum thickness. *LV* left ventricular, *EDV* end-diastolic volume, *EF* ejection fraction, *BSA* body surface area, *GPT-4* Generative Pre-trained Transformer 4, *CMR* cardiovascular magnetic resonance .

Table 5
Distribution of the subgroups according to cases with myocarditis, no myocarditis, and all cases regarding the availability (yes = available, no = not available) of mapping, laboratory values, and structured reports.

	Mapping		Laboratory values		Structured reports	
	No	Yes	No	Yes	No	Yes
Myocarditis	58/396 (14.6%)	113/396 (28.5%)	74/396 (18.7%)	97/396 (24.5%)	65/396 (16.4%)	106/396 (26.8%)
No myocarditis	88/396 (22.2%)	137/396 (34.6%)	156/396 (39.4%)	69/396 (17.4%)	85/396 (21.5%)	140/396 (35.4%)
Total	146/396 (36.9%)	250/396 (63.1%)	230/396 (58.1%)	166/396 (41.9%)	150/396 (37.9%)	246/396 (62.1%)

A structured report is a method of clinical documentation in standardized formats.

Table 6
Performance of GPT-4 and radiologists with 1 (resident 1), 2 (resident 2), and 4 years (resident 3) of experience regarding the availability (yes = available, no = not available) of laboratory values.

	Accuracy		Precision		Recall (sensitivity)		F1 score		Specificity	
Laboratory values	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
GPT-4	0.85 (195/230)	0.81 (135/166) (p = 0.44)	0.71 (67/95)	0.81 (87/108)	0.91 (67/74)	0.90 (87/97)	0.79 (134/169)	0.85 (174/205)	0.82 (128/156)	0.70 (48/69)
Resident 1	0.84 (194/230)	0.89 (148/166) (p = 0.22)	0.72 (63/88)	0.88 (91/103)	0.85 (63/74)	0.94 (91/97)	0.78 (126/162)	0.91 (182/200)	0.84 (131/156)	0.83 (57/69)
Resident 2	0.87 (201/230)	0.91 (151/166) (p = 0.34)	0.81 (59/73)	0.94 (88/94)	0.79 (59/74)	0.91 (88/97)	0.80 (118/147)	0.92 (176/191)	0.91 (142/156)	0.91 (63/69)
Resident 3	0.89 (205/230)	0.94 (156/166) (p = 0.13)	0.89 (56/63)	0.97 (90/93)	0.76 (56/74)	0.93 (90/97)	0.82 (112/137)	0.95 (180/190)	0.96 (149/156)	0.96 (66/69)

GPT-4 Generative Pre-trained Transformer 4, *F1 score* measurement of predictive performance.
The difference in accuracy with and without the availability of laboratory values is shown as p values.

Table 7
Performance of GPT-4 and radiologists with 1 (resident 1), 2 (resident 2), and 4 years (resident 3) of experience regarding the availability (yes = available, no = not available) of mapping.

	Accuracy		Precision		Recall (sensitivity)		F1 score		Specificity	
Mapping	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
GPT – 4	0.79 (115/146)	0.86 (215/250) (p = 0.08)	0.67 (53/79)	0.81 (101/124)	0.91 (53/58)	0.89 (101/113)	0.77 (106/137)	0.85 (202/237)	0.70 (62/88)	0.83 (114/137)
Resident 1	0.79 (116/146)	0.90 (226/250) **(p = 0.004)**	0.69 (50/72)	0.87 (104/119)	0.86 (50/58)	0.92 (104/113)	0.77 (100/130)	0.90 (208/232)	0.75 (66/88)	0.89 (122/137)
Resident 2	0.85 (124/146)	0.91 (228/250) (p = 0.08)	0.81 (47/58)	0.92 (100/109)	0.81 (47/58)	0.88 (100/113)	0.81 (94/116)	0.90 (200/222)	0.88 (77/88)	0.93 (128/137)
Resident 3	0.86 (126/146)	0.94 (235/250) **(p = 0.02)**	0.87 (45/52)	0.97 (101/104)	0.78 (45/58)	0.89 (101/113)	0.82 (90/110)	0.93 (202/217)	0.92 (81/88)	0.98 (134/137)

GPT-4 Generative Pre-trained Transformer 4, *F1 score* measurement of predictive performance.
The difference in accuracy with and without the mapping sequences is shown as p values, bold indicates statistical significance.

Table 8
Performance of GPT-4 and radiologists with 1 (resident 1), 2 (resident 2), and 4 years (resident 3) of experience regarding the availability of structured report (yes = structured report, no = free-text report).

	Accuracy		Precision		Recall (sensitivity)		F1 score		Specificity	
Structured report	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
GPT-4	0.81 (122/150)	0.85 (208/246) (p = 0.49)	0.75 (55/73)	0.76 (99/130)	0.85 (55/65)	0.93 (99/106)	0.80 (110/138)	0.84 (198/236)	0.79 (67/85)	0.78 (109/140)
Resident 1	0.86 (129/150)	0.87 (213/246) (p = 0.99)	0.80 (59/74)	0.81 (95/117)	0.91 (59/65)	0.90 (95/106)	0.85 (118/139)	0.85 (190/223)	0.82 (70/85)	0.84 (118/140)
Resident 2	0.93 (139/150)	0.87 (213/246) (p = 0.09)	0.92 (59/64)	0.85 (88/103)	0.91 (59/65)	0.83 (88/106)	0.91 (118/129)	0.84 (176/209)	0.94 (80/85)	0.89 (125/140)
Resident 3	0.91 (137/150)	0.91 (224/246) (p > 0.99)	0.96 (54/56)	0.92 (92/100)	0.83 (54/65)	0.87 (92/106)	0.89 (108/121)	0.89 (184/206)	0.98 (83/85)	0.94 (132/140)

GPT-4 Generative Pre-trained Transformer 4, *F1 score* measurement of predictive performance.
The difference in accuracy with and without structured reports is shown as p values.

support by providing accurate differential diagnosis of imaging patterns [14]. However, its performance to provide a final diagnosis is unknown.

The present study evaluates a new approach that utilizes GPT-4 as a text-processing AI model to aid in the decision-making process for the diagnosis of myocarditis. Furthermore, demographic data and clinical symptoms as well as laboratory values, if available, were provided to GPT-4 to reflect the real-world clinical scenario for the assessment of the presence of myocarditis.

Based on these findings, GPT-4 showed potential as an auxiliary tool for text-based diagnosis of myocarditis for inexperienced readers by yielding comparable accuracy. However, its performance was inferior to experienced readers, who showed a higher diagnostic accuracy. Of note, the availability of T1- and T2-mapping sequences as part of the reports improved the diagnostic performance of GPT-4 and of the human readers, for first- and fourth-year residents with statistical significance. These findings underline the necessity of mapping sequences for myocarditis diagnosis as indicated in previous studies comparing the original and 2018 LLC, which showed a higher diagnostic performance when implementing mapping sequences [32,33]. Interestingly, GPT-4 showed a lower specificity when laboratory values were available potentially due to increased cardiac biomarkers not associated with myocarditis misleading the LLM into a wrong diagnosis, indicating necessary improvement of GPT-4 in the future. Despite not yielding statistical significance, GPT-4 had a higher diagnostic performance regarding the diagnosis of myocarditis when assessing structured reports. These findings emphasize the usefulness of structured reporting in radiology, leading to enhanced communication and facilitating collaboration among physicians [34].

Previous studies investigating the usefulness of GPT-4 in radiology mainly focused on data from a single center [14,15,18,19,31]. However, LLMs tend to show dependency on textual information and the language style of text information reports [35]. Furthermore, as shown in this study, there is a large variance in study protocols for CMR in suspected myocarditis. To this end, we decided to conduct the present study as a multi-center investigation by incorporating data sets from eight different institutions, including different styles of reporting and study protocols. Consequently, the present study gives insight into the real-world application of mapping sequences for suspected myocarditis 5 years after the introduction of the 2018 LLC with a third of examinations still being performed without the acquisition of mapping sequences [6]. Furthermore, despite not including image data, this work highlights the potential of incorporating the clinical setting (symptoms, laboratory results) for the final radiological assessment in terms of AI-supported combined diagnostics.

5. Limitations

The aforementioned strengths of this study are offset by some limitations, mostly related to the AI model itself. AI-based aspects, e.g. GPT-4, are considered language models that merely provide

information but are not capable of critically questioning, understanding, and interpreting facts [18,19]. Another limitation is due to the uncertain sources of the GPT-4s training dataset. This problem can lead to inconsistent and contradictory results. Future research should focus on LLMs with built-in capabilities to transparently disclose the exact sources or guidelines underlying their decision-making processes enabling the verification and critical evaluation of these references. Furthermore, the restricted access of GPT-4, potentially requiring the sharing of sensitive data with third parties, represents an additional limitation of the model. In contrast, competing models, e.g. Large Language Model Meta AI, Meta Platforms, Menlo Park, California, USA (LLaMA) [36], offer hospitals the potential to be applied within their infrastructure, abolishing the necessity to transfer data to external servers. Furthermore, the retrospective design and the binary diagnostic approach have to be regarded as limitations of this study since the latter does not reflect a real-world scenario with radiology reports occasionally including differential diagnosis. The chosen dichotomous yes or no approach regarding the presence of myocarditis most likely led to a selection bias potentially influencing the results. As in every study investigating AI-based diagnostic tasks, the chosen reference standard can be seen as a limitation of the study design. Despite implementing a consensus reading by two experts in cardiovascular imaging and given that the diagnostic criteria for myocarditis in terms of the 2018 LLC are somewhat straightforward, there still was a slight disagreement in 2% of cases between the official reports and the consensus reading, underlining the complexity of correct text-based interpretation. As another possible limitation of the study design, T1- and T2-mapping as well as an LGE may be described in ways that are suggestive of myocarditis, thus biasing GPT-4 and human readers.

In conclusion, this proof-of-concept study indicates the potential use of GPT-4 to assist radiology residents and radiologists inexperienced in cardiovascular imaging in diagnostic tasks, assuming the information in the body of the report is correct. However, future research, improvements, and specifications of LLMs are required to improve diagnostic performance and serve as a daily support or training tool.

Funding

This work was partly funded by NUM 2 (Netzwerk Universitätsmedizin, Berlin, Germany) (FKZ: 01KX2121).

Author contributions

Andreas Hagendorff: Investigation, Data curation. **Lukas Goertz:** Methodology, Data curation. **Roman J. Gertz:** Data curation. **Alexander Christian Bunck:** Project administration. **David Maintz:** Project administration. **Thorsten Persigehl:** Project administration. **Kenan Kaya:** Writing – original draft, Investigation, Conceptualization. **Jan Borggrefe:** Investigation, Data curation. **Ricarda von Krüchten:** Data curation. **Katharina Müller-Peltzer:** Data curation. **Constantin Ehrengut:** Data curation. **Timm Denecke:** Data curation. **Jonathan Kottlors:** Writing – original draft, Investigation, Conceptualization. **Yannic Elser:** Data curation. **Patrick Krumm:** Investigation, Data curation. **Jan M. Brendel:** Data curation. **Konstantin Nikolaou:** Investigation, Data curation. **Nina Haag:** Data curation. **Simon Lennartz:** Visualization, Supervision, Project administration. **Carsten Gietzen:** Writing – original draft, Investigation, Conceptualization. **Julian A. Luetkens:** Supervision, Project administration. **Robert Hahnfeldt:** Data curation. **Astha Jaiswal:** Visualization, Validation, Project administration, Methodology. **Maher Zoubi:** Data curation. **Lenhard Pennig:** Writing – original draft, Investigation, Conceptualization. **Tilman Emrich:** Investigation, Data curation. **Moritz C. Halfmann:** Investigation, Data curation. **Malte Maria Sieren:** Data curation. **Andra Iza Iuga**: Investiagation, Conceptualization.

Declaration of competing interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: David Maintz received speaker's honoraria from Philips Healthcare. Jan Borggrefe received speaker's honoraria from Siemens Healthineers. Simon Lennartz is a member of Editorial Board of Radiology and a Senior Deputy Editor of Radiology in Training. Otherwise, the authors declare no conflicts of interest and had full control over all data, and guarantee correctness.

References

[1] Phillips M, Robinowitz M, Higgins JR, Boran KJ, Reed T, Virmani R. Sudden cardiac death in Air Force recruits. A 20-year review. JAMA 1986;256:2696–9.
[2] Liu PP, Mason JW. Advances in the understanding of myocarditis. Circulation 2001;104:1076–82. https://doi.org/10.1161/hc3401.095198.
[3] Caforio AL, Pankuweit S, Arbustini E, Basso C, Gimeno-Blanes J, Felix SB, et al. Current state of knowledge on aetiology, diagnosis, management, and therapy of myocarditis: a position statement of the European Society of Cardiology Working Group on Myocardial and Pericardial Diseases. Eur Heart J 2013;34:2636–48. https://doi.org/10.1093/eurheartj/eht210.
[4] Friedrich MG, Sechtem U, Schulz-Menger J, Holmvang G, Alakija P, Cooper LT, et al. Cardiovascular magnetic resonance in myocarditis: a JACC White Paper. J Am Coll Cardiol 2009;53:1475–87. https://doi.org/10.1016/j.jacc.2009.02.007.
[5] Kotanidis CP, Bazmpani MA, Haidich AB, Karvounis C, Antoniades C, Karamitsos TD. Diagnostic accuracy of cardiovascular magnetic resonance in acute myocarditis: a systematic review and meta-analysis. JACC Cardiovasc Imaging 2018 Nov;11:1583–90. https://doi.org/10.1016/j.jcmg.2017.12.008.
[6] Ferreira VM, Schulz-Menger J, Holmvang G, Kramer CM, Carbone I, Sechtem U, et al. Cardiovascular magnetic resonance in nonischemic myocardial inflammation. J Am Coll Cardiol 2018;72:3158–76. https://doi.org/10.1016/j.jacc.2018.09.072.
[7] Feisst A, Kuetting DLR, Dabir D, Luetkens J, Homsi R, Schild HH, et al. Influence of observer experience on cardiac magnetic resonance strain measurements using feature tracking and conventional tagging. IJC Heart Vasc 2018;18:46–51. https://doi.org/10.1016/j.ijcha.2018.02.007.
[8] Gore JC. Artificial intelligence in medical imaging. Magn Reson Imaging 2020;68:A1–4. https://doi.org/10.1016/j.mri.2019.12.006.
[9] Kelly BS, Judge C, Bollard SM, Clifford SM, Healy GM, Aziz A, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol 2022;32:7998–8007. https://doi.org/10.1007/s00330-022-08784-6.
[10] Kriza C, Amenta V, Zenié A, Panidis D, Chassaigne H, Urbán P, et al. Artificial intelligence for imaging-based COVID-19 detection: systematic review comparing added value of AI versus human readers. Eur J Radiol 2021;145:110028. https://doi.org/10.1016/j.ejrad.2021.110028.
[11] Matsoukas S, Scaggiante J, Schuldt BR, Smith CJ, Chennareddy S, Kalagara R, et al. Accuracy of artificial intelligence for the detection of intracranial hemorrhage and chronic cerebral microbleeds: a systematic review and pooled analysis. Radiol Med 2022;127:1106–23. https://doi.org/10.1007/s11547-022-01530-4.
[12] Soffer S, Klang E, Shimon O, Barash Y, Cahan N, Greenspana H, et al. Deep learning for pulmonary embolism detection on computed tomography pulmonary angiogram: a systematic review and meta-analysis. Sci Rep 2021;11:15814. https://doi.org/10.1038/s41598-021-95249-3.
[13] Zheng Q, Yang L, Zeng B, Li J, Guo K, Liang Y, et al. Artificial intelligence performance in detecting tumor metastasis from medical radiology imaging: a systematic review and meta-analysis. EClinicalMedicine 2021;31:100669. https://doi.org/10.1016/j.eclinm.2020.100669.
[14] Kottlors J, Bratke G, Rauen P, Kabbasch C, Persigehl T, Schlamann M, et al. Feasibility of differential diagnosis based on imaging patterns using a large language model. Radiology 2023;308. https://doi.org/10.1148/radiol.231167.
[15] Gertz RJ, Bunck AC, Lennartz S, Dratsch T, Iuga AI, Maintz D, et al. GPT-4 for automated determination of radiologic study and protocol based on radiology request forms: a feasibility study. Radiology 2023;307:1–3. https://doi.org/10.1148/radiol.230877.
[16] Mañas-García A, González-Valverde I, Camacho-Ramos E, Alberich-Bayarri A, Maldonado JA, Marcos M, et al. Radiological structured report integrated with quantitative imaging biomarkers and qualitative scoring systems. J Digit Imaging 2022;35:396–407. https://doi.org/10.1007/s10278-022-00589-9.
[17] Cornacchia S, Errico R, Balzano RF, Fusco V, Maldera A, Pierpaoli E, et al. Medical radiological procedures: which information would be chosen for the report? Radiol Med 2019;124:783–93. https://doi.org/10.1007/s11547-019-01032-w.
[18] The Lancet Digital Health. ChatGPT: friend or foe? Lancet Digit Health 2023;5:e102. https://doi.org/10.1016/S2589-7500(23)00023-7.
[19] Biswas S. ChatGPT and the Future of Medical Writing. Radiology 2023;307:e223312. https://doi.org/10.1148/radiol.223312.
[20] Petroski Such F, Madhavan V, Liu R, Wang R, Pablo Samuel C, Li Y, et al. An Atari model zoo for analyzing, visualizing, and comparing deep reinforcement learning agents. Neural Evol Comput 2019:1–6. https://doi.org/10.48550/arXiv.1812.07069.
[21] OpenAI, Achiam J., Adler S., Agarwal S., Ahmad L., Akkaya I., et al. GPT-4 Technical Report; 2023: 1-100. doi.org/10.48550/arXiv.2303.08774.. OpenAI. GPT-4 Technical Report; 2023.

[22] Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. Minds Mach (Dordr) 2020;30:681–94. https://doi.org/10.1007/s11023-020-09548-1.
[23] Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312. https://doi.org/10.2196/45312.
[24] Bankier AA, Levine D, Halpern EF, Kressel HY. Consensus interpretation in imaging research: is there a better way? Radiology 2010;257:14–7. https://doi.org/10.1148/radiol.10100252.
[25] van Leeuwen KG, de Rooij M, Schalekamp S, van Ginneken B, Rutten MJCM. How does artificial intelligence in radiology improve efficiency and health outcomes? Pediatr Radiol 2022:2087–93. https://doi.org/10.1007/s00247-021-05114-8.
[26] Brady AP, Neri E. Artificial intelligence in radiology—ethical considerations. Diagnostics 2020;10:231. https://doi.org/10.3390/diagnostics10040231.
[27] van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol 2021;31:3797–804. https://doi.org/10.1007/s00330-021-07892-z.
[28] Salam B, Kravchenko D, Nowak S, Sprinkart AM, Weinhold L, Odenthal A, et al. Generative Pre-trained Transformer 4 makes cardiovascular magnetic resonance reports easy to understand. J Cardiovasc Magn Reson 2024;26:101035. https://doi.org/10.1016/J.JOCMR.2024.101035.
[29] Hasani AM, Singh S, Zahergivar A, Ryan B, Nethala D, Bravomontenegro G, et al. Evaluating the performance of Generative Pre-trained Transformer-4 (GPT-4) in standardizing radiology reports. Eur Radiol 2024 Jun;34(6):3566–74. https://doi.org/10.1007/s00330-023-10384-x.
[30] Gertz RJ, Dratsch T, Bunck AC, Iuga AI, Hellmich MG, Persigehl T, et al. Potential of GPT-4 for detecting errors in radiology reports: implications for reporting accuracy. Radiology 2024;311(1):e232714. https://doi.org/10.1148/radiol.232714.
[31] Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. Radiology 2023;307:e230725. https://doi.org/10.1148/radiol.230725.
[32] Cundari G, Galea N, De Rubeis G, Frustaci A, Cilia F, Mancuso G, et al. Use of the new Lake Louise Criteria improves CMR detection of atypical forms of acute myocarditis. Int J Cardiovasc Imaging 2021;37:1395–404. https://doi.org/10.1007/s10554-020-02097-9.
[33] Luetkens JA, Faron A, Isaak A, Dabir D, Kuetting D, Feisst A, et al. Comparison of original and 2018 Lake Louise Criteria for diagnosis of acute myocarditis: results of a validation cohort. Radiol Cardiothorac Imaging 2019;1:e190010. https://doi.org/10.1148/ryct.2019190010.
[34] Larson DB, Towbin AJ, Pryor RM, Donnelly LF. Improving consistency in radiology reporting through the use of department-wide standardized structured reporting. Radiology 2013;267:240–50. https://doi.org/10.1148/radiol.12121502.
[35] Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med 2023;3:141. https://doi.org/10.1038/s43856-023-00370-1.
[36] Touvron H., Lavril T., Izacard G., Martinet X., Lachaux M.A., Lacroix T., et al. LLaMA: open and efficient foundation language models; 2023:1-27. doi.org/10.48550/arXiv.2302.13971.

4. Diskussion

4.1. Zusammenfassung der Ergebnisse

In dieser Studie wurde die Leistungsfähigkeit von GPT-4 für die text-basierte Diagnose von Myokarditis basierend auf Befundberichten der CMR, der klinischen Symptome und Blutwerte, falls vorhanden, untersucht und mit der Beurteilung radiologischer Assistenzärzte mit unterschiedlicher Erfahrung verglichen. Unter Berücksichtigung eines Konsensusreadings von zwei Experten für kardiovaskuläre Bildgebung als Referenzstandard erzielte GPT-4 eine diagnostische Leistung (Genauigkeit 83%, Sensitivität 90%, Spezifität 78%), die mit der eines Assistenzarztes im ersten Jahr vergleichbar (86%, 90%, 84%, p=0,14), aber niedriger als das der erfahrenen Ärzte (89%, 86%, 91%, p=0,007 bzw. 91%, 85%, 96%, p<0,001) war. Während die Verfügbarkeit von Laborwerten die Genauigkeit der GPT-4 basierten Diagnose verringerte, verbesserten strukturierte radiologische Befundtexte und verfügbare Mapping-Sequenzen die diagnostische Leistung, obwohl dies statistisch nicht signifikant war.

Um die Einflussfaktoren auf die diagnostische Leistung des LLMs und der Radiologen zu untersuchen, wurden Subgruppen basierend auf der Verfügbarkeit von (a) T1- und T2-Mapping-Sequenzen im Befundtext, (b) Laborwerten und (c) strukturierten Befundberichten gebildet. Laborergebnisse wurden als verfügbar definiert, wenn alle folgenden Werte verfügbar waren: CRP, CK, CK-MB und Hs-cTn. Erwähnenswert ist in diesem Kontext, dass die Verfügbarkeit von T1- und T2-Mapping-Sequenzen im Befundtext die diagnostische Leistung von GPT-4 und der Assistenzärzte im ersten und vierten Jahr verbesserte, für die Radiologen mit statistischer Signifikanz (p=0,004 bzw. p=0,02). Diese Ergebnisse heben die Bedeutung von Mapping-Sequenzen für die Myokarditis-Diagnose hervor. Frühere Studien, welche die ursprünglichen Lake-Louise-Kriterien, mit denen von den aktualisierten Kriterien von 2018 verglichen, zeigten ebenfalls eine höhere diagnostische Leistung der CMR, wenn Mapping-Sequenzen verwendet wurden[30,103]. Die vorliegende Studie bietet somit Einblicke in die praktische Anwendung von Mapping-Sequenzen bei Verdacht auf eine Myokarditis, sechs Jahre nach Einführung der Lake-Louise-Kriterien von 2018. Dabei wurde deutlich, dass ein Drittel der Untersuchungen an deutschen Universitätsklinika nach wie vor ohne die Erfassung von T1- und T2-Mapping-Sequenzen durchgeführt wird[8].

Interessanterweise zeigte GPT-4 eine geringere Spezifität, wenn Laborparameter verfügbar waren (42% der Patienten). Diese Beobachtungen sind primär darauf zurückzuführen, dass die erhöhten kardialen Biomarker, die nicht zwingend mit einer Myokarditis in Verbindung stehen müssen, sondern vielmehr aufgrund anderer kardialer Pathologien erhöht sind, das Modell zu einer falsch-positiven Diagnose führten. Dies deutet darauf hin, dass GPT-4 in Zukunft vor allem hinsichtlich der Spezifität weiter verbessert werden muss.

Die Relevanz der strukturierten Befunderstellung in der Radiologie ist weithin anerkannt. Zahlreiche Studien heben hervor, wie sie die Kommunikation verbessert, die Zusammenarbeit zwischen Gesundheitsfachkräften fördert und die Vereinheitlichung der Befundsprache über verschiedene Institutionen hinweg unterstützt[94,104]. So konnte in einer Studie von Dimarco et al. gezeigt werden, dass strukturierte CT-Befunde bei Patienten mit duktalem Adenokarzinom des Pankreas signifikant die Anzahl fehlender morphologischer Merkmale des Tumors und dessen Gefäßbeteiligung im Gegensatz zu Freitextbefunden reduzierte und zudem die Übereinstimmung zwischen den menschlichen Lesern verbesserte[105]. In einer weiteren Studie von Wetterauer et. al konnten erfahrene Urologen mittels strukturierter MRT-Befunde bei Patienten mit Prostatakarzinom die Lokalisation einzelner malignitätssuspekter Läsionen genauer beurteilen, was die chirurgische Planung erleichtert. Somit konnte hier eine erhöhte Zufriedenheit der überweisenden Ärzte durch die strukturierte Befunderstellung erzielt werden[106]. Obwohl keine statistische Signifikanz erreicht wurde, zeigte GPT-4 eine bessere diagnostische Leistung bei der Diagnose von einer Myokarditis, wenn strukturierte Befundtexte verwendet wurden (62% der Patienten). Diese Ergebnisse unterstreichen die Vorteile einer strukturierten Befundung, da sie wie oben beschrieben die interdisziplinäre Zusammenarbeit fördern[104].

Frühere Studien zur Anwendung von GPT-4 in der Radiologie basierten überwiegend auf Daten aus nur einem einzigen Zentrum[14,94,95]. LLMs tendieren jedoch dazu, stark von der Textstruktur und dem Sprachstil der Berichte abhängig zu sein[107]. Um die generelle Anwendbarkeit von GPT-4 zu untersuchen, haben wir uns daher entschieden die vorliegende Studie als multizentrische Studie durchzuführen und Datensätze von acht verschiedenen Standorten einzubeziehen, um somit ein realistisches Szenario zu schaffen. Dies wird durch die unterschiedlichen Befundstile und CMR-Protokolle des Studienkollektivs deutlich.

4.2. Limitationen

Die aufgeführten Stärken dieser Studie werden durch einige Einschränkungen abgeschwächt, die hauptsächlich mit dem KI-Modell selbst zusammenhängen. LLMs wie GPT-4 können zwar Informationen bereitstellen, sind jedoch nicht in der Lage, Fakten kritisch zu hinterfragen, zu verstehen oder zu interpretieren[108,109]. Eine weitere Einschränkung ergibt sich aus den unsicheren Quellen des Trainingsdatensatzes von GPT-4, was zu inkonsistenten und widersprüchlichen Ergebnissen führen kann. Zukünftige Forschung sollte sich auf LLMs konzentrieren, die über eingebaute Mechanismen verfügen, um die genauen Quellen oder Leitlinien, die ihren Entscheidungsprozessen zugrunde liegen, transparent offenzulegen. Dies

würde die Überprüfung und kritische Bewertung dieser Referenzen ermöglichen[16]. Zudem stellt der eingeschränkte Zugang zu GPT-4, der möglicherweise die Weitergabe sensibler Daten an Dritte erfordert, eine zusätzliche Hürde dar. Darüber hinaus müssen der retrospektive Charakter und der binäre diagnostische Ansatz dieser Studie als Einschränkungen betrachtet werden, da letzterer nicht die Realität widerspiegelt, in der radiologische Befundtexte häufig Differentialdiagnosen beinhalten. Der gewählte dichotome Ja- oder Nein-Ansatz hinsichtlich des Vorliegens einer Myokarditis hat wahrscheinlich zu einem Selektionsbias geführt, der die Ergebnisse potenziell beeinflusst hat. Wie in jeder Studie, die KI-basierte diagnostische Aufgaben untersucht, kann auch der gewählte Referenzstandard als Schwäche des Studiendesigns betrachtet werden. Trotz der Konsensusbewertung durch zwei Experten für kardiovaskuläre Bildgebung mit mehrjähriger Erfahrung und der relativ klaren diagnostischen Kriterien für die Diagnose einer Myokarditis gemäß den Lake-Louise-Kriterien von 2018, gab es in 2% der Fälle Unstimmigkeiten zwischen den offiziellen Befundberichten der Universitätsklinika und der Konsensusbewertung. Dies unterstreicht die Komplexität der korrekten textbasierten Interpretation. Eine weitere mögliche Einschränkung des Studiendesigns ist, dass T1- und T2-Mapping sowie LGE-Sequenzen möglicherweise auf eine Weise beschrieben werden, die auf eine Myokarditis hindeutet (z.B. subepikardiales Verteilungsmuster), was zu einem Bias bei GPT-4 und den menschlichen Lesern führen könnte.

4.3. Ausblick

Moderne LLMs wachsen stetig in Größe und Komplexität, und es wird erwartet, dass ihre Leistungsfähigkeit weiter zunimmt. Proprietäre Modelle wie das Gemini Ultra von Google (Alphabet, Mountain View, Kalifornien, USA) und das kommende GPT-5 von OpenAI werden voraussichtlich die bisherigen Grenzen erweitern. Allerdings bringt die fehlende Transparenz von Closed-Source-Modellen und damit nicht für die Öffentlichkeit einsehbaren Programmiercodes und Verschlüsselungen erhebliche Herausforderungen für die Implementierung in der Medizin mit sich[110].

Hinsichtlich der weiteren Anwendungsbereiche der LLMs in der Radiologie ist zu konstatieren, dass derzeit Radiologen, um zusätzliche Informationen zur medizinischen Vorgeschichte oder zu Laborergebnissen der Patienten zu erhalten, sich separat in die elektronische Patientenakte im RIS und in das Bildarchivierungssystem (Picture Archiving and Communication System, PACS) einloggen müssen. Da die meisten im PACS enthaltenden Bildgebungsaufträge knappgehalten sind und keine umfassende Zusammenfassung der Krankengeschichte enthalten, muss der Radiologe oft zwischen den Systemen hin und her wechseln, was sehr

zeitaufwändig sein kann. Dieser Prozess könnte durch LLMs unterstützt oder sogar vollständig übernommen werden, indem automatisch eine Zusammenfassung der Krankengeschichte und Befunde des Patienten bereitgestellt wird[111]. Eine weitere potenzielle Anwendung von LLMs ist ihre Nutzung als fortschrittliche klinische Entscheidungshilfesysteme. Diese Modelle sollten mit Richtlinien und Empfehlungen der einzelnen Fachgesellschaften gezielt angepasst werden, um automatisch evidenzbasierte Empfehlungen aus Radiologieberichten zu generieren, zum Beispiel Nachsorgeempfehlungen gemäß der Fleischner Society für solide Lungenrundherde[112].

LLMs könnten auch eine wichtige Rolle bei der Ausbildung der nächsten Generation von Radiologen spielen[113]. Da die Ausbildung derzeit durch hohe Arbeitsbelastung beeinträchtigt werden kann, könnten LLMs durch eine Integration ins PACS eine personalisierte, interaktive und effektive Lernumgebung schaffen. Sie könnten ähnliche Fälle aus den Archiven bereitstellen, die mit dem Fall übereinstimmen, an dem der Auszubildende arbeitet, zusätzliche Diagnose-Ressourcen empfehlen oder ein realistisches klinisches Szenario vollständig simulieren, um die Auszubildenden auf Nachtschichten vorzubereiten[76].

Trotz des offensichtlichen Potenzials von LLMs in der Radiologie gibt es eine Reihe von Einschränkungen und Herausforderungen, die im Verlauf der weiteren Forschung berücksichtigt werden müssen. Eine der größten Herausforderungen bei der Anwendung von LLMs in der Medizin ist weiterhin der Datenschutz[114]. Zudem ist die zügige Einführung dieser Werkzeuge in der Radiologie ist aufgrund fehlender Vorschriften und ethischer Unklarheiten noch unsicher. Voraussichtlich werden diese Werkzeuge in der EU und den USA ähnlich wie andere klinische Entscheidungsunterstützungssysteme in der Zukunft reguliert werden[115].

4.4. Fazit

Zusammenfassend verdeutlicht diese Machbarkeitsstudie das Potenzial von GPT-4 Radiologen mit wenig Erfahrung in der kardiovaskulären Bildgebung bei der text-basierten Diagnose der Myokarditis zu unterstützen, vorausgesetzt, dass die Informationen in den vorliegenden Befundberichten korrekt sind. Zukünftige Forschung, Verbesserungen und Spezifikationen von LLMs sind jedoch erforderlich, um die diagnostische Leistung zu verbessern um als tägliches Unterstützungs- und Trainingswerkzeug zu dienen.

5. Literaturverzeichnis

1 Virani SS, Alonso A, Aparicio HJ, *et al.* Heart Disease and Stroke Statistics—2021 Update. *Circulation* 2021; **143**. DOI:10.1161/CIR.0000000000000950.

2 Liu PP, Mason JW. Advances in the Understanding of Myocarditis. *Circulation* 2001; **104**: 1076–82.

3 Phillips M, Robinowitz M, Higgins JR, Boran KJ, Reed T, Virmani R. Sudden Cardiac Death in Air Force Recruits: A 20-Year Review. *JAMA* 1986; **256**: 2696–9.

4 Dai H, Lotan D, Much AA, *et al.* Global, Regional, and National Burden of Myocarditis and Cardiomyopathy, 1990–2017. *Front Cardiovasc Med* 2021; **8**. DOI:10.3389/fcvm.2021.610989.

5 Kindermann I, Barth C, Mahfoud F, *et al.* Update on Myocarditis. *J Am Coll Cardiol* 2012; **59**: 779–92.

6 Fung G, Luo H, Qiu Y, Yang D, McManus B. Myocarditis. *Circ Res* 2016; **118**: 496–514.

7 Caforio ALP, Pankuweit S, Arbustini E, *et al.* Current state of knowledge on aetiology, diagnosis, management, and therapy of myocarditis: a position statement of the European Society of Cardiology Working Group on Myocardial and Pericardial Diseases. *Eur Heart J* 2013; **34**: 2636–48.

8 Ferreira VM, Schulz-Menger J, Holmvang G, *et al.* Cardiovascular Magnetic Resonance in Nonischemic Myocardial Inflammation. *J Am Coll Cardiol* 2018; **72**: 3158–76.

9 Kotanidis CP, Bazmpani M-A, Haidich A-B, Karvounis C, Antoniades C, Karamitsos TD. Diagnostic Accuracy of Cardiovascular Magnetic Resonance in Acute Myocarditis. *JACC Cardiovasc Imaging* 2018; **11**: 1583–90.

10 Feisst A, Kuetting DLR, Dabir D, *et al.* Influence of observer experience on cardiac magnetic resonance strain measurements using feature tracking and conventional tagging. *IJC Heart & Vasculature* 2018; **18**: 46–51.

11 Phillips M, Marsden H, Jaffe W, *et al.* Assessment of Accuracy of an Artificial Intelligence Algorithm to Detect Melanoma in Images of Skin Lesions. *JAMA Netw Open* 2019; **2**: e1913436.

12 Kelly BS, Judge C, Bollard SM, *et al.* Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). *Eur Radiol* 2022; **32**: 7998–8007.

13 Saha A, Bosma JS, Twilt JJ, *et al.* Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study. *Lancet Oncol* 2024; **25**: 879–87.

14 Gertz RJ, Bunck AC, Lennartz S, *et al.* GPT-4 for Automated Determination of Radiologic Study and Protocol Based on Radiology Request Forms: A Feasibility Study. *Radiology* 2023; **307**. DOI:10.1148/radiol.230877.

15 Laukamp KR, Terzis RA, Werner J-M, *et al.* Monitoring Patients with Glioblastoma by Using a Large Language Model: Accurate Summarization of Radiology Reports with GPT-4. *Radiology* 2024; **312**. DOI:10.1148/radiol.232640.

16 Bhayana R. Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. *Radiology* 2024; **310**. DOI:10.1148/radiol.232756.

17 Maximilian Reiser, Fritz-Peter Kuhn, Jürgen Debus. Duale Reihe Radiologie, 4. Auflage. Thieme, 2017.

18 Pooley RA. Fundamental Physics of MR Imaging. *RadioGraphics* 2005; **25**: 1087–99.

19 Gallagher TA, Nemeth AJ, Hacein-Bey L. An Introduction to the Fourier Transform: Relationship to MRI. *American Journal of Roentgenology* 2008; **190**: 1396–405.

20 Elster AD. Gradient-echo MR imaging: techniques and acronyms. *Radiology* 1993; **186**: 1–8.

21 Jung BA, Weigel M. Spin echo magnetic resonance imaging. *Journal of Magnetic Resonance Imaging* 2013; **37**: 805–17.

22 Bitar R, Leung G, Perng R, *et al.* MR Pulse Sequences: What Every Radiologist Wants to Know but Is Afraid to Ask. *RadioGraphics* 2006; **26**: 513–37.

23 Mitchell C, Rahko PS, Blauwet LA, *et al.* Guidelines for Performing a Comprehensive Transthoracic Echocardiographic Examination in Adults: Recommendations from the American Society of Echocardiography. *Journal of the American Society of Echocardiography* 2019; **32**: 1–64.

24 Pontone G, Rossi A, Guglielmo M, *et al.* Clinical applications of cardiac computed tomography: a consensus paper of the European Association of Cardiovascular Imaging—part I. *Eur Heart J Cardiovasc Imaging* 2022; **23**: 299–314.

25 Panda A, Homb AC, Krumm P, *et al.* Cardiac Nuclear Medicine: Techniques, Applications, and Imaging Findings. *RadioGraphics* 2023; **43**. DOI:10.1148/rg.220027.

26 Neumann F-J, Sousa-Uva M, Ahlsson A, *et al.* 2018 ESC/EACTS Guidelines on myocardial revascularization. *Eur Heart J* 2019; **40**: 87–165.

27 Hundley WG, Bluemke DA, Bogaert J, *et al.* Society for Cardiovascular Magnetic Resonance (SCMR) guidelines for reporting cardiovascular magnetic resonance examinations. *Journal of Cardiovascular Magnetic Resonance* 2022; **24**: 29.

28 Zugwitz D, Fung K, Aung N, *et al.* Mitral Annular Disjunction Assessed Using CMR Imaging. *JACC Cardiovasc Imaging* 2022; **15**: 1856–66.

29 Ioannou A, Cappelli F, Emdin M, *et al.* Stratifying Disease Progression in Patients With Cardiac ATTR Amyloidosis. *J Am Coll Cardiol* 2024; **83**: 1276–91.

30 Luetkens JA, Faron A, Isaak A, *et al.* Comparison of Original and 2018 Lake Louise Criteria for Diagnosis of Acute Myocarditis: Results of a Validation Cohort. *Radiol Cardiothorac Imaging* 2019; **1**: e190010.

31 Kramer CM, Barkhausen J, Bucciarelli-Ducci C, Flamm SD, Kim RJ, Nagel E. Standardized cardiovascular magnetic resonance imaging (CMR) protocols: 2020 update. *Journal of Cardiovascular Magnetic Resonance* 2020; **22**: 17.

32 Ridgway JP. Cardiovascular magnetic resonance physics for clinicians: part I. *Journal of Cardiovascular Magnetic Resonance* 2010; **12**: 71.

33 Bieri O, Scheffler K. Fundamentals of balanced steady state free precession MRI. *Journal of Magnetic Resonance Imaging* 2013; **38**: 2–11.

34 Benjelloun H, Cranney GB, Kirk KA, Blackwell GG, Lotan CS, Pohost GM. Interstudy reproducibility of biplane cine nuclear magnetic resonance measurements of left ventricular function. *Am J Cardiol* 1991; **67**: 1413–20.

35 Schulz-Menger J, Bluemke DA, Bremerich J, *et al.* Standardized image interpretation and post processing in cardiovascular magnetic resonance: Society for Cardiovascular Magnetic Resonance (SCMR) Board of Trustees Task Force on Standardized Post Processing. *Journal of Cardiovascular Magnetic Resonance* 2013; **15**: 35.

36 Francone M, Carbone I, Agati L, *et al.* Utility of T2-weighted short-tau inversion recovery (STIR) sequences in cardiac MRI: an overview of clinical applications in ischaemic and non-ischaemic heart disease. *Radiol Med* 2011; **116**: 32–46.

37 Friedrich MG, Sechtem U, Schulz-Menger J, *et al.* Cardiovascular Magnetic Resonance in Myocarditis: A JACC White Paper. *J Am Coll Cardiol* 2009; **53**: 1475–87.

38 Galea N, Francone M, Fiorelli A, *et al.* Early myocardial gadolinium enhancement in patients with myocarditis: Validation of “Lake Louise consensus” criteria using a single bolus of 0.1 mmol/Kg of a high relaxivity gadolinium-based contrast agent. *Eur J Radiol* 2017; **95**: 89–95.

39 Kellman P, Arai AE. Cardiac imaging techniques for physicians: Late enhancement. *Journal of Magnetic Resonance Imaging* 2012; **36**: 529–42.

40 Ordovas KG, Higgins CB. Delayed Contrast Enhancement on MR Images of Myocardium: Past, Present, Future. *Radiology* 2011; **261**: 358–74.

41 Vermes E, Carbone I, Friedrich MG, Merchant N. Patterns of myocardial late enhancement: Typical and atypical features. *Arch Cardiovasc Dis* 2012; **105**: 300–8.

42 Kuruvilla S, Adenaw N, Katwal AB, Lipinski MJ, Kramer CM, Salerno M. Late Gadolinium Enhancement on Cardiac Magnetic Resonance Predicts Adverse Cardiovascular Outcomes in Nonischemic Cardiomyopathy. *Circ Cardiovasc Imaging* 2014; **7**: 250–8.

43 Wong TC, Piehler K, Puntil KS, *et al.* Effectiveness of late gadolinium enhancement to improve outcomes prediction in patients referred for cardiovascular magnetic resonance after echocardiography. *Journal of Cardiovascular Magnetic Resonance* 2013; **15**: 6.

44 Messroghli DR, Moon JC, Ferreira VM, *et al.* Clinical recommendations for cardiovascular magnetic resonance mapping of T1, T2, T2* and extracellular volume: A consensus statement by the Society for Cardiovascular Magnetic Resonance (SCMR) endorsed by the European Association for Cardiovascular Imaging (EACVI). *Journal of Cardiovascular Magnetic Resonance* 2016; **19**: 75.

45 Messroghli DR, Radjenovic A, Kozerke S, Higgins DM, Sivananthan MU, Ridgway JP. Modified Look-Locker inversion recovery (MOLLI) for high-resolution T_1 mapping of the heart. *Magn Reson Med* 2004; **52**: 141–6.

46 Pennig L, Luetkens J, Nähle CP. Technik und klinische Bedeutung des kardialen Mappings – was der Radiologe wissen sollte. *Radiologie up2date* 2021; **21**: 135–52.

47 Sprinkart AM, Luetkens JA, Träber F, *et al.* Gradient Spin Echo (GraSE) imaging for fast myocardial T2 mapping. *Journal of Cardiovascular Magnetic Resonance* 2015; **17**: 12.

48 Luetkens JA, Homsi R, Sprinkart AM, *et al.* Incremental value of quantitative CMR including parametric mapping for the diagnosis of acute myocarditis. *Eur Heart J Cardiovasc Imaging* 2016; **17**: 154–61.

49 Ferreira VM, Piechnik SK, Dall'Armellina E, *et al.* Non-contrast T1-mapping detects acute myocardial edema with high diagnostic accuracy: a comparison to T2-weighted cardiovascular magnetic resonance. *Journal of Cardiovascular Magnetic Resonance* 2012; **14**: 53.

50 Leone O, Veinot JP, Angelini A, *et al.* 2011 Consensus statement on endomyocardial biopsy from the Association for European Cardiovascular Pathology and the Society for Cardiovascular Pathology. *Cardiovascular Pathology* 2012; **21**: 245–74.

51 Friedrich MG, Strohm O, Schulz-Menger J, Marciniak H, Luft FC, Dietz R. Contrast Media–Enhanced Magnetic Resonance Imaging Visualizes Myocardial Changes in the Course of Viral Myocarditis. *Circulation* 1998; **97**: 1802–9.

52 Kawai C. From Myocarditis to Cardiomyopathy: Mechanisms of Inflammation and Cell Death. *Circulation* 1999; **99**: 1091–100.

53 Mewton N, Liu CY, Croisille P, Bluemke D, Lima JAC. Assessment of Myocardial Fibrosis With Cardiovascular Magnetic Resonance. *J Am Coll Cardiol* 2011; **57**: 891–903.

54 Blauwet LA, Cooper LT. Myocarditis. *Prog Cardiovasc Dis* 2010; **52**: 274–88.

55 Kühl U, Pauschinger M, Schwimmbeck PL, *et al.* Interferon-β Treatment Eliminates Cardiotropic Viruses and Improves Left Ventricular Function in Patients With Myocardial Persistence of Viral Genomes and Left Ventricular Dysfunction. *Circulation* 2003; **107**: 2793–8.

56 Schultz JC, Hilliard AA, Cooper LT, Rihal CS. Diagnosis and Treatment of Viral Myocarditis. *Mayo Clin Proc* 2009; **84**: 1001–9.

57 Kontorovich AR, Patel N, Moscati A, *et al.* Myopathic Cardiac Genotypes Increase Risk for Myocarditis. *JACC Basic Transl Sci* 2021; **6**: 584–92.

58 Lota AS, Hazebroek MR, Theotokis P, *et al.* Genetic Architecture of Acute Myocarditis and the Overlap With Inherited Cardiomyopathy. *Circulation* 2022; **146**: 1123–34.

59 Ammirati E, Moslehi JJ. Diagnosis and Treatment of Acute Myocarditis: A Review. *JAMA* 2023; **329**: 1098–113.

60 Ammirati E, Frigerio M, Adler ED, *et al.* Management of Acute Myocarditis and Chronic Inflammatory Cardiomyopathy: An Expert Consensus Document. *Circ Heart Fail* 2020; **13**: e007405.

61 JCS Joint Working Group. Guidelines for Diagnosis and Treatment of Myocarditis (JCS 2009). *Circulation Journal* 2011; **75**: 734–43.

62 Prepoudis A, Koechlin L, Nestelberger T, *et al.* Incidence, clinical presentation, management, and outcome of acute pericarditis and myopericarditis. *Eur Heart J Acute Cardiovasc Care* 2022; **11**: 137–47.

63 Sagar S, Liu PP, Cooper LT. Myocarditis. *The Lancet* 2012; **379**: 738–47.

64 Lauer B, Niederau C, Kühl U, *et al.* Cardiac Troponin T in Patients With Clinically Suspected Myocarditis. *J Am Coll Cardiol* 1997; **30**: 1354–9.

65 Di Bella G, Florian A, Oreto L, *et al.* Electrocardiographic findings and myocardial damage in acute myocarditis detected by cardiac magnetic resonance. *Clinical Research in Cardiology* 2012; **101**: 617–24.

66 Biesbroek PS, Beek AM, Germans T, Niessen HWM, van Rossum AC. Diagnosis of myocarditis: Current state and future perspectives. *Int J Cardiol* 2015; **191**: 211–9.

67 Cooper LT, Baughman KL, Feldman AM, *et al.* The Role of Endomyocardial Biopsy in the Management of Cardiovascular Disease. *J Am Coll Cardiol* 2007; **50**: 1914–31.

68 Ma P, Liu J, Qin J, *et al.* Expansion of Pathogenic Cardiac Macrophages in Immune Checkpoint Inhibitor Myocarditis. *Circulation* 2024; **149**: 48–66.

69 Pan JA, Lee YJ, Salerno M. Diagnostic Performance of Extracellular Volume, Native T1, and T2 Mapping Versus Lake Louise Criteria by Cardiac Magnetic Resonance for Detection of Acute Myocarditis: A Meta-Analysis. *Circ Cardiovasc Imaging* 2018; **11**: e007598.

70 Luetkens JA, Doerner J, Thomas DK, *et al.* Acute Myocarditis: Multiparametric Cardiac MR Imaging. *Radiology* 2014; **273**: 383–92.

71 Nakaura T, Higaki T, Awai K, Ikeda O, Yamashita Y. A primer for understanding radiology articles about machine learning and deep learning. *Diagn Interv Imaging* 2020; **101**: 765–70.

72 Cheng PM, Montagnon E, Yamashita R, *et al.* Deep Learning: An Update for Radiologists. *RadioGraphics* 2021; **41**: 1427–45.

73 Chen C, Qin C, Qiu H, *et al.* Deep Learning for Cardiac Image Segmentation: A Review. *Front Cardiovasc Med* 2020; **7**. DOI:10.3389/fcvm.2020.00025.

74 Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *Eur Radiol Exp* 2018; **2**: 35.

75 Richardson ML, Garwood ER, Lee Y, *et al.* Noninterpretive Uses of Artificial Intelligence in Radiology. *Acad Radiol* 2021; **28**: 1225–35.

76 Akinci D'Antonoli T, Stanzione A, Bluethgen C, *et al.* Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagnostic and Interventional Radiology* 2024; **30**: 80–90.

77 Zhao XW, Zhou K, Li J, Tang T. A Survey of Large Language Models. 2023; published online Nov 24.

78 Wei J, Tay Y, Bommasani R, *et al.* Emergent Abilities of Large Language Models. 2022; published online June 15.

79 Hoffmann J, Borgeaud S, Mensch A, *et al.* Training Compute-Optimal Large Language Models. 2022; published online March 29.

80 Wei J, Wang X, Schuurmans D, *et al.* Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. 2022; published online Jan 27.

81 Vaswani A, Shazeer N, Parmar N, *et al.* Attention Is All You Need. 2017; published online June 12.

82 Lu Y, Bartolo M, Moore A, Riedel S, Stenetorp P. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. 2021; published online April 18.

83 OpenAI, Achiam J, Adler S, *et al.* GPT-4 Technical Report. 2023; published online March 15.

84 Bubeck S, Chandrasekaran V, Eldan R, Gehrke J. Sparks of Artificial General Intelligence: Early experiments with GPT-4. 2023; published online April 13.

85 Ouyang L, Wu J, Jiang X, Almeida D. Training language models to follow instructions with human feedback. 2022; published online March 4.

86 Schick T, Dwivedi-Yu J, Dessì R, Raileanu R. Toolformer: Language Models Can Teach Themselves to Use Tools. 2023; published online Feb 9.

87 Bautista AB, Burgos A, Nickel BJ, Yoon JJ, Tilara AA, Amorosa JK. Do Clinicians Use the American College of Radiology Appropriateness Criteria in the Management of Their Patients? *American Journal of Roentgenology* 2009; **192**: 1581–5.

88 Sheng AY, Castro A, Lewiss RE. Awareness, Utilization, and Education of the ACR Appropriateness Criteria: A Review and Future Directions. *Journal of the American College of Radiology* 2016; **13**: 131–6.

89 Rau A, Rau S, Zöller D, *et al.* A Context-based Chatbot Surpasses Radiologists and Generic ChatGPT in Following the ACR Appropriateness Guidelines. *Radiology* 2023; **308**. DOI:10.1148/radiol.230970.

90 Castillo C, Steffens T, Sim L, Caffery L. The effect of clinical information on radiology reporting: A systematic review. *J Med Radiat Sci* 2021; **68**: 60–74.

91 Wassermann TB, Straus CM. A Failure to Communicate? *Acad Radiol* 2018; **25**: 943–50.

92 Trivedi H, Mesterhazy J, Laguna B, Vu T, Sohn JH. Automatic Determination of the Need for Intravenous Contrast in Musculoskeletal MRI Examinations Using IBM Watson's Natural Language Processing Algorithm. *J Digit Imaging* 2018; **31**: 245–51.

93 Sun Z, Ong H, Kennedy P, *et al.* Evaluating GPT4 on Impressions Generation in Radiology Reports. *Radiology* 2023; **307**. DOI:10.1148/radiol.231259.

94 Adams LC, Truhn D, Busch F, *et al.* Leveraging GPT-4 for Post Hoc Transformation of Free-text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. *Radiology* 2023; **307**. DOI:10.1148/radiol.230725.

95 Kottlors J, Bratke G, Rauen P, *et al.* Feasibility of Differential Diagnosis Based on Imaging Patterns Using a Large Language Model. *Radiology* 2023; **308**. DOI:10.1148/radiol.231167.

96 Liu Z, Zhong A, Li Y, *et al.* Radiology-GPT: A Large Language Model for Radiology. 2023; published online June 14.

97 Fink MA, Bischoff A, Fink CA, *et al.* Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. *Radiology* 2023; **308**. DOI:10.1148/radiol.231362.

98 Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low Health Literacy and Health Outcomes: An Updated Systematic Review. *Ann Intern Med* 2011; **155**: 97.

99 Ayers JW, Poliak A, Dredze M, *et al.* Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med* 2023; **183**: 589.

100 Salam B, Kravchenko D, Nowak S, *et al.* Generative Pre-trained Transformer 4 makes cardiovascular magnetic resonance reports easy to understand. *Journal of Cardiovascular Magnetic Resonance* 2024; **26**: 101035.

101 Li H, Moon JT, Iyer D, *et al.* Decoding radiology reports: Potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports. *Clin Imaging* 2023; **101**: 137–41.

102 Jiang S-T, Xu Y-Y, Lu X. ChatGPT in Radiology: Evaluating Proficiencies, Addressing Shortcomings, and Proposing Integrative Approaches for the Future. *Radiology* 2023; **308**. DOI:10.1148/radiol.231335.

103 Cundari G, Galea N, De Rubeis G, *et al.* Use of the new Lake Louise Criteria improves CMR detection of atypical forms of acute myocarditis. *Int J Cardiovasc Imaging* 2021; **37**: 1395–404.

104 Larson DB, Towbin AJ, Pryor RM, Donnelly LF. Improving Consistency in Radiology Reporting through the Use of Department-wide Standardized Structured Reporting. *Radiology* 2013; **267**: 240–50.

105 Dimarco M, Cannella R, Pellegrino S, *et al.* Impact of structured report on the quality of preoperative CT staging of pancreatic ductal adenocarcinoma: assessment of intra- and inter-reader variability. *Abdominal Radiology* 2020; **45**: 437–48.

106 Wetterauer C, Winkel DJ, Federer-Gsponer JR, *et al.* Structured reporting of prostate magnetic resonance imaging has the potential to improve interdisciplinary communication. *PLoS One* 2019; **14**: e0212444.

107 Clusmann J, Kolbinger FR, Muti HS, *et al.* The future landscape of large language models in medicine. *Communications Medicine* 2023; **3**: 141.

108 Biswas S. ChatGPT and the Future of Medical Writing. *Radiology* 2023; **307**. DOI:10.1148/radiol.223312.

109 The Lancet Digital Health. ChatGPT: friend or foe? *Lancet Digit Health* 2023; **5**: e102.

110 Toma A, Senkaiahliyan S, Lawler PR, Rubin B, Wang B. Generative AI could revolutionize health care — but not if control is ceded to big tech. *Nature* 2023; **624**: 36–8.

111 Madden MG, McNicholas BA, Laffey JG. Assessing the usefulness of a large language model to query and summarize unstructured medical notes in intensive care. *Intensive Care Med* 2023; **49**: 1018–20.

112 Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI Responds to Common Lung Cancer Questions: ChatGPT versus Google Bard. *Radiology* 2023; **307**. DOI:10.1148/radiol.230922.

113 Lourenco AP, Slanetz PJ, Baird GL. Rise of ChatGPT: It May Be Time to Reassess How We Teach and Test Radiology Residents. *Radiology* 2023; **307**. DOI:10.1148/radiol.231053.

114 Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* 2023; **6**: 120.

115 Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. *Nat Med* 2023; **29**: 2396–8.

6. Anhang

6.1. Abbildungsverzeichnis

7. Vorabveröffentlichungen von Ergebnissen

Basis für diese kumulative Dissertation ist der 2024 in der Fachzeitschrift „Journal of Cardiovascular Magnetic Resonance“ veröffentlichte Artikel „Generative Pre-trained Transformer 4 analysis of cardiovascular magnetic resonance reports in suspected myocarditis: A multicenter study“ mit dem Doktoranden als Erstautor. Im genannten Artikel wurden die Ergebnisse dieser Promotion in Rücksprache mit dem Betreuer vorabveröffentlicht. „Journal of Cardiovascular Magnetic Resonance“ ist eine PubMed gelistete Fachzeitschrift, die Artikel erst nach Peer Review veröffentlicht. Es ist eine offizielle Fachzeitschrift der Society for Cardiovascular Magnetic Resonance (SCMR). Der Doktorand hatte bei der Erstellung der Publikation den wichtigsten Anteil. So wurde das Manuskript durch den Doktoranden verfasst. Auch Hypothesenstellung, Datensammlung und Datenauswertung erfolgte durch den Doktoranden; hier in Rücksprache mit den Co-Autoren des Papers.