Original-Titel:
Comparing the number and relevance of false activations between 2 artificial intelligence computer-aided detection systems: the NOISE study

NOISE-Studie: Vergleich der absoluten Anzahl und Relevanz von falsch-positiven Signalen (Fehl-Aktivierungen) von 2 unterschiedlichen KI-Systemen bei der automatisierten Polypen-Detektion.

Autoren:
Marco Spadaccini, MD; Cesare Hassan, MD, PhD; Ludovico Alfarone, MD; Michael B. Wallace, MD; Prateek Sharma, MD; Alessandro Repici, MD et al. | Gastrointest Endosc 2022; 95:975-81

Kommentar:
PD Dr. Axel Eickhoff, Medizinische Klinik II, Klinikum Hanau, 23.05.2022

Die künstliche Intelligenz (KI) bzw. artificial intelligence (AI) ist aus unserem modernen digitalen Leben nicht mehr wegzudenken und hält seit 3 Jahren breiten Einzug auch in den medizinischen Sektor. Sämtliche bildgebenden Verfahren (CT/MRT, Endoskopie, Sonographie) machen sich heute immer mehr der bestechenden Vorteile der sogenannten visuellen künstlichen Intelligenz (VKI) zu nutzen.

Dabei geht es zunächst um die eigentliche Mustererkennung (Detektion) und im zweiten, viel faszinierenderen Schritt um die Mustervorhersage, also der Charakterisierung von Objekten und Bildern. Dabei werden anhand einer Bildserie vorhergesagt, um was es sich handelt („Charakterisierung“) und zukünftig auch was zu tun ist („doing“).

Die entscheidenden Qualitätsparameter bei der KRK-Vorsorge durch Screening-Koloskopie sind die Adenom-Detektions-Rate (ADR) und die Übersehens Rate (Adenoma-Miss-rate: AMR). Diese ist abhängig von Untersuchungs-/ und Untersucherrelevanter Faktoren wie Darmsauberkeit, Rückzugszeit bei Koloskopie, Geräte und Hilfstechniken sowie vor allem der Erfahrenheit des Untersuchers. Nur mit einer ausreichend hohen Detektionsrate kann das Auftreten des Kolorektalen Karzinoms (KRK) nachhaltig gesenkt werden. Die ADR sollte daher heute nach Empfehlung der Fachgesellschaften mindestens 30% betragen. Moderne Verfahren wie die künstlichen Intelligenz (KI) bieten hier natürlich faszinierende neue Möglichkeiten, diese müssen aber durch Studien und eine gute Evidenz abgesichert und auch gegen mögliche Nachteile und Schwächen abgewogen werden. Der Wert der KI bei der Polypen Erkennung (ADR) ist inzwischen wissenschaftlich in einer Vielzahl prospektiver und randomisierter Studien, sowohl aus Asien als auch der westlichen Hemisphäre, untersucht worden. Die sogenannte computergestützte Erkennung (CADe), hat sich dabei als überaus wirksam bei der Polypen Erkennung erwiesen, es konnten Sensitivitätswerten zwischen 75 % und 100 % dokumentiert werden. Also alles klar und KI/ CADe ist der „game changer“ in der Vorsorge und Koloskopie? Wahrscheinlich ja, wir sollten uns trotzdem der möglichen Schwächen und Nachteile bewusst sein. Welche könnten das sein?

- Die KI ist nur so gut, wie die Endoskopiker/ Experten die sie „trainiert“ haben
- Nur die real-sichtbare Oberfläche wird beurteilt (Problem Verschmutzung, Falten, Flexuren)
- Heterogenität der verschiedenen Systeme und deren Trennschärfen
- „Verdummung“ der Ärzte und Endoskopiker („Tunnelblick“)
- Reizüberflutung und Übermüdung/ Aufmerksamkeitsreduktion durch falsch-positive Signale/ Alarme

Insbesondere die mögliche Heterogenität der verschiedenen Systeme und die Rate falsch-positive Signale/ Alarme (FP-Rate) stellen ein mögliches Problem vor genereller Implementierung der KI bei der Vorsorge da. Die Effizienz der KI würde unter einer hohen falsch-positiv Rate leiden, denn durch sie erhöht sich neben Untersuchungs- und Rückzugszeit auch die Zahl unnötiger Polypektomien und es kommt darüber hinaus auch zur möglichen „Reizüberflutung“ der Untersucher.

Sowohl aus medizinischer als auch der Patientenperspektive muss somit vor breiter Einführung der KI in die klinische Praxis neben der Minimierung der Heterogenität und Standardisierung der verschiedenen Systeme auch ein eindeutiges, messbares „benchmark tool“ definiert werden, um die verschiedenen kommerziell verfügbaren Techniken vergleichen zu können. Ein möglicher Weg zur skalierbaren Vergleichbarkeit von KI/ CADe in der klinische Praxis könnte ebendiese FP-Rate an falsch-positiven Signalen/ Alarmen sein. Dies ist das Ziel der vorliegenden Studie, die publiziert wurde von einer ausgewiesenen Gruppe von Koloskopie-KI Experten aus Italien und USA.

Was wurde untersucht und wie stellen sich die Ergebnisse dar?

Prospektiv evaluiert wurden zwei kommerzielle KI-Systeme (CAD Eye Fujifilm Corp. und GI-Genious, Medtronic Corp.) hinsichtlich ihrer FP-Rate. Dabei sollte evaluiert werden, ob die FP-Rate und die aktuell diskutierte NOISE Definition-/ Klassifikation zur besseren Vergleichbarkeit der KI-Systeme herangezogen werden kann. Dabei beschreibt die NOISE Klassifikation die Ursachen von falsch-positiven KI-Signalen bei der Koloskopie. Unterschieden werden dabei erstens Artefakte der Darmwand selbst (Falten, IC-Klappe, Appendix, Saug-/Quetschartefakte etc.) und zweitens Ursachen im Darmlumen (Stuhlreste, Schleim, Blasen, Wasser, Blut). Die klinische Relevanz definiert über den Zeitzuwachs der Untersuchung durch Klärung FP-Alarme wurde ebenfalls hinreichend evaluiert. Sie reicht von mild (<1 sek.), moderat (1-3 sek.) bis hin zu ausgeprägtem (>3 sek.) Zeitzuwachs zur Klärung. Hierfür wurden im Zeitraum 09/20 bis 11/20 insgesamt 40 komplette Prozeduren und Videosequenzen aus einem KI-Referenzzentrum in Italien für die KI-Datenbank generiert und vom System fortlaufend analysiert (CADe System A= Cade Eye Fujifilm). Diese wurden mit 40 Sequenzen aus einer bereits publizierten Studie aus 2020 (AID Trial) verglichen (CADe System B= GI Genious). Auf der Basis der NOISE Klassifikation wurden alle 80 Sequenzen zunächst von 2 unabhängigen erfahrenen Reviewern beurteilt und nachfolgend noch durch einen dritten „Senior“ Reviewer objektiviert.

Die Ergebnisse in Kürze:

- FP-Rate CADe System A 1021 gesamt, pro Koloskopie 25.5 +/-12.2 versus CADe System B 1028 gesamt, pro Koloskopie 25.7 +/-13.2 (p=0.53) nicht signifikant
- Ursache Darmwand: CADe- System A 22.9 (89.8%) versus 22.1 (86%) CADe-System B
- Ursache Darmlumen: 2.6 (10.2% in CADe A) und 3.5 (14% in CADe B)
- Klinische Relevanz: CADe-A Zeitzuwachs plus 1.6 +/-1.0 pro FP pro Koloskopie versus CADe-B plus 1.8 +/-1.2 pro FP und Koloskopie und somit auch ohne signifikanten Unterschied
- Prozentual waren dies 0.8% (CADe-A) und 0.9% (CADe-B) relativer Zeitverlust beim Rückzug und dieser somit verschwindend gering für die Gesamtprozedur
- Nur in ca. einem Viertel (27.7% und 26.4%) aller FP's war der Zeitverlust >3 Sek. und damit höhergradig (Stufe 3)

Als wichtigste Kernbotschaft aus der Studie kann mitgenommen werden, dass beide CADe-/ KI-Systeme (Fujifilm und GI Genious) eine vergleichbare Effektivität und Sensitivität in der Adenom-Detektionsrate (ADR) aufzuweisen scheinen. Somit scheint die postulierte Heterogenität der KI-Systeme sehr wahrscheinlich doch eher theoretischer Natur zu sein. Zweitens konnte die Arbeitsgruppe sehr schön zeigen, dass die NOISE-Klassifikation basierend auf der FP-Rate ein valides Mittel zu sein scheint, auch zukünftig KI-Systeme hinsichtlich ihrer Effektivität und klinischen relevant zu vergleichen und zu bewerten.