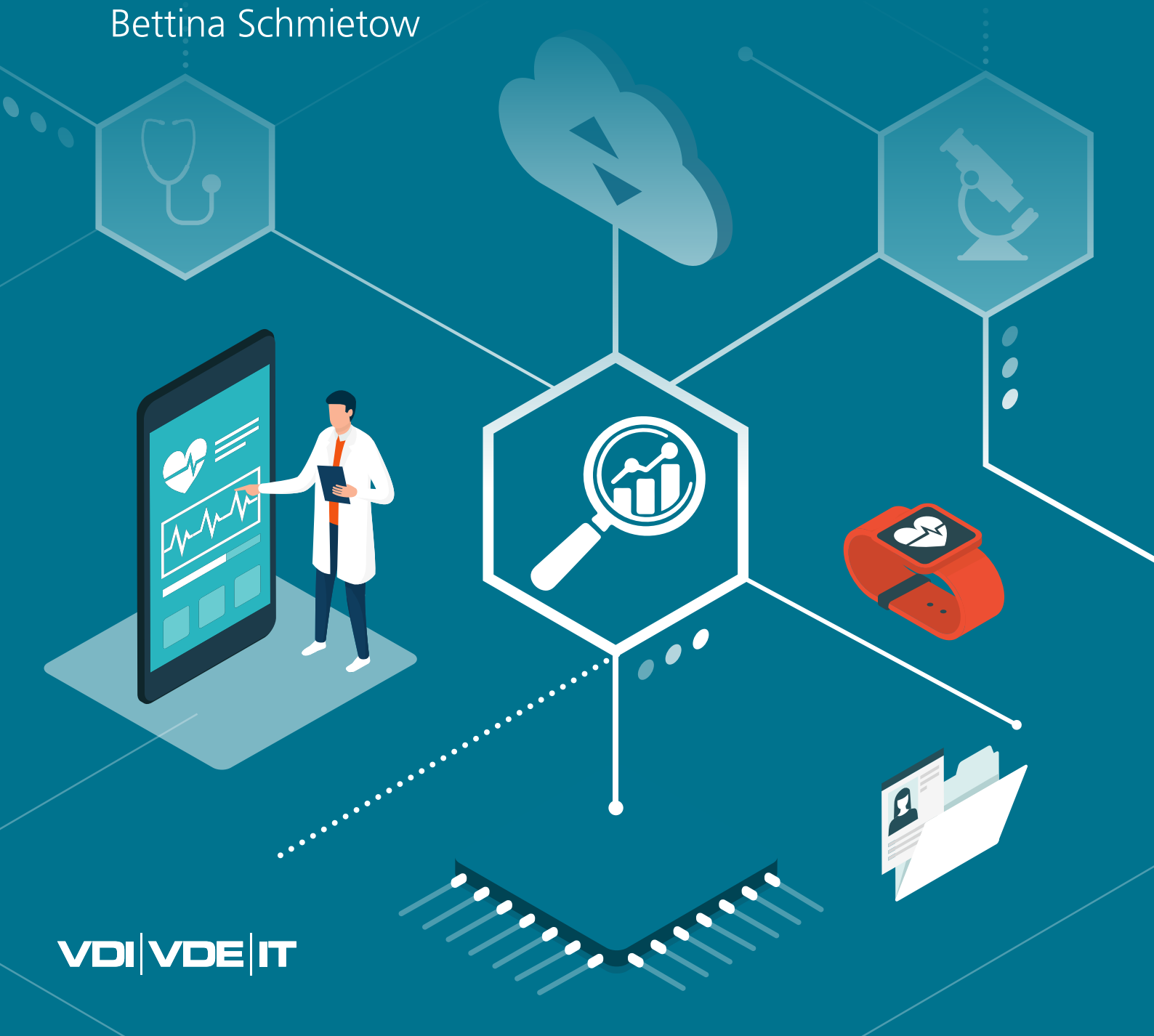
Von einzelnen Metriken zu dynamischer Evidenz – Aktuelle Herausforderungen und Perspektiven in der Evaluation KI-basierter Medizintechnik

Katharina Dassel, Benedikt Krieger, Maxie Lutze, Bettina Schmietow



VDI|VDE|IT

Von einzelnen Metriken zu dynamischer Evidenz – Aktuelle Herausforderungen und Perspektiven in der Evaluation KI-basierter Medizintechnik

Künstliche Intelligenz hält zunehmend Einzug in die Medizintechnik und verspricht Fortschritte in Diagnostik, Therapie und Versorgungsprozessen. Mit der wachsenden Verbreitung KI-gestützter Systeme verändert sich jedoch auch der Nachweis von Sicherheit, Wirksamkeit und Nutzen. Dieser Beitrag beleuchtet aktuelle Herausforderungen und Ansätze der Evaluation KI-basierter Medizintechnik und zeigt auf, wie technische Leistungsfähigkeit, klinische Evidenz und Real-World-Daten zu einem ganzheitlichen Bewertungsrahmen zusammengeführt werden können. Evaluation wird dabei nicht nur als regulatorische Pflicht verstanden, sondern als zentrales Instrument für Vertrauen, Akzeptanz und erfolgreichen Transfer in die Versorgung.

1. Warum die Evaluation von KI-Systemen entscheidend ist

KI-Innovationen verändern die Medizintechnik in bisher ungesehenem Tempo. Gleichzeitig ändern sich die rechtlichen Rahmenbedingungen für den Einsatz von KI-gestützter Medizintechnik. In der Europäischen Union trat 2024 der AI Act in Kraft, während Deutschland mit der DiGA-Verordnung und der Medizininformatik-Initiative internationale Maßstäbe auch für digitale Medizintechnik und Datennutzung setzt. Doch KI ist nicht gleich KI: Entscheidungsbäume, Machine-Learning-Modelle, Deep-Learning-Architekturen oder Transformer-Systeme unterscheiden sich nicht nur technisch, sondern auch in ihren Risiken, Einsatzbedingungen und Anforderungen an Transparenz und Robustheit. Diese Vielfalt ist nicht nur eine technologische, sondern eine evaluative Herausforderung. Denn die Art der KI bestimmt, welche Nachweise erforderlich sind, um ihre medizinische Wirksamkeit, Sicherheit und ihren Nutzen belegen zu können. Damit verbindet sich die definitorische Frage ‚Was ist KI?' unmittelbar mit der Evaluationsfrage: Welche Evidenz braucht welches System in welchem Kontext?

Auch von Grundlagenforschung über präklinische und klinische Forschung bis zu Routineversorgung erfordert jede Etappe unterschiedliche Evaluationen. Deshalb will dieser Artikel der Frage nachgehen: Wie lässt sich in einzelnen Anwendungen belegen, dass KI-Systeme nicht nur technisch funktionsfähig, sondern auch medizinisch wirksam und gesellschaftlich nützlich sind? Denn das Zielbild ist klar: Evaluation als Gesamtkonzept sollte darauf einzahlen, dass es effektive wie effiziente neue Anwendungen in die Versorgungspraxis schaffen. Evaluation ist also Mittel zum Zweck der Translation und des Transfers. Unter Evaluation verstehen wir dabei den systematischen und mehrdimensionalen Nachweis, dass KI-gestützte Medizintechnik (KI-MT) im klinischen Kontext tatsächlich zur Verbesserung von Gesundheit, Sicherheit und Qualität sowie Ressourceneffizienz beiträgt. Sie ist damit nicht nur regulatorische Pflicht, sondern auch Voraussetzung für Vertrauen, Akzeptanz, Versorgungsgüte und Markterfolg. Dieser Beitrag beleuchtet zentrale Dimensionen und aktuelle Ansätze einer ganzheitlichen KI-Medizintechnik-Evaluation – von der Metrik zur Evidenz.

2. Technische Evaluation: Leistungsfähigkeit, Robustheit und Generalisierbarkeit von KI

In der frühen Entwicklungsphase ist die technische Validierung der KI entscheidend: Sensitivität, Spezifität oder AUC-Werte (Area under the Curve) belegen die Modellgüte. Doch diese Kennzahlen beschreiben zunächst die Leistung im Labor, die sich im klinischen Alltag erst noch unter Beweis stellen muss. Ein zentrales Problem ist die Übertragbarkeit: Wie stabil bleibt die KI, wenn sich Daten, Geräte, Anwender:innen oder die Patientengruppen verändern? Die Herausforderungen von Data-, Concept- und übergeordnet Model-Drift – also dem Umstand, dass sich Eingabedaten, Zusammenhänge und Modellverhalten über die Zeit verschieben – sind wissenschaftlich zwar anerkannt, methodisch aber noch unzureichend adressiert. In der Praxis existieren bislang kaum verbindlich standardisierte Messgrößen für reale Leistungsstabilität und nur begrenzte Verfahren zur prospektiven Drift-Überwachung, wobei erste Methoden bereits diskutiert werden (Dolin et al. 2025, FDA 2025). Dabei variiert das Drift- und Fehlerprofil je nach Modelltyp erheblich: Während bei regelbasierten Systemen Veränderungen in Datenqualität und -verteilung besser nachvollzogen werden können,

reagieren Deep-Learning- oder Transformer-Modelle nicht absehbar auf solche Veränderungen.

Die Good Machine Learning Practice (GMLP) der Food and Drug Administration (FDA, U.S.), Medicines and Healthcare products Regulatory Agency (MHRA, U.K.) und Health Canada (2021) fordert deshalb kontinuierliches Monitoring und geeignete Kontrollmaßnahmen, um derartigen Risiken zu begegnen.

Diese Anforderungen lassen sich in einem gestuften Prüfkonzept abbilden, das unterschiedliche Testarten entlang des Lebenszyklus eines KI-Systems unterscheidet. Die in der folgenden Abbildung dargestellte Gewichtung der Testarten verdeutlicht, dass formative Tests, summative Validierung, Site Acceptance Tests und kontinuierliches Monitoring je nach Typ des KI-Systems, seinem Risikoprofil und seinem Lernverhalten eine unterschiedliche Relevanz besitzen. Dabei macht die Abbildung deutlich, dass Evaluationsanforderungen nicht einheitlich, sondern risikoadäquat und kontextsensitiv ausgestaltet werden müssen.

Es lassen sich vier Arten von Tests unterscheiden: Formative Tests werden während der Entwicklung durchgeführt, um das Modell iterativ zu verbessern. Der Summative Test dient der finalen Validierung, bevor eine Installation in der Zielumgebung erfolgt. Er soll mit unabhängigen Testdaten die Leistungsfähigkeit und Sicherheit des Systems nachweisen. Der Site Acceptance Test (SAT) erfolgt in einer dem Einsatzort entsprechenden Testumgebung mit realitätsnahen Daten und identischen IT-Komponenten. Sowohl der Summative Test als auch der SAT fungieren als Gatekeeper für die Freigabe zum Betrieb. Nur wenn diese bestanden sind, sollte ein KI-System in Betrieb genommen werden. Nach der Inbetriebnahme eines KI-Systems ist dessen Leistungsfähigkeit und Sicherheit kontinuierlich zu überwachen. Diese Überwachung umfasst Maßnahmen wie eine regelmäßige Re-Validierung (zum Beispiel zur Erkennung von Modell-Drifts), ein kontinuierliches Monitoring (zum Beispiel Fehlerraten durch Overrides, Performancekennzahlen) sowie strukturierte Reaktionen auf Vorfälle und kritische Ereignisse. Die Relevanz der einzelnen Testarten variiert in Abhängigkeit vom Typ des KI-Systems, seinem Risikoprofil und seinem Einsatzkontext.

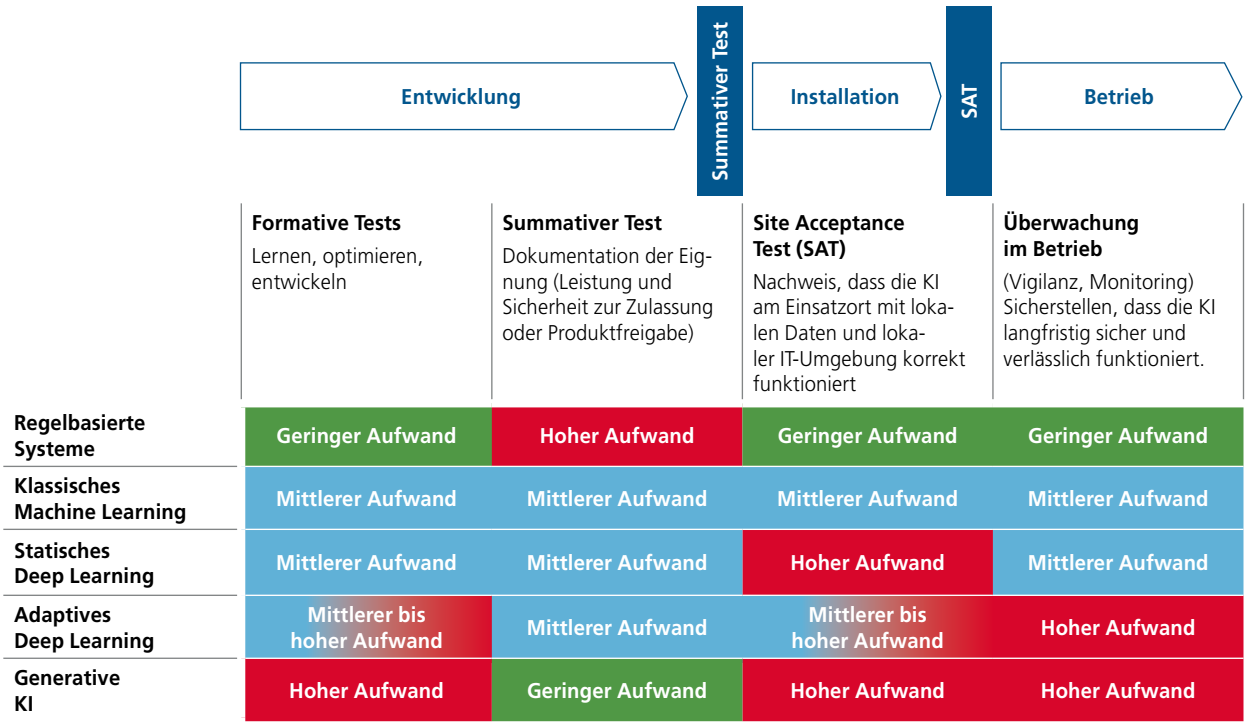


Abbildung 1: Gestuftes Prüfkonzept für KI-gestützte Medizintechnik, das die Relevanz der verschiedenen Tests von KI-Systemen in Abhängigkeit von der Art der eingesetzten KI-Methoden zeigt. (Quelle: Abbildung Lipprandt M, Storck M, Röhrig R, Technologie- und Methodenplattform für die vernetzte medizinische Forschung e. V., Arbeitsgruppe Medizinische Software und Medizinprodukte[1]

1 **Regelbasierte Systeme** (Symbolische KI, Expertensysteme): Wissensbasierte, regelgeleitete Systeme mit deterministischem Verhalten. / **Klassisches Machine Learning:** Datengetriebene Modelle, die Muster in historischen Beispielen lernen (z. B. Regressionsmodelle). **Statisches Deep Learning:** Komplexe neuronale Netze, die nach dem Training unverändert eingesetzt werden. Adaptives Deep Learning: Die Modelle lernen nach der Inbetriebnahme weiter, / passen sich selbständig an neue Daten an (Online Learning). **Generative KI:** Modelle, die neue Inhalte erzeugen können – z. B. Daten, Text, Bilder, (z. B. LLMs, Diffusion Modelle).

3. Klinische Evaluation: Von der Modellleistung zum medizinischen Nutzen mit Versorgungseffekten

Technische Performanz ist Voraussetzung, aber erst Robustheit und Nachbeobachtung (Post-Market Surveillance (PMS) und Post-Market Clinical Follow-up (PMCF)) sichern klinische Verlässlichkeit. Dabei bestimmt nicht nur der Einsatzkontext, sondern auch der Modelltyp, welche klinischen Nachweise angemessen sind: Statische ML-Modelle lassen sich anders evaluieren als adaptiv lernende Deep-Learning- oder Transformer-Systeme, die sich ohne expliziten Model-Freeze während der Nutzung weiterentwickeln können. Klinische Evaluation muss den Übergang vom Labor zur Versorgung abbilden und zeigen, ob Systeme unter realen Bedingungen wirksam sind und inwieweit sich ihr Nutzen auch in effizienteren Prozessen widerspiegelt. Viele Publikationen zeigen eine algorithmische Überlegenheit, nicht aber bessere Outcomes in der Versorgungspraxis (Wilhelm et al. 2025). So kann etwa ein Algorithmus Tumore schneller erkennen, ohne dass dies zu früheren oder besseren Therapien führt. Um entsprechende Zusammenhänge belastbar zu prüfen, braucht es geeignete Studiendesigns, die die Dynamik von KI-MT berücksichtigt. Klassische RCTs gelten zwar weiter als Goldstandard, sind aber aufwändig und stoßen gerade bei lernenden oder workflow-integrierten Systemen an Grenzen. Besonders komplexe, kontinuierlich lernende Modellarchitekturen verändern ihr Entscheidungsverhalten im Versorgungsalltag, was klassische Studiendesigns nur unzureichend abbilden können. Der Großteil der Evidenz bleibt auf retrospektive Modellstudien beschränkt, während robuste prospektive, in-clinico-Evaluationen fehlen (Weissmann 2025). Folglich braucht es flexiblere Studiendesigns, etwa adaptive, pragmatische und Cluster-RCTs, die Evaluation unter realen Versorgungsbedingungen erlauben, ohne wissenschaftliche Strenge aufzugeben.

Die genannten methodischen Herausforderungen prägen zunehmend auch Regulierungs- und Bewertungsinstitutionen und verschärfen sich bezogen auf KI-MT. Leitlinien des International Medical Device Regulators Forum (IMDRF) zu Software as a Medical Device sowie das NICE Evidence Standards Framework adressieren diese Dynamik, indem sie risikobasierte Anforderungen an Evidenzgenerierung, klinische Leistungsnachweise und den Umgang mit algorithmischer Adaptivität formulieren. Während das IMDRF primär regulatorische Leitplanken für die Zulassung lernender Systeme setzt, fungiert NICE als Brücke zwischen Regulierung und Health-Technology-Assessment (HTA), indem regulatorische Anforderungen in versorgungsrelevante klinische und ökonomische Bewertungslogiken übersetzt werden. In den USA ergänzt das ICER-PHTI-Framework (2023) diese Perspektive: Es versteht sich als HTA-Instrument, das digitale Technologien nach Funktion und Risiko einordnet und damit erstmals differenziert, welche Art von Evidenz bei welchem Risikoprofil als ausreichend gilt. Neben klinischer Wirksamkeit integriert es Usability, Privacy & Security, Health Equity und Economic Impact. So kann der Beitrag von KI-MT zu Zugängen, Workflows und Gesundheitsverhalten adäquat abgebildet werden.

In Europa existiert ein mehrschichtiger Bewertungsrahmen für KI-Systeme: Auf regulatorischer Ebene bilden MDR/IVDR in Verbindung mit KI-spezifischen Auslegungen zu Software as a Medical Device die Grundlage für Sicherheits-, Leistungs- und Post-Market-Anforderungen. Die EU-HTA-Verordnung (EU) 2021/2282 etabliert ergänzend einen rechtlichen Rahmen für gemeinsame klinische Bewertungen, deren Ergebnisse künftig auch für KI-basierte Medizinprodukte nutzbar sein sollen, ohne nationale Erstattungsentscheidungen zu harmonisieren. Methodisch adressiert EDiHTA diese Lücke, indem es ein DHT-spezifisches HTA-Framework bereitstellt, das KI-Eigenschaften wie Adaptivität, Implementierungskontexte und unterschiedliche Reifegrade systematisch berücksichtigt. In Deutschland konkretisieren die DiGA- und DiPA-Verfahren des BfArM diese Ebenen zu einem nationalen Erstattungszugang, bei dem regulatorische Konformität vorausgesetzt wird, während der Fokus der Bewertung auf positiven Versorgungseffekten, Nutzennachweisen unter Real-World-Bedingungen und der Einbettung in bestehende Versorgungsprozesse liegt.

4. Akteursperspektiven: Wer braucht welche Evidenz?

Die Vielzahl der Dimensionen in Evaluationsframeworks ist mitunter durch die Vielzahl von Akteuren – von Entwickler:innen und Herstellern über Ärzt:innen und Pflegepersonal bis hin zu Regulierungsbehörden, Kostenträgern, Patient:innen und Gesundheitseinrichtungen – zu begründen. Jede dieser Gruppen verfolgt eigene Erkenntnisinteressen und nutzt Evidenz auf unterschiedliche Weise. Während Entwickler:innen vor allem Nachweise zur Leistungsfähigkeit und Sicherheit benötigen, um Zulassung und Marktzugang zu erlangen, interessieren sich Kliniker:innen für Alltagstauglichkeit, Arbeitsentlastung und Patientensicherheit. Kostenträger wiederum fordern Nachweise für Versorgungseffizienz und Wirtschaftlichkeit, und Patient:innen erwarten Technologien, die sicher, nachvollziehbar und zugänglich sind. Sie sehen in KI-Systemen zugleich Hoffnung auf präzisere Diagnostik und personalisierte Therapien, aber auch Risiken für Datenschutz, algorithmische Fairness und die Qualität der Arzt-Patient-Interaktion (Hogg et al. 2023). Gesundheitsorganisationen und Führungskräfte wiederum achten auf Workflow-Integration, Schulungsbedarf, Interoperabilität und Investitionssicherheit (Wang und Beecy 2025). KI-MT kann nur

dann erfolgreich implementiert werden, wenn Prozessänderungen mitgedacht werden, das Personal unterstützt statt überfordert und langfristig betriebliche Stabilität gewährleistet wird. Entsprechend haben auch organisatorische Kriterien – etwa Change-Management, Schnittstellenkompatibilität oder Wartungsaufwand – einen Einfluss auf das Ergebnis der Nutzenbewertung.

Diese unterschiedlichen Perspektiven treffen auf ein umfassendes, international heterogenes Regulierungsumfeld. Zwar besteht breiter Konsens über grundlegende Prinzipien wie Sicherheit, Wirksamkeit und Nutzen, doch die konkrete Auslegung variiert erheblich zwischen Behörden und Bewertungssystemen. So akzeptieren etwa FDA und MHRA zunehmend Real-World-Evidence als Teil klinischer Nachweise, während europäische Instanzen auf formale Studienstrukturen und kontrollierte Designs setzen. Die MDR lässt Spielräume zu, die von den benannten Stellen unterschiedlich interpretiert werden. Für Hersteller bedeutet dies, Evidenz mehrfach und in verschiedenen Formaten erbringen zu müssen. Für Regulierungs- und Bewertungsinstitutionen wiederum stellt sich die Frage, wie sich nationale Anforderungen mit globalen Standards abstimmen lassen, ohne Patientensicherheit oder Transparenz zu gefährden (Reddy 2025). Evidenz ist daher nicht nur eine wissenschaftliche, sondern auch eine kommunikative, werte- und interessengeleitete Währung zwischen Akteuren. Was beispielsweise für Regulierer als ausreichend gilt, kann für Kostenträger zu wenig und für Kliniker:innen zu abstrakt sein als auch andersherum.

5. Partizipative Evaluation in der KI-Medizintechnik: Neue Rollen für Patient:innen und Fachpersonal

Patient:innen rücken zunehmend ins Zentrum der Evaluation von KI-MT. Sie sind nicht mehr nur Datenspender:innen oder Beobachtete klinischer Studien, sondern aktive Mitgestalter:innen der Evidenz und damit auch der sicheren und fairen Nutzung. Ihr Nutzungserleben, ihr Vertrauen in die Technologie und ihre Rückmeldungen im Versorgungsalltag prägen unmittelbar, wie KI-Systeme performen, wahrgenommen werden und sich über ihren Lebenszyklus hinweg entwickeln. Nur wenn Patient:innen nachvollziehen können, wie und warum ein System eine Entscheidung empfiehlt, lässt sich Vertrauen aufbauen und die Nutzung langfristig sichern. Gleiches gilt für Kliniker:innen und Pflegepersonal: Ihre Rolle verschiebt sich von der reinen Anwendung hin zur Co-Evaluation. Damit bewerten sie nicht nur Ergebnisse eines technischen Systems, sondern auch dessen Nutzungsformen, Workflow-Integration und Sicherheit.

Diese Verschiebung spiegelt sich auch methodisch wider. Neben klassischen klinischen Endpunkten gewinnen subjektive und kontextbezogene Messgrößen an Bedeutung: Patient Reported Outcome Measures (PROMs), Patient Reported Experience Measures (PREMs) oder qualitative Feedbackschleifen liefern wertvolle Hinweise auf Erklärbarkeit, Alltagstauglichkeit und wahrgenommenen Nutzen oder potenzielle Verzerrungen in Versorgungsalltag. In Kombination mit Real-World-Evidence (RWE) entsteht so ein kontinuierlicher Evaluations- und Lernprozess, der technische Leistungsfähigkeit, soziale Wirkung und Nutzungskontexte gemeinsam abbildet. Partizipative Ansätze, etwa Fokusgruppen in der Zieldefinition, partizipative Studienprotokolle oder Rückkopplung über digitale Plattformen ermöglichen es, dass Patient:innen und Fachpersonal ihre Perspektiven frühzeitig einbringen. So erweitert sich Evaluation von einem technischen Validierungsprozess zu einem Aushandlungsprozess geteilter Methoden und Praktiken in Bezug auf Vertrauen, Akzeptanz und tatsächlicher Wirkung in der Versorgung.

6. Daten der Real-World-Evidence als Fundament lernender Evaluation

Die Qualität der Daten bestimmt die Qualität der Evidenz, das gilt insbesondere für KI-Systeme. Damit verschiebt sich der Fokus von der einmaligen Datenerhebung hin zu einer kontinuierlichen, kontextsensitiven Datennutzung. Real-World-Evidence beschreibt Erkenntnisse, die aus realen Versorgungsdaten gewonnen werden, also aus der Nutzung von zum Beispiel elektronischen Patientenakten, Registern oder Routinedaten. Im Gegensatz zu kontrollierten Studiendaten spiegelt dies die Vielfalt realer Bedingungen wider: unterschiedliche Patientengruppen, Geräte, Klinikumgebungen und Verhaltensmuster.

Die Entwicklung und Evaluation von KI-MT hängt stark von großen, realen klinischen Datenbeständen und mehrphasiger, kontextbezogener Forschung ab, da nur so Leistung, Fairness und Robustheit solcher Systeme in realen Versorgungsszenarien belastbar bewertet werden können. Der geschützte Rahmen der Forschung spielt dabei eine zentrale Rolle, weil er es ermöglicht, KI-Systeme über längere Zeiträume zu entwickeln, zu erproben und mit wachsenden Datenmengen systematisch zu evaluieren. Für die Bewertung von KI-MT ist dabei die Verbindung zwischen Forschungsdaten und Versorgungsdaten entscheidend. Forschungsdaten stammen aus kontrollierten Studien und dienen der Entwicklung und Validierung. Sie gewährleisten Präzision und Vergleichbarkeit, bilden aber reale Nutzungssituationen nur begrenzt ab. Versorgungsdaten hingegen entstehen im klinischen Alltag und zeigen, ob ein System unter realen Bedingungen stabil, sicher und wirksam ist. Erst die Kombination beider Datenwel-

ten – kontrollierte Forschung (interne Validität) und reale Versorgung (externe Validität) – ermöglicht eine belastbare Evidenzkette von der Idee bis zur Anwendung. Je nach Modelltyp kann der Bedarf an RWE erheblich variieren: Während manche statischen Modelle bereits mit begrenzten klinischen Daten ausreichend validiert werden können, sind kontinuierlich lernende oder auf Human-AI-Teaming angewiesene Konzepte auf regelmäßige Rückkopplung aus Versorgungsdaten (etwa im Rahmen von PMS und PMCF) angewiesen, um Wirksamkeit und Sicherheit langfristig zu gewährleisten.

Die Idee der Real-World-Evidence ist international gereift und strukturell verankert: Bei der US-amerikanischen FDA als Erweiterung der GMLP entwickelt, wurde sie inzwischen von EMA, MHRA und der europäischen HTA-Verordnung (2021/2282) aufgegriffen. In Deutschland wird RWE etwa im DiGA- und DiPA-Verfahren des BfArM sowie in der Versorgungsforschung des Innovationsfonds oder durch Initiativen wie die Medizininformatik-Initiative und HiGHmed umgesetzt. Das neue Forschungsdatenzentrum Gesundheit (§ 303a–f SGB V) schafft zudem rechtliche und technische Rahmenbedingungen, um pseudonymisierte Versorgungsdaten für evidenzbasierte Forschung nutzbar zu machen.

RWE ermöglicht damit weit mehr als nachträgliche Evaluation. Sie wird zum Bindeglied und zur gemeinsamen Evidenzsprache zwischen Regulierungs-, Bewertungs-, Versorgungs- und Entwicklungsebene:

- Für Regulierungsbehörden liefert sie kontinuierliche Sicherheits- und Leistungsdaten im Rahmen der Post-Market-Surveillance.
- Für Kliniken und Fachpersonal bietet sie eine Rückkopplungsschleife, um Nutzung und Integration zu verbessern.
- Für Kostenträger und HTA-Gremien schafft sie die Grundlage für eine dynamische Nutzenbewertung unter Alltagsbedingungen.
- Für Patient:innen dokumentiert sie, ob Systeme tatsächlich Selbstständigkeit, Teilhabe und Versorgungsqualität fördern.

Dennoch bleibt RWE methodisch und organisatorisch herausfordernd. Fragen des Datenschutzes, der Datenhoheit und der Standardisierung sind noch unzureichend harmonisiert. Unterschiedliche Definitionen, Qualitätskriterien und Kodierungen erschweren den internationalen Vergleich und Datenaustausch. Für die KI-MT-Forschung kommt weiterhin ein Zielkonflikt hinzu: Sie muss bei der prospektiven Erhebung von Versorgungsdaten berücksichtigen, nicht als zulassungsrelevante klinische Prüfung nach MDR eingeordnet zu werden. Die damit verbundenen regulatorischen, organisatorischen und finanziellen Anforderungen stellen insbesondere für universitäre Forschungseinrichtungen eine erhebliche Hürde dar.

7. Lernende Evaluation: Offene Fragen und Ableitungen

a. Dateninteroperabilität ist der Schlüssel zur Real-World-Evidence

Eine belastbare Evaluation KI-gestützter Medizintechnik erfordert einen integrierten Bewertungsrahmen, der Forschungs- und Versorgungsdaten mit regulatorischen Anforderungen und gesellschaftlichen Zielsetzungen verbindet. Evaluation darf dabei nicht auf die Zulassungsphase beschränkt bleiben, sondern muss Entwicklung, Markteintritt und Anwendung im Versorgungskontext über den gesamten Lebenszyklus hinweg begleiten, wie es sowohl der europäische AI Act als auch aktuelle HTA-Rahmenwerke vorsehen.

Initiativen wie die Medizininformatik-Initiative oder das Forschungsdatenzentrum Gesundheit zeigen, wie interoperable Datenräume als Grundlage für kontinuierliche Evidenzgenerierung aufgebaut werden können. Gemeinsame Referenzstrukturen auf Basis interoperabler Datenstandards (zum Beispiel FHIR, DICOM) ermöglichen es, KI-MT über verschiedene Versorgungssettings hinweg vergleichbar, sicher und langfristig zu evaluieren, unabhängig von ihrer technischen Komplexität.

Auf dieser Basis wird eine kontinuierliche, datengetriebene Überwachung von Leistungs- und Sicherheitsparametern möglich. Rückmeldungen aus klinischen Informationssystemen oder Telemedizinplattformen können frühzeitig Hinweise auf Veränderungen in Nutzung, Performance und Sicherheitsprofil liefern. Solche Feedback-Schleifen sind insbesondere für lernende und adaptive KI-Systeme essenziell, um Risiken früh zu erkennen und eine kontinuierliche Evaluation im klinischen Einsatz zu gewährleisten.

b. Evaluations-Frameworks international harmonisieren und an KI-spezifische Anforderungen anpassen

Langfristig erfordert die Evaluation von KI-MT eine harmonisierte Evidenzlogik, die technische Leistungsnachweise, klinische Wirksamkeit und ethische Anforderungen strukturiert zusammenführt. Abgestimmte Bewertungsmodelle müssen die regulatorischen Mindestanforderungen (MDR, IMDRF) mit der gesundheitsökonomischen Nutzenbewertung (zum Beispiel NICE, ICER) verbinden. Auf dieser Grundlage lassen sich internationaler Vergleich-

barkeit, transparente Entscheidungsprozesse und eine geteilte Verantwortung für den Einsatz lernender KI-MT im Gesundheitswesen realisieren.

Damit wird in Zukunft auch eine schnellere Dissemination von Innovationen ermöglicht. Entwickler:innen können sich auf optimierte Funktion statt Bürokratieerfüllung fokussieren und Gesundheitssysteme auf effektive und effiziente Leistungserbringung statt auf Pilotierungen ohne nachhaltige Wirkung.

c. Evidenz als gemeinsame Sprache verstehen

Evidenz für KI-MT muss den Nutzen aus Sicht aller beteiligten Akteure über den gesamten Lebenszyklus hinweg abbilden. Nur dann kann sie helfen, KI-basierte Medizintechnik im Sinne aller Stakeholder einzusetzen. Das bedeutet auch, dass Zulassung und Wirksamkeitsprüfung nicht am Markteintritt enden dürfen, sodass Nutzenanforderungen von Kostenträgern bis zu Patient:innen berücksichtigt werden. Da KI-Systeme im Einsatz weiterlernen, muss die Evaluation über den gesamten Lebenszyklus durch kontinuierliche Datenerhebung, PMS und RWE fortgesetzt werden.

d. Wer Transfer will, muss auch Strukturen schaffen

Der Blick ins internationale Umfeld zeigt, dass unterstützende institutionelle Strukturen bei Zulassungs- und Bewertungsbehörden einen entscheidenden Einfluss auf die Geschwindigkeit und Qualität des Transfers von KI-MT haben. Regulatorische Instrumente wie die im AI Act vorgesehenen AI Sandboxes sind hierfür ein wichtiger Baustein, reichen jedoch allein nicht aus, da sie primär Übergangsphasen adressieren. Für einen nachhaltigen Transfer ist es erforderlich, längere translationsorientierte Entwicklungsphasen zu ermöglichen, in denen KI-Systeme unter Forschungsbedingungen reifen können, bevor sie in regulierte Versorgungsprozesse überführt werden. Dies setzt Behördenstrukturen voraus, die auf die Bedarfe klein- und mittelständischer Unternehmen und Startups sowie universitärer und nicht-industrieller Akteure ausgerichtet sind und Beratung, Transparenz und Proportionalität im Umgang mit regulatorischen Anforderungen gewährleisten. Nur so lassen sich Innovationspotenziale systematisch in Versorgung überführen, ohne dass Forschung an regulatorischer Komplexität scheitert.

Herausgeber:
VDI/VDE Innovation + Technik GmbH
Steinplatz 1 | 10623 Berlin
www.vdivde-it.de

Bildnachweis:
elenabs/istockphoto

VDI|VDE|IT