UGESKRIFT FOR LÆGER

Statusartikel

AI-etik i sundhedsvæsenet

Anne Gerdes[1], Iben Fasterholdt[2] & Benjamin S.B. Rasmussen[1, 3]

1) CAI-X – Centre for Clinical Artificial Intelligence, Odense Universitetshospital, 2) CIMT – Center for Innovativ Medicinsk Teknologi, Odense Universitetshospital, 3) Radiologisk Afdeling, Odense Universitetshospital

Ugeskr Læger 2024;186:V09230600. doi: 10.61409/V09230600

HOVEDBUDSKABER

- Kunstig intelligens (AI) betragtes som en løsning på ressourceudfordringer i sundhedsvæsenet.
- Hvis AI skal bruges intelligent i sundhedsvæsenet, er det en ressourcekrævende øvelse, der fordrer etisk indsigt og involvering af kliniske domæneeksperter samt patienter.
- Forventningspresset til AI risikerer at hæmme grundlaget for etisk ansvarlig udvikling og ibrugtagning.

Kunstig intelligens (AI) har potentiale til at forbedre diagnostik og behandling og forventes også at kunne afhjælpe konsekvenserne af personalemangel i sundhedsvæsenet. For at indfri de høje forventninger til AI-beslutningsunderstøttende værktøjer kræves en betydelig interdisciplinær indsats, hvor patienten og den kliniske domæneekspert er centralt placeret.

Vi står over for et komplekst dilemma, hvor der er modsatrettede hensyn mellem pres for hurtig implementering af AI i sundhedsvæsenet og behovet for grundig udvikling og testning af AI-værktøjer før ibrugtagning. Etikken spiller en afgørende rolle til sikring af, at AI-anvendelser i sundhedsvæsenet er retfærdige, transparente og respekterer patienternes rettigheder samt privathed. Patienten skal være i centrum for anvendelse af AI i sundhedsvæsenet, og beslutninger, der påvirker behandling og sundhed, skal tages med afsæt i patientens interesse.

På den baggrund giver denne artikel overblik over etiske udfordringer i tilknytning til udvikling og ibrugtagning af AI i sundhedsvæsenet. Ligeledes præsenteres og diskuteres forskellige guidelines og standarder, der proaktivt søger at imødegå etiske udfordringer.

Selv inden for lovende AI-anvendelsesområder som billeddiagnostik viser systematiske reviews [1, 2], at mange studier om AI-værktøjer, der performer bedre end klinikere, ikke er velunderbyggede. Heldigvis er der også fremskridt at spore, som i [3]. Men manglen på prospektive og randomiserede kliniske undersøgelser (RCT'er) og standardiseret rapportering samt inddragelse af underrepræsenterede grupper udgør udfordringer, der understreger betydningen af etik og inklusion i AI-udviklingen inden for sundhedsvæsenet [4].

Det er vigtigt at sikre, at AI-værktøjer er nøje afstemt med de etiske principper, der er relevante for patientpleje og diagnostik, og at de ikke fører til forringelse af patientens sundhed eller kliniske resultater. Derfor er det afgørende at afstemme AI-optimismen med et kritisk perspektiv for at fremme en tilgang til AI i sundhedsvæsenet, hvor patientens velbefindende og etiske principper prioriteres for at sikre ansvarlig og bæredygtig anvendelse af AI i patientpleje og behandling.

Kunstig intelligens i sundhedsvæsenet – etiske udfordringer

AI kan betragtes som en samlebetegnelse, som dækker over systemer, der kan analysere, lære og træffe beslutninger og genererer nyt på samme måde, som vi mennesker gør. AI kan i sundhedsvæsenet alt afhængigt af kontekst fungere med forskellige grader af autonomi fra diagnostisk beslutningsstøtte til brugen af sprogmodeller.

Der er en række etiske problemstillinger i tilknytning til anvendelse af AI-beslutningsstøtteværktøjer i sundhedsvæsenet [5]. F.eks. kan AI-værktøjer være baseret på mangelfulde eller skævvredne datasæt og dermed give anledning til fejlbehæftede output, der diskriminerer bestemte patientgrupper. Et eksempel er, når et dermatologisk AI-værktøj til detektering af malignt melanom trænes på billedmateriale med overvejende lyse hudtyper og derved underdiagnosticerer andre hudtyper [6]. Tilsvarende kan AI-værktøjer bidrage til at forstærke ulighed i sundhed, som det påvises i [7], hvor et amerikansk AI-værktøj fejlagtigt tilskriver en for lav risikoscore til patientgrupper med mørke hudtyper, der påviseligt er mere syge end tilsvarende patienter med lyse hudtyper. Problemet opstår, fordi algoritmen baserer sig på en prædiktion af sundhedsudgifter i stedet for sygdom. Et sådant greb betragtes traditionelt som et effektivt pejlemærke, dvs. en proxy, for præcision i prædiktioner. Men dette gælder ikke for alle patientgrupper, da der historisk grundet systemiske fordomme har været store forskelle i behandlingstilbud i det amerikanske sundhedsvæsen. Når et generelt effektivt pejlemærke anvendes ukritisk, kan dette greb afføde bias. Men det er også en kompleks udfordring at afdække sociale determinanters betydning for helbredet og dernæst afgrænse datasæt samt udvikle modeller til AI-værktøjer, der afhjælper ulighed i sundhed og understøtter inklusion. AI-værktøjer kan således utilsigtet gentage og forstærke forskelsbehandling i sundhedsvæsenet. Endelig kan det, selv med grundighed og de bedste intentioner, være vanskeligt på forhånd at afgøre, hvordan skævvredne modeller kan undgås.

Når AI-værktøjer bidrager med input til beslutninger, er det et afgørende legitimitetshensyn, at

AI-bidraget er gennemskueligt, og ansvar kan placeres, ligesom de fleste sætter højere etiske standarder for AI-systemer end mennesker og f.eks. kræver, at AI systemer skal være langt mindre fejlbarlige end klinikere [8]. Transparens betragtes derfor som væsentligt for pålidelig AI, hvilket afspejles i lovgivning og reguleringstiltag [9, 10] samt guidelines [11-13]. Her fremhæves, at vi skal kunne forstå, hvad der foregår i en black-box-model for at kunne have tillid til et AI-systems bidrag til beslutninger.

Det er især vigtigt, at AI-værktøjer ikke giver korrekte svar på et fejlagtigt grundlag. Det kan forekomme, hvis et betydningsløst mærke på røntgenbilleder i et datasæt tilfældigvis medvirker til at generere et korrekt output.

Til fremme af transparens ses »explainable AI« (XAI) som en potentiel løsning [14-16]. Her kan benyttes »black-box explainers«, dvs. modeller, der f.eks. post hoc præsenterer en forklaring af en kompleks models output [17, 18]. Sådanne forklaringsmodeller gengiver ikke 1:1 komplekse modellers indre virkningsmekanismer, eftersom disse ikke afspejler menneskelige måder at ræsonnere på. Men hvis et intransparent AI-værktøj er grundigt empirisk valideret efter RCT-forbillede, så kan en idealiseret eller tilnærmet forklaringsmodel dog siges at tilvejebringe et nødvendigt legitimeringsgrundlag, der understøtter klinikerens ansvar for behandlingen samt bidrager til et informeret beslutningsgrundlag i samråd med patienten.

Omvendt hævder nogle [19, 20], at eftersom XAI er en stående udfordring, så bør AI-performance have forrang over et forlangende om forklarlighed. Sat på spidsen hævder argumentet, at det er vigtigere at redde liv end at kunne tilvejebringe en forklaring af en black-box-model. Imidlertid er det svært at forestille sig og tillige velkendt fra studier [21], at klinikere og patienter vil finde det legitimt, hvis AI-systemer anbefaler en behandling alene med henvisning til, at et AI-system er valideret og pålideligt. Således præsenterer [22] et scenarie, hvor et intransparent, men valideret til at være diagnostisk præcist AI-system rangordner flere behandlingsmuligheder for en patient med brystkræft og anbefaler, at begge bryster fjernes kirurgisk. Her er det ikke tilstrækkeligt blot at vide, at AI-systemet er valideret og pålideligt. Uden yderligere information om, hvordan systemet kom frem til anbefalingen, vil klinikeren i samråd med patienten ikke på ansvarlig vis kunne træffe en beslutning vedrørende behandling.

En kombination af empirisk validerede AI-værktøjer og forklaringsmodeller kan derfor udgøre et tilstrækkeligt legitimeringsgrundlag for ibrugtagning af AI-beslutningsstøttesystemer. Imidlertid er der en konflikt mellem presset for at implementere AI i sundhedsvæsenet versus behovet for omfattende og tidskrævende empirisk validering af sundhedsteknologier, herunder AI-værktøjer. Derfor skal det understreges, at validering af AI-modeller kræver omhyggelige evalueringer og, under implementering af AI, et stramt kontrolregime.

Standarder og guidelines til etisk ansvarlig udvikling

Der findes et væld af tilgange, der anviser, hvordan man proaktivt adresserer etiske

problemstillinger og sikrer brugerinddragelse i forbindelse med udvikling, ibrugtagning og evaluering af AI-systemer. På EU-plan er der udviklet retningslinjer og såkaldte Ethics by Design-guidelines med afsæt i grundlæggende rettigheder og etiske principper om respekt for autonomi, forebyggelse af skade, retfærdighed og forklarlighed [12]. Herunder udmøntes krav til udvikling af pålidelig AI, der omfatter krav om menneskelig kontrol, robusthed og sikkerhed, sikring af privathed og datastyring, gennemsigtighed, diversitet, ikkediskrimination, retfærdighed samt ansvarlighed [11]. Tilsvarende fremmer AI-forordningen disse krav, ligesom forordningen indeholder et appendiks med krav til teknisk dokumentation for AI-systemer [10]. WHO fremhæver ligeledes en række værdibaserede designmetoder [13]. Men værdibaserede tilgange er ikke i sig selv nogen garanti for succes. F.eks. har en af de fremhævede tilgange, value sensitive design, kun i ringe omfang bidraget til egentlig systemudvikling [23]. Ofte finder systemudviklere sådanne tilgange alt for overordnede og foretrækker rene tekniske tilgange til f.eks. biashåndtering og privathedsfremmende design [24]. Men når etiske udfordringer alene ses i en teknisk optik, nedtones sociotekniske problemstillinger og kontekstuelle faktorers betydning i klinisk beslutningstagen.

Derfor er den største udfordring for værdibaserede designtilgange og Ethics by Design-guidelines, at der kræves interdisciplinært samarbejde med inddragelse af humanistiske såvel som tekniske samfunds- og sundhedsfaglige kompetencer. I en række systemudviklingssammenhænge kan man indskyde en brobyggerfunktion, der kan bistå i afstemningen af de mange forskelligartede behov og hensyn. Men i kliniske sammenhænge er det desuden nødvendigt, at kliniske domæneeksperter er tæt knyttet til AI-systemudviklingen.

Patienternes inddragelse i AI-udviklingsprocessen er ligeledes af afgørende betydning. Deres perspektiver kan medvirke til at identificere potentielle etiske udfordringer og bidrage til en retfærdig og ansvarlig udvikling og implementering af AI i sundhedsvæsenet. Patientinddragelse kan resultere i, at patientens behov bedre tilgodeses, og at offentlighedens tiltro til forskningen stiger [25]. Ligeledes fremhæver [26] effekt på forskningsprocessen i form af påvirkning af forskningsspørgsmål, og at patientrelevante outcomes medtages. *Vogsen et al* [27] nævner forbedring af patientinformation og patientrekruttering som konsekvens af patientinddragelse, mens *Karlsson et al* [28] finder, at patientinddragelse kan løfte kvaliteten af projektet og være med til at give ny viden til forskerne samt en større forståelse for patient- og pårørendeperspektivet. Ud over de nævnte gevinster argumenterer [25] for, at offentligheden bør inddrages ud fra demokratiske og moralske bevæggrunde – udvikling uden dem, det handler om, anses her for uetisk.

Der foreligger desuden en række standarder til dokumentation af kliniske studier, som har tilføjet AI-guidelines, såsom The STARD-AI protocol [29] og The CONSORT-AI extension [15]. Sådanne standarder fremmer transparent og pålidelig AI ved at anskueliggøre kvalitetskrav til arbejdet med data samt træning og validering af modeller. Desuden gives der overblik og guidelines på sitet »Future AI: Best Practices for trustworthy AI in medicine« [30]. Slutteligt kan det nyligt udviklede danske Model for ASsessing the value of AI-evalueringsredskab [16] give administrative og

sundhedsprofessionelle ledere en holistisk vurdering af værdien af nye AI-løsninger.

Konklusion

Siden de tidlige 00'ere har vi set en række succesfulde AI-gennembrud båret frem af lettilgængelige enorme datamængder på nettet, forøget computerregnekraft og øget lagringsplads samt gamle og nye teknikker inden for AI. Sådanne succeser kan ikke umiddelbart transformeres til kritiske domæner med komplekse beslutningsprocesser. Det er ressourcekrævende at udvikle og anvende AI i højrisikoområder, hvor fejl er fatale, og data en sparsom ressource.

Ved at integrere kliniske domæneeksperter, patienter og etiske principper i udviklingen og implementeringen af AI i sundhedsvæsenet kan vi skabe pålidelige AI-løsninger, der opfylder de virkelige behov i patientpleje og behandling.

Korrespondance *Benjamin S.B. Rasmussen*. E-mail: bsr@rsyd.dk

Antaget 31. maj 2024

Publiceret på ugeskriftet.dk 8. juli 2024

Interessekonflikter Der er anført potentielle interessekonflikter. Forfatternes ICMJE-formularer er tilgængelige sammen med artiklen på ugeskriftet.dk

Referencer findes i artiklen publiceret på ugeskriftet.dk. Fuld referenceliste kan fås ved henvendelse hos forfatterne

Artikelreference Ugeskr Læger 2024;186:V09230600

doi 10.61409/V09230600

SUMMARY

AI ethics in healthcare

Artificial Intelligence (AI) holds promise in improving diagnostics and treatment. Likewise, AI is anticipated to mitigate the impacts of staff shortages in the healthcare sector. However, realising the expectations placed on AI requires a substantial effort involving patients and clinical domain experts. Against this setting, this review examines ethical challenges related to the development and implementation of AI in healthcare. Furthermore, we introduce and discuss various approaches, guidelines, and standards that proactively aim to address ethical challenges.

REFERENCER

1. Nagendran M, Chen Y, Lovejoy CA et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689.

https://doi.org/10.1136/bmj.m689

2. Freeman K, Geppert J, Stinton C et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ. 2021;374:n1872. https://doi.org/10.1136/bmj.n1872
3. Lång K, Josefsson V, Larsson A-M et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. Lancet Oncol. 2023;24(8):936-44. https://doi.org/10.1016/S1470-2045(23)00298-X
4. Plana D, Shung DL, Grimshaw AA et al. Randomized clinical trials of machine learning interventions in health care: a systematic review. JAMA Netw Open. 2022;5(9):e2233946. https://doi.org/10.1001/jamanetworkopen.2022.33946
5. Lekadir K, Quaglio G, Garmendia TA, Gallin C. Artificial intelligence in healthcare: applications, risks, and ethical and societal impacts, 2022. https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729512/EPRS_STU(2022)729512_EN.pdf (3. apr 2024).
6. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol. 2018;154(11):1247-1248. https://doi.org/10.1001/jamadermatol.2018.2348
7. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453. https://doi.org/doi:10.1126/science.aax2342
8. Lenskjold A, Nybing JU, Trampedach C et al. Should artificial intelligence have lower acceptable error rates than humans? BJR Open. 2023;5(1):20220053. https://doi.org/10.1259/bjro.20220053
9. European Parliament, Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). OJ L. 119, 32016R0679, 4.5.2016:1-88. http://data.europa.eu/eli/reg/2016/679/oj (3. apr 2024).
10. European Commission. Proposal for a regulation of the European Parliament and the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. EUR-Lex – 52021PC0206, 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELLAR:e0649735-a372-11eb-9585-01aa75ed71a1 (3. apr 2024).
11. High-Level Expert Group on Artificial Intelligence set up by the European Commission. Ethics guidelines for trustworthy AI, 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (3. apr 2024).
12. Dainow B, Brey P. Ethics By Design and Ethics of Use Approaches for Artificial Intelligence. Secondary Ethics By Design and Ethics of Use Approaches for Artificial Intelligence 2021. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf (3. apr 2024).
13. Ethics and Governance of Artificial Intelligence for Health: WHO guidance. Secondary Ethics and Governance of Artificial Intelligence for Health: WHO guidance 2021. https://www.who.int/publications/i/item/9789240029200 (3. apr 2024).
14. Gunning D, Vorm E, Wang JY, Turek M. DARPA's explainable AI (XAI) program: a retrospective. Applied AI letters. 2021;2(4):e61. https://doi.org/10.1002/ail2.61

15. Liu X, Rivera SC, Moher D et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digital Health. 2020;2(10):e537-e548. https://doi.org/10.1016/S2589-7500(20)30218-1
16. Fasterholdt I, Kjølhede T, Naghavi-Behzad M et al. Model for ASsessing the value of Artificial Intelligence in medical imaging (MAS-AI). Int J Technol Assess Health Care. 2022;38(1):e74. https://doi.org/10.1017/S0266462322000551
17. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. arXiv:1705.07874. https://doi.org/10.48550/arXiv.1705.07874
18. Gerdes A. Dialogical guidelines aided by knowledge acquisition: enhancing the design of explainable interfaces and algorithmic accuracy. I: Arai K, Kapoor S, Bhatia R, red. Proceedings of the future technologies conference (FTC) 2020, Volume 1. Springer, 2021:243-57.
19. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-e750. https://doi.org/10.1016/S2589-7500(21)00208-9
20. London AJ. Artificial intelligence and black-box medical decisions accuracy versus explainability. Hastings Cent Rep. 2019;49(1):15-21. https://doi.org/10.1002/hast.973
21. Ploug T, Sundby A, Moeslund TB, Holm S. Population preferences for performance and explainability of artificial intelligence in health care: choice-based conjoint survey. J Med Internet Res. 2021;23(12):e26611. https://doi.org/10.2196/26611
22. Bjerring JC, Busch J. Artificial intelligence and patient-centered decision-making. Philos Technol. 2021;34(2):349-71. https://doi.org/10.1007/s13347-019-00391-6
23. Gerdes A, Frandsen TF. A systematic review of almost three decades of value sensitive design (VSD): what happened to the technical investigations? Ethics Inf Technol. 25(26):2023. https://doi.org/10.1007/s10676-023-09700-2
24. FAT/ML. Fairness, accountability, and transparency in machine learning, 2022. https://www.fatml.org/ (3. apr 2024).
25. Blackburn S, Clinch M, de Wit M et al. Series: Public engagement with research. Part 1: the fundamentals of public engagement with research. Eur J Gen Pract. 2023;29(1):2232111. https://doi.org/10.1080/13814788.2023.2232111
26. Bird M, Ouellette C, Whitmore C et al. Preparing for patient partnership: a scoping review of patient partner engagement and evaluation in research. Health Expect. 2020;23(3):523-539. https://doi.org/10.1111/hex.13040
27. Vogsen M, Geneser S, Rasmussen ML et al. Learning from patient involvement in a clinical study analyzing PET/CT in women with advanced breast cancer. Res Involv Engagem. 2020;6(1):1. https://doi.org/10.1186/s40900-019-0174-y
28. Karlsson AW, Kragh-Sørensen A, Børgesen K et al. Roles, outcomes, and enablers within research partnerships: a rapid review of the literature on patient and public involvement and engagement in health research. Res Involv Engagem. 2023;9(1):43. https://doi.org/10.1186/s40900-023-00448-z
29. Sounderajah V, Ashrafian H, Golub RM et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. BMJ Open. 2021;11(6):e047709. https://doi.org/10.1136/bmjopen-2020-047709
30. Future AI. FUTURE-AI: Best practices for trustworthy AI in medicine, 2024. https://future-ai.eu/ (2. apr

2024).

