Leidraad voor kwalitatieve diagnostische en prognostische toepassingen van AI in de zorg

Versie 1.1 16-08-2023

Auteurs

Maarten van Smeden, Carl Moons, Lotty Hooft (fase 1 t/m 3)

Ilse Kant, Hine van Os, Niels Chavannes (fase 4 t/m 6)

Namens de werkgroepleden Medische AI

In opdracht van het ministerie van Volksgezondheid, Welzijn en Sport

Waardevolle AI

Inhoud

Dankwoord

In 2020 is door het Universitair Medisch Centrum Utrecht (UMCU) en Leiden Universitair Medisch Centrum (LUMC) een literatuuronderzoek uitgevoerd en een rapport geschreven waarin een overzicht is gemaakt van de beschikbare nationale en internationale richtlijnen en criteria voor de ontwikkeling, validatie, evaluatie en implementatie van een *Artificial Intelligence Prediction Algorithm* (AIPA) in de medische sector, inclusief de publieke gezondheidszorg. Op basis van dit literatuuronderzoek en de daaruit volgende werkgroepen hebben het UMCU en LUMC een leidraad opgesteld met eisen en criteria voor ontwikkeling, validatie, evaluatie en implementatie van een AIPA in de zorg.

Deze leidraad is tot stand gekomen door de inzet van vele betrokkenen. In het bijzonder danken wij *Rosalie van Oostrom* en *Pieter Boone* van VWS voor het organiseren en coördineren van de totstandkoming van deze leidraad, *Roy Tomeij* voor het voorzitten van de werkgroepen en *Rachel Peeters* voor de ondersteuning tijdens en na de werkgroepen. Wij danken alle actieteamleden en de werkgroepleden voor hun actieve deelname, inzet en betrokkenheid in de werkgroepen die hebben geleid tot deze leidraad. Daarnaast danken wij alle reviewers en de deelnemers aan de praktijktoetsen voor de vele feedbackcommentaren die hebben geleid tot een verbetering van de conceptversies, en KPMG voor het coördineren en organiseren van dit proces in samenwerking met de redactieraad. Verder danken we de NEN Nederlandse AI Medical Device expert Group, Patiëntenfederatie Nederland en Inspectie Jeugd en Gezondheid voor hun input.

<u>Actieteamleden</u>

Jan Jaap Baalbergen (NFU), Robert Geertsma (RIVM, Dennis Japink (ZN), Carl Moons (UMCU), Rozemarijn Pennings (InEen), Marlies Schijven (Amsterdam UMC), Jaap Schrieke (GGZ Nederland), Inge Steinbuch (ActiZ), Jos Schimmelpennink (Nederlandse Vereniging van Ziekenhuizen), Stefan Visscher (Federatie Medisch Specialisten) en Laurine Keulemans (Ministerie van VWS).

<u>Werkgroepleden</u>

Fase 1: Paul Agra, Amy Eikelenboom, Christian van Ginkel, Martine de Vries, Saskia Haitjema, Andre Dekker.

Fase 2: Daniel Oberski, Desy Kakiay, Kicky van Leeuwen, Joran Lokkerbol, Evangelos Kanoulas, Gabrielle Davelaar.

Fase 3: Wouter Veldhuis, Bart-Jan Verhoeff, Vincent Stirler, Daan van den Donk, Huib Burger.

Fase 4: Giovanni Cina, Martijn van der Meulen, Maurits Kaptein, Floor van Leeuwen, Egge van der Poel, Marcel Hilgersom.

Fase 5: Teus Kappen, Sade Faneyte, René Verhaart, Jonas Teuwen, Ewout Steyerberg, Leo Hovestadt, René Drost.

Fase 6: Anne de Hond, Bart Geerts, Nynke Breimer, Karen Wiegant, Laure Wynants, Lysette Meuleman.

Reviewers (op alfabetische volgorde)

Annemarie van 't Veen, Charlotte Brouwer, Daniel Vijlbrief, Elise Quik, Jan-Jaap Visser, Jan-Kees van Wijnen, Jan-Willem Wasmann, Jean-Paul Kleijnen, Joris van Dijk, Leon Doorn, Lieke Poot, Maaike van Mourik, Mark Scheper, Martin van Buuren, Merel Huisman, Richard Bartels, Rimmert Brandsma, Rob Tolboom, Roel Streefkerk, Roel van Est, Wouter Bulten.

Praktijktoets deelnemers:

RetCAD, Thirona - Mark van Grinsven
HUME, Mentech – Reon Smits en Erwin Meinders
U-Prevent, Ortec – John Jacobs
Risicotaxatie agressie in de psychiatrie, UMCU – Karin Hagoort
Covid-19 severity score, Maasstad ziekenhuis – Sade Faneyte

Redactieraad

Maarten van Smeden, Carl Moons, Lotty Hooft, Ilse Kant, Hine van Os, Niels Chavannes, onder begeleiding van Ylja Remmits en Alexander Boer (KPMG).

Introductie

Auteurs

Maarten van Smeden, Ilse Kant, Alexander Boer

Namens de werkgroepleden medische AI

Status van dit document

Deze leidraad is een uitdrukking van wat er in het werkveld als goed professioneel handelen wordt beschouwd bij het ontwikkelen, toetsen en toepassen van een *Artificial Intelligence Prediction Algorithm* (AIPA) in de medische sector inclusief publieke gezondheidszorg. De mate waarin zij dwingend is wordt door het werkveld bepaald. Aan het naleven van de leidraad kunnen derhalve geen rechten worden ontleend. De ambitie is dat deze leidraad als breed gedragen veldnorm wordt geaccepteerd.

In voorkomende gevallen bespreekt de leidraad verplichtingen uit van toepassing zijnde wet- en regelgeving, bijvoorbeeld uit de *Algemene Verordening Gegevensbescherming* (AVG), of eisen die in de *Medical Device Regulation* (MDR) of *In-Vitro Diagnostic Medical Device Regulation* (IVDR) gesteld worden aan medische hulpmiddelen (*medical devices*). De leidraad probeert geen uitputtende opsomming te zijn van toepasselijke wet- en regelgeving, of van richtlijnen en ISO-normen die gebruikelijk zijn in medische informatietechnologie of medische hulpmiddelen sector. De leidraad is dus een aanvulling op bestaande wet- en regelgeving, richtlijnen en normen op basis van wat in het werkveld als goed professioneel handelen wordt beschouwd. De leidraad is ook geen uitwerking van het liggende voorstel voor een verordening voor kunstmatige intelligentie van de Europese Commissie en de documentatieverplichtingen die daaraan verbonden zullen zijn (in bijlage IV van dat voorstel) of het wetsvoorstel Elektronische gegevensuitwisseling in de Zorg (Wegiz) dat de mogelijkheid zal creëren organisaties te verplichten bepaalde data elektronisch uit te wisselen. Beide voorstellen worden zeer relevant geacht voor de toekomst van het werkveld. In samenvatting, deze leidraad moet als aanvulling worden gezien op bestaande wet- en regelgeving, richtlijnen en normen.

De inrichting van toezicht op de ontwikkeling, toetsing en toepassing van het AIPA in het algemeen, valt buiten het toepassingsgebied van de leidraad. De leidraad is geen recept, geen toetsingsinstrument en geen risicoanalyse-instrument. De leidraad gaat in de eerste plaats over goed professioneel handelen, en niet over hoe een organisatie in de diverse omgevingen waarin de leidraad gebruikt kan worden goed professioneel handelen vorm kan geven en controleren. Voor de uiteindelijke keuze voor de vorm van het toezichtproces en de documentatie is het van groot belang of het AIPA bijvoorbeeld onderdeel wordt van een medisch hulpmiddel als bedoeld in de MDR, en of deze in de Europese Unie in de handel wordt gebracht (art. 5 MDR), of binnen een enkele zorginstelling wordt vervaardigd en gebruikt (de uitzondering in art. 5 lid 5 MDR). Ook is van groot belang in welke risicoklasse van de MDR het hulpmiddel zal gaan vallen (volgens regel 11 van Annex VIII van de MDR). De leidraad laat deze vragen open. De leidraad bevat aanbevelingen die voor markttoelating conform de MDR feitelijk dwingend zullen zijn. In andere gevallen verruimt de leidraad

bestaande vereisten naar een breder toepassingsbereik. Om de bruikbaarheid van de leidraad te verhogen wordt er waar toepasselijk geregeld verwezen naar de deze reeds bestaande wet- en regelgeving.

Toepassingsbereik

Deze leidraad is van toepassing op het ontwikkelen, toetsen en toepassen van een AIPA die deel uit maak van een hulpmiddel bedoeld voor gebruik in de gezondheidszorgverlening, waaronder tevens de thuis- en zelfzorg vallen. Onder hulpmiddelen bedoeld voor gebruik in de gezondheidszorgverlening verstaan we tenminste medische hulpmiddelen als bedoeld in de MDR. Het hulpmiddel kan zelfstandige software zijn, of een hulpmiddel dat software bevat. Voor een definitie van AI wordt verwezen naar art. 3(1) van het voorstel voor een verordening voor kunstmatige intelligentie van de Europese Commissie of de uitgebreidere, maar inhoudelijk vergelijkbare, definitie van 18 december 2018 van de *AI High Level Expert Group on Artificial Intelligence* van de Europese Commissie in box 1.

Box 1: Definitie AI van de AI HLEG van 18 december 2018

"Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.

As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)."

Onder de term AIPA wordt in deze leidraad verstaan:

Algoritmen die leiden tot een voorspelling van een gezondheidsuitkomst bij individuele personen. Dit betreft tenminste het voorspellen van de kans op, of classificatie van, het hebben (diagnostisch) of het in de tijd optreden (prognostisch) van gewenste of ongewenste gezondheidsuitkomsten.

Diagnostische AIPA's voorspellen de kans op bijvoorbeeld het hebben van een aandoening of ziekte bij bepaalde symptomen of klachten, of de kans op het hebben van een

onderliggende aandoening zonder dat men enige symptomen of klachten heeft bij individuele personen in de algemene bevolking (screening).

Prognostische AIPA's voorspellen de kans op het optreden in de tijd van gezondheidsuitkomsten bij, bijvoorbeeld, patiënten met een bepaalde aandoening of ziekte, of ze voorspellen de kans op het moeten ondergaan van bepaalde behandeling of ziekenhuisopname, of ze voorspellen bij individuele personen in de algemene bevolking of ze op termijn een bepaalde aandoening of bepaalde kwaliteit van leven zullen ontwikkelen.

De door het AIPA voorspelde gezondheidsuitkomsten betreft gezondheidsuitkomsten bij de *individuele* patiënt, cliënt of burger, maar kan ook gezondheidsuitkomsten bij derden omvatten, bijvoorbeeld in de geestelijke gezondheidszorg lijdensdruk bij familieleden of mantelzorgers. Toepassingen van AI met een andere functie dan het voorspellen van *individuele* gezondheidsuitkomsten vallen buiten het toepassingsbereik van deze leidraad, zoals bijvoorbeeld navigatietoepassingen in de robotica, toepassingen die op populatieniveau patiëntenstromen voorspellen ten behoeve van capaciteitsplanning, of toepassingen die voor puur beschrijvende classificatie en segmentatie ingezet worden zonder dat daarmee direct een diagnostische of prognostische functie gediend wordt. Wel kan een AIPA in voorkomende gevallen deel uitmaken van dergelijke toepassingen. Alleen het AIPA valt dan binnen het toepassingsbereik van de leidraad. Software voor het *verstrekken van informatie* die gebruikt wordt bij het nemen van beslissingen op basis van een AIPA voor bijvoorbeeld, diagnostische, prognostische, therapeutische or preventieve doeleinden, inclusief leefstijlaanpassing, zoals bedoeld in de eerdergenoemde regel 11, valt in de regel binnen het toepassingsbereik van de leidraad.

In de leidraad wordt soms van een *medische context* gesproken, waarmee alle denkbare context en interacties in de gezondheidszorgverlening worden bedoeld, ongeacht of er sprake is van een tussenkomst van een zorgverlener of zorginstelling. In deze leidraad omvat de term medische context zowel de cure, care als preventiesector, ofwel toepassing van een AIPA in de nulde, eerste, tweede en derde lijn (inclusief thuis- en zelfzorg). Daar waar expliciet gesproken wordt van medisch handelen wordt wel verwezen naar voorbehouden handelingen als bedoeld in de *Wet op de beroepen in de individuele gezondheidszorg*.

Betrokken partijen

De eisen en aanbevelingen in de leidraad zijn direct geadresseerd aan ontwikkelaars en testers van het AIPA, de fabrikant van de software waar het AIPA deel van uit maakt, en de zorgorganisatie die die software implementeert in de organisatie. De beoogde lezers zijn fabrikanten van hulpmiddelen waar een AIPA deel van uitmaakt, onderzoekers die een AIPA

ontwikkelen en testen, zorgorganisaties en zorgverleners die dergelijke hulpmiddelen inkopen en inzetten, en instanties die de kwaliteit, inzetbaarheid en vergoeding van het AIPA mede bepalen. De leidraad benoemt wat zorgverleners, burgers, patiënten, patiëntvertegenwoordigers, verzekeraars en beleidsmakers (zoals het Zorginstituut en de Nederlandse Zorgautoriteit) van een AIPA ontwikkelaar of fabrikant mogen verwachten, als zij dergelijke hulpmiddelen inkopen, toepassen of als het op hen toegepast wordt.

Onder de AIPA-*ontwikkelaar* of -*tester* verstaan we de persoon die beroepshalve of vrijwillig bij de ontwikkeling en testen van het AIPA betrokken is en goed professioneel handelen nastreeft, bijvoorbeeld onderzoekers, datamanagers, dataleveranciers, ontwikkelaars en data scientists.

Onder de *fabrikant* verstaan we een eventuele (rechts)persoon die het hulpmiddel met software die een AIPA bevat vervaardigt of volledig reviseert, of dat laat doen, en het onder zijn naam of merk verhandelt, als gedefinieerd in de MDR. De aanwezigheid van een fabrikant als bedoeld in de MDR is niet maatgevend voor de toepasbaarheid van deze leidraad, maar de ontwikkelaar kan wel aan de plichten van de fabrikant gebonden worden.

Onder de *zorgorganisatie* verstaan we een eventuele rechtspersoon die het hulpmiddel dat het AIPA bevat beschikbaar stelt aan eindgebruikers en verplichtingen heeft jegens die eindgebruikers. De ontwikkelaar of tester van het hulpmiddel kan zijn werk namens de zorgorganisatie uitvoeren die het AIPA zelf gaat inzetten. In dat geval ontbreekt een fabrikant en vallen de rollen van fabrikant en zorgorganisatie voor de interpretatie van de eisen en aanbevelingen samen. De zorgorganisatie is meestal een zorgaanbieder waarin zorgverleners zorg verlenen, bijvoorbeeld een ziekenhuis, verpleeghuis, GGZ-instelling of eerstelijnspraktijk voor verblijf en behandeling, maar kan ook een welzijnsorganisatie of een gemeente zijn. De precieze invulling van het begrip wordt opzettelijk opengelaten.

Onder de *zorgverlener* verstaan we de persoon die, beroepshalve of vrijwillig, zorg verleent en daarbij het hulpmiddel dat het AIPA bevat toepast of gebruikt als eindgebruiker. De zorgverlener kan ook een individuele zorgaanbieder zijn als de zorgverlener niet namens een organisatie zorg aanbiedt, bijvoorbeeld een individuele huisarts, tandarts of psycholoog. In dit geval ontbreekt een zorgorganisatie als tussenliggende partij.

Onder de *patiënt, cliënt* of *burger* verstaan we de persoon waarop de voorspelling van het AIPA betrekking heeft. Deze persoon kan ook zelf direct de eindgebruiker van het hulpmiddel zijn dat het AIPA bevat. In dat geval spreken we in de regel van zelfzorg, zoals bijvoorbeeld in de (primaire) preventiesetting.

Onder *stakeholders* worden alle partijen en personen verstaan die betrokken zijn bij de ontwikkeling, validatie of gebruik van het AIPA of anderszins belanghebbende zijn. Hieronder

vallen alle bovenstaande categorieën zoals ontwikkelaars en gebruikers (bijv. zorgprofessionals, patiënten, burgers) en ook controlerende en toezichthoudende of certificerende partijen (bijv. privacydeskundigen, notified bodies, METCs) en de uiteindelijke doelgroep bij of voor wie de voorspellingen gedaan worden (bijv. patiënten en burgers, afhankelijk van de beoogde doelgroep van het AIPA).

Pas toe of leg uit

De leidraad maakt onderscheid tussen eisen en aanbevelingen voor goed professioneel handelen. De eisen worden aangegeven met **moet**. Aanbevelingen worden aangegeven met **aanbevolen** of **sterk aanbevolen**. Gebruik van de leidraad veronderstelt een *pas toe of leg uit* (*comply or explain*) benadering, waarbij de keuze om aanbevelingen wel of niet uit te voeren gebaseerd is op een risico-afweging die alleen met het oog op een specifieke toepassing van het AIPA gemaakt kan worden. Deze risico-afweging wordt expliciet gemaakt en is uitlegbaar aan derden. Met een goede uitleg kan de leidraad nageleefd worden, zonder dat alle aanbevelingen gevolgd worden. In sommige gevallen wordt bij de aanbeveling duidelijk gemaakt welke risico's of omstandigheden tenminste tot een keuze voor uitleg kunnen leiden.

Impact van de voorspelling op de patiënt, cliënt of burger is een belangrijke overweging bij het inschatten van risico's van het gebruik, inzetten en vergoeden van AIPA's. Ook bij een verwachte lage impact van het AIPA op de patiënt, cliënt of burger blijft goed professioneel handelen van belang, en blijven die delen van de leidraad die geen betrekking hebben op de rechten en plichten van eindgebruikers als leidraad voor goed professioneel handelen van toepassing. Dit is tenminste het geval als het AIPA geen deel uit gaat maken van een hulpmiddel.

Fasering

De leidraad is ingedeeld in zes fasen:

- Fase 1: Verzameling en beheer van de data
- Fase 2: Ontwikkeling van het AIPA
- Fase 3: Validatie van het AIPA
- Fase 4: Ontwikkeling van de benodigde software
- Fase 5: Effectbeoordeling van het AIPA in combinatie met de software
- Fase 6: Implementatie en gebruik van het AIPA met software in de dagelijkse praktijk.

De daarin veronderstelde chronologie is niet bedoeld als dwingend, en past lang niet altijd op de feitelijke of meest efficiënte volgorde van handelingen. Het dient wel als handvat voor intern toezicht, waarbij de fases kunnen dienen als structuur voor het organiseren van documentatie.

Een AIPA die slechts medisch-wetenschappelijk onderzoek als doel heeft, eindigt in de regel na fase 3. De leidraad biedt ook hiervoor een goede leidraad voor professioneel handelen. Fase 4 tot en met 6 betrekken nadrukkelijker de fabrikant van het hulpmiddel, de zorgorganisatie die het AIPA gaat inzetten en de eindgebruikers en belanghebbenden zoals de zorgverleners, patiënten en burgers.

Fase 0: Voorbereiding van het ontwikkelproces

Een ontwikkelproces gericht op inzet in gezondheidszorgverlening of zelfzorg begint in de praktijk niet bij voorbereiding en beheer van de gegevens in fase 1, maar bij de voorbereiding van de beslissing om het AIPA te gaan ontwikkelen en daar middelen voor in te gaan zetten. Daarom is het belangrijk in dit kader ook *Fase 0: Voorbereiding van het ontwikkelproces* te benoemen.

In fase 0 wordt in samenspraak met domeinexperts en eindgebruikers bepaald of het nodig is om een AIPA te ontwikkelen voor het beoogde probleem en de haalbaarheid van een idee voor ontwikkeling van het AIPA getoetst. Veelal zullen deze afwegingen gemaakt worden aan de hand van experimenten of een *proof-of-concept (PoC).* Ook wordt een eerste informele risicoanalyse verricht en een plan van aanpak gekozen in multidisciplinair verband, inclusief de benodigde risicobeheersingsmaatregelen en intern toezicht. Uiteindelijk kunnen dan de totale kosten en baten van het uitvoeren van het plan ingeschat worden.

Om de overwegingen die centraal staan in fase 0 op een gedegen manier te kunnen maken zal inzicht moeten worden verkregen in welke partijen en personen belang hebben bij het te ontwikkelen AIPA. Het is met name raadzaam om naast gebruikers ook patiënten, patiëntvertegenwoordigers, cliënten of burgers al in deze fase te betrekken bij de ontwikkeling.

Fase 0 is geen onderdeel van de leidraad. Wel is het van belang om in deze fase al na te denken over de invulling van specifieke normen en aanbevelingen uit de leidraad. Risico's en medisch-ethische overwegingen bepalen bijvoorbeeld welke aanbevelingen uit deze leidraad zullen gaan worden toegepast, en praktische overwegingen bepalen vervolgens of het toepassen van die aanbevelingen zakelijk gezien een haalbaar plan oplevert. Daarom is in deze fase voorsorteren op het *pas toe of leg uit* principe van de leidraad belangrijk, omdat hier in deze fase vaak al vorm aan wordt gegeven.

Voor fase 0 zijn geen expertsessies in de vorm van werkgroepen georganiseerd en er zijn derhalve ook geen eisen of aanbevelingen opgesteld voor deze fase. Fase 0 valt daarmee buiten het toepassingsbereik van deze leidraad.

Totstandkoming van de leidraad

Voorbereidend systematisch literatuuronderzoek samengevat in het artikel *Guidance and Quality Criteria for Artificial Intelligence based Prediction Algorithms in healthcare: a scoping review*[2]; uitgevoerd in opdracht van Ministerie van Volksgezondheid, Welzijn en Sport vormde het startpunt van deze leidraad.

Op basis van dit uitgebreide literatuuronderzoek is vervolgens voor elk van de bovengenoemde zes fasen een multidisciplinaire werkgroep samengesteld met vele experts uit het veld, waaronder zorgverleners, normexperts, epidemiologen, datamanagers, ethici, statistici, beleidsmedewerkers, kwaliteitsmedewerkers, data-scientists en AI-experts werkzaam bij onder andere overheden, academische en non-academische ziekenhuizen en bedrijven. De werkgroepen hebben binnen elke fase diverse relevante onderwerpen geagendeerd, besproken, geprioriteerd en verder uitgewerkt om te komen tot een leidraad met minimale eisen en aanbevelingen. De leidraad is vervolgens door belanghebbenden uit het veld van commentaar voorzien, welke via brede publieke oproepen vanuit het ministerie van volksgezondheid zijn geworven. Verder heeft een praktijktoets aan de hand van vijf verschillende AIPA's plaatsgevonden en hebben de Patiëntenfederatie Nederland, de Nederlandse AI Medical Device expert Group van het NEN en Inspectie Gezondheidszorg en Jeugd input geleverd.

Referenties

1 A definition of Artificial Intelligence: main capabilities and scientific disciplines | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines (24 June 2021)

2 Anne A.H. de Hond*, Artuur M. Leeuwenberg*, Lotty Hooft, Ilse M.J. Kant, Steven W.J. Nijman, Hendrikus J. A. van Os, Jiska J. Aardoom, Thomas P.A. Debray, Ewoud Schuit, Maarten van Smeden, Johannes B. Reitsma, Ewout W. Steyerberg, Niels H. Chavannes‡ and Karel G.M. Moons‡. *Shared first author, ‡ shared last author. Guidance and Quality Criteria for artificial intelligence-based prediction models in healthcare: a scoping review. Under review.

1 Verzameling en beheer van de data

Auteurs

Maarten van Smeden, Ilse Kant, Lotty Hooft, Paul Algra, Pieter Boone, Andre Dekker, Amy Eijkelenboom, Christian van Ginkel, Saskia Haitjema, Martine de Vries, Hine van Os, Niels Chavannes, Carl Moons

Fase 1 beslaat het verzamelen en beheren van de benodigde data voor fasen 2 t/m 6. Data kan beschikbaar worden gesteld ten behoeve van de ontwikkeling van het AIPA (fase 2), de externe validatie van het AIPA (fase 3), de softwareontwikkeling (fase 4), de effectbeoordeling van het AIPA in combinatie met de software (fase 5) en de implementatie in het dagelijks gebruik (fase 6). Fase 1 speelt dus een overkoepelende rol in het gehele traject richting implementatie en gebruik van het AIPA (met software) in de dagelijkse praktijk. De specifieke eisen die aan de data worden gesteld kunnen verschillen per fase. Hoe besloten wordt welke data specifiek zouden moeten worden verzameld of gebruikt voor ontwikkeling, validatie en implementatie van een AIPA, wordt in deze leidraad niet behandeld.

De invulling van fase 1 draait om het opstellen, beheren en uitvoeren van een zogenoemd datamanagementplan. In dit plan worden afspraken, (verwerkings)overeenkomsten en procedures vastgelegd over het verzamelen van benodigde (meta)data, het opslaan van deze (meta)data en de toegankelijkheid ervan.

Eén AIPA, meerdere datamanagementplannen

De verschillende AIPA ontwikkel-, test-, en implementatiefasen vragen om verschillende vormen van data die verzameld en/of beheerd moeten worden. Bij de dataverzameling zijn in verschillende fasen vaak verschillende partijen betrokken. Bij een AIPA die alle fasen heeft doorlopen zullen waarschijnlijk meerdere datamanagementplannen een rol spelen, omdat in elke fase andere (juridische en ethische) eisen en andere onderzoeksvormen een rol spelen. Bij het opstellen van een datamanagementplan is het verstandig rekening te houden met de verwachtingen van toezichthouders en certificerende instanties die aanlevering van een datamanagement plan eisen, bijvoorbeeld notified bodies en ethische commissies, in het kader van toelating op de markt voor medische hulpmiddelen.

De precieze invulling van het datamanagementplan is afhankelijk van velerlei factoren. In algemeenheid kunnen vier kerndomeinen van het datamanagementplan worden onderscheiden: juridische randvoorwaarden, dataverzameling, metadata en beschikbaarheid data.

In elke fase waar data wordt verzameld **moet** vooraf door de ontwikkelaar een (nieuwe versie van een) datamanagementplan zijn opgesteld. **(1a)**

Met verzameling wordt ieder verzamelen of samenbrengen van data tot een dataset ten behoeve van de ontwikkeling of evaluatie van het AIPA bedoelt, ook als het een verzameling uit reeds bestaande registers of interne databronnen betreft. Ook de informatie die door het

AIPA in fase 5 en 6 benodigd is en beschikbaar wordt gemaakt aan de eindgebruiker, valt uiteindelijk als dataverzameling onder het datamanagementplan.

1.1 Juridische randvoorwaarden

De juridische randvoorwaarden en context **moeten** door de ontwikkelaar worden benoemd in het datamanagementplan, door middel van beschrijving of verwijzing. **(1.1a)**

Daarbij **moet** tenminste worden beschreven welke nationale en Europese wet- en regelgeving van toepassing zijn op de data en de daarop gebaseerde AIPA. **(1.1b)**

In deze context kan worden gedacht aan onder meer de *Medical Device Regulation* (MDR), de *Wet op de Geneeskundige Behandelovereenkomst* (WGBO), de *Wet beveiliging netwerk- en informatiesystemen (WBNI)*, de *Wet medisch-wetenschappelijk onderzoek met mensen* (WMO), en de *Algemene Verordening Gegevensbescherming* (AVG). Verdere juridische randvoorwaarden hangen af van de doelstelling van het AIPA en de vorm waarin deze ingezet gaat worden.

Daarnaast **moet**, in het geval van een samenwerking tussen organisaties of gebruik van data van derde partijen, worden beschreven welke overeenkomsten (bijv. verwerkersovereenkomsten met externe partijen) zijn gesloten of worden gesloten, welke afspraken in deze overeenkomsten zijn opgenomen (bijv. wat betreft informatiebeveiliging en bewaartermijnen) en welke afspraken er worden gemaakt wat betreft intellectueel eigendomsrecht. **(1.1c)**

Daarnaast wordt **sterk aanbevolen** bestaan en werking van algemene informatiebeveiligingsmaatregelen omtrent toegang tot data die dienen ter naleving van de wet vast te leggen door naar passende documentatie, zoals bijv. een ISO 27001 of NEN 7510 certificering te verwijzen. **(1.1d)**

1.2 Dataverzameling

De eigenschappen van de dataverzameling **moeten** nauwkeurig en gedetailleerd door de ontwikkelaar worden vastgelegd in het datamanagementplan dat betrekking heeft op de specifieke AIPA ontwikkel-, evaluatie-, of implementatie-fase. **(1.2a)**

Voor een dataverzameling, **moeten** daarbij tenminste worden vastgelegd:

(i) De herkomst van de data, zoals de (verwachte) begin- en (verwachte) einddatum van de dataverzameling, locatie(s) van verzameling (bijv. of uit ziekenhuizen of registers data werd verzameld),

(ii) Het originele doel en de context van de dataverzameling incl. de toegepaste in- en exclusiecriteria van de beoogde doelgroep (bijv. patiënten of burgers), en in die gevallen dat dataverwerking berust op expliciete toestemming van patiënt,

cliënt of burger, de voorwaarden waaronder de patiënt of burger toestemming heeft verleend (o.a. het verwerkingsdoel),

(iii) De procedures van metingen en registratie van data, zoals het ontwerp (design) van de datacollectie (bijv. cohortonderzoek, routinematig verzamelde zorggegevens), de timing van metingen waarmee data van individuen wordt verzameld (bijv. meting van patiënten direct na ziekenhuisopname, periodieke herhalingen van metingen) en indien van toepassing de technische eigenschappen van meetinstrumenten (bijv. fabrikant, type nummer en sensitiviteit/responsiviteit). **(1.2b)**

Het uitgangspunt van fase 1 (en de bijbehorende datamanagementplannen) is dat deze beschrijvingen voldoende gedetailleerd zijn om de dataverzameling en/of data-extractie in beginsel te kunnen reproduceren, al dan niet door een derde partij.

1.2.1 Privacy en herleidbaarheid

Ten aanzien van privacy is de geldende regelgeving (de huidige AVG) leidend, ongeacht of de data betrekking heeft op ingezetenen van de Europese Unie.

De privacy van personen waar data van is verkregen **moet** door de ontwikkelaar worden gerespecteerd en gewaarborgd. **(1.2.1a)**

Herleidbaarheid van data naar personen **moet** worden voorkomen (anonimisering) of beperkt (pseudonimisering). (**1.2.1b)**

Daarnaast **moet** het principe van dataminimalisatie gevolgd worden, daarmee wordt bedoeld dat niet meer data per subject wordt vastgelegd dan nodig voor de ontwikkeling of het gebruik van het AIPA. **(1.2.1c)**

Metadata kan in voorkomende gevallen worden gebruikt om, in combinatie met data zelf, personen te identificeren. **Aanbevolen wordt** de mogelijkheid op reidentificatie van personen door middel van combinatie van de data over de persoon en de metadata over de data te onderzoeken en de resultaten van dit onderzoek bij de keuze voor vastlegging van metadata mee te wegen. **(1.2.1d)**

Daarnaast **moet**, indien van toepassing, door de AIPA-ontwikkelaar of -tester expliciet in het datamanagementplan worden vastgelegd hoe om wordt gegaan met eventuele toevalsbevindingen (bevindingen die aan het licht komen tijdens een onderzoek wat een ander doel dient) en het recht op vernietiging van data van personen waar data van is verkregen. **(1.2.1e)**

In verschillende fasen kan dataverzameling onder de WMO vallen. Onderzoek dat onder de WMO valt **moet**, afhankelijk van het type onderzoek, vooraf door een erkende medisch-ethische toetsingscommissie (METC) of de Centrale Commissie Mensgebonden Onderzoek (CCMO) worden getoetst. Daarnaast is vaak ook een gegevensbeschermingseffectbeoordeling (GEB) in het kader van de AVG **vereist**.

Sterk aanbevolen wordt de plannen rondom privacy en herleidbaarheid door een gegevensbeschermingseffectbeoordeling (GEB) te laten toetsen of een privacy-deskundige of METC te benaderen, ook wanneer hier geen juridische verplichting toe bestaat. **(1.2.1e)**

1.3 Metadata

Het datamanagementplan **moet** een gedetailleerde beschrijving geven van metadata. **(1.3a)**

Metadata is data die de karakteristieken van de verzamelde data inzichtelijk maakt – data over data – en beschrijft onder meer de verzameling, rapportage en toegankelijkheid van de verzamelde data. Het gaat hier in essentie om het vastleggen van de in algemene zin omschreven eigenschappen en processen (zoals beschreven in 1.2) in de dataverzameling zelf. De nadruk moet daarbij liggen op het verschaffen van transparantie en duidelijkheid over de verzamelde data.

Sterk aanbevolen wordt metadata vast te leggen op de volgende niveaus:

- *Data provenance*[1] (ofwel data lineage): bevat informatie over de herkomst van de verzamelde data(punten), eventuele veranderingen en transformaties aan de data inclusief classificatie van het doel van de verandering en overige details die informatie kunnen geven over de validiteit van de verzamelde data, voor zover verenigbaar met de aanbevelingen omtrent herleidbaarheid in 1.2.1.
- *Medische context:* informatie over het ontwerp (design) van de datacollectie en de populatiecontext (bijv. consecutieve patiënten bij de huisarts met huidklachten, ziekenhuispatiënten verwezen voor een CT vanwege verdenking longembolie, gezonde mensen in de algemene bevolking van 70 jaar of ouder waarbij nagegaan wordt hoe groot hun kans op een bepaalde kanker is). Daarnaast beschrijft dit ook de fysieke en sociale omgevingsdeterminanten van de geïncludeerde populatie, indien relevant voor de toepassing van het AIPA.
- *Eigenschappen en beschrijvende statistiek van de data,* zoals de eenheden, gemiddelden, ranges van waarden, beschrijving ontbrekende waarnemingen en eventuele verschuivingen of trends in relatie tot de tijd. **(1.3b)**

De keuze voor metadata en de omschrijving van metadata **moet** gebaseerd zijn op een inventarisatie van de belangen van de verschillende stakeholders die inzage in de metadata

zouden moeten kunnen krijgen, in het bijzonder controlerende of certificerende instanties en in het geval van samenwerkingsverbanden, partner (zorg)organisaties. **(1.3d)**

Indien er meerdere databronnen in een bepaalde fase worden gebruikt, bijvoorbeeld verschillende datasets uit verschillende dataverzamelingsprocessen voor het valideren (fase 3) van het AIPA, wordt **sterk aanbevolen** om de metadata voor elke databron apart te presenteren en te specificeren hoe de bronnen gekoppeld zijn. **(1.3e)**

1.4 Beschikbaarheid data

Het datamanagementplan **moet** duidelijke informatie verschaffen over de beschikbaarheid van de data, voor belanghebbenden en derden. **(1.4a)**

Voor het (intern of extern) beschikbaar stellen van data wordt **sterk aanbevolen** om de FAIR-principes[2] te volgen. **(1.4b)**

FAIR is een acroniem voor *Findable* (vindbaar), *Accessible* (toegankelijk), *Interoperable* (uitwisselbaar) en *Reusable* (herbruikbaar). De FAIR-principes zijn richtlijnen voor de beschrijving, opslag en publicatie van (meta)data.

In het geval dat data beschikbaar gesteld wordt aan partners of derden **moet** in het datamanagementplan worden vastgelegd welke afspraken bestaan over de opslag van de gebruikte data. Daarbij moet tenminste worden vermeld: de vorm waarin data wordt opgeslagen, de locatie(s) van dataopslag, de planning van incidentele en periodieke data back-ups, afspraken over mogelijk incidenten zoals data lekken en de (resterende) bewaartermijn van de data. **(1.4c)**

Hierbij **moet** waar van toepassing worden vastgelegd hoe wordt voldaan aan de geldende nationale en internationale wet- en regelgeving omtrent verwerking van persoonsgegevens, dataopslag en databeveiliging, zoals beschreven in o.a. de Algemene Verordening Gegevensbescherming (AVG) en de Wet beveiliging netwerk- en informatiesystemen (WBNI) en daarop berustende richtlijnen. **(1.4d)**

Daarnaast wordt **aanbevolen** de data beschikbaar te maken in vormen die aansluiten bij informatiestandaarden die gebruikelijk zijn in digitale informatie-uitwisseling in de gezondheidszorg. **(1.4e)**

Daarbij kan voor Nederland bijvoorbeeld gedacht worden aan de diverse informatiestandaarden voor informatie-uitwisseling in de zorg die voor de Nederlandse zorgcontext beheerd worden door Nictiz[3] en een belangrijke rol spelen bij de uitwisseling van patiëntdata tussen verschillende zorginstellingen en zorgverleners. Internationaal kan bijvoorbeeld gedacht worden aan het medisch terminologiestelsel SNOMED[4].

Voor invulling van bovenstaande eisen kan in het datamanagementplan worden verwezen naar de gesloten verwerkingsovereenkomsten, in zoverre de overeenkomsten afspraken over al deze punten bevatten.

1.5 Versiebeheer en beschikbaarheid van het datamanagementplan

Het datamanagementplan **moet** door de ontwikkelaar beschikbaar worden gesteld aan de bij de dataverzameling of -verwerking betrokken partijen. **(1.5a)**

Aanbevolen wordt het datamanagementplan openbaar of op aanvraag toegankelijk te maken, bijvoorbeeld door het te plaatsen op een openbaar toegankelijke website. **(1.5b)**

Deze aanbeveling mag afgewogen worden tegen commerciële belangen.

Voor alle onderdelen van het datamanagementplan **moet** versiebeheer geïmplementeerd worden. **(1.5c)**

Dit betekent dat eventuele veranderingen aan het datamanagementplan over de tijd nauwkeurig moeten worden geregistreerd en vastgelegd. Daarmee is het datamanagementplan een levend document, dat in de opvolgende fasen regelmatig bijgewerkt wordt of opnieuw wordt opgesteld in een volgende fase.

1.6. Referenties

1. Gupta A. Data Provenance. In: Liu L, Özsu MT, eds. *Encyclopedia of Database Systems* Boston, MA: Springer US; 2009. P. 608–608
2. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, Silva Santos LB da, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth rowth P, Goble C, Grethe JS, Heringa J, Hoen PAC 't, Hooft R, Kuhn T, Kok R, Kok J, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018.
3. Nictiz. Standaardisatie van digitale gegevensuitwisseling in de zorg. https://www.nictiz.nl/overig/standaardisatie-van-digitale-gegevensuitwisseling-in-de-zorg/. Published December 5, 2018. Accessed December 8, 2021.
4. SNOMED. SNOMED International. https://www.snomed.org/. 2021. Accessed December 8, 2021.

2 Ontwikkeling van het AIPA

Auteurs

Maarten van Smeden, Ilse Kant, Lotty Hooft, Gabrielle Davelaar, Desy Kakiay, Evangelos Kanoulas, Kicky van Leeuwen, Joran Lokkerbol, Daniel Oberski, Hine van Os, Niels Chavannes, Carl Moons

Fase 2 beslaat het ontwikkelen van het AIPA model. Het model is het geheel aan algoritme-specifieke datastructuren dat in combinatie met een algoritme het AIPA vormt, en is het resultaat van analyse van de trainingsdata. In dit document wordt geen concreet stappenplan voor de analytische AIPA modelontwikkeling gegeven, de lezer wordt daarvoor verwezen naar bestaande literatuur.[1–6].

Sterk aanbevolen wordt om een gestandaardiseerd stappenplan te gebruiken voor een complete vastlegging van de ontwikkelingsstappen en de procedures en resultaten van interne validatie (zie onder) van het AIPA. **(2a)**

De TRIPOD reporting guidelines[7–9] (www.tripod-statement.org) dienen daarbij als leidraad, en een specifieke TRIPOD-AI reporting guideline is bijna voltooid.

1.1. Uitleg doelgebruik

De ontwikkelaar van het model **moet** duidelijk het doelgebruik van het AIPA definiëren en vastleggen. **(2.1a)**

In het vastgelegde doelgebruik **moet** tenminste duidelijk worden gemaakt:

i) Voor welke medische- of gezondheidstoepassing het AIPA bedoeld is (bijv. bij welke medische context, indicatie of doelpopulatie) en wie de beoogde eindgebruiker is (bijv. een specifiek specialisme, eerstelijnszorgverlener, of de patiënt, cliënt of burger zelf);

ii) Welk medisch of gezondheidszorgproces beïnvloed beoogt te worden door het AIPA en wat de verwachte meerwaarde t.o.v. het huidige proces is (bijv. het maken van een snellere diagnose een betere inschatting van iemands prognose, of indicatie voor aanpassing van een leefstijlgewoonte);

iii) Wat de beoogde momenten van het AIPA gebruik ofwel van het voorspellen zijn (bijv. bij opname in ziekenhuis of Intensive Care, bij moment van diagnose met kanker, bij verwijzing voor CT-scan, of bij constatering van symptomen of klachten, of controle van het suikergehalte in het bloed);

iv) Of het een diagnostische, prognostische, monitoring, screening of ander type gezondheidszorgtoepassing betreft;

v) Wat de predictiehorizon van het AIPA is (in geval van prognostische voorspellingen: hoever in de tijd het AIPA voorspelt). **(2.1b)**

Sterk aanbevolen wordt om bij de definitie van doelgebruik stakeholders zoals gebruikers en patiënten, cliënten of burgers te betrekken. **(2.1c)**

2.1.1 Dataset(s) en doelgebruik

In fase 1 is reeds een precieze beschrijving vastgelegd van de oorsprong van de dataset(s) (bijv. tijd/plaats) die worden gebruikt voor ontwikkeling van het AIPA model, het design van dataverzameling (bijv. opeenvolgende patiënten), meet- en registratieprocedures, eventuele selecties, in- en exclusiecriteria van deelnemers of datapunten in het onderzoek.

In algemeenheid wordt **sterk aanbevolen** een representatieve steekproef uit de doelpopulatie (zoals vastgelegd in het doelgebruik, zie sectie 2.1.) te gebruiken voor de ontwikkeling van het AIPA. **(2.1.1a)**

Indien het vermoeden bestaat dat de gebruikte data niet (volledig) representatief is **moet** dit worden gedocumenteerd en inhoudelijk worden onderbouwd. **(2.1.1b)**

2.2 Analyse- en modeleringstappen

De ontwikkelaar van het model **moet** alle analyse- en modelontwikkeling stappen vastleggen. Daarbij horen alle voorbereidingsstappen (bijv. initiële data analyse[10], feature engineering), gebruikte modelleringstechniek (bijv. neuraal netwerk, *random forest, time-to-event*, logistische regressie), alle modeleringstappen (bijv. modelselectie, tuning, (her-) kalibratie). **(2.2a)** Het uitgangspunt is dat de achtereenvolgende analyse- en modeleringstappen voldoende gedetailleerd zijn zodat een derde partij op basis van de beschrijving alle analyse- en modelstappen exact zou kunnen reproduceren [7–9, 11].

2.3 Interne evaluatie van het model

2.3.1 Interne validatie

Interne validatie is een belangrijk onderdeel in het proces van ontwikkeling van het AIPA. De interne validatie heeft als doel realistische schattingen van de voorspelkracht van het AIPA te kwantificeren. Een adequate schatter van de voorspelkracht van het model (bijv. de C(oncordance)-statistic en kalibratiecurve[8,12]) kan verschillen tussen soorten toepassingen en eindpunten (bv binair, multi-categorie, *time-to-event*), zie ook sectie 3.1.2. Expliciete minimale criteria voor voorspelkracht worden niet gegeven in dit document omdat minimale voorspelkracht context-afhankelijk is.

Sterk aanbevolen wordt om de voorspelkracht zoveel mogelijk in context te beschrijven, bijv. door vergelijking met andere voorspelmodellen of AIPA's voor dezelfde medische context of doelpopulatie, of door vergelijking met een benchmark relevant voor de medische context zodat de meerwaarde ten opzichte van de huidige medische praktijk beoordeeld kan worden. **(2.3.1a)**

Om tot realistische schattingen van de voorspelkracht te komen **moeten** adequate maatregelen genomen worden om *voorspelkracht optimisme*[1,2,5,6] te minimaliseren. **(2.3.1b)**

Dit betekent dat de interne validatie strikt moet worden gescheiden van de modelontwikkeling, zoals de variabele- en modelselectie en tuning van het model (d.w.z. waken voor *leakage*). Bijvoorbeeld door *nested cross-validation*, waarin het uitvoeren van alle modelontwikkelstappen (*inner loop*) wordt gescheiden van het intern valideren van het model (*outer loop*).

Sterk aanbevolen wordt om statistisch efficiënte interne validatiemethoden toe te passen (bijv. cross-validatie, bootstrap), waarin alle data die beschikbaar zijn voor de ontwikkeling worden gebruikt voor de ontwikkeling van het model, boven inefficiënte interne validatiemethoden (bijv. enkele train-test splits)[13]. **(2.3.1c)**

Indien hiervan wordt afgeweken, bijvoorbeeld omdat het computationeel niet haalbaar is, moet dit inhoudelijk worden onderbouwd.

2.3.2 Analyse van mogelijke (negatieve) impact van het model

Naast een realistische schatting van de voorspelkracht is het van belang doorlopend in de ontwikkeling vooruit te kijken naar de (mogelijke) toepassing van het AIPA in de praktijk, zodat het AIPA in ontwikkeling aangesloten blijft bij de medisch context en het op te lossen probleem.

Aanbevolen wordt om een geloofwaardige en transparante analyse van de mogelijke negatieve impact van het gebruik of invoer van het AIPA uit te voeren en deze vast te leggen als onderdeel van de beoordeling van de meerwaarde van het AIPA ten opzichte van de huidige medische praktijk. **(2.3.2a)**

Bijvoorbeeld door een analyse van de voorspelfouten van het model (d.w.z. *error analyse*) te verrichten en deze expliciet te relateren aan het doelgebruik.

Aanbevolen wordt om samen met stakeholders uit de medische context die bij het doelgebruik beoogd wordt, een inschatting te maken van *fairness* risico's. Zie sectie 3.3 voor een gedetailleerdere uitwerking. **(2.3.2b)**

Voorspelkracht is niet altijd voor alle deelpopulaties waarover gegeneraliseerd wordt gelijk.

Daarom wordt **sterk aanbevolen** om zoveel mogelijk *heterogeniteit* in de geschatte voorspelkracht van het AIPA in kaart te brengen, bijvoorbeeld door gebruik van data uit meerdere locaties (bijv. verschillende medische centra) of andere patiënt relevante contexten – bijvoorbeeld met behulp van *internal-external cross-validation*[14,15]. **(2.3.2c).**

Indien hiervan wordt afgeweken moet dit worden onderbouwd met verwijzing naar de doelstelling van het model en een afweging van de risico's voor de robuustheid van het model.

Ook wordt **sterk aanbevolen** om de verwachtte meerwaarde in de medische context van het model te onderzoeken en vast te leggen. **(2.3.2d)**

Dit kan bijvoorbeeld met *een decision curve analysis*[16]. Een andere, meer robuuste en uitgebreidere, manier om de impact op de medische praktijk in een vroeg stadium van de ontwikkeling van een AIPA te onderzoeken is met een *early Health Technology Assessment (eHTA)* van het AIPA[17,18] .

2.4 Technische Robuustheid

Bij de ontwikkeling van het AIPA **moet** ook de technische robuustheid van het model worden onderzocht en bevindingen transparant worden vastgelegd, tenminste voor die modellen die gebruikt worden in de externe validatie (fase 3). **(2.4a)**

Sterk aanbevolen wordt om technische robuustheid, naast voorspelkracht (zoals bedoeld in 2.3.1), te gebruiken als criterium voor modelselectie. **(2.4b)**

Om de robuustheid te onderzoeken wordt aanbevolen om diverse sensitiviteitsanalyses uit te voeren. Hierbij kan worden gedacht aan analyses van de:

- *Architectuur robuustheid:* het herhalen van de analyse stappen op dezelfde data leidt tot een model dat niet significant afwijkt van het oorspronkelijke model.
- *Consistentie van modelvoorspellingen:* het herhalen van de analysestappen op dezelfde data leidt tot modellen met voorspellingen die niet veel afwijken van de voorspellingen uit het oorspronkelijke AIPA.
- *Adversarial robuustheid:* de invloed van het (opzettelijk) verstoren van de inputvariabelen van het model op de voorspellingen en/of de architectuur.
- *Domeinshift en outliers:* de invloed van eventuele *outliers* in de data en/of bewuste verandering van de dataset (bijv. bewust bepaalde groepen in- of excluderen) op de modelvoorspellingen en/of de architectuur (bijv. *outlier rejection* analyse). Zie ook fase 4 en 6 voor aanvullende activiteiten.

Daarnaast kan, om de transparantie van het AIPA te vergroten, ervoor worden gekozen om de invloed van bepaalde inputvariabelen op de voorspelling inzichtelijk te maken met behulp van bijvoorbeeld *feature importance* methoden (d.w.z. *explainable* AI[19]).

Naast het onderzoeken van de technische robuustheid van het AIPA model tijdens de ontwikkeling, dient ook de robuustheid van het model in combinatie met de software waar het model deel van uit maakt te worden onderzocht. Zie hiervoor fase 4.

2.5 Grootte van de dataset voor ontwikkeling van het AIPA

Het uitgangspunt voor het kiezen van de grootte van de dataset voor ontwikkeling van het model is: hoe groter, hoe beter. Wel moet dit uitgangspunt gewogen worden tegen medisch-ethische overwegingen en de eis van dataminimalisatie uit fase 1. In algemeenheid wordt de minimaal benodigde grootte van de dataset groter naarmate de incidentie (of prevalentie) van de te voorspellen uitkomst verder weg van 50% af ligt (d.w.z. hogere *class imbalance*), naarmate er minder sterke voorspellers zijn voor de uitkomst in de input (lagere verklaarde variantie in de uitkomst door de input variabelen) en naarmate het model meer inputvariabelen bevat en/of computationeel complexer is. Voor op regressie gebaseerde modellen bestaan er expliciete regels en formules die kunnen worden toegepast om de minimale grootte van de dataset te kunnen berekenen[20,21]. Voor complexere modellen bestaan dit soort regels voor *a priori* berekeningen van minimale grootte van de dataset (nog) niet. Wel bestaan er *a posteriori* sample size criteria[22], bijv. zogenaamde *learning curves*[23], waarmee kan worden geëvalueerd of de dataset aan minimale criteria voldoet, zoals beperkt risico op *overfitting* en precieze schatting van de geïndividualiseerde kansen op de uitkomst.

Het gebruik van *a priori* of *a posteriori* methoden om te evalueren of de grootte van de dataset aan minimale criteria voldoet wordt **sterk aanbevolen**. **(2.5a)**

Een METC zal om een rechtvaardiging van de grootte van de dataset vragen.

2.6 Vastlegging, beschikbaarheid en versiebeheer

2.6.1 Vastlegging, reproduceerbaarheid en repliceerbaarheid

Reproduceerbaarheid en repliceerbaarheid zijn belangrijke uitgangspunten voor de ontwikkeling van het AIPA.

Om reproduceerbaarheid (d.w.z. de mogelijkheid van heruitvoering van de ontwikkeling met andere data) te garanderen **moeten** alle analyse stappen (zie ook eis 2.2a) en interne validatiestappen en analyse van technische robuustheid volledig worden vastgelegd. **(2.6.1a)**

Hierbij is het uitgangspunt opnieuw dat de vastlegging voldoende gedetailleerd is voor derden om de ontwikkelstappen te kunnen reproduceren. De TRIPOD reporting guidelines[7-9] (www.tripod-statement.org) kunnen daarbij gebruikt worden als leidraad.

Daarnaast wordt **aanbevolen** om waar van toepassing, gevonden resultaten te publiceren in een wetenschappelijk tijdschrift. **(2.6.1b)**

Ook wordt **aanbevolen** om computercodes die gebruikt zijn voor het AIPA ontwikkeling openbaar of op aanvraag beschikbaar te stellen zodat deze door derden gebruikt kunnen worden voor een onafhankelijke validatie van het AIPA model[24]. **(2.6.1c)**

Daarnaast wordt **aanbevolen** om de repliceerbaarheid (d.w.z. heruitvoering van de modelontwikkeling met dezelfde data) door derden te waarborgen door waar het kan de data op aanvraag beschikbaar te stellen. **(2.6.1d)**

Daarbij dient natuurlijk rekening te worden gehouden met de huidige wet- en regelgeving omtrent privacy, en daaruit voortkomende beperkingen en risico's zoals de kans op identificatie van betrokkenen. Ook mogen deze aanbevelingen afgewogen worden tegen commerciële belangen, en kan als argument aangevoerd worden dat externe validatie door een vertrouwde derde plaats zal vinden.

2.6.2 Versiebeheer en beschikbaarheid van model

De versiegeschiedenis (het uiteindelijke model en eventuele updates van het model) **moet** volledig vastgelegd worden, bijvoorbeeld door het toekennen van een versienummer. **(2.6.2a)**

Deze versiegeschiedenis van het model is een aanvulling op de versiegeschiedenis van de software, zoals bijv. geëist door de MDR.

Daarnaast wordt **sterk aanbevolen** het feitelijke model en/of modellen (bijv. modelcoëfficiënten indien van toepassing, nomogrammen, computercode (met het feitelijk model)) openbaar toegankelijk te maken, indien beschikbaar inclusief (minimale) software rondom het model ter demonstratie. **(2.6.2b)**

Indien hiervan wordt afgeweken, moet dit worden onderbouwd. Commerciële belangen kunnen in deze afweging doorslaggevend zijn.

2.7 Referenties

1. Harrell FE. Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.
2. Steyerberg EW. Clinical Prediction Models. Cham: Springer International Publishing; 2019.
3. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning The Elements of Statistical LearningData Mining, Inference, and Prediction, Second Edition. Springer Ser. Stat. 2009.
4. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, Massachusetts: The MIT Press; 2016.
5. Riley RD, van der Windt D, Croft P, Moons KGM. Prognosis Research in Health Care: Concepts, Methods, and Impact. Oxford University Press, 2019.
6. Moons KG, Kengne AP, Woordward M, Royston P, Vergouwe Y, Altman DG, Grobbee DE. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012;98:683-690.
7. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med* 2015;162:55.
8. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann Intern Med* 2015;162:W1–W73.
9. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019; 393:1577-1579.
10. Huebner M, Vach W, Cessie S le. A systematic approach to initial data analysis is good research practice. *J Thorac Cardiovasc Surg* 2016;151:25–27.
11. Collins GS, van Smeden M, Riley RD. COVID-19 prediction models should adhere to methodological and reporting standards. European respiratory journal 2020; 56: 2002643.
12. Van Calster B, McLernon DJ, Smeden M van, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;17:230.

13. Debray TP, Riley RD, Rovers MM, Reitsma JB, Moons KG; Cochrane IPD Meta-analysis Methods group. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. *PloS Med.* 2015; 12:e1001886.

14. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. *J Clin Epidemiol* 2016;69:245–247.

15. Debray TP, Damen JA, Riley RD, Snell K, Reitsma JB, Hooft L, Collins GS, Moons KG. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res.* 2019;28:2768-2786.

16. Vickers AJ, Calster B van, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019;3:18.

17. Jenniskens K, Lagerweij GR, Naaktgeboren CA, Hooft L, Moons KGM, Poldervaart JM, Koffijberg H, Reitsma JB. Decision analytic modeling was useful to assess the impact of a prediction model on health outcomes before a randomized trial. *J Clin Epidemiol* 2019;115:106-115.

18. Van Giessen A, Wilcher B, Peters J, Hyde C, Moons KG, de Wit GA, Koffijberg H. Health economic evaluation of diagnostic and prognostic prediction models. A systematic review. *Value in Health* 2014;17:A560.

19. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion* 2020;58:82–115.

20. van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, Reitsma JB. Sample size for binary logistic prediction models: beyond events per variable criteria. *Statistical methods in medical research* 2019;28:2455-2474.

21. Riley RD, Ensor J, Snell KI, Harrell FE, Martin GP, Reitsma JB, Moons KG, Collins GS, van Smeden M. Calculating the sample size required for developing a clinical prediction model. *Bmj* 2020;368: m441.

22. Balki I, Amirabadi A, Levman J, Martel AL, Emersic Z, Meden B, Garcia-Pedrero A, Ramirez SC, Kong D, Moody AR, Tyrrell PN. Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review. *Can Assoc Radiol J* 2019;70:344–353.

23. Christodoulou E, Smeden M van, Edlinger M, Timmerman D, Wanitschek M, Steyerberg EW, Van Calster B. Adaptive sample size determination for the development of clinical prediction models. *Diagn Progn Res* 2021;5:6.

24. Community TTW, Arnold B, Bowler L, Gibson S, Herterich P, Higman R, Krystalli A, Morley A, O'Reilly M, Whitaker K. The Turing Way: A Handbook for Reproducible Data Science. Zenodo; 2019.

3 Validatie van het AIPA

Auteurs

Maarten van Smeden, Ilse Kant, Lotty Hooft, Huib Burger, Daan van den Donk, Vincent Stirler, Bart Jan Verhoeff, Wouter Veldhuis, Hine van Os, Niels Chavannes, Carl Moons

Fase 3 beslaat het (extern) valideren van de in fase 2 ontwikkelde AIPA. Met externe validatie wordt de evaluatie van de voorspellingen van het AIPA model bedoeld met data die niet is gebruikt voor de ontwikkeling (of doorontwikkeling) in fase 2[1–3]. We maken daarbij onderscheid tussen de evaluatie van de statistische ofwel voorspellende waarde, de evaluatie van de (meer)waarde ten opzichte van de huidige zorgpraktijk en de evaluatie van *fairness en algoritmische bias*.

De overgang van fase 2 naar fase 3 is gebaseerd op de aanname dat de ontwikkeling van het AIPA model voltooid is. Wel kan het voorkomen dat een kleine set van kandidaat modellen in fase 3 gevalideerd worden om tot een finale keuze van een model te komen. Met externe validatie wordt expliciet niet bedoeld: het (her-)trainen of (her-)tunen van een model. Complete of gedeeltelijke her-training van een ontwikkelde AIPA kan wel een consequentie zijn van een externe validatie. Dit wordt ook wel *model updating* genoemd[4–6]. Het is dan nodig (een gedeelte van) fase 1 en 2 nogmaals te doorlopen en vast te leggen.

3.1 Evaluatie voorspellende (statistische) eigenschappen van het AIPA

3.1.1 Doelpopulatie en -context

Voor een goede evaluatie van de voorspellende waarde van een AIPA **moet** de tester een andere dataset gebruiken voor externe validatie, dan voor de AIPA ontwikkeling in fase 2 is gebruikt, maar wel een dataset die representatief is voor de doelpopulatie en context[3]. **(3.1.1a)**

Als sprake is van een zogenaamde *holdout* dataset dan zijn de eigenschappen van deze dataset reeds in fase 1 volledig vastgelegd. Anders zijn in ieder geval de eigenschappen van het dataverzamelingsproces vastgelegd.

Het gebruik van onderzoeksdesigns waarin uitsluitend data van gezonde controles wordt gebruikt die niet representatief zijn voor de beoogde context of doelpopulatie, leidt veelal tot een te optimistische evaluatie van de voorspellende waarde van het AIPA model (*spectrum bias*[7]).

Sterk aanbevolen wordt geen onderzoeksdesign te gebruiken waarin uitsluitend data van zogenaamde gezonde controles wordt gebruikt. **(3.1.1b)**

Een precieze beschrijving van de oorsprong van de data (bijv. tijd en plaats), de manier van dataverzameling (bijv. opvolgende patiënten), meet- en registratieprocedures, eventuele selecties en in- en exclusiecriteria **moeten** zijn vastgelegd in het datamanagementplan om de voorspellende eigenschappen in context te kunnen plaatsen (zie fase 1). **(3.1.1c)**

Daarbij dient het uiteindelijke doelgebruik van het AIPA ook scherp in het oog worden gehouden, zoals vastgelegd in fase 2.

Aanbevolen wordt exclusie van individuen die wel tot de doelpopulatie of –context behoren te vermijden. **(3.1.1d)**

Eventuele onvoorziene exclusie van data of individuen door bijvoorbeeld mislukte metingen of door het intrekken van toestemming worden nauwkeurig vastgelegd en, bij voorkeur, per casus of per groep beschreven.

De doelpopulatie of -context bij externe validatie kan in het kader van generaliseerbaarheid bewust afwijken van de doelpopulatie zoals geformuleerd en gebruikt voor ontwikkeling van het AIPA model (fase 2).

Bij structurele verschillen tussen de ontwikkeling (fase 2) en externe validatie (fase 3) in design van dataverzameling, meet- en registratieprocedures, eventuele selecties en in- en exclusiecriteria, **moet** de reden voor verschil in doelpopulaties tussen fase 2 en 3 worden vastgelegd. Daarnaast wordt de aard van de verschillen ook duidelijk vastgelegd in het datamanagementplan.[8,9] **(3.1.1e)**

Daarnaast wordt **aanbevolen** om, indien mogelijk, baseline karakteristieken (bijv. verdelingen van leeftijd, geslacht, comorbiditeit) te vergelijken en statistisch te toetsen tussen de data gebruikt voor de ontwikkeling (fase 2) en voor externe validatie in fase 3 (ook wanneer de doelpopulaties hetzelfde zijn)[10]. **(3.1.1f)**

Hiermee kan in latere fasen eventuele *data drift* in kaart worden gebracht.

3.1.2 Voorspelkracht van het AIPA

Een realistische evaluatie van de voorspelkracht, d.w.z. de overeenkomstigheid van de uitkomst die wordt voorspeld en de uitkomst die is geobserveerd, is een belangrijk onderdeel van de externe validatie van het AIPA model.

Bij de evaluatie van voorspelkracht **moet** bij de keuze voor schatters rekening worden gehouden met de schaal waarop de voorspellingen worden gedaan. **(3.1.2a)**

Een juiste schatter of maat van voorspelkracht kan verschillen tussen een binair eindpunt (bijv. gezondheidsuitkomst aanwezig versus afwezig), multi-categorie eindpunt (bijv. zeker aanwezig, waarschijnlijk aanwezig, waarschijnlijk afwezig, zeker afwezig), een *survival* eindpunt (met mogelijke *censoring*) of een uitkomst op een continue schaal.

Ook **moet** bij de keuze voor schatters van voorspelkracht rekening worden gehouden met de voorspelde output van het AIPA model. **(3.1.2b)**

Bijvoorbeeld: voor voorspellende modellen die alleen binaire (ja versus nee of aanwezig versus afwezig) classificaties als output geven, ligt de nadruk vaak op de accuratesse van classificaties en daaraan verwante maten als de F1-score en *C-statistic*, terwijl voor een

voorspelmodel met kansen/risico output, de nadruk vaak op de kalibratie en discriminatie (*C-statistic*) van het model ligt. Voor een leidraad voor vastlegging van deze keuzes wordt verwezen naar de TRIPOD guidelines[8,9].

Sterk aanbevolen wordt om, net als bij de interne validatie in fase 2, de schattingen van voorspelkracht zoveel mogelijk in de beoogde context te plaatsen, bijvoorbeeld door vergelijking van voorspelkracht met vergelijkbare voorspelmodellen voor dezelfde context of doelpopulatie of een relevante benchmark voor de medische setting die in het doelgebruik beoogd wordt. **(3.1.2c)**

Sterk aanbevolen wordt ook de schattingen van voorspelkracht van het AIPA bij externe validatie te vergelijken met de voorspelkracht gerapporteerd na interne validatie tijdens ontwikkeling (fase 2). **(3.1.2d)**

Een grote discrepantie in voorspelkracht die gevonden wordt in fase 3 bij externe validatie kan onder andere duiden op *overfitting* van het model tijdens de ontwikkeling in fase 2[5,6].

3.2 Evaluatie medische eigenschappen en verwachtingen voor implementatie van het AIPA

Bij een externe validatie van het AIPA model **moet** net als in fase 2 bij de interne validatie ook naar de medische eigenschappen, of de prestaties in de beoogde medische setting, worden gekeken. **(3.2a)**

Aanbevolen wordt een analyse van voorziene kosten en baten te maken. **(3.2b)**

Dit kan bijvoorbeeld, net als in fase 2, worden geïmplementeerd door een *decision curve analysis*[11]. Een andere, veelal uitgebreidere, manier om de impact ten opzichte van de huidige medische praktijk te onderzoeken is met een early Health Technology Assessment (eHTA[12,13]), zoals eerder genoemd in fase 2.

Aanbevolen wordt een inschatting te maken van verwachtte barrières voor implementatie van het AIPA en die in deze fase vast te leggen, ten behoeve van fase 5 en 6. **(3.2c)**

Bijvoorbeeld door een beschrijving van beperkingen veroorzaakt door beschikbaarheid van data, een inschatting van de tijd (bijv. invoer van data of de rekentijd van het model) en kosten (bijv. metingen) die nodig zijn om het model te gebruiken in de praktijk, en verwachtte barrières met betrekking tot de inpassing van het AIPA in de huidige processen van de beoogde medische praktijk.

Sterk aanbevolen wordt stakeholders uit de medische setting die bij het doelgebruik beoogd wordt (zowel eindgebruikers als patiënten) bij de evaluatie van medische eigenschappen, de analyse van kosten en baten en de inschatting van barrières te betrekken. **(3.2d)**

3.3 Fairness en algoritmische bias

Bij de externe validatie van het AIPA model dient men verder te kijken dan alleen naar de voorspelkracht en medische waarde. Ook evaluatie van eerlijkheid (*fairness*[17]) en bias is van groot belang.

Ongelijke behandeling ontstaat meestal door een vorm van algoritmische bias. Verschillende vormen van algoritmische bias kunnen worden onderscheiden. Het begrippenkader van Suresh & Guttag (2020)[14] wordt hierin als leidraad genomen. Zij onderscheiden zes vormen van bias:

- *Historical bias*: ongewenste modeluitkomsten of- voorspellingen door de data uit de wereld zoals het is of zoals het was. Dit kan bijvoorbeeld worden veroorzaakt doordat het AIPA model werd ontwikkeld op data waar systematische onder- of overdiagnose een rol speelde.
- *Representation bias*: ongewenste modeluitkomsten of -voorspellingen door in de data onder-gerepresenteerde subgroepen. Dit kan bijvoorbeeld worden veroorzaakt doordat het AIPA model werd ontwikkeld op data die niet representatief was voor de doelpopulatie of context.
- *Measurement bias*: ongewenste modeluitkomsten of -voorspellingen doordat het AIPA werd getraind op data waarbij de uitkomstvariabele misclassificaties bevatte (zie sectie 3.4) of door verschillen tussen de (nauwkeurigheid van de) meting van voorspellers/features voor de ontwikkeling van het AIPA en de externe validatie/toepassing. Dit wordt ook wel meetheterogeniteit genoemd[15,16].
- *Aggregration bias*: ongewenste modeluitkomsten of -voorspellingen voor bepaalde subgroepen. Dit kan bijvoorbeeld worden veroorzaakt doordat het AIPA een veel slechtere voorspelkracht heeft in (vaak onder-representeerde) subgroepen.
- *Evaluation bias*: vertekende statistische evaluatie door externe validatie van het AIPA op een dataset die niet representatief is voor de doelpopulatie. Hierbij kan men denken aan het gebruik van een AIPA in de eerstelijns zorg die getraind werd met data uit de tweede lijn waarin meer ernstige ziekte voorkomt.
- *Deployment bias*: mismatch tussen het probleem dat het AIPA probeert op te lossen en de manier waarop het gebruikt wordt door anderen.

Fairness van een algoritme wordt in de regel risicogericht onderzocht: eerst worden hypotheses geformuleerd over groepen die mogelijk ongelijk behandeld zouden kunnen worden door het AIPA, bijvoorbeeld door een systematische analyse in de aanpak begeleidingsethiek[24], en vervolgens worden deze hypotheses onderzocht tijdens de externe

validatie. Hierbij wordt tenminste, maar zeker niet uitsluitend, rekening gehouden met groepen die op grond van de bijzondere persoonsgegevens uit de AVG onderscheiden kunnen worden.

De aanwezigheid van bepaalde soorten bias (zoals *algoritmische bias*) die kan leiden tot ongunstige uitkomstenongelijkheden voor bepaalde groepen in de populatie, **moet** worden onderzocht en vastgelegd. **(3.3a)**

Daarnaast **moet** het risico op ongelijke behandeling of ongewenste uitkomstongelijkheden voor bepaalde groepen in de populatie worden onderzocht en vastgelegd. **(3.3b)**

Sterk aanbevolen wordt stakeholders zoals eindgebruikers en patiënten te betrekken bij deze evaluatie van fairness-risico's, bijvoorbeeld m.b.v. de aanpak begeleidingsethiek[24]. **(3.3c)**

3.4 Vaststellen van de uitkomstvariabele (labeling)

Het accuraat vaststellen van de te voorspellen uitkomst in de externe validatie dataset is een belangrijke factor voor de validiteit van de statistische voorspelkracht en de medische waarde. In de geneeskunde zijn er veel situaties waarin geen gouden standaard voorhanden is voor het meten van de uitkomstvariabele (bijv. voor sommige diagnosen, classificaties of oorzaak-specifiek overlijden), wat mogelijk tot misclassificatie van de uitkomstvariabele leidt[18,19]. Daarom wordt vaak de term referentiestandaard gehanteerd. In sommige situaties is beoordeling van een expert of groep experts nodig om tot een oordeel per casus te komen (bijv. de beoordeling van een tumor op een CT scan[20]).

Het zogenaamde *labelen* van uitkomsten in de dataset voor externe validatie **moet** in deze fase zo accuraat mogelijk te worden gedaan en zo transparant mogelijk worden vastgelegd en verantwoord. **(3.4a)**

Daarbij wordt **aanbevolen** precies bij te houden en te rapporteren welke experts betrokken waren bij het labelen (bijv. opleiding, expertise), in welke omstandigheden (bijv. aantal experts per casus, beschikbare tijd), en hoe eventuele discrepanties tussen labels werden opgelost. **(3.4b)**

Sterk aanbevolen wordt om de kwaliteit van het labelen te kwantificeren (bijv. door middel van *measures of agreement* (bijv. de *kappa statistic* of ICC), of door de accuraatheid van het labelen in te schatten[21]). **(3.4c)**

3.5 Grootte van de dataset voor externe validatie

Het uitgangspunt voor het kiezen van de grootte van de dataset voor externe validatie is: hoe groter, hoe beter. Hoe groter de dataset, hoe preciezer de schattingen die worden gebruikt voor de statistische en medische evaluatie en hoe beter algoritmische bias kan worden

onderzocht. Wel dient het belang van accurate labels (zie sectie 3.4) in het oog gehouden te worden. Voor een berekening van de minimale grootte van een dataset verwijzen we naar de literatuur[22,23].

De grootte van de dataset voor externe validatie **moet** worden beargumenteerd. **(3.5a)**

Berekening van de minimale grootte van de dataset, indien mogelijk (zie sectie 2.5), wordt **aanbevolen**. **(3.5b)**

3.6 Vastlegging, reproduceerbaarheid en repliceerbaarheid

Net als in fase 2 zijn reproduceerbaarheid en repliceerbaarheid belangrijke uitgangspunten voor externe validatie van het AIPA.

Om *reproduceerbaarheid* (d.w.z. heruitvoering van de externe validatie met andere data) te garanderen **moet** het gevolgde proces en de gebruikte data voor externe validatie volledig en transparant worden vastgelegd[8,9], ook in het geval van negatieve resultaten. **(3.6a)** De TRIPOD reporting guidelines[7-9] (www.tripod-statement.org) dienen daarbij als leidraad.

Ook wordt **aanbevolen** om de computercodes die worden gebruikt voor de externe validatie openbaar beschikbaar te stellen. **(3.6b)**

Om de repliceerbaarheid (d.w.z. heruitvoering van de externe validatie met zelfde data) te vergroten, wordt **aanbevolen** om de data (openbaar) beschikbaar te stellen. **(3.6c)**

Daarbij dient natuurlijk rekening te worden gehouden met de regelgeving omtrent privacy en daaruit voortkomende beperkingen. Ook mogen deze aanbevelingen afgewogen worden tegen commerciële belangen, en kan als argument aangevoerd worden dat externe validatie door een vertrouwde derde heeft plaatsgevonden.

3.7 Referenties

1. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;**19**:453–473.
2. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;**338**:b605–b605.
3. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, Collins GS. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016;i3140.
4. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, Woodward M. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;**98**:691–698.
5. Harrell FE. Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.
6. Steyerberg EW. Clinical Prediction Models. Cham: Springer International Publishing; 2019.
7. Usher-Smith JA, Sharp, SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. *BMJ* 2016;i3139.
8. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med* 2015;**162**:55.
9. Moons KGM, Altman DG, Reitsma JB, Ioannidis JP a, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann Intern Med* 2015;**162**:W1–W73.
10. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015;**68**:279–289.
11. Vickers AJ, Calster B van, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019;**3**:18.
12. Jenniskens K, Lagerweij GR, Naaktgeboren CA, Hooft L, Moons KGM, Poldervaart JM, Koffijberg H, Reitsma JB. Decision analytic modeling was useful to assess the impact of a prediction model on health outcomes before a randomized trial. *J Clin Epidemiol* 2019;**115**:106–115.
13. Van Giessen A, Wilcher B, Peters J, Hyde C, Moons KG, Wit GA de, Koffijberg H. Health economic evaluation of diagnostic and prognostic models. A systematic review. *Value in Health* 2014;**17**:A560.

14. Suresh H, Guttag JV. A Framework for Understanding Unintended Consequences of Machine Learning. *ArXiv190110002 Cs Stat* 2020;
15. Luijken K, Groenwold RHH, Van Calster B, Steyerberg EW, Smeden M. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective. *Stat Med* 2019;sim.8183.
16. Luijken K, Wynants L, Smeden M van, Van Calster B, Steyerberg EW, Groenwold RHH, Timmerman D, Bourne T, Ukaegbu C. Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *J Clin Epidemiol* 2020;**119**:7–18.
17. Ethics guidelines for trustworthy AI | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (25 June 2021)
18. Rutjes AWS, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PMM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess* Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academical Medical Center, University of Amsterdam, The Netherlands; 2007;**11**:ix–51.
19. Naaktgeboren CA, Groot JAH de, Rutjes AWS, Bossuyt PMM, Reitsma JB, Moons KGM. Anticipating missing reference standard data when planning diagnostic accuracy studies. *BMJ* 2016;i402.
20. Bertens LCM, Broekhuizen BDL, Naaktgeboren CA, Rutten FH, Hoes AW, Mourik Y van, Moons KGM, Reitsma JB. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. *PLoS Med* 2013;**10**:e1001531.
21. Jenniskens K, Naaktgeboren CA, Reitsma JB, Hooft L, Moons KGM, Smeden M van. Forcing dichotomous disease classification from reference standards leads to bias in diagnostic accuracy estimates: A simulation study. *J Clin Epidemiol* 2019;**111**:1–10.
22. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, Smeden M, Snell KIE. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med* 2021;sim.9025.
23. Archer L, Snell KIE, Ensor J, Hudda MT, Collins GS, Riley RD. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Stat Med* 2021;**40**:133–146.
24. ECP. Handleiding aanpak begeleidingsethiek voor AI in de zorg. https://begeleidingsethiek.nl/publicaties/handleiding-aanpak-begeleidingsethiek-voor-ai-in-de-zorg/. Published January 15, 2021. Accessed December 8, 2021.

4 Ontwikkeling van de benodigde software

Auteurs

Ilse Kant, Maarten van Smeden, Hine van Os, Egge van der Poel, Floor van Leeuwen, Pieter Boone, Giovanni Cina, Maurits Kaptein, Marcel Hilgersom, Martijn van der Meulen, Lotty Hooft, Carl Moons, Niels Chavannes

In fase 2 en fase 3 zijn de ontwikkeling en de validatie van het AIPA als voorspellend model behandeld. Fase 4 beslaat de verdere ontwikkeling van de software rondom het AIPA door de fabrikant. Dat wil zeggen: het ontwerp, de ontwikkeling, de gebruikerstesten en de bijbehorende systeemeisen aan de software waar het AIPA deel van uitmaakt (hierna ook: AIPA software) vallen onder deze fase. Onderdeel van deze systeemeisen zijn in te bouwen voorzieningen in het kader van de inrichting van een kwaliteitsmanagementsysteem. Daarnaast behoort de informatie die wordt geleverd bij de software tot deze fase.

Opgemerkt dient te worden dat de ontwikkelaar en de zorgorganisatie waar het AIPA geïmplementeerd wordt dezelfde partij kunnen zijn. In die gevallen is er geen sprake van interactie tussen een fabrikant en een zorgorganisatie. In dat geval dienen de eisen en aanbevelingen gericht aan de fabrikant als eisen en aanbevelingen aan de ontwikkelende zorgorganisatie gelezen te worden, en kunnen schijnbare duplicaties die daardoor ontstaan genegeerd worden.

Vanaf deze fase wordt verondersteld dat de software waar het AIPA deel van uitmaakt bedoeld is voor inzet in de beoogde medische context, inclusief zelfzorg. Fase 4 zal vaak na de externe evaluatie (fase 3) plaatsvinden, maar kan ook na fase 2 reeds plaatsvinden. In voorkomende gevallen vinden beide gelijktijdig plaats, bijvoorbeeld omdat de evaluatie van het AIPA alleen in combinatie met de software of het software-bevattende hulpmiddel waar het AIPA deel van uitmaakt kan plaatsvinden.

4.1 Uitlegbaarheid, transparantie, design en informatie

4.1.1 Uitlegbaarheid, transparantie en ontwerp van de AIPA software

De uitkomsten van het AIPA model worden in de software op een transparante en uitlegbare wijze gepresenteerd. In de presentatie van de uitkomsten van het AIPA model in de software wordt onderscheid gemaakt tussen een inherent uitlegbaar en een complex model.

Een inherent uitlegbaar model (bijv. een beslisboom of algoritme waarin de gewichten van de inputvariabelen inzichtelijk zijn) is een model waarvan direct te interpreteren is hoe de modelvoorspellingen (of -classificaties) tot stand zijn gekomen.

In het geval van een inherent uitlegbaar model, **moet** de fabrikant informatie over de interpretatie van het model en de modelvoorspellingen beschikbaar maken voor de beoogde eindgebruikers. **(4.1.1a)** Dit **moet** in een de op de eindgebruiker gerichte presentatie van de modelvoorspellingen door de software, dit geldt in het bijzonder als op basis van de modelvoorspellingen medische beslissingen worden genomen. **(4.1.1b)**

Dit kan bijvoorbeeld worden ingericht door een uitleg te geven in de gebruikersinterface waar ook de voorspelling of uitkomst gepresenteerd wordt.

In het geval van een complex model (bijv. een algoritme op basis van *deep learning*), is de relatie tussen inputvariabelen en voorspelde uitkomsten dusdanig complex is dat deze niet meer te overzien is (zogenoemde 'black-box' algoritmen). Post-hoc informatie en interpretatie van het model heeft dan extra aandacht nodig in de presentatie van het model door de software[1-3] .

Voor complexe modellen **moet** het volgende worden onderbouwd: 1) waarom er niet voor een uitlegbaar model is gekozen, en 2) als er wordt gekozen voor een post-hoc uitleg, waarom deze passend is bij het model en de bedoelde eindgebruiker. **(4.1.1c)**

Sterk aanbevolen wordt om in beide gevallen een aangepaste modelpresentatie en -uitleg per eindgebruiker te ontwerpen, en onderwijsmiddelen op het gebied van interpretatie van het AIPA te ontwikkelen en beschikbaar te stellen (fase 6), om verkeerde model interpretaties bij de bedoelde eindgebruiker te voorkomen. **(4.1.1d)**

Hierbij dient rekening gehouden te worden met: bruikbaarheid en testen van de software waar het AIPA model deel van uit maakt volgens de bestaande normen en regelgeving (zie sectie 4.4), presentatie van de voorspelde uitkomsten inclusief informatie over (on)zekerheid (bijvoorbeeld door een betrouwbaarheidsinterval weer te geven), taalgebruik passend bij de eindgebruiker (bijvoorbeeld voor een zorgverlener anders dan voor een patiënt), intuïtieve visualisatie, de integratie van de software in het zorg- en werkproces en voor zover van toepassing de mogelijkheid tot interactie met het AIPA model.

Sterk aanbevolen wordt om bij de het ontwerp van de modelpresentatie en -uitleg stakeholders zoals gebruikers en patiënten, cliënten of burgers te betrekken. **(4.1.1e)**

4.1.2 Informatie behorend bij de software

Het is voor gebruikers van de software van belang dat zij de eigenschappen van de software kennen en begrijpen. De informatiebehoefte van de eindgebruikers is dan ook leidend voor de wijze waarop de eigenschappen van de software helder en eenduidig moeten worden toegelicht.

Het doel van het verstrekken van deze informatie is om:

- De eindgebruiker inzicht te geven in het beoogde doel en de werking van de software en daarmee vertrouwen te wekken;
- De eindgebruiker van voldoende informatie te voorzien om de essentie van het AIPA te kunnen uitleggen aan derden (bijv. de patiënt, cliënt of burger indien deze geen eindgebruiker is), dat wil zeggen de betekenis van de output te kunnen interpreteren in de beoogde medische context;
- De betrouwbaarheid van de software te kunnen onderbouwen.

Sterk aanbevolen wordt om een digitale bijsluiter op te stellen voor de eindgebruiker met informatie over het gebruik van het AIPA in de software. **(4.1.2a)**

Hierin wordt aanbevolen om het volgende op te nemen:

- Voor wie de output van het AIPA in de software bedoeld is, en in welke medische context, bijv.:
    - Zorgorganisatie (niet-zorginhoudelijk)
    - Zorgverlener (intercollegiaal)
    - Zorgverlener (in gesprek met patiënt, cliënt)
    - Patiënt, cliënt of burger (al dan niet in hun eigen omgeving)
- Op welke wijze (periodiek) de informatiebehoefte van de eindgebruiker is en wordt bepaald;
- De informatiebehoefte van de eindgebruiker en het voorzien in deze informatie, bijv. door het beantwoorden van de volgende vragen:
    - Zorgverlener (en tot op zekere hoogte ook patiënt, cliënt en burger):
        - Is deze AIPA software al elders toegepast?
        - Waarop is de voorspelling van het AIPA gebaseerd? Welke inputvariabelen zijn gebruikt en welke methodiek is gebruikt om deze data te verwerken tot een AIPA? Waar vind ik informatie over gebruikte trainingsdata, en het beoogde doelgebruik (zie ook fase 1 t/m 3)?
        - Zijn er dominante inputvariabelen aan te wijzen in relatie tot de voorspellingen van het AIPA?
        - Hoe zeker zijn de voorspellingen (zie ook fase 2 en 3)? Waar kan ik aanvullende informatie vinden over welke validatiedata is gebruikt, welke processen en methodiek hierbij gevolgd zijn en wat de resultaten waren (zie ook fase 3)?
    - Patiënt, cliënt of burger:
        - Welke invloed heeft de toepassing van de AIPA software op het proces met mijn zorgverlener?
        - Wat betekent deze voorspelling voor mij als persoon?
        - Waar kan ik terecht als ik een vraag of klacht heb of meer informatie wil over de software of het AIPA?
        - Hoe zeker kan men van de voorspellingen zijn (zie ook fase 2 en 3)?

4.2 Voorzieningen voor continue monitoren

Continue monitoren van het AIPA is een belangrijk onderdeel van kwaliteitsmanagement en een vereiste vanuit de MDR. Voor de inrichting van een kwaliteitsmanagementsysteem conform MDR verwijzen we naar *ISO 13485 Medical devices - Quality management systems - Requirements for regulatory purposes.*

Sterk aanbevolen wordt om in deze fase een monitoringsplan op te stellen, zodat de AIPA software hierop kan worden ingericht (zie sectie 6.2.1). **(4.2a)**

De fabrikant maakt in het monitoringplan onderscheid tussen de eigen informatiebehoeftes en die van de zorgorganisatie.

Sterk aanbevolen wordt om de mogelijkheid om gebruikte data, het model en gebruik van het AIPA te monitoren na introductie in de praktijk (zie fase 6), te faciliteren in de software zodat in ieder geval de zorgorganisatie hiervan gebruik kan maken. **(4.2b)**

Van deze aanbeveling kan afgeweken worden als deze functionaliteit geen toegevoegde waarde heeft voor het beoogd doelgebruik.

Aanbevolen wordt een mogelijkheid in de software in te bouwen om te registreren of de eindgebruiker daadwerkelijk de voorspelling (of classificatie of behandeladvies) van het AIPA volgt of niet (en zo niet, waarom niet) zodat in ieder geval de zorgorganisatie hier gebruik van kan maken. Let hierbij ook op bijvoorbeeld de mogelijkheid van inzet voor schaduwdraaien of testen. **(4.2c)**

Van deze aanbeveling kan afgeweken worden als deze functionaliteit geen toegevoegde waarde heeft voor het beoogd doelgebruik.

Sterk aanbevolen wordt om input data geautomatiseerd te valideren in de software. **(4.2d)**

Bijvoorbeeld door *out-of-domain detection*, waarbij bepaalde voorspellingen die buiten het domein vallen niet worden gepresenteerd aan de eindgebruiker.

Daarnaast wordt **sterk aanbevolen** om in de software te monitoren op systematische verschuivingen in de data. **(4.2e)**

Bijvoorbeeld door detectie van *data drift* waarbij de software automatisch aangeeft aan de fabrikant en aan de beheerder van de software wanneer er systematische verschuivingen in de data plaats lijken te vinden.

De fabrikant van de AIPA software **moet** als onderdeel van het kwaliteitsmanagementsysteem een mogelijkheid voor terugkoppeling faciliteren, waar mogelijk in de software, voor feedback van eindgebruikers en voor rapportage van technische problemen. **(4.2f)**

De inrichting en hoeveelheid informatie die wordt gevraagd in deze terugkoppeling verschilt per toepassing en kan verschillen per fase van de ontwikkeling van de AIPA software. Bijvoorbeeld: in het begin is de mogelijkheid tot een handmatige controle en verificatie van voorspellingen van het AIPA sterk aan te raden.

4.3 Beveiliging

In het algemeen zijn de bestaande normen en regelgeving voor de beveiliging van software dekkend voor AIPA software in de medische context, daarom wordt verwezen naar de bestaande normen, regelgeving en richtsnoeren voor beveiliging van (medische) software, in het bijzonder:

- MDS2 statement: manufacturer disclosure statement for medical device security, ISO27002, NEN 7510.
- MDCG 2019-16 European Commission: Guidance on Cybersecurity for Medical Devices.
- UL 2900-1 ANSI/CAN/UL Standard for Software Cybersecurity for Network-Connectable Products.
- IMDRF/CYBER WG/N60 Principles and Practices for Medical Device Cybersecurity.
- FDA (laatste guidance 2014 https://www.fda.gov/medical-devices/digital-health-center-excellence/cybersecurity, draft 2019).
- Algemene Verordening Gegevensbescherming (AVG).
- ISO/IEC TS 7110, ISO/IEC 27032, ISO/IEC 27014.

Voor AIPA software is specifiek van belang dat input- en output data in toenemende mate in grote hoeveelheden zal worden opgeslagen en gebruikt, met name bij het opnieuw trainen en opnieuw kalibreren van modellen.

Sterk aanbevolen wordt daarom erop toe te zien dat eindgebruikers die werken met de software of het hulpmiddel dat het AIPA bevat, door de zorgorganisatie getraind worden op het gebruik van beveiligde (cloud) systemen, en de normen en regelgeving op het gebied van data-delen, -veiligheid en -privacy. **(4.3a)**

Er **moet** gewerkt worden met versiemanagement voor de software en de benodigde en gebruikte trainings- en testdatasets voor de ontwikkeling, het valideren en het aanpassen van het AIPA moeten samen met de corresponderende versie worden opgeslagen (zie ook fase 1 en 2). **(4.3b)**

4.4 Software testen

In het algemeen zijn de bestaande normen en regelgeving voor het testen van software dekkend voor AIPA software in de medische context, daarom wordt verwezen naar de volgende bestaande normen en regelgeving voor testen van (medische) software. Het uitgangspunt is hierbij dat er een volledige traceerbaarheid moet zijn van de vertaling van het beoogd doeleind *(intended use)* naar de software eisen en design. Deze vertaling moet vervolgens worden geverifieerd en gevalideerd. Hiervoor wordt verwezen naar de volgende bestaande normen en richtsnoeren:

- IEC 62304 - Medical device software - Software life-cycle processes.
- IEC 82304-1 - Health software - Part 1: General requirements for product safety.
- IEC 62366-1 - Medical devices - Part 1: Application of usability engineering to medical devices.
- ISO 14971 - Medical devices - Application of risk management to medical devices.
- FDA, General principles of software validation, 2002.
- FDA, off-the-shelf software use in medical devices, 2019.

Wanneer componenten van de AIPA software zijn ontwikkeld door een derde partij, die niet door de fabrikant beheerd worden (ook wel off-the-shelf of OTS software genoemd) wordt **sterk aanbevolen** die componenten lokaal te testen volgens bestaande standaarden. **(4.4a)**

Zie bijvoorbeeld de richtlijn van de FDA, off-the-shelf software use in medical devices uit 2019.

Daar waar het AIPA model zelf kan worden aangemerkt als off-the-shelf component wordt **sterk aanbevolen** om fase 3, Validatie van het AIPA, (opnieuw) uit te voeren. **(4.4b)**

4.5 Referenties

1. Magrabi F, Ammenwerth E, McNair JB, et al. Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications. *Yearb Med Inform* 2019;28(1):128-34. doi: 10.1055/s-0039-1677903 [published Online First: 2019/04/26]
2. Cearns M, Hahn T, Baune BT. Recommendations and future directions for supervised machine learning in psychiatry. *Transl Psychiatry* 2019;9(1):271. doi: 10.1038/s41398-019-0607-2 [published Online First: 2019/10/24]
3. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17(1):195. doi: 10.1186/s12916-019-1426-2 [published Online First: 2019/10/31]

5 Effectbeoordeling van het AIPA in combinatie met de software

Auteurs

Ilse Kant, Maarten van Smeden, Hine van Os, Teus Kappen, Jonas Teuwen, Leo Hovestadt, Ewout Steyerberg, René Drost, Sade Faneyte, René Verhaart, Lotty Hooft, Carl Moons, Niels Chavannes

Fase 5 beslaat het bepalen van de impact of meerwaarde van het gebruik van het AIPA als onderdeel van de software op respectievelijk de beoogde medische praktijk of context, het medisch handelen en de gezondheidsuitkomsten van de beoogde doelgroep (bijv. de patiënt, cliënt of burger). Ook een een Health technology assessment vindt plaats in deze fase[2]. Het bepalen van impact en meerwaarde is een verantwoordelijkheid van de fabrikant (of in het geval van interne ontwikkeling, de ontwikkelende zorgorganisatie), maar vindt in de regel in samenwerking met ontwikkelaars, zorgorganisaties en eindgebruikers plaats. Impact of meerwaarde van het gebruik van het AIPA, inclusief de benodigde software, op en in de medische praktijk kan op verschillende manieren worden bereikt. Bijvoorbeeld door het ondersteunen van de zorgverlener, patiënt, cliënt, of burger bij het maken van behandel- of leefstijlbeslissingen of door een verandering in het zorgproces die efficiënt (kostenbesparend) werkt.

Op dit moment zijn er nog relatief weinig AIPA's in gebruik in de dagelijkse medische praktijk, en daarmee is ook van relatief weinig AIPA's de impact onderzocht[3-6]. De aansluiting van het AIPA bij de dagelijkse medische praktijk en zorg blijkt een struikelblok. De hoop en verwachting is dat de toepassing in de komende jaren zal toenemen. Voor een overzicht van de stand van zaken omtrent de toepassing van AI in de medische context en praktijk wordt verwezen naar het rapport 'Inventarisatie AI in gezondheid en zorg'[7].

5.1 Effectbeoordeling en bijbehorende studie opzetten

De fabrikant **moet** een effectbeoordeling van het AIPA (als onderdeel van de software) binnen het beoogde doelgebruik uitvoeren. **(5.1a)**

Dit is nodig om de potentiële toegevoegde waarde van het AIPA in de dagelijkse medische praktijk te toetsen. Naast de generieke methoden voor het empirisch toetsen van de (meer)waarde van (digitale) innovaties en predictiemodellen in de gezondheidszorg[8-13], worden enkele (AI-specifieke) stappen behandeld die in het bijzonder van belang zijn voor een AIPA[14-16].

De ontwikkeling van de software waar het AIPA deel van uitmaakt en de bijbehorende effectbeoordeling **moet** een proces zijn, waarbij de fabrikant ervoor zorgt dat eindgebruikers (bijv. zorgverleners) en patiënten, cliënten of burgers zo vroeg mogelijk worden betrokken en meerdere contactgelegenheden krijgen. **(5.1b)**

Wanneer de effectbeoordeling plaatsvindt binnen een zorgorganisatie kan een effectbeoordeling als een vorm van implementatie worden beschouwd.

In het geval dat de effectbeoordeling (deels) binnen het zorgproces wordt uitgevoerd en daarmee dus wordt geïmplementeerd, wordt **sterk aanbevolen** een implementatieplan, zoals beschreven in sectie 6.1, op te stellen en relevante onderdelen, zoals het invullen van

het implementatieteam en een lokale evaluatie, een pilot of run-in periode, uit te voeren vóórdat een grootschalige empirische studie wordt gestart. **(5.1c)**

In het geval van ontwikkeling binnen een zorgorganisatie **moet** aanbeveling 5.1c worden gevolgd. **(5.1d)**

Eventuele benodigde wijzigingen aan de software kunnen dan nog worden doorgevoerd voeren voorafgaand aan de interventieperiode als onderdeel van de effectbeoordeling.

De effectbeoordeling wordt besproken aan de hand van een stappenplan. Dit stappenplan beschrijft het in kaart brengen van de te verwachten effecten van het gebruik van software met een AIPA in de medische praktijk tot aan het ontwerp van een vergelijkende empirische studie om de verwachte effecten aan te kunnen tonen t.o.v. .de huidige zorg(processen) in de beoogde context. Hierbij wordt uitgegaan van het beoogd doelgebruik van het AIPA (zie fase 2) en de bijbehorende *indications* of *claims* die worden gemaakt door de ontwikkelaar (zie fase 4). Dit laatste is vooral van belang om aan te sluiten bij huidige wet- en regelgeving (MDR), en daarmee voldoende bewijs op meerwaarde te leveren dat nodig is om in de medische praktijk te kunnen worden geïntroduceerd.

De stappen zijn de volgende:

1. Verwachte effecten: breng aan de hand van het beoogd doeleind in kaart hoe men verwacht dat het AIPA effect zal hebben op het beoogde medische zorgproces en gezondheidsuitkomsten in de beoogde medische context
2. Risico-inventarisatie: schat mogelijke risico's en onbedoelde effecten voorafgaand aan implementatie van het AIPA in de dagelijkse praktijk in;
3. Mens-machine interactie: Expliciteer hoe zorgproces en zorgverlener interacteren met de software voor een empirische studie uitgevoerd wordt.

5.1.1 De verwachte effecten

Er **moet** duidelijk gemaakt worden hoe het AIPA opereert: zelfstandig of adviserend volgens het niveau van automatisering (beter bekend als *level of automation*; box 5.1). **(5.1.1a)**

Het niveau van automatisering kan invloed hebben op de classificatie in termen van de eerder genoemde regel 11 in bijlage VIII van de MDR.

In fase 2 is het beoogd doelgebruik vastgelegd dat na ontwikkeling van de software wordt opgenomen in een digitale bijsluiter, zie sectie 4.1.2.

In aanvulling op het vastgelegde beoogd doelgebruik, **moet** uitgebreider worden vastgesteld wat de te verwachten effecten van het AIPA gebruik zijn op mogelijke relevante (gezondheids- en proces) uitkomsten (m.a.w. definieer het *beoogd doeleind* ofwel *intended use* van het AIPA). **(5.1.1b)**

Deze vaststelling valt te onderscheiden in twee lagen: 1) De verwachte beslissingen van de eindgebruiker op basis van de voorspelde uitkomsten of classificaties; 2) de verwachte effecten en consequenties van deze voorspellingen en beslissingen op latere (gezondheids)uitkomsten van de patiënt, cliënt, burger, en/of op de lokale zorgcontext of de maatschappij.

Box 5.1: niveau van automatisering

Op het hoogste niveau van automatisering maakt een AIPA zelfstandig een voorspelling en neemt zelfstandig een medische beslissing, waarbij de bijbehorende handeling wordt uitgevoerd en niet wordt gecontroleerd door een mens. Dit komt in de klinische praktijk nagenoeg niet voor, hoewel ook in de klinische praktijk op dit vlak wel ontwikkelingen zijn met name op het gebied van radiologische beeldvorming, waarbij de mens niet altijd meer betrokken wordt in de beoordeling van de radiologische testbeelden[1].

Een niveau lager zijn *human-monitored* oplossingen. Hierbij maakt de software zelfstandig een voorspelling of classificatie, neemt zelfstandig de medische vervolgkeuze of beslissing en voert deze uit, maar laat zich - daarvoor of daarna - controleren door een mens (veelal zorgverlener of patiënt). Wanneer de menselijke beslissing wordt gemaakt op basis van het advies van de software, spreken we van *joint*- of *computer-assisted decision making*. Als in de zorg over *clinical decision support* gesproken wordt, gaat het meestal over dit niveau van automatisering.

Figuur 1: het niveau van automatisering. Verreweg de meeste AIPA's in softwaretoepassingen bevinden zich tot nu toe in de *computer assisted*' of *joint decision making* categorie.

Bij het vaststellen van de mogelijke uitkomsten wordt **aanbevolen** rekening te houden met de doelen zoals gedefinieerd in het *quadruple aim model* voor het verbeteren van de zorgsector, dat wordt toegepast in waardegedreven zorg[18]:

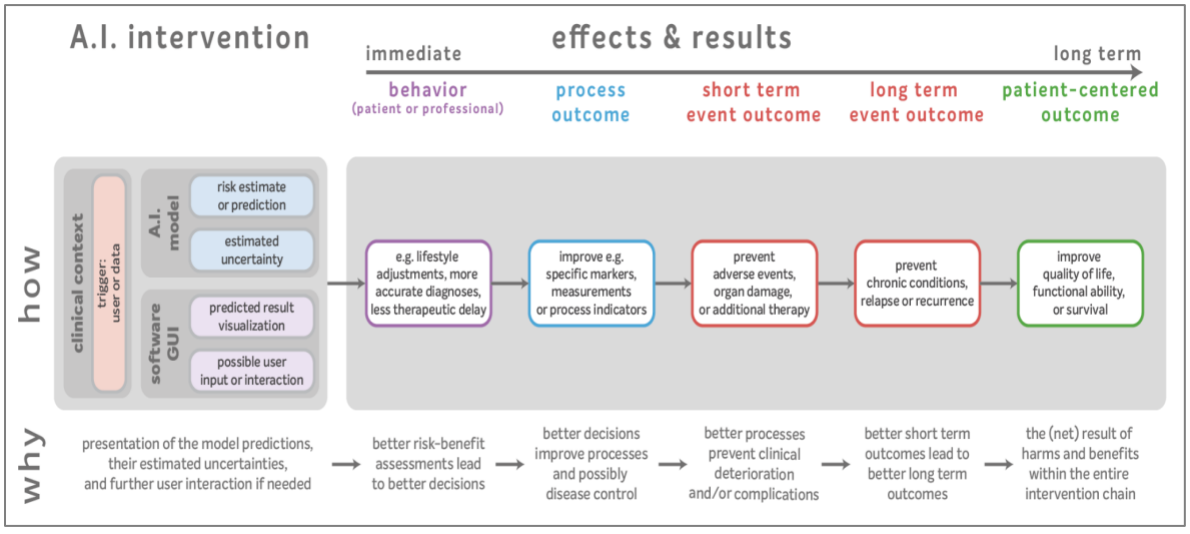
1) het verbeteren van de patiënt/burgerervaring m.b.t. de kwaliteit van zorg;
2) het verlagen van de zorgkosten;
3) het verbeteren van de gezondheid van de bevolking;

4) het verbeteren van de beleving van de zorgverlener. **(5.1.1c)**

Mogelijke uitkomsten waaraan gedacht kan worden zijn[12 14 17]:

- *Procesuitkomsten en gebruiksvriendelijkheid*
  Leidt de introductie van het AIPA tot directe veranderingen in het zorgproces voor de zorgverlener (bijv. het sneller of beter stellen van een diagnose of prognose, mogelijkheid tot eerder ziekenhuisontslag of -opname, verbeterde werkprocessen zoals thuis- of zelfmonitoring van de patiënt)?
- *Korte termijn gezondheidsuitkomsten*
  Leidt het veranderde proces na introductie van het AIPA tot verbeterde gezondheidsuitkomsten bij de individuele patiënten/burgers op de korte termijn?
- *Lange termijn gezondheidsuitkomsten*
  Leidt het veranderde proces na introductie van het AIPA tot verbetering van lange termijn gezondheidsuitkomsten bij de individuele patiënten/burgers op de lange termijn, of tot een verbeterde preventie (bijv. betere overleving, hogere kwaliteit van leven, of betere dagelijkse functionaliteit)?
- *Maatschappij*
  Leidt het veranderde proces na introductie van het AIPA tot veranderingen op maatschappelijk niveau (bijv. kosteneffectiviteit)?

Teken hiervoor de verwachte keten van het medische zorgproces waar de software en het AIPA deel van gaat uitmaken uit. Ter illustratie daartoe zie figuur 2.



Figuur 2: voorbeeld van het uittekenen van de verwachte keten van gebeurtenissen en het zorg- of werkproces bij het introduceren van een AIPA in de beoogde medische context.

Sterk aanbevolen wordt om de inschattingen van te verwachten effecten zoals gevraagd in 5.1.1b, te maken in een multidisciplinair verband, in samenspraak met de eindgebruiker(s) en patiënten, cliënten of burgers. **(5.1.1d)**

Kijk hierbij naar elke stap in de zorgketen van gebeurtenissen waar het AIPA deel van gaat uitmaken zoals uitgetekend in figuur 2. Houd rekening met de *waarschijnlijkheid* (hoe groot is de kans dat dit effect optreedt?) en de mogelijke *consequenties* (positief en negatief). Deze inschattingen kunnen worden gebruikt om risico mitigerende maatregelen te nemen (volgende stap), en een vergelijkende empirische studie op te zetten.

5.1.2 Risico-inventarisatie

Er **moet** een risico-inventarisatie worden uitgevoerd om de mogelijke risico's van het gebruik van het AIPA in de dagelijkse medische praktijk in kaart te brengen. **(5.1.2a)**

Hiertoe behoren de verwachte onbedoelde beslissingen en effecten in het gehele zorgproces (zie sectie 5.1.1) en redelijkerwijs voorzienbaar verkeerd gebruik.

De mogelijke ongewenste effecten (risico's) van implementatie van het AIPA in het zorgproces **moeten** per onderdeel van het zorgproces in kaart worden gebracht in de risico-inventarisatie in nauwe samenwerking met de stakeholders (bijv. eindgebruikers en patiënten). **(5.1.2b)**

Op basis van de in de risico-inventarisatie geïdentificeerde risico's **moeten** risicomitigerende maatregelen worden gekozen en geïmplementeerd. **(5.1.2c)**

Bronnen van onzekerheid zoals benoemd in 5.3.1a **moeten** worden meegenomen in de risico-inventarisatie. **(5.1.2d).**

Geïdentificeerde risico's in de risico-inventarisatie **moeten** worden meegenomen in de uitkomsten van de empirische studie (zie sectie 5.1.4). **(5.1.2e)**

Sterk aanbevolen wordt om bij de risico-inventarisatie stakeholders zoals gebruikers en patiënten, cliënten of burgers te betrekken. **(5.1.2f)**

De hierboven beschreven risico inventarisatie door de fabrikant is vergelijkbaar met een prospectieve risico-inventarisatie (PRI) door een zorgorganisatie. Het verschil is dat bovenstaande risico-inventarisatie een algemene inventarisatie betreft, gericht op de gehele breedte van mogelijk gebruik, zoals in het beoogd doelgebruik beschreven is, in plaats van een PRI gericht gebruik van het AIPA in een specifieke zorgorganisatie. In het geval van

ontwikkeling door de zorgorganisatie kan een PRI worden uitgevoerd in plaats van bovengenoemde inventarisatie.

5.1.3 Mens-machine interactie

De effectiviteit van de interactie van de eindgebruiker met de software is van groot belang voor de impact van de AIPA software.

De AIPA software **moet** voordat een empirische studie wordt uitgevoerd zo goed mogelijk aansluiten bij de huidige medische zorgprocessen en bijbehorende medische besluitvorming (zie figuur 2 in sectie 5.1.1). **(5.1.3a)**

Daartoe **moeten** meerdere eindgebruikers in het lokale implementatieteam worden betrokken (zie ook fase 6 voor de verdere samenstelling van een implementatieteam). **(5.1.3b)**

Daarnaast **moeten** verwachte veranderingen in de zorgcontext (bijv. veranderingen in het werkproces) door interactie met de software in kaart gebracht worden, bij voorkeur in samenspraak met de beoogde gebruiker en patiënt, cliënt of burger. **(5.1.3c)**

Vanzelfsprekend moeten lokale medische richtlijnen die van toepassing zijn worden gevolgd. Indien nieuwe informatie beschikbaar is gekomen die aantoont dat door invoering van het AIPA het volgen van de huidige richtlijnen niet langer wenselijk blijkt te zijn (bijv. door een nieuw werkproces met behulp van het AIPA, dat zorgt voor minder risico voor patiënten of burgers), is dit aanleiding om een verandering aan te vragen in de lokale medische richtlijnen bij de verantwoordelijke organisatie (bijv. een medische beroepsvereniging).
Sterk aanbevolen wordt om de gewenste presentatie van de uitkomsten van het AIPA in de software (zie fase 4) en het bijbehorende werkproces waarin het AIPA gebruikt wordt (zie sectie 5.1.2) aan de eindgebruikers te demonstreren alvorens een empirische studie te starten, bij voorkeur in de vorm van een pilot, zie ook aanbeveling 5.1c. **(5.1.3d)**

Overweeg hierbij een directieve presentatie waarbij een advies tot handelen wordt uitgebracht aan de eindgebruiker[19].

Sterk aanbevolen wordt om te toetsen dat de opmaak van de software, de benodigde input en eventuele handelingen die worden gevraagd van de eindgebruiker passen bij de huidige workflow om optimaal gebruik aan te moedigen. Deze moeten onder meer aansluiten bij de *werkdruk* van de eindgebruiker. **(5.1.3e)**

Sterk aanbevolen wordt om hierbij ook de zogenaamde *faciliterende factoren en barrières* rondom de implementatie van het AIPA in de praktijk in kaart te brengen, door gebruik te maken van kwalitatieve onderzoeksmethoden zoals focusgroepen en vragenlijsten. Betrek in

dit onderzoek naast de ontwikkelaar van het AIPA en de fabrikant van de software altijd meerdere eindgebruikers en patiënten, cliënten of burgers[10 17]. **(5.1.3f)**

5.1.4 Vergelijkende studie

Om de meerwaarde van de implementatie van een AIPA in de context van de dagelijkse medische praktijk op valide wijze te kwantificeren **moet** een vergelijkende studie worden uitgevoerd, waarin de (gewenste en ongewenste) effecten van het gebruik van het AIPA (zie sectie 5.1.1) worden afgezet tegen eenzelfde context, waarin zoveel mogelijk dezelfde zorgprocessen worden toegepast zonder gebruik van het AIPA[9 11 12 17]. **(5.1.4a)**

Het ideale vergelijkende design is een gerandomiseerd *(adaptief) vergelijkend studiedesign* waarin twee groepen worden vergeleken: een groep waarin na randomisatie de standaard zorg wordt uitgevoerd (*controlegroep*) zonder gebruik van het AIPA en een groep waar het advies van het AIPA kenbaar wordt gemaakt aan de eindgebruiker en waarop gehandeld kan worden in de zorgpraktijk (*interventiegroep*). In het ideale geval gebruikt men hiervoor een gerandomiseerde studie (gerandomiseerd op individueel of clusterniveau). Indien hiervan wordt afgeweken moet dit worden onderbouwd. Redenen om af te wijken kunnen tenminste zijn: logistieke onhaalbaarheid om te randomiseren, het risico op contaminatie van de controlegroep, lange duur van een studie (bijv. in het geval van zeldzame ziekten of langetermijneffecten), onhaalbaar hoge kosten, of een onhaalbaar te includeren aantal patiënten[9 11-13].

Alternatieven voor een gerandomiseerde studie zijn: een gecontroleerde prospectieve voor-na studie, onderbroken tijd series, geografische vergelijking, of een cross-sectionele gerandomiseerde studie met de behandelingsbeslissing (i.p.v. de individuele gezondheidsuitkomsten) als effectmaten.

De keuze voor de populatie en context waarin de AIPA software wordt bestudeerd **moet** worden onderbouwd. **(5.1.4b)**

Sterk aanbevolen wordt om te kiezen voor een redelijkerwijs vergelijkbare populatie met de doelpopulatie waarvoor de software is ontwikkeld (fase 2). **(5.1.4c)**

Ook de keuze voor de controlegroep maakt deel uit van de keuze voor deze populatie. Sterk aanbevolen wordt om te kiezen voor een redelijkerwijs vergelijkbare groep met de interventiegroep, die dus ook direct representatief is voor de doelpopulatie. Waar onderwijs nodig is voor de interventieperiode (dat wil zeggen: het gebruik van het AIPA in praktijk), zal er voor de controlegroep eenzelfde hoeveelheid onderwijs moeten worden gegeven om het leereffect zoveel mogelijk te voorkomen[10].

Voor richtlijnen omtrent de rapportage van *randomized trials* waarin gebruik gemaakt wordt van kunstmatige intelligentie wordt verwezen naar de SPIRIT-AI en CONSORT-AI statements[20 21].

5.2 Health technology assessment

Sterk aanbevolen wordt om in fase 5 ook een modelmatige impactstudie, ofwel een modelmatige Health Technology Assessment (HTA), uit te voeren. **(5.2a)**

Dat wil zeggen, dat men door middel van een mathematisch model (bijv. een Markov model) een objectieve analyse van de verwachte kosten en baten (meerwaarde) van introductie van het AIPA in de medische praktijk maakt t.o.v. de huidige reguliere zorg als benchmark of controle[2 22 23]. Het resultaat van dergelijke HTA zal in toenemende mate een rol spelen in het goedkeuren van digitale gezondheidszorg in Nederland en de EU. Wanneer vergoeding noodzakelijk is, is een passende HTA daarom nodig om voor deze vergoeding, of conditionele vergoeding, van het gebruik in aanmerking te komen. In het rapport 'Waardevolle AI voor gezondheid' is een routekaart voor de uitvoering van een HTA voor AIPA software bijgevoegd. Deze routekaart geeft een overzicht van de kosten en de mogelijke financieringsbronnen van HTA onderzoek voor AI en dus ook voor een AIPA[24 25].

5.3 Onzekerheid, risico's en onverwachte uitkomsten

5.3.1 Onzekerheid in voorspellingen

Bronnen van onzekerheid van toepassing in fase 5 kunnen zijn: de toepasbaarheid van het AIPA in een andere medische context dan waarin het AIPA model oorspronkelijk is ontwikkeld (fase 2) of gevalideerd (fase 3), veranderingen in de lokale zorg- of werkprocessen en een systematische verandering in de mens-machine interactie zoals beschreven in 5.1.1, stap 3.

De fabrikant **moet** expliciteren welke bronnen van onzekerheid bestaan na uitvoeren van de effectbeoordeling en welke mitigerende maatregelen er zijn genomen om deze onzekerheden die men kan tegenkomen bij introductie in de dagelijkse zorgpraktijk te minimaliseren. **(5.3.1a)**

De ontwikkelaar en eindgebruikers dienen hierbij in het bijzonder aandacht besteden aan de transporteerbaarheid van het AIPA naar een andere medische setting en/of context[8 11 12 26].

5.3.2 Onverwachte uitkomsten, vigilantie

Onverwachte uitkomsten tijdens de effectbeoordeling **moeten** in alle gevallen worden vastgelegd en daarnaast gemeld in overeenstemming met wet- en regelgeving. **(5.3.2a)** Daarbij dient voor zover van toepassing in de zorgcontext de reeds bestaande wet- en regelgeving omtrent vigilantie en het lokale veiligheidsmanagement systeem[27] (VMS)

gevolgd te worden: MDR Artikel 5.5 (in huis ontwikkelde software), MDR Artikel 10 (kwaliteitssysteem fabrikant), Convenant Veilige Toepassing van Medische Technologie in de Medisch Specialistische Zorg (hierna: CMT)[28], MDR Artikel 80 (rapporteren van adverse events tijdens klinisch onderzoek), in het geval van een product met CE-markering: MDR Artikel 87 (veiligheidsrapportages). Fabrikanten van een AIPA zijn verplicht een systeem te handhaven voor risicomanagement, en een systeem voor het melden van incidenten en corrigerende acties in verband met de veiligheid in het veld, zowel tijdens de impactstudie (fase 5) als tijdens en na implementatie (fase 6). Voor de inrichting van een kwaliteitsmanagementsysteem conform MDR verwijzen we naar *ISO 13485 Medical devices - Quality management systems - Requirements for regulatory purposes.*

5.4 Referenties

1. Rodriguez-Ruiz A, Lang K, Gubern-Merida A, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *Eur Radiol* 2019;29(9):4825-32. doi: 10.1007/s00330-019-06186-9 [published Online First: 2019/04/18]
2. van Giessen A, Peters J, Wilcher B, et al. Systematic Review of Health Economic Impact Evaluations of Risk Prediction Models: Stop Developing, Start Evaluating. *Value Health* 2017;20(4):718-26. doi: 10.1016/j.jval.2017.01.001 [published Online First: 2017/04/15]
3. van Leeuwen KG, Schalekamp S, Rutten M, et al. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol* 2021;31(6):3797-804. doi: 10.1007/s00330-021-07892-z [published Online First: 2021/04/16]
4. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17(1):195. doi: 10.1186/s12916-019-1426-2 [published Online First: 2019/10/31]
5. Usher-Smith JA, Silarova B, Schuit E, et al. Impact of provision of cardiovascular disease risk estimates to healthcare professionals and patients: a systematic review. *BMJ Open* 2015;5(10):e008717. doi: 10.1136/bmjopen-2015-008717 [published Online First: 2015/10/28]
6. Kleinrouweler CE, Cheong-See FM, Collins GS, et al. Prognostic models in obstetrics: available, but far from applicable. *Am J Obstet Gynecol* 2016;214(1):79-90 e36. doi: 10.1016/j.ajog.2015.06.013 [published Online First: 2015/06/14]
7. KPMG. Rapport Inventarisatie AI in gezondheid en zorg in Nederland: KPMG, 2020.
8. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21(1):128-38. doi: 10.1097/EDE.0b013e3181c30fb2 [published Online First: 2009/12/17]
9. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10(2):e1001381. doi: 10.1371/journal.pmed.1001381 [published Online First: 2013/02/09]
10. Kappen TH, van Klei WA, van Wolfswinkel L, et al. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagn Progn Res* 2018;2:11. doi: 10.1186/s41512-018-0033-6 [published Online First: 2019/05/17]

11. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;98(9):691-8. doi: 10.1136/heartjnl-2011-301247 [published Online First: 2012/03/09]
12. Moons KG, Altman DG, Vergouwe Y, et al. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* 2009;338:b606. doi: 10.1136/bmj.b606 [published Online First: 2009/06/09]
13. Riley RD, van der Windt DA, Croft P, et al. Prognosis research in healthcare: concepts, methods, and impact. Oxford, United Kingdom: Oxford University Press 2019.
14. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 2020;368:l6927. doi: 10.1136/bmj.l6927 [published Online First: 2020/03/22]
15. Magrabi F, Ammenwerth E, McNair JB, et al. Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications. *Yearb Med Inform* 2019;28(1):128-34. doi: 10.1055/s-0039-1677903 [published Online First: 2019/04/26]
16. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019;25(9):1337-40. doi: 10.1038/s41591-019-0548-6 [published Online First: 2019/08/21]
17. Kappen TH, van Loon K, Kappen MA, et al. Barriers and facilitators perceived by physicians when using prediction models in practice. *J Clin Epidemiol* 2016;70:136-45. doi: 10.1016/j.jclinepi.2015.09.008 [published Online First: 2015/09/25]
18. Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider. *Ann Fam Med* 2014;12(6):573-6. doi: 10.1370/afm.1713 [published Online First: 2014/11/12]
19. Kappen TH, Vergouwe Y, van Wolfswinkel L, et al. Impact of adding therapeutic recommendations to risk assessments from a prediction model for postoperative nausea and vomiting. *Br J Anaesth* 2015;114(2):252-60. doi: 10.1093/bja/aeu321 [published Online First: 2014/10/03]
20. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med* 2020;26(9):1351-63. doi: 10.1038/s41591-020-1037-7 [published Online First: 2020/09/11]
21. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020;26(9):1364-74. doi: 10.1038/s41591-020-1034-x [published Online First: 2020/09/11]

22. Haverinen J, Keränen N, Falkenbach P, et al. Digi-HTA Health technology assessment framework for digital healthcare services. *Finnish Journal of eHealth and Welfare* 2019;11(4):326-41.
23. Jenniskens K, Lagerweij GR, Naaktgeboren CA, et al. Decision analytic modeling was useful to assess the impact of a prediction model on health outcomes before a randomized trial. *J Clin Epidemiol* 2019;115:106-15. doi: 10.1016/j.jclinepi.2019.07.010 [published Online First: 2019/07/23]
24. iMTA. Routekaart HTA onderzoek. Ministerie van Volksgezondheid, Welzijn en Sport: Erasmus University Rotterdam, 2021.
25. iMTA. Waardevolle AI voor gezondheid. Ministerie van Volksgezondheid, Welzijn en Sport: Erasmus University Rotterdam, 2021.
26. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130(6):115-524.
27. VMS. VMS Handleiding Prospectieve Risico Inventarisatie.
28. (NVZ) NVvZ, (NFU) NFvUMC. Convenant Veilige Toepassing van Medische Technologie in de medisch specialistische zorg, 2e druk, 2016.

6 Implementatie en gebruik van het AIPA met software in de dagelijkse praktijk

Auteurs

Ilse Kant, Maarten van Smeden, Hine van Os, Karen Wiegant, Laure Wynants, Anne de Hond, Bart Geerts, Nynke Breimer, Lysette Meuleman, Lotty Hooft, Carl Moons, Niels Chavannes

Fase 6 beslaat de implementatie en het gebruik van de AIPA in de gezondheidszorgverlening. Centrale thema's zijn in deze fase implementatie, monitoring en educatie door de zorgorganisatie. Voor een AIPA die deel uit maakt van een medisch hulpmiddel als bedoeld in de MDR, moet er in fase 6 ook rekening worden gehouden met de wettelijke eisen aan *post-market surveillance* voor fabrikanten. Opgemerkt dient te worden dat de ontwikkelaar en de zorgorganisatie waar het AIPA geïmplementeerd wordt dezelfde partij kunnen zijn. In die gevallen is er geen sprake van interactie tussen een fabrikant en een zorgorganisatie. In dat geval dienen de eisen en aanbevelingen gericht aan de fabrikant als eisen en aanbevelingen aan de ontwikkelende zorgorganisatie gelezen te worden en kunnen schijnbare duplicaties die daardoor ontstaan genegeerd worden.

Ook is het mogelijk dat het AIPA niet als zodanig bij een zorgorganisatie geïmplementeerd wordt, bijvoorbeeld wanneer een AIPA direct door patiënten, cliënten of burgers in de thuissituatie gebruikt wordt, zonder tussenkomt van een zorgverlener. In die gevallen zijn de subhoofdstukken 6.1, 6.2.2 en 6.3.2 niet van toepassing.

6.1 Implementatieplan

Wanneer een AIPA geïmplementeerd en toegepast wordt binnen een zorgorganisatie **moet** een *implementatieplan* worden opgesteld door de zorgorganisatie. **(6.1a).**

Een implementatieplan omvat zowel de technische implementatie van het AIPA en de software in de bestaande (IT) infrastructuur als de inbedding van het gebruik van het AIPA in bestaande werkprocessen. Voor een algemene leidraad voor de implementatie wordt verwezen naar Convenant Medische Technologie[9] en Leidraad Nieuwe Interventies in de Klinische Praktijk[10]

Als onderdeel van het lokale implementatieproces **moet** de betrouwbaarheid en toepasbaarheid van het AIPA geëvalueerd worden door middel van een beoordeling van (de resultaten van) eerdere studies uitgevoerd als onderdeel van fase 3 en 5. **(6.1b)**

Indien deze resultaten onvoldoende indicatie geven van de betrouwbaarheid en toepasbaarheid van het AIPA binnen de lokale context, kan een aanvullende validatie van het AIPA worden uitgevoerd (zie ook fase 3).

Daarnaast **moet** het AIPA en het bijbehorende werkproces waarin het AIPA gebruikt wordt op gecontroleerde wijze worden geïntroduceerd in het zorgproces, bijvoorbeeld in de vorm van een pilot, run-in periode of door middel van het schaduwdraaien van het AIPA naast het traditionele zorgproces. **(6.1c)**

Er **moet** een prospectieve risico-inventarisatie (PRI) worden uitgevoerd om de mogelijke risico's van het gebruik van het AIPA in de dagelijkse medische praktijk in kaart te brengen. **(6.1d)**

In een PRI wordt geëxpliciteerd welke medische relevantie iedere verwachte fout heeft en hoe waarschijnlijk het is dat deze fout op kan treden. Een PRI is onderdeel van het *veiligheidsmanagementsysteem* van een specifieke zorgorganisatie. Voor de uitvoering van een PRI wordt verwezen naar de VMS handleiding Prospectieve Risico Inventarisatie (VMS)[1].

Sterk aanbevolen wordt om de bevindingen uit de effectbeoordeling (*impact assessment*) en specifiek de risico-inventarisatie zoals uitgevoerd door de fabrikant in fase 5, indien van beschikbaar, expliciet op te nemen in de PRI. **(6.1e)**

De in de PRI geïdentificeerde risico's **moeten** worden geëvalueerd, op basis hiervan worden risico's expliciet geaccepteerd worden of risicomitigerende maatregelen worden gekozen. Zowel acceptatie van risico's als eventuele risico mitigerende maatregelen worden opgenomen in het implementatieplan. **(6.1f)**

Sterk aanbevolen wordt een gegevensbeschermingseffectbeoordeling (GEB) in het kader van de AVG uit te voeren, ook wanneer hier geen wettelijke of beleidsmatige vereiste toe bestaat. **(6.1g)**

In het implementatieplan **moet** ook het beoogde implementatieteam worden opgenomen. **(6.1h)**
Het is van belang een speciaal aangewezen implementatieteam samen te stellen met een multidisciplinaire achtergrond. Dit team bestaat bij voorkeur uit:

- Minimaal twee eindgebruikers, bij voorkeur zijn eindgebruikers vanaf fase 1 betrokken bij de ontwikkeling van het AIPA;
- Data scientist (of vergelijkbaar, bijv. klinisch fysicus, statisticus, epidemioloog, biomedisch technoloog, technisch geneeskundige);
- IT-specialist (met kennis van het AIPA en software-integratie in bestaande systemen);
- Projectleider met bedrijfskundige achtergrond (of vergelijkbaar, bijv. gezondheidswetenschapper, beleidsmaker, bestuurskundige).

Sterk aanbevolen wordt het implementatie plan op te stellen in samenwerking met patiënten of cliënten. **(6.1i)**

Het implementatieteam **moet** worden ondersteund door de ambitie en bewuste keuze voor AI door betrokken bestuurders van de zorgorganisatie. Zij zorgen voor beleid op het gebied van de volgende zaken: voldoende middelen voor de IT afdeling, protocollen voor

toepassing, rapportage en monitoring, uitvoer van de kosteneffectiviteits- en impactanalyse, eventuele samenwerking met een groter ziekenhuis en/of zorgorganisatie bij gebrek aan kennis in huis, certificering door de fabrikanten kennis van inkopers. **(6.1j)**

De betrokken bestuurders kunnen op al deze facetten bepalen of ze *AI-ready* zijn, bijvoorbeeld door gebruik te maken van beschikbare tools als de *AI-readiness assessment* (https://www.ai-routekaart.nl/#readiness-assessment-intro).

6.2 Monitoring

6.2.1 Verantwoordelijkheden van de fabrikant of ontwikkelende zorgorganisatie

Door de fabrikant **moet** worden gemonitord op technische fouten in het AIPA en de bijbehorende software, op foutief gebruik, op foutieve voorspellingen, op fairness en onverwachte neveneffecten van gewoon gebruik van de software in de dagelijkse praktijk. **(6.2.1a)**

Dit is van belang om veilig en effectief gebruik van het AIPA op de lange termijn te kunnen waarborgen[2-7].

Dit is in het bijzonder van belang voor toepassen van een AIPA, omdat sommige fouten onvoorspelbaar, moeilijk te detecteren of nieuw zijn, of pas optreden op een termijn die niet in de effectbeoordeling onderzocht was. Dergelijke fouten kunnen een grote invloed op de zorg (en maatschappij) hebben als het AIPA op grote schaal gebruikt wordt[8].

Voor de ontwikkelaar en fabrikant van de AIPA zijn er twee sporen om de productveiligheid te borgen: via een *post market surveillance systeem* en via vigilantie (“waakzaamheid”) betreffende incidentrapportages en veiligheidswaarschuwingen (‘Field Safety Notices’). Voor verdere details aangaande deze systemen wordt verwezen naar de MDR en het CMT (MDR Artikel 87, Convenant Medische Technologie[9]). Voor het inrichten van een post-market surveillance (PMS) plan wordt opnieuw verwezen naar de MDR.

Aanvullend op bestaande eisen aan het PMS plan **moet** ten minste aandacht te worden besteed aan:

- Monitoring op en analyse van foute voorspellingen van het AIPA (werkelijke prestaties of uitkomsten tonen een verschil t.o.v. verwachte voorspellende prestaties en uitkomsten o.b.v. ontwikkel en/of validatiestudies[8]). Bijvoorbeeld door monitoring op (afhankelijk van type toepassing):
    - Miscalibratie (bijv. de voorspelde prognose is niet accuraat, de voorspelde kans op sterfte door cardiovasculaire aandoeningen is een over- of onderschatting);

    - Foutpositieve classificatie (bijv. het AIPA geeft aan dat een tumor kwaadaardig is, blijkt goedaardig). Te meten met de positief voorspellende waarde (in %), bij voorkeur per relevante subgroep.;
    - Foutnegatieve classificatie (bijv. het AIPA geeft aan dat een tumor goedaardig is, maar deze blijkt kwaadaardig);
    - De foutenmarge door de tijd en kwaliteit van gebruikte data.
- Monitoring op technische fouten (bijv. de AIPA software geeft geen output voor bepaalde individuen/patiënten, is niet toegankelijk, is niet goed geïntegreerd in bestaande IT-systemen, de responstijd is niet acceptabel). De software moet een mogelijkheid bieden tot geven van feedback door de eindgebruiker op slechte prestaties van de software aan de verantwoordelijke beheerder (bijv. de IT-afdeling) en/of de fabrikant (zie fase 4).
- Monitoring op fairness: om fairness (en bias) te kunnen meten moet voor kwetsbare groepen zoals geïdentificeerd in fase 3 getest worden of er geen ongewenste verschillen voorkomen in de effecten en uitkomsten van een AIPA, ook na introductie van het AIPA in de medische praktijk. Hierbij is het van belang dat essentiële variabelen die socio-economische determinanten van gezondheid vangen voor zover mogelijk verzameld worden.
- Monitoring op de in de risico-inventarisatie (zie sectie 5.1.2) en analyse van onzekerheden (zie sectie 5.3.1) onderkende risico's;
- Monitoring op *deployment bias, zie aanbeveling 6.2.1c*. **(6.2.1b)**

Voor het vaststellen van *deployment bias* wordt **aanbevolen**:

- Automatisch in de software te documenteren of de gebruiker het advies van het algoritme volgt of niet (en waarom niet). Indien niet mogelijk, onderbouw waarom;
- Regelmatig een evaluatie uit te voeren of het AIPA wordt gebruikt voor de doelgroep waarop ook getraind en getest is;
- Regelmatig een evaluatie uit te voeren van de wijze van gebruik door eindgebruikers om te onderzoeken of het AIPA wordt gebruikt in overeenstemming met het beoogd doelgebruik. **(6.2.1c)**

6.2.2 Verantwoordelijkheden van de zorgorganisatie

Als een AIPA wordt gebruikt binnen een zorgorganisatie, heeft de zorgorganisatie de verantwoordelijkheid om juiste werking en gebruik van het AIPA blijvend te monitoren.

Er **moet** een lokaal monitoringsplan worden opgesteld door de zorgorganisatie waar de software die het AIPA bevat wordt geïmplementeerd. **(6.2.2a)**

In het monitoringplan **moeten** minimaal de volgende elementen beschreven staan:

- Monitoring of het doel en gewenste effect bereikt worden.
- Monitoring op en analyse van foute voorspellingen van het AIPA (werkelijke prestaties of uitkomsten tonen een verschil t.o.v. verwachte voorspellende prestaties en uitkomsten o.b.v. ontwikkel en/of validatiestudies[8]), bijvoorbeeld op miscalibratie, foutpositieve en foutnegatieve resultaten (zie monitoring op en analyse van voorspellingen van het AIPA in eis 6.2.1c voor een toelichting). Dit kan in samenwerking met de fabrikant worden ingericht;
- Monitoring op technische fouten (bijv. de AIPA software geeft geen output voor bepaalde individuen/patiënten, is niet toegankelijk, is niet goed geïntegreerd in bestaande IT-systemen, of de responstijd is niet acceptabel);
- Monitoring op onverwachte effecten voor de zorgverlener, patiënt, organisatie, de zorgprocessen en/of de maatschappij. Deze effecten moeten worden gemeld door zorgverlener, zorgorganisatie en/of de fabrikant;
- Monitoring op foutief gebruik van het AIPA:
    - Automation (confirmation) bias: de voorspellingen van een AIPA wegen te zwaar mee in de beslissingen van een zorgverlener. Monitor daarnaast op *'deskilling'*: het verleren van medische handelingen door automatisering; en
    - *Deployment bias*: oneigenlijk gebruik (bijv. toepassing bij patiënten waarvoor de AIPA software niet ontwikkeld is, niet naleven van model aanbevelingen) of een foute interpretatie van de uitkomst. Zie aanbeveling 6.2.2d.
- Een analyse van welke medische relevantie iedere verwachte fout heeft en hoe waarschijnlijk het is dat deze fout op kan treden, op basis van de prospectieve risico inventarisatie (PRI, zie eis 6.1d). Monitoring op de in de PRI onderkende risico's;
- Onderbouwing van welke gegevens (data) worden verzameld ten behoeve van monitoring, en hoe dit in overeenstemming is gebracht met eindgebruikers; en
- De frequentie van monitoren is en waarom hiervoor is gekozen.
- Expliciete benoeming van een meldplicht van de eindgebruiker aan de fabrikant en vice versa bij een onverwachte uitkomst waarvan de oorzaak niet te herleiden is (vergelijk met een *serious adverse event* (SAE) bij een experimentele behandeling) **(6.2.2b)**

Aanbevolen wordt om als onderdeel van het monitoringplan individuele ervaringen van stakeholders (bijv. van de patiënt en zorgverlener) waar mogelijk op te vragen. **(6.2.2c)**

Voor het vaststellen van lokale deployment bias wordt **sterk aanbevolen**:

- De aansluiting van de software bij zorgprocessen blijvend of periodiek te monitoren;
- Te registreren hoe vaak de gebruiker het AIPA gebruikt, en of er sprake is van een zekere leercurve;
- Te monitoren op de doelgroep: bewaak dat het AIPA voor de voorbereiding van medisch handelen wordt gebruikt voor de doelgroep waarop ook getraind en getest is;
- Regelmatig een controle uit te voeren van de wijze van gebruik door eindgebruikers om correct gebruik en zinvolle interpretatie van resultaten te garanderen;
- Data-invoer en -opslag te controleren op mogelijke meetfouten, ook na implementatie en ingebruikname algoritme. **(6.2.2d)**

6.3 Educatie

6.3.1 Eindgebruiker

De eindgebruiker (bijv. een patiënt of de zorgverlener) **moet** toegang hebben tot informatie over de onderwerpen beschreven in box 6.1, te leveren door de ontwikkelaar of fabrikant. **(6.3.1a)**

Als de eindgebruiker een zorgverlener is **moet** de zorgverlener toegang hebben tot onderwijs over de onderwerpen beschreven in box 6.2. **(6.3.1b)**

Dit heeft als doel om de zorgverlener te sterken in de mogelijkheid om het AIPA bewust bekwaam te gebruiken in de medische praktijk.

Sterk aanbevolen wordt om dit onderwijs met enige regelmaat te herhalen, afhankelijk van de toepassing en de medische context. **(6.3.1c)**

Dit kan worden ondersteund door de eindgebruiker van informatie te voorzien (door de fabrikant of ontwikkelaar) over het volgende: de resultaten van de validatie en effectbeoordeling zoals uitgevoerd in fasen 3 en 5, de populatie waarop getraind en gevalideerd is (fase 2 en 3), de prestaties van het AIPA (fase 3), de verwachte foutmarge, statistische kennis, en uitlegbaarheid van het algoritme.

Box 6.1: Educatie van de eindgebruiker

De eindgebruiker moet toegang hebben tot onderwijs op de volgende punten:

- Het bedoelde gebruik van het specifieke AIPA, beperkingen aan dit (bedoelde) gebruik en de bijbehorende gebruikershandleiding zoals verplicht is gesteld door de MDR.
- Interpretatie van de uitkomsten van het AIPA.
- De mogelijke fouten in gebruik die kunnen worden gemaakt zoals beschreven in sectie 6.1. Speciale aandacht voor transporteerbaarheid en generaliseerbaarheid van het AIPA naar de lokale medische omgeving. Zorg dat het AIPA eigen grenzen kan aangeven door te signaleren wanneer een datapunt te erg afwijkt van train populatie, in bepaalde gevallen geen voorspelling te geven, of betrouwbaarheidsintervallen te geven (zie ook fase 4), train eindgebruikers op interpretatie hiervan.
- De mogelijke winst die is te behalen door het toepassen van het AIPA.
- Instructie over (definities en kwaliteit van) data-invoer die wordt verwacht van de eindgebruiker.

6.3.2 Zorgorganisatie

De zorgorganisatie **moet** toegang hebben tot informatie en/of onderwijs op de onderwerpen beschreven in box 6.2, te leveren door de fabrikant of ontwikkelaar. **(6.3.2a)**

Dit heeft als doel om de zorgorganisatie *bewust* te maken van de verantwoordelijkheden die komen bij de inkoop of ontwikkeling in eigen huis van een AIPA die zal worden geïmplementeerd en gebruikt in de medische praktijk. De zorgorganisatie draagt hierbij de verantwoordelijkheid en moet zich informeren; het is een randvoorwaarde om de benodigde kennis in huis te hebben en/of halen betreffende implementatie van AIPA's.

Eindgebruikers (bijv. zorgverleners) **moeten** de ruimte en mogelijkheid krijgen voor onderwijs betreffende het AIPA. **(6.3.2b)**

Box 6.2: Educatie van de zorgorganisatie

- Wetgeving van toepassing op AIPA (o.a. MDR, zie ook *fase 4*).
- Benodigd technisch beheer. Gedacht kan worden aan (niet-uitputtende lijst): beheer van de cloud, ondersteuning van IT infrastructuur, opslag en service grote datastromen, communicatie met beheerder elektronisch patiënten dossier.
- *D*e resultaten van uitgevoerde evaluaties en beoordelingen als onderdeel van fase 3 en 5Het effect van AI op de zorginstelling: er moet aandacht zijn voor de verandering van werkprocessen en autonomie van de zorgverleners.
- Kosten/baten: ontwikkeling van een business case, waarbij kosten niet alleen in aanschaf, maar ook in de benodigde werkzaamheden op het gebied van educatie, implementatie en beheer liggen.
- Ethische overwegingen (zie ook bias en fairness, fase 3)

6.4 Rechten en plichten

De betrokken stakeholders bij de invoer en het gebruik van AIPA software in de medische praktijk hebben elk *rechten* en *plichten* die volgen uit deze leidraad om zorg te dragen voor een verantwoordelijke introductie, dat wil zeggen lettend op de veiligheid en toegevoegde waarde van AIPA software in de praktijk. Er is onderscheid gemaakt tussen vier stakeholders: de zorgverlener, de zorgorganisatie, de patiënt en de fabrikant.

Het gebruik van de lijst rechten en plichten van zorgverleners, zorgorganisatie, patiënt/burger en fabrikant om te controleren en vast te stellen dat alle rechten en plichten effectief uitgeoefend kunnen worden, wordt **sterk aanbevolen**. **(6.4a)**

6.4.1 Zorgverlener

Rechten:

- Ondersteund te worden in kennis over specifieke AIPA door zorgorganisatie en fabrikant, begrijpelijke eindgebruiker informatie en training op gebruik en monitoring;
- Transparante communicatie (door fabrikant van de AIPA software en/of derden die studies hebben uitgevoerd ter validatie) over eerdere studies;
- Terugkoppeling op gemelde incidenten.

Plichten:

- Bewust bekwaam zijn in het gebruik van het AIPA;
- Het bedoelde gebruik naleven in het belang van de patiënt en fabrikant;
- Implementatie volgens het implementatieplan van de zorgorganisatie;

- Transparantie over het gebruik van AI in prognose of diagnose (aan zorgorganisatie en patiënt, bijvoorbeeld in het dossier);
- Terug melden incidenten in het kwaliteitsmanagementsysteem;
- Transparantie over het gebruik van patiëntdata ter verbetering van AI aan patiënt, en waar nodig het verkrijgen van geïnformeerde toestemming.

6.4.2 Zorgorganisatie:

Rechten, in het geval dat een AIPA van een fabrikant wordt toegepast:

- Ondersteund te worden door de fabrikant in kennis over het specifieke AIPA;
- Transparante communicatie (door de fabrikant) over eerdere studies;
- Terugkoppeling op gemelde incidenten;
- Kwaliteitsmanagementsysteem dat meldingen faciliteert (vanuit de fabrikant).

Plichten, zowel bij toepassing van AIPA van fabrikant and eigen ontwikkeling:

- Ondersteuning verlenen aan medewerkers die werken met het AIPA;
- Borgen bekwaamheid van (AI) betrokken medewerkers;
- Het bedoelde gebruik naleven in belang van de patiënt en fabrikant;
- In het algemeen vermelden dat AI gebruikt wordt;
- Borgen dat incidentrapportages en veiligheidswaarschuwingen (‘Field Safety Notices’) te allen tijde kunnen worden ontvangen en verwerkt
- Terug melden incidenten (aan fabrikant, zorgorganisatie, patiënt en in dossier);
- Zorgen dat er een implementatieplan is dat wordt uitgevoerd in lijn met de leidraad (zie sectie 6.1);
- Transparantie over het gebruik van patiëntgegevens ter verbetering van het AIPA;
- Werken met een kwaliteitsmanagementsysteem.

6.4.3 Patiënt, cliënt of burger

Rechten

- Het bedoelde gebruik van het AIPA wordt nageleefd door de zorgverlener en zorgorganisatie;
- Transparantie over het gebruik van AI in de zorgorganisatie;
- Transparantie over het gebruik van patiëntdata ter verbetering van AI;
- Ondersteuning fabrikant bij rechtstreekse relatie in zelfzorgtoepassingen (bijv. *e-health* toepassingen);
- Terugkoppeling door fabrikant op individuele door patiënt zelf gemelde incidenten;
- Terugkoppeling van bugs en kwaliteitsbewaking;

- Het recht om op ieder gewenst moment te stoppen met het delen van data. Het recht om vergeten te worden (als bedoeld in AVG artikel 17);
- *Mogelijk: t*ransparantie over informatie over het gebruikte AIPA (bijv. resultaten van eerdere studies).

Plichten

- Indien akkoord gebruik, correct invullen/aanleveren van data, zolang er gebruik wordt gemaakt van de AIPA software;
- Terug melden van incidenten (aan zorgverlener, zorgorganisatie of direct aan de fabrikant);
- Bedoeld gebruik naleven (bij rechtstreekse relatie, bijv. correcte data-invoer en datameting op aangegeven momenten).

6.4.4 Fabrikant of ontwikkelende zorgorganisatie

Rechten

- Verkrijgen informatie over meldingen van de zorgorganisatie, zorgverlener of patiënt.

Plichten

- Ondersteuning aan zorgverlener en patiënt voor onderhoud en kennis van juist gebruik van het AIPA (zoals verplicht door de MDR);
- Transparante communicatie (aan zorgorganisatie/eindgebruiker) over eerder uitgevoerde studies;
- Aanbod van informatie over monitoring en kwaliteitsmanagementsysteem;

Behandeling en terugkoppeling aan zorgverlener of patiënt op door hen gemelde incidenten (waarbij ook verplichtingen omtrent vigilantie als besproken in 5.3.2).

6.5 Referenties

1. VMS. VMS Handleiding Prospectieve Risico Inventarisatie.

2. Buruk B, Ekmekci PE, Arda B. A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Med Health Care Philos* 2020;23(3):387-99. doi: 10.1007/s11019-020-09948-1 [published Online First: 2020/04/03]

3. Floridi L, Cowls J, King TC, et al. How to Design AI for Social Good: Seven Essential Factors. *Sci Eng Ethics* 2020;26(3):1771-96. doi: 10.1007/s11948-020-00213-5 [published Online First: 2020/04/05]

4. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17(1):195. doi: 10.1186/s12916-019-1426-2 [published Online First: 2019/10/31]

5. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019;25(9):1337-40. doi: 10.1038/s41591-019-0548-6 [published Online First: 2019/08/21]

6. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 2020;368:l6927. doi: 10.1136/bmj.l6927 [published Online First: 2020/03/22]

7. Magrabi F, Ammenwerth E, McNair JB, et al. Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications. *Yearb Med Inform* 2019;28(1):128-34. doi: 10.1055/s-0039-1677903 [published Online First: 2019/04/26]

8. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020;26(9):1364-74. doi: 10.1038/s41591-020-1034-x [published Online First: 2020/09/11]

9. (NVZ) NVvZ, (NFU) NFvUMC. Convenant Veilige Toepassing van Medische Technologie in de medisch specialistische zorg, 2e druk, 2016.

10. Zorginstituut Nederland, Federatie Medische Specialisten (voorheen OMS). Leidraad Nieuwe Interventies in de Klinische Praktijk. https://www.demedischspecialist.nl/sites/default/files/Leidraad%20Nieuwe%20interventies%20in%20de%20klinische%20praktijk%20def.pdf. Published Oktober 2014. Accessed December 8, 2021

Toekomstperspectieven

Auteurs

Auteurs: Ilse Kant, Maarten van Smeden, namens de werkgroepen medische AI

De werkgroepen ‘veldnorm medische AI’ hadden als doel om gemeenschappelijk opgestelde, publiek inzichtelijke criteria op te stellen om medische AIPA software te kunnen evalueren en toetsen, in opdracht van het *Ministerie van Volksgezondheid, Welzijn en Sport*. Gezien de snelle ontwikkelingen in het veld moet deze leidraad gezien worden als een dynamische norm. Zo zijn enkele generieke onderwerpen en toekomstige ontwikkelingen reeds door de werkgroepleden gesignaleerd en wordt het door hen van belang geacht dat hier in het bijzonder aandacht aan moet worden besteed bij toekomstige doorontwikkelingen.

Dynamisch updaten van AIPA’s

AI technologie maakt een snelle ontwikkeling door die de huidige normen en regelgeving op enkele punten voorbij dreigt te streven. De werkgroepleden signaleren als probleem dat het naleven van de MDR in zijn huidige vorm óf zeer onzeker óf zeer bewerkelijk is voor een AIPA die zeer frequent hertraind wordt met nieuwe data om het AIPA continue tijdens gebruik te verbeteren en up-to-date te houden. Daarmee belemmert onzekerheid bij ontwikkelaars over de toepassing van de MDR in zijn huidige vorm vele mogelijke (toekomstige) AI toepassingen. Om geen valse verwachtingen over naleving te scheppen is dit onderwerp buiten de huidige versie van de leidraad gehouden. Het wordt zeer wenselijk geacht dit onderwerp, en specifiek de evaluatie van dit soort updates, in toekomstige versies van de norm op te nemen.

Kosteneffectiviteitsevaluaties

In de huidige versie van de norm wordt met markttoelating en de route tot opname van het AIPA in zorgpakketten en vergoeding vanuit zorgverzekeringen niet expliciet rekening gehouden. De uitkomsten van impactanalyses d.m.v. modelmatige HTA analyse van het AIPA (zie fase 5) zullen in toenemende mate een rol gaan spelen in de markttoelating en de mogelijkheden tot het opnemen van AI in zorgpakketten van verzekeraars. Dit aspect zal daarom in de toekomst moeten worden opgenomen in de leidraad.

Data delen

De werkgroepleden onderstrepen het belang van (inter)nationale samenwerking op het gebied van data delen en AIPA’s om toekomstige ontwikkelingen in de gezondheidszorg aan te moedigen en ondersteunen. Voorbeelden van initiatieven die hier al aan werken zijn de stichting NICE[1] en de stichting NEED[2] databases voor respectievelijk intensive care data en spoedeisende hulp data, Health-RI en de Personal Health Train[3]. De Nederlandse AI coalitie werkgroep *data delen* heeft een leidraad gepubliceerd voor het omgaan met data in samenwerkingsverbanden[4]. Een voorbeeld van een nieuwe ontwikkeling die kan worden toegepast in dit soort samenwerkingsverbanden is *federated learning*. Federated learning maakt grootschalige samenwerking van meerdere instituten mogelijk zonder het gebruik van

een centrale database en is daarmee dus zeer privacy-vriendelijk. Momenteel is dit onderwerp niet opgenomen in de leidraad. Het wordt wenselijk geacht dit in toekomstige versies van de norm op te nemen.

Vroegtijdig multidisciplinair werken

De werkgroepleden signaleren diverse struikelblokken die de toepassing van een AIPA in de medische praktijk in de weg kunnen staan. In het medisch-specialistische veld, met name in de radiologie, neemt het aantal toepassingen snel toe. Maar nog steeds stranden veel projecten in eerdere fasen (fase 2 of 3 van de leidraad). Door de werkgroepleden wordt het vroegtijdig betrekken van meerdere eindgebruikers van het AIPA wordt regelmatig als oplossing aangedragen. Andere struikelblokken die zijn gesignaleerd door de werkgroepleden gaan over de transporteerbaarheid en generaliseerbaarheid van voorspellende modellen naar een andere medische context (bijv. een ander ziekenhuis, land of zelfs medisch werkproces). Hierover is nog veel onbekend. Meer onderzoek is nodig om (vooraf) tot betere inschattingen te kunnen komen of, en wanneer, een AIPA in een andere context opnieuw gevalideerd en/of getraind moet worden alvorens te kunnen worden toegepast. Daarom zijn hier geen specifieke aanbevelingen over opgenomen in de huidige versie van de leidraad. Daarnaast is het wenselijk specifieke aanbevelingen op te nemen over de omgang met een verruiming van het beoogd doelgebruik van een bestaande AIPA.

Monitoren in praktijk

In fase 6.2 van de veldnorm wordt monitoring na invoer van AIPA software in de *real world* (dat wil zeggen: buiten studieverband) medische praktijk beschreven. De werkgroepleden signaleren als probleem dat na implementatie van AIPA software in de praktijk, waar veelal een behandelkeuze is gekoppeld aan de uitkomst van het AIPA (bijv. bij beslisondersteuning), het monitoren op *miscalibratie* op basis van data uit de praktijk problematisch is. Over dit vraagstuk is nog veel onbekend en meer onderzoek is gewenst. Het wordt wenselijk geacht om hier in toekomstige versies van de veldnorm meer duidelijkheid over te scheppen.

Educatie

De snelle opkomst van digitale gezondheidszorg kan mogelijk zorgen voor een disruptieve verandering in de medische zorgprocessen. De werkgroepleden signaleren echter een algemeen gebrek aan basiskennis over AI bij zorgorganisaties en op de werkvloer, dat wil zeggen bij artsen, verpleegkundigen en andere betrokkenen die in toenemende mate met AI te maken zullen krijgen in hun dagelijkse zorgwerkzaamheden. In voorkomende gevallen is het zelfs de patiënt die over deze basiskennis zou moeten beschikken. Om AIPA technologie effectief en verantwoordelijk te kunnen introduceren in de dagelijkse medische praktijk zal

deze kennisachterstand moeten worden ingehaald, ten eerste in het onderwijs van zorgprofessionals die momenteel werkzaam zijn, en ten tweede in het onderwijs van nieuwe zorgprofessionals. Daarnaast zullen ontwikkelaars en fabrikanten een basiskennis moeten opbouwen op het gebied van het veilig en effectief introduceren van een AIPA, waarbij deze leidraad richtinggevend kan zijn. Tijdens doorontwikkeling van deze norm moet er aandacht blijven voor (toekomstige) beheersbaarheid en uitvoerbaarheid van de norm voor ontwikkelaars en fabrikanten.

Referenties

1. Stichting NICE [Available from: https://www.stichting-nice.nl/dd/#start.
2. Stichting NEED [Available from: https://www.stichting-need.nl/.
3. Health-RI [Available from: https://www.health-ri.nl.
4. Klauw Kvd, Bastiaansen H, Ette Fv. Verantwoord datadelen voor AI: Nederlandse AI coalitie, 2020.

Colofon

Deze Leidraad is ontwikkeld door experts uit de breedte van het zorgveld, ondersteund door een actieteam binnen het programma waardevolle AI voor gezondheid. Deze leidraad is een eerste versie en is een uitdrukking van wat er in het werkveld als goed professioneel handelen wordt beschouwd bij het ontwikkelen, toetsen en toepassen van AI in de zorg. De ambitie is dat deze leidraad als breed gedragen veldnorm wordt geaccepteerd.

Voor inhoudelijke vragen of opmerkingen kunt u terecht bij de auteurs:

dr. Maarten van Smeden - M.vanSmeden@umcutrecht.nl
dr. Ilse Kant – I.M.J.Kant@lumc.nl

Waardevolle AI