Hjerterom for etisk AI

Sluttrapport fra sandkasseprosjektet til Ahus (EKG AI)

Tema: Algoritmeskjevhet og rettferdige algoritmer

Februar 2023

Datatilsynet

Innhold

Hva er sandkassa?

I sandkassa utforsker deltakere sammen med Datatilsynet personvernrelaterte spørsmål, for å bidra til at tjenesten eller produktet deres etterlever regelverket og ivaretar personvernet på en god måte.

Datatilsynet tilbyr veiledning i dialog med deltakerne, og konklusjonene fra prosjektene er ikke vedtak eller forhåndsgodkjenning. Deltakerne står fritt i valget om å følge rådene de får.

Sandkassa er en verdifull metode for å utforske problemstillinger der jussen har få praktiske eksempler å vise til, og vi håper konklusjoner og vurderinger i rapporten kan være til hjelp for andre med liknende problemstillinger.

Februar 2023

Denne pdf-en tilsvarer den første versjonen av rapporten, slik den ble publisert på Datatilsynets sider februar 2023. Teknologien og jussen er stadig i utvikling, så det kan være behov for å justere eller presisere rapportene med tiden. Dersom denne pdf-en skiller seg fra det som står på Datatilsynets nettsider, kan du ta utgangspunkt i at det er nettsidens tekst som er gjeldende råd.

Sammendrag

Målet med dette sandkasseprosjektet har vært å utforske begrepene «rettferdighet» og «algoritmeskjevhet» i et konkret helseprosjekt, EKG AI. Ahus er i gang med å utvikle en algoritme for å predikere sannsynligheten for hjertesvikt hos pasienter. På sikt skal den brukes som et beslutningsstøtteverktøy[1] for helsepersonell for en forbedret og mer effektiv behandling og oppfølging av pasienter. Vi har i sandkasseprosjektet diskutert mulighetene for at det eksisterer skjevheter i EKG AI, og konkretisert forslag til tiltak for å forhindre diskriminering.

Oppsummering av resultatene:

- **Hva er rettferdighet?** Begrepet «rettferdighet» har ingen legaldefinisjon i personvernforordningen, men er ifølge artikkel 5 et sentralt personvernprinsipp. Rettferdighetsprinsippet står sentralt i andre lovverk, og vi har derfor sett til likestillings- og diskrimineringsloven (ldl.) for å klargjøre hva som ligger i prinsippet. I dette prosjektet har vi vurdert EKG AIs grad av rettferdighet ut i fra krav til ikke-diskriminering og åpenhet, den registrertes forventninger, og etiske betraktninger om hva samfunnet anser som rettferdig.

- **Hvordan avdekke algoritmeskjevhet?** For å sikre at algoritmen er rettferdig må det undersøkes om EKG AI-algoritmen gir mindre treffsikre prediksjoner for noen pasientgrupper. I dette prosjektet valgte vi å se nærmere på diskrimineringsgrunnlagene «kjønn» og «etnisitet». Når man skal kontrollere algoritmen for diskriminering, vil det som regel være behov for behandling av nye personopplysninger, herunder særlige kategorier av personopplysninger. I den forbindelse må man vurdere kravene til behandlingens lovlighet og dataminimeringsprinsippets krav til en forholdsmessig og nødvendig behandling av personopplysninger.

- **Hvilke tiltak kan redusere algoritmeskjevhet?** Arbeidet i sandkasseprosjektet har synliggjort en potensiell risiko for at EKG AI-algoritmen kan diskriminere enkelte pasientgrupper. Skjevheter kan reduseres gjennom tekniske eller organisatoriske tiltak. Aktuelle tiltak for EKG AI er å sikre et representativt datagrunnlag og sørge for god informasjon til, og opplæring av, helsepersonell slik at prediksjonene brukes riktig i praksis. I tillegg, vil Ahus etablere en mekanisme som skal overvåke treffsikkerheten til algoritmen og som sørger for at algoritmen etterlæres ved behov.

Veien videre

Ahus ønsker å prøve ut algoritmen i klinisk drift i starten av 2024. Klinisk beslutningsstøtteverktøy basert på kunstig intelligens (KI) anses som medisinsk-teknisk utstyr og må CE-merkes av Statens legemiddelverk for å kunne tas i bruk i klinisk virksomhet.

Arbeidet i sandkasseprosjektet har synliggjort en potensiell risiko for at EKG AI kan diskriminere enkelte pasientgrupper. Ahus vil se på muligheten for å gjennomføre en klinisk studie for å undersøke om algoritmen har dårligere treffsikkerhet og prediksjoner for pasienter med ulike etniske bakgrunner (i denne rapporten brukt om genetiske opphav). Resultatene fra studien vil vise om det må iverksettes korrigerende tiltak i etterlæringsfasen til algoritmen.

I løpet av prosjektperioden har vi sett at det ikke finnes en felles og omforent metode for å avdekke algoritmeskjevhet. Dersom vi hadde hatt mer tid i prosjektet ville vi ha utviklet en egen metode, basert på erfaringer fra prosjektperioden. I tillegg ville det ha være interessant å gå enda dypere ned i de etiske kravene knyttet til bruk av kunstig intelligens i helsesektoren.

[1] Beslutningsstøtteverktøy = i forarbeidene til helsepersonelloven står det at begrepet skal forstås vidt, og det omfatter alle typer kunnskapsbaserte hjelpemidler og støttesystemer som kan gi råd og støtte og veilede helsepersonell ved ytelse av helsehjelp.

Om prosjektet EKG AI

Akershus universitetssykehus (Ahus) er et lokal- og områdesykehus med 12 000 ansatte. Ahus har ansvar for omtrent 594 000 innbyggere i Follo, Romerike, Kongsvinger-regionen samt de nordligste bydelene i Oslo. Ahus er Norges største akuttsykehus med et pasientilbud innenfor somatikk, psykisk helsevern og rusbehandling.

Ahus har utviklet et beslutningsstøtteverktøy basert på kunstig intelligens, EKG AI, som kan predikere hjertesvikt hos pasienter. Beslutningsstøtteverktøyet utvikles ved å koble EKG-data med spesifikke diagnoser, såkalt veiledet læring[2]. Etter trening, test og validering resulterer dette i en algoritme som skal predikere sannsynligheten for hjertesvikt. Det er ikke tidligere utviklet lignende verktøy som er tatt i bruk i klinisk drift. Med det store pasientgrunnlaget til Ahus, har prosjektet gode forutsetninger for å utvikle et verktøy som har høy treffsikkerhet og som også kan tas i bruk ved andre helseforetak i Norge.

Med EKG AI ønsker Ahus å:

- Øke effektiviteten på diagnostisering og behandling av hjertesvikt.
- Forbedre diagnostisering av hjertesvikt, og kunne fastslå dette på et tidligere stadium enn tidligere.
- Mindre liggetid, raskere behandling og redusere dødelighet.
- Kommersialisere algoritmen til andre helseaktører.

Datakilder og dataflyt

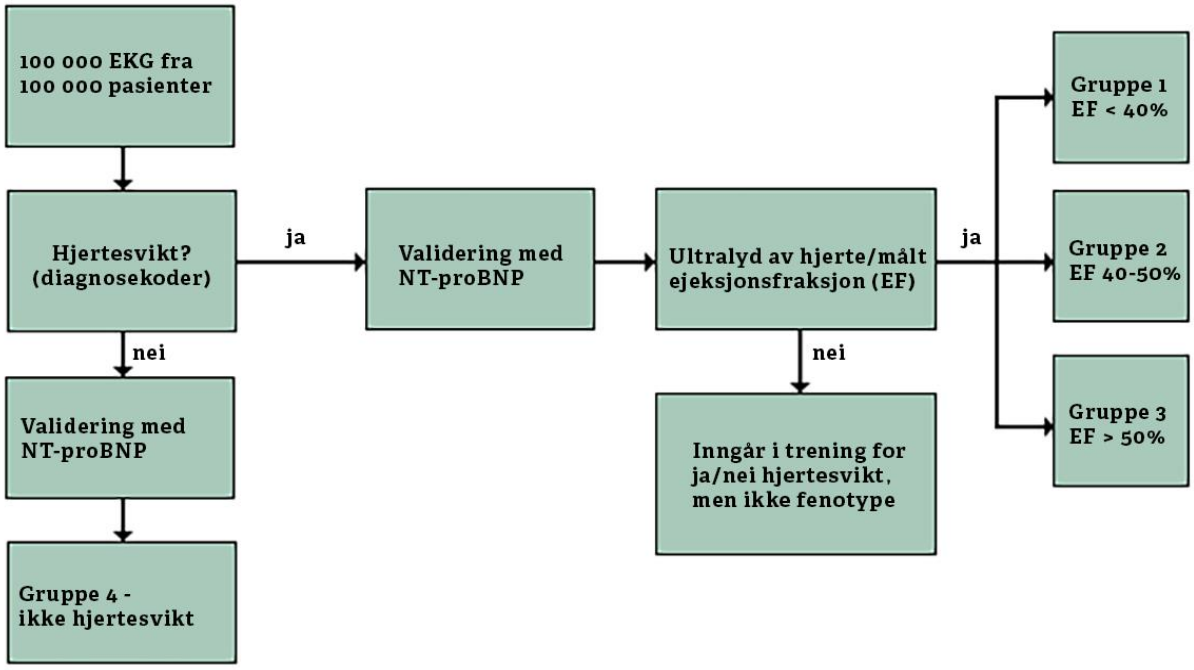
Beslutningsstøtteverktøyet skal utvikles på Google Cloud platform med autoML-verktøyet Vertex AI.

Det anvendes omtrent 100 000 EKG-undersøkelser fra pasienter som har vært innlagt på Ahus de siste årene. Det finnes tre ulike fenotyper for hjertesvikt som krever ulik behandling. EKG-undersøkelsene grupperes basert på dette:

- Gruppe 1: hjertesvikt med redusert pumpefunksjon
- Gruppe 2: hjertesvikt med middels redusert pumpefunksjon
- Gruppe 3: hjertesvikt med bevart pumpefunksjon
- Gruppe 4: ikke hjertesvikt

Inndelingen gjøres ut fra diagnosekoder, blodprøven NT-proBNP og hjerteultralyd-målingen ejeksjonsfraksjon (EF). Se figur på neste side.

[2] Veiledet læring betyr at det benyttes kategoriserte data. Veiledningen skjer da i form av de merkelappene som følger med dataene. Hentet fra Kunstig intelligens og personvern, Datatilsynet, 2017, s. 7



Dataene hentes fra to ulike kilder: EKG-arkivet ComPACS (EKG-målinger og hjertefunksjon (EF)) og journalsystemet DIPS (diagnosekoder og blodprøven NT-proBNP). Etter inndelingen i de fire gruppene pseudonymisere alle EKG-undersøkelsene. Etter dette overføres de til Google Cloud platform for trening, testing og validering med Vertex AI.

> **Pseudonymisering**
>
> Å avidentifisere personopplysninger slik at de ikke kan knyttes til en bestemt person uten bruk av tilleggsopplysninger (for eksempel en koblingsnøkkel) som lagres adskilt og tilstrekkelig sikkert. Pseudonymiserte personopplysninger er ikke anonyme.

Mål for sandkasseprosjektet

Ahus hadde allerede startet utvikling av EKG AI-algoritmen da de ble plukket ut som deltaker i Datatilsynets regulatoriske sandkasse, våren 2022. Ahus ønsket å diskutere algoritmeskjevhet og hvordan sikre at EKG AI-algoritmen gir rettferdige prediksjoner.

Det finnes begrenset rettspraksis om kravet til rettferdige algoritmer, og personvernforordningen gir ikke klare svar på hvordan prinsippet skal tolkes i praksis. Målet med dette sandkasseprosjektet har derfor vært å utforske hva som ligger i rettferdighetsprinsippet i personvernforordningen artikkel 5 og hvordan det skal forstås i møte med et konkret KI-prosjekt.

Sandkasseprosjektet har som mål å undersøke om det eksisterer skjevheter i EKG AI, og komme med forslag til konkrete tiltak for å redusere en eventuell diskriminering. Tiltakene har til hensikt å utvikle algoritmer som fremmer likebehandling og forhindrer diskriminering. I den forbindelse har Likestillings- og diskrimineringsombudet (LDO) bidratt inn i prosjektet med spisskompetanse på diskrimineringsregelverket.

Hensikten med sluttrapporten er at diskusjonene og resultatene har overføringsverdi til andre helseprosjekter som benytter seg av kunstig intelligens.

Sandkasseprosjektets mål:

1. **Hva er rettferdighet og algoritmeskjevhet?** Få bedre forståelse for begrepene «rettferdighet», «algoritmeskjevhet» og «diskriminerende algoritmer» samt redegjøre for hvilke regelverk som er relevant i disse tilfellene.
2. **Hvordan avdekke algoritmeskjevhet?** Undersøke hvorvidt, og i hvilken grad, algoritmeskjevhet eksisterer, og eventuelt kan oppstå, i EKG AI-algoritmen til Ahus.
3. **Hvilke tiltak kan redusere algoritmeskjevhet?** Komme med forslag til tekniske og organisatoriske tiltak for å redusere og korrigere eventuelle skjevheter i algoritmen.

Av kapasitetshensyn har sandkasseprosjektet ikke vurdert krav til rettslig grunnlag for behandling av personopplysninger i dette prosjektet. Det kan kort nevnes at bruk av pasientopplysninger i utviklingen av beslutningstøtteverktøyet ble godkjent av Helsedirektoratet i januar 2022 i henhold til helsepersonelloven § 29. Et slikt vedtak gir en dispensasjon fra taushetsplikten og et supplerende rettslig grunnlag for behandling av personopplysninger etter personvernforordningen. For nærmere beskrivelse av rettslig grunnlag for behandling av helseopplysninger i forbindelse med utvikling av algoritmer i helsesektoren, se sluttrapport i sandkasseprosjektet Helse Bergen, publisert høsten 2022[3].

Etter personvernforordningen artikkel 35 stilles det krav til vurdering av personvernkonsekvenser (DPIA) dersom behandlingen av personopplysninger medfører «høy risiko» for fysiske personers rettigheter og friheter.[4] Før Ahus startet utvikling av algoritmen, EKG AI, utformet de en DPIA, men dette har ikke vært et fokusområde for sandkasseprosjektet. Elementer fra rapporten, særlig metode for å avdekke algoritmeskjevhet samt tiltak for å redusere risiko for algoritmeskjevhet, er imidlertid relevante å ha med i en DPIA.

---

[3] https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/ferdige-prosjekter-og-rapporter/helse-bergen-sluttrapport-kunstig-intelligens-i-oppfolging-av-sarbare-pasienter/

[4] Se mer om kravene til DPIA på Datatilsynets nettsider: https://www.datatilsynet.no/rettigheter-og-plikter/virksomhetenes-plikter/vurdere-personvernkonsekvenser/vurdering-av-personvernkonsekvenser/

Rettferdighet og algoritmeskjevhet

I dagligtalen brukes ordet rettferdighet om hvorvidt det finnes en lik og rettferdig fordeling av byrder og goder i samfunnet. Begrepet «rettferdighet» har ingen legaldefinisjon i personvernforordningen, men fremkommer av artikkel 5 nr. 1 bokstav a som et sentralt personvernprinsipp. Der står det at «personopplysninger skal behandles på en lovlig, rettferdig og åpen måte med hensyn til den registrerte».

Innholdet i rettferdighetsprinsippet er dynamisk, som betyr at det endrer seg over tid i takt med samfunnsoppfatningen. Det Europeiske personvernrådet (EDPB) presiserer i sin veileder om innebygd personvern[5] at rettferdighetsprinsippet omfatter ikke-diskriminering, den registrertes forventninger, behandlingens bredere etiske problemstillinger og respekt for den registrertes rettigheter og friheter. Rettferdighetsprinsippet har, med andre ord, et vidt anvendelsesområde.

En tilsvarende beskrivelse finnes i fortalen[6] til personvernforordningen, som uttrykker at prinsippet innebærer at all behandling av personopplysninger skal gjøres med respekt for den registrertes rettigheter og innenfor den registrertes rimelige forventninger om hva opplysningene skal brukes til. Åpenhet og transparens i behandlingen av personopplysninger henger altså nøye sammen med kravet til rettferdighet. Tilstrekkelig informasjon er avgjørende for at behandlingen skal være forutsigbar for den registrerte og for å kunne ivareta sin rett til rettferdig behandling av sine personopplysninger. Denne sluttrapporten vil ikke gå nærmere inn på problemstillinger knyttet til kravet til informasjon, men henviser til sluttrapporten til sandkasseprosjektet til Helse Bergen[7] som har redegjort for dette tema.

Fortalen[8] til personvernforordningen uttrykker videre at den behandlingsansvarlige skal forhindre forskjellsbehandling av den enkelte på grunnlag av rasemessig eller etnisk opprinnelse, politisk oppfatning, religion eller filosofisk overbevisning, fagforeningsmedlemskap, genetisk status, helsetilstand eller seksuell orientering. Fortalen er ikke juridisk bindende, men kan benyttes som veiledende i tolkning av rettsreglene i personvernforordningen. I Ahus prosjektet har vi blant annet stilt spørsmål ved hvorvidt EKG AI-algoritmen vil gi alle pasienter lik tilgang, og like god helsehjelp, uavhengig av om pasienten er mann eller kvinne, eller har en annen etnisk bakgrunn enn den etniske majoriteten av pasientene.

Rettferdighetsprinsippet står sentralt i flere andre lovverk, blant annet ulike menneskerettighetsbestemmelser og likestillings- og diskrimineringsloven (ldl.). Disse regelverkene får betydning for tolkning av begrepet rettferdighet, der kravene i noen tilfeller kan være strengere og mer spesifikke enn personvernregelverket.

Hva sier likestillings- og diskrimineringsloven?

Likestillings- og diskrimineringsloven (ldl.) forbyr diskriminering på grunn av «kjønn, graviditet, permisjon ved fødsel eller adopsjon, omsorgsoppgaver, etnisitet[9], religion, livssyn, funksjonsnedsettelse, seksuell orientering, kjønnsidentitet, kjønnsuttrykk, alder eller kombinasjoner av disse grunnlagene», jf. § 6.

Diskriminering defineres som usaklig forskjellsbehandling og kan forekomme «direkte» eller «indirekte», se henholdsvis ldl. §§ 7 og 8.

Direkte forskjellsbehandling betyr at personer med diskrimineringsvern behandles dårligere enn andre sammenlignbare personer, jf. § 7, mens indirekte forskjellsbehandling betyr at en tilsynelatende nøytral bestemmelse eller praksis fører til at personer med diskrimineringsvern stilles dårligere enn andre, jf. § 8. Indirekte diskriminering kan for eksempel oppstå ved at en tilsynelatende nøytral algoritme benyttes ukritisk på alle pasientgrupper. Som følge av at forekomsten av hjertesvikt har vært gjennomgående høyere blant menn enn

---

[5] Guidelines 4/2019 on Article 25 Data Protection by Design and by Default | European Data Protection Board (europa.eu)

[6] Fortalepunkt 39

[7] https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/ferdige-prosjekter-og-rapporter/helse-bergen-sluttrapport-kunstig-intelligens-i-oppfolging-av-sarbare-pasienter/

[8] Fortalepunkt 71

[9] Det følger av bestemmelsen at med «etnisitet» menes blant annet nasjonal opprinnelse, avstamning, hudfarge og språk

kvinner[10], vil kvinnelige hjertesviktpasienter i mindre grad være representert i datagrunnlaget som kan gi kvinner mindre treffsikre prediksjoner. Både direkte og indirekte diskriminering krever årsakssammenheng mellom forskjellsbehandlingen og diskrimineringsgrunnlaget, altså at en person stilles dårligere på grunn av vedkommende kjønn, alder, funksjonsnedsettelse, etnisitet og ligenende.

Offentlige myndigheter har videre en aktivitetsplikt i ldl. § 24 til å arbeide aktivt, målrettet og planmessig for å fremme likestilling og hindre diskriminering. Arbeidet i sandkassa kan benyttes som et eksempel på et prosjekt som har til hensikt å fremme likestilling og forhindre diskriminering gjennom konkrete tiltak i offentlig helsetjeneste.

Rettferdighet som etisk prinsipp

Rettferdighet er også et etisk prinsipp, som innebærer at etiske vurderinger står sentralt i både tolkning av, og anvendelse av rettferdighetsprinsippet i personvernforordningen. I «Etiske retningslinjer for pålitelig kunstig intelligens»[11] utarbeidet av en ekspertgruppe oppnevnt av EU-kommisjonen, nevnes det tre hovedprinsipper for ansvarlig kunstig intelligens, nemlig lovlig, etisk og sikker kunstig intelligens. De samme prinsippene gjenspeiles i regjeringens Nasjonale strategi for kunstig intelligens [12] fra 2020.

I etikken reiser man spørsmål om hvordan man *burde* oppføre seg og handle for å minimere de etiske konsekvensene som kan oppstå ved bruk av kunstig intelligens. Selv om noe er innenfor loven i rettslig forstand, kan man likevel spørre seg om det er etisk riktig å utføre handlingen. Etiske refleksjoner burde stilles i alle faser av en algoritmes liv, henholdsvis i utviklingsfasen, når algoritmen brukes i praksis og i etterlæringsfasen.

I etikken ønsker man å besvare spørsmål som «hva er bra og dårlig, godt og ondt, riktig og galt eller rettferdig og lik behandling for alle?». En etisk tilnærming til EKG AI-prosjektet vil være å stille seg spørsmål om hvorvidt algoritmen gir like gode prediksjoner for alle pasienter. Ved økt bruk av algoritmer i klinisk behandling i fremtiden, vil det være relevant å stille spørsmål om de generelle fordelene kunstig intelligens bringer med seg kommer alle pasienter til gode.

Algoritmeskjevhet

Løsninger som er basert på en programkode vil, naturlig nok, inneholde feil eller unøyaktigheter. Dette gjelder både for svært avanserte systemer og for enkle løsninger, men jo mer omfattende koden er, jo større er sjansen for feil. I maskinlæringsløsninger vil feilene vanligvis føre til at prediksjonen algoritmen gir, blir mindre nøyaktig eller gir galt resultat. En feil som systematisk gir mindre nøyaktige, eller gale, prediksjoner for enkelte grupper vil være eksempel på det vi kaller en algoritmeskjevhet.

Når to pasienter har et tilnærmet likt helsebehov, skal disse få like omfattende helsehjelp – uavhengig av etnisitet, kjønn, funksjonsnedsettelse, seksuell orientering og lignende. Hvis en algoritme derimot anbefaler at de skal motta ulik grad av bistand, kan det være grunn til å mistenke en form for diskriminering.

Det finnes mange mulige årsaker til algoritmeskjevhet og årsakene er ofte sammensatte. I dette prosjektet har vi konsentrert oss om fire årsaker til algoritmeskjevhet. Begrunnelsen er at disse fire årsakene har vært mest aktuelle for vårt prosjekt, men er også vanlige for kunstig intelligens generelt. Beskrivelsen og illustrasjonen nedenfor baserer seg på

Kunstig intelligens

Kunstig intelligens ble i Regjeringens nasjonale strategi for kunstig intelligens definert som «systemer som utfører handlinger, fysisk eller digitalt, basert på tolkning og behandling av strukturerte eller ustrukturerte data, i den hensikt å oppnå et gitt mål. Enkelte KI-systemer kan også tilpasse seg gjennom å analysere og ta hensyn til hvordan tidligere handlinger har påvirket omgivelsene.» Begrepet algoritme brukes ofte om programkode i et system, det vil si oppskriften på hva systemet skal finne løsningen på.

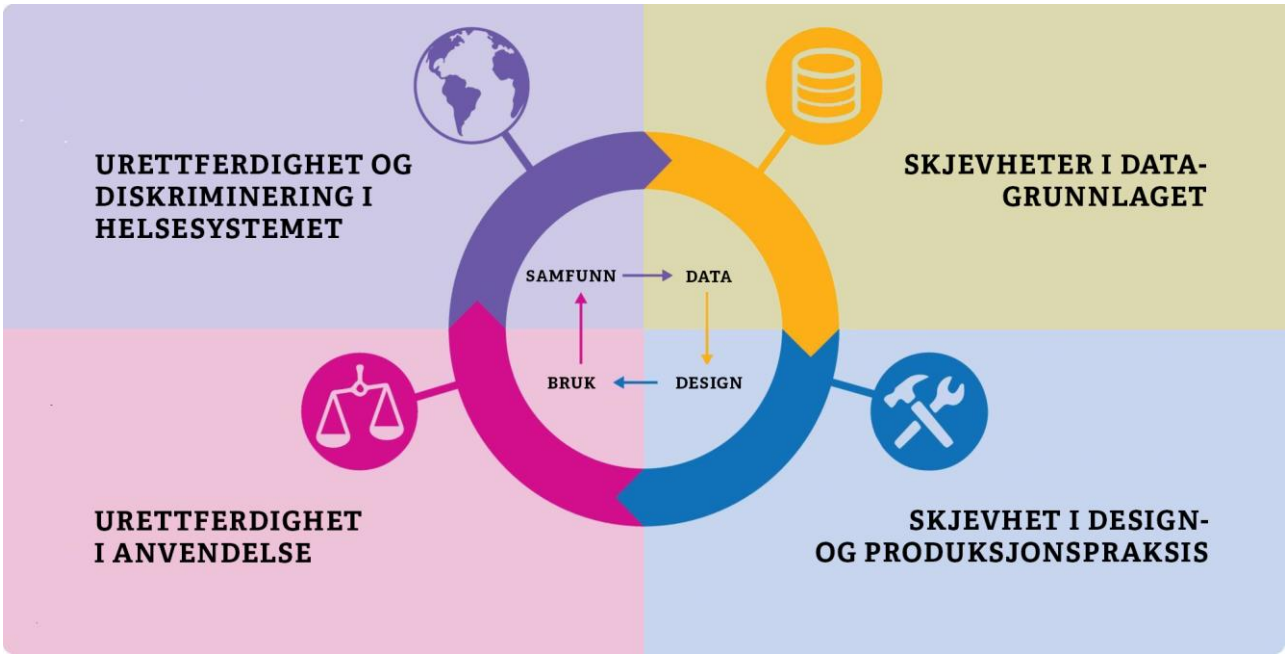
---

[10] Hjerte- og karregisteret: Rapport for 2012–2016. Hentet fra https://www.fhi.no/globalassets/dokumenterfiler/rapporter/2016/hjerte--og-karregisteret.-rapport-for-2012-2016.pdf

[11] Etiske retningslinjer for pålitelig kunstig intelligens , 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[12] Nasjonal strategi for kunstig intelligens, 2020, <https://www.regjeringen.no/no/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/>

en fremstilling i «British medical journal», som har et særlig fokus på algoritmeskjevhet i helsesektoren. Vi ønsker å understreke at flere av årsakene kan føre til samme skjevhet, og at det eksisterer en overlapp mellom de forskjellige årsakene.



1 Urettferdighet og diskriminering i helsevesenet

En maskinlæringsalgoritme lager prediksjoner basert på statistisk sannsynlighet for visse utfall. Statistikken vil basere seg på historiske data som gjenspeiler virkeligheten, herunder eksisterende urettferdighet og diskriminering i helsevesenet. Det kan være snakk om ekskluderende systemer, helsepersonell med fordommer eller ulik tilgang til helsehjelp. Dette videreføres og forsterkes i algoritmen.

2 Skjevheter i datagrunnlaget

Urettferdighet i samfunnet vil gjenspeile seg i datagrunnlaget som tilføres algoritmen under trening. Dersom det foreligger skjevheter i faktagrunnlaget vil også algoritmens prediksjoner gjenspeile disse skjevhetene. Har man ikke et representativt mangfold i treningsdataene vil ikke algoritmen være i stand til å gi presise prediksjoner for underrepresenterte individer eller grupper. Dette kan eksempelvis oppstå ved at ulike befolkningsgrupper har ulik tilgang til helsehjelp av sosioøkonomiske årsaker.

3 Skjevhet i design- og produksjonspraksis

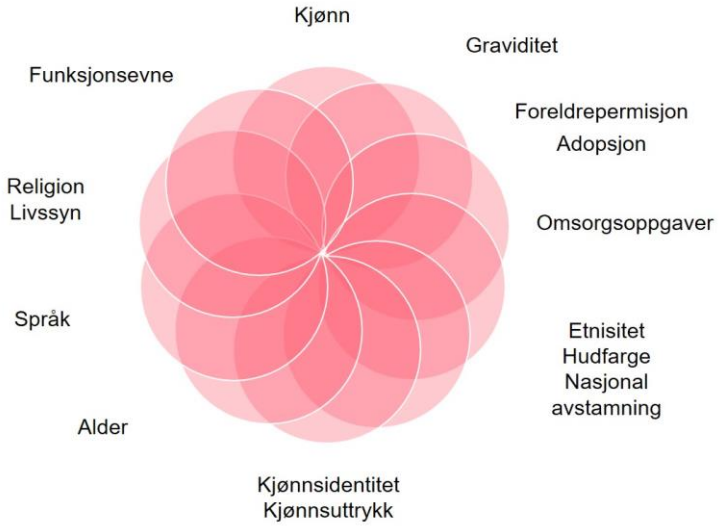
Skjevheter i algoritmen kan også springe ut av utviklernes fordommer og valg som tas underveis i utviklingen. En begrunnelse kan være at utviklerne har manglende kompetanse om, eller forståelse for, mulige diskriminerende utfall av designvalgene de gjør. En annen mulighet er at det ikke eksisterer systemer for å fange opp utilsiktet diskriminering når algoritmen benyttes i praksis.

4 Urettferdighet i anvendelse

En algoritme som inneholder skjulte skjevheter vil forsterke en allerede diskriminerende praksis. I tilfeller der algoritmen lærer av egne prediksjoner, vil skjevheten forsterkes ytterligere i algoritmen. Bruk av en algoritme som er programert med feil formål vil også kunne føre til diskriminering i praksis, eksempelvis at formålet om å avdekke en persons helsebehov begrunnes med hvor mye penger personen har brukt på helsetjenester.

Hvordan avdekke algoritmeskjevhet?

Kunstig intelligens har et stort potensiale til å forbedre diagnostikk og behandling av sykdom. Samtidig, har man i flere tilfeller sett at kunstig intelligens i stor grad viderefører og forsterker eksisterende diskriminering i samfunnet. Sandkasseprosjektet har ønsket å se nærmere på hvordan man kan avdekke potensielle skjevheter i EKG AI-algoritmen. For å sikre en rettferdig algoritme ønsket vi å undersøke om EKG AI gir mindre treffsikre prediksjoner for enkelte pasientgrupper. Med god veiledning fra LDO, tok vi utgangspunkt i diskrimineringsgrunnlagene[13] oppstilt i ldl. § 6. Vi ble oppmerksomme på at det i noen tilfeller ikke er tilstrekkelig å se på diskrimineringsgrunnlagene isolert, men krysskontrollere diskrimineringsgrunnlagene, som eksempelvis «minoritetskvinne».



I Norge har vi et gratis helsetilbud for alle, som legger et grunnlag for et representativt datagrunnlag med en stor variasjon av pasienter. Det kan likevel forekomme tilfeller der enkelte pasientgrupper ikke har hatt den samme tilgangen til en spesifikk behandling, og på denne måten utgjør en kilde til skjevhet i algoritmen. Vi har i prosjektet blant annet valgt å se nærmere på hvorvidt EKG AI gir varierende grad av treffsikre prediksjoner for pasienter med ulik etnisk bakgrunn. En av grunnene til at vi valgte å se nærmere på diskrimineringsgrunnlaget etnisitet var fordi det finnes flere eksempler[14] på algoritmer som systematisk diskriminerer etniske minoriteter. Fra et helseperspektiv kan ulik etnisk bakgrunn gi varierende symptomer på hjertesvikt, varierte EKG-målinger og blodprøver. I denne sammenheng er det viktig å presisere at begrepet «etnisitet» brukes som en referanse for pasientens biologiske eller genetiske opphav.

Utfordringen for Ahus er at det finnes ingen eller begrenset data om pasientenes etnisitet. Når Ahus ikke har tilgang til disse opplysningene vil det heller ikke være mulig å kontrollere hvorvidt algoritmen har mindre presise prediksjoner for denne pasientgruppen. For å være i stand til å avdekke om det eksisterer en slik skjevhet, må Ahus gjennomføre en klinisk studie. I denne studien kan opplysninger om etnisitet samles inn basert på samtykke og det vil være mulig å kontrollere om etniske minoriteter kommer systematisk dårligere ut enn majoriteten av pasientene algoritmen er trent på.

---

[13] For at forskjellsbehandlingen skal anses som diskriminering etter ldl. må den ha sin årsak i ett eller flere diskrimineringsgrunnlag. Les mer om de ulike diskrimineringsgrunnlagene på Diskrimineringsnemda sin nettside: https://www.diskrimineringsnemnda.no/klagegrunnlag/diskriminering

[14] Blant annet en algoritme som diskriminerer etniske minoriteter i helsevesenet: https://www.wired.com/story/how-algorithm-favored-whites-over-blacks-health-care/
Og Amazon sin rekrutteringsalgoritme som systematisk diskriminerte kvinnelige søkere: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

For å kunne kontrollere om det eksisterer skjevheter i EKG AI, må det først fastsettes en grenseverdi for statistisk usikkerhet i algoritmen. Som følge av at det alltid vil eksistere fordommer og diskriminering i den virkelige verden, vil også algoritmer inneholde feil og skjevheter. Det vesentlige har derfor vært å finne ut av hvor grensen går for en uakseptabel forskjellsbehandling i EKG AI.

Ønsket vs. uønsket forskjellsbehandling

Selve formålet med EKG AI er å prioritere de pasientene som har en pågående hjertesvikt, fremfor de som ikke har hjertesvikt. Dette er en form for forskjellsbehandling. Det avgjørende i vurderingen etter diskrimineringsregelverket er om det skjer en usaklig forskjellsbehandling.

For å kunne fastsette en slik grenseverdi må det gjennomføres en klinisk studie som undersøkeler om algoritmen har en redusert nøyaktighet for enkelte pasientgrupper basert på et eller flere diskrimineringsgrunnlag. Viser det seg for eksempel at EKG AI prioriterer pasientgrupper på grunnlag av etnisitet istendefor medisinske faktorer som EKG, diagnosekoder etc., vil dette anses som en usaklig forskjellsbehandling etter diskrimineringsregelverket.

Når man skal kontrollere om det eksisterer diskriminering i algoritmer, kreves det som regel innsamling og behandling av nye personopplysninger. Et nytt formål krever derfor en ny vurdering av behandlingens rettslig grunnlag etter artikkel 6, og eventuelt artikkel 9, i personvernforordningen.

Det er interessant å merke seg at EU-kommisjonens forslag til ny forordning om kunstig intelligens (AI act) artikkel 10-5 oppstiller en adgang til å behandle særlige kategorier av personopplysninger, som for eksempel helseopplysninger, dersom det er strengt nødvendig, til formål om å overvåke, oppdage og korrigere algoritmeskjevhet. Blir denne bestemmelsen vedtatt har man et lovgrunnlag for å i enkelte tilfeller kunne undersøke om det eksisterer diskriminering i algoritmer ved behandling av særlig kategorier av personopplysninger.

Dataminimeringsprinsippet i personvernforordningen artikkel 5 vil likevel sette en begrensning for hvilke personopplysninger man kan behandle. Prinsippet krever at opplysningene som behandles skal være adekvate, relevante og begrenset til det som er nødvendig for å oppnå formålet med behandlingen.[15] Nødvendighetskravet inneholder også en vurdering av behandlingens forholdsmessighet. I vurderingen av om Ahus skal samle inn og behandle opplysninger om pasienters etnisitet, har vi derfor vurdert om det er forholdsmessig sett opp mot konsekvensene en potensiell skjevhet i algoritmen vil ha for den enkelte pasient.

Prediksjonene fra EKG AI brukes kun som én av mange informasjonskilder i helsepersonellets vurdering av den videre oppfølging av pasienten. Dette innebærer at konsekvensen av en eventuell skjevhet i algoritmen vil utgjøre en liten skaderisiko. Dersom algoritmen for eksempel ikke fanger opp hjertesvikt hos en pasient (falsk negativ), vil likevel hjertesvikten kunne avdekkes med ultralyd av hjertet og etterfølgende blodprøver. Ahus må derfor stille seg spørsmål om innsamling og behandling av opplysninger om etnisitet, som er særlige kategorier av personopplysninger, er en forholdsmessig behandling sett opp mot å avdekke en eventuell diskriminering i algoritmen.

Det finnes ingen klare svar på hvor en slik grense skal settes. I noen tilfeller er det først etter at opplysningene er samlet inn og behandlet at man kan si noe om behandlingen har vært nødvendig for å avdekke diskriminering. I medisinsk sammenheng er det imidlertid like viktig å få avdekket hva som ikke er relevant, som hva som anses relevant for forsvarlig helsehjelp.

---

[15] personvernforordningen artikkel 5 nr. 1 bokstav c

Tiltak som kan redusere algoritmeskjevhet

I sandkasseprosjektet har vi diskutert hvordan Ahus kan redusere algoritmeskjevhet både i selve modellen, gjennom tekniske tiltak, og korrigere algoritmeskjevhet etter at algoritmen er tatt i bruk, gjennom organisatoriske tiltak.

Det presiseres i personvernforordningen fortalepunkt 71 at en rettferdig behandling av personopplysninger innebærer «gjennomføring av egnede tekniske og organisatoriske tiltak for særlig å sikre at faktorer som fører til uriktige personopplysninger, rettes opp og at risikoen for feil minimeres, sikre personopplysningene på en måte som tar hensyn til den registrertes interesser og rettigheter, og hindre blant annet forskjellsbehandling av fysiske personer (...)».

Likestillings- og diskrimineringsombudets kommentar:

Manglende presisjon i algoritmen for definerte pasientgrupper er ikke ensbetydende med at Ahus som rettssubjekt diskriminerer i sin praksis. Det avgjørende i henhold til likestillings- og diskrimineringsloven er at personer med diskrimineringsvern ikke «behandles dårligere enn andre», jf. § 7. Det avgjørende i denne sammenheng er om forskjellsbehandlingen fører til «skade eller ulempe for den som forskjellsbehandles, for eksempel at forskjellsbehandlingen fører til tap av fordeler, økonomisk tap eller færre muligheter sammenlignet med andre i en tilsvarende situasjon. Forholdet må ha en konkret og direkte betydning for bestemte fysiske personer.»[16]

Ombudet er av den oppfatning at Ahus kan kompensere for algoritmeskjevhet, ved å supplere med andre medisinske metoder i undersøkelse av pasientgruppene algoritmen ikke gir like gode prediksjoner for. Det avgjørende for å oppnå like god behandling for alle, er å vite *hvilke* pasientgrupper algoritmen er mindre presis for, og *iverksette tiltak* slik at disse pasientene får et like godt tilbud som andre.

Et generelt eksempel på tekniske tiltak er å pseudonymisere, kryptere eller anonymisere personopplysninger for å minimere personverninngrepet overfor den registrerte. Organisatoriske tiltak kan for eksempel være innføring og gjennomføring av rutiner og praksis som bidrar til at regelverket etterleves. Mer informasjon om implementering av egnede tiltak i programvareutvikling finnes på Datatilsynets nettside.[17]

Tre typer tiltak

Vi har særlig fokusert på tre ulike typer tiltak i dette prosjektet:

1. Analysere og kontrollere datagrunnlaget til algoritmen (teknisk tiltak)
2. Etablere rutiner som sørger for opplæring av helsepersonell i bruk av beslutningsstøttesystemet (organisatorisk tiltak)
3. Etablere en overvåkingsmekanisme for etterlæring (teknisk tiltak)

[16] Prop. 81 L (2016-2017), kap. 30, kommentar til § 7

[17] Datatilsynet nettside om innebygget personvern, https://www.datatilsynet.no/rettigheter-og-plikter/virksomhetenes-plikter/innebygd-personvern/programvareutvikling-med-innebygd-personvern/

Teknisk tiltak: Analysere og kontrollere datagrunnlaget til algoritmen

I Nasjonal strategi for kunstig intelligens[18] fremheves særlig skjevhet i datagrunnlaget som et hinder for inkludering og likebehandling. Dette forklares med at «datasett som brukes til å trene opp KI-systemer kan inneholde historiske skjevheter, være ufullstendige eller uriktige». Dårlig datakvalitet og feil i datagrunnlaget til EKG AI vil dermed forplante seg og forsterkes i algoritmen og kan føre til uriktige og diskriminerende resultater. Dårlig datakvalitet vil eksempelvis kunne forårsakes av gjentatte feildiagnostiseringer av helsepersonell. For at algritmen skal klare å finne et korrekt mønster i datasettet må opplysningene være konsistente og samsvare med virkelige fakta.

Ahus har benyttet historiske helsedata som har en statistisk sammenheng med risikoen for hjertesvikt. Datagrunnlaget til algoritmen består av resultater fra omtrent 100 000 elektrokardiogrammer (EKG-målinger), ICD-10 diagnosekoder for hjertesvikt og opplysninger om hjertets pumpeevne. Denne informasjonen er hentet fra kardiologisystemet og journalsystemet DIPS på Ahus.

I diskusjonen har vi særlig sett på to ulike metoder for å minimere algoritmeskjevhet i datagrunnlaget:

- **Metode 1** tilfører ikke ny data til algoritmen, men man "blåser opp" de pasientgruppene som er underrepresenterte i datagrunnlaget. Utfordringen med denne metoden er at den gir bedre treffsikkerhet for enkelte pasientgrupper, men vil medføre dårligere presisjon for majoritetsgruppen av pasienter.

- **Metode 2** tilfører algoritmen flere datapunkter (labels) knyttet til den underrepresenterte pasientgruppen. Man vil med denne teknikken måtte akseptere større grad av diskriminering i startfasen, men som over tid vil gi algoritmen bedre treffsikkerhet. Utfordringen med denne metoden er at man ikke nødvendigvis har opplysningene som er nødvendig for å rette opp i en avdekket skjevhet.

Det er først etter man har dokumentert en representasjonsskjevhet i algoritmen, at man kan iverksette korrigerende tiltak. Det har vært sentralt i diskusjonene å avdekke hvilke samfunnsgrupper som i liten grad er representert i datagrunnlaget til algoritmen. Vi har særlig diskutert representasjon av kjønn og etnisitet, da det finnes eksempler[19] på tidligere algoritmer som har vist seg å diskriminere på disse grunnlagene.

Likestillings- og diskrimineringsombudets synspunkt: I Ahus' EKG AI-prosjekt er ombudet av den oppfatning at det er størst risiko for diskriminering knyttet til representasjonsskjevhet i datagrunnlaget. Forskningslitteraturen[20] på hjertesvikt peker på at hjerterytmen varierer for ulike etnisiteter. Etniske minoriteter vil altså kunne ha varierende EKG-kurver enn majoritetsbefolkningen i Norge, noe som det bør tas høyde for både i utviklingen og anvendelsen av algoritmen.

Opplysninger om etnisitet registreres ikke i pasientjournalen, og finnes heller ikke tilgjengelig i andre nasjonale kilder. Det gir Ahus begrensede muligheter for å kontrollere om treffsikkerheten til algoritmen er dårligere for etniske minoriteter enn den etniske majoriteten av pasientene. For at Ahus skal få tilgang til disse opplysningene må det gjennomføres en klinisk studie[21] basert på frivillig innsamling av data, og på denne måten få kontrollert algoritmens prediksjoner for denne pasientgruppen. I løpet av sandkasseprosjektet har Ahus argumentert for et behov for å registrere opplysninger om pasienters genetiske opphav for å sikre at helsetjenesten yter forsvarlig helsehjelp til alle. Registrering av etnisitet er et sensitivt tema, og hvorvidt, og hvordan dette kan gjennomføres i praksis er opp til nasjonale myndigheter å vurdere.

---

[18] Nasjonal strategi for kunstig intelligens, 2020, kapittel 5, <https://www.regjeringen.no/no/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/>
[19] Diskriminering av etniske minoriteter i helsevesenet: https://www.wired.com/story/how-algorithm-favored-whites-over-blacks-health-care/
Og Amazon sin rekrutteringsalgoritme som systematisk diskriminerte kvinnelige søkere: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
[20] Blant annet: JCDD | Free Full-Text | The Impact of Ethnicity on Athlete ECG Interpretation: A Systematic Review | HTML (mdpi.com)
og
Assessing and Mitigating Bias in Medical Artificial Intelligence: The Effects of Race and Ethnicity on a Deep Learning Model for ECG Analysis - PMC (nih.gov)
[21] Ifølge Helsenorge.no er kliniske studier definert som «forskning på effekten av nye legemidler eller nye behandlingsmetoder, og på om bivirkningene er akseptable.», hentet fra: https://www.helsenorge.no/kliniske-studier/om/

Opplysninger om pasientens kjønn er en del av datagrunnlaget til EKG, kategorisert som «mann» og «kvinne». Hvorvidt algoritmen har lavere treffsikkerhet for pasienter som ikke definerer seg som mann eller kvinne, eller har gjennomgått kjønnsskorrigerende operasjon, finnes det ikke tall på. Dette må det i så fall opprettes en klinisk studie for å undersøke. Forekomsten av førstegangstilfeller av hjertesvikt har vært gjennomgående høyere blant menn enn kvinner.[22] I følge Ahus, krever ulike typer av hjertesvikt ulik behandling og oppfølging, der kvinner er høyest representert i en av de tre typene av hjertesvikt. Som følge av at kvinner historisk sett har vært underrepresentert i medisinsk forskning, foreligger det en generell risiko for at kvinner kommer dårligerer ut enn menn når algoritmen er utviklet på historiske helseopplysninger. Ahus kan imidlertid bekrefte at det finnes et godt datagrunnlag for kvinnelige hjertesviktpasienter i EKG AI og at det derfor er liten sannsynlighet for diskriminering av kvinner i EKG AI.

Dersom algoritmen i fremtiden selges videre til andre aktører i, eller utenfor, Norge, oppstår det en fare for at datagrunnlaget ikke representerer den nye pasientgruppen algoritmen skal gi prediksjoner for. For å opprettholde en høy treffsikkerhet må algoritmen ettertrenes på ny og lokalt tilpasset pasientinformasjon, som vil bli behandlet nærmere under tiltak tre «Overvåkingsmekanisme for etterlæring».

Organisatorisk tiltak: Algoritme møter helsepersonell

Fram til nå har utviklingen av beslutningsstøtteverktøyet EKG AI foregått på laboratoriet. Det er nødvendig å gjennomføre en klinisk studie for å kontrollere algoritmens treffsikkerhet og presisjon på virkelige data før den tas i bruk i klinikk. For at beslutningsstøtteverktøyet skal fungere optimalt, må resultatet fra algoritmens beregninger presenteres for helsepersonellet på en måte som gjør at prediksjonen fungerer som tiltenkt, altså som et beslutningsstøtteverktøy for raskere diagnostisering av hjertesvikt enn i dag. Resultatet må nå frem til mottaker umiddelbart, helsepersonell må forstå svaret og kunne anvende det på riktig måte. Det overordnede målet er at EKG AI skal gi større treffsikkerhet og presisjon enn hva helsepersonell alene har kapasitet til, slik at den bidrar til forbedret og raskere helsehjelp.

Artikkel 22 i personvernforordningen oppstiller et forbud mot beslutninger som kun er basert på automatisert behandling. For at et slikt forbud skal gjøre seg gjeldende, må det være tale om en automatisert beslutning uten noe form for menneskelig innblanding. EKG AI skal brukes som et beslutningsstøtteverktøy og faller derfor ikke innenfor forbudet i artikkel 22. God informasjon og opplæring til helsepersonell om bruk av beslutningsstøtteverktøyet vil bidra til en reell menneskelig innblanding i beslutningsprosessen, og redusere risikoen for at helsepersonell legger algoritmens prediksjoner helt ukritisk til grunn i praksis.

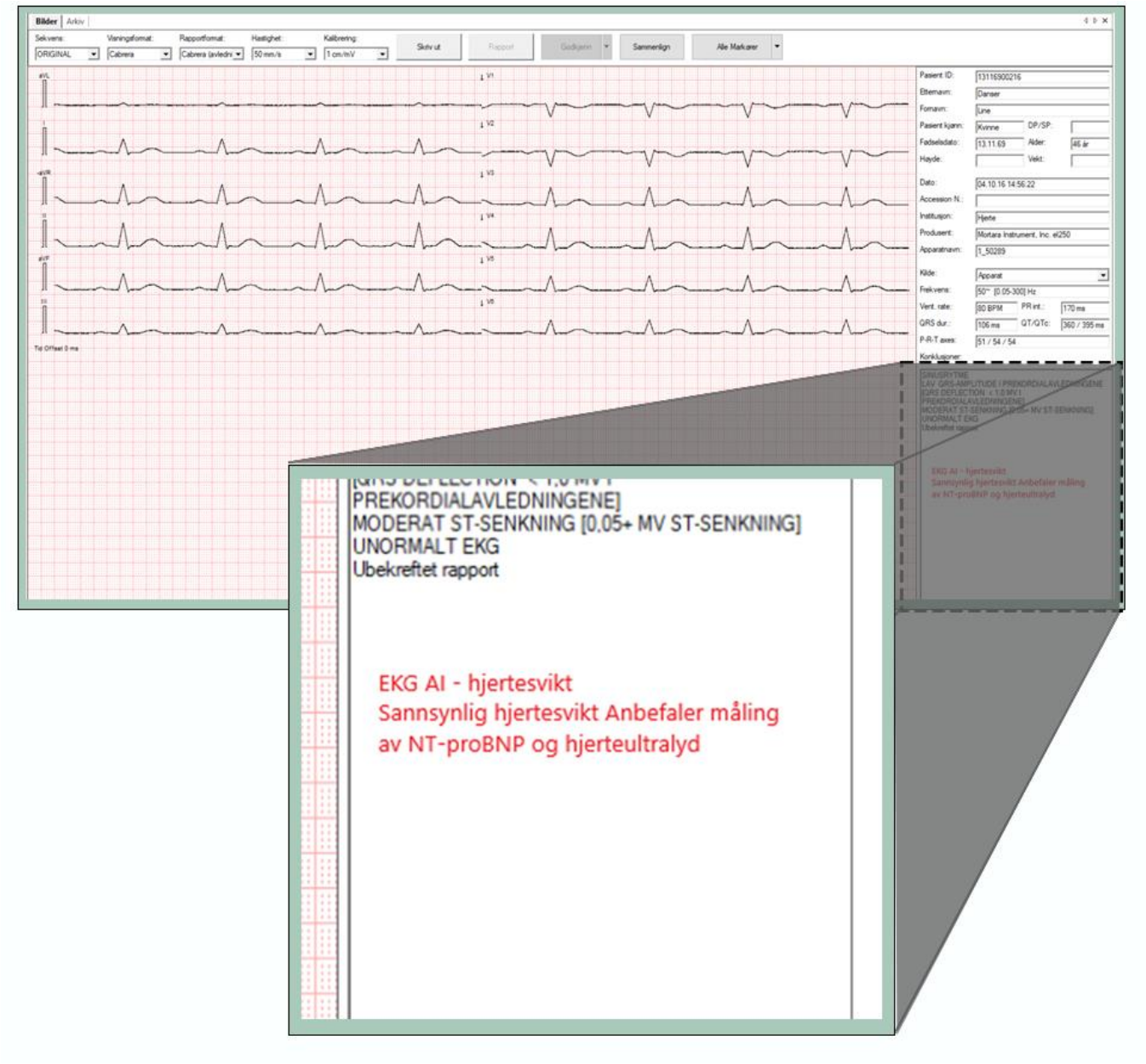
Algoritmer er avanserte tekniske systemer som forutsetter både kunnskap hos, og opplæring av, helsepersonell. Det krever først og fremst at helsepersonell får en forklaring på hvordan algoritmen skal brukes, men også foreliggende risikofaktorer og feilmargin i prediksjonene. Helsepersonell må selv kunne forstå algoritmens funksjon og bakgrunn for prediksjonene, for å forhindre mistillit til verktøyet. Innsikt og forståelse for modellens virkemåte er også avgjørende for at helsepersonell kan vurdere prediksjonen på et selvstendig grunnlag. Dersom det i fremtiden viser seg at algoritmens treffsikkerhet varierer mellom ulike pasientgrupper, må dette også kommuniseres til helsepersonell. Opplæringen kan inkludere informasjon om diskrimeringsgrunnlag og at helsepersonell selv må være oppmerksom på disse.

Helsepersonellet skal kun ta i bruk EKG AI algoritmen dersom det foreligger mistanke om hjertesvikt hos pasienten. Resultatet fra algoritmen sendes tilbake til EKG-arkivet (ComPACS) på Ahus. Svaret presenteres for helsepersonellet i et tekstlig format i et felt ved siden av pasientens EKG-måling. Ahus er klar over risikoen for alarmtrøtthet hos helsepersonell dersom mange varslinger, og mye informasjon, dukker opp samtidig. I og med at EKG AI skal brukes i en akuttsituasjon vil det særlig være viktig med klar og presis informasjon. I sandkasseprosjektet har vi derfor diskutert ulike måter prediksjonen kan presenteres på. Prediksjonen kan presenteres som prosenttall, kategoriene «lav», «middels» og «høy», eller det kan settes en grense for akutt/ikke akutt oppfølging. I denne prosessen vil det være hensiktsmessig å involvere helsepersonell for å få innspill til hvordan resultatene kan presenteres på en god måte. Hvordan prediksjonen helt konkret skal formuleres vil Ahus

---

[22] Hjerte- og karregisteret: Rapport for 2012–2016. Hentet fra https://www.fhi.no/globalassets/dokumenterfiler/rapporter/2016/hjerte--og-karregisteret.-rapport-for-2012-2016.pdf

undersøke nærmere i en klinisk studie. Her vil man kunne teste ut beslutningsstøttverktøyet i praksis på et antall pasienter og med utvalgte helsepersonell.



Illustrasjon som viser hvordan resultatet fra EKG AI kan presenteres for helsepersonell.

Som følge av at resultatet fra algoritmen lagres i EKG arkivet (ComPacs), vil det foreligger en risiko for at resultatet ikke blir oppdaget av helsepersonell umiddelbart. Ahus vil derfor se på alternative muligheter for å varsle, for eksempel via skjermer på sykehuset eller på helsepersonellets jobbtelefon. Slike varslingsløsninger forutsetter imidlertid bedre infrastruktur enn det som finnes på Ahus idag.

I forbindelse med opplæring av helsepersonell vil Ahus etablere rutiner og protokoller for bruk av verktøyet. Når algoritmen videreutvikles er det naturlig at rutiner, protokoller og selve opplæringsløpet også oppdateres. Ahus har et eksisterende avvikssystem hvor helsepersonell melder inn avvik som kan brukes til å forbedre algoritmen i fremtiden.

Teknisk tiltak: Overvåkingsmekanisme for etterlæring

Med tiden og skiftende forhold i samfunnet vil EKG AIs prediksjoner bli mindre nøyaktige. Lavere treffsikkerhet vil naturligvis oppstå når en befolkningsmasse endrer seg. For eksempel kan det komme til nye pasientgrupper på grunn av økt flyktningestrøm fra et land algoritmen ikke tidligere har datagrunnlag på. Når treffsikkerheten ikke lenger er tilfredsstillende vil det være aktuelt med etterlæring av algoritmen. Etterlæring vil innebære trening med nye opplysninger, testing og validering av algoritmen.

EKG AI algoritmen vil ikke etterlæres kontinuerlig, som innebærer at nøyaktigheten ikke automatisk tilpasser seg fremtidige endringer. Ahus planlegger istedenfor å implementere en overvåkningsmekanisme som skal varsle når algoritmens treffsikkerhet faller under en forhåndsbestemt grenseverdi og algoritmen har behov for etterlæring. For å validere en slik grenseverdi vil Ahus gjennomføre en klinisk studie samtidig som algoritmen testes ut i klinikk.

Helt praktisk vil overvåkningsmekanismen sammenligne algoritmens prediksjoner med den diagnosen helsepersonellet stiller pasienten. På denne måten vil man kunne vurdere i hvilken grad algoritmen predikerer riktig i forhold til pasientens faktiske medisinske tilstand. Grenseverdien for treffsikkerhet vil deretter avgjøre når, og hvorvidt, det er behov for etterlæring av algoritmen.

Veien videre

Dersom prosjektet lykkes i å utvikle en god prediksjonsmodell, er målet å prøve den ut i klinisk drift i starten av 2024. Neste steg er å få algoritmen CE-merket og godkjent av Statens legemiddelverk.

Klinisk beslutningsstøtteverktøy basert på kunstig intelligens anses som medisinsk-teknisk utstyr og må godkjennes for å kunne brukes i klinisk virksomhet. Et viktig skille mellom alminnelig medisinsk-teknisk utstyr og et beslutningsstøttesystem basert på kunstig intelligens, er at sistnevnte må retrenes jevnlig for å forhindre at treffsikkerheten reduseres. Statens legemiddelverk har ikke praksis i dag for å godkjenne medisinsk-teknisk utstyr som må retrenes. Ahus har derfor mottatt prosjektstøtte fra Live Science Growth House for å, sammen med DNV, utforske mulighetene som finnes for å CE-merke en algoritme som retrenes.

Arbeidet i sandkasseprosjektet har synliggjort en potensiell risiko for at EKG AI kan diskriminere enkelte pasientgrupper som i dag ikke er like godt representert i datagrunnlaget algoritmen baserer seg på. Ahus vil gjennomføre en klinisk studie for å undersøke om algoritmen gir dårligere prediksjoner for pasienter med ulik etnisk bakgrunn og eventuelt andre aktuelle diskrimineringsgrunnlag. Resultatene vil vise om det er behov for å iverksette korrigerende tiltak.

I løpet av prosjektperioden har vi sett at det ikke finnes en felles og omforent metode for å avdekke algoritmeskjevhet. Dersom vi hadde hatt mer tid i prosjektet ville vi ha utviklet en egen metode, basert på erfaringer fra prosjektperioden. I tillegg ville det ha være interessant å gå enda dypere ned i de etiske kravene knyttet til bruk av kunstig intelligens i helsesektoren.

Datatilsynets regulatoriske sandkasse for ansvarlig kunstig intelligens

Besøksadresse:
Trelastgata 3, Oslo

Postadresse:
Postboks 458 Sentrum
0105 Oslo

sandkasse@datatilsynet.no
Telefon: +47 22 39 69 00

datatilsynet.no/sandkasse
personvernbloggen.no
twitter.com/datatilsynet