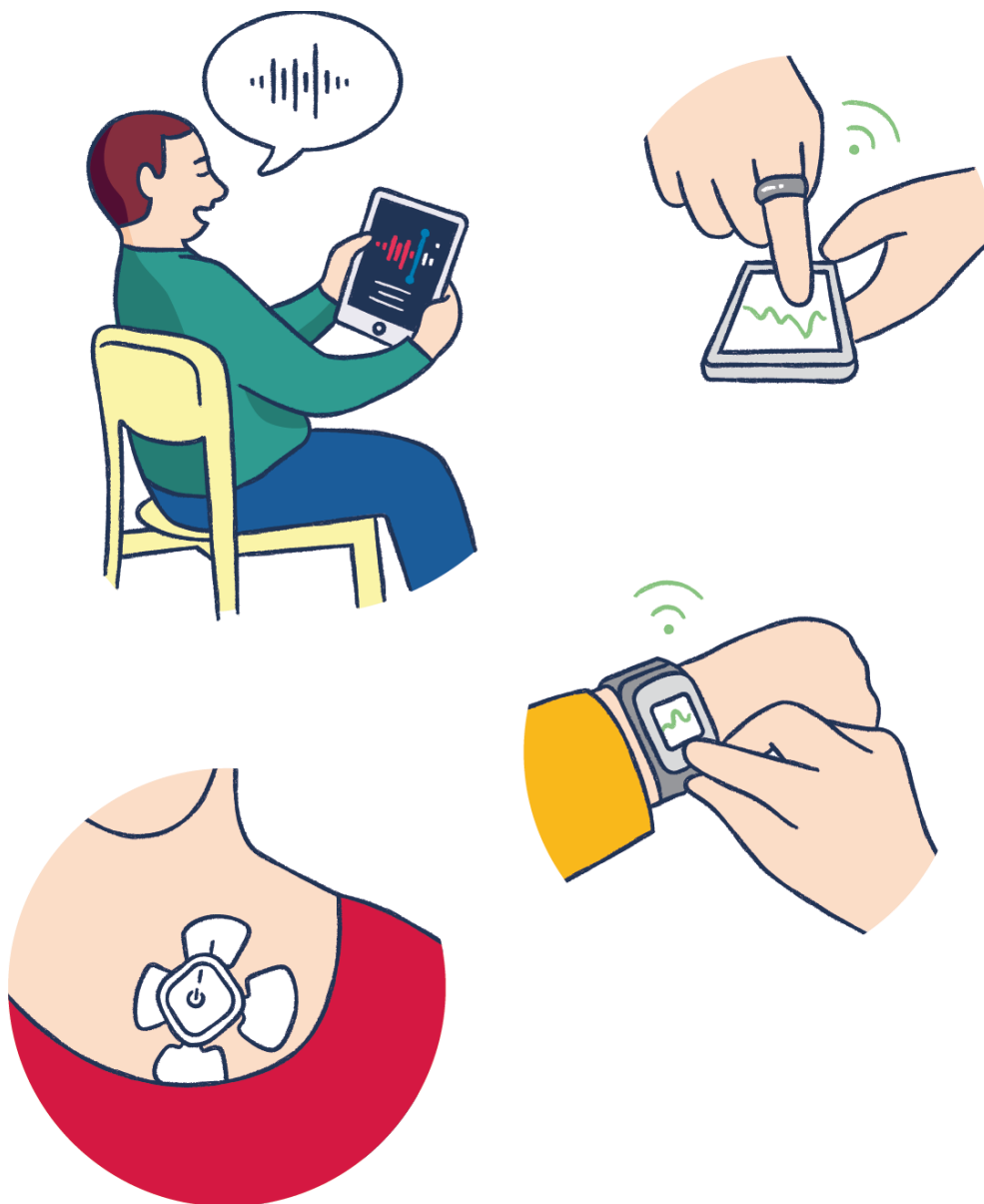


Erfaringer fra pilotforsøg i PreCareKlinikken

December 2023



Erfaringer fra pilotforsøg i PreCareKlinikken

© Region Sjælland, 2023

Forsideillustration: Urgent.Agency

Projekt: PreCare II - "Person-centric Health Innovation: New Service Models for Preventive Care Delivery to Elderly and Citizens with Chronical Conditions or at Risk of getting Chronical Conditions
PreCare Phase II"

Rapporten og analyserne er udarbejdet af Region Sjælland

Region Sjælland

Alléen 15, 4180 Sorø

Region Sjællands publikationer kan frit citeres med tydelig kildeangivelse.

Indhold

1. Baggrund og formål	5
2. Resume	5
2.1 Opmærksomhedspunkter	9
3. Pilotforsøg med Garmin Smart Watches	14
3.1 Metode	14
3.2 Analyse af data fra tilstandsmålinger	15
3.3 Resultater fra tilstandsmålingerne	15
3.4 Evaluering på baggrund af borgererfaringer	16
3.5 Perspektivering	17
4. Pilotforsøg med Oura-ringe	19
4.1 Metode	19
4.2 Data fra API	19
4.3 Analyse af data fra tilstandsmålinger	21
4.3.1 Logistisk regressionsmodellering	23
4.3.2 Opsummering	24
4.4 Observationer	25
4.5 Evaluering på baggrund af borgererfaringer	26
4.5.1 Resultat af evaluering - borgererfaring	26
4.5.2 Diskussion og konklusion – borgererfaringer	27
4.6 Perspektivering	28
5. Pilotforsøg med BioBeat-plastre	30
5.1 Metode	30
5.2 Data	31
5.4 Evaluering på baggrund af erfaringerne i hjemmeplejen og den fremskudte visitation	32
5.5 Evaluering på baggrund af erfaringerne fra borgerne	37
6. Pilotforsøg med PPG-teknologi	42
6.1 Metode	42
6.2 Data	43
6.3 Opmærksomhedspunkter	44
7. Pilotforsøg med stemmegenkendelse	47
7.1 Metode	47
7.2 Data	47
7.3 Analyse af data	49

8. Evaluering af KOL-algoritme i PreCareKlinikken	52
8.1 Metode	52
8.1.1 Baseline: Den nuværende KOL-algoritme	52
8.2 Resultater	53
8.2.1 Analyser af logisk regressionsmodel med fravælgelse af variable	53
8.4 Konklusion	54
8.5 Drøftelse af analyseresultaterne med det sundhedsfaglige personale	55
8.6 Perspektivering	58
9. Analyse af målerutiner	61
9.1 Baggrund	61
9.2 Data og metode	62
9.3 Resultater	64
9.3.1 Triagering af borgernes egenmålinger	64
9.3.2 Variansanalyse	64
9.3.3 Evaluering af andre modeller	65
9.3.4 Forudsigelse af forværringer tidligere	66
Bilag 1. Gennemgang af præliminære analyser af Garmin data 18.03.2022	69
Bilag 2. Pilotafprøvning af Oura-ringe – vurdering af datakvalitet, modellering og generelle overvejelser	77
Bilag 3. Relevante variable fra MEDCOM-beskeder	101
Bilag 4. Tekster oplæst i pilotforsøg med stemmegenkendelse	105
Bilag 5. Analyser af KOL-algoritme	110
Bilag 6. Målerutiner	126

Forord

En af hovedmålsætningerne i PreCare-projektet og i PreCareKlinikken er at kunne overdrage så meget ansvar og styring som muligt til borgerne selv. Det foregår gennem empowerment-fremmende tiltag og velfærdsteknologiske hjælperedskaber, der gør borgeren i stand til at monitorere sig selv under løbende supervision af PreCareKlinikken.

Udgangspunktet har været, at alle ikke har samme behov, samme kompetencer og samme tilstand, men at alle skal have mulighed for gennem målrettet anvendelse af teknologi og data at blive mødt af et proaktivt sundhedsvæsen.

Et vigtigt fokusområde i PreCare-projektet og i PreCareKlinikken har været at afprøve nye typer af digitale devices samt anvende avancerede analysemetoder på de mange data, som de inkluderede borgere er leverandører af, når de foretager egen-målinger af deres helbredstilstand.

Et af formålene har bl.a. været at gøre det nemt at foretage egen-målinger for så mange mennesker med kroniske sygdomme som muligt, så digitale kompetencer eller fysisk funktionsevne ikke spænder ben for at foretage en forebyggende indsats fremfor en akut indsats, når det er relevant.

PreCareKlinikken har med det for øje testet nogle digitale løsninger og AI-modeller i mindre pilotprojekter, og deler i denne rapport nogle af erfaringerne fra disse.

Rigtig god læsning.

Anne Bernth
Enhedschef, Next Generation Technology
Region Sjælland

Stine Bræstrup Arthur
Chefkonsulent, Omsorg og Sundhed
Odsherred Kommune

1. Baggrund og formål

I PreCare-projektet afprøves bl.a. telemedicinske løsninger for KOL-borgere. Det indbefatter egenmålinger af lungefunktion, iltmætning, temperatur samt besvarelse af tre spørgsmål om borgerens tilstand. Den telemedicinske løsning forudsætter, at borgeren er i stand til at måle sig selv/håndtere de udleverede devices. Det betyder, at borgere med kognitive udfordringer eller anden funktionsnedsættelse, der gør, at de ikke kan måle sig selv, ekskluderes fra projektet på trods af, at de har KOL.

Der er derfor i projektet afprøvet forskellige andre telemedicinske løsninger, der alle overordnet har haft til formål af gøre tilbuddet tilgængeligt for en bredere gruppe af borgere.

Derudover er der foretaget analyser på baggrund af data fra borgernes egenmålinger og medicinske behandling i forbindelse med forværringer for at undersøge, hvorvidt det er muligt at differentiere målefrekvens og antallet af måleparametre til at forudsige forværringer efter individuelle karakteristika. Formålet er at gøre det lettere at foretage egenmålinger og dermed benytte sig af de tilbudte services i PreCareKlinikken.

Denne rapport opsamler den viden, der er opnået gennem forskellige afprøvninger i mindre pilotforsøg/feasibility studier i PreCareKlinikken og analyser af data fra borgernes egenmålinger og tilknyttede services i PreCareKlinikken.

Der har været gennemført syv forskellige pilotforsøg samt evaluering og udvikling af algoritme i PreCareKlinikken. Det drejer sig om:

1. Afprøvning af Garmin Smart Watches
2. Afprøvning af Oura-ringe
3. Afprøvning af BioBeat-plastre
4. Afprøvning af kamerateknologi
5. Indsamling af stemmedata
6. Evaluering og udvikling af KOL-algoritme
7. Analyse af målerutiner

Hver pilotforsøg beskrives nedenfor ift. formål, resultatopnåelse samt perspektivering.

Fælles for pilotforsøgene har været at teste løsninger, der gennem viden om en borgers tilstand, kan gøre, at det er muligt at reagere så tidligt som muligt i forhold til forværringer og dermed tage dem i opløbet og om muligt undgå at behandlingen kræver akut indlæggelse på sygehus.

2. Resume

I PreCare-projektets indledende fase blev telemedicinske løsninger for KOL-borgere afprøvet med tekniske midler, der bl.a. krævede, at borgerne var tilstrækkeligt selvhjulpne til at kunne håndtere en tablet, spirometer og saturationsmåler mv. Det betød, at borgere med kognitive udfordringer i form af synsproblemer o.l. blev ekskluderet fra projektet på trods af, at de havde KOL.

Der blev derfor iværksat en række pilot-projekter for at finde metoder til at afhjælpe udfordringerne hos borgere, der ikke vil være i stand til at anvende det planlagte telemedicinske udstyr. Projekterne har haft fokus på at undersøge om det kan gøres nemmere at foretage egenmålinger, og om data kan pege i retning af, hvor hyppigt borgernes skal foretage målingerne for, at de opsamlede værdier kan anvendes til en forudsigelse og identifikation af forværringer, og dermed give borger og sundhedspersonale mulighed for i god tid til at reagere på kommende forværringer.

En prædiktionsmodel på data indsamlet uden aktive handlinger fra borgeren selv muliggør en udvidelse af PreCareKlinikkens målgruppe til også at inkludere borgere, der ikke selv kan varetage indsendelse af egenmålinger, herunder demente eller svagtseende borgere.

PreCare-projektet har derfor afprøvet fem forskellige "devices": Tre forskellige typer af wearables (Garmin Smart Watch, Oura Ring, BioBeat-Plaster), kamerateknologi til scanning af finger samt app til indsamling af stemmedata.

Derudover har PreCare-projektet i to forskellige pilotprojekter haft fokus på muligheden for udvikling af beslutningsstøtteværktøjer både til det sundhedsfaglige personale og borgerne mhp. at kunne identificere forværringer ved brug af kunstig intelligens (AI).

I analyserne er det undersøgt, om AI-modeller kan forudsige triagering af indsendte målinger på baggrund af færre indsamlede værdier i målingerne eller om flere data kan give en bedre prædiktionsdag forinden en forværring.

Slutteligt har formålet også været at undersøge, hvorvidt det er muligt at opstille modeller for hvor hyppigt forskellige borgere har behov for at måle sig.

Garmin Smart Watch og Oura ring

Fra december 2021 til foråret 2022 blev der gennemført et pilotprojekt, hvor data blev opsamlet og videresendt fra Garmin Smart Watch. De opsamlede data gav gode forhåbninger om, at værdier som respirationsfrekvens kunne have en prædiktiv værdi for kommende akutte forværringer, men herudover var erfaringerne, at uret var stort, remmen generende og, at lyset fra skærmen holdt deltagere vågne om natten.

Ovenpå erfaringerne med smarturet blev der igangsat et pilotprojekt i januar 2023, hvis formål var at undersøge om et device i form af en Oura smartring kunne erstatte eller supplere den egenmonitorering, der foregik i PreCareKlinikken for borgere med KOL, herunder om "aktivitet" og "søvn-mønstre" kunne inddrages som forklarende variable ift. borgerens tilstand/forværringer.

Analyser af Oura-data viste, at der var for stor ubalance i datasættet til, at en model kunne trænes til at skelne tilstrækkeligt mellem prædiktionsmålene. Modellerne viste også en høj falsk positiv procent, hvilket i en driftssituation vil besværliggøre arbejdet for de sygeplejersker, der skal tage stilling til borgernes triagerede målinger. Vurderingen i pilotprojektet var, at mere data (flere igangsættelser eller forværringer i datasættet) formentlig vil resultere i bedre modellering, men det kan ikke vides med sikkerhed.

Der har endvidere vist sig en del begrænsninger i anvendelsen af Oura-ringen som alternativ eller supplement til de manuelle målinger, borgerne foretager. Oura-ringen forudsætter som Garmin Smart Watch løbende opladning, og at ringen passer godt på borgerens finger.

Det tilgængelige data blev bearbejdet af det software, der fulgte med ringen og kunne vanskeligt anvendes direkte af sygeplejerskerne i forbindelse med vurdering af borgerens tilstand. Det vil forudsætte en integration af data til de løsninger, som sygeplejerskerne benytter sig af i den daglige monitoreringer af borgernes tilstand.

Oura-ringen har dog vist sig anvendelig i forbindelse med en borgers udredning for søvnapnø. Ringen har således kunne anvendes i forbindelse med søvnmønstre. Den vil evt. også kunne anvendes i forbindelse med øget fokus på aktivitet hos borgeren.

BioBeat-plaster og KOL-app

I februar 2023 blev der igangsat et pilotprojekt med anvendelse af digitale "plastre" i hjemmesygeplejen i Odsherred Kommune.

Formålet med pilotprojektet var at undersøge om hyppige indlæggelser kan forebygges for ældre borgere i Odsherred Kommune ved at udføre manuelle målinger eller foretage målinger gennem digital device (BioBeat-plaster) på udvalgte borgere efter udskrivning fra sygehus. Plastret målte løbende bl.a. temperatur, respirationsfrekvens, iltmætning i blodet, puls og blodtryk ved anvendelse af reflekterende fotoplethysmografi (PPG), der er en ikke-invasiv optisk metode.

Plastret er godkendt som medicinsk udstyr af FDA i USA. Data har i forbindelse med afprøvningen alene været anvendt til at kontrollere, om plastret var tilsluttet korrekt, og om der var forbindelse til den tablet hvor app'en, der indsamler data, er installeret. Borgerens tilstand er vurderet på baggrund af de manuelt foretagne målinger af det sundhedsfaglige personale i hjemme- og sygeplejen i Odsherred Kommune de dage, plastret var påsat borgeren.

Overordnet tyder den mere tætte opfølgning på borgerens centrale helbredsverdier lige efter udskrivning fra sygehus, at forværringer, der kan følge efter udskrivning for borgere med hyppige indlæggelser, kan identificeres hurtigere, og at en indsats kan blive iværksat hurtigere og dermed i højere grad virke proaktivt i stedet reaktivt.

Borgerne oplever generelt en tryghed ved, at der bliver fulgt tættere op på deres tilstand efter udskrivning fra sygehus, og kun en enkelt borger oplever gener fra skarpe kanter på plastret.

Muligheden for at anvende fotoplethysmografi (PPG) er også blevet udnyttet i et pilotprojekt, der påbegyndte udviklingen af en app i juni 2023. Teknologien måler ændringer i blodvolumen i væv, der bl.a. kan anvendes til at estimere hjertefrekvens, variation i hjertefrekvens, blodgennemstrømning og andre hæmodynamiske parametre. KOL-app'en blev i første gang udviklet til måling af iltmætning og puls. Den har været afprøvet på godt 15 borgere, der både gennemførte de almindelige egenmålinger med pulsoximeter og på samme tid anvendte KOL-app'en til at måle iltmætning og puls.

Målingen foregår ved, at borgeren placerer sin finger på tablettens kamera. Når målingerne skal indsamles fra kameraet, er der brug for, at fingeren holdes på kameraet indtil, der er indsamlet nok informationer til, at målingerne bliver stabile.

Den første version af den udviklede KOL-app har vist, at der er en systematisk, men heterogen bias i målinger af kamerateknologi sammenlignet med den normale måling, borgerne foretager med pulsoximeteret.

Det har vist sig at være vanskeligt at anvende KOL-appen for borgere, der har krogede fingre eller er farveblinde, idet anvisninger for, hvorledes fingeren skal placeres korrekt på kamera, er vist med forskellige farveangivelser.

Der er på nuværende tidspunkt også en udfordring ift., hvor lang tid, at borgeren skal holde sin finger på KOL-app'en sammenlignet med, hvor lang tid det tager at sætte pulsoximetret på en finger og sende/aflæse værdier for puls og iltmætning.

Stemmegenkendelse

I december 2022 blev der igangsat et pilotforsøg med henblik på at undersøge om stemmedata fra borgere med KOL kan erstatte eller supplere borgernes manuelle målinger særligt ift. borgere med begrænsede digitale kompetencer.

Indsamlingen af lyddata er foregået ved, at der er udviklet en simpel app, som er blevet installeret på en tablet. Borgeren og evt. pårørende som kontrolgruppe, har læst den samme tekst op så ofte som muligt, dog højst én gang dagligt.

Der er indsamlet ca. 3.000 lydfiler, hvoraf 1.855 af dem er blevet transskriberet. Forskellen mellem det totale antal og transskriberet lydfiler skyldes fejloptagelser. En gennemgang af kvaliteten af lyd-filerne viser, at kvaliteten af lyddataene varierer mellem deltagerne, hvor nogle indsendte filer er påvirket af forstyrrende baggrundsstøj, såsom radiolyd eller artefaktstøj fra elektrisk interferens. Disse udgør dog kun en mindre del af datasættet. Artefaktstøjen er vurderet til ikke at have effekt på den overordnede kvalitet.

Analyser af kvaliteten af rådata viser, at den overordnede kvalitet generelt er god. De eksempler hvor fejlen er markant, er grundet meget støj i baggrunden typisk fra en radio. Et næste skridt ville være at se på lyddatas evne til at prædiktere forværringer, men der har været et meget begrænset antal forværringer. Det vurderes, at der er for få data vedr. forværringer til det videre arbejde. Der vil således fremadrettet være behov for indsamling af flere data over en længere periode.

AI-analyser: KOL-algoritme og målerutiner

I foråret 2023 blev der foretaget analyser af mulighederne for at forbedre den eksisterende KOL-algoritme, således at den bedre kan bruges til at forudse, hvornår borgernes tilstand forværres og undersøge om det er muligt at bruge færre eller andre inputvariable end dem, der anvendes i den nuværende KOL-algoritme.

Det kan generelt konkluderes, at det er muligt at forbedre baseline KOL-algoritmen. Der ses dog en tendens til, at en optimering af de falske negative og sande positive værdier medfører en forværring med flere falske positive. En idriftsættelse af en anden model forudsætter en drøftelse af balancen med det sundhedsfaglige personale, idet modellen vil være et dagligt værktøj for personalet og skal give mening for dem.

Analyserne har også vist, at der ikke er en stor forskel på performance af de logistiske regressio-ner versus mere komplekse maskinlæringsmodeller. Det betyder, at hvis det ønskes, at modellen skal være transparent, vil den logistiske regression være en fin løsning. Dog kan der altid anvendes Explainable AI for at håndtere black-box problemet med maskinlæringsmodellerne.

I efteråret 2023 blev der igangsat analyser af borgernes målerutiner.

Formålet var bl.a. at undersøge det reelle målebehov, og komme med et databaseret estimat af den optimale målefrekvens, der resulterer i mindst mulig intervention i borgerens hverdag uden tab af information, samt at undersøge om borgerne kan segmenteres i flere grupper med hver deres optimale måleregime.

Det har ikke været muligt at sige noget om den optimale målefrekvens for borgerne, da datakvaliteten generelt ikke er tilfredsstillende særligt ift. lungefunktionsmålingerne. Hvis datakvaliteten generelt forbedres, vurderes det, at der er et stort potentiale for at udvikle en maskinlæringsmodel, der er i stand til at prædiktere forværringer hos borgere flere dage i forvejen. Derved kan målefrekvensen potentielt også sættes ned, således at borgere ikke behøver at udføre egenmålinger dagligt, men at borgeren derimod promptes for flere målinger såfremt maskinlæringsmodellen returnerer en forhøjet sandsynlighed for en forværring inden for de nærmeste dage.

Analyserne har også vist den prædiktive kraft i tre tilstandsspørgsmål, der indsamles som supplement til egenmålingerne. Det peger på et potentiale for at anvende disse til den daglige tilstandsevaluering, som kan danne grundlag for, at borgeren promptes for mere omfattende egenmålinger. Tilstandsspørgsmålene er mindre intervenerende end de nuværende egenmålinger, og den prædiktive kraft kan muligvis tilmed forbedres, hvis der inkluderes flere spørgsmål i den daglige tilstandsevaluering.

2.1 Opmærksomhedspunkter

På tværs af de syv forskellige pilotforsøg kan uddrages opmærksomhedspunkter inden for forskellige områder om bl.a. borgernes oplevelse og interaktion med devices/apps samt data, datakvalitet og hensyn, man skal være opmærksom på i udviklingen af algoritmer.

Erfaringerne skal ses i lyset af, at nogle af de afprøvede devices ikke direkte er udviklet til det formål, som devicen er anvendt til i pilotprojektet. Eksempelvis har Gamin Smart Watch og Oura-ring en anden typisk målgruppe end ældre kroniske borgere. Det samme gælder BioBeat-plastret, som er udviklet til hospitalsafdelinger og ikke hospital-at-home, som tilnærmelsesvis er afprøvet i et af pilotprojekterne.

Pilotprojekterne skal derfor også mere ses som feasibility-studier – altså om det praktisk kan lade sig gøre at indsamle og bearbejde data, og om det giver mening for borgerne og for det sundhedsfaglige personale – end som effektstudier af afprøvning af den enkelte device.

Borgernes praktiske erfaringer med device

Borgerne er meget opmærksomme på størrelse og det materiale en bærbar device er lavet af. Smart Watches opfattes særligt af kvinder som meget stort og klodset. Mens kvinderne omvendt synes bedre om at bære en Smart (Oura) ring.

Det forholder sig i grad høj modsat for mænd, der primært har fravalgt at deltage i Oura-pilotprojektet. Det har bl.a. skyldtes, at de ikke ønsker at bære en ring, men også at de har været bekymret for, at de kom til at hænge fast med ringen, når de arbejder fx på et værksted. Nogle opfattede også ringen som for bred.

Synlige wearable kan også sende forskellige signaler, som mennesker med kronisk sygdom ikke ønsker at blive identificeret med. Synlige wearables kan både signalere, at man er optaget af sit helbred (Biohacker), men omvendt også, at man er kronisk syg og har behov for løbende at monitorere sin helbredstilstand.

Derudover oplever nogle borgere at få forstyrret nattesøvn, hvis wearable lyser op om natten eller sender notifikationer.

Materialet på wearable kan også medvirke til, at wearable ikke sidder godt fast, og dermed foretager nogle ustabile målinger/upålidelige målinger, der afviger fra de faktiske værdier målt med fx et pulsoximeter, og således ikke er anvendelige i en prædiktionsammenhæng.

Opladning af device har også betydning. Det drejer sig både om den tid, det tager at oplade den pågældende device, og hvor ofte device skal oplades. Oura-ring blev oplevet som lettere og hurtigere at oplade end Garmin Smart Watch. Derudover holdt opladningen længere.

Teknik og digitale kompetencer

Brugen af digitale devices med tilhørende apps kan være udfordrende ift. borgernes digitale kompetencer. Det stiller krav til uddannelse af borgerne ved ibrugtagning af device, men undersøgelser i PreCareKlinikken peger også på, at det er muligt at oplære borgerne i brugen af digitale devices. Nogle borgere har også adgang til hjælp fra børn eller børnebørn.

Der har generelt været en interesse fra borgerne i at følge med i data, men borgerne har været udfordret af, om frontenden har været på dansk eller engelsk og hvor kompleks præsentation af deres egenmålinger har været.

Borgernes egenmålinger i PreCareKlinikken bliver fulgt op af en dialog mellem sygeplejerske og borger, hvis målingerne er røde og dermed afviger fra borgerens habitualværdier. I pilotprojekterne er borgerne blevet gjort opmærksom på, at de også skulle foretage deres almindelige målinger, og sygeplejerskerne alene ville forholde sig til disse målinger. Det har dog vist sig, at borgere kan blive bekymrede over data, der viser abnorme værdier, selvom de almindelige egenmålinger ikke har vist samme udsving.

Fremtidige projekter skal derfor have meget klare beskrivelser af, hvilke parametre borgerne får udstillet, og hvad de betyder herunder gode danske oversættelser. Herudover skal der laves et brugerinterface, der kun udstiller sikre og validerede data i letforståeligt sprog.

Datakvalitet, herunder dækningsgrad og kompletthed af data

Pilotprojekternes gennemførelse som feasibility-studier på mindre grupper af deltagere i kortere perioder har haft betydning for mængden af indsamlede data og de hændelser, som det ønskes at kunne prædiktere.

Dækningsgraden af data har således været udfordret af antallet af deltagende borgere på mellem 10-25 borgere og den begrænsede tidshorisont på 2-5 måneder, som studierne er gennemført på.

Målet for udvikling af prædiktioner har haft fokus på forudsigelse af forværringer. Da PreCare-projektet i sig selv har fokuseret på at imødegå akutte indlæggelser i forbindelse med forværringer, er der blevet gjort en aktiv indsats for netop at nedbringe antallet af forværringer, og dermed også den

hændelse, som algoritmerne skal forudsige. Antallet af data om forværringer har således været begrænset og har betydet et ubalanceret datasæt.

Ubalancer i datasættet har været en tilbagevendende problematik i forbindelse med prædiktion af forværringer i de forskellige pilotprojekter.

Devicen indsamler kun data, når deviceen er opladet. Manglende opladning af device efterlader således huller i data. Det gælder tilsvarende devices, der har strøm men bliver slukket ved et uheld af borgere. Det gjorde sig gældende i flere tilfælde i forbindelse med BioBeat-plastret, som har en let tilgængelig adgang til tænd-sluk-knappen.

En anden problemstilling, der relaterer sig til opladningen er, at deltagerne ofte glemmer at tage device på efter opladning. Det efterlader større huller i datasættet, hvilket bl.a. ses i afprøvningen af Oura-ring. Et ufuldstændigt datasæt gør det vanskeligt at anvende data til prædiktionsmodeller af fx forværringer.

På tværs af pilotprojekterne har andre forhold også afstedkommet huller i dataindsamlingen eller datasættet. BioBeat-plastret forudsætter, at der maksimalt er 8 meter til den enhed, som via en app på en tablet uploader data. I et driftsset-up vil det betyde, at den monitoreringsansvarlige vil opleve det i en realtidsrapportering. Hvorvidt alle data på et senere tidspunkt, når borger igen er inden for 8 meter til app og tablet, bliver uploadet er uklart, da det afhænger af, hvor meget data BioBeat-plastret kan rumme. Hvis datamængden er større end hvad BioBeat-plastret kan rumme, slettes de ældste data.

Analysen af borgernes egenmålinger har understreget behovet for straks-validering af indtastede og målte værdier, således at målingerne bliver mere valide og antallet af forkerte triageringer mindskes.

AI-modeller

I forbindelse med de gennemførte AI-analyser, hvor fokus har været på at optimere udfaldet af så lave falske negative værdier og så høje sande positive værdier som muligt, ses en tendens til, at en sådan optimering medfører flere sande negative og flere falske positive.

Der har således vist sig et trade off mellem at få så mange korrekte igangsættelser af borgere og få flere borgere, der i dag ikke laver røde målinger, men som vil lave røde målinger, der ikke giver anledning til igangsættelse af behandling, dvs. falsk positive. Det betyder flere opkald til borgerne om deres målinger for at være sikker på deres tilstand ikke er kritisk og øget bekymring hos borgere for en potentiel forværring. Omvendt får en større andel af borgere med røde målinger igangsat en behandling end i dag.

Dette kompromis er noget, der skal tages stilling til, hvis en endelig model skal idriftsættes. En drøftelse af balancen bør ske med det sundhedsfaglige personale, idet modellen vil være et dagligt værktøj for personalet og skal give mening for dem.

Analyser af lungefunktionsmålingerne har vist en stor prædiktiv kraft, men samtidig er de også årsag til mange falske positive værdier. Det skyldes, at der er stor varians i borgernes egenmålte lungefunktion. Dette kan mindskes gennem anvendelse af andre spirometre uden propel, men også gennem øget fokus på at lære borgerne at udføre lungefunktionsmålingen korrekt eller eventuelt at gentage en måling, såfremt den afviger meget fra et normalområde eller den sidst udførte måling.

Analyserne har endvidere understreget den prædiktive kraft i de tre tilstandsspørgsmål på KOL-området: Øget åndenød, tiltagende hoste, farvet sput. Det peger på et potentiale for at anvende disse til den daglige tilstandsevaluering, som kan danne grundlag for, at borgeren promptes for mere omfattende egenmålinger.

Tilstandsspørgsmålene er mindre intervenerende end de nuværende egenmålinger, og den prædiktive kraft kan muligvis tilmed forbedres, hvis der inkluderes flere spørgsmål i den daglige tilstandsevaluering.

Etiske overvejelser i forbindelse med driftliggørelse

Indsamling af data ved brug af digitale devices rejser forskellige etiske overvejelser.

Det skal være muligt at sikre borgernes privatliv og datasikkerhed for disse data ift. at forhindre uautoriseret adgang og misbrug af data.

Der er behov for i forbindelse med brug af samtykke i forbindelse med pilotprojekter at sikre at borgerne er fuldt ud informeret om, hvordan deres sundhedsdata vil blive anvendt, herunder også omfanget samt konsekvenser af overvågningen af borgerens vitale værdier.

Der kan endvidere være tekniske begrænsninger eller fejl i devices, der kan føre til unøjagtige data. Anvendes unøjagtige data kan det fx føre til forkerte medicinske beslutninger, der kan påvirke borgerens velbefindende.

Det er også vigtigt at finde en balance mellem fordelene ved monitorering af borgernes egenmålinger eller målinger fra wearables og borgerens oplevelse af indtrængen i borgeren liv. Det er desuden vigtigt, at muligheden for overvågning af vitale værdier ikke afhænger af borgerens socioøkonomiske status, men er en lige adgang for alle.

Det er behov for opmærksomhed på, at kontinuerlig overvågning kan medvirke til en stigmatisering af borgere. Det skal derfor være muligt at trække sit samtykke til kontinuert måling tilbage.

Borgere, der bærer wearables, forventer at blive monitoreret, og det kan muligvis skabe en falsk tryghed, hvis der er risiko for, at data i perioder ikke opsamles af forskellige årsager. For at der ikke skabes falsk tryghed ved fremtidige afprøvninger, er det vigtigt, at det sikres, at borgerne altid indsender de normale tilstandsmålinger, som klinikken leverer sine ydelser ud fra.

Målerutiner

Målerutiner kan også indebære brug af forskellige typer af devices på forskellige tilstandstidspunkter, og dermed en segmentering af borgerne. Hvis det er muligt at finde wearables med mere valide data kunne disse fx anvendes i "fredstid"/i den habituelle tilstand uden forværringer, som så kunne anvendes som baseline for vurderingen af begyndende forværringer. Ved tegn på at borgerens tilstand forværres kan aftales, om der er behov for at overgå til det manuelle/ordinære medicinske udstyr, herunder antallet af målinger fx daglige målinger i en periode. Denne tilgang kan dog ikke anvendes i alle tilfælde. For nogle borgere sker forværringerne meget hurtigt og dermed er ovenstående tilgang ikke brugbar.

GARMIN SMART WATCH



3. Pilotforsøg med Garmin Smart Watches

I slutningen af december 2021 blev der igangsat et pilotprojekt i PreCareKlinikken, hvor formålet var at undersøge om smart watches kan erstatte eller supplere den device baserede egenmonitorering, der foregår i PreCareKlinikken i dag for borgere med KOL.

Formålet med projektet var at sammenligne data fra borgernes egenmålinger med de indsamlede værdier fra Garmin Smart Watches for at belyse grundlaget for på sigt at udvikle en model til prædiktionsmodel på data indsamlet uden interaktion fra borger muliggør en udvidelse af PreCareKlinikkens målgruppe til også at inkludere borgere, der ikke selv kan varetage indsendelse af egenmålinger, herunder demente eller svagtseende borgere.

Udover sammenligning af data blev borgernes umiddelbare erfaringer med at indgå i pilotprojektet med et smart ur også indhentet for ved senere lignende projekter at kunne adressere mulige barrierer hos målgruppen.

3.1 Metode

I projektet blev 26 KOL-borgere udvalgt fra en liste over de borgere, der måler sig hyppigst til dagligt, og som alle havde oplevet en tilstandsforværring inden for det sidste halve år op til projektet. Derudover var der ingen andre inklusionskriterier end lyst til at deltage i projektet, da eneste interaktion med uret er opladning og evt. opdatering, men med mulighed for teknisk assistance fra projektsygeplejerske i PreCareKlinikken. Borgerne blev opfordret til at bære uret mest muligt og lade det op hver 2.-3. dag.

Garmin uret var af typen Garmin Vivoactive 4S. Uret kan bl.a. måle iltmætning, hjerterytme, respirationsfrekvens, gang og søvn. Uret blev udleveret i uge 50 og 51 i 2021 og blev indsamlet i uge 12 i 2022. 2 borgere døde under dataindsamlingen, en få dage efter påsætning af uret og en halvvejs i projektet. Førstnævnte indgik ikke i analyserne. Indsamling af data mislykkedes fra et enkelt ur. Analyser på data er derfor foretaget på i alt 24 ure.

Iltmætning, hjerterytme og respirationsfrekvens blev løbende opsamlet af uret hvert minut i alle døgnets timer, under forudsætning af at uret sad korrekt på armen. Data blev via Bluetooth overført til tablet og uploadet til en sikker database. Det vil sige, at hvert ur kunne indsende op til 1.440 målinger pr. døgn. De 24 ure, der indsendte data, indsendte data på gennemsnitligt 62 dage (fra 35 til 75 dage) i alt. I samme periode indsendte borgere, hvis ur indgik i analyserne gennemsnitligt over 71 tilstandsmålinger (fra 31 til 93 målinger) via app til PreCareKlinikken i alt.

Ved afslutning af dataindsamlingen udleverede RKC-sygeplejersken fra PreCareKlinikken i forbindelse med afhentning af urene et kort spørgeskema til borgerne. Formålet med spørgeskemaet var at indhente borgernes dagligdagserfaringer med brug af et forholdsvist stort ur udviklet med henblik på fitness og aktivitets tracking. Forståelse for målgruppens evt. udfordringer og gener ved brug af Smart Watches vil kunne hjælpe i planlægning og udførsel af lignende projekter i fremtiden.

24 borgere fik udleveret spørgeskemaet, jf. boks 3.1. 19 borgere besvarede spørgeskemaet, hvilket svarer til en svarprocent på ca. 80 pct.

Boks 3.1. Spørgeskema udleveret i forbindelse med Garmin Smart Watch pilotforsøg

Spørgeskemaet bestod af 4 spørgsmål med 2-3 svarmuligheder og tilhørende kommentarfelt, samt 2 åbne spørgsmål:

- 1) Er uret nemt at betjene? (ja/nej/delvist) + kommentar
- 2) Er det nemt at oplade uret? (ja/nej/delvist) + kommentar
- 3) Har du haft gener ved at have uret på? (ja/nej) + hvis ja, hvilke:
- 4) Har du interesseret dig for de data uret viser? (ja/nej) +kommentar
- 5) På hvilke tidspunkter har du taget uret af? F.eks. i bad, når du sover, været dårlig, etc.
- 6) Er der andet, du gerne vil dele om din oplevelse med at have uret på?

Kilde: PreCareKlinikken

3.2 Analyse af data fra tilstandsmålinger

Der blev gennemført præliminære analyser af data. Dels blev der forsøgt at udføre machine learning med afsæt i Garmin data, og dels blev der udført klassisk t-test, mixed model og logistisk regression på data.

Da antallet af påbegyndte exacerbationsbehandlinger var ret lavt, blev der fokuseret på triageringen i PreCareKlinikken på baggrund af KOL-algoritmen for tilstandsmålinger (grøn, gul, rød). Et event blev identificeret som minimum to på hinanden røde eller tre gule målinger, før det blev karakteriseret som et event.

Med afsæt i tidspunktet for dette event summeredes data op på de to foregående døgn for at undersøge, om der var noget særligt at observere ud fra data ved dagene op til et event. Sammenligningerne blev foretaget ved at identificere sammenhængende grønne perioder som sammenligningsgrundlag for de perioder, der ledte op til et event.

Metoden betød, at der ved events og kontroller udelukkende kunne inkluderes perioder i analyserne, hvis der var grønne målinger. Eksempel; hvis der var en rød event lørdag, ville dette kun kunne bruges i analyserne, hvis der forelå grønne målinger både torsdag og fredag. Tidsperioden ville blive ekskluderet, hvis en grøn måling ikke forelå enten torsdag eller fredag.

3.3 Resultater fra tilstandsmålingerne

De præliminære analyser viste, at iltmætningen fremstod som den mindst betydende faktor sammenlignet med puls og respirationsfrekvens.

Der var indikation på, at respirationsfrekvensen i tiden op til tre på hinanden gule målinger eller en rød måling kan have særlig stor betydning forud for et event.

Pilotforsøgets korte tidshorisont gjorde, at der ikke var data nok til at kunne træne og teste en machine learning baseret prædiktionsmodel, men de præliminære analyser giver indikation af, at respirationsfrekvens især kan have en prædiktiv værdi ift. gule og røde målinger, men der er også tegn på, at puls kan have prædiktiv værdi.

I de indledende machine learning analyser pegede modellen ydermere på, at Garmin Stress indikation kan have betragtelig betydning, når eventet blev defineret som 'igangsat behandling'. Stress er i Garmins algoritme afledt af Heart Rate Variability (HRV), altså hvor stor varians der er mellem de enkelte hjerteslag.

Årsagen til iltmætningens lave forklaringsgrad kan muligvis forklares ved, at den iltmætning som borgeren måler med PreCareKlinikkens udstyr og den iltmætning, som blev målt med Garmin Smart Watch, ikke havde en systematisk sammenhæng, dvs. ikke var perfekt korreleret.

Det er forventningen, at devices godkendt til medicinsk behandling og dermed valideret ift. udstyrets præcision i målingerne, der indsamler værdier om respirationsfrekvens, iltmætning og puls og HRV, vil kunne have en væsentlig større prædiktiv værdi.

3.4 Evaluering på baggrund af borgererfaringer

Alle svarpersoner har angivet, at der var få oplevede problemer, der relaterer sig til teknik. Alle fandt det nemt at oplade uret, og de fleste fandt det nemt eller delvis nemt at betjene det.

Der peges dog på udfordringer med at finde tilbage til udgangsskærm, hvis man er kommet ind i andre funktioner. Desuden kan det både være svært at huske at få uret opladt og huske at tage det på efter opladning. Derudover er der tekniske gener ved lys fra skærm. En gene, der fik et par brugere til at undlade at have ur på om natten.

4 svarpersoner (21 pct.) angiver, at de har haft gener i forbindelse med brug af uret, heraf 3 der relaterer sig til udslæt, og en der oplevede gene pga. andre helbredsproblemer ved håndled. Der beskrives også gene ved urets størrelse og remmens materiale. En enkelt angiver at tage uret af, da vedkommende blev dårlig, jf. tabel 3.1.

11 svarpersoner (58 pct.) har angivet, at de har interesseret sig for urets data. Der er angivet, at de kræver forklaring og er svære at forstå. En enkelt angiver, at det er mere overskueligt på pc'en. Det nævnes af et par personer, at 'skridt' havde deres interesse.

19 af de 24 borgere, der var i live ved afslutning af dataindsamling, har besvaret spørgeskemaet om deres oplevelser med uret.

Alle personer fandt det let at oplade, men det tyder på, at der kan være behov for hjælp til at huske både at få uret opladet og få uret på efter opladning.

Borgerne i PreCareKlinikken var motiverede for projektet og interesserede i data, men det kræver mere forklaring og indføring, da data var svære at forstå.

Derudover var der helbredsmæssige udfordringer, der skal tages højde for ved lignende projekter, da en større andel oplever gener ved brug af uret, der relaterer sig til materiale, størrelse og funktion af lys, der kan forstyrre søvnkvalitet i en gruppe, hvor søvn i forvejen er udfordret af deres sygdom.

Tabel 3.1. Besvarelser fordelt på svarmuligheder

Spørgsmål	Ja	Nej	Delvist	Kommentarer
Nemt at betjene	9	3	7	Svært at finde tilbage, kunne kun se klokken
Nemt at oplade	19			Når det huskes
Gener ved ur	4	15		Udslæt, eksem, fordybning i huden, lidt for stort, svært ved gamle skader og psoriasis, remmen generer – varm, lys fra ur generer om natten
Interesse for data fra ur*	11	7		Data på pc mest overskuelige, havde brug for forklaring, svære at forstå, skridt havde interesse
Taget uret af				I bad, ved opladning og efter pga. glemsel, om natten, om morgenen, når dårlig
Andet om oplevelse				Får tryghed, håber det gavner senere

Note: *en svarperson har ikke svaret. Grå felter – ingen svarmulighed

Kilde: PreCareKlinikken

3.5 Perspektivering

Det er vurderingen, at afprøvning af mere nøjagtige digitale devices, vil kunne forbedre prædiktio-
nen og dermed anvendeligheden ift. borgere, der ikke selv kan lave daglige målinger. Man kunne
således med fordel afprøve andre typer af wearables og urremme, der sikrer at uret sidder bedre
fast på armleddet.

OURA RING



4. Pilotforsøg med Oura-ringe

Primo januar 2023 blev der igangsat et projekt med anvendelse af Oura-ringe til borgere med KOL. Deltagende havde mulighed for at aflevere Oura-ringen tilbage medio juni eller beholde den på til medio september 2023.

Formålet med pilotprojektet var at undersøge om en Oura-ring kan erstatte eller supplere den egenmonitorering, der foregår i PreCareKlinikken i dag for borgere med KOL. Derudover blev det testet om "aktivitet" og "søvn mønstre" var forklarende variable ift. borgerens tilstand/forværringer.

4.1 Metode

I projektet blev 10 KOL-borgere udvalgt fra en liste over de borgere, der måler sig hyppigst til dagligt, og der alle havde oplevet en tilstandsforværring inden for det sidste halve år op til projektets start. Derudover var der ingen andre inklusionskriterier end lyst til at deltage i projektet, da eneste interaktion med ringen er opladning og evt. opdatering, men med mulighed for teknisk assistance fra projektsygeplejerske i PreCareKlinikken. Borgerne blev opfordret til at bære ringen mest muligt og lade den op hver 2. dag.

Der var med baggrund i erfaringer fra smart watch afprøvningen tilknyttet en sygeplejerske i PreCareKlinikken, der stod til rådighed for borgerne ved tekniske spørgsmål eller ved nysgerrighed om data. Sygeplejersken fulgte op på borgerne ved manglende data og minimum en gang om måneden for at høre om evt. vanskeligheder.

Oura-ringen måler iltmætning, søvn mønstre, fysisk aktivitet, puls, pulsvariation, respirationsfrekvens og temperatur. Iltmætningen måles kun om natten.

Oura-ringen måler puls, bevægelse og kropstemperatur, mens man sover for at bestemme, hvornår personen er faldet i søvn, og hvilket søvnstadium personen er i. Dette er muligt, fordi hvert af de 4 søvnstadier (vågen, lys søvn, REM og dyb søvn) er karakteriseret ved forskellige biosignaler.

Der er forskellige størrelser ring, så den skal passe til den enkelte borger/deltager. Efter borgers accept af deltagelse i projektet, har projektsygeplejersken testet ringstørrelse, og den korrekte størrelse ring er bestilt til borger.

Oura har et application programming interface (API), der kan bruges til at trække data fra flere brugere. Som standard har API-applikationer en grænse på ti brugere. Hvis der skal inkluderes flere end 10 brugere skal API-applikation registreres og godkendes af Oura. Der er ingen grænse, når først en ansøgning er godkendt. Derudover skal Oura-brugere give samtykke til at dele hver datatype, som en API-applikation har adgang til. Det er årsagen til, at der alene indgår 10 borgere i pilotforsøget.

Udover data til sammenligning med egenmålinger indhentes også borgernes umiddelbare erfaringer med at anvende en Oura-ring, for ved senere lignende projekter at kunne adressere mulige barrierer hos målgruppen.

4.2 Data fra API

Det samlede datagrundlag består af 1.339 observationer (dage) opsamlet i perioden 30. januar 2023 til 23. juni 2023. Projektet var planlagt at afslutte primo juni, hvor 3 borgere valgte at aflevere ringene igen, men det blev aftalt, at projektet fortsat kunne samle data fra borgere, der valgte at

beholde ringen herefter frem til medio september 2023, dog uden opfølgning fra projektsygeplejersken.

Alle borgere har dage med non-wear tid, dvs. perioder hvor ringen ikke har været på og derfor ikke har opsamlet data. Antallet af timer med non-wear varierer fra 43 til 230 timer, mens den gennemsnitlige non-wear tid ligger mellem 1,3 og 5,7 timer på de dage, hvor borgeren har haft ringen af. 8 borgere har mellem 0 og 2 dage helt uden data i hele perioden. Kun 2 borgere har mere end 2 dage uden data (henholdsvis 13 og 68 dage).

Datakvaliteten af den opsamlede data må antages at være påvirket af non-wear tid - særligt for nogle variable. Eksempelvis de variable, der indikerer noget om borgerens aktivitetsniveau såsom antal skridt og antal forbrændte kalorier. Andre variable, der er baseret på et gennemsnit af de målte værdier over en længere periode, bliver formentlig mindre påvirket af non-wear tid.

Non-wear tid kan dog også resultere i, at der mangler målinger for nogle variable. Eksempelvis hvis borgeren har taget ringen af, eller den er løbet tør for strøm om natten, opsamles data ikke, hvilket medfører, at de variable, der er udledt af data fra søvn-perioden ikke findes.

En vurdering af datakvaliteten viser, at stor del af observationerne ikke er komplette. 904 ud af 1.339 observationer har manglende værdier i mindst én variabel, svarende til 67,5 pct. Variablen aktivitetsscore mangler bl.a. for en del af observationerne (64 pct.).

Oura-ringene opsamler en lang række variable, nogle variable er rådata, og nogle er kun tilgængelige gennem API'en som afledte værdier, der dannes på baggrund af andre variable. Værdier for temperatur svarer eksempelvis ikke til rådata, idet der foretages en form for bearbejdning af de målte værdier fra Ouras side. Det vanskeliggør sammenligning af Oura-ring data med borgerens manuelle egenmålinger.

Selvom Oura-ringene måler kontinuert, er de fleste af de opsamlede variable ligeledes kun tilgængelige gennem API'en som gennemsnitsmål over den enkelte dag.

Enkelte variable forventes at korrelere med borgernes egenmålinger, herunder egenmålinger af lungefunktion med respirationsfrekvens målt af Oura-ringene, samt borgernes egenmåling af puls og hjertefrekvens målt af Oura-ringene.

Borgernes egen lungefunktionsmåling ser ud til at korrelere negativt med den gennemsnitlige respirationsfrekvens målt af Oura-ringene med en korrelationskoefficient på -0,60. Borgernes egne pulsmålinger ser endvidere ud til at korrelere positivt med den gennemsnitlige hjertefrekvens målt af Oura-ringene med en korrelationskoefficient på 0,59.

Den egenmålte puls er generelt højere end puls-målingerne fra Oura-ringene, hvilket er forventeligt, da borgernes foretager deres egenmålinger i vågen tilstand, mens Oura-ringene måler puls i løbet af natten.

Sammenligningen af disse data fra ringene med borgernes egenmålinger viser således en korrelation, men ikke en perfekt korrelation. Dog er det vigtigt at huske, at borgernes egenmålinger ikke kan anses som en guld-standard, da disse også kan være fejlbetonede, jf. bilag 2 og 6.

4.3 Analyse af data fra tilstandsmålinger

For at vurdere om data fra Oura-ringene har nogen prædiktativ kraft, er det nødvendigt at træne og evaluere en model på datasættet. Prædiktionsmodellen skal forudsige sandsynligheden for en igangsættelse af en borger baseret på data fra Oura-ringen.

Udfordringen ved at træne sådan en model på det nuværende datasæt er, at der er meget få igangsatte behandlinger, dvs. den variable som skal prædikteres. Der er i alt 13 igangsættelser hos de 10 borgere i perioden for dataindsamlingen, hvilket besværliggør modellering af data. Dog kan modellering muligvis give en indikation af den prædiktive værdi i det tilgængelige data.

Det er initialt set på seks forskellige maskinlæringsmodeller:

- Logistic Regression
- Decision Tree (første test) / KNN (anden test)
- Random Forest
- Gradient Boosted Trees
- Support Vector Machine (SVM)
- Neural Network

For hver af maskinlæringsmodellerne, er modellen trænet på 60 pct. af datasættet og evalueret på de resterende 40 pct. Modellerne er først trænet alene på inputvariablene fra Oura-ringene, og efterfølgende er de trænet på input variablene fra Oura-ringene beriget med de tre tilstandsspørgsmål¹, som borgerne svarer på i forbindelse med de ordinære tilstandsmålinger i PreCareKlinikken mhp. at evaluere effekten af at medtage spørgsmålene.

Tabel 4.1. Maskinlæringsmodeller ved anvendelse af inputvariablene fra Oura-ringdatasættet

Model	Målinger	Præcision, pct.	Sensitivitet, pct.
Logistic Regression	Ikke-igangsættelse	99	77
	Igangsættelse	0	0
Decision Tree	Ikke-igangsættelse	99	76
	Igangsættelse	0	0
Random Forest	Ikke-igangsættelse	99	88
	Igangsættelse	0	0
Gradient Boosted Trees	Ikke-igangsættelse	99	99
	Igangsættelse	0	0
Support Vector Machine	Ikke-igangsættelse	99	75
	Igangsættelse	2	50
Neural Network	Ikke-igangsættelse	99	99
	Igangsættelse	0	0

¹ De tre tilstandsspørgsmål borgerne svarer på efter måling af lungefunktion, puls, iltmætning og temperatur er: 1. Har dit opspyt ændret farve fra i går til at blive mere grønt eller grønligt? 2. Har du mere åndenød i dag end i går? 3. Hoster du mere i dag end i går?

Test datasættet indeholder kun 2 igangsættelser. Alle seks modeltyper har svært ved at identificere de to igangsættelser. Support Vector Machine-modellen (SVM) identificerer én af de to igangsættelser, mens de resterende modeller ikke identificerer nogen igangsættelser korrekt. Til gengæld klassificerer modellerne mellem 2 og 52 ikke-igangsættelser som igangsættelser.

Evalueringen af de seks modeltyper viser altså, at modellerne har svært ved at lære de mønstre i træningsdata, der er karakteristiske for igangsættelserne, og derved ikke er i stand til at klassificere dem korrekt i testdatasættet, jf. tabel 4.1.

Tabel 4.2 viser de seks udvalgte modeller, hvor de tre tilstandsspørgsmål medtages som prædiktorer. Det fremgår, at de fleste af modellerne har svært ved korrekt at klassificere de to igangsættelser i testdatasættet. SVM-modellen klassificerer den ene af de to igangsættelser korrekt, Logistic regression modellen klassificerer begge igangsættelser korrekt, mens de resterende modeller alle klassificerer de to igangsættelser som ikke-igangsættelser. Modellerne klassificerer ligeledes mellem 2 og 42 ikke-igangsættelser som igangsættelser, jf. bilag 2.

Tabel 4.2. Maskinlæringsmodeller ved anvendelse af inputvariablene fra Oura-ringdatasættet samt de tre tilstandsspørgsmål

Model	Målinger	Præcision, pct.	Sensitivitet, pct.
Logistic Regression	Ikke-igangsættelse	100	90
	Igangsættelse	5	100
KNN	Ikke-igangsættelse	99	92
	Igangsættelse	0	0
Random Forest	Ikke-igangsættelse	100	100
	Igangsættelse	0	0
Gradient Boosted Trees	Ikke-igangsættelse	100	96
	Igangsættelse	0	0
SVM	Ikke-igangsættelse	100	94
	Igangsættelse	4	50
Neural Network	Ikke-igangsættelse	100	98
	Igangsættelse	0	0

Modellernes performance ser ud til at stige lidt, når de tre spørgsmål medtages som prædiktorer sammenlignet med modellerne uden de tre tilstandsspørgsmål. Da formålet med modelleringen er at kunne finde de igangsættelser borgerne har, udvælges Logistic Regression Modellen til yderligere eksperimenter.

Logistisk regressionsmodeller er ligeledes mulige at forklare sammenlignet med de andre modeltyper, der bærer mere præg af at være black-box modeller.

4.3.1 Logistisk regressionsmodellering

Da antallet af igangsættelser er meget lavt i forhold til det totale antal af observationer, er anvendt *Leave-One-Out krydsvalidering*² til evaluering af modellens generelle evne til at korrekt at klassificere testdatapunkterne³.

Modellen klassificerer 5 af de 13 igangsættelser korrekt (svarende til en sensitivitet på 38 pct.), mens 1.196 ud af de 1.314 ikke-igangsættelser klassificeres korrekt (svarende til en sensitivitet på 91 pct.). Det betyder, at når modellen klassificerer en måling som en igangsættelse, vil kun 5 pct. af disse være faktiske igangsættelser, mens 99 pct. af de klassificerede ikke-igangsættelser er faktiske ikke-igangsættelser. Overordnet har modellen en nøjagtighed på 91 pct.

Tabel 4.3. Performance for Logistisk Regression fordelt på ikke-igangsættelser og igangsættelser

Klasse	Præcision, pct.	Sensitivitet, pct.	Datapunkter
Ikke-igangsæt	99	91	1314 (99 pct.)
Igangsæt	4	38	13 (1 pct.)

På trods af at minoritetsgruppen er upsamlet i træningsdata, tyder resultaterne på, at modellen har svært ved at lære de mønstre, der er karakteristiske for de målinger, som fører til en igangsættelse, hvilket er forventeligt, når ubalancen i datasættet er så signifikant.

Det er endvidere undersøgt om modellen er bedre til at klassificere de triagerede målinger i henholdsvis rød, gul og grøn fremfor igangsættelser. Target for klassifikationsmodellen defineres som røde målinger eller grønne og gule målinger, og modelperformance evalueres igen med *Leave-One-Out krydsvaliderings-metoden*, hvor minoritetsklassen er upsamlet, således at træningsdata indeholder lige mange af de henholdsvis grønne/gule målinger og røde målinger.

Modellen klassificerer 18 af de 32 røde målinger korrekt (svarende til en sensitivitet på 56 pct.), mens 1.048 ud af de 1.222 grønne/gule målinger klassificeres korrekt (svarende til en sensitivitet på 81 pct.). Det betyder, at når modellen klassificerer en måling som rød, vil kun 7 pct. af disse være faktiske røde målinger, mens 99 pct. af de klassificerede grønne/gule målinger er faktiske grønne/gule målinger, jf. tabel 4.4. Overordnet har modellen en nøjagtighed på 80 pct.

Tabel 4.4. Performance for Logistisk Regression fordelt på grøn/gule målinger og røde målinger (binær model)

Klasse	Præcision, pct.	Sensitivitet, pct.	Datapunkter
Grøn/gul måling	99	81	1295 (98 pct.)
Rød måling	7	56	32 (2 pct.)

Resultaterne tyder på, at modellen har svært ved at lære de mønstre, der er karakteristiske for de røde målinger selvom, der er foretaget en upsampling af minoritetsgruppen i træningsdata.

² [sklearn.model_selection.LeaveOneOut — scikit-learn 1.3.1 documentation](#)

³ De få igangsættelser resulterer i meget stor ubalance i datasættet, hvorfor der anvendes SMOTE upsampling af minoritetsklassen (igangsættelser) i træningsdatasættet, således træningsdatasættet indeholder lige mange igangsættelser, som ikke-igangsættelser.

Slutteligt er det undersøgt om modellen vil kunne klassificere triageringen bedre, hvis der tages udgangspunkt i et træningsdata som indeholder lige mange grønne, gule og røde målinger igen ved brug af Leave-One-Out krydsvaliderings-metoden.

Modellen klassificerer 18 af de 32 røde målinger korrekt (svarende til en sensitivitet på 56 pct.), 95 ud af de 173 gule målinger klassificeres korrekt (svarende til en sensitivitet på 55 pct.), mens 757 ud af de 1.122 grønne målinger klassificeres korrekt (svarende til en sensitivitet på 67 pct.).

Det betyder, at når modellen klassificerer en måling som rød, vil 11 pct. af disse være faktiske røde målinger, og hvis modellen klassificerer en måling som gul, vil 28 pct. af disse målinger være faktiske gule målinger, mens 93 pct. af de klassificerede grønne målinger er faktiske grønne målinger, jf. tabel 4.5. Overordnet har modellen en nøjagtighed på 66 pct.

Tabel 4.5. Performance for Logistisk Regression fordelt på grøn, gule målinger og røde målinger (multiklasse model)

Klasse	Præcision, pct.	Sensitivitet, pct.	Datapunkter
Grøn måling	93	67	1.122 (85 pct.)
Gul måling	28	55	173 (13 pct.)
Rød måling	11	56	32 (2 pct.)

Evalueringen af multiklassemodellen tyder ligeledes på, at modellen har svært ved at skelne mellem de mønstre, der er karakteristiske for de grønne, gule og røde målinger henholdsvis, og mange af målingerne klassificeres forkert.

4.3.2 Opsummering

De indledende eksperimenter med forskellige modeltyper viser tydeligt udfordringen ved det ubalancerede datasæt med kun 13 igangsatte behandlinger ud af i alt 1.327 målinger, da alle seks testede modeltyper har svært ved at klassificere målingerne korrekt.

Den logistiske regressionsmodel så ud til at performe en smule bedre end de resterende modeltyper, og sammenholdt med at denne modeltype har en stor grad af forklarbarhed, blev den valgt til yderligere eksperimenter. Dog skal det understreges, at de seks modeltyper blev evalueret på et testdatasæt med kun to igangsættelser, og derfor er den opnåede performance ikke nødvendigvis repræsenterende for et datasæt med flere/andre målinger, der har været ledsaget af en igangsættelse.

Leave-One-Out krydsvalidering blev anvendt for at give et mere generelt estimat af den logistiske regressionsmodels performance på alle målinger i datasættet, og tre forskellige mål for prædiktionen blev testet: igangsættelse/ikke-igangsættelse, rød måling/grøn eller gul måling, samt grøn måling/gul måling/rød måling.

Uanset hvilke prædiktionsmål modellen har, indikerer resultaterne, at der er for stor ubalance i datasættet til, at en model kan trænes til at skelne tilstrækkeligt mellem prædiktionsmålene. Udover at modellerne har svært ved at klassificere igangsættelser eller røde målinger korrekt (mange falske negative), klassificerer modellerne desuden mange ikke-igangsættelser eller grønne målinger som igangsættelser eller gule/røde målinger, og falsk positiv procenten er derfor høj. En stor andel af falske positive besværliggør arbejdet for de sygeplejersker, der skal tage stilling til borgernes triagerede målinger, hvorfor det er ligeså vigtigt, at en model ikke resulterer i for mange falske positive, som at modellen er i stand til at finde alle igangsættelser/forværringer eller den korrekte triagering af målingerne.

På baggrund af den afprøvede modellering, vurderes det, at det nuværende datagrundlag ikke er tilstrækkeligt til at træne en model med acceptabel performance, grundet den store ubalance i datasættet og de få igangsættelser. Mere data (flere igangsættelser eller forværringer i datasættet) vil formentlig resultere i bedre modellering, men dette kan ikke siges med sikkerhed.

En alternativ tilgang til modellering af målingerne kunne være at bruge outlier detection metoder, da der er så få igangsættelser/forværringer i datasættet. Tidsserie-modellering er også en mulighed, da Oura-ringene opsamler data fortløbende, hvorfor det potentielt er muligt at finde mønstre i udviklingen af målepunkterne, der kan forudsige en forværring hos borgeren.

4.4 Observationer

Navnlig mænd har fravalgt at deltage i projektet. Det kan bl.a. skyldes, at de ikke ønsker at bære en ring, men også at de er bekymret for, at de kan hænge fast med ringen, når de arbejder fx på et værksted. Ringen bliver af nogen opfattet som for bred, og kan både signalere, at man er optaget af sit helbred og omvendt, at man er kronisk syg.

De 10 borgere vurderes at være blandt de mest ressourcestærke af de inkluderede borgere i PreCareKlinikken. De fleste tilhører den gruppe af PreCare-borgere, der er teknisk dygtigst eller har fået hjælp af børn eller børnebørn.

Opladning af device har betydning. Oura-ring er lettere og hurtigere at oplade end Garmin Smart Watch. Derudover holder opladningen længere.

Borgerne har været interesseret i at følge med i data, men har været udfordret af, at frontenden var på engelsk med sammenvejede score.

Undervejs i pilotforsøget er der løbende blevet fulgt op på, om der er kommet målinger fra de deltagende borgere. Der har vist sig følgende udfordringer:

- Der mangler data i længere perioder for enkelte borgere
- Ringen tages ofte af op til 4,5 timer ad gangen, idet borgerne glemmer at tage ringen på igen efter opladning
- Ringen sidder løst på fingeren, hvilket gør at der ikke registreres data
- Data opdateres først, når Oura-app'en på tablet åbnes
- Hvis der går for lang tid imellem at app'en åbnes, går data tabt, da ringen i sig selv kun kan indeholde en vis mængde data, hvorefter det bliver slettet, hvis ikke det kan overføres via åbning af app'en
- Ring tages af ved sengetid og dermed mangler der søvndata.

På baggrund af drøftelser undervejs med det sundhedsfaglige personale og på baggrund af analysearbejde er der endvidere noteret følgende overvejelser ift. anvendelsen af Oura-ringe til monitorering af centrale værdier:

- Hvis borgere bærer ringen, forventer de at blive monitoreret, og føler derfor muligvis en falsk tryghed, hvis der er risiko for at data i perioder ikke opsamles (af forskellige årsager).
- Kvaliteten af det data der opsamles af ringen er betinget af, at ringen sidder korrekt på fingeren af borgeren. Flere ting kan medføre ændringer i tykkelsen af fingeren, herunder Prednisolon behandling, hvorved ringen ikke længere passer optimalt, eller endda ikke kan anvendes af borgeren i en periode
- Data opsamles via en iPad, og overføres kun hvis der er forbindelse til WIFI. Dette kan medføre at noget data går tabt, da hukommelsen i Oura-ringene er begrænset

- Data opsamles og opbevares i Ouras API, men her er lagerpladsen også begrænset, hvorfor noget data tilsyneladende slettes efter en periode. Dette gør sig gældende for puls-data, der opsamles i løbet af dagen (Beats per minute (BPM)), hvor data kun er tilgængeligt for de foregående to dage. Dette betyder, at man skal overføre data fra Ouras API til Databanken minimum hver anden dag, for at sikre at data ikke går tabt
- Data udstilles i bearbejdet form i API, hvilket begrænser analysearbejdet og mulighed for at vurdere målinger som evt. erstatning for egenmålinger.

4.5 Evaluering på baggrund af borgererfaringer

Ved afslutning af dataindsamlingen udleverede projektsygeplejersken i forbindelse med planlagt afslutning og afhentning af ringene, et kort spørgeskema til borgerne med titlen "Evalueringsskema i forbindelse med afprøvning af Oura-ring". Formålet med spørgeskemaet var at indhente borgerne erfaringer med brug af et tætsiddende device i form af en forholdsvis stor ring i dagligdagen. Forståelse for målgruppens evt. glæder, udfordringer og gener ved brug af ringen vil, sammen med lignende erfaringer med pilotprojektet med Garmin smarture, kunne hjælpe i planlægning og udførsel af lignende projekter i fremtiden.

Boks 4.1. Spørgeskema udleveret i forbindelse med Oura-ring pilotforsøg

Spørgeskemaet bestod af 5 spørgsmål med 2-3 svarmuligheder og tilhørende kommentarfelt, samt 3 åbne spørgsmål:

- 1) Er ringen nemt at betjene? (ja/nej/delvist) + kommentar
- 2) Er det nemt at oplade ringen t? (ja/nej/delvist) + kommentar
- 3) Har du haft gener ved at have ringen på? (ja/nej) + hvis ja, hvilke:
- 4) På hvilke tidspunkter har du taget ringen af? F.eks. i bad, når du sover, røntgen/undersøgelser, været dårlig, osv. Kommentar
- 5) Har du interesseret dig for de data ringen viser? (ja/nej) + hvis ja: Har data fra ringen vist dig noget, du ikke vidste før?
- 6) Hvad har det betydet for dig, at app'en er på engelsk? Kommentar
- 7) Får data fra ringen dig til at gøre noget andet, end du plejer i forhold til dit helbred? (ja/nej) + hvis ja, hvad gør du anderledes
- 8) Vil du anbefale ringen? (ekstra spørgsmål)

Kilde: PreCareKlinikken

Ud af 10 borgere, der initialt fik påsat en ring, har alle svaret.

4.5.1 Resultat af evaluering - borgererfaring

Alle svarpersoner har angivet, at ringen var let at betjene herunder oplade. Borgerne har kun taget ringen af i forbindelse med opladning og få i forbindelse med personlig pleje samt ved røntgen/CT-scanninger. Der er få rapporterede gener med ringen i forbindelse med hævede fingre, som en person angiver relaterer sig til prednisolon behandling. 6 ud af de 10 fandt det svært, at app'en var på engelsk.

8 af de 10 borgere har interesseret sig for data fra ringen. Det er især data vedrørende søvn-mængde og søvnkvalitet, der har haft borgernes interesse, men også iltmætning og aktivitet er fulgt. En enkelt borger har på baggrund af søvndata været hos øre-, næse-, halslæge og blevet diagnosticeret med søvnapnø. Derudover angives ren nysgerrighed ved at 'holde øje' med de indkomne data. En enkelt angiver, at ringen får vedkommende til at gøre noget anderledes ved at 'optimere' og 'reflektere' over især søvn og aktivitetsdata.

I forbindelse med spørgsmål om interesse for data og det ekstra spørgsmål om borgeren vil anbefale ringen – angives 'Tryghed' af flere som en faktor. Der er en borger, der sammenligner ringen med smart watches, og pointerer at det er et 'stort plus ikke at blive forstyrret hele tiden'. En af de få borgere, der ikke har interesseret sig for data, angiver, at vedkommendes eget Apple Watch har langt flere funktioner herunder nødkald samt EKG og, at data fra ringen ikke er interessante at følge. 5 ud af 6 adspurgte vil anbefale ringen.

Tabel 4.6. Besvarelser fordelt på svarmuligheder

Spørgsmål	Ja	Nej	Delvist	Kommentarer
Nemt at betjene	10			Stort plus ikke forstyrrer hele tiden
Nemt at oplade	10			
Gener ved ring	1	9		Hævede fingre, sæbe nødvendig til aftagning
Ringen taget af ved				Ved opladning (et par angiver de har glemt at tage på efterfølgende), ved hævede fingre, personlig hygiejne og ved røntgen/CT-scanning
Interesse for data fra ring	8	2		Søvn som interessant, herunder: - overrasket over nøjagtighed - fokus på power naps, herunder virkning på parametre - opmærksomhed på dårlig søvn Tjekker søvn, SAT og aktivitet Bare nysgerrig 'holdt øje' IKKE interesse, da har apple-watch med flere data og funktioner
Svært ved engelsk app				6 borgere kommenterede at det var vanskeligt i forskellige grad
Gøre noget andet	1	9		Reflekterer over middagslure, søvn og aktivitet. Forsøger at optimere.
Ekstra: anbefale ringen	5	1		Skaber tryghed i at den måler hele tiden For at få overblik af – ikke leve efter

Note: Grå felter – ingen svarmulighed

Kilde: PreCareKlinikken

4.5.2 Diskussion og konklusion – borgererfaringer

Baseret på indsamlede borgererfaringer kan det konstateres, at der ingen overordnede problemer er med at bruge ringen i det daglige, og borgerne rapporterer stor nysgerrighed om data, hvoraf især parametre omkring søvn-mængde og søvnkvalitet har vækket borgernes interesse. Idet søvn er korreleret med helbred og livskvalitet og en kendt udfordring for KOL-patienter, kan muligheden for at følge søvn-mønstre ved forskellige søvnoptimerende tiltag, være relevant ved udredning og afhjælpning af borgers dårlige søvnkvalitet i klinikken. Da der ikke rapporteres gener ved brug af ringen om natten, som det var tilfældet med Garmin Smart Watch, giver ringen en god mulighed for brug i PreCareKlinikken, hvis de senere dataanalyser viser konsistens i søvn-data.

Da prednisolon er standardbehandling ved moderat og svær exacerbation, er det en vigtig erfaring, at en tætsiddende ring er problematisk ved hævede fingre. En kendt bivirkning af prednisolon er

netop ødem i varierende grad. Derfor kan ringen ikke anses for at være en god løsning som eneste monitoreringsdevice til indsamling af data under forværringer behandlet med prednisolon.

Antallet af observationer og de indsamlede svar på spørgeskemaet viser en fortsat motivation til at indgå i pilotprojekter hos borgerne i PreCareKlinikken. Borgerne er nysgerrige på data. En vigtig indsigt, der fremkom på læringsworkshop i klinikken var, at borgere kan blive bekymrede over data der udstilles i appen, når de viser abnorme værdier.

Der er også flere der angiver tryghed ved ringens kontinuerlige målinger, til trods for at borgerne var blevet oplyst om at det var sideløbende dataindsamling, og at der ikke ville blive monitoreret på data under projektet.

Fremtidige projekter skal have meget klare beskrivelser af, hvilke parametre borgerne får udstillet, og hvad de betyder herunder gode danske oversættelser, eller der skal laves et brugerinterface, der kun udstiller sikre og validerede data i letforståeligt sprog. For at der ikke skabes falsk tryghed ved fremtidige afprøvninger, er det vigtigt, at det sikres, at borgerne samtidigindsender de normale tilstandsmålinger, som klinikken leverer sine ydelser ud fra.

4.6 Perspektivering

Der har vist sig en del begrænsninger i anvendelsen af Oura-ringen som alternativ eller supplement til de manuelle målinger borgerne foretager. Oura-ringen forudsætter som Garmin Smart Watch løbende opladning, og at ringen passer godt på borgerens finger.

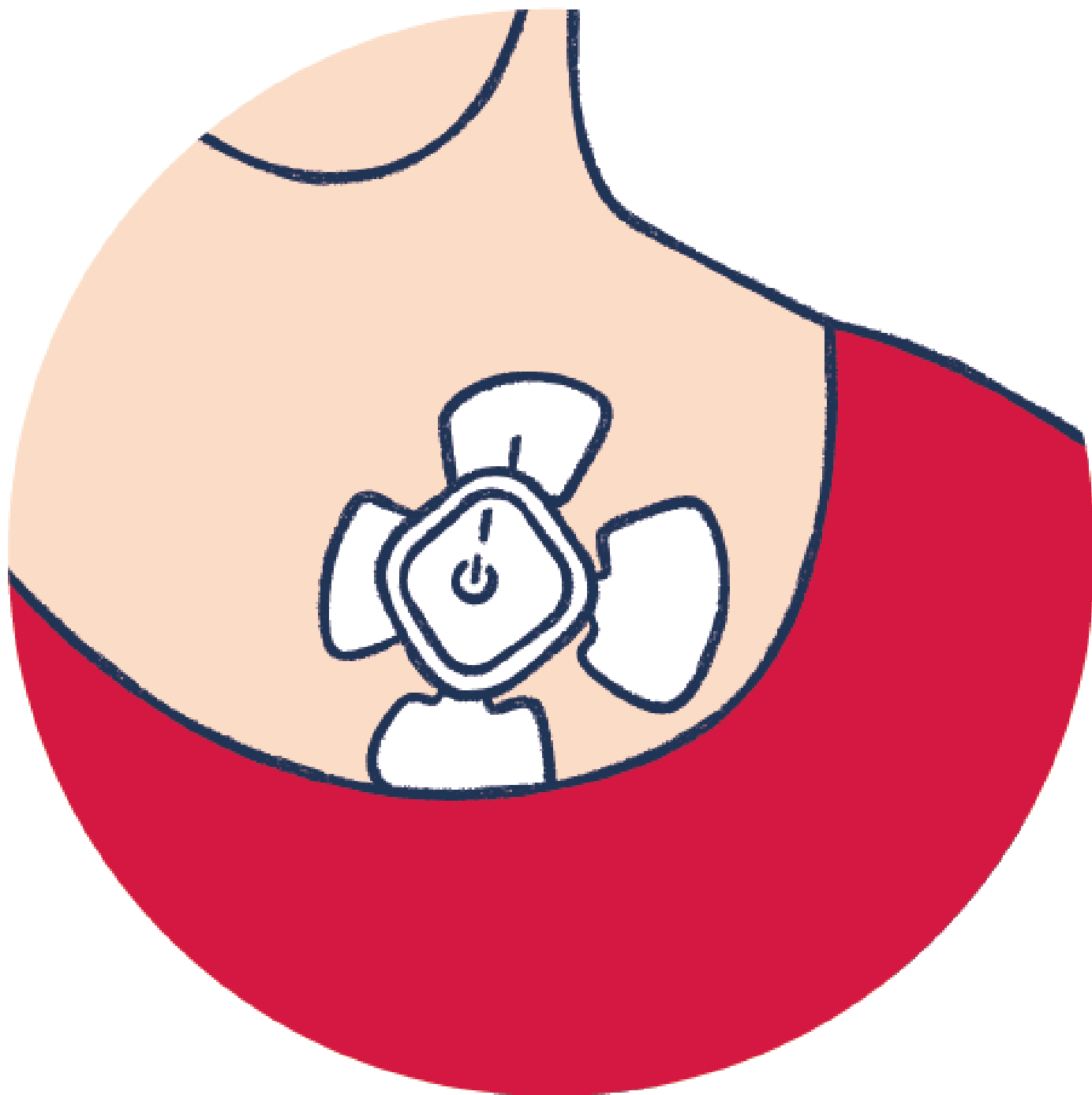
Den tilgængelige data er bearbejdet data, og dermed vanskelig direkte at anvende for sygeplejerskerne i forbindelse med vurdering af borgerens tilstand.

Oura-ringen har vist sig anvendelig i forbindelse med en borgers udredning for søvnapnø. Ringen har således kunne anvendes i forbindelse med søvnmønstre.

Kunne evt. anvendes i forbindelse med øget fokus på aktivitet.

Frontenden kan med fordel laves på dansk med udvalgte parametre fx i opsætningsmenuen, hvor brugeren peger på, hvilke parametre der ønskes fulgt. Også gerne med mulighed for at se, hvordan man ligger ift. opstillede mål og ros/nudging i form af "virtuelle kongekroner", når man overperformer.

BIOBEAT PLASTER



5. Pilotforsøg med BioBeat-plastre

I februar 2023 blev der igangsat et pilotprojekt med anvendelse af digitale "plastre" i hjemmesygeplejen i Odsherred Kommune.

Formålet med pilotprojektet er at undersøge om indlæggelser kan forebygges for ældre borgere i Odsherred Kommune ved at udføre manuelle målinger eller foretage målinger gennem digital device (BioBeat-plaster) på udvalgte borgere. Plasteret måler løbende bl.a. temperatur, respirationsfrekvens, iltmætning i blodet, puls og blodtryk.

I pilotforsøget skulle det også undersøges:

- Om en digitalisering af målingerne kan bidrage til en bedre arbejdsgang for personalet
- Om målinger tilnærmelsesvis kan opnås ved brug af et digitalt plaster, som borgeren anvender døgnet rundt frem for, at det er personalet, der skal foretage målinger en eller flere gange i døgnet
- Borgernes oplevelse af at have et plaster på. Hvilke erfaringer giver det ift. fremtidige muligheder for anvendelse af forskellige devices?
- Om en data-understøttet udvælgelse (algoritme) af borgere kan hjælpe med at prioritere, hvem der skal være omfattet af målinger.

5.1 Metode

Pilotforsøget blev gennemført på 10 borgere, der får kommunale hjemmesygepleje ydelser i eget hjem i Odsherred Kommune, og som har været indlagt på Holbæk Sygehus og blev vurderet i risiko for at blive indlagt igen.

Odsherred Kommunes fremskudte visitation på Holbæk Sygehus har udvalgt, hvilke borgere der er blevet tilbudt at deltage i pilotforsøget. Kriterierne for udvælgelse af borgerne har været:

- +70 år
- Ikke-misbruger
- Ikke demens sygdom
- Ikke udskrevet til palliation
- Har infektion eller været igennem stor operation.

Plastret er blevet påsat ved udskrivning fra sygehus efter samtykke fra borgeren. Plastret kan sidde på i op til 5 dage. Når plastret er blevet fjernet, er borgeren blevet stillet nogle spørgsmål om plastret. Svarene er medgået ved evalueringen af det digitale plaster.

Der er blevet udleveret en tablet til låns, så plastret via tabletten kan indsamle data digitalt. En forudsætning for at modtage data på tablet er, at den er opladet.

Borgeren har også fået målt temperatur, respirationsfrekvens, iltmætning, puls og blodtryk én gang dagligt af en sygeplejerske eller social- og sundhedsassistent fra Odsherred Kommune i de dage, plasteret har været påsat. Hjemmesygeplejen har taget stilling til borgerens helbredstilstand ud fra de målinger, som de har foretaget en gang dagligt.

De medarbejdere fra hjemmesygeplejen, der har foretaget målinger, er også blevet stillet spørgsmål om det digitale plaster, og data indsamlet fra plasteret. Svarene er også medgået i vurderingen af denne form for medicinsk udstyr.

Derudover skulle der i pilotforsøget udvikles en algoritme til segmentering af borgere med risiko for gentagne akutte indlæggelse på baggrund af kommunale data og MedCom-beskeder, og dermed fungere som et beslutningsstøtteværktøj for sygeplejerskerne i den fremskudte visitation på Holbæk Sygehus.

Det har dog efterfølgende efter en længere afklaringsperiode vist sig, at det vil tage op til 10 måneder at få adgang til MedCom-beskederne fra Center for Innovativ Medicinsk Teknologi (CIMT) i Region Hovedstaden. Pilotforsøget indeholder derfor ikke en udvikling af en algoritme.

Boks 5.1. Data fra MEDCOM-beskeder

Dataindholdet ift. MEDCOM-beskeder er i løbet af foråret blevet afklaret med CIMT. Der er tale om et omfattende dataudtræk af hensyn til den eksplorative natur data-science metoderne har. CIMT har fået fremsendt skærbilleder af, hvor oplysningerne indtastes i sundhedsplatformen (SP). Skærbillederne har CIMT anvendt til at identificere pågældende variable i SP.

Data hidrører fra 5 forskellige typer af MedCom-meddelelser:

- Indlæggelsesrapport
- Plejeforløbsplan
- Udskrivningsrapport
- Genoptræningsplan
- Melding om færdigbehandling

For hver type er der udfærdiget en liste over de nødvendige variable. Variabellisterne er gengivet i bilag 3.

Kilde: Koncern Digitalisering, Region Sjælland

5.2 Data

Plastret foretager målinger hvert 15. min. døgnet rundt i op til 5 dage.

Plastret er godkendt som medicinsk udstyr af FDA i USA.

Data har i forbindelse med anvendelsen af plastret alene været brugt til at følge op på, om plastret virkede, og der var forbindelse til den tablet, hvor app'en, der indsamler data, er installeret. Borgers tilstand er vurderet på baggrund af de manuelt foretagne målinger.

5.3 Opfølgning efter udskrivning fra sygehus

Biobeat-plastret – og den manuelle måling af centrale værdier – har været medvirkende til, at borgere med hyppige indlæggelser efter udskrivning fra sygehus hurtigere har fået identificeret en forværring af deres tilstand og dermed på et tidligere tidspunkt er påbegyndt en udredning og/eller behandling.

I boks 5.2 fremgår nogle eksempler på anvendelse af en mere proaktiv opfølgning efter udskrivning fra sygehus for borgere, der i forvejen modtager kommunale ydelser, og som har hyppige sygehusindlæggelser.

Boks 1.2. Eksempler på proaktivitet i forbindelse med brugen af BioBeat-plastret

Forskellige cases har vist at:

- BioBeat-plastret – og den tilhørende manuelle måling – har medvirket til, at en borger efter operation og udskrivning fra sygehus tidligt i et forværringsforløb kan indlægges på en temperatur over

habiltualværdien. Hvis hjemmesygeplejen ikke havde foretaget de tidlige målinger, ville der været gået længere tid, før der var blevet taget kontakt til udskrivende afdeling på Holbæk Sygehus.

- BioBeat-plastret – og den manuelle opfølgning – har bidraget til tidlig opsporing og rette behandling. Eksempelvis betød den tætte observation af borgers vitale værdier, at der efter udskrivning registreres stigende temperatur, lavt blodtryk og højere respirationsfrekvens, som fører til indlæggelse. Med den hurtige indlæggelse undgår borgeren, at der opstår yderligere forværring og svækkelse, som fx. exacerbation eller dehydrering med risiko for efterfølgende faldulykke.
- Proaktivitet og tæt samarbejde mellem kommunens hjemmesygepleje og kommunens akutfunktion understøtter opmærksomhed og observation af afvigelser i borgers målinger fra habiltualværdierne. Eksempelvis blev kommunens akutfunktion bedt om at tilse borger efter udskrivning på baggrund af opfølgende målinger og yderligere opmærksomhed på værdier fra hjemmesygeplejen. Borger indlægges herefter på baggrund af målinger foretaget af kommunens akutfunktion, der bekræfter de tidligere målinger.
- Potentielle forværringer kan forebygges ved at monitorere borgers vitale værdier og evt. forværringer eksempelvis for borgere, der inden for kort tid har gennemgået en eller flere større operationer
- BioBeat-plastret og den manuelle opfølgning har også bidraget til en hurtigere involvering af egen læge og dermed en tidligere indsats ift. forværring

Kilde: PreCareKlinikken/Odsherred Kommune.

I forbindelse med påsætning af de første to plastre har der været udfordringer med, at der ikke var adgang til data i perioder (som om device var slukket). Efter dialog med leverandør synes den eneste forklaring på afbrydelsen af device at være, at der er kommet en betonvæg/murstensvæg i vejen for signalet mellem plastret og tablet.

Flere borgere har givet udtryk for, at det har givet rigtig god mening med tæt monitorering. Det er i flere tilfælde ført til identifikation af en forværring tidligere end ellers og en efterfølgende indsats på sygehus.

Der har været et ønske fra kommunen om, at det er sygeplejersker i stedet for social- og sundhedsassistenterne, der varetager måling af de vitale værdier. Dette ønske er afstedkommet af, at ikke alle social- og sundhedsassistenterne har de fornødne kompetencer til at reagere på en tidlig opsporing af begyndende sygdom (TOBS)-scoring.

5.4 Evaluering på baggrund af erfaringerne i hjemmeplejen og den fremskudte visitation

Personalet i hjemmesygeplejen er blevet stillet spørgsmål om deres oplevelser med plastret og evt. muligheder for fremtidig anvendelse jf. boks 5.2.

Boks 5.2. Spørgsmål til personale i hjemmesygeplejen i Odsherred Kommune

Spørgeskemaet til hjemmesygeplejen bestod af 6 spørgsmål:

- 1) Hvad har du oplevet i forbindelse med plasteret?
- 2) Hvad tænker du at sådan et plaster kan bruges til i fremtiden?
- 3) Hvilken betydning kunne det have for dit arbejde i hverdagen?
- 4) Hvad er potentialet i, at der er adgang til nogle af borgerens centrale værdier?
- 5) Har du ideer til hvornår og hvordan, man skal følge op på kontinuerede målinger, som fx BioBeat-plasteret
- 6) Har du forslag til, hvad der kunne gøre det til en endnu bedre oplevelse at gå med sådan et plaster?

Generelt er tilbagemeldingerne fra medarbejderne fra visitationen og hjemmesygeplejen positive i forhold til deres oplevelser med plasteret og deres dialog med de borgere, der havde fået påsat plasteret i forsøgsperioden.

Teknik

Der er nævnt bekymring i forhold til, at plasteret i visse tilfælde har været ustabil i målinger, og at det ligeledes forekom, at tallet for målingen ikke fremstod på monitoren, men i stedet blev vist som en graf.

Sygeplejerskerne fremhæver behovet for stabilitet i plasterets målinger som en vigtig forudsætning for borgerens oplevelse og tryghed forbundet med anvendelse af plasteret til måling af vitale værdier fremfor de traditionelle kontaktformer og interaktion med hjemmesygeplejen i forbindelse med behov for målinger af tilstand.

Erfaringer med anvendelse på borger

I forhold til selve plasteret er der i enkelte tilfælde oplevet rødme og begyndende irritation på huden ved plasteret. Erfaringer under forsøget har vist, at det er vigtigt at affedte huden med sprit, inden plasteret sættes fast for således at sikre, at det sidder ordentligt fast på huden.

Behov for rammer for arbejdet

Kontinuerede målinger gør det muligt løbende at følge borgerens vitale værdier. Det rejser spørgsmål om, hvorvidt afvigelser fra habitualværdier og opgørelse af score for en tidlig observeret forandring fx TOBS/Early Warning Score (EWS)-lignede⁴ opgørelse dog uden måling/opgørelse af graden af bevidsthed, også løbende skal opgøres og handles på.

De fleste af borgerne, som har deltaget i pilotprojektet, har peget på, at plasteret giver en tryghed, idet der bliver holdt øje med borgerens vitale værdier, om end borgeren er blevet oplyst om, at hjemmesygeplejen alene vil reagere på de daglige manuelle målinger, som hjemmesygeplejen gennemfører. Forventningen fra borger og formentlig også pårørende vil muligvis være en løbende monitorering.

⁴ TOBS: Tidlig Opsporing af Begyndende Sygdom. EWS: Early Warning Score.

I PreCareKlinikken monitoreres borgernes målinger fast – dog gennemfører borgerne i langt de fleste tilfælde kun én måling om dagen. Målingerne triageres, så sygeplejerskerne nemt har et overblik over de målinger, i dette tilfælde røde målinger og tre på hinanden gule målinger, som sygeplejerskerne skal følge op på.

Der kunne på samme måde gives en triageringsfarve til forskellige niveauer af score (aggrerering af de vitale værdier samt evt. nyopstået store spring) og en handlingsalgoritme, som det kendes fra sygehus.

Der gennemføres på intensiv afdeling på sygehusene systematiske målinger af patienternes vitale værdier. Der anvendes bl.a. en EWS, som er en aggregeret score, der udregnes på baggrund af en række vitale værdier. I handlingsalgoritmen for EWS i Region Hovedstaden og Region Sjælland er fastlagt minimumsobservationsintervaller på baggrund af størrelsen af EWS-scoren, jf. tabel 5.1.

Tabel 5.1. Early Warning Scores (EWS) handlingsalgoritme (kortfattet)

EWS score	Minimums-observationsinterval	Handlingsalgoritme	
0-1	Hver 12. time (+/- 1 time)	Fortsæt scoring hver 12. time. Scoringshyppigheden kan øges.	Ved en enkelparameter-score på 3 (nyopstået eller uden dokumenteret handlingsplan) skal lægen straks tilse patienten eller der kan kaldes Mobil Akut Team
2	Hver 6. time (+/- 30 min.)	Plejepersonale udfører ABCDE	
3-5	Hver 4. time	Sygeplejerske udfører ABCDE. Sygeplejerske vurderer om lægen skal orienteres eller tilkaldes.	
6	Hver 4. time	Sygeplejerske udfører ABCDE. Sygeplejerske tilkalder vagthavende læge, som lægger behandlingsplan.	
7-8	Hver time	Sygeplejerske udfører ABCDE. Sygeplejersken tilkalder vagthavende læge, som lægger behandlingsplan. Planen konfereres med bagvagten.	
9 eller der-over	Hver 30. minut	Sygeplejerske udfører ABCDE. Sygeplejerske tilkalder vagthavende læge, som straks tilser patient og konfererer med bagvagten eller Mobil Akut Team/anæstesiologisk assistance. Der lægges plan for behandling.	

Note: ABCDE: Systematisk og prioriteret fremgangsmåde til vurdering og behandling af patienter med akut sygdom eller skade. A: Airway (luftveje), B: Breathing (respiration), C: Cirkulation (kredsløb), D: Disability (bevidsthedsniveau), E: Exposure (ekstern påvirkning og/eller feber).

Kilde: Region Sjællands Dokumentportal: <http://dok.regionsjaelland.dk/view.aspx?DokID=510759>

En automatisk beregning af scoren ud fra en handlingsalgoritme og en forskellig markering/intensitet i farven af scoren kunne give et enkelt overblik ift. opfølgning. Opfølgningen kunne evt. suppleres af en fast gennemgang ved vagtskifte af borgere i højeste risiko.

Det fremhæves fra hjemmesygeplejefagligt hold, at der ved anvendelse af plasteret skal foreligge retningslinjer for, hvor hyppigt personalet skal kontrollere monitoreringerne herunder også en retningslinje for hvilke målinger, der afstedkommer henholdsvis kontakt til borger, fysisk besøg hos borger fra sygeplejerske, social- og sundhedsassistent eller lægekontakt.

Graden af akutniveau herunder værdien af de foretagne målinger vil være afgørende for behov for telefonisk og/eller fysisk besøg og skal ses i tæt sammenhæng med behovet for retningslinjer.

Der påpeges endvidere behov for en retningslinje, der beskriver hvor hurtigt personalet skal kunne være hos borger, jf. behov retningslinje for målinger og tilknyttede aktiviteter i forhold til hjemmesygepleje og lægekontakt.

Klinisk

Sygeplejerskerne pointerer, at adgangen til borgeres data over distancen giver flere muligheder. Bl.a. kan data bruges som senere reference i forhold til udvikling af borgers tilstand, fx kan borgers tilstand følges på afstand, mens borger er under behandling med antibiotika.

Plasteret vil med fordel kunne anvendes ved ustabile borgere, der har behov for hyppig vurdering af tilstand og vitale værdier. Plasteret vil ligeledes med fordel kunne anvendes ved borgere, der skal have udført målinger af hjemmeblodtryk. I begge tilfælde fremhæves det ressourcebesparende perspektiv i forhold til både kommunale hjemmeplejeressourcer og i forhold til forstyrrelser af borgerens hverdag.

Samarbejde og arbejdsmiljø

Det fremhæves som tryghedsskabende faktor for medarbejderne, at der er flere kolleger, der følger med i borgerens vitale værdier, og at sygeplejersker og social- og sundhedsassistenterne samarbejder om dette.

Overvejelser i forhold til plasterets anvendelse

Sygeplejerskerne er af den opfattelse, at det overordnet set er godt at have mulighed for at anvende plasteret til at foretage målinger, da det er mindre indgribende for borgeren af få foretaget målingerne via et plaster, end det vil være ved almindelig målepraksis med besøg og forstyrrelser i borgerens hverdag.

Undtagelser kan være i forbindelse med:

- Terminale borgere med meget kort levetid. I disse tilfælde bør plasteret ikke anvendes
- Borgere med udprægede kognitive svækkelser. I disse tilfælde kan det være svært at sikre at borgeren forstår, hvorfor de får plasteret på. Der kan således opstå et behov for at vurdere, hvornår der er tilstrækkeligt samtykke. Borgeren kan en ene dag måske sige ja til at lægen skal have målinger og ikke huske dette, når plasteret skal sættes på dagen efter
- Borgere der vil have tendens til at hive plasteret af. I disse tilfælde bør plasteret ikke anvendes.

Bæredygtighedsprincippet i forhold til anvendelse af plasteret

I PreCare-projektet er der opmærksomhed på bæredygtighedstiltag. I pilotforsøget med plastrene er disse blevet kasseret efter endt brug. Ved permanent indførelse af anvendelse af plasteret til måling af borgeres vitale værdier, bør det overvejes hvilke muligheder, der er for genanvendelse. I disse overvejelser bør indgå beregning af ressourceforbrug forbundet med bl.a. afhentning af materiellet, afspritning og transport forbundet med genanvendelsen.

Opmærksomhedspunkter om betydning af nærhed og geografisk afstand i forhold til anvendelse af plasteret

Som nævnt ovenfor var der i en række tilfælde udfald i målingerne, hvilket har givet anledning til følgende input fra sygeplejerskerne:

- Hvis borgerne overvåges, skal der tages stilling til, hvor hurtigt man skal kunne nå ud til borgeren, hvis målinger viser en akutsituation

- Hvis borger er længere væk end 8 meter fra tablet, gemmer plaster data, og uploader data, når borger er tæt på tablet igen. Der bør derfor være tydelighed i aftalen med borgeren om betydningen af hele tiden at være tæt på tablet i de tilfælde, hvor det er nødvendigt med konstant overvågning af borgerens vitale værdier.

Den fremskudte visitation på Holbæk Sygehus er ligeledes blevet stillet spørgsmål om plastret og dets anvendelighed, jf. boks 5.3.

Boks 5.3. Spørgsmål til personale i den fremskudte visitation på Holbæk Sygehus

Spørgeskemaet til den fremskudte visitation bestod af 6 spørgsmål.

- 1) Hvem skal have plasteret på?
- 2) Hvilke overvejelser gjorde du dig i den forbindelse?
- 3) Oplevede du, at der var borgere, der havde forbehold for at få plasteret på?
- 4) Hvad handlede disse forbehold om?
- 5) Ville det for nogle være bedre at plasteret først blev sat på efterfølgende?
- 6) Hvilke andre overvejelser og input har du til anvendelse af BioBeat-plaster på borgere?

Det fremhæves af den fremskudte visitation på Holbæk Sygehus, at alle borgere, der fik påsat plasteret i forbindelse med pilotforsøget, var interesserede i at hjælpe med at afprøve plasteret som muligt fremtidigt måleudstyr.

Plasterets egnethed til målgrupper

Den fremskudte visitation påpeger, at det er vigtigt, at borgerne, der skal have påsat plasteret, er klare og relevante, så de er i stand til at modtage og forstå information om brug af plasteret herunder, hvordan det virker, og hvad der kommer til at foregå i den periode, borgeren bærer plasteret.

Erfaringerne fra Holbæk Sygehus er, at en del af de borgere, der blev udskrevet i forsøgsperioden ikke kunne deltage, da de var i delir eller havde hukommelsesudfordringer.

Det optimale ville være, at plasteret kan bruges på alle borgere, der skal udskrives og ikke kun begrænses til en bestemt aldersgruppe eller borgergruppe.

Som en del af løsningen i forhold til så bred anvendelse som muligt bør det overvejes, hvornår det er det bedste tidspunkt at sætte plasteret på borgeren.

Udskrivelsessituationen

Den fremskudte visitation beskriver, hvorledes det ofte er tilfældet, når en borger udskrives fra sygehuset, at der for borgeren ofte er forbundet en del stress med selve udskrivelsessituationen. Borgeren er ofte hverken helt rask eller udhvilet efter indlæggelsen og skal forholde sig til skiftet fra sygehus til fx eget hjem og kan være påvirket af den usikkerhed, der følger med ikke længere at være på sygehuset.

På baggrund af erfaringer og observationer i forsøgsperioden har den fremskudte visitation gjort sig overvejelser om følgende to løsningsmodeller i forhold til at opnå en så bred anvendelse af plasteret som muligt.

- At plasteret påsættes af en medarbejder på sygehuset. Erfaringer fra pilotforsøget tyder på, at det kan komplicere udskrivelsessituationen yderligere og være svært at overskue for den

enkelte borger, at det er en ekstern person, som fx den fremskudte visitation fra kommunen, der skal give informationen om plasteret og sætte det på samtidig med, at borgeren får information om medicin og lignende i forbindelse med udskrivelsen.

Hvis det er en medarbejder fra sygehuset, der forestår information og påsætning af plasteret, er der en indbygget fleksibilitet i forhold til pludselige ændringer i planen for borgeren fx, hvis udskrivelsen udskydes i sidste øjeblik.

- At plasteret påsættes efter udskrivelsen fra sygehuset af hjemmesygeplejen, når borgeren er installeret i eget hjem. Der vil sandsynligvis være mindre uro og stress forbundet med situationen, hvor borgeren skal have information om, hvordan plasteret virker, og hvad det betyder, at gå med plasteret.

For at sikre at gevinsten ved proaktiv og kontinuert måling af borgerens vitale værdier efter indlæggelse, opfordrer den fremskudte visitation til at hente inspiration fra kommuner, der arbejder med såkaldte "tidsvinduer" i forhold til, hvor hurtigt efter udskrivelsen en borger skal modtage besøg af hjemmesygeplejen.

I Kalundborg Kommune er tidsvinduet på 6 timer. Det vil sige, at den ny udskrevne borger modtager besøg af hjemmesygeplejen inden for 6 timer efter udskrivelsen, hvor det i de fleste tilfælde må formodes, at risikoen for voldsom forværring af tilstand i løbet af de 6 timer vil være forholdsvis begrænset. Det vil ofte være aftenvagten fra hjemmesygeplejen, der vil tilse borgeren og kunne påsætte plasteret og give den nødvendige information, inden borgeren skal gå til ro for natten.

5.5 Evaluering på baggrund af erfaringerne fra borgerne

Borgere, der har deltaget i pilotforsøget og afprøvet plastret, har efterfølgende svaret på spørgsmål om, hvordan de har oplevet plastret, hvad de synes om plastret og om det kunne være relevant at tage i brug i hjemmesygeplejen i Odsherred Kommune, jf. boks 5.4.

Boks 5.4. Spørgsmål til borgere

Spørgsmål til borgerne bestod af 10 spørgsmål.

- 1) Hvordan oplevede du plasteret?
- 2) Hvordan føltes det at have plasteret på?
- 3) Oplevede du gener i forbindelse med at have plasteret på? (Udslæt, kløe, andet?)
- 4) Kunne plasteret sidde fast efter bad eller gled det af?
- 5) Har du forslag til, hvad der kunne gøre det til en endnu bedre oplevelse at gå med sådan et plaster?
- 6) Kunne du tænke dig at plasteret blev taget i brug i hjemmeplejen?
- 7) Hvad tænker du om at plasteret giver mulighed for at sundhedspersonale på afstand kan følge med i din tilstand?
- 8) Hvilke muligheder kunne det give dig som borger, hvis du gik med sådan et plaster?
- 9) Kunne det betyde noget for dine pårørende, at du har et plaster på?
- 10) Har du andre forslag i forbindelse med plasteret?

Deltagerne fremhæver oplevelsen af tryghed som den primære gevinst ved at bære plasteret. Flere af deltagerne opfordrer til, at plasterets evne til at klæbe forbedres, da de har oplevet, at plasteret falder af i forbindelse med bad. Endvidere påpeger nogle af borgerne, at "tænd-sluk-knappen" skal flyttes, så man ikke ved et uheld får trykket på den.

Tabel 5.1. Besvarelser fordelt på svarmuligheder

Spørgsmål	Kommentarer
Hvordan oplevede du plasteret?	Ingen gener; skarpe kanter, der kan skære hul på huden; Udmærket – godt – tryghed; generede det 1 døgn tid derefter ok; Ikke generende men ustabil ift. at data ikke blev transmitteret konsekvent – hyppige udfald; Uden gener, ingen forandring; Ikke mærket til det; Ingen problemer; Mærkede det ikke; men det lyste om natten, hvor man lægger mest mærke til det; Ikke til besvær; Ikke mærket det;
Hvordan føltes det at have plasteret på?	Ingen gener; ok – mærkede det ikke; Udmærket; det har været helt fint; Hæmmede mig ikke men var opmærksom på om natten, at det ikke skulle slippe hudkontakt; Uden gener; Ingen problemer; Ingen gener; har kunne mærke plasteret, når ligget på siden ellers ikke; Normalt, generede ikke; Det har ikke siddet ordentligt fast – har forstærket med ekstra plaster ovenpå; Ikke mærket det
Oplevede du gener i forbindelse med at have plasteret på?	Ingen gener; nej; ingen gener; Nej; Nej; Minus; Nej; Nej; Nej; Lidt kløe; Nej
Kunne plasteret sidde fast efter bad eller gled det af?	Det gled af ved vask - påsat igen, ikke relevant; Har kun klatvasket mig, jeg etagebadede; Har siddet løst, men det blev trykket på plads igen et par gange; Ja; Minus; Nej, gled af, fik hjælp til at sætte nyt på; Nej, det har siddet fast under og efter bad; Det blev siddende; Har ikke været i bad, men gledet af alligevel; Ikke været i bad
Har du forslag til, hvad der kunne gøre det til en endnu bedre oplevelse at gå med sådan et plaster?	Nej, Nej - oplevelsen var generel god; Personalet har en enkelt gang udstrålet utrygheden ved plastret (SSA); Nej; Tættere kontakt mellem hud og sensor, bedre klisterfunktion for at undgå udfald af data; Nej; Nej; har ikke haft problemer; Tænd og sluk knap skal flyttes om på siden, så plasteret ikke kan slukkes ved tryk foran; Nej; Bedre klister; Nej
Kunne du tænke dig at plasteret blev taget i brug i hjemmeplejen?	Ja; ja – det viste sig at være ret godt; Ja; Ja, det kunne være godt, hvis det virker; Ja; Det er op til kommunen; Ja; Ja; Det kunne være en god ide, hvis der blev holdt øje med målingerne, det giver tryghed; Ja; Ja
Hvad tænker du om at plasteret giver mulighed for at sundheds-personale på afstand kan følge med i din tilstand?	God idé; Det er fint – kommer hurtig afsted til behandling; Tryghed; Det er en god ting; Supergodt, giver tryghed for mig, kan udbredes til andre i ringere konstitution end mig; Det har været ok; God ting; Glimrende ide; Synes det er godt; Det er ok – en god ide; Ja;
Hvilke muligheder kunne det give dig som borger, hvis du gik med sådan et plaster?	Ved ikke; Der er tjek på sundheden – skal ikke selv vurdere, hvordan man har det; ?; Ved ikke; Tryghed, døgnovervågning øger tryghed; Vides ikke, kunne være godt hvis det var forebyggende; Blive fulgt bedre; Tidligere kontakt ved symptomer; Monitorering af værdier – tryghed efter udskrivelse; Tryghed; Tryghed

Kunne det betyde noget for dine pårørende, at du har et plaster på?	Ikke umiddelbart; Det er en sikkerhed for pårørende; Der er i hjemmet ikke talt om plastret; Ved ikke; En øget tryghed; Det tror jeg ikke; Ja tryghed; Det er en sikkerhed for de pårørende; Ja stor tryghed og aflastning af pårørendes ansvar; Ja øget tryghed; Ja,
Har du andre forslag i forbindelse med plasteret?	Nej; Godt det ikke skal oplades; skarpe kanter; Nej; Nej; Nej; Nej; Kan være forebyggende; Nej; Nej; Hvordan man nemt kan dække det af i forbindelse med bad; Nej

5.6 Perspektivering

Måling af blodtryk med BioBeat er blevet sammenlignet med invasiv blodtryksmåling (gold standard) og med "almindelig" non-invasiv blodtryksmåling med sphygmomanometer, og der er fundet en god overensstemmelse.

Det burde således være muligt at tage BioBeat i brug til blodtryksmåling i forskellige situationer uden forudgående større undersøgelser af, hvorvidt metoden er præcis. Det kunne være:

- Døgnblodtryksmåling (mere behagelig)
- Blodtryksmåling under arbejdstest (eliminere problemer med upræcise målinger grundet fysisk aktivitet)
- Ortostatisk blodtryksmåling (undersøgelsens præcision forringes hvis blodtryksmåling tager for lang tid)
- Blodtryksmåling under indlæggelse på fx isolationsstuer eller ved behov for hyppig kontrol.

Dernæst kan tæt blodtrykskontrol iværksættes som hjemmemåling i andre situationer, hvor det kan betragtes som et nyt tilbud, der reducerer indlæggelser eller ambulant fremmøde. Eksempelvis:

- Monitorering af gravide med præeklampi

BioBeat kan måle adskillige andre hæmodynamiske parametre samt ekg (dog kun thorax-devicet). Det giver potentiale for:

- Fjernmonitorering af indlagte patienter, fx patienter på kirurgisk eller non-kardiologisk afdeling, der kan blive overvåget af mediciner/ kardiolog med potentiale for bedre monitorering og færre flytninger af patienter
- Observation af patienter på isolationsstue eller med andet behov for tæt overvågning
- Omlægning af invasiv hæmodynamisk monitorering til non-invasiv og derved mulighed for at monitorere patienter tættere på lavere indikation
- Monitorering af hvorvidt patienter bliver hæmodynamisk påvirkede af arytmi (indlagte eller hjemme)
- Monitorering af hvorvidt patienter med hjertesvigt:
 - o Oplever forværring, der kræver handling (fx kronikere der kan monitoreres hjemme)
 - o Er i bedring og kan udskrives efter indlæggelse.

Der er publiceret et studie af personer, der monitoreres hæmodynamisk i forbindelse med influenzasmitte, hvor det vises at flere parametre, ændres før klinisk sygdom. Det giver mulighed for:

- Tidlig opsporing af sygdom ved fx forventning om sygdom eller risiko for recidiv; det kunne være KOL-patienter, der måske kan diagnosticeres med exacerbation tidligere end ved nuværende vitalparametre og selvrapporterede symptomer

- Triagering under epidemi, fx patienter med positiv COVID-test, der kan monitoreres hjemme og kun udredes/behandles yderligere ved afvigende parametre.

KOL-APP



6. Pilotforsøg med PPG-teknologi

Der er i juni 2023 igangsat et pilotforsøg med at undersøge anvendelighed af kamerateknologi til måling af vitale værdier og dermed identifikation af forværringer hos KOL-borgere i PreCareKlinikken.

Teknologien anvender kamerabaseret photoplethysmografi (PPG) til bestemmelse af real-time værdier, og projektet vil afprøve om metoden kan erstatte eller supplere den egenmonitorering, der foregår i PreCareKlinikken i dag for borgere med KOL.

Pilotprojektet skal overordnet:

- Vurdere om målgruppen kan anvende teknologien, herunder forudsætninger for, at PPG kan fungere i eget hjem med eller uden støtte
- Afprøve om kamerateknologien kan anvendes til at indhente valide værdier med de i PreCareKlinikken allerede anvendte devices som benchmark

6.1 Metode

Teknologien baserer sig på photoplethysmografi (PPG), der er en ikke-invasiv metode til at måle ændringer i blodvolumen i væv. Det indebærer optagelse af lysabsorption og refleksion i vævet, normalt ved hjælp af en lyskilde og en fotodetektor.

Når lyset passerer gennem vævet, absorberes en del af det af det røde blodpigment (hæmoglobin) og reflekteres tilbage til fotodetektoren. Disse ændringer i lysabsorption og refleksion er relateret til blodets pulsationer og kan bruges til at estimere hjerterefrekvens, variation i hjerterefrekvens, blodgennemstrømning og andre hæmodynamiske parametre. PPG-teknologien er blevet anvendt i en bred vifte af kommercielt tilgængelige medicinske apparater til måling af blodtryk, iltmætning, puls og respirationsfrekvens. I første omgang måles iltmætning og puls.

Teknologien er ikke på nuværende godkendt som en del af Appinix KOL-app.

I pilotprojektet deltager 19 borgere, der allerede er tilknyttet PreCareKlinikken pga. deres lungesygdom. Det er således deltagere, der er bekendte med medarbejdere og arbejdsgange i PreCareKlinikken herunder telemedicinsk behandling, eget ansvar for at udføre målinger, betjening af tablets, deltagelse i telefoniske konsultationer med klinikkens sygeplejersker mm.

Målgruppen er inkluderede borgere, der regelmæssige gennemfører egenmålinger, og som har haft akutte forværringer i de sidste 3 måneder.

Formålet med testen er, at undersøge om der er andre metoder til at få tilstandsmålinger fra borgere på end ved de sædvanlige målinger, borgerne i dag anvender i PreCare-projektet med en saturation(SAT)-måler, termometer og spirometer.

I pilotforsøget er kamerafunktionen indsat i det almindelige KOL-tilstandsmålings-flow i app'en, således at borgeren laver sine almindelige målinger og samtidig foretager den nye måling ved hjælp af kameraet.

Målingen med kameraet foretages samtidig med målingen med SAT-måleren⁵. Borgeren tager SAT-måleren på den ene finger og holder en anden finger på tabletens kamera. Når begge målinger er kommet ind på skærmen, trykkes send, og herefter fortsætter resten af målingerne i app'en som de plejer.

Når målingerne skal indsamles fra kameraet, er der brug for at fingeren holdes på kameraet indtil, der er indsamlet nok informationer til, at målingerne kan opgøres. For at støtte borgeren i at holde fingeren korrekt vises en linje, som er rød, hvis fingeren ikke holdes rigtigt, gul hvis den næsten holdes rigtigt, og grøn når fingeren sidder korrekt på kameraet.

Fingeren skal mindst holdes 30 sekunder korrekt på kameraet for, at målingerne kan foretages. Hvis fingeren ikke sidder korrekt på kameraet (gul eller rød) i starten af de 30 sekunder, starter app'en forfra med at tælle 30 sekunder. Derfor kan borgeren komme til at sidde længere en 30 sekunder for at få en måling ind fra kameraet.

6.2 Data

Målingerne der kommer ind via kameraet er:

- Puls/Ilt lige som SAT-måleren
- Respiration
- Blodtryk Systolisk/Diastolisk

I pilotprojektet er der primært opmærksomhed på, om det er muligt at anvende denne form for teknologi hos den målgruppe af borgere, der er inkluderet i PreCareKlinikken. Altså kan de sætte fingeren korrekt på kameraet, og kan kameraet lave målinger på ældre fingre.

Dernæst testes hvilke målinger der kommer ind og de sammenlignes med de konventionelle målinger.

Det forventes, at målinger fra kameraet systematisk ligger for højt eller for lavt, og derfor er der behov for at sammenligne dem med de konventionelle målinger. I pilotforsøget kan målinger fra kameraet ikke direkte anvendes til vurdering af borgerens helbredstilstand. Men testen skal føre til en vurdering af, om målinger fra kamera på sigt kan anvendes.

Hvis der er en systematisk afvigelse fra de almindelige målinger, vil der kunne tages højde for det i algoritmen.

I perioden fra den 28. oktober 2023 til den 23. november 2023 har 17 borgere indsendt 217 målinger med i alt 1.085 målepunkter læst med kamerateknologi. Målepunkterne dækker over puls, iltmætning, respirationsfrekvens, systolisk og diastolisk blodtryk.

Borgerne har i perioden indsendt mellem 1 og 23 målinger, gennemsnit og median er 13.

⁵ Pulsoximeter, der måler iltmætning og puls ved PPG via fingerspidsen.

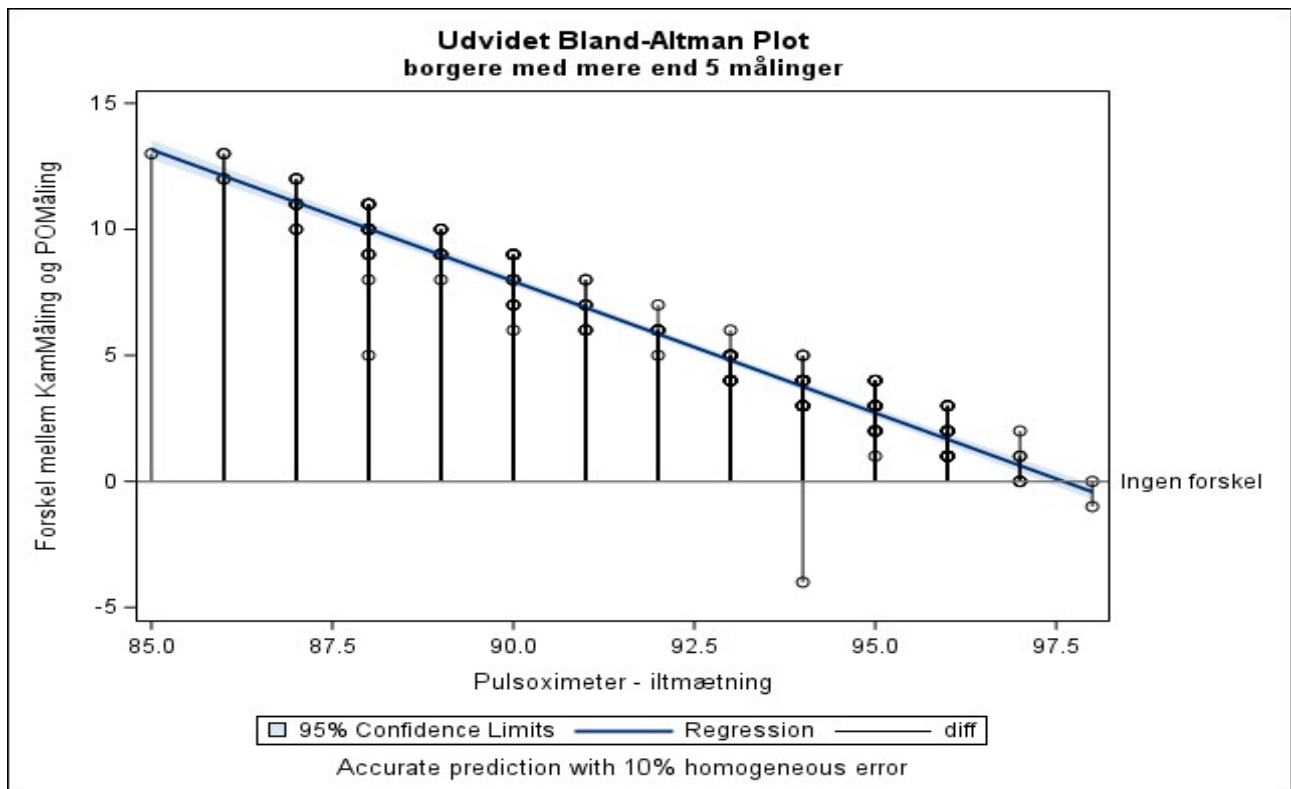
Der er 13 borgere, der har målt sig efter begge metoder samme dag i mindst 6 dage. Tilsammen har de lavet 172 målinger.

6.3 Resultater

Der er en systematisk, men heterogen bias i målinger af kamerateknologi sammenlignet med den normale måling, borgerne foretager med SAT-måleren (pulsoximeteret). Det betyder, at kamera-teknologien lige nu ikke måler konstant og ensartet for alle observationer, men varierer mellem forskellige områder og individer inden for datasættet.

Umiddelbart ser det ud til, at forskellen mellem de to målemetoder er størst for de målinger, hvor målingerne med pulsoximeteret har lavest værdi, dvs. de dårligste borgere.

Figuren nedenfor viser alle målinger for borgere, der på samme dag har målt sig med begge metoder, og som i perioden har målt sig mindst 6 gange.



Kilde: Appinux

På individniveau ses for alle de deltagende borgere en nedadgående trend, men også en stor variation mellem de enkelte individer.

6.3 Opmærksomhedspunkter

For at kameraet kan anvendes til målingerne skal der være en blitz.

Hvis fingeren lægges på blitzen, kan varmen fra pæren gøre ondt på fingeren. Derfor skal borge-ren gøres opmærksom på kun at holde fingeren på kameraet og ikke hen over blitzen.

App'ens kamerafunktion skal tilpasses det enkelte type device. Pilotforsøget har kun anvendt én type af Samsung tablet. Hvis flere typer forskellige devices skal anvendes, forudsætter det at kamerafunktionen kalibreres til de anvendte devices kamera-specifikationer.

App'en har desværre været for vanskelig at håndtere for de fleste af borgerne, da den i nogle tilfælde har krævet en genstart af tabletten, hvilket har udfordret nogle af deltagende borgere. Det betyder, at nogle af borgerne ikke får målt med kameraet, hver gang de foretager målinger med pulsoximeteret.

App'en er svært anvendelig for borgere, der er farveblinde, da de ikke kan se, hvornår fingeren ligger rigtigt på kameraet. Det samme gælder for borgere med meget krogede fingre, hvor kameraet ikke kan lave en måling.

Data har vist sig som ventet at være biased, men heterogeniteten udfordrer på nuværende tidspunkt mulighed for en simpel korrektion i algoritmen.

STEMMEGENKENDELSE



7. Pilotforsøg med stemmegenkendelse

I december 2022 blev der igangsat et pilotforsøg med henblik på at undersøge om stemmedata fra borgere med KOL kan erstatte eller supplere borgernes manuelle målinger. Hvis dette er tilfældet kan det være muligt at udvikle digitale tilbud til borgere med KOL ved brug af lyd/stemme-gengivelse. Det digitale tilbud er tænkt særligt til borgere med begrænsede digitale kompetencer.

7.1 Metode

Pilotforsøget skal indsamle lyddata fra ca. 25 udvalgte borgere med KOL, der er inkluderet i PreCareKlinikken, og der i dag foretager målinger af egen tilstand, og som ønsker at deltage i dette projekt. Perioden for indsamling af lyddata ønskes så lang som mulig. Hvis muligt bedes pårørende også at deltage, og dermed fungere som kontrolgruppe.

Indsamlingen af lyddata er foregået ved, at der er udviklet en simpel app, der er blevet installeret på en tablet og udleveret til borgeren. Borgeren og evt. pårørende har fået 4 forskellige tekster at vælge imellem, jf. bilag 4. De skal vælge én tekst, og læse den op alle dagene, pilotforsøget kører.

Herudover er der ikke sket ændringer i borgernes tilknytning til klinikken. De deltagende borgere er fortsat med at foretage de nuværende tilstandsmålinger, som vil blive fulgt op af PreCareKlinikkens personale, hvis der er tegn på forværringer.

Der er desuden ophængt en tablet i Regionshuset i Sorø/Ringsted, hvor medarbejdere har mulighed for at oplæse en tekst og dermed fungere som kontrolgruppe.

7.2 Data

De deltagende borgere og pårørende er blevet opfordret til at oplæse en tekst hver dag. Der er pr. 29. juni 2023 25 deltagende borgere og 6 pårørende og kontroller.

Der er indsamlet ca. 3.000 lydfiler, hvoraf 1.855 af dem er blevet transskriberet og er blevet benyttet i den videre analyse.

Blandt de transskriberet filer er 1.481 fra borgere med KOL og 374 lydfiler er fra kontrolgruppen uden KOL.

Forskellen mellem det totale antal transskriberede lydfiler hhv. for KOL-borgere og kontrolgruppen skyldes fejloptagelser.

Borgerne med KOL har foretaget de almindelige egenmålinger og oplæst tekster samme dag i alt 1.477 gange.

Der ses en lidt lavere aktivitet i weekenden end i hverdagen. Det samme ses for borgernes almindelige egenmålinger.

Hyppigheden af oplæsning af tekster sammenlignet med egenmålingerne har ændret sig over tid. Efter de første tre måneder læste ca. 63 pct. af de deltagende borgere oftere tekster op end de måler sig, mens omkring 17 pct. næsten læser tekster op lige så ofte, som de måler sig. De resterende 20 pct. af de deltagende borgere læser teksterne op færre gange, end de måler sig.

Efter ca. seks måneder læste ca. 20 pct. af de deltagerende borgere oftere tekster op end de måler sig, mens 15 pct. næsten lige så ofte læser tekster op, som de måler sig. De resterende ca. ca. 64 pct. af de deltagende borger læser teksterne op færre gange end de måler sig.

Der er i perioden fra 29. juni 2023 til 21. november 2023 foretaget i alt 2.297 egenmålinger. I samme periode blev der igangsat i alt 10 behandlinger.

Der er stor spredning i, hvor ofte KOL-borgerne oplæser tekster, jf. tabel 7.1 Det typiske antal oplæsninger i perioden er 61 gange. Borgerne har typisk oplæst mindst 1 gang om ugen i 20 uger, mens oplæsning af tekster mindst 5 gange om ugen typisk er sket i 6 uger, heraf har tre borgere ikke oplæst 5 gange om ugen på noget tidspunkt i løbet af perioden. Årsagen til få oplæsninger skyldes primært borgernes helbredstilstand.

Tabel 7.1. Antal oplæsninger af tekst pr. borger, herunder hyppigheden af oplæsningerne

Borger	Antal	Relativ frekvens	Uge med én eller flere oplæsninger	Uge med 5 eller flere oplæsninger
A	147	9,92	25	20
B	38	2,56	12	3
C	98	6,61	18	13
D	117	7,89	22	16
E	61	4,12	23	1
F	115	7,76	23	16
G	110	7,42	21	15
H	89	6,01	23	6
I	8	0,54	2	1
J	37	2,5	9	4
K	108	7,29	21	15
L	42	2,83	13	3
M	105	7,09	20	15
N	101	6,82	22	13
O	49	3,31	11	6
P	60	4,05	21	2
Q	12	0,81	5	0
R	95	6,41	20	16
S	27	1,82	12	1
T	50	3,37	21	0
U	12	0,81	4	0
Sum	1.481	100	348	166
Median	61	4,12	20	6

Kilde: PreCareKlinikken

7.3 Analyse af data

Den indledende analyse har primært koncentreret sig om data fra borgerne med KOL. Lyddataenes kvalitet er blevet bedømt ved at benytte en kombineret tilgang, der inkluderer lytteprøver for at verificere lydens gennemtrængelighed, samt anvendelse af en sprogmodel udviklet med kunstig intelligens (AI). Sprogmodellen transskriberer lyd til tekst.

Dette gør det muligt at beregne fejlmarginen ved sammenligning af det oplæste materiale med AI-modellens transskription. Derudover giver det mulighed for at få indsigt i blandt andet, hvor lang tids stilhed, der er imellem hvert sagt ord, hvilket muligvis kan give et signal omkring forværringer af KOL.

Kvaliteten af lyddataene varierer mellem deltagerne, hvor nogle indsendte filer er påvirket af forstyrrende baggrundsstøj, såsom radiolyd eller artefaktstøj fra elektrisk interferens. Disse udgør dog kun en mindre del af datasættet. Artefaktstøjen vurderes ikke at have effekt på den overordnet kvalitet. Radiolyden udgør omkring 1,6 pct. af datasættet. Det er muligt manuelt at fjerne den transskriberet radiolyd fra en stor del af oplæsningerne, der indeholder radiolyd.

Den manuelle behandling af transskriberet radiolyd er dog ikke foretaget på oplæsninger endnu. Vurderingen af kvaliteten er således baseret på det rå datasæt uden nogen yderligere behandling.

Den overordnet kvalitet af den transskriberet tekst blev vurderet ud fra 'Word Error Rate' (WER). WER sammenligner den transskriberede tekst med den tekst, som borgeren skulle læse op. Hvis der ikke er nogen fejl, er tallet 0. Hvis tallet er mellem 0 og 1, betyder det, at der er nogle fejl. For eksempel betyder et tal på 0,2, at der er 20 fejl ved hvert 100 ord. Hvis tallet er over 1, er der ingen ord, der passer. Det betyder, at der er flere ord, som er skrevet ned, end der var i teksten, borgeren skulle sige. Det kan ske, hvis der for eksempel er radiolyd i baggrunden.

Figur 9.1 viser den overordnet kvalitet, hvor det kan ses, at WER-medianen er 0,24. Det er generelt en god kvalitet, da ordfejlene er fejl som "da" og "der", "din" og "dit", "regnbuen" og "regnbogen" ol.

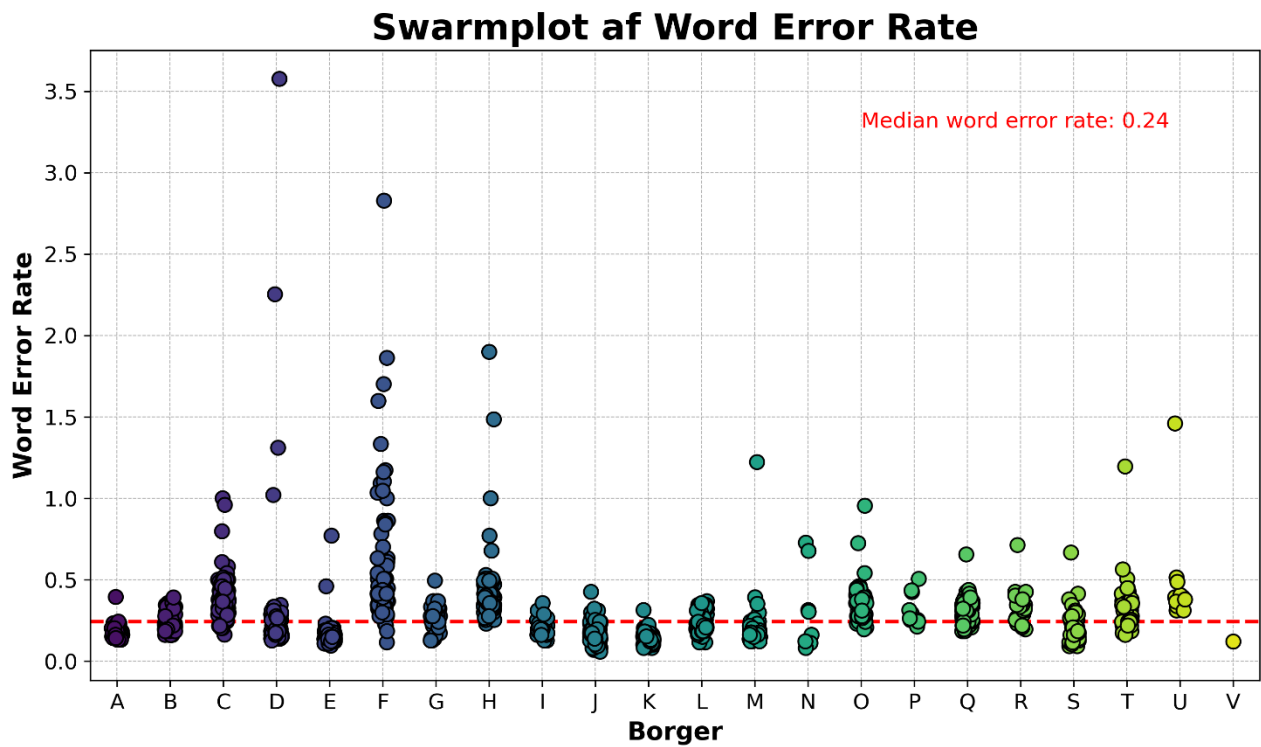
De eksempler, hvor fejlen er meget høj, skyldes meget støj i baggrunden typisk fra en radio.

Det er interessant at undersøge små fejl, da det kan indikere forskelle i udtalelse, eksempelvis hvis man har svært ved at få luft.

Det er 24 tilfælde, dvs. ca. 1.6 pct. af det totale datasæt, hvor WER er på 1 eller derover, dvs. at der ingen ord, der passer.

Der arbejdes videre med at undersøge om lyddata kan have potentiale til at forudsige forværringer i borgernes tilstand.

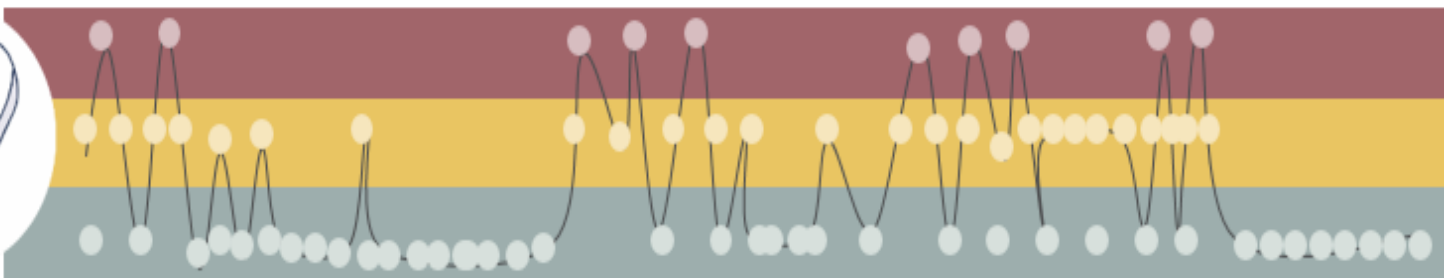
Figur 9.1 Swarmplot der viser Word Error Rate pr. transskriberet lydfil for borgere med KOL.



Note: Den stiplede linje indikerer medianværdien.

Kilde: PreCareKlinikken

KOL-ALGORITME



8. Evaluering af KOL-algoritme i PreCareKlinikken

I foråret 2023 er der foretaget analyser af mulighederne for:

- at forbedre KOL-algoritmen, således at den bedre kan bruges til at forudse, hvornår borgerens tilstand forværres
- at undersøge om det er muligt at bruge færre eller andre inputvariable end dem der anvendes for nuværende.

8.1 Metode

I arbejdet med analysen af egenmålingerne er der blevet struktureret to forskellige analysedatasæt fra dataene. De to datasæt bruger de samme 11 variable, jf. bilag 2. De 11 variable består af 7 tilstandsmålinger og 4 habituelle værdier.

Til begge datasæt er der tilføjet vejrdato fra vejrstationen i Københavns lufthavn.

De to primære evaluerings-metrikker, der er anvendt, er en "confusion matrix" og "ROC curve".

En confusion matrix er en metrik til at fortælle følgende fire ting:

1. Hvor mange **"ikke igangsættelser"**⁶ fanger algoritmen.
2. Hvor mange **"igangsættelser"** fanger algoritmen.
3. Hvor ofte prædikerer algoritmen en **"ikke igangsættelse"**, hvor målingen faktisk resulterede i en "igangsættelse".
4. Hvor ofte prædikerer algoritmen en **"igangsættelser"** hvor målingen faktisk ikke resulterede i en igangsættelse.

En perfekt algoritme resulterer i at have 100 pct. sande positive og sande negative. Men i virkeligheden vil det yderst sjældent forekomme, og accepten af en model vil derfor afhænge af en vurdering af modellens egenskaber og accepterede antal af falske positive og falske negative udfald.

En ROC-kurve er en grafisk fremstilling af ydeevne af en binær klassifikationsmodel, når man varierer modellens træfsikkerhedstærskel. Den illustrerer modellens evne til at skelne mellem den positive og negative klasse, og viser forholdet mellem modellens følsomhed og specificitet.

8.1.1 Baseline: Den nuværende KOL-algoritme

Baseline for analyserne er den nuværende KOL-algoritme, der anvendes i PreCareKlinikken. Der er foretaget en analyse af den nuværende algoritme for hver af de to opstillede datasæt, jf. bilag 5.

Datasættet med en række pr. måling viser, at der næsten ingen falske positive (3,02 pct.) er, mens performance i forhold til sande positive er mere middelmådig (44,09 pct.). Den nuværende KOL-algoritme er god til at klassificere ikke-igangsættelserne korrekt, men har problemer med at klassificere de målinger, der faktisk fører til igangsættelser. Det vil sige, at relativt mange igangsættelser ikke er baseret på en måling ud fra KOL-algorithmens triagering.

⁶ Igangsættelser af medicinske behandlinger af KOL-forværringer.

8.2 Resultater

Alle modellerne i det nedenstående er trænet på 80 pct. af data og evalueret på 20 pct. Derfor er de absolutte tal forskellige fra baseline, hvorfor det udelukkende er procentangivelserne, der kan sammenlignes.

Der er opstillet tre forskellige modeller:

- Logistisk regression
- Random Forest
- XGBoost

Alle tre modeller har en del bedre balanceret accuracy end baselinemodellen. XGBoost er bedst med 82 pct. accuracy, hvor der er 74,14 pct. sande positive imod de 44,09 pct. som baseline modellen rammer. Til gengæld har denne model også flest falske positive, jf. tabel 8.1.

Tabel 8.1. Evaluering af de forskellige maskinlæringsmodeller

Model	Balanceret accuracy	Sande positive	Falske positive
Baseline	71 pct.	44,09 pct.	3,02 pct.
Logistisk regression	78 pct.	62,76 pct.	7,26 pct.
Random Forest	80 pct.	68,97 pct.	8,39 pct.
XGBoost	82 pct.	74,14 pct.	9,29 pct.

Kilde: Appinux

For at eksemplificere hvad det betyder, at den bedste maskinlæringsmodel (XGBoost) har en større antal falske positive forudsigelser er der taget udgangspunkt i målinger fra ugen 1. marts til 7. marts 2023. I denne uge var der 858 målinger i alt, hvor 35 af målingerne er triageret røde og 5 røde målinger gav anledning til en igangsættelse. Der var i alt 11 igangsættelser, hvor 9 kunne knyttes til en måling.

Hvis XGBoost anvendes til at forudsige igangsættelser, vil det resultere i ca. 87-100 målinger triageres som røde målinger, hvor ca. 7 af igangsættelserne vil have en rød triagering. Dette betyder, at modellen vil øge antallet af røde målinger til 87-100 i forhold til den nuværende KOL-algoritme, hvor der er 35 røde målinger.

8.2.1 Analyser af logisk regressionsmodel med fravælgelse af variable

Der er foretaget analyser med udgangspunkt i en logistisk regressionsmodel, hvor det undersøges om fjernelse af nogle af variablene har betydning for modellens evne til at forudsige igangsættelser. Der er analyseret følgende modeller:

- Logistisk regression alene på variable, der har signifikant information (tre tilstandsspørgsmål)
- Logistisk regression, hvor alle variable indgår med undtagelse af lungefunktionsmåling og habitualværdien for lungefunktionen
- Logistisk regression alene ved brug af værdier for besvarelse af spørgsmålet om åndenød

I modellen alene baseret på de tre tilstandsspørgsmål ses, at den balancerede accuracy er identisk med den for den fulde model, jf. tabel 8.2. Dog er der lidt forskel på fordelingen af sande positive og falske positive, hvor der er en smule flere sande positive i denne model end den fulde model. Denne forskel er dog ikke signifikant.

Tabel 8.2. Analyse af reducerede maskinlæringsmodeller

Model	Balanceret accuracy	Sande positive	Falske positive
Baseline	71 pct.	44,09 pct.	3,02 pct.
difficultyBreathing alene	73 pct.	49,66 pct.	4,32 pct.
Logistisk regression fuld model	78 pct.	62,76 pct.	7,26 pct.
Logistisk regression reduceret model	78 pct.	61,38 pct.	6,23 pct.
Logistisk regression uden lunge funktion	78 pct.	62,76 pct.	7,19 pct.
Random Forest på alle variable	80 pct.	68,97 pct.	8,39 pct.
XGBoost på alle variable	82 pct.	74,14 pct.	9,29 pct.

Kilde: Appinix

Resultaterne for modellen uden lungefunktionsmålinger er stort set identiske med den fulde model. Det betyder, at man burde kunne undlade lungefunktionsmålingerne og opnå den samme forudsigelse som i en model med lungefunktionsmålinger.

Resultaterne fra en model alene baseret på besvarelse af spørgsmålet om åndenød ses det, at denne variabel alene med en balanceret accuracy på 73 pct. kan prædiktere bedre end KOL-algoritmen, som har en accuracy på 71 pct. Andelen af sande positive på 49,66 pct. er en smule højere end de 44,09 pct. for KOL-algoritmen, og andelen af falske positive er med 4,32 pct. er også en del højere end KOL-algorithmens på 3,02 pct.

Overordnet viser analyserne, at når features med ikke-signifikante parametre fjernes, har det ikke givet anledning til væsentligt dårligere performance, ligesom det heller ikke har givet anledning til dårligere performance, når lungefunktionsmålingerne er blevet fjernet.

8.4 Konklusion

Det kan generelt konkluderes, at det er muligt at forbedre baseline KOL-algoritmen. Analyserne har fokuseret på at optimere "balanced accuracy" metrikken, hvor der tydeligt ses en forbedring med identificering af sande positive og falske negative observationer, dvs. der er valgt at fokusere på at optimere udfaldet af så lave falske negative værdier og høje sande positive værdier som muligt.

Dog ses en tendens til, at en optimering af de falske negative og sande positive værdier kommer på bekostning af en model med flere falske positive. Dette kompromis er noget, der skal tages stilling til, hvis en endelig model skal idriftsættes. En drøftelse af balancen bør ske med det sundhedsfaglige personale, idet modellen vil være et dagligt værktøj for personalet og skal give mening for dem.

Fokus har været på at få så mange korrekt triagerede målinger forud for medicinske behandlinger af borgere, med den risiko at der vil være flere borgere, der i dag ikke laver røde målinger, men som vil få røde målinger i den nye model, og som stadig ikke giver anledning til igangsættelse af behandling. Dette vil bl.a. gøre, at der skal foretages flere opkald til borgerne om deres målinger for at være sikker på deres tilstand ikke er kritisk, men træfsikkerheden vil være større, og flere borgere med røde målinger vil få igangsat en behandling end i dag.

Resultaterne viser, at ved at bruge den balanceret accuracy som metrik, er det muligt at fjerne lungefunktionsmålingerne og stadig have en balanceret accuracy på 78 pct. Dog er det svært at sige, hvorvidt dette vil ændre sig med tiden efterhånden, som mere data kommer i spil.

Resultaterne viser også, at der kan laves en solid model udelukkende med variabelen baseret på besvarelse af tilstandsspørgsmål om åndenød. Dette skyldes, at værdien er den første i beslutningstræet/baseline algoritmen og derfor har enorm stor tyngde i triageringen. Den tendens antages ikke som en kompleks beslutning og kan evt. ses om et problem i baseline algoritmen, da en beslutning kan være baseret 100 pct. på baggrund af en variabel.

Analyse har også vist, at der ikke er en stor forskel på performance af de logistiske regressioner versus mere komplekse maskinlæringsmodeller. Det betyder, at hvis det ønskes, at modellen skal være transparent, vil den logistiske regression være en fin løsning. Dog kan der altid anvendes Explainable AI for at håndtere black-box problemet med maskinlæringsmodellerne.

8.5 Drøftelse af analyseresultaterne med det sundhedsfaglige personale

Analyseresultaterne har været drøftet med sygeplejerskerne i PreCareKlinikken. Drøftelserne tog bl.a. udgangspunkt i tabel 8.3, som viser et eksempel på, hvordan forudsigelser fra maskinlæringsalgoritmen kunne støtte sygeplejerskerne i monitoreringen af de inkluderede borgere målinger.

Tabel 8.3. Illustration af et beslutningsstøtteværktøj i forbindelse med monitorering af borgernes egenmålinger

Eksempel på information som beslutningsredskabet vil kunne bidrage med:

BORGER	IGANGSÆTTELSES SANDSYNLIGHED	FORKLARING
Knud	95,7%	Knud har en <i>øget temperatur på 5% og farvet sput.</i>
Erik	86,6%	Erik har <i>10% lavere lungefunktion.</i>
Bodil	54,7%	Bodil har en <i>øget temperatur på 2% og en tiltagende hoste.</i>
.	.	.
.	.	.
.	.	.
Jens	12,2%	Jens har en <i>øget puls på 10%</i>

Kilde: Data og udviklingsstøtte

Tabellen var et forslag til, hvordan forudsigelserne fra modellen vil kunne give en igangsættelses-sandsynlighed for hver borger, som er rangeret efter den højeste sandsynlighed for en igangsættelse. Formålet var at skabe en prioritering over borgerne og deres tilstand i et set-up med mange inkluderede borgere.

Beslutningsstøtteværktøjet skal ikke anvendes til automatisk at tage beslutninger, men som et redskab til at hjælpe sygeplejerskerne med en prioritering af borgerne. Det er op til sygeplejersker at beslutte, hvem de ringer til først. Redskabet tilbyder sygeplejerskerne en data-baseret rangering af borgernes tilstand – og qua analyserne ovenfor er det tydeligt, at modellen naturligvis ikke er 100 pct. korrekt.

Som opfølgning på mødet blev der bl.a. arbejdet videre med elementerne i beslutningsstøtteværktøjet. Der blev tilføjet farver på sandsynligheden/risiko for en igangsættelse og en kolonne med de 3 forrige målinger, jf. tabel 8.4.

Tabel 8.4. Justeret illustration af et beslutningsstøtteværktøj i forbindelse med monitorering af borgernes egenmålinger**Scenario:**

- Knud har fået en risiko procent tildelt samt en forklaring på "hvorfor" Knud har en høj risici.
- Knud's forrige 3 målinger var gule.

BORGER	RISIKO	FORRIGE 3 MÅLINGER	FORKLARING
Knud	95,7%	3 gule	Knud har en <i>øget temperatur på 1,2 grader og farvet spyt.</i>
Erik	86,6%	1 gul	Erik har <i>10% lavere lungfunktion.</i>
Bodil	54,7%	2 gule	Bodil har en <i>øget temperatur på 0,4 grader og en tiltagende hoste.</i>
...	...		...
Jens	52,2%	0 gule	Jens har en <i>øget puls på 10%</i>

Kilde: Data og udviklingsstøtte

Der blev også udarbejdet to konkrete cases ved anvendelse af den foreslåede maskinlæringsalgoritme. Casene tog udgangspunkt i 6 borgeres egenmålinger på to udvalgte dage i marts. Sygeplejerskerne blev hver især bedt om at rangere de 6 målinger (2 røde og 4 gule målinger) ud fra oplysningerne om målingerne ift., hvor dårlig borgeren forventedes at være.

På baggrund af rangeringen var der en drøftelse af de enkelte sygeplejerskers rangering og begrundelsen for rangeringen.

Tabel 8.5. Rangering af egenmålinger fra 20. marts 2023

Borger	Sgpl. A	Sgpl. B	Sgpl. C	Sgpl. D	Sgpl. E	Algoritme	Forklaring på rangering
2	1	1	2	1	2	3	
3	Ej kontakt	Ej kontakt	Ej kontakt	Ej kontakt	Ej kontakt	5	Kun én gul måling
4	3	Ej kontakt	1	1	Ej kontakt	2	Borger nyopstart
5	Ej kontakt	Ej kontakt	Ej kontakt	Ej kontakt	Ej kontakt	4	Kun én gul måling
6	2	2	3	3	1	1	7 gule målinger i træk samt åndenød

Note: Borger 1 kunne ikke slås op pga. manglende oplysninger.

Kilde: Data og udviklingsstøtte

Tabel 8.6. Illustration af beslutningsstøtteværktøjet for 20. marts 2023 sammenholdt med den nuværende KOL-algoritme

SCENARIO 1 – D. 20 MARTS - 2023						
Målinger for perioden						
88 23 2						
Nuværende triageringsmetode			Vores triageringsmetode			
BORGER	Måling	FORRIGE 3 DAGS MÅLINGER	BORGER	RISIKO	FORRIGE 3 MÅLINGER	FORKLARING
<u>1</u>		Grøn, gul, gul	<u>6</u>	94%	Gul, gul, gul	...
<u>2</u>		Ingen forrige målinger	<u>4</u>	84%	Grøn, gul, gul	...
<u>3</u>		Grøn, grøn	<u>2</u>	66%	Ingen forrige målinger	...
<u>4</u>		Grøn, gul, gul	<u>5</u>	49%	Ingen forrige målinger	...
<u>5</u>		Ingen forrige målinger	<u>1</u>	47%	Grøn, gul, gul	...
<u>6</u>		Gul, gul, gul	<u>3</u>	3%	Grøn, grøn	...

Kilde: Data og udviklingsstøtte

Efterfølgende blev den forbedrede algoritmes resultater præsenteret. Der var en vis overensstemmelse mellem sygeplejerskerne rækkefølge og algoritmens rækkefølge, der var forskellig fra den tidsmæssige rækkefølge, som målingerne almindeligvis kommer ind i klinikken, jf. tabel 8.5 og 8.7. Overensstemmelsen er lidt større ift. rangeringen den 21. marts end den foretagne rangering for den 20. marts 2023.

Tabel 8.7. Rangering af egenmålinger fra 21. marts 2023

Borger	Sgpl. A	Sgpl. B	Sgpl. C	Sgpl. D	Sgpl. E	Algoritme	Forklaring på rangering
2	1	1	1	1	3	1	
3	2	2	2	2	4	2	
4	3	Ej kontakt	Ej kontakt	Ej kontakt	1	5	Lav SAT og går nemt i panik
5	Ej kontakt	3	3	Ej kontakt	2	3	8 gule målinger i træk
6	Ej kontakt	Ej kontakt	Ej kontakt	Ej kontakt	Ej kontakt	4	

Note: Borger 1 kunne ikke slås op pga. manglende oplysninger.

Kilde: Data og udviklingsstøtte

Tabel 8.8. Illustration af beslutningsstøtteværktøjet for 21. marts 2023 sammenholdt med den nuværende KOL-algoritme

SCENARIO 2 – D. 21 MARTS - 2023						
Målinger for perioden						
93 9 3						
Nuværende triageringsmetode			Vores triageringsmetode			
BORGER	Måling	FORRIGE 3 DAGS MÅLINGER	BORGER	RISIKO	FORRIGE 3 MÅLINGER	FORKLARING
<u>1</u>		Grøn, gul, rød	<u>2</u>	97%	Ingen forrige målinger	...
<u>2</u>		Ingen forrige målinger	<u>1</u>	93%	Grøn, gul, rød	...
<u>3</u>		Ingen, ingen, grøn	<u>3</u>	55%	Ingen, ingen, grøn	...
<u>4</u>		Grøn, grøn, gul	<u>5</u>	50%	Gul, gul, gul	...
<u>5</u>		Gul, gul, gul	<u>6</u>	33%	Rød, rød, grøn	...
<u>6</u>		Rød, rød, grøn	<u>4</u>	10%	Grøn, grøn, gul	...

Kilde: Data og udviklingsstøtte

For begge cases blev der drøftet generelle tilgange ift. målingerne og dermed også begrundelser for prioritering fra sygeplejerskerne.

Generelt prioriterer sygeplejerskerne de røde målinger (som en del af PreCareKlinikkens service-mål) og derefter gule målinger dog afhængig af tidspunktet for målingen.

Sygeplejerskerne oplever, at der kan komme fejlmålinger ind fx ift. lungefunktionsmålingen. Hvis lungefunktionsmålinger afviger fra de øvrige målinger og besvarelse af spørgsmål, afventer sygeplejerskerne et par minutter for at se om, der kommer en ny måling fra borgeren med en anden værdi for lungefunktionsmålingen. I lyset af nogle af borgernes udfordringer med at have tilstrækkeligt med luft til at udføre en lungefunktionsmåling vurderer sygeplejerskerne, at disse målinger ikke er lige så vigtige som de øvrige målinger og besvarelse af spørgsmål.

Sygeplejerskerne tillægger besvarelsen af tilstandsspørgsmålene om øget åndenød, tiltagende hoste og farvet opspyt som vigtige parametre ift. borgernes tilstand.

Et andet element som sygeplejerskerne peger på er deres kendskab til borgerne, herunder deres mentale tilstand, og hvordan de reagerer i forbindelse med forværringer.

8.6 Perspektivering

Der er i forbindelse med analyserne gjort overvejelser om næste skridt for at forbedre modellens performance mhp. en senere idriftsættelse, *jf. bilag 5 for en fuld og uddybet gennemgang af alle overvejelserne.*

Beslutning om flere falske positive og falske negative

Der skal tages en beslutning om, hvorvidt der ønskes at optimere de falske positive eller de falske negative, da der umiddelbart i analyserne ikke er en stabil balance mellem de to. Balancen skal drøftes med det sundhedsfaglige personale, idet modellen vil være et værktøj, der påvirker sygeplejerskernes arbejde.

Mere evaluering mellem modellerne (krydsvalidering)

Da der er mange forskellige modeller, vil det være hensigtsmæssigt med mere evaluering mellem modellerne for med højere sikkerhed at vurdere, hvilken model der bedst kan anbefales.

Modelkalibrering

Modelkalibrering refererer til processen med at justere forudsigelserne eller output-sandsynlighederne i en maskinlæringsmodel for at stemme overens med de sande sandsynligheder eller resultater i den virkelige verden. Kalibrering er vigtig i forskellige anvendelser, især når output-sandsynlighederne bruges til beslutningstagning eller risikovurdering. Hvis en model er dårligt kalibreret, kan de forudsagte sandsynligheder muligvis ikke nøjagtigt afspejle sandsynligheden for, at en begivenhed faktisk forekommer, hvilket fører til forkerte beslutninger eller upålidelig risikovurdering.

I de gennemførte analyser er der foretaget en hurtig kalibrering af modellen, og derfor vil det være hensigtsmæssigt at kigge dybere ind i dette.

Anvendelse af Explainable AI (XAI)

Explainable AI (XAI) er en del af maskinlæring, hvor komplekse modeller kan blive gjort transparente. Ved at anvende XAI kan der ses meget mere specifikt på, hvorfor modellerne er kommet til de forudsigelser, som de er. Det giver mulighed for at optimere både datasæt og model.

For borgerne specifikt, vil det være muligt at lave en "forklaring" i form af fx "Knud er forudset til at blive igangsat indenfor de næste 4 dage, fordi Knud har haft farvet sput, høj temperatur og der har været høj fugtighed. Der er også mange andre måder XAI kan blive brugt på, fx. at gruppere "observationer/borgere" og identificere bestemte trends for grupper af borgere, så der kan optimeres ift. behandlinger m.m.

MÅLERUTINER



9. Analyse af målerutiner

Der blev i september 2023 igangsat analyser af borgernes målerutiner.

Formålet med pilotprojektet var at gennemføre analyser, der kan hjælpe med at svare på, om der er hændelser og i givet fald hvilke, der tilsiger behov for ændringer i borgernes målerutiner. Det kan fx ske ved at være opmærksom på generelle hændelser så som vejrskifte, pollensæson, årstider eller individuelle hændelser så som indlæggelse for KOL-forværringer eller borgerens stamdata i form af fx flere kroniske sygdomme, hurtig ændring i habituel tilstand eller graden af lungefunktionsnedsættelse.

Borgere med KOL har en forskellig risikoscore, baseret på hyppighed af forværringer og oplevede symptomer og dermed også behov for forskellige tilgange ift. monitorering/målerutiner. Der er ca. 30.000 borgere med KOL i Region Sjælland. Tidligere undersøgelser har vist, at fordelingen af borgere med KOL på risikoscore A-D er 39 pct. med risikoscore A, 53 pct. med risikoscore B, 5 pct. med risikoscore C og 3 pct. med risikoscore D.

Udgangspunkt for nedenstående analyse var initialt at undersøge det reelle målebehov, og potentielt at komme med et databaseret estimat af den optimale målefrekvens, der resulterer i mindst mulig intervention i borgerens hverdag uden tab af information, samt at undersøge om borgerne kan segmenteres efter behov risikogrupper med hver deres optimale måleregime.

9.1 Baggrund

Borgerne i PreCareKlinikken bliver opfordret til at måle sig dagligt. I gennemsnit måler KOL-borgerne sig 4-5 gange om ugen. Det gennemsnitlige antal ugentlige målinger har stort set ligget stabilt på dette niveau fra etableringen af PreCareKlinikken ultimo 2018.

Når tilstandsmålingerne og tilstandsvurderingerne bliver registreret, bliver de automatisk triageret og tildelt en samlet vurdering i form af farven grøn, gul eller rød. Klassificeringen foregår ved hjælp af en algoritme, som er baseret på værdier for den enkelte borgers habitueltilstand. Røde målinger indikerer, at borgeren muligvis er på vej til at få en KOL-forværring. Det samme gør 3 på hinanden følgende gule målinger. 79 pct. af de gennemførte tilstandsmålinger er grønne, mens 17 pct. er gule og 4 pct. er røde.

Sygeplejerskerne i PreCareKlinikken kontakter alle borgere med røde målinger inden for maksimalt 1 time efter målingerne er registreret. Tilsvarende kontakter sygeplejerskerne de borgere, der har 3 på hinanden følgende gule målinger. På baggrund af målingerne og sygeplejerskernes dialog med den enkelte borgere, vurderes behovet for at igangsætte en behandling for en KOL-forværring. Hvis sygeplejerskerne er i tvivl, kontakter de elægen, der rådgiver ift. behandling.

Det er dog langt fra, at alle røde målinger fører til et behandlingsforløb. Der igangsættes behandlingsforløb for godt 17 pct. af de røde målinger, og 3 pct. af alle gule målinger resulterer i et behandlingsforløb. Ca. 16 pct. af alle behandlingsforløb igangsættes enten på baggrund af en grøn tilstandsmåling eller uden nogen forudgående tilstandsmåling.

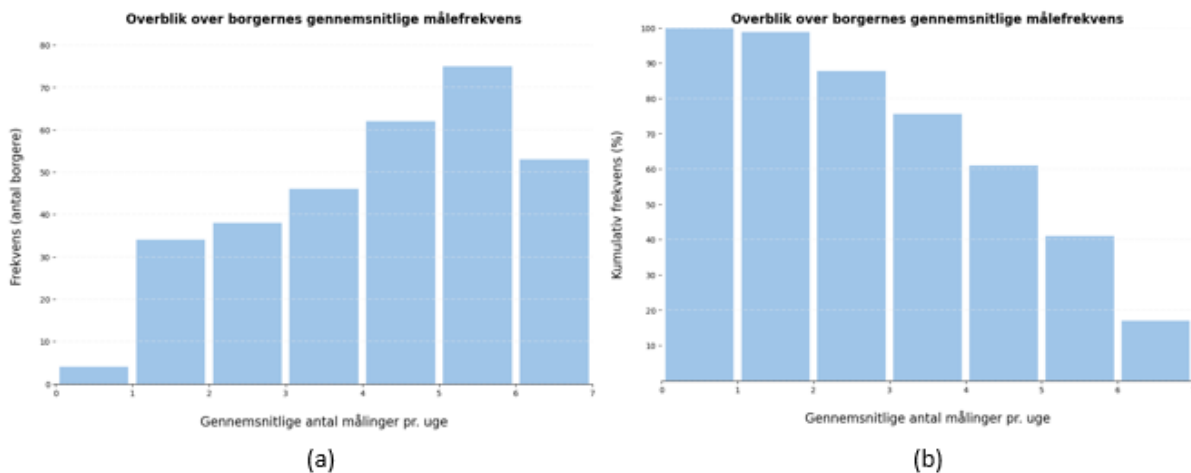
Det er dog vigtigt at holde sig for øje, at succeskriteriet i PreCareKlinikken er at holde borgerne i deres habituelle tilstand, og kun i det omfang, at det er nødvendigt at igangsætte behandlinger, sker det.

9.2 Data og metode

Analysen er baseret på indsamlet data i PreCareKlinikken fra borgernes tilstandsmålinger fra 5. oktober 2018 frem til 27. september 2023 for borgere, der har været inkluderet 90 dage eller mere i PreCareKlinikken.

Borgernes målerutiner er illustreret i Figur 9, som viser, at tæt på 100 pct. af alle borgere måler sig mellem 1 og 2 gange om ugen gennemsnitligt, over 85 pct. af borgerne måler sig 2-3 gange om ugen gennemsnitligt, mens mere end 75 pct. af borgerne måler sig 3-4 gange om ugen gennemsnitligt.

Figur 1. Overblik over borgernes gennemsnitlige antal målinger pr. uge



Note: (a) viser den absolutte frekvens af borgere fordelt på antallet af gennemsnitlige målinger per uge, mens (b) viser den kumulative frekvens af borgere fordelt på antallet af gennemsnitlige målinger per uge.

Kilde: PreCareKlinikken

Af tabel 9.1 fremgår et overblik over de inkluderede borgeres egenmålinger, triageringsfarve samt igangsættelse af behandling.

Datagrundlaget indeholder i alt 22.195 gule målinger og 4.358 røde målinger fordelt på de 312 borgere, og langt hovedparten af disse målinger har ikke været ledsaget af en medicinsk igangsættelse.

Tabel 9.1. Overblik over borgernes inklusionstid, målerutiner, medicinske igangsættelser samt fordeling af gule og røde målinger i perioden 5. oktober 2018 til 27. september 2023

	Middel ($\pm$ std.)	Min.	25%-kvar-til	50%-kvar-til	75%-kvar-til	Max.
Inklusionstid (dage)	744 ($\pm$ 434)	90	368	689	1038	1771
Gennemsnitligt antal ugentlige målinger	4,3 ($\pm$ 1,6)	0,2	3,0	4,5	5,6	7,0
Antal igangsættelser i hele perioden	5,8 ($\pm$ 6,5)	0	1	4	8	41
Gennemsnitligt antal igangsættelser pr. inklusions-år*	3,1 ($\pm$ 2,7)	0,0	0,9	2,4	4,6	13,5
Antal gule målinger i hele perioden	77,9 ($\pm$ 73,5)	3	34	57	98	470
Gennemsnitligt antal gule målinger pr. inklusions-år*	44,0 ($\pm$ 35,4)	2,1	21,0	36,5	55,8	329,3
Antal røde målinger i hele perioden	14,1 ($\pm$ 15,7)	0	4	10	18	120
Gennemsnitligt antal røde målinger pr. inklusions-år*	9,1 ($\pm$ 10,7)	0,0	2,7	5,5	11,9	95,9

Note: * Gennemsnitligt antal målinger/igangsættelser pr. inklusionsår er beregnet som et mål for nemmere sammenligning af alle borgere, uanset varighed af inklusionstiden. For borgere, der har været inkluderet mindre end et år, vil dette derfor afspejle et estimat af antal målinger/igangsættelser pr. år med den antagelse, at den gennemsnitlige rate pr. år er uændret ift. raten i inklusionsperioden.

Kilde: Appinux.

I tabel 9.2 er foretaget en opgørelse af de samme variable som i tabel 9.1, men grupperet på borgernes risikoscore ved inklusion i PreCareKlinikken. Det fremgår, at borgernes gennemsnitlige målerutiner er sammenlignelige på tværs af risikogrupperingerne. Det fremgår også, at antallet af gule triageringer gennemsnitligt er større, des lavere risikogruppe. Samme tendens ses med antallet af røde triageringer samt antal medicinske igangsættelser.

Tabel 9.2. Overblik over borgernes inklusionstid, målerutiner, medicinske igangsættelser samt fordeling af gule og røde målinger i perioden for dataopsamlingen grupperet på risikoscore ved inklusion

Risikogruppe	A	B	C	D
Antal borgere	19	85	70	134
Inklusionstid (dage)	849 ($\pm$ 398)	717 ($\pm$ 384)	971 ($\pm$ 424)	617 ($\pm$ 421)
Gennemsnitligt antal ugentlige målinger	4,1 ($\pm$ 1,6)	4,2 ($\pm$ 1,7)	4,3 ($\pm$ 1,6)	4,4 ($\pm$ 1,6)
Antal igangsættelser i hele perioden	2,4 ($\pm$ 2,2)	4,3 ($\pm$ 4,8)	7,4 ($\pm$ 6,8)	6,3 ($\pm$ 7,0)
Gennemsnitligt antal igangsættelser pr. inklusions-år*	1,0 ($\pm$ 1,1)	2,4 ($\pm$ 2,1)	3,0 ($\pm$ 2,7)	3,9 ($\pm$ 2,9)
Antal gule målinger i hele perioden	65,6 ($\pm$ 47,7)	65,2 ($\pm$ 48,9)	91,8 ($\pm$ 81,5)	81,1 ($\pm$ 84,0)
Gennemsnitligt antal gule målinger pr. inklusions-år*	29,6 ($\pm$ 17,7)	37,4 ($\pm$ 25,8)	38,8 ($\pm$ 29,9)	53,4 ($\pm$ 42,5)

Antal røde målinger i hele perioden	8,2 ($\pm$ 6,2)	13,0 ($\pm$ 14,4)	13,9 ($\pm$ 15,3)	15,6 ($\pm$ 17,5)
Gennemsnitligt antal røde målinger pr. inklusions-år*	3,7 ($\pm$ 3,6)	7,6 ($\pm$ 7,0)	6,3 ($\pm$ 7,1)	12,4 ($\pm$ 13,7)

Note: * Gennemsnitligt antal målinger/igangsættelser pr. inklusionsår er beregnet som et mål for nemmere sammenligning af alle borgere, uanset varighed af inklusionstiden. For borgere, der har været inkluderet mindre end et år, vil dette derfor afspejle et estimat af antal målinger/igangsættelser pr. år med den antagelse, at den gennemsnitlige rate pr. år er uændret ift. raten i inklusionsperioden.

Kilde: Appinux

Der er i den anvendte triagering en stor andel af de røde eller minimum 3 på hinanden følgende gule målinger, som er falske positive, dvs. de ikke er ledsaget af en igangsættelse. Af de 3.734 røde målinger, der ikke er ledsaget af en igangsættelse, er årsagen til den røde triagering fald i lungefunktion for 2.749 (74 pct.) af egenmålingerne, mens fald i lungefunktion er årsag til 10.938 (51 pct.) ud af de 21.564 gule egenmålinger, der ikke er ledsaget af en igangsættelse. Lungefunktion er altså den hyppigste årsag til gule og røde triageringer uden medicinsk igangsættelse (falske positive).

9.3 Resultater

9.3.1 Triagering af borgernes egenmålinger

Analysen af triageringen af borgernes egenmålinger viser, at den nuværende Appinux-algoritme fanger ca. 58 pct. af igangsættelserne, men også resulterer i en del falske positive (en falsk-positivrate på 4,1 pct.), og at årsagen til de falske positive hovedsageligt skyldes lav lungefunktion (74 pct. af de røde egenmålinger, der ikke er ledsaget af en medicinsk igangsættelse skyldes fald i lungefunktion, mens fald i lungefunktion er årsag til 51 pct. af de gule målinger, der ikke er ledsaget af en medicinsk igangsættelse). Lungefunktion er altså den hyppigste årsag til gule og røde triageringer uden medicinsk igangsættelse (falske positive).

Triagerings-algoritmen blev afprøvet uden lungefunktionsmålingen, hvorved andelen af falske positive blev reduceret til 1,5 pct., men samtidigt faldt andelen af sande positive også til 25 pct. Dette indikerer, at store udsving i borgernes målte lungefunktion er årsag til mange falske positive, men at lungefunktionsmålingen stadigvæk har en stor kraft i at opdage forværringer, der leder til medicinsk igangsættelse hos borgerne med den nuværende Appinux triagerings-algoritme.

9.3.2 Variansanalyse

Variansen i borgernes egenmålinger er analyseret, hvor det blev særligt tydeligt, at borgerne har svært ved at udføre lungefunktionsmålingerne korrekt. Borgernes lungefunktionsmålinger (målt som pct. af forventet FEV1) har i gennemsnit en standardafvigelse på 3,8 pct. Det vil sige, at alle målingerne i gennemsnit afviger med 3,8 pct. fra middelværdien (og nogle målinger altså afviger mere, og andre målinger afviger mindre).

For borgere med en generel dårlig lungefunktion, vil de observerede almindelige udsving i den målte lungefunktion hurtigt resultere i en gul eller endda rød triagering af egenmålingen på trods af, at borgeren ikke oplever en reel tilstandsforværring. De store udsving i lungefunktionsmålingerne gør det udfordrende at bruge disse målinger i en valid triagering af borgernes tilstand.

Analysen af variansen i lungefunktionsmålingerne indikerer desuden, at variansen bliver en smule mindre des flere dage de enkelte borgere har været inkluderet i PreCareKlinikken, hvilket kan tyde på, at borgerne muligvis bliver bedre til at udføre korrekte egenmålinger med tiden. Generelt bely-

ser analysen et behov for øget fokus på at undervise borgerne i at udføre korrekte lungefunktionsmålinger, eller at assistere dem virtuelt, eksempelvis i form af en digital hjælper, der guider borgeren undervejs, og beder borgeren gentage en måling, såfremt den virker urealistisk.

Der observeres også en smule varians i de resterende egenmålte variable, hvor puls i gennemsnit varierer med ca. 8 slag i minuttet, iltmætningen varierer gennemsnitligt med 1,85 pct., og temperaturen varierer gennemsnitligt med 0,35°C, men variationen i lungefunktionsmålingerne vurderes at være den mest kritiske i forhold til validiteten af egenmålingen.

9.3.3 Evaluering af andre modeller

Til evaluering af om triageringsmetoden kan forbedres ved at anvende maskinlæring, er der foretaget forskellige tests af modellering af data. Der er afprøvet forskellige maskinlæringsmodeller på de egenmålte variable samt variable, der er afledt af disse. Desuden er forskellige eksterne variable medtaget i afprøvningen for at evaluere effekten af eksterne faktorer på borgernes tilstand.

En række modeller med varierende antal inputvariable er blevet trænet og evalueret for at undersøge, om triageringen kan forbedres ved brug af maskinlæring, samt om triageringen kan forbedres ved at medtage eksterne faktorer såsom vejr, pollental og sæson.

Først er der trænet en model på de rå egenmålinger (de 3 tilstandsspørgsmål, lungefunktion, puls, temperatur og iltmætning), og derefter er der trænet en række andre modeller med forskellige inputvariable. Et overblik over inputvariablene kan ses i tabel 9.3, sammen med den balancerede nøjagtighed, sand-positivrate og falsk-positivrate for de forskellige modeller, når de er evalueret på testdatasættet.

Tabel 9.3. Oversigt over inputvariable i de afprøvede maskinlæringsmodeller af forskellig kompleksitet samt den balancerede nøjagtighed (F1-score), sand-positivrate (TPR) og falsk-positivrate (FPR).

Model	Inkluderede parametre	Balanceret nøjagtighed ⁷	TPR	FPR
Base-line	Tilstandsspørgsmål Lungefunktion Temperatur Iltmætning Puls	94%	62,7%	9,6%
XXSmall	Tilstandsspørgsmål	96%	61,0%	5,0%
XSmall	Tilstandsspørgsmål Temperatur relativt til habituelværdi Iltmætning relativt til habituelværdi Puls relativt til habituelværdi	94%	62,0%	8,8%
Small	Tilstandsspørgsmål Vejrdata ⁸ Pollendata ⁹ Måned Sæson	95%	52,6%	8,1%
Medium	Tilstandsspørgsmål Lungefunktion (% af forventet) relativt til habituelværdi Temperatur relativt til habituelværdi	94%	73,2%	9,6%

⁷ Målt som F1-score, der er det harmoniske gennemsnit af præcision og sensitivitet ($F1 = 2 * (\text{præcision} * \text{sensitivitet}) / (\text{præcision} + \text{sensitivitet})$).

⁸ Vejrdato inkluderer dagens gennemsnitlige målte temperatur, luftfugtighed og lufttryk.

⁹ Pollendata inkluderer dagens målte pollental for El, Hassel, Elm, Birk, Græs, Alternaria, Cladosporium.

	Iltmætning relativt til habituelværdi Puls relativt til habituelværdi			
Large	Tilstandsspørgsmål Lungefunktion (% af forventet) relativt til habituelværdi Temperatur relativt til habituelværdi Iltmætning relativt til habituelværdi Puls relativt til habituelværdi Euclidian distance ¹⁰ Lag 1	94%	68,8%	8,7%
XLarge	Tilstandsspørgsmål Lungefunktion (% af forventet) relativt til habituelværdi Temperatur relativt til habituelværdi Iltmætning relativt til habituelværdi Puls relativt til habituelværdi Euclidian distance ¹¹ Lag 1 - 5 ¹²	95%	62,0%	7,2%

Kilde: Appinix og Københavns lufthavns vejrstation.

Baseret på resultaterne i tabel 9.3 tyder det på, at modellens evne til at kunne prædiktere igangsættelser ikke nødvendigvis bliver bedre med modelkompleksiteten. På trods af at de største modeller (Large og XLarge) har høj nøjagtighed skyldes dette, at antallet af falske positive er lavt, men antallet af sande positive er til gengæld også lavt. De største modeller er altså ikke ideelle til at fange igangsættelserne.

Den model med den højeste sand-positivrate er Medium-modellen, der er bygget på de 3 tilstandsspørgsmål samt FEV1 (pct. af forventet), temperatur, iltmætning og puls relativt til de respektive habituel-værdier. Denne model er altså bygget på de samme inputvariable som Appinix-algoritmen, men opnår en væsentligt højere sand-positivrate på 73 pct. (sammenlignet med 58 pct. for Appinix-algoritmen). Dog ses der også en lidt højere falsk-positivrate på 10 pct. (sammenlignet med 4 pct. for Appinix-algoritmen).

Tabel 9.3 viser også at XXSmall-modellen, der alene er baseret på de 3 tilstandsspørgsmål, opnår en sand-positivrate på 61 pct., mens falsk-positivraten holdes nede på 5 pct. XSmall-modellen, der inkluderer de samme parametre som Medium-modellen undtagen lungefunktion resulterer i en sand-positivrate på 62 pct. og en falsk-positivrate på 9 pct.

Når de eksterne faktorer, såsom vejr og sæson, inkluderes i modellen, observeres et fald i sand-positivraten.

9.3.4 Forudsigelse af forværringer tidligere

I den nuværende datamodel er det kun den ene måling på dagen for en igangsættelse, der får tildelt en positiv label. Dette betyder, at modellen ikke er trænet til at genkende eventuelle mønstre i data for de målinger, der er foretaget i dagene op til en medicinsk igangsættelse.

Det ideelle vil være at kunne prædiktere en sygdomsforværring i så god tid, at sygeplejerskerne eller borgerne selv kan initiere tiltag, der potentielt kan forhindre en medicinsk igangsættelse. Derfor testes forskellige prædiktionsvinduer (antal dage forud for en igangsættelse), for at undersøge, om det er muligt at prædiktere igangsættelserne flere dage, før de forekommer.

¹⁰ Euclidian distance er beregnet som summen af lungefunktion, temperatur, iltmætning og puls (relativt til de respektive habituel-værdier) kvadreret.

¹¹ Beregnet som summen af lungefunktion, temperatur, iltmætning og puls (relativt til habituel-værdierne) kvadreret.

¹² Lag er medtaget for lag 1 til lag 5 (altså i alt 5 individuelle lag features)

Prædiktionsvinduer på 0 til 7 dage er testet med XXSmall-modellen og Medium-modellen. Resultaterne af modeevalueringen for de 7 afprøvede prædiktionsvinduer er vist i tabel 9.4.

Tabel 9.4. Overblik over resultater af evaluering af XXS og M modellerne med prædiktionsvinduer mellem 0 og 7 dage

Prædiktions-vindue	XXSmall model			Medium model		
	Balanceret nøjagtighed	TPR	FPR	Balanceret nøjagtighed	TPR	FPR
0 dage	96%	61,0%	5,0%	94%	73,2%	9,6%
1 dag	96%	58,2%	5,0%	91%	79,4%	13,6%
2 dage	96%	55,0%	4,9%	90%	79,0%	15,3%
3 dage	96%	50,7%	4,9%	89%	76,0%	17,1%
4 dage	96%	46,7%	4,9%	88%	74,2%	17,9%
5 dage	95%	43,7%	4,9%	87%	71,5%	19,6%
6 dage	95%	41,2%	4,9%	86%	68,0%	20,7%
7 dage	95%	38,7%	4,9%	85%	68,0%	22,0%

Kilde: Appinux.

For XXSmall-modellen forbliver falsk-positivraten stort set uændret uanset prædiktionsvinduet, mens sand-positivraten falder, når prædiktionsvinduet udvides. Dette indikerer, at det er svært at forudsige sygdomsforværringer flere dage i forvejen alene baseret på de 3 tilstandsspørgsmål.

For Medium-modellen stiger sand-positivraten derimod, når prædiktionsvinduet udvides til en dag, hvorefter den falder en smule i takt med, at prædiktionsvinduet udvides yderligere. Falsk-positivraten stiger dog også i takt med, at prædiktionsvinduet udvides. Det tyder altså på, at mange af sygdomsforværringerne kan forudsiges flere dage forud for en igangsættelse på trods af, at modellen også returnerer en større andel af falske positive.

Når prædiktionsvinduet udvides, overlapper de sande igangsættelser og ikke-igangsættelser i stigende grad. Trods den faldende grad af adskillelse indikerer analyserne, at man til en vis grad kan forudsige igangsættelser, men med stigende grad af sikkerhed jo tættere man kommer på selve igangsættelsen.

9.4 Perspektivering

På baggrund af det nuværende datagrundlag er det svært at sige noget om den optimale målefrekvens for borgerne, da datakvaliteten generelt ikke er tilfredsstillende. Hvis datakvaliteten generelt forbedres (særligt hvis borgerne bliver bedre til at udføre valide lungefunktions-målinger), vurderes det, at der er et stort potentiale for at udvikle en maskinlæringsmodel, der er i stand til at prædiktere forværringer hos borgere flere dage i forvejen.

Derved kan målefrekvensen potentielt også sættes ned, således at borgere ikke behøver at udføre egenmålinger dagligt, men at borgeren derimod promptes for flere målinger, såfremt maskinlæringsmodellen returnerer en forhøjet sandsynlighed for en forværring inden for de nærmeste dage.

Analysen af borgernes egenmålinger har belyst et behov for straks-validering af indtastede og målte værdier, således at egenmålingerne bliver mere valide og antallet af forkerte triageringer mindskes.

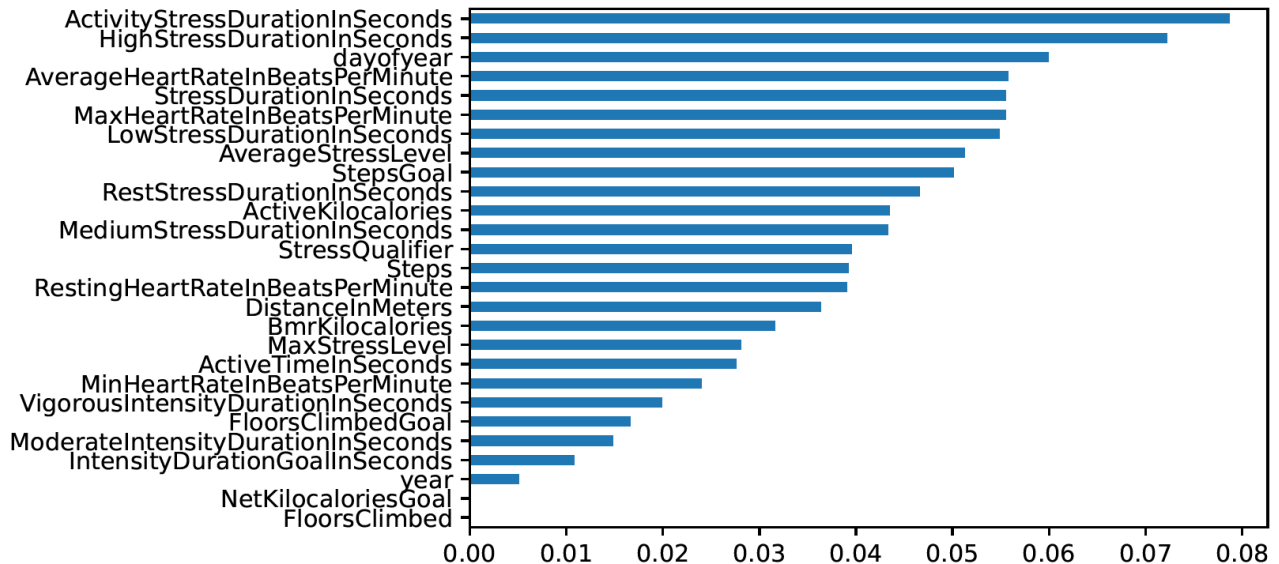
Derudover har analysen vist, at lungefunktionsmålingerne har en stor prædiktiv kraft, men samtidig er årsag til mange falske positive, da der er stor varians i borgernes egenmålte lungefunktion, hvilket belyser, at der er et behov for øget fokus på at lære borgerne at udføre lungefunktionsmålingen

korrekt eller eventuelt at gentage en måling, hvis den afviger meget fra et normalområde eller den sidst udførte måling.

Analysen har ligeledes understreget den prædiktive kraft i de tre tilstandsspørgsmål, hvilket peger på et potentiale for at anvende disse til den daglige tilstandsevaluering, som kan danne grundlag for, at borgeren promptes for mere omfattende egenmålinger. Tilstandsspørgsmålene er mindre intervenserende end de nuværende egenmålinger, og den prædiktive kraft kan muligvis tilmed forbedres, hvis der inkluderes flere spørgsmål i den daglige tilstandsevaluering.

Bilag 1. Gennemgang af præliminære analyser af Garmin data 18.03.2022

Machine Learning feature importance på dailies data



- Machine Learning med afsæt i de data Garmin summerer dagligt (dailies) viser en betragtelig betydning af stress. Stress er i Garmins algoritme afledt af Heart Rate Variability (HRV), altså hvor stor varians der er mellem de enkelte hjerteslag.

Machine Learning feature importance på rå biometriske data

SP02Value_MAX	0.002853
SP02Value_MIN	0.006513
SP02Value_AVG	0.013755
BreathsPerMin_MIN	0.064308
BreathsPerMin_MAX	0.075314
BreathsPerMin_AVG	0.099238
BeatsPerMin_MAX	0.183999
BeatsPerMin_MIN	0.241980

- Her fokuseres der på henholdsvis SP02, vejtrækning og puls.
- Disse bliver løbende sendt hvert minut i alle døgnets timer, under forudsætning af at uret sidder korrekt på armen. Altså kan der være op til 1.440 målinger pr. døgn.
- I denne tabel er der data aggregeret på 15 minutters intervaller, når AVG er anført.
- Det overraskende resultat i denne sammenhæng er, at iltmætningen fremstår som den mindst betydende faktor, sammenlignet med henholdsvis vejtrækning og puls.

Hvilke forudsætninger har vi sat for de videre analyser herefter:



Afsæt i farvekoder fra KOL skema



Sammenhængede events



Tidsperiode -48 timer



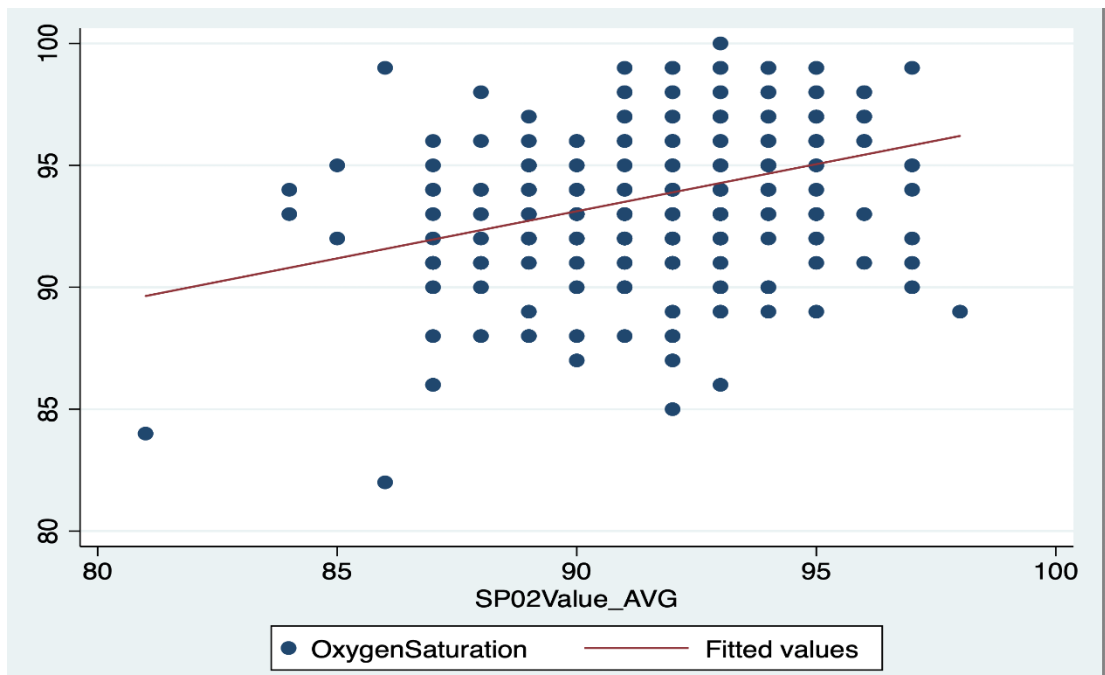
Sammenhængende målinger



Fokus på tiden før events

- Da antallet af påbegyndte exacerbationsbehandlinger var ret lavt, blev det i stedet aftalt at fokusere på farvemarkeringer (grøn, gul, rød) fra patienternes KOL-spørgeskema.
- Herefter blev der opstillet en forudsætning om, at events af enten gul eller rød karakter ikke kunne stå alene – altså skulle der være minimum to på hinanden røde eller tre gule målinger, før vi karakteriserede det som et event.
- Med afsæt i dette event summerede vi data op på de to foregående døgn, for at undersøge om der var noget særligt ved dagene op til et event. For at sammenligne udvalgte vi sammenhængende grønne perioder, som kunne agere sammenligningsgrundlag for de perioder, der ledte op til et event.
- Derfor kunne vi både ved events og kontroller udelukkende inkludere perioderne i analyserne, hvis der var grønne målinger. Eksempel; hvis der var et rødt event lørdag, ville dette kun bruges i analyserne, hvis der forelå grønne målinger både torsdag og fredag. Derfor ville tidsperioden blive ekskluderet såfremt en grøn måling ikke forelå enten torsdag eller fredag.

Hvordan er sammenhængen mellem Garmin målt SP02 og SP02 målingerne fra KOL-spørgeskemaet?



Source	SS	df	MS	Number of obs	=	248
Model	303.209346	1	303.209346	F(1, 246)	=	33.58
Residual	2220.96807	246	9.0283255	Prob > F	=	0.0000
				R-squared	=	0.1201
				Adj R-squared	=	0.1165
Total	2524.17742	247	10.2193418	Root MSE	=	3.0047

OxygenSatur~n	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
SP02Value_AVG	.3863968	.0666754	5.80	0.000	.2550693	.5177244
_cons	58.34189	6.129583	9.52	0.000	46.26873	70.41505

- Der er en sammenhæng mellem de to respektive SP02 målinger, men de er ikke perfekt korrelerede som anført i ovenstående scatterplot. Dette er umiddelbart ikke af betydning for udarbejdningen af en machine learning baseret prædiktionsmodel, men kan eventuelt forstås som en klinisk forklaring på den overraskende lave feature importance.
- Med afsæt i et statistisk perspektiv kan vi dog se at sammenhængen mellem de to målinger alt andet lige er statistisk signifikant.

Er der korrelation mellem puls, hjertefrekvens og SP02 fra Garmin uret?

```
. regress SP02Value_AVG BeatsPerMin_AVG
```

Source	SS	df	MS	Number of obs	=	109,090
Model	1403.6341	1	1403.6341	F(1, 109088)	=	101.03
Residual	1515533.8	109,088	13.8927637	Prob > F	=	0.0000
				R-squared	=	0.0009
				Adj R-squared	=	0.0009
Total	1516937.44	109,089	13.9055032	Root MSE	=	3.7273

SP02Value_AVG	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
BeatsPerMin_AVG	.0101885	.0010136	10.05	0.000	.0082018	.0121752
_cons	90.07554	.0751917	1197.94	0.000	89.92816	90.22291

- SP02 og puls: Statistisk signifikant positiv sammenhæng, hvor det estimeres at en stigning i puls på 1 vil resultere i en forøgelse af SP02 på 0.01.

```
. regress SP02Value_AVG BreathsPerMin_AVG
```

Source	SS	df	MS	Number of obs	=	106,905
Model	153854.063	1	153854.063	F(1, 106903)	=	12300.18
Residual	1337172.63	106,903	12.5082797	Prob > F	=	0.0000
				R-squared	=	0.1032
				Adj R-squared	=	0.1032
Total	1491026.69	106,904	13.9473424	Root MSE	=	3.5367

SP02Value_AVG	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
BreathsPerMin_AVG	-.4007555	.0036135	-110.91	0.000	-.4078379	-.3936732
_cons	97.38476	.0601408	1619.28	0.000	97.26689	97.50264

- SP02 og vejrtrækning: Igen statistisk signifikant sammenhæng. Denne gang dog en negativ sammenhæng, hvor det estimeres at en stigning i vejrtrækningsfrekvens på 1 vil medføre en reduktion i SP02 på 0.4. Altså at iltmætningen er dårligere, når vejrtrækningen stiger.

. regress BeatsPerMin_AVG BreathsPerMin_AVG

Source	SS	df	MS	Number of obs	=	106,893
Model	4108.28216	1	4108.28216	F(1, 106891)	=	33.41
Residual	13144039.2	106,891	122.966753	Prob > F	=	0.0000
				R-squared	=	0.0003
				Adj R-squared	=	0.0003
Total	13148147.5	106,892	123.004037	Root MSE	=	11.089

BeatsPerMin_AVG	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
BreathsPerMin_AVG	.0654986	.0113317	5.78	0.000	.0432886 .0877086
_cons	72.13788	.1886037	382.48	0.000	71.76822 72.50754

- Puls og vejtrækning: Igen en statistisk signifikant sammenhæng. Denne er også positiv, hvor det estimeres at en stigning i vejtrækningsfrekvens på 1 vil føre til en puls, der er 0.07 højere (meget lavt).

Herefter undersøger vi med en klassisk t-test, om der er forskel i vejtrækning i perioden op til et event og i perioden der ikke leder op til et event.

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
0	78	14.14344	.1674401	1.478791	13.81003 14.47686
1	76	14.59969	.1903615	1.659533	14.22047 14.97891
combined	154	14.36861	.1274871	1.582073	14.11674 14.62047
diff		-.4562492	.2531424		-.9563811 .0438826

diff = mean(0) - mean(1) t = -1.8023
 Ho: diff = 0 degrees of freedom = 152
 Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
 Pr(T < t) = 0.0367 Pr(|T| > |t|) = 0.0735 Pr(T > t) = 0.9633

- Her ser vi noget der er grænsende til statistisk signifikant (p=0.07). Det er dog værd at holde for øje, at dette alene er beregnet på cirka 150 tidsperioder, og at det muligvis kan opspores i tilfælde af større datasæt.
- Testen viser som forventet, at tendensen peger i retning af at der er højere vejtrækningsfrekvens op til eventet. Det kan muligvis give mening, da vi tidligere ser at SP02 værdien netop er lavere ved højere vejtrækningsfrekvens. OBS på kausalitet – hønen og ægget.

Fokus bliver nu igen på tiden før event, men bliver nu undersøgt på to andre måder – stadig med fokus på vejtrækning, da der ikke er statistisk forskel ved puls og SP02.

BreathsPerMin_AVG		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Time							
2		.2278671	.3954078	0.58	0.564	-.5471179	1.002852
3		.8310689	.3951719	2.10	0.035	.0565462	1.605592
1.Event		1.065503	.4116987	2.59	0.010	.2585884	1.872418
Time#Event							
2 1		-.6422893	.568162	-1.13	0.258	-1.755866	.4712879
3 1		-1.389111	.5620245	-2.47	0.013	-2.490658	-.287563
_cons		13.70005	.3231739	42.39	0.000	13.06664	14.33345

- Denne analyse er en mixed model, hvor tidsperioderne bliver delt i henholdsvis 1) grønne perioder og 2) perioder op til event.
- Den rødmarkerede række viser at der er statistisk signifikant højere vejrtækningsfrekvens i perioden op til event.

Vi holder lige fast i det fund og graver videre

. logistic State3 c.BreathsPerMin_AVG1

logistic regression	Number of obs	=	50
	LR chi2(1)	=	8.94
	Prob > chi2	=	0.0028
_log likelihood = -30.146712	Pseudo R2	=	0.1291

State3	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
BreathsPerMin_AVG1	2.104671	.6593559	2.38	0.018	1.138985	3.889112
_cons	.0000224	.0001002	-2.39	0.017	3.51e-09	.1431695

Note: _cons estimates baseline odds.

- Her viser en logistisk regression at risikoen for et event vil stige med en odds-ratio på 2.1 (CI: 1.4 – 3.9) for hver vejtrækningsforøgelse på 1 pr min i perioden fra 48-24 timer før event.

Forskel på Odds Ratio og RR.

	Event	Ikke event	Odds	OR	Risk	Relativ R
Høj vejtrækning	2	10	$2/10 = 0.2$	$0.2 / 0.09$	$2/12 = 0.17$	$0.17 / 0.08$
Lav	1	11	$1/11 = 0.09$	$= 2.22$	$1/12 = 0.08$	$= 2.125$

Den generelle opsummering med det mest essentielle

1. Det ønskes på sigt at oparbejde en machine learning baseret prædiktionsmodel, der kan prædiktere KOL-relaterede events, før de opstår.
2. Vi har fået afklaret at Garmin-urene opsamler data og sender tilfredsstillende videre til oparbejdet SQL-database. Dermed kan vi sige at opsamlingen af data fungerer.
3. Præliminære Machine Learning analyser indikerer at både Garmin puls- og vejtrækningsdata bærer større prædiktiv værdi end Garmin SP02 data.
4. Præliminære analyser med afsæt i statusfarvekoder fra KOL-spørgeskema indikerer at vejtrækningsfrekvens i tiden op til event kan være af særlig stor betydning.
5. Vi har på nuværende tidspunkt ikke data nok til at kunne træne og teste en machine learning baseret prædiktionsmodel. Dog giver vores præliminære analyser gode forhåbninger for, at det data der bliver opsamlet og videresendt fra Garmin-uret kan have prædiktiv værdi.

Bilag 2. Pilotafrøvning af Oura-ringe – vurdering af datakvalitet, modellering og generelle overvejelser

1. Formål

Formålet med denne rapport er at beskrive datagrundlaget for smart device pilotafrøvningen i PreCare Klinikken. 10 borgere fra PreCare-projektet er udvalgt til at deltage i pilotprojektet, der har til formål at afsøge om Oura-ringe kan erstatte eller supplere den egenmonitorering borgerne i PreCare-projektet i dag foretager. Et af inklusionskriterierne i PreCare-projektet er, at borgeren selv er i stand til at udføre egenmonitorering, herunder lungefunktionsmåling, måling af iltmætning og temperatur samt besvarelse af 3 spørgsmål. Det betyder, at borgere med kognitive udfordringer, der gør at de ikke kan anvende de udleverede devices til egenmonitorering, ekskluderes fra projektet på trods af, at de har KOL. Pilotprojektet har overordnet til formål at:

- Afprøve om Oura-ringe kan anvendes til at forudsige akutte forværringer med henblik på at inkludere borgere, som kognitivt er udfordret i forhold til at gennemføre egenmålinger
- Afprøve om måling af fysiske aktivitet/søvn mønstre kan være med til at forudsige forværringer i habitustilstanden.

Denne rapport beskriver datakvaliteten for det data, der er opsamlet via Oura-ringene på de 10 udvalgte borgere, med henblik på at vurdere, om datakvaliteten er tilstrækkelig til potentielt at kunne anvende disse til at forudsige akutte forværringer i borgernes tilstand. Rapporten indeholder desuden en sammenligning af udvalgte variable fra Oura-ringene og borgerens egenmålinger.

2. Beskrivelse af Oura-data

De 10 udvalgte borgere fik udleveret og påsat Oura-ringen primo 2023, og data er opsamlet indtil ultimo juni 2023. Oura-ringen opsamler en lang række variable, hvoraf nogle er udvalgt som potentielle prædiktorer (se tabel 2).

Oura-ringene opsamler data fra borgeren, når denne har ringen på, og opsamler derfor ikke data, når ringen oplades, eller af anden årsag ikke sidder på fingeren. Ringene opsamler information om, hvor mange timer dagligt borgeren ikke bærer ringen. I tabel 1 ses et overblik over hvor mange dataobservationer, der er opsamlet for hver borger, samt hvor mange dage/timer borgeren ikke har haft ringen på. En dataobservation svarer til én dag.

Som det fremgår af tabel 1, varierer antallet af dage, hvor borgeren tager ringen af i en periode, samt den gennemsnitlige varighed af disse perioder. Nogle borgere tager sjældent ringen af og kun i kortere perioder, mens andre tager ringen af oftere og i længere perioder. Det må forventes, at den opsamlede data samt potentielt kvaliteten heraf påvirkes af, hvor meget borgeren har ringen på. Eksempelvis vil de variable, der fortæller noget om borgerens aktivitetsniveau på en given dag, formegentlig være påvirket i en negativ retning, hvis borgeren har haft en periode uden ringen på.

Tabel 2. Overblik over non-wear tid fordelt på de 10 borgere

Borger	Start	Slut	Antal dage	Dage u. data	Dage m. non-wear	Total non-wear (timer)	Middel non-wear (timer)
1	23-01-30	23-06-26	148	0	32	88	2.7
2	23-01-31	23-06-24	145	13	40	230	5.7
3	23-01-30	23-06-26	148	0	33	48	1.4
4	23-01-24	23-06-23	151	1	43	155	3.6
5	23-02-08	23-06-22	135	68	26	85	3.3
6	23-01-30	23-06-25	147	2	63	79	1.3
7	23-01-31	23-06-10	131	1	32	159	5.0
8	23-01-24	23-06-23	151	0	140	223	1.6
9	23-01-31	23-06-25	146	0	43	91	2.1
10	23-01-31	23-06-01	122	0	27	43	1.6

Note: Start og slut indikerer dato for henholdsvis første og sidste dato, hvor data er opsamlet. Antal dage er antallet af unikke datoer mellem start og slut, mens dage u. data, indikerer hvor mange af de pågældende dage, hvor der ikke er opsamlet data. Dage med non-wear tid er antallet af dage, hvor borgeren i en periode (på under 24 timer) ikke har haft ringen på, mens total non-wear tid er det totale antal timer, hvor borgeren ikke har haft ringen på, og middel non-wear tid er den gennemsnitlige non-wear tid for de pågældende dage med non-wear tid (de dage, hvor der slet ikke er opsamlet data er ikke medtaget i beregningen af antallet af gennemsnitlig og total non-wear tid).

2.1 Udvælgelse af attributter

Oura-ringene opsamler en lang række variable, nogle variable er rådata, og nogle er afledt på baggrund af andre variable. For at begrænse mængden af potentiel støj i en modellering, har vi gennemgået listen af variable, og udvalgt nogle variable som potentielle prædiktorer. En beskrivelse af de udvalgte variable kan ses i tabel 2.

2.2 Datakvalitet for udvalgte attributter

Kvaliteten af de udvalgte attributter vurderes for at evaluere pålideligheden og fuldstændigheden af data.

2.2.1 Data fuldstændighed

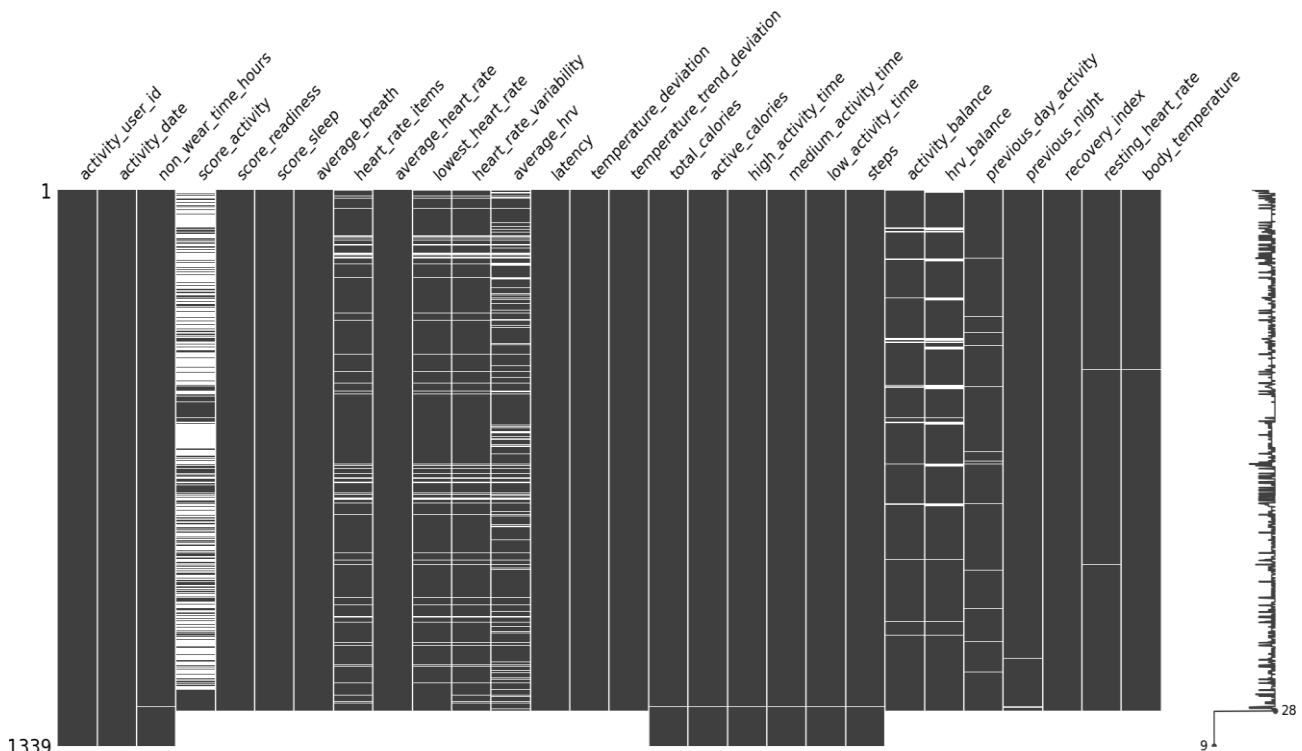
Data fuldstændighed indikerer andelen af manglende data i datasættet og for de enkelte observationer. Som angivet i tabel 2, er der en række variable, der har værdien 0, når data ikke er tilgængeligt. For de pågældende variable, er alle 0-værdier ændret til NaN-værdier (Not a Number), for at kunne undersøge den samlede andel af manglende data.

Figur 1 viser et matrixplot, der giver et overblik over manglende data (NaN- værdier). Y-aksen repræsenterer hver observation, mens de udvalgte variable er repræsenteret på x-aksen. Når data er til stede for en given variabel, er kolonnen skraveret, mens den er hvid, såfremt data mangler. Ud fra figuren kan vi se, at en stor del af observationerne har manglende værdier i mindst én variabel. 904 ud af 1.339 observationer har manglende værdier i mindst én variabel, svarende til 67,5%.

Figuren viser desuden, at der er en korrelation mellem nogle af variablene i forhold til manglende data, da en del af observationerne (85 observationer) kun har data i hhv. 9 variable ud af de 25 variable vist i figuren. Vi kan også se, at variabelen *activity score* mangler for en stor del af observationerne (64%). Procentdelen af manglende datapunkter for hver variabel er angivet i tabel 3.

Tabel 3. Overblik og beskrivelse af udvalgte variable fra Oura-ringene

Variabel	Beskrivelse
Non -wear Time	Number of hours during the day when the user was not wearing the ring.
Readiness Score	Readiness tells how ready you are for the day. It is calculated as a weighted average of readiness score contributors. Range: 1-100, or 0 if not available.
Activity Score	A measure of recent physical activity. It is calculated as a weighted average of activity score contributors. Range: 1-100, or 0 if not available.
Sleep Score	Overall sleep quality during the sleep period. It is calculated as a weighted average of sleep score contributors. Range: 1-100, or 0 if not available.
Average Breath	Average respiratory rate in breaths per minute registered during the sleep period.
Heart Rate Items	A list of detected heart rate in 5-minute intervals during the sleep period.
Average Heart Rate	The average heart rate registered during the sleep period.
Lowest Heart Rate	Lowest detected hearts rate registered during the sleep period.
Heart Rate Variability	A list of detected heart rate variability in 5-minute intervals during the sleep period.
Average HRV	Average Heart rate variability registered during the sleep period..
HRV Balance	Contribution of heart rate variability balance to the Readiness score. Range: 1-100, or 0 if not available.
Resting Heart Rate	Contribution of resting heart rate to the Readiness score. Range: 1-100, or 0 if not available.
Latency	Detected latency from bedtime to the beginning of the first five minutes of persistent sleep in secs.
Body Temperature	Contribution of Skin temperature to the Readiness score
Temperature Deviation	Skin temperature deviation from the long-term temperature average in Celcius.
Temperature Trend Deviation	Skin temperature deviation from the temperature trend.
Total Calories	Total energy consumption during the day including Basal Metabolic Rate in kilocalories.
Active Calories	Energy consumption caused by the physical activity of the day in kilocalories.
High Activity Time	Number of minutes during the day with high intensity activity (e.g. running).
Medium Activity Time	Number of minutes during the day with medium intensity activity (e.g. walking).
Low Activity Time	Number of minutes during the day with low intensity activity (e.g. household work).
Steps	Number of steps detected during the day.
Activity Balance	Measures a person's balance between physical activity and rest. Range: 1-100, or 0 if not available.
Previous Day Activity	Contribution of heart rate variability balance to Readiness score. Range: 1-100, or 0 if not available.
Previous Night	Contribution of a person's sleep from the previous night to the Readiness score. Range: 1-100, or 0 if not available.
Recovery Index	Contribution of recovery Index to the Readness score. Recovery index represents the amount of time you've slept after your resting heart rate (RHR) stabilizes during the night. Range: 1-100, or 0 if not available.

Figur 2. Matrix plot med overblik over manglende data for alle observationer

Note: De enkelte observationer er repræsenteret på y-aksen, mens hver variabel findes på x-aksen. Når der er data til stede for en given variabel, er kolonnen skraveret med grå, mens kolonnen er hvid, såfremt data mangler for den pågældende variabel.

2.2.2 Data pålidelighed

Data pålidelighed fortæller noget om, hvorvidt de målte værdier er korrekte. Tabel 3 viser gennemsnittet, standardafvigelsen, medianen, lavest og højest målte værdi for hver variabel. Fordelingen af datapunkterne for udvalgte variable er også illustreret som box-plot diagrammer i figur 2. Hvert box-plot illustrerer hvordan observationerne er fordelt for den pågældende variabel. Kassen illustrerer i hvilket område 50% af observationerne ligger, og den nederste og øverste kant i kassen repræsenterer derfor den nedre og øvre kvartil hhv. antennerne under og over kassen illustrerer variabiliteten af data omkring den nedre og øvre kvartil, og observationerne, der er plottet som individuelle punkter udenfor antennerne, er derfor observationer, der varierer signifikant fra resten af datasættet.

Ud fra de udvalgte boks-plots vist i figur 2 kan vi se, at mange af variablene har en del observationer, der ligger numerisk langt fra den overordnede datafordeling, hvilket kan indikere outliers i form af målefejl, eller det kan være et udtryk for korrekte målinger, der blot varierer signifikant fra normalområdet. For nogle variable er det nemmere at gennemskue, om et målepunkt i ekstrem-området er udtryk for en målefejl, eksempelvis *average heart rate*, *lowest heart rate*, *heart rate variability* og *average breath*, da udfaldsrummet for mulige værdier er rimelig veldefinerede. For andre variable er de maksimale grænseværdier mindre veldefinerede, hvorfor det kan være svært at identificere målefejl, herunder *latency*, *steps* og *activity time*.

Tabel 4. Statistik over variablerne

Name	Missing (%)	Mean	Std.	Median	Min.	Max.
non_wear_time_hours	0	0.9	2.4	0.0	0.0	19.32
score_activity	64	79.57	9.11	81.0	47.0	95.0
score_readiness	6	80.46	8.52	81.0	36.0	98.0
score_sleep	6	80.24	10.01	82.0	33.0	97.0
average_breath	6	17.72	1.38	17.5	14.75	21.625
heart_rate_items	14	N/A	N/A	N/A	N/A	N/A
average_heart_rate	6	74.06	7.62	73.25	56.875	108.75
lowest_heart_rate	14	71.93	14.72	70.0	51.0	255.0
heart_rate_variability	14	N/A	N/A	N/A	N/A	N/A
average_hrv	21	38.93	41.52	25.0	8.0	229.0
latency	6	506.15	529.66	360.0	0.0	4110.0
temperature_deviation	6	-0.01	0.33	0.01	-1.8	2.4
temperature_trend_deviation	6	0.12	0.24	0.11	-1.0	1.91
total_calories	0	2068.1	382.13	2018.0	1257.0	3291.0
active_calories	0	223.09	122.32	232.0	0.0	829.0
high_activity_time	0	134.09	287.86	60.0	0.0	3480.0
medium_activity_time	0	1637.68	1638.39	1200.0	0.0	13260.0
low_activity_time	0	14053.51	6498.37	14460.0	0.0	30600.0
steps	0	5575.96	3041.39	5546.0	0.0	21237.0
activity_balance*	9	84.11	9.69	85.0	49.0	100.0
hrv_balance*	13	79.83	14.52	85.0	1.0	100.0
previous_day_activity*	8	90.92	9.54	92.0	26.0	100.0
previous_night*	7	78.43	17.94	82.0	4.0	100.0
recovery_index*	6	72.76	28.86	82.0	1.0	100.0
resting_heart_rate*	6	79.32	22.96	86.0	1.0	100.0
body_temperature*	6	91.15	12.56	96.0	1.0	100.0

Note: Missing er den procentvise andel af manglende værdier, mean er gennemsnitsværdien, std. er standardafvigelsen, min. og max. er den laveste og højeste målte værdi hhv. *Bemærk værdien er udtryk for hvor meget variablen bidrager til beregningen af readiness score, og påtager en værdi mellem 1-100.

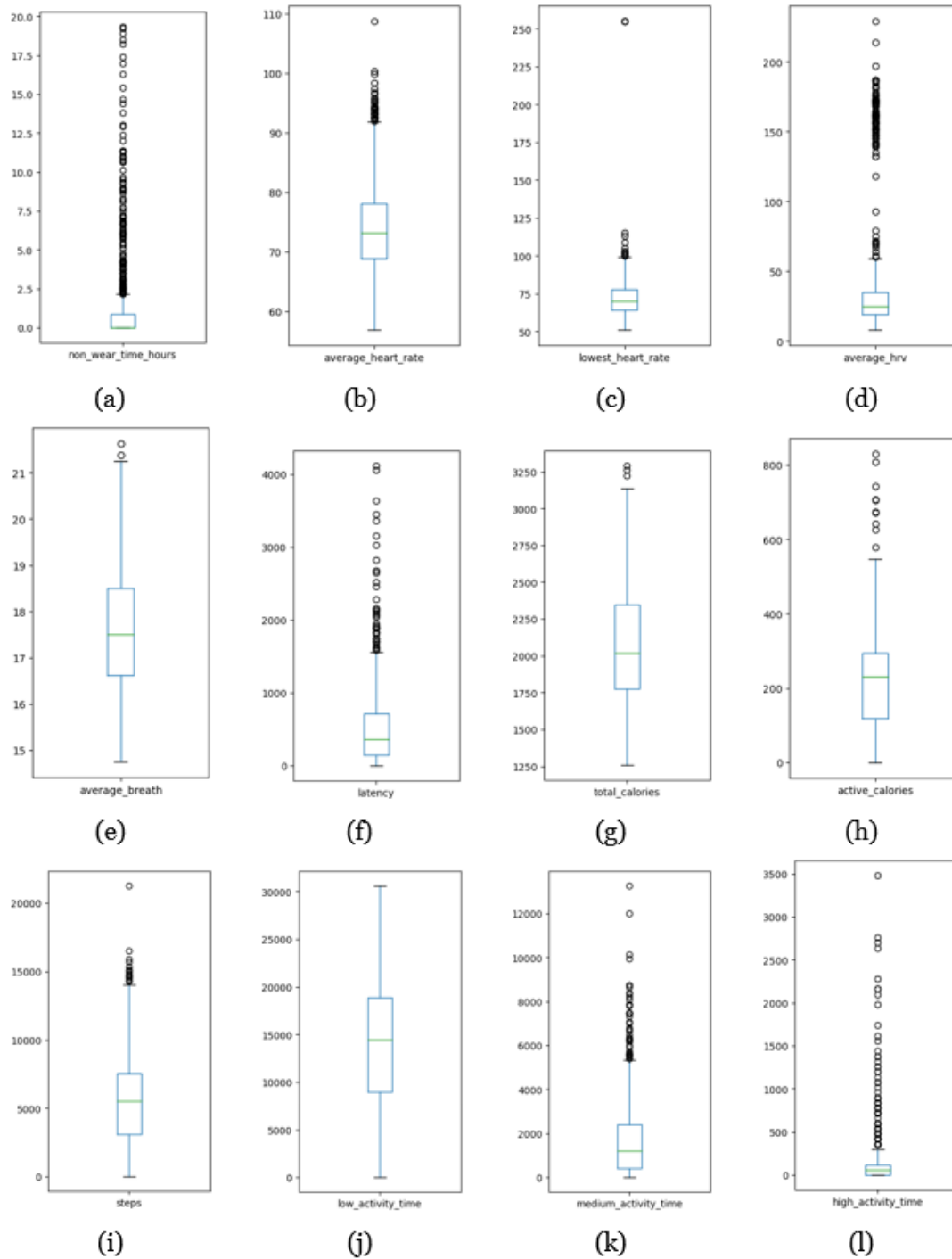
3. Sammenligning af Oura-data med Appinux-data

Da formålet med pilotprojektet er at undersøge, hvorvidt Oura-ringene kan erstatte eller supplere borgernes egenmålinger, er det interessant at sammenligne målingerne fra ringene med. egenmålingerne. For 1069 af Oura-datapunkterne findes der tilhørende egenmålinger (Appinux). Sammenligningen af data fra Oura-ringene med Appinux-data er lavet på dette udsnit af data. Borgernes egenmålinger omfatter temperatur, lungefunktion, puls og iltmætning. Korrelationen mellem egenmålingerne og målingerne fra Ouraringene (udvalgte variable) er vist i korrelations-matricen i figur 3.

3.1 Egenmålinger af lungefunktion og puls

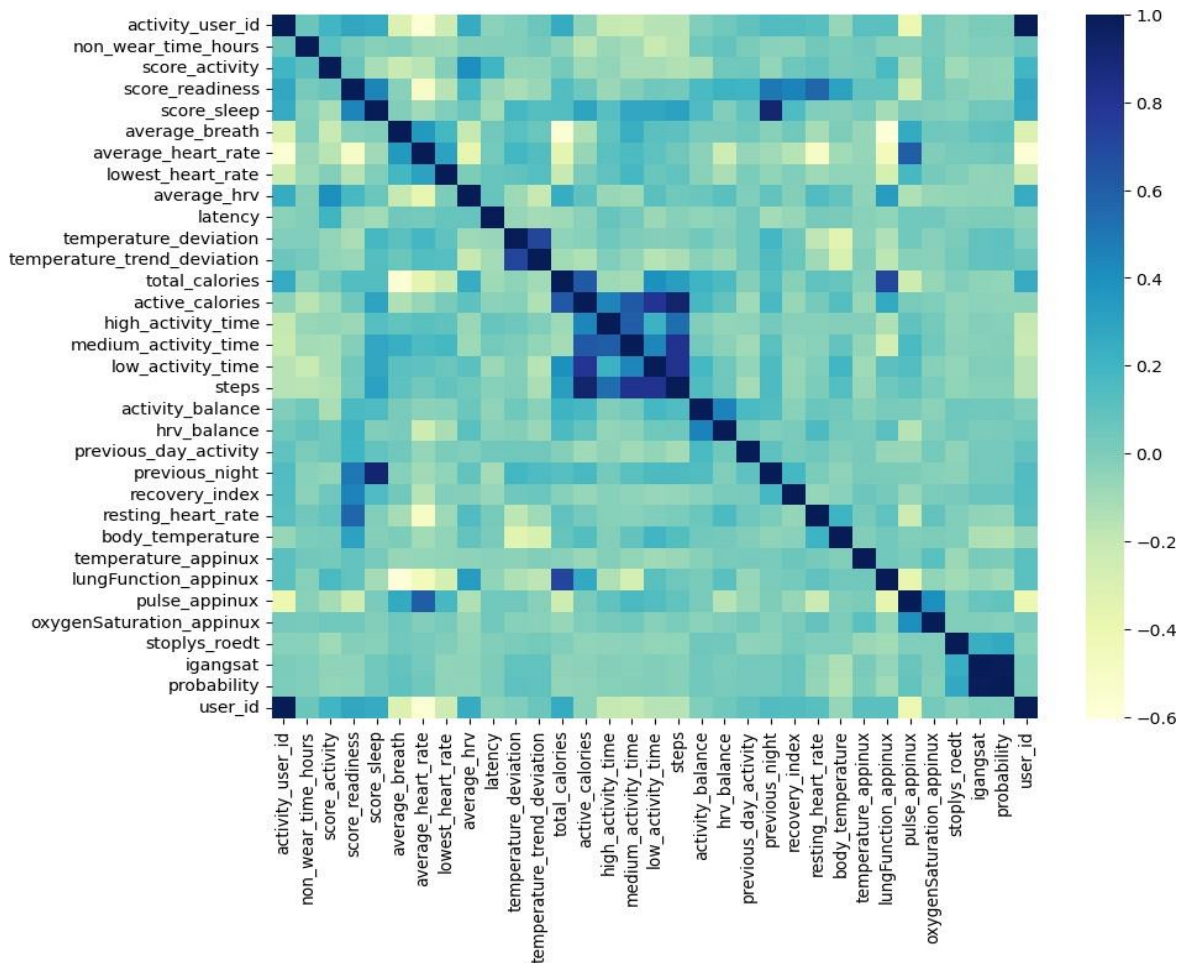
I korrelationsmatricen (figur 3) kan vi se, at borgernes egen lungefunktionsmåling (lungFunction appinux) ser ud til at korrelere negativt med den gennemsnitlige åndedræts-rate målt af Oura-ringene (average breath) med en korrelationskoefficient på -0,60. Lungefunktionsmålingerne og gennemsnitsåndedrætsraten er plottet i figur 4a for at illustrere korrelationen mellem de to variable.

Figur 3. Box-plots for udvalgte variable



Note: a) Non-wear time, b) Average heart rate, c) Lowest heart rate, d) Average heart rate variability, e) Average breath, f) Latency, g) Total calories, h) Active calories, i) Steps, j) Low activity time, j) Medium activity time, l) High activity time.

Figur 4. Korrelation mellem variable



Lungefunktionsmålingerne ser desuden ud til at korrelere positivt med den totale mængde forbrændte kalorier (*total_calories*) med en korrelationskoefficient på 0,71. Lungefunktionsmålingerne og den totale mængde forbrændte kalorier er ligeledes plottet for at visualisere korrelationen (fig. 4b).

Borgernes egne pulsmålinger (*pulse_appinux*) ser ud til at korrelere positivt med den gennemsnitlige hjertefrekvens målt af OURA-ringene (*average_heart_rate*) med en korrelationskoefficient på 0,59. Figur 4c illustrerer korrelationen mellem puls og den gennemsnitlige hjertefrekvens. Den egenmålte puls er generelt højere end pulsmålingerne fra Oura-ringene, hvilket er forventeligt, da borgernes foretager deres egenmålinger i vågen tilstand, mens Oura-ringene måler puls i løbet af natten.

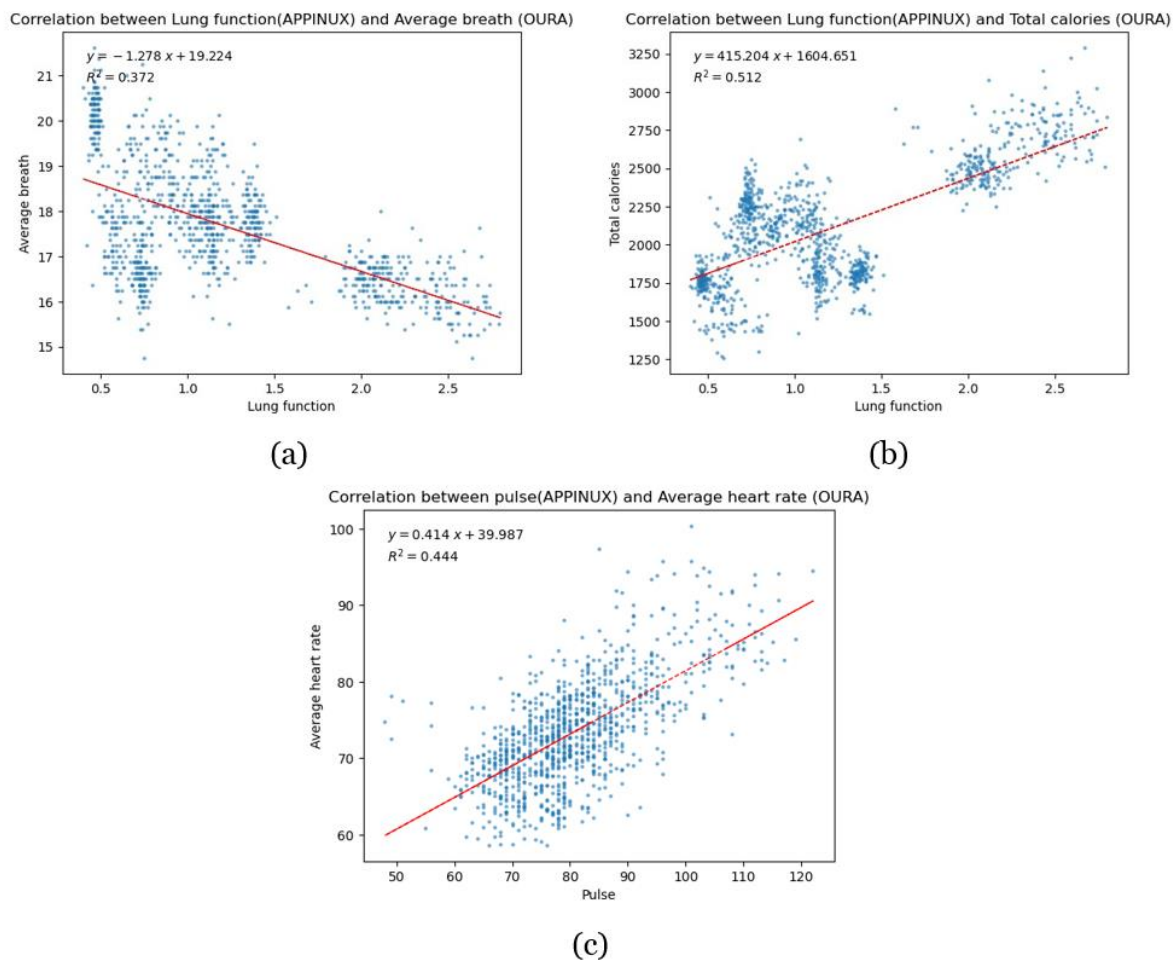
Figur 7 i bilag 1 viser ligeledes korrelationen mellem borgernes egenmålinger og målingerne fra Oura-ringene for de udvalgte parametre, der også er illustreret i figur 4. I figur 7 er punkterne illustreret i forskellige farver for de 10 borgere. På trods af, at vi ser en generel tilnærmelsesvis lineær korrelation mellem egenmålingerne og målingerne fra Oura-ringene, viser figur 7 at vi ser en grad af gruppering af borgernes målinger. Særligt i figur 7a (Lungefunktion versus gennemsnitlig ånde-drætsrate) samt figur 7b (lungefunktion versus totale mængde forbrændte kalorier) ses grupperingen tydeligt, mens grupperingen er mindre tydelig i figur 7c (puls versus gennemsnitlig hjertefrekvens). Figur 8 i bilag 2 viser korrelationen mellem den egenmålte puls og hjertefrekvensmålingen

fra Oura-ringen for hver borger separat, hvor det er tydeligt, at den lineære tendens i korrelationen er mindre signifikant med R^2 -værdier mellem 0,03 og 0,47.

Den egenmålte iltmætning (*oxygenSaturation_appinux*) samt temperatur (*temperature_appinux*) ser ikke ud til at have nogen signifikant korrelation med nogen af målingerne fra Oura-ringene.

Oura-ringene måler også temperatur, men vi kan ikke tilgå temperaturdataene i deres rå form, og har derfor kun variablene temperature deviation og temperature trend deviation som udtryk for den målte temperatur, hvilket ikke er direkte sammenligneligt med borgernes egne temperaturmålinger (sammenligning af disse med den egenmålte temperatur findes i næste afsnit).

Figur 5. Korrelation mellem udvalgte variable



Note: a) korrelation mellem Lung function og Average breath, b) korrelation mellem Lung function og total calories, c) korrelation mellem Pulse og Average heart rate.

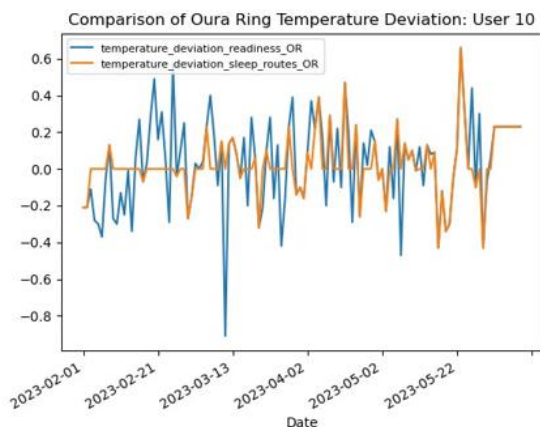
3.2 Egenmåling af temperatur

"Readiness" og "sleep routes" kategorierne har begge to målinger, som giver en indikation af den udvendige temperatur fra en person. Der er en generel afvigelse og en tendens afvigelse. Det vides ikke hvordan afvigelserne er beregnet, da Oura-ing ikke giver selve temperaturmålingen men kun afvigelsen baseret på en ukendt baseline.

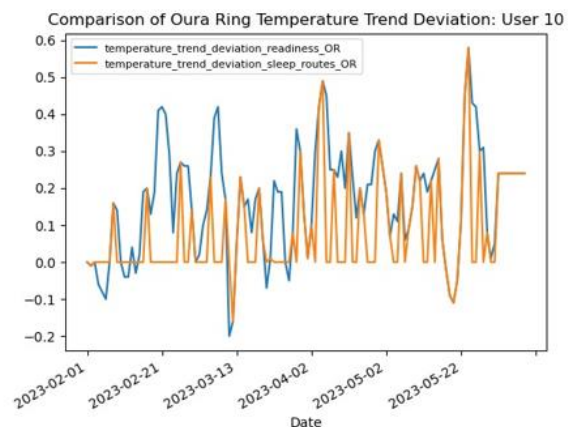
Der er kigget på datakvaliteten af temperatúrafgigelsen ved at dykke ned i dataene for de forskellige borgere og sammenligne med deres egenmåling (en indvendig temperaturmåling), som borgerne foretager dagligt. Første eksempel kommer fra "readiness" og "sleep route"s generelle temperatúrafgivelse, hvor der ses på figur 5a en sammenligning over tid på denne afvigelse for borger 10. Et problem med "sleep routes" målingen, den orange linje, er at der er mange manglende værdier. Dette er grundet to ting, 1) der er mange dage med en afvigelse på 0, da intet er registreret den dag, 2) der er ingen ændring i temperatúrafgigelsen. Disse to årsager ses ved en flad linje ved 0.0. Hvorvidt der faktisk ingen værdi er registreret eller der ingen daglig afvigelse er, vides ikke uden at kigge på værdierne i dataene.

Det andet eksempel ses på figur 5b med samme kategorier "readiness" og "sleep routes", men denne gang sammenlignes afvigelsen givet en *ukendt* tendens.

Figur 6. Temperatúrafgivelse og Temperatur Tendens afvigelse for kategorierne "readiness" og "sleep routes"



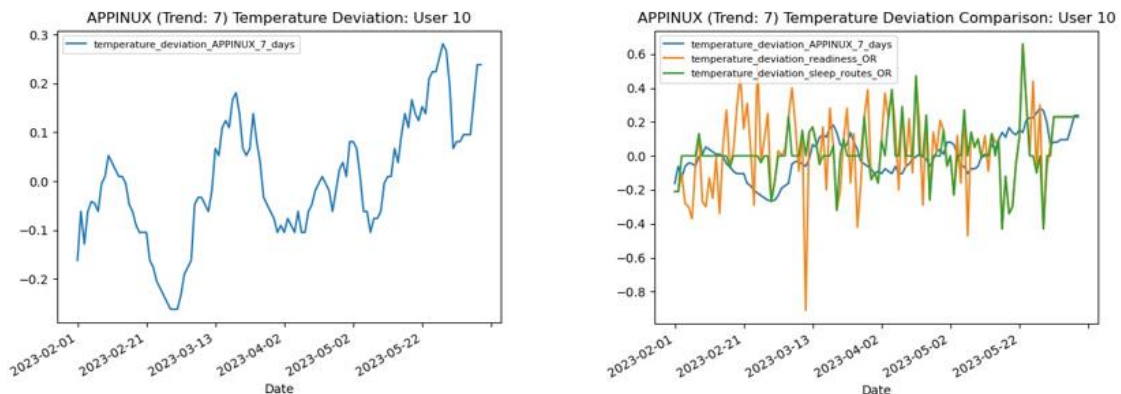
(a) Temperatur afvigelse fra "readiness" og "sleep routes" for borger 10.



(b) Temperatur trend afvigelse fra "readiness" og "sleep routes" for borger 10.

I dette tredje eksempel kigges der på egenmålingerne, og hvor der er lavet en "custom" afvigelse baseret på temperaturmålingerne for at kunne sammenligne egenmålingerne med Oura-ring og analysere kvaliteten af dataene. På figur 6a ses den isolerede afvigelse for temperatur egenmålingen. Figur 6b viser den isolerede afvigelse for temperatur egenmålingen sammen med "readiness" og "sleep routes" temperatúrafgivelse. Der kan ses på dette eksempel at Oura-ring afvigelse følger hinanden tættere ift. temperatúrafgivelse fra egenmålingen. Dette er en generel trend for alle borgere, og den lineære trend er ikke tilstedeværende eller meget svag på individuelt niveau.

Figur 7. Isoleret afvigelse for borger 10 givet Appinux- målingerne og en sammenligning mod "readiness" og "sleep routes" kategorierne



(a) Isoleret afvigelse fra APPINUX KOL-målingerne, udregnet over 7 dage.

(b) Sammenligning af APPINUX afvigelse mod "readiness" og "sleep routes"

Ydermere, for hver borger, er der kigget på hvilke to målinger ligger tættest på hinanden for at identificere, hvilke kategorier ligger tættest. Kategorierne som ligger i samme pulje er ikke sammenlignet, det vil sige at Temperatur Tendens Afvigelse "readiness" er ikke sammenlignet med Temperatur Afvigelse "readiness" og så videre. I Tabel 4 ses sammenligningen, hvor der kan identificeres at "readiness" og "sleep routes" ligger generelt tættere på hinanden end Appinuzmålingerne. Det noteres også at Appinux-målingerne og "sleep routes" for borger 5,6 og 8 er undtagelserne, men det tyder på at det er grundet mange afvigelser på 0 i "sleep routes" for disse borgere.

Tabel 5. Sammenligning af de tætteste afvigelser på tværs af kategorier for hver borger

Borger	Kategori 1	Kategori 2
1	Temperatur Tendens Readiness	Temperatur Tendens Sleep Routes
2	Temperatur Tendens Readiness	Temperatur Tendens Sleep Routes
3	Temperatur Tendens Readiness	Temperatur Tendens Sleep Routes
4	Temperatur Tendens Readiness	Temperatur Tendens Sleep Routes
5	Temperatur Tendens APPINUX	Temperatur Tendens Sleep Routes
6	Temperatur Tendens APPINUX	Temperatur Sleep Routes
7	Temperatur Tendens Readiness	Temperatur Tendens Sleep Routes
8	Temperatur Tendens APPINUX	Temperatur Tendens Sleep Routes
9	Temperatur Tendens Readiness	Temperatur Tendens Sleep Routes
10	Temperatur Tendens Readiness	Temperatur Tendens Sleep Routes

4. Modellering af data fra Oura-ringene

For at vurdere om data fra Oura-ringene har nogen prædiktiv kraft, er det nødvendigt at træne og evaluere en model på datasættet. Outcome af sådan en model kan være sandsynligheden for en igangsættelse af en borger baseret på data fra Oura-ringen. Udfordringen ved at træne sådan en model på det nuværende datasæt er, at der er meget få igangsættelser - altså det outcome vi er interesserede i at prædiktere. Der er i alt 13 igangsættelser hos de 10 borgere i perioden for dataindsamlingen, hvilket besværliggør modellering af data. Dog kan modellering muligvis give en indikation af den prædiktive værdi i det tilgængelige data.

Til vurdering af den prædiktive værdi har vi udvalgt en række variable som input variable til modellering:

- Sleep score
- Temperature deviation
- Total sleep
- Recovery index
- Resting heart rate
- Readiness score
- Latency
- HRV balance
- Previous night
- Previous day activity
- Activity balance
- Steps

Derudover har vi medtaget de 3 spørgeskema spørgsmål fra AppinuxX:

- Har dit opspyt ændret farve fra i går til at blive mere grønt eller grønligt? (Ja/nej)
- Har du mere åndenød i dag end i går? (Ja/nej)
- Hoster du mere i dag end i går? (Ja/nej)

4.1 Initiale eksperimenter

6 forskellige maskinlæringsmodeller er udvalgt til initiale eksperimenter:

- Logistic Regression
- Decision Tree (første test) / KNN (anden test)
- Random Forest
- Gradient Boosted Trees
- SVM
- Neural Network

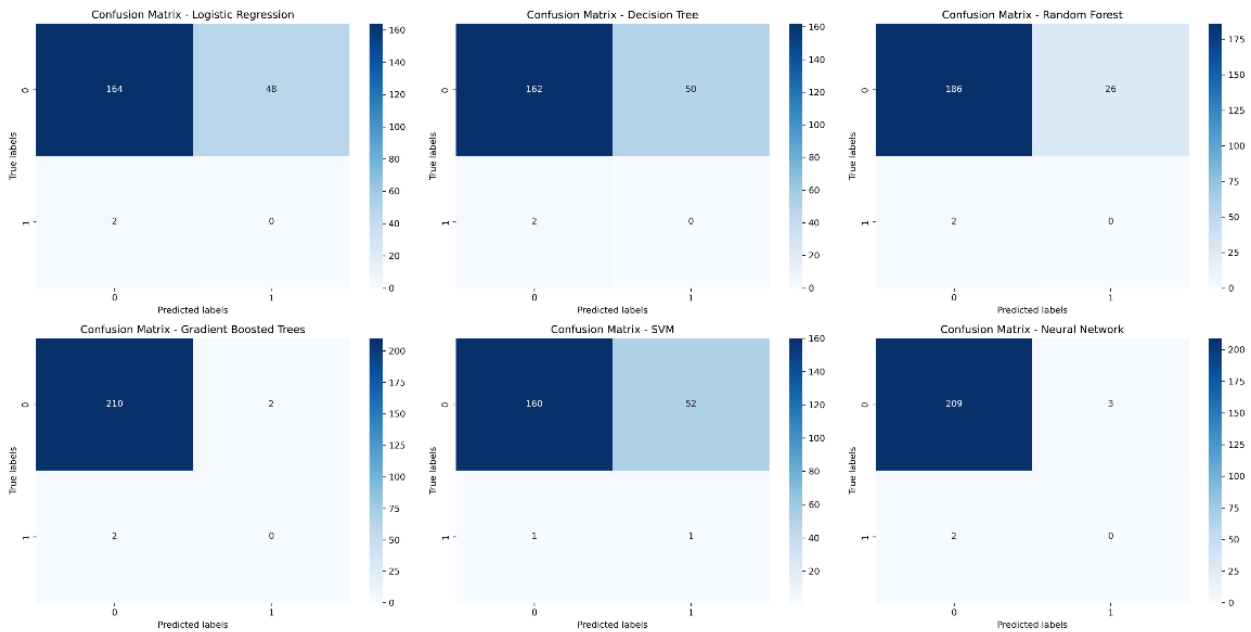
For hver af de ovenstående maskinlæringsmodeller, er modellen trænet på 60% af datasættet og evalueret på de resterende 40%. Modellerne er først trænet alene på inputvariablene fra Oura-ringene, og efterfølgende er de trænet på input variablene fra Oura-ringene samt de 3 spørgsmål fra Appinux for at evaluere effekten af at medtage disse.

4.1.1 Resultater af modeller uden de tre Appinux-spørgsmål

Herunder ses individuelle fejlmatrixer (figur 7) for de 6 modeller evalueret på testdatasættet.

Fejlmatrixerne i figur 7 viser antallet af korrekt og forkert klassificerede igangsættelser og ikke igangsættelser hhv. for testdatasættet. Testdatasættet indeholder kun 2 igangsættelser, og figur 7 viser, at de 6 modeltyper alle har svært ved at identificere disse. SVM identificerer én af de to igangsættelser mens de resterende modeller ikke identificerer nogen igangsættelser korrekt. Til gengæld klassificerer de 6 modeller mellem 2 og 52 ikke-igangsættelser som igangsættelser. Evalueringen af de 6 forskellige modeltyper viser altså, at modellerne har svært ved at lære de mønstre i træningsdata, der er karakteristiske for igangsættelserne, og derved ikke er i stand til at klassificere dem korrekt i testdatasættet.

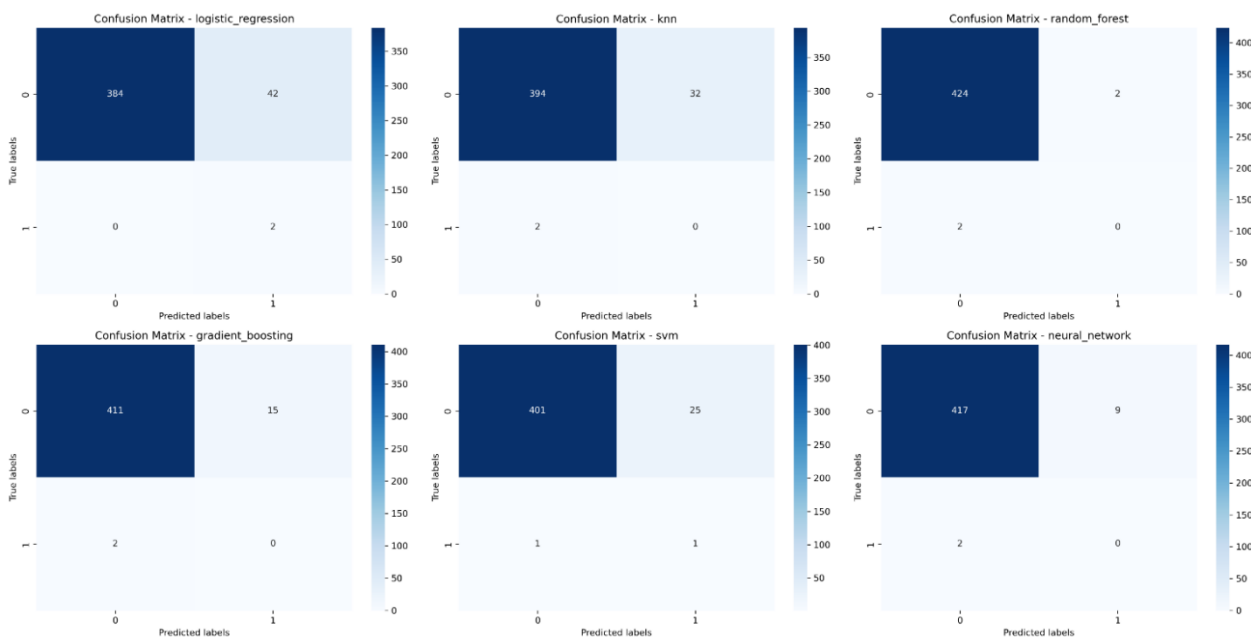
Figur 8. Fejlmatrixer for testdatasættet for de 6 udvalgte maskinlæringsmodeller, hvor input variablene består af Oura-data variablene alene



4.1.2 Resultater af modeller med de 3 Appinux-spørgsmål

Nedenfor ses individuelle fejlmatrixer for de 6 modeller ved evaluering på testdatasættet.

Figur 8. Fejlmatrixer for testdatasættet for de 6 udvalgte maskinlæringsmodeller, hvor input variablene består af både Oura-data variablene og de 3 Appinux-spørgsmål



Figur 8 viser antallet af korrekt og forkert klassificerede igangsættelser og ikke-igangsættelser hhv. for de 6 udvalgte modeller, når de 3 spørgsmål fra Appinux medtages som prædiktorer. Igen kan vi se, at de fleste af modellerne har svært ved korrekt at klassificere de to igangsættelser i testdatasættet. SVM-modellen klassificerer den ene af de to igangsættelser korrekt, logistic regression modellen klassificerer begge igangsættelser korrekt, mens de resterende modeller alle klassificerer de to igangsættelser som ikke-igangsættelser. Modellerne klassificerer ligeledes mellem 2 og 42 ikke-igangsættelser som igangsættelser.

Modellernes performance ser ud til at stige lidt, når de tre Appinux-spørgsmål medtages som prædiktorer sammenlignet med modellerne uden de tre spørgsmål. Da formålet med modelleringen er at kunne fange de igangsættelser borgerne har, udvælges logistic regression modellen til yderligere eksperimenter. Logistisk regressionsmodeller er ligeledes mulige at forklare sammenlignet med de andre modeltyper, der bærer mere præg af at være black-box modeller.

4.2 Logistisk regressions-modellering

Resultaterne af de initiale eksperimenter indikerede, at logistisk regressionsmodellen var bedst til at identificere igangsættelserne i testdatasættet, men da testdatasættet kun indeholder 2 igangsættelser, er resultatet af de initiale eksperimenter alene en indikator for modellens performance, som muligvis ikke er generaliserbar til et andet datasæt/en anden fordeling af trænings- og testdata. Derfor er denne ene modeltype udvalgt til yderligere eksperimenter, for at evaluere performance af modellen lidt mere omfattende.

Ud over at forsøge at klassificere målingerne i igangsæt eller ikke igangsæt, evalueres en binær logistisk regressionsmodel, hvor targetet er røde målinger (trafiklys), samt en multiklassemodel hvor målingerne klassificeres som grønne, gule eller røde målinger hhv. Evalueringen af modellerne med hver af de 3 valgte targets er beskrevet i de følgende underafsnit.

4.2.1 Binær klassifikationsmodel: igangsæt eller ikke-igangsæt

Da antallet af igangsættelser er meget lavt i forhold til det totale antal af observationer, anvendes *Leave-One-Out krydsvalidering*¹³ til evaluering af modellens generelle evne til at klassificere, hvor ét enkelt datapunkt fratages som testdatapunkt, der trænes en model på det resterende datasæt, som anvendes til at klassificere testdatapunktet. Efterfølgende fratages et andet testdatapunkt, der trænes en ny model på det resterende data, som anvendes til klassificering af det nye testdatapunkt. Denne procedure fortsættes indtil hvert enkelt datapunkt har fungeret som testdatapunkt. Altså trænes n modeller for et datasæt med n datapunkter, og modellens generelle performance defineres som antallet af korrekt klassificerede testdatapunkter.

De få igangsættelser resulterer i meget stor ubalance i datasættet, hvorfor der anvendes SMOTE upsampling¹⁴ af minoritetsklassen (igangsættelser) i træningsdatasættet, således at træningsdatasættet indeholder lige mange igangsættelser, som ikke-igangsættelser.

Tabel 5 viser modellens performance (præcision og sensitivitet) for testdata, når *Leave-One-Out* krydsvalidering anvendes, mens figur 9 viser fejl-matricen. Modellen klassificerer 5 af de 13 igangsættelser korrekt (svarende til en sensitivitet på 38%), mens 1196 ud af de 1314 ikke-igangsættelser

¹³ [sklearn.model_selection.LeaveOneOut — scikit-learn 1.3.1 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.LeaveOneOut.html)

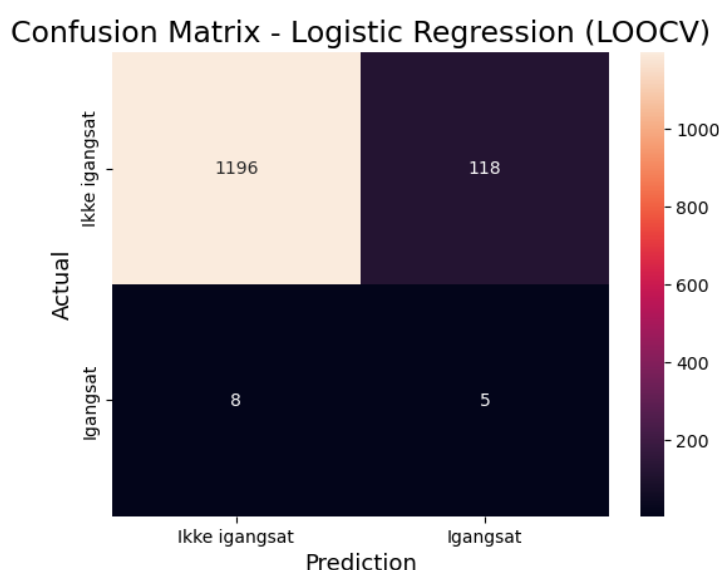
¹⁴ [SMOTE — Version 0.11.0 \(imbalanced-learn.org\)](https://imbalanced-learn.org/stable/transformations/minority.html)

klassificeres korrekt (svarende til en sensitivitet på 91%). Det betyder altså, at når modellen klassificerer en måling som en igangsættelse, vil kun 5% af disse være faktiske igangsættelser, mens 99% af de klassificerede ikke-igangsættelser er faktiske ikke-igangsættelser. Overordnet har modellen en nøjagtighed på 91%.

Tabel 6. Overblik over model performance for modellen ved LOOCV-evaluering fordelt på de 2 klasser (ikke-igangsat eller igangsat)

Klasse	Præcision	Sensitivitet	Datapunkter
Ikke-igangsat	99%	91%	1314 (99%)
Igangsat	4%	38%	13 (1%)

Figur 9. Fejlmatrixe for testdatapunkterne med modellen (Leave One Out krydsvalidering) hvor target er ikke-igangsat eller igangsat



På trods af at minoritetsgruppen er upsamlet i træningsdata, tyder resultaterne på, at modellen har svært ved at lære de mønstre der er karakteristiske for de målinger, der fører til en igangsættelse, hvilket er forventeligt når ubalancen i datasættet er så signifikant.

4.2.2 Binær klassifikationsmodel: Rød måling eller grøn/gul måling

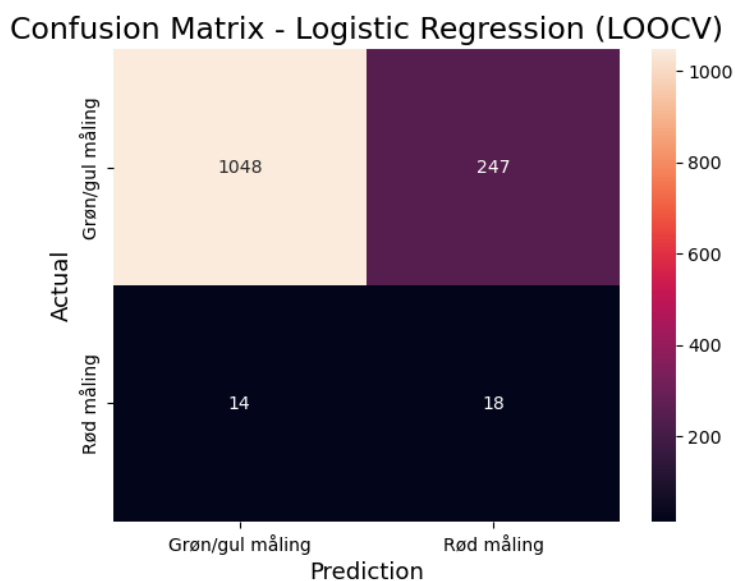
Target for klassifikationsmodellen defineres som røde målinger eller grønne og gule målinger (baseret på Appinix-trafiklysalgoritmen), og modelperformance evalueres igen med Leave-One-Out krydsvaliderings-metoden, hvor minoritetsklassen er upsamlet, således at træningsdata indeholder ligeså mange grønne/gule målinger som røde målinger.

Tabel 6 viser modellens performance (præcision og sensitivitet) for testdata, mens figur 10 viser fejlmatrixen. Modellen klassificerer 18 af de 32 røde målinger korrekt (svarende til en sensitivitet på 56%), mens 1048 ud af de 1222 grønne/gule målinger klassificeres korrekt (svarende til en sensitivitet på 81%). Det betyder altså, at når modellen klassificerer en måling som rød, vil kun 7% af disse være faktiske røde målinger, mens 99% af de klassificerede grønne/gule målinger er faktiske grønne/gule målinger. Overordnet har modellen en nøjagtighed på 80%.

Tabel 7. Overblik over model performance for modellen ved LOOCV-evaluering fordelt på de 2 klasser (grøn/gul eller rød måling)

Klasse	Præcision	Sensitivitet	Datapunkter
Grøn/gul måling	99%	81%	1295 (98%)
Rød måling	7%	56%	32 (2%)

Figur 10. Fejlmatrixe for testdatapunkterne med modellen (Leave-One-Out krydsvalidering) hvor target er grøn/gul eller rød måling



Igen observeres der, at på trods af upsampling af minoritetsgruppen i træningsdata, tyder resultaterne på, at modellen har svært ved at lære de mønstre der er karakteristiske for de røde målinger.

4.2.3 Multiklasse klassifikationsmodel: grøn måling, gul måling eller rød måling (trafiklys)

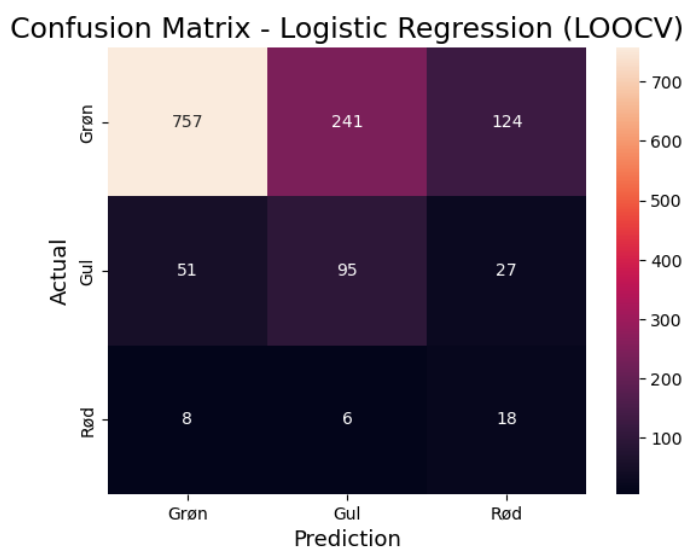
Endeligt defineres target for klassifikationsmodellen som henholdsvis grøn måling, gul måling eller rød måling (baseret på Appinix-trafiklysalgoritmen), og modelperformance evalueres på ny med Leave-One-Out krydsvalideringsmetoden, hvor minoritetsklasserne upsamples, således at træningsdata indeholder lige mange grønne, gule og røde målinger.

Tabel 7 viser modellens performance (præcision og sensitivitet) for testdata, mens figur 11 viser fejlmatrixen. Modellen klassificerer 18 af de 32 røde målinger korrekt (svarende til en sensitivitet på 56%), 95 ud af de 173 gule målinger klassificeres korrekt (svarende til en sensitivitet på 55%), mens 757 ud af de 1122 grønne målinger klassificeres korrekt (svarende til en sensitivitet på 67%). Det betyder altså, at når modellen klassificerer en måling som rød, vil kun 11% af disse være faktiske røde målinger, og hvis modellen klassificerer en måling som gul, vil 28% af disse målinger være faktiske gule målinger, mens 93% af de klassificerede grønne målinger er faktiske grønne målinger. Overordnet har modellen en nøjagtighed på 66%.

Tabel 8. Overblik over model performance for modellen ved LOOCV-evaluering fordelt på de 3 klasser (grøn, gul eller rød måling)

Klasse	Præcision	Sensitivitet	Datapunkter
Grøn måling	93%	67%	1122 (85%)
Gul måling	28%	55%	173 (13%)
Rød måling	11%	56%	32 (2%)

Figur 11. Fejlmatrice for testdatapunkterne med modellen (Leave-One-Out krydsvalidering) hvor target er grøn, gul eller rød måling



Evalueringen af multiklassemodellen tyder ligeledes på, at modellen har svært ved at skelne mellem de mønstre, der er karakteristiske for de grønne, gule og røde målinger henholdsvis, og mange af målingerne klassificeres forkert.

5. Opsummering og generelle refleksioner

5.1 Datagrundlaget

Det samlede datagrundlag består af 1339 observationer (dage) opsamlet fra 10 borgere i perioden 30-01-2023 til 26-06-2023. Alle borgerne har dage med non-wear tid, altså perioder hvor ringen ikke har været på, og derfor ikke har opsamlet data (kan skyldes opladning, forglemmelse eller andet). Den totale non-weartid varierer fra 43 til 230 timer, mens den gennemsnitlige non-wear tid (svarende til den gennemsnitlige længde af tidsperioden, hvor ringen ikke er på) ligger mellem 1,3 og 5,7 timer for de 10 borgere. Særligt 2 borgere skiller sig ud (borger 2 og 5), da disse har hhv. 13 og 68 dage hvor der ikke er opsamlet data.

Det må antages, at datakvaliteten af den opsamlede data bliver påvirket af non-wear tid - særligt for nogle variable. Eksempelvis de variable, der indikerer noget om borgerens aktivitetsniveau såsom antal skridt og antal forbrændte kalorier. Andre variable, der er baseret på et gennemsnit af de målte værdier over en længere periode, bliver formegentlig mindre påvirket af non-wear tid.

Non-wear tid kan dog også resultere i, at der mangler målinger for nogle variable. Eksempelvis hvis borgeren tager ringen af eller den løber tør for strøm om natten, opsamles data ikke, hvilket medfører at de variable, der er udledt af data fra søvnperioden ikke findes.

Pålideligheden af data fra Oura-ringene er svær at vurdere, da vi ikke har et direkte sammenligningsgrundlag. Enkelte variable forventes at korrelere med borgernes egenmålinger, herunder egenmålt lungefunktion med vejtrækningsfrekvens målt af Oura-ringene, samt borgernes egenmåling af puls og hjerterefrekvens målt af Oura-ringene. Sammenligningen af disse data fra ringene med borgernes egenmålinger viser også en korrelation, end ikke en perfekt korrelation. Dog er det vigtigt at huske, at borgernes egenmålinger ikke kan anses som en guld-standard, da disse også kan være fejlbetonede. Derfor er det svært at sige noget endegyldigt om pålideligheden af data fra Oura-ringene.

5.2 Modellering med data fra Oura-ringene

De indledende eksperimenter med forskellige modeltyper viser tydeligt udfordringen ved det ubalancerede datasæt med kun 13 igangsættelser ud af i alt 1327 målinger, da alle 6 testede modeltyper har svært ved at klassificere målingerne korrekt. Den logistisk regressionsmodel så ud til at performe en smule bedre end de resterende modeltyper, og sammenholdt med at denne modeltype har en stor grad af forklarbarhed, blev den valgt til yderligere eksperimenter. Dog skal det understreges at de 6 modeltyper blev evalueret på et testdatasæt med kun 2 igangsættelser, og derfor er den opnåede performance ikke nødvendigvis repræsenterende for et datasæt med flere/andre målinger, der har været ledsaget af en igangsættelse.

Leave-One-Out krydsvalidering blev anvendt for at give et mere generelt estimat af den logistiske regressionsmodels performance på alle målinger i datasættet, og 3 forskellige targets blev testet: igangsættelse/ikke-igangsættelse, rød måling/grøn eller gul måling, samt grøn måling/gul måling/rød måling. Uanset hvilket target klassifikationsmodellen har, indikerer eksperimenterne, at der er for stor ubalance i datasættet, til at en model kan trænes til at skelne tilstrækkeligt mellem klasserne. Udover at modellerne har svært ved at klassificere igangsættelser eller røde målinger korrekt (mange falske negative), klassificerer modellerne desuden mange ikke-igangsættelser eller grønne målinger som igangsættelser eller gule/røde målinger henholdsvis, og falsk-positiv raten er derfor høj. En stor andel af falske positive besværliggør arbejdet for de sygeplejersker, der skal tage stilling til borgernes målinger, hvorfor det er lige så vigtigt, at en model ikke resulterer i for mange falske positive, som at modellen er i stand til at finde alle igangsættelser/forværringer.

På baggrund af den afprøvede modellering, vurderes det, at det nuværende datagrundlag ikke er tilstrækkeligt til at træne en model med acceptabel performance, grundet den store ubalance i datasættet og de få igangsættelser. Mere data (flere igangsættelser eller forværringer i datasættet) vil formegentlig resultere i bedre modellering, men dette kan ikke siges med sikkerhed.

En alternativ tilgang til modellering af målingerne kunne være at bruge outlier detection metoder, såsom LOF i stedet, da der er så få igangsættelser/forværringer i datasættet. Tidsserie-modellering er også en mulighed, da Oura-ringene opsamler data fortløbende, hvorfor det potentielt er muligt, at finde mønstre i udviklingen af målepunkterne, der kan forudsige en forværring hos borgeren.

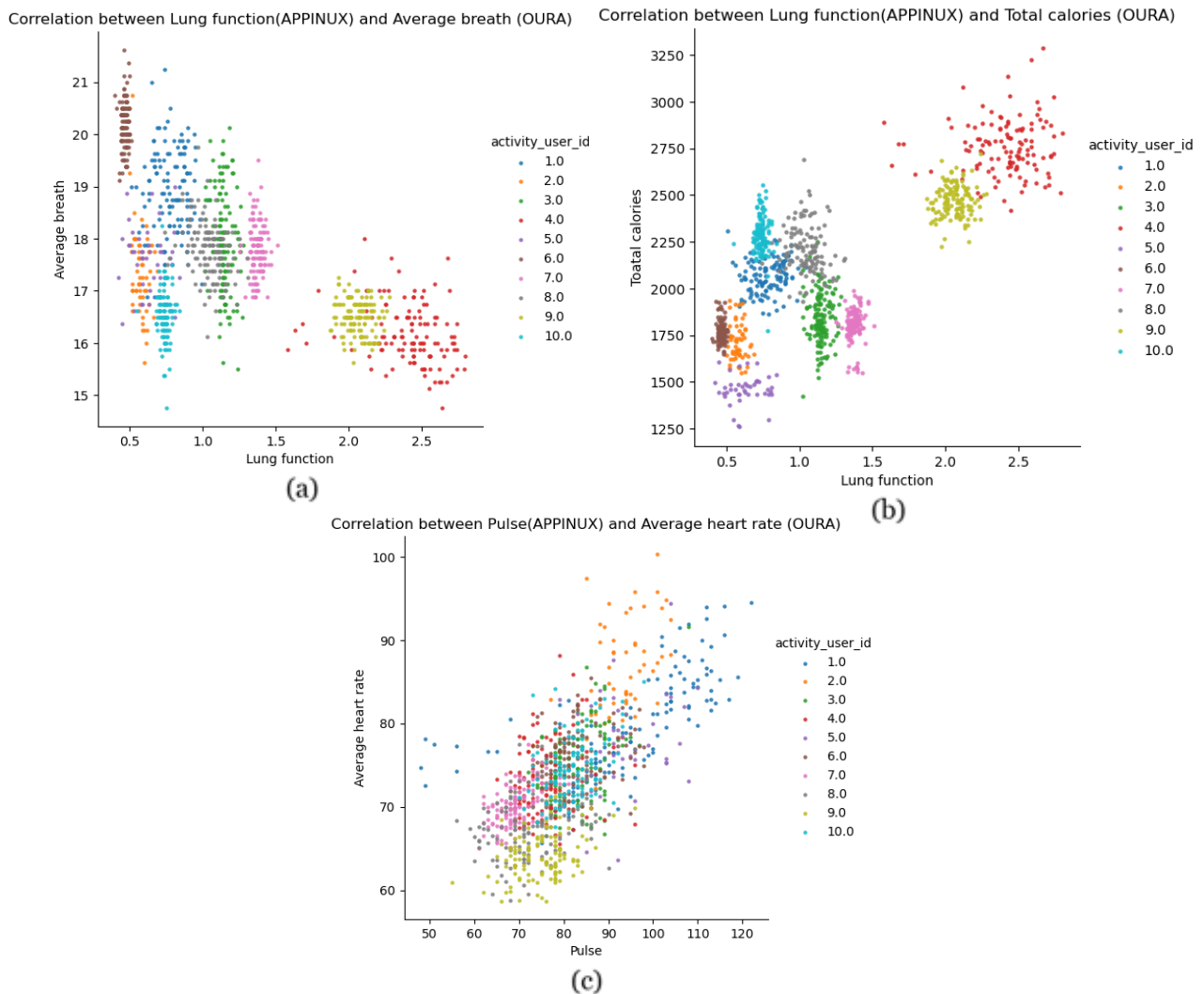
5.3 Praktisk anvendelse af ringene

Overvejelser:

- Hvis borgere bærer ringen, forventer de at blive monitoreret og føler derfor muligvis en falsk tryghed, hvis der er risiko for at data i perioder ikke opsamles (af forskellige årsager).
- Kvaliteten af det data der opsamles af ringen er betinget af, at ringen sidder korrekt på fingeren af borgeren. Flere ting kan medføre ændringer i tykkelsen af fingeren, herunder prednisolon behandling, hvorved ringen ikke længere passer optimalt, eller endda ikke kan anvendes af borgeren i en periode
- Data opsamles via en iPad, og overføres kun såfremt der er forbindelse til WIFI. Dette kan medføre at noget data går tabt, da hukommelsen i Oura-ringene er begrænset
- Data opsamles og opbevares i Ouras API, men her er lagerpladsen også begrænset, hvorfor noget data tilsyneladende slettes efter en periode. Dette gør sig gældende for puls-data, der opsamles i løbet af dagen (BPM), hvor data kun er tilgængeligt for de foregående to dage. Dette betyder, at man skal overføre data fra Ouras API til Databanken minimum hver anden dag, for at sikre at data ikke går tabt.

Bilag A. Korrelation mellem egenmålinger og målinger fra Oura-ringene

**Figur 12: Korrelation mellem egenmålt lungefunktion (Appinux) og gennemsnitlig ånde-
drætsrate (Oura)**

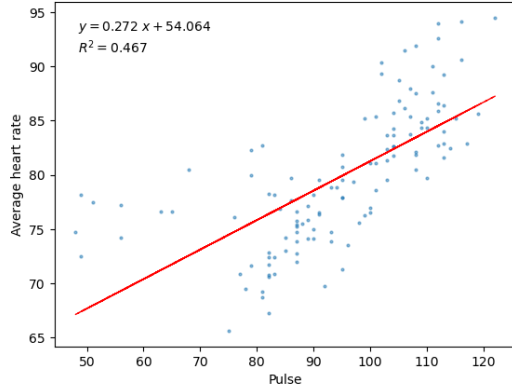


Note: (a), egenmålt lungefunktion (Appinux) og totale kalorier (Oura) (b), og egenmålt puls (Appinux) og gennemsnitlig hjerterefre-
kvens (Oura) (c) for alle 10 borgere. Hver borger er visualiseret med forskellige farver, for at adskille datapunkterne fra hinanden
baseret på hvilken borger målingen kommer fra.

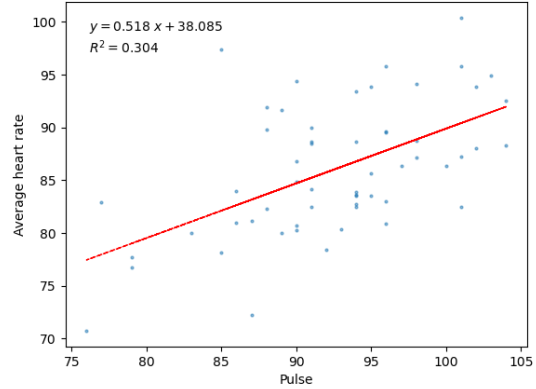
Bilag B. Korrelation mellem egenmålt puls og pulsmålinger fra Oura-ringe

Figur 13. Korrelation mellem egenmålt puls (Appinux) og gennemsnitlig hjertefrekvens (Oura) plottet for hver af de 10 borgere separat (figuren fortsætter på næste side)

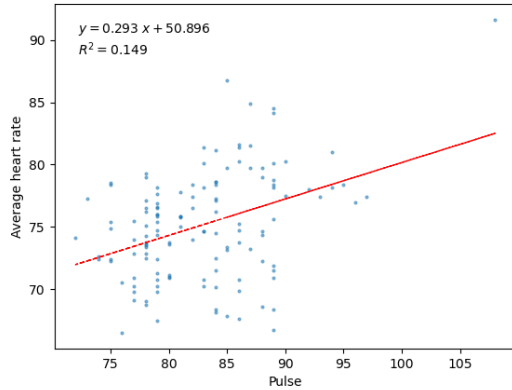
Correlation between pulse(APPINUX) and Average heart rate (OURA) for user 1



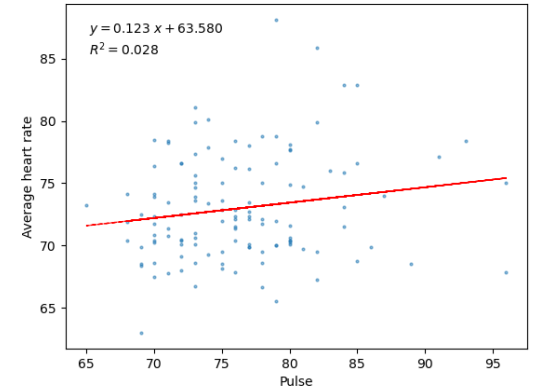
Correlation between pulse(APPINUX) and Average heart rate (OURA) for user 2



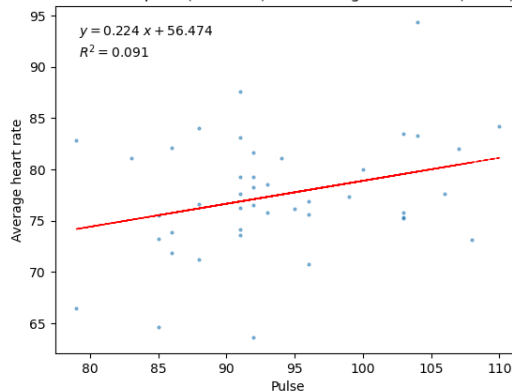
Correlation between pulse(APPINUX) and Average heart rate (OURA) for user 3



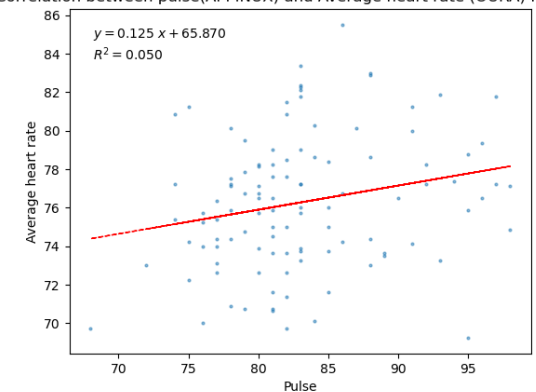
Correlation between pulse(APPINUX) and Average heart rate (OURA) for user 4



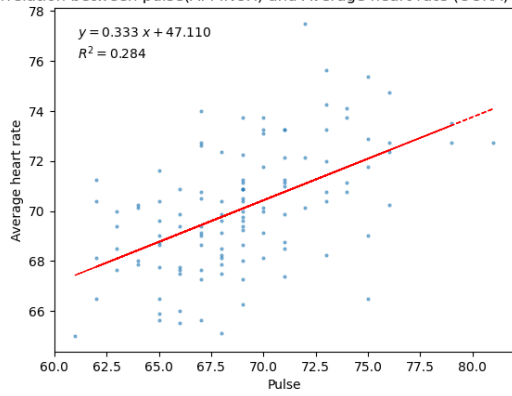
Correlation between pulse(APPINUX) and Average heart rate (OURA) for user 5



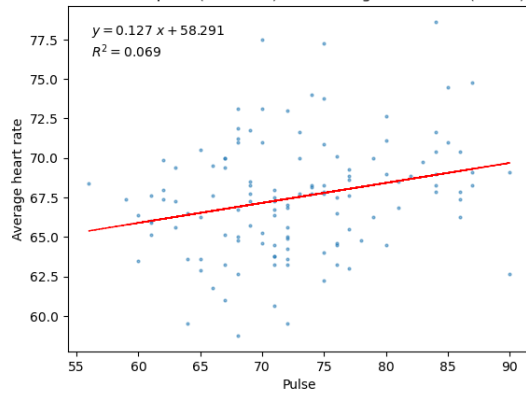
Correlation between pulse(APPINUX) and Average heart rate (OURA) for user 6



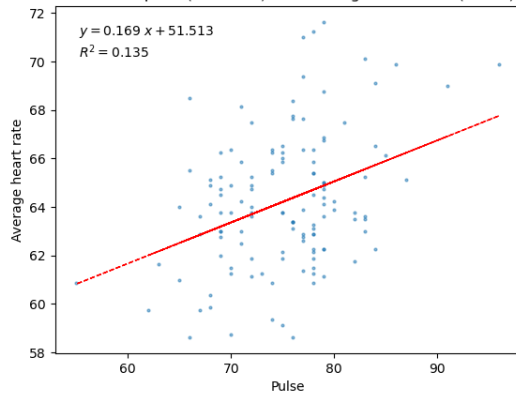
Correlation between pulse(APPINUX) and Average heart rate (OURA) for user 7



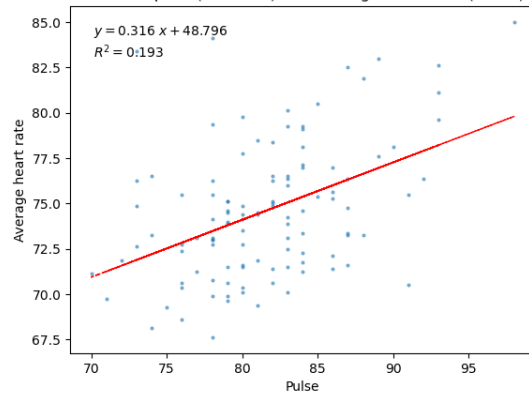
Correlation between pulse(APPINUX) and Average heart rate (OURA) for user 8



Correlation between pulse(APPINUX) and Average heart rate (OURA) for user 9



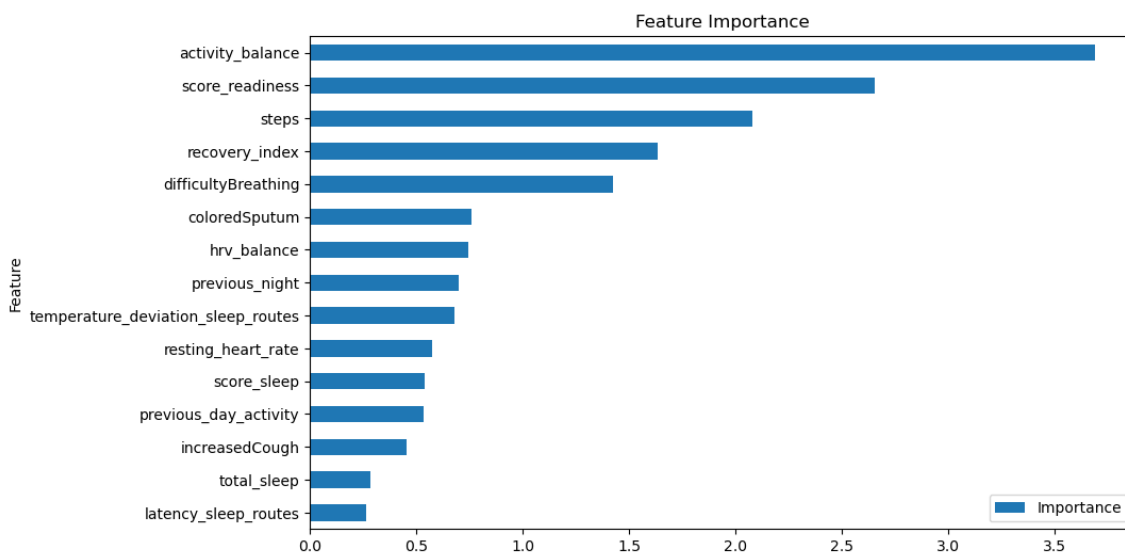
Correlation between pulse(APPINUX) and Average heart rate (OURA) for user 10



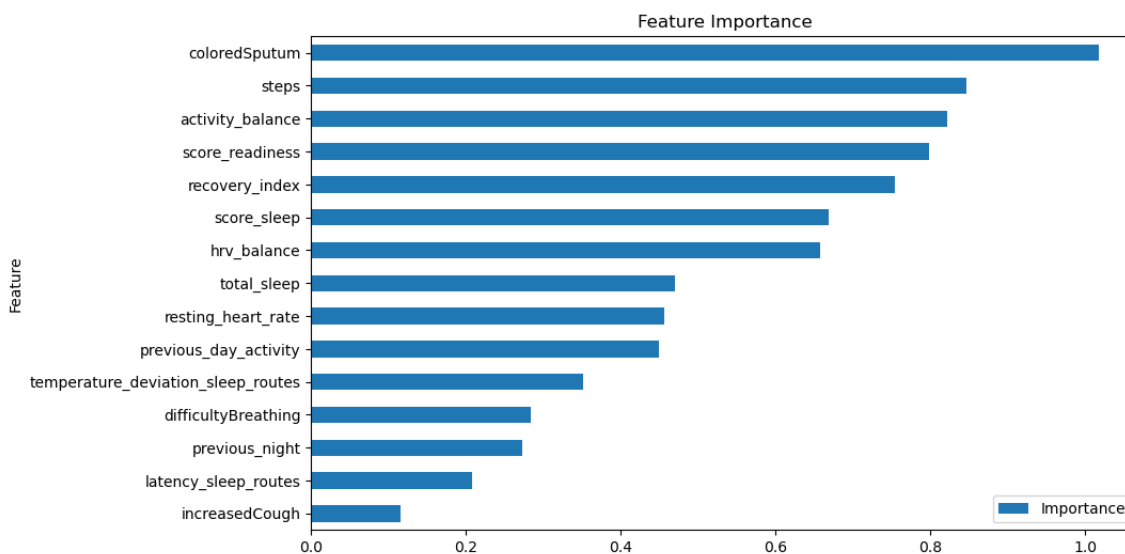
Bilag C. Feature Importance plots

Nedenfor ses feature importance plots for de 3 modeller beskrevet i afsnit 4.2. Hver plot viser vigtigheden af de anvendte prædiktorer i sorteret rækkefølge, således er de vigtigste prædiktorer vist øverst i plottet. For de binære modeller er der vist et enkelt feature importance plot, der beskriver vigtigheden af de prædiktorer, der er udslagsgivende for klasse 1 (igangsættelse eller rød måling hhv.). For multiklassemodellen, er der vist 4 plots: første plot er den generelle vigtighed af prædiktorerne i modellen, og derudover er der vist 3 plots, der beskriver vigtigheden af de prædiktorer, der er udslagsgivende for hver enkelt klasse (grøn, gul eller rød måling).

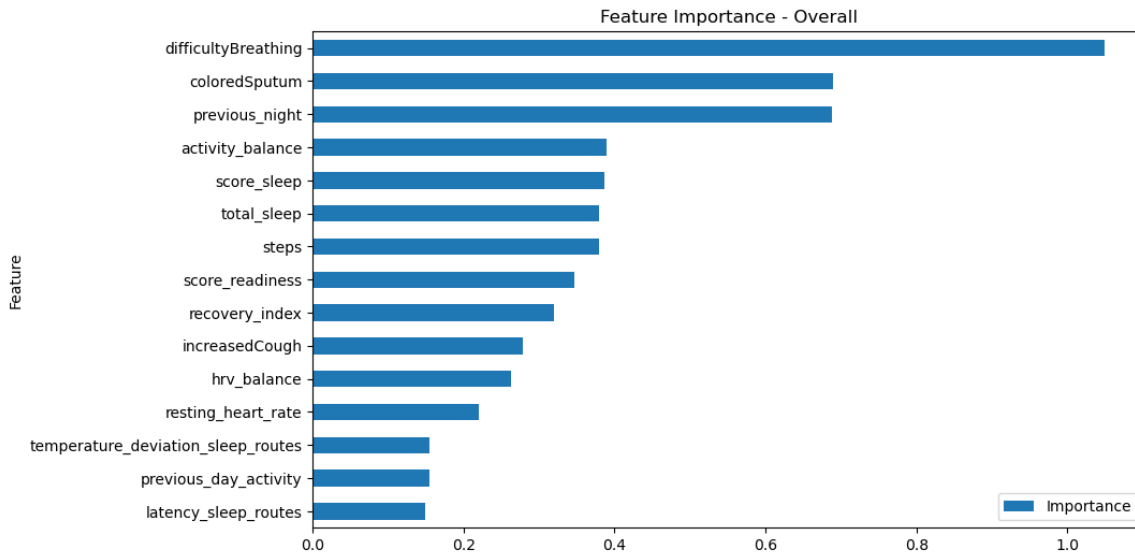
Figur 14. Feature Importance plot for den binære logistiske regressionsmodel med igangsat/ikke igangsat som target



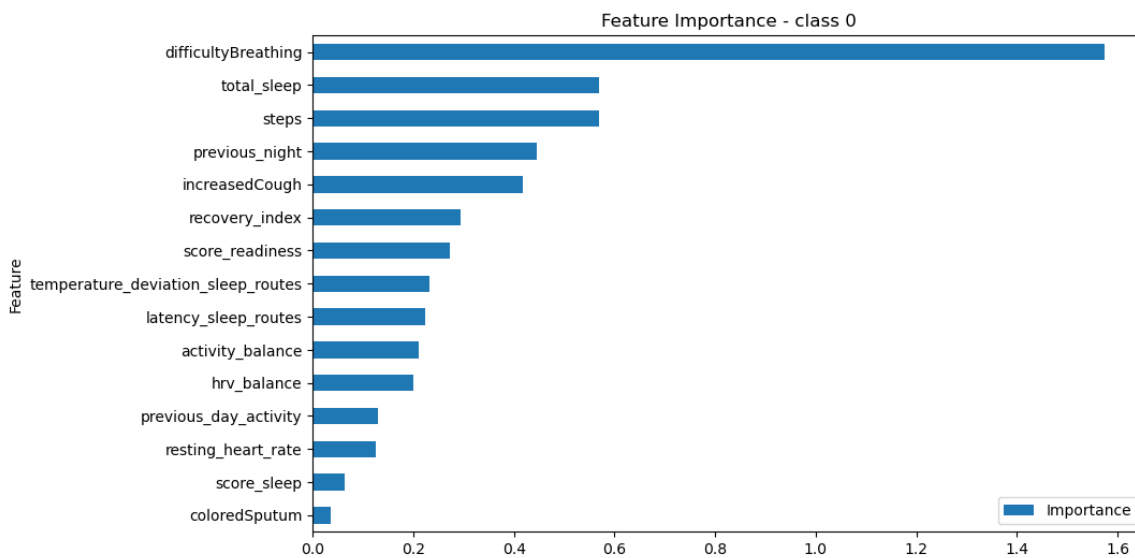
Figur 15. Feature Importance plot for den binære logistiske regressionsmodel med rød måling/gul og grøn måling som target



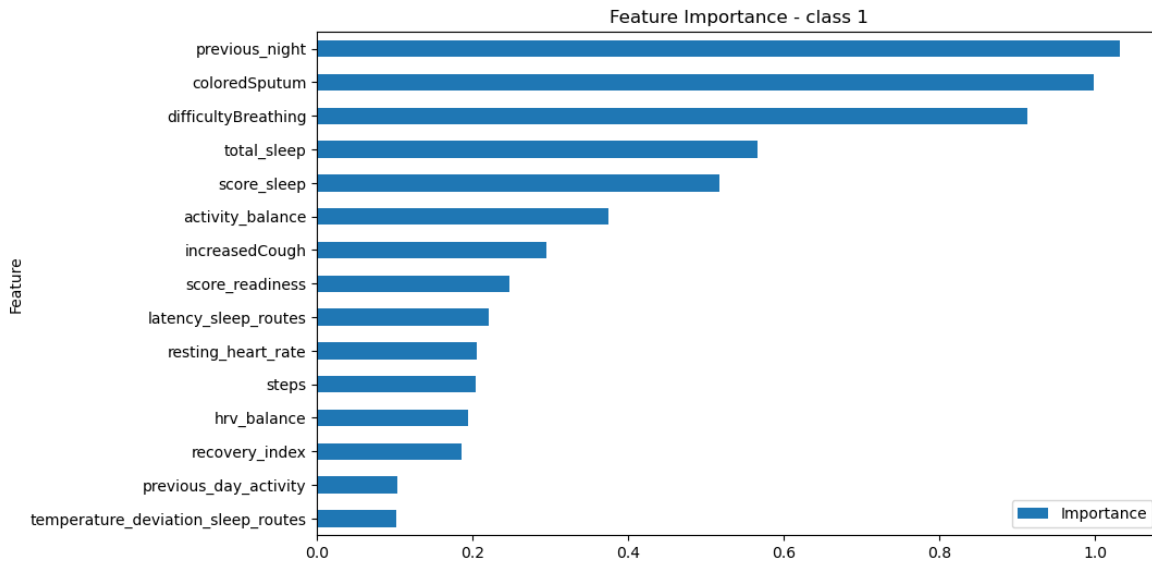
Figur 16. Overordnet Feature Importance plot for multiklasse logistiske regressionsmodel med rød, gul eller grøn måling som target



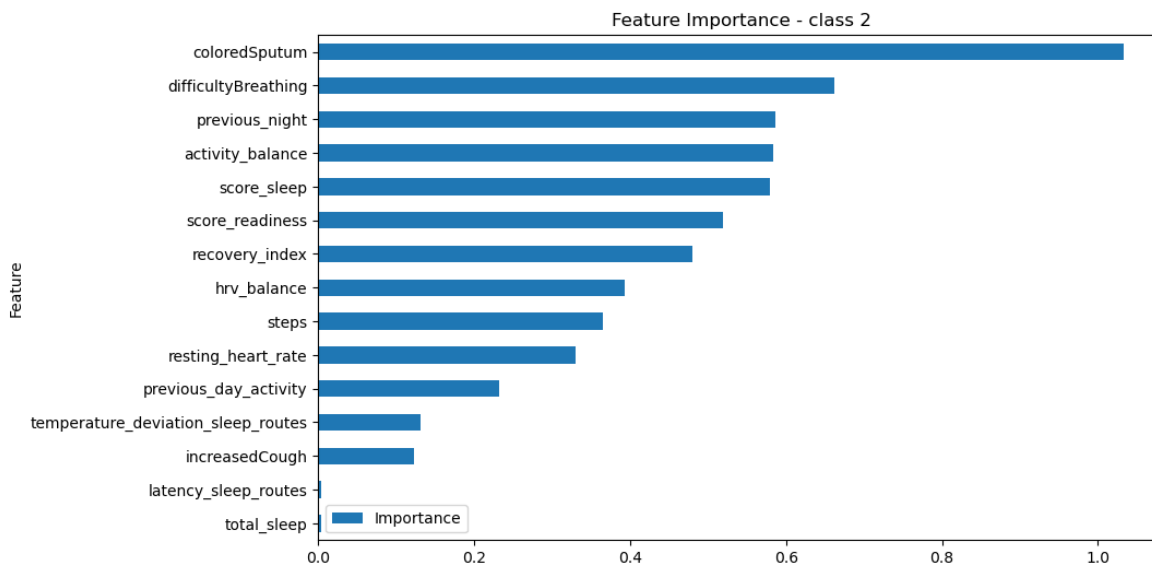
Figur 17. Feature Importance plot for klasse 0 (grønne målinger) i multiklasse logistiske regressionsmodellen



Figur 18. Feature Importance plot for klasse 1 (gule målinger) i multiklasse logistiske regressionsmodellen



Figur 19. Feature Importance plot for klasse 2 (røde målinger) i multiklasse logistiske regressionsmodellen



Bilag 3. Relevante variable fra MEDCOM-beskeder

Variable i indlæggelsesrapporterne	datatype
Cpr	tal
Kommentar til pårørende	fritekst
Sygeplejeydelse - sårbehandling	fritekst
Sygeplejeydelse - medicinadministration	fritekst
Sygeplejeydelse - plasterskift	fritekst
Plejeydelse - personlig hygiejne	fritekst
Plejeydelse - ernæring	fritekst
Praktisk hjælp - rengøring	fritekst
Praktisk hjælp - indkøb	fritekst
Træning - vedligeholdelsestræning af færdigheder	fritekst
Øvrige ydelser - madleverance	fritekst
Uspecificerede ydelser - værkstedsaktivitet	fritekst
Boligtype - ældrebolig	fritekst
kommentar til ydelser	fritekst
Hjælpemidler/behandlingsredskaber - beskrivelse	fritekst
Kommentar til hjælpemidler	fritekst
Aktuel medicin - lokal ordination	fritekst
Cave	fritekst
kommentar til aktuel medicin	fritekst
Depot medicin	fritekst
PN (efter behov) medicin	fritekst
Dosisdispenseret	ja/nej
Medicinansvar	ja/nej
Årsag til aktuel indlæggelse	fritekst
Smittorisiko	fritekst
Funktion vaske sig	score/tal
Funktion af- og påklædning	score/tal
Funktion gå på toilet	score/tal
Funktion forflytte sig	score/tal
Funktion færden i forskellige omgivelser	score/tal
Funktion drikke	score/tal
Funktion spise	score/tal
Sygeplejefagligt problemområde - bevægelsesapparat	fritekst
Sygeplejefagligt problemområde - ernæring	fritekst
Sygeplejefagligt problemområde - hud og slimhinder	fritekst
Sygeplejefagligt problemområde - kommunikation	fritekst
Sygeplejefagligt problemområde - psykosociale forhold	fritekst
Sygeplejefagligt problemområde - respiration og cirkulation	fritekst
Sygeplejefagligt problemområde - seksualitet	fritekst
Sygeplejefagligt problemområde - smerter og sanseindtryk	fritekst

Sygeplejefagligt problemområde - søvn og hvile	fritekst
Sygeplejefagligt problemområde - viden og udvikling	fritekst
Sygeplejefagligt problemområde - udskillelse af affaldsstoffer	fritekst

Variable i plejeforløbsplanerne	datatype
Cpr	tal
Tidspunkt for indlæggelse	tids-punkt/dato
Forventet udskrivnings-/færdigbehandlingsdato	tids-punkt/dato
Kommentar til udskrivningsdato	fritekst
Aftalt udskrivningsdato	tids-punkt/dato
Kommentar til pårørende	fritekst
Smittorisiko	fritekst
Madpakke gives med	ja/nej
Aftalt indkøb på udskrivningsdagen	ja/nej
Ændring på bolig	fritekst
Hjælpemidler skal aftales	fritekst
Behov for genoptræning	fritekst
Behov for hjemmesygepleje	fritekst
Behov for hjælp til medicinadministration	fritekst
Behov for medicin til medicinindtagelse	fritekst
Behov for anden kommunal indsats	fritekst
Behov for psykiatrisk udskrivningsaftale/koordineringsplan	fritekst
Funktion vaske sig	score/tal
Funktion af- og påklædning	score/tal
Funktion gå på toilet	score/tal
Funktion forflytte sig	score/tal
Funktion færden i forskellige omgivelser	score/tal
Funktion drikke	score/tal
Funktion spise	score/tal
Sygeplejefagligt problemområde - bevægelsesapparat	fritekst
Sygeplejefagligt problemområde - ernæring	fritekst
Sygeplejefagligt problemområde - hud og slimhinder	fritekst
Sygeplejefagligt problemområde - kommunikation	fritekst
Sygeplejefagligt problemområde - psykosociale forhold	fritekst
Sygeplejefagligt problemområde - respiration og cirkulation	fritekst
Sygeplejefagligt problemområde - seksualitet	fritekst
Sygeplejefagligt problemområde - smerter og sanseindtryk	fritekst
Sygeplejefagligt problemområde - søvn og hvile	fritekst
Sygeplejefagligt problemområde - viden og udvikling	fritekst
Sygeplejefagligt problemområde - udskillelse af affaldsstoffer	fritekst

Variable i udskrivningsrapporterne	datatype
Cpr	tal
Kommentar til pårørende	fritekst
Tidspunkt for indlæggelse	tids-punkt/dato
Tidspunkt for færdigbehandling	tids-punkt/dato
Tidspunkt for udskrivning	tids-punkt/dato
Årsag til aktuell indlæggelse	fritekst
Indlæggelsesforløb	fritekst
Smitterisiko	fritekst
Diagnoser	SKS/ICPC
Funktion vaske sig	score/tal
Funktion af- og påklædning	score/tal
Funktion gå på toilet	score/tal
Funktion forflytte sig	score/tal
Funktion færden i forskellige omgivelser	score/tal
Funktion drikke	score/tal
Funktion spise	score/tal
Sygeplejefagligt problemområde - bevægelsesapparat	fritekst
Sygeplejefagligt problemområde - ernæring	fritekst
Sygeplejefagligt problemområde - hud og slimhinder	fritekst
Sygeplejefagligt problemområde - kommunikation	fritekst
Sygeplejefagligt problemområde - psykosociale forhold	fritekst
Sygeplejefagligt problemområde - respiration og cirkulation	fritekst
Sygeplejefagligt problemområde - seksualitet	fritekst
Sygeplejefagligt problemområde - smerter og sanseindtryk	fritekst
Sygeplejefagligt problemområde - søvn og hvile	fritekst
Sygeplejefagligt problemområde - viden og udvikling	fritekst
Sygeplejefagligt problemområde - udskillelse af affaldsstoffer	fritekst
Midlertidige hjælpemidler/behandlingsredskaber	fritekst
Kommentar til hjælpemidler	fritekst
Depot medicin	fritekst
PN (efter behov) medicin	fritekst
Medsendt medicin (bemærkning)	fritekst
Dosisdispensering genbestilt	ja/nej
Cave	fritekst
Kommentar til medicin	fritekst
Madpakke gives med	ja/nej
Aftalt indkøb på udskrivningsdagen	ja/nej
Fremtidige aftaler	fritekst

Variable i genoptræningsplanerne	datatype
Ccpr	Tal
Starttidspunkt for aktuel sygehuskontakt	tidspunkt/dato
Forventet afslutningstidspunkt for aktuel sygehuskontakt	tidspunkt/dato
Information om behandling til patient og pårørende	fritekst
Praktiske oplysninger	fritekst
Patientsikkerhedsmæssige forhold	fritekst
Tidsfrist for start af genoptræning	fritekst
Dato for aftaler om kontrol og opfølgning	dato
Helbredsforhold inkl. Beskrivelse af behandlingsforløb - sammenfatning	fritekst
Helbredsrelateret funktionsevne på udskrivningstidspunktet - sammenfatning	fritekst
Udvidet beskrivelse af helbredsrelateret funktionsevne på udskrivningstidspunktet	fritekst
Patientens genoptræningsbehov og potentiale	fritekst
Aktionsdiagnoser	diagnosenomenklatur
ICF-koder	funktionsevnekoder

Variable i melding om færdighedsbehandling	datatype
Cpr	tal
Tidspunkt for indlæggelse	tids-punkt/dato
Tidspunkt for færdighedsbehandling	tids-punkt/dato
Tidspunkt for genoptaget behandling	tids-punkt/dato

Bilag 4. Tekster oplæst i pilotforsøg med stemmegenkendelse

Jeg Vil La' Lyset Brænde

Sang af Ray Dee Ohh

Vers 1

Stille sommeraftenvind
Månen trækker i din sjæl
Kalder ud, der hvor regnbuen ender
La' mig hviske til farvel

Omkvæd

Jeg vil la' lyset brænde
Og la' min dør stå på klem
Indtil du har set, hvad du ville
Indtil du er hos mig igen
Jeg vil la' lyset brænde
Og la' min dør stå på klem
Se, mit hjerte gløder i natten
For at du kan finde vejen hjem

Vers 2

Stille sommermorgenregn
Gennem byen med mig selv
Vinker højt mod din flyver på himlen
Hør mig hviske til farvel

Den første gang jeg så dig
Sang af Kim Larsen & Kjukken

Vers 1

Den første gang, jeg så dig, det var en sommerdag
En sommerdag da solen lyste klart
Og alle engens blomster de stod i stakkevis
Og vuggede som unge brudepar

Bro 1

Og vinden kom så stille ned til stranden, hvor vi stod
Og bølgerne de klukkede og kyssede din fod

Omkvæd 1

Den første gang jeg så dig, det var en sommerdag
Den første gang du holdt mig i hånden

Vers 2

Den første gang jeg så dig, der skinned' sommer sky
Så blendende som syv vilde svaner
Jeg mødte dig i vrimlen et sted midt ind' i byen
Ofelia og den sidste mohikan

Jeg Plukker Fløjlsgræs

Skrevet af Sigfred Pedersen & Knud Vad-Thomsen

Vers 1

En vår er kommet så mild og stille
Med drømmedufte om mark og eng
På sølvstreng klimprer den friske kilde
Og fuld af sødme af fuglens sang

Omkvæd

Jeg plukker fløjlsgræs og ridderspore
Jeg plukker mandstro og jomfrusko
Men blodet banker i mine årer
Mit lille hjerte har ingen ro
Ja, blodet banker i mine årer
Mit lille hjerte har ingen ro

Vers 2

Jeg synger ikke, jeg danser ikke
Der er så meget, jeg ikke når
Er jeg bedrøvet? Jeg ved det ikke
Jeg drømmer bare, og tiden går

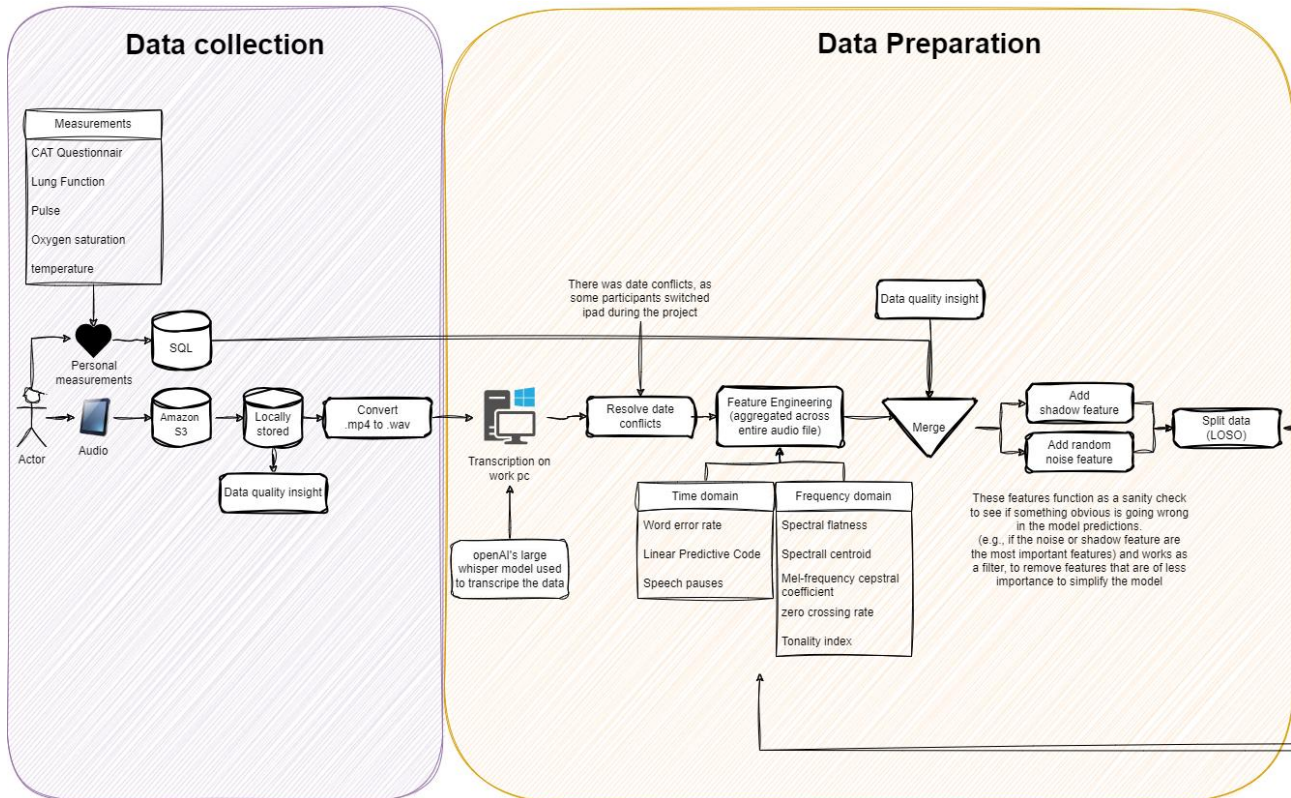
Sommerdrøm – under et blomstrende æbletræ
Kumbel: Gruk – 19. samling

Under en æblekrone,
som bierne summer i,
går jeg i middagshi.
Ned fra den lyse zone
lyder en summetone,
sommerens melodi.

Mellem de hvide grene
løber en luftig sti.
Her er de to alene,
blomstrende træ og bi.
Dette er livets scene
i sommerligt sceneri.

Biens arbejdsomme færden
blir til en alegori
i drømmen den overgår i.
Let som en ordleg er den:
dét er den hvide verden,
og dét er den kække bi!

Workflow for lyddata



Bilag 5. Analyser af KOL-algoritme

Evaluering af KOL-algoritmen i PreCareKlinikken

PreCareKlinikken tilbyder bl.a. borgere med KOL gennem egenmålinger at få kontrollen tilbage i eget sygdomsforløb og undgå akutte indlæggelser.

De tilknyttede borgere med KOL måler følgende centrale værdier (tilstandsmålinger):

- Puls
- Iltmætning
- Lungefunktion (FEV1)
- Temperatur

Desuden bliver borgerne bedt om at tage stilling til, om de oplever (tilstandsvurderinger):

- Øget vejrtrækningsbesvær
- Tiltagende hoste
- Farvet spyt

Borgerne bliver opfordret til at måle sig dagligt. Når tilstandsmålingerne og tilstandsvurderingerne bliver registreret i Appinux-systemet, bliver de automatisk tildelt en samlet vurdering i form af farven grøn, gul eller rød. Klassificeringen foregår ved hjælp af en algoritme (beslutningstræ), som er baseret på værdier for den enkelte borgers habitualtilstand svarende til habitualtilstanden.

Hvis eksempelvis borgerens iltmætning (oxygenSaturation) i en måling er under 85% af habitualværdien, bliver "stoplyset" rødt. Se nedenstående tabel.

Parameter	Interval 1 (normalt)	Interval 2 (risiko)	Interval 3 (forværring)
SAT	>90	85-90%	<85%
LFU	<10% (ift. normalt)	<15-20% (ift. normalt)	<20% (ift. normalt)
Puls	<100	100-120	>120
Temp (C)	<37,4	37,4-38	>38
Øget åndenød	Nej	Ja	
Tiltagende hoste	Nej	Ja	
Purulent ekspektorat	Nej	-	Ja
Alert Score			

I nærværende rapport analyseres mulighederne for 1) at forbedre algoritmen, således at den bedre kan bruges til at forudse hvornår borgernes tilstand forværres 2) at undersøge om det er muligt at bruge færre eller andre inputvariable end dem, der anvendes for nuværende.

Analysedatasæt

I arbejdet med analysen af egenmålingerne er der blevet struktureret to forskellige analysedatasæt fra dataene. De to datasæt bruger de samme 11 variabler, som ses for neden:

Tilstandsmålinger	Habituelle Værdier
pulse	pulse_habitualVaerdi
oxygenSaturation	oxygenSaturation_habitual
lungFunction	lungfunction_habitualVaerdi
temperature/temperatur	temperature_habitualVaerdi
Tilstandsvurderinger	
	difficulty_breathing
	increasedCough
	ColoredSputum

Det er vigtigt at pointere her, at de 11 variabler er de 7 tilstandsmålinger + de 4 habituelle værdier.

Nedenfor ses en beskrivelse af hver analysedatasæt, samt et eksempel på strukturen af tabellen. Responsvariablen til de to analysedatasæt er noteret som "y_true", hvor en **"ikke igangsættelse"** er en falsk værdi, 0, og en **"igangsættelse"** er en sand værdi, 1.

Igangsættelser og behandlingsforløb

Igangsættelserne, som fungerer som targetværdier i analyserne, er sat sammen i behandlingsforløb, hvis de ligger med højst 10 dage mellemrum. Dette gøres, fordi det vurderes, at der ikke er tale om egentlige behandlingsigangsættelser, men derimod typisk er tale om ændringer i dosis på allerede igangværende behandlinger etc. I 10 dage fra behandlingsigangsættelsen er borgerens egenmålinger taget ud af datasættene, idet det vurderes, at borgeren stadig er i et behandlingsforløb, hvilket betyder, at borgerens egenmålinger ikke kan resultere i nye behandlingsigangsættelser.

Analysedatasæt med én række per måling

Det første analysedatasæt er designet til at indeholde én række per måling. Datasættet er lavet ved at tage udgangspunkt i målingerne. Hvis der er på en dag, er flere målinger, tages alene dagens sidste måling med. Hvis der ikke er nogen målinger på en given dag, er der ikke nogen række for denne dag.

Behandlingsforløbene sættes på den måling, der ligger senest før igangsættelsen, blandt de målinger der ligger højst en time efter igangsættelsen og ligger efter klokken 17:00 på dagen inden igangsættelsen. Der findes igangsættelser uden klokkeslæt, men med dato i databasen. Disse er sat til klokken 17, for at tillade at alle målinger foretaget inden for almindelig arbejdstid på den pågældende dag ligger inden igangsættelsen.

Hvis ikke behandlingsforløbet har kunnet kobles til en måling, er det ikke medtaget. Det vil sige, for hver borger, er der registeret en række med deres daglig måling med den tilsvarende respons variabel ("y_true").

Måling ID	pulse	oxygenSaturation	...	increasedCough	ColoredSputum	y_true
0	75	97	.	True	False	0
1	89	92	.	False	True	1
2	69	150	.	False	False	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Analysedatasæt med én række per dag per borger inkl. behandlingsforløb uden måling og historik

Det andet analysedatasæt er designet til at indeholde én række per dag per borger, uanset om borgeren har foretaget egenmålinger eller om der er foretaget behandlingsindsatser. Det adskiller sig fra det første analysedatasæt, idet der findes rækker uden indhold i input- eller outputvariablene.

Dette datasæt er dannet med henblik på bedre at kunne simulere den empiriske situation, hvor borgerens tilstand må prædikeres uanset, om der indsendes måling eller ej. I dette datasæt er der således også inkluderet behandlingsforløb, som ikke har kunnet kobles til egenmålinger efter den samme regel som også anvendtes i det foregående datasæt.

Der er i dette datasæt også medtaget observationer fra de foregående tre dages egenmålinger. Det vil sige, for en enkelt borger, vil der være en måling per dag, hvor kolonner er tilføjet med udfaldet af de tre forrige dages målinger. Dette er illustreret i tabellen forneden, hvor pulse_0 er dagens måling, pulse_1 er forrige dages måling, pulse_2 er –2 dages måling. Dette bliver så gjort for alle variablerne på de forrige 3 dage.

Dato	Borger ID	pulse_0	Pulse_1	Pulse_2	...	y_true
20-01-2023	1	62	68	61	.	0
20-01-2023	5	56	76	87	.	1
20-01-2023	10	77	75	73	.	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

I de tilfælde hvor der ikke er målinger på en af de fire indgående dage bruges gennemsnittet af værdierne fra de dage, hvor der har været foretaget målinger. Hvis en måling ikke har været foretaget alle fire dage, bruges gennemsnittet over alle målinger af den givne type i datasættet.

Vejrdata

Til begge datasæt er der tilføjet vejrdato fra vejrstationen i København Lufthavn (stationID: 06156). Vejrdato registreres hver time døgnet rundt. Vi har valgt at medtage beregnede gennemsnit, minimum, maksimum og median af temperatur, luftfugtighed og lufttryk for hver dag, samt en "vejrkode", som indikerer hvor godt/skidt vejret har været den dag.

Evalueringsmetrikker

De to primære evalueringsmetrikker der bruges er en "confusion matrix" og "ROC curve".

Confusion Matrix

En confusion matrix er en metrik til at fortælle følgende fire ting:

5. Hvor mange **“ikke igangsættelser”** fanger algoritmen. Det vil sige, når det er sandt, at der ikke er en igangsættelse på baggrund af en måling, hvor ofte fanger algoritmen dette. Denne værdi er også benævnt som “Sande Negative / True Negatives (TN)”.
6. Hvor mange **“igangsættelser”** fanger algoritmen. Det vil sige, når det er sandt, at der er en igangsættelse på baggrund af en måling, hvor ofte fanger algoritmen dette. Denne værdi er også benævnt som “Sande Positive / True Positives (TP)”.
7. Hvor ofte prædikterer algoritmen en **“ikke igangsættelser”**, hvor målingen faktisk resulterede i en “igangsættelse”. Det vil sige når en borger har indgivet en måling og derefter er blevet igangsat, men algoritmen ikke har fanget dette. Denne værdi er også benævnt som “Falske Negative / False Negatives (FN)”.
8. Hvor ofte prædikterer algoritmen en **“igangsættelser”** hvor målingen faktisk *ikke* resulterede i en igangsættelse. Det vil sige, når en borger har indgivet en måling, men ikke har fået igangsat en behandling, men algoritmen vurderer at der vil blive igangsat behandling. Denne værdi er også benævnt som “Falske Positive / False Positive (FP)”.

		Predicted	
		Negative (N) -	Positive (P) +
Actual	Negative -	True Negative (TN)	False Positive (FP) Type I Error
	Positive +	False Negative (FN) Type II Error	True Positive (TP)

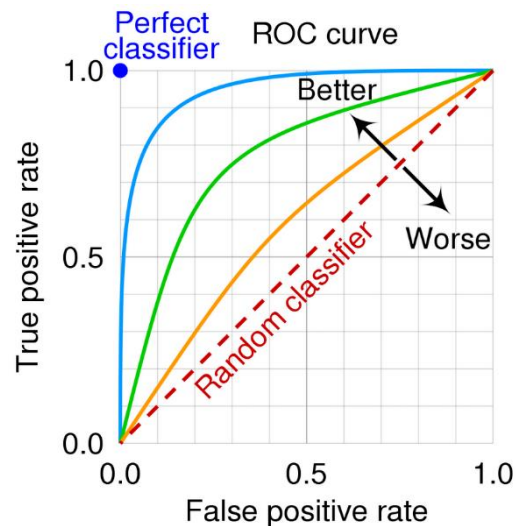
En perfekt algoritme resulterer i at have 100% sande positive og sande negative. Men i en mere realistisk verden vil der oftest være et kompromis mellem antallet af falske positive og falske negative.

ROC (Receiver Operating Characteristic) kurve

En ROC-kurve er en grafisk fremstilling af ydeevne af en binær klassifikationsmodel, når man varierer modellens træfsikkerhedstærskel. Den illustrer modellens evne til at skelne mellem den positive og negative klasse og viser forholdet mellem modellens følsomhed og specificitet.

Følsomheden er også kendt som “True Positive Rate (TPR)”, som er en måling af, hvor godt modellen kan identificere positive tilfælde. Specificiteten også kendt som “True Negative Rate (TNR)” måler, hvor godt modellen kan identificere negative tilfælde. Som ses forinden, plottes ROC-kurven som en graf med følsomheden på y-aksen og 1-specificiteten på x-aksen. Jo tættere kurven er på øverste venstre hjørne af grafen, jo bedre er modellens ydeevne.

AUC (Area Under the curve) er et tal, der angiver det samlede område under ROC-kurven, og er en måling af den samlede ydeevne af klassifikationsmodellen. AUC varierer mellem 0 og 1, hvor en værdi på 0 indikerer, at modellen er helt ineffektiv til at skelne mellem de to klasser, "ikke igangsat" og "igangsat", mens en værdi på 1 indikerer en perfekt ydeevne. AUC bruges til at sammenligne ydeevnen af forskellige klassifikationsmodeller, og oftest vælges den model, som har den bedste AUC score, da den er bedst til at skelne mellem de to klasser.

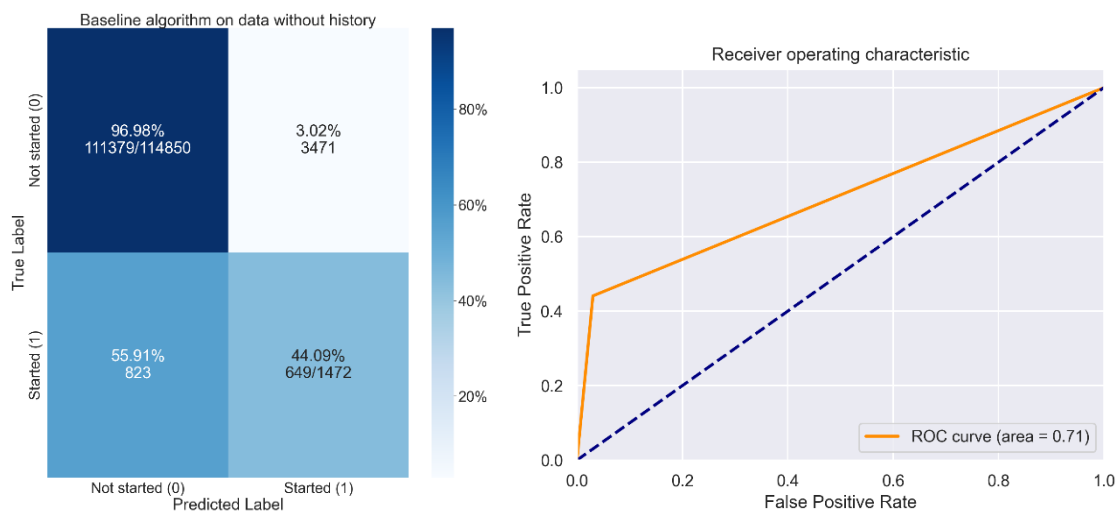


Baseline: Den nuværende KOL-algoritme

For at kunne sammenligne resultaterne med den nuværende KOL-algoritme, som benævnes som baselinen, skal baselinen først analyseres. Da der er to analysedatasæt, vil der tilsvarende være to baselines.

Baseline for analysedatasættet med én række per måling

Den første baseline bruger analysedatasættet "én række per måling", hvor der ses følgende evaluering af baseline algoritmen:



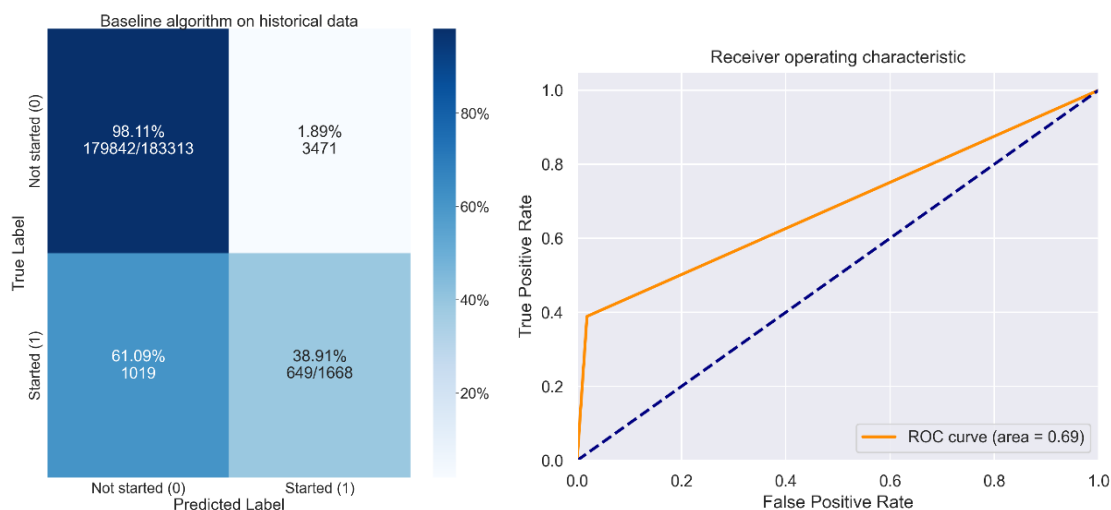
Til højre ses confusion matrix hørende til stoplys-algoritmen, når den er anvendt på alle målinger, og der kun er medtaget behandlingsigangsættelser, som kunne knyttes til en måling. Tallene i øverste venstre hjørne indikerer, at der har været 114.850 egenmålinger i hele datasættet, der *ikke* har udløst en behandlingsigangsættelse, mens der har været 1.472 egenmålinger, som har udløst en behandlingsigangsættelse, jf. nederste højre hjørne. Dette giver i alt 116.322 egenmålinger. Af de 1.472 behandlingsigangsættelser kan 649 af dem kobles til en egenmåling, der har resulteret i et rødt stoplys. Dette svarer til 44,09%.

ROC-kurven til højre viser forholdet mellem andelen af sande positive versus andelen af falske positive. Arealet under ROC-kurven er også tilsvarende balanceret accuracy med en værdi på 71 pct. Som det kan ses af ROC-kurven er der næsten ingen falske positive (3,02%), mens performance i forhold til sande positive er mere middelmådig (44.09%). Denne baseline er rigtig god til at klassificere ikke-igangsættelserne korrekt, men har problemer med at klassificere de målinger der faktisk fører til igangsættelser. Det vil sige at relativt mange målinger, der fører til en igangsættelse, ikke fanges af stoplyset.

Baseline for analysedatasættet med én række per dag per borger

Den anden baseline er lavet på analysedatasættet "*én række per dag per borger*". Her ses det, at der sammenlagt er 184.999 dage, hvor borgerne har deltaget i KOL-projektet, hvoraf de har leveret egenmålinger i 114.850 af dagene.

I 1.686 af dagene har der været en behandlingsigangsættelse, hvilket er 214 flere end i datasættet hvor der kun er medtaget igangsættelser med tilknyttet egenmåling. Dette betyder også, at der i dette datasæt er 1019 (61,09%) igangsættelser, der ikke er forårsaget af en egenmåling, der har ført til rødt stoplys, hvilket vil sige at kun 38,91% af igangsættelserne er knyttet til et rødt stoplys. Til sammenligning var det 44,09 pct. i det første datasæt, jf. ovenfor. Forskellen mellem de to skyldes at der i det andet datasæt er medtaget igangsættelser, der ikke er knyttet til en måling. Disse er i sagens natur heller ikke knyttet til et stoplys, hvorfor der ses en yderligere forringelse af stoplysets evne til at prædiktere igangsættelserne.



Modeller med analysedatasæt på fuldt data med "én række per måling"

Alle modellerne i det nedenstående er trænet på 80% af data og evalueret på 20%. Derfor er de absolutte tal forskellige fra baseline, hvorfor det udelukkende er procentangivelserne, der kan sammenlignes.

Der opstilles tre forskellige modeller.

- Logistisk regression
- Random Forest
- XGBoost

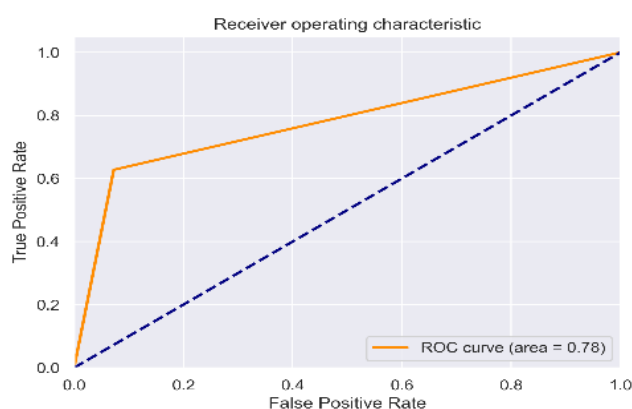
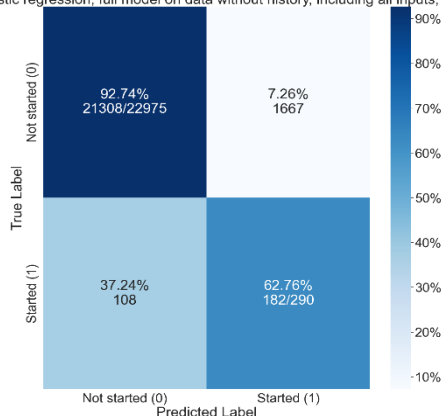
Model 1: Logistisk regression

Som basismodel er valgt at udføre en række logistiske regressioner på data. Logistisk regression er en transparent algoritme, hvilket vil sige den ikke er særlig kompleks. Den fungerer ved at der tages en linearkombination af de indgående parametre, som derpå transformeres med en sigmoid-kurve, som angiver sandsynligheden for positivt outcome. Formålet med at bruge denne metode er at lave en række af de simpleste modeller, for at få et overblik over, hvad en minimal model vil kunne prædiktere.

I den fulde logistiske regressionsmodel, indgår alle de indgående variable, samt enkelte kombinationer af variablene: alle egenmålingerne, vejrdata, forholdet mellem egenmåling og habitualværdi, samt anden potens af egenmålingerne. Alle variablene er listet her:

Kontinuerte Variabler			
lungFunction	lungFunctionHabitualVaerdi / lungFunction	lungFunction*lungFunction	
pulse	pulseHabitualVaerdi / pulse	pulse*pulse	
temperature	temperatureHabitualVaerdi / temperature	temperature*temperature	
oxygenSaturation	oxygenSaturationHabitualVaerdi / oxygenSaturation	oxygenSaturation*oxygenSaturation	
humidityaverage	humiditymax	humiditymedian	humiditymin
pressureaverage	pressuremax	pressuremedian	pressuremin
humidityaverage	humiditymax	humiditymedian	humiditymin
tempaverage	tempmax	tempmedian	tempmin
Kategoriske Variabler			
increasedCough	difficultyBreathing	coloredSputum	weathercode

Logistic regression, full model on data without history, including all inputs.



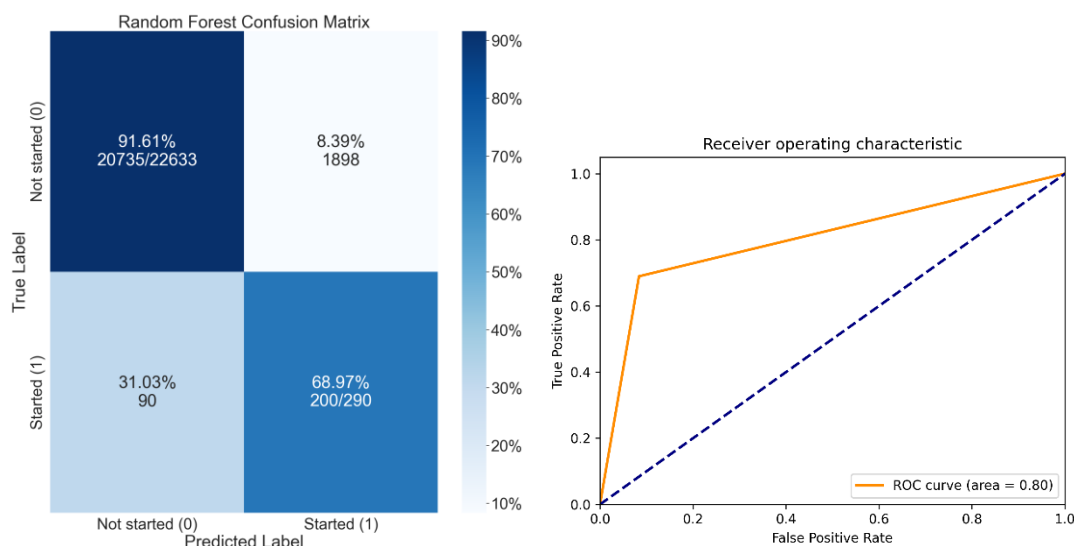
Logistisk regression med alle indgående features, inkl. habitualværdier og oplysninger om vejret.

Som det ses på diagrammet med ROC-kurven, er balanceret accuracy 78% som skal sammenlignes med stoplys baseline på 71%. 62,76% af igangsættelserne blev prædikeret. Bemærk at der er væsentligt flere falske positive (7,26%) sammenlignet med baseline på 3,02%.

Model 2: Random Forest

En mere kompleks klasse af modeller er Random Forest. En Random Forest model er en maskinlærings-algoritme, som bruger mange individuelle beslutningstræer. Hvert beslutningstræ bliver trænet på forskellige tilfældige dele af dataene. Resultaterne af de individuelle træer kombineres derefter for at generere en samlet forudsigtelse. Random Forest er ikke en transparent model, som et beslutningstræ, men er i stand til at håndtere store datamængder og giver mere præcise resultater end en enkelt beslutningstræ-algoritme, da den er i stand til at finde komplekse sammenhæng i dataene.

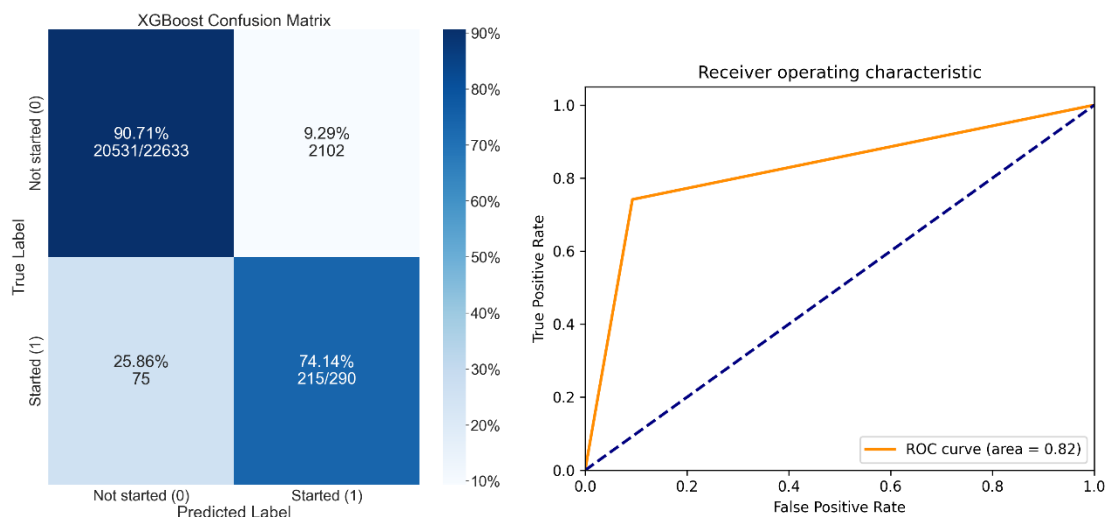
Med denne model har vi fået en balanceret accuracy på 80%. I denne model bliver 68,97 pct. af igangsættelserne prædikeret, hvilket er mere end i modellen med logistisk regression. Der er 8,39 pct. falske positive, hvilket er mere end i baseline og i modellen med logistisk regression.



Model 3: XGBoost

Endnu en kompleks model er XGBoost. XGBoost (eXtreme Gradient Boosting) er en maskinlærings-algoritme, som også benytter sig af beslutningstræer, som gradvist forbedrer modellens præcision ved hjælp af gradient baserede optimeringsmetoder. I modsætning til Random Forest, så bliver beslutningstræerne forbundet i en serie, og for hvert efterfølgende træ forsøger algoritmen at rette op på de fejl, der blev gjort af tidligere træer. XGBoost er heller ikke en transparent model, men bruges ofte i sammenligning med Random Forest.

Med denne model har vi fået en balanceret accuracy på 82%. I denne model bliver 74,14 pct. af igangsættelserne prædikeret. Der er 9,29 pct. falske positive, hvilket er større end i baseline og de to øvrige modeller.



Evaluering af analysedatasæt på fuldt data med “én række per måling”

Efter forsøg med de tre ovenstående maskinlæringsmodeller, er det muligt at sammenligne og evaluere modellerne op imod den tilsvarende baseline algoritme.

Model	Balanceret accuracy	TPR	FPR
Baseline	71%	44,09%	3,02%
Logistisk regression	78%	62,76%	7,26%
Random Forest	80%	68,97%	8,39%
XGBoost	82%	74,14%	9,29%

Som det ses, har alle tre modeller en del bedre balanceret accuracy end baselinemodellen (fremhævet med orange). XGBoost (fremhævet med grønt) er bedst med 82% accuracy, hvor der er hele 74,14% sande positive imod de 44,09% som baseline modellen rammer. Til gengæld har denne model også flest falske positive.

For at eksemplificere hvad det betyder at vores bedste maskinlæringsmodel (XGBoost) har en større antal falske positive forudsigelser har vi et eksempel fra ugen d.1-7/3-2023. I denne uge var der 858 målinger i alt, hvor 35 af målingerne er triageret røde, og 5 røde målinger gav anledning til en igangsættelse. Dermed var der 11 igangsættelser, hvor 9 kunne knyttes til en måling.

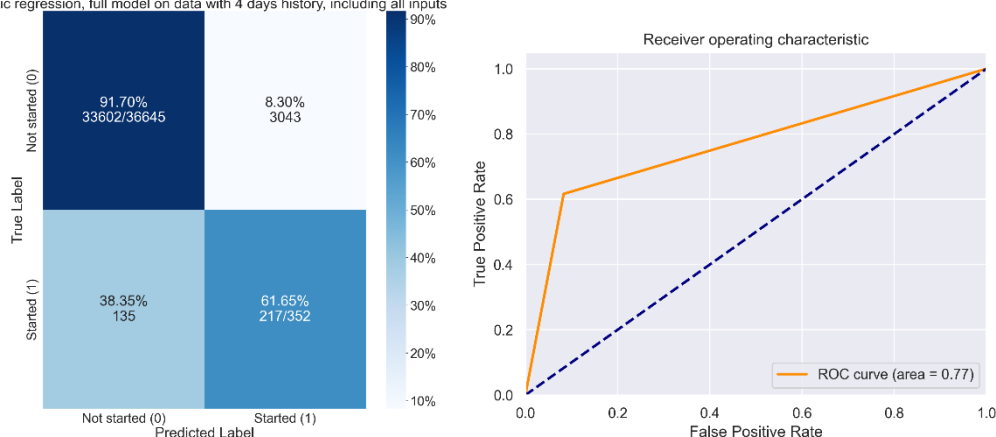
Hvis vi bruger XGBoost til at forudsige igangsættelser, ville det resultere i ca. 87-100 målinger triageret som røde målinger, med ca. 7 af behandlingsigangsættelserne, der ville have en rød triagering. Dette betyder, at med det nuværende stoplys er det 35 røde målinger, og vores model ville øge dette tal til 87-100 røde målinger. Med dette i tankerne kan der tænkes meget over hvorvidt det er realistisk, og hvorvidt der vil være en måde at implementere den avancerede maskinlæringsmodel hvor der håndteres de mange falske positive forudsigelser ved, fx. at lave en risiko for forværring (rangering) af forudsigelserne.

Modeller med analysedatasæt på fuldt data med "én række per dag per borger"

Model 1: Logistisk regression

I den fulde logistisk regressionsmodel, indgår alle de indgående variable samt enkelte kombinationer af variablene: alle egenmålingerne samt forholdet mellem egenmålinger og habitualværdier for dagen, samt de sidste tre seneste dage, vejrdata for den pågældende dag samt anden potens af egenmålingerne. I alt 77 indgående variable.

Logistic regression, full model on data with 4 days history, including all inputs



Logistisk regression med alle indgående features, inklusive habitualværdier og oplysninger om vejret, samt data fra de foregående tre dage

Evaluering af analysedatasæt på fuldt data med "én række per dag per borger"

Sammenlignes baseline for datamodellen med "én række per dag per borger" med baseline for datamodellen med "én række per måling", ses at den balancerede accuracy er omkring to procentpoint lavere her. Dette skyldes det øgede antal rækker uden observation, samt de igangsættelser, der ikke har kunnet knyttes til en måling.

Model	Balanceret accuracy	TPR	FPR
Baseline	69%	38,91%	1,89%
Logistisk regression	77%	61,65%	8,30%

Her er den logistiske regression 8 procentpoint bedre end baseline i forhold til den anden datamodel, hvor forskellen var 7 procentpoint. Denne forskel kan muligvis tilskrives de yderligere data (fra de tre foregående dage), men kan nok også tilskrives statistisk støj.

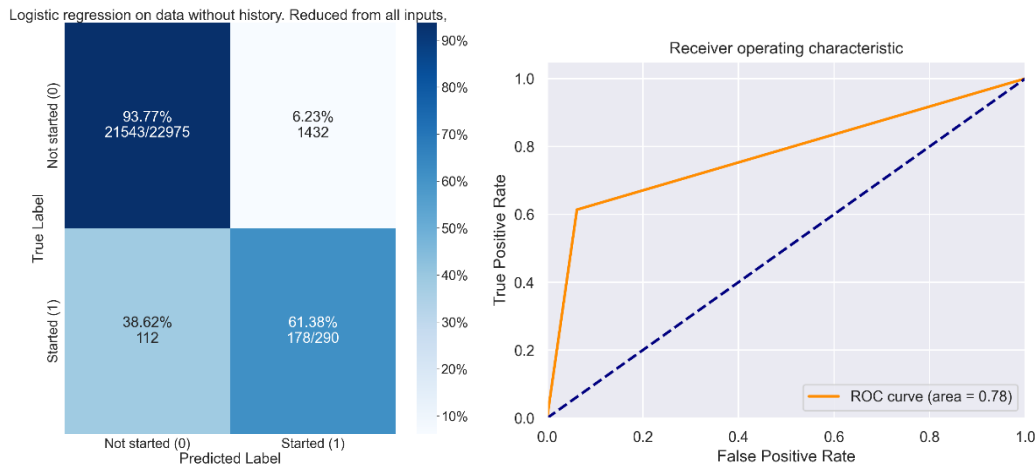
Analyse af fravælgelse af variabler

Efter disse modeller og test er fuldført, analyseres nu hvorvidt det er muligt at vælge de bedste modeller fra hvert analysedatasæt og fravælge variabler. Det vil sige, hvorvidt vi er i stand til at undlade variabler så som lung funktion og eventuel stadig opnå et bedre resultat end baseline.

Fravælgelse af variabler på analysedatasæt "én række per måling"

Model: Logistisk regression reduceret model *Logistisk regression med alle indgående features, inklusive habitualværdier og oplysninger om vejret, men hvor der kun er medtaget features med signifikante parametre.*

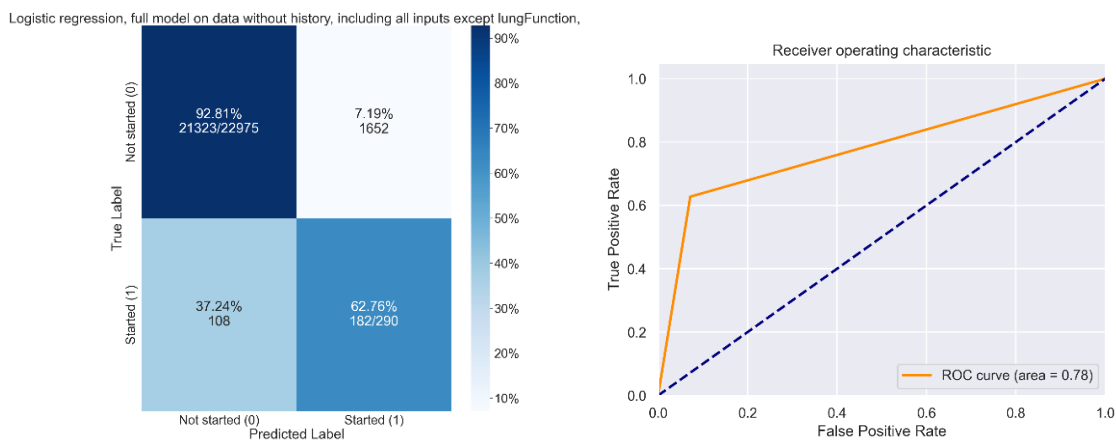
Hvor den første model medtager alle indgående features, er der her fjernet alle features, som ikke bærer signifikant information. Pointen med denne model er at vise, at man kan få en lige så god model ud af kun at medtage variablene `increasedCough`, `coloredSputum` samt `difficultyBreathing`, som er de tre variable, hvor borgeren selv vurderer sin egen tilstand. Disse variable er medtaget fordi de har statistisk signifikant effekt på modellens evne til at forudsige outcome.



Som det kan ses, er den balancerede accuracy identisk med den for den fulde model. Dog er der lidt forskel på fordelingen af sande positiver vs. falske positiver, hvor der er en smule flere sande positiver i denne model end den fulde model. Denne forskel er dog næppe signifikant.

Model: Logistisk regression uden lungefunktion

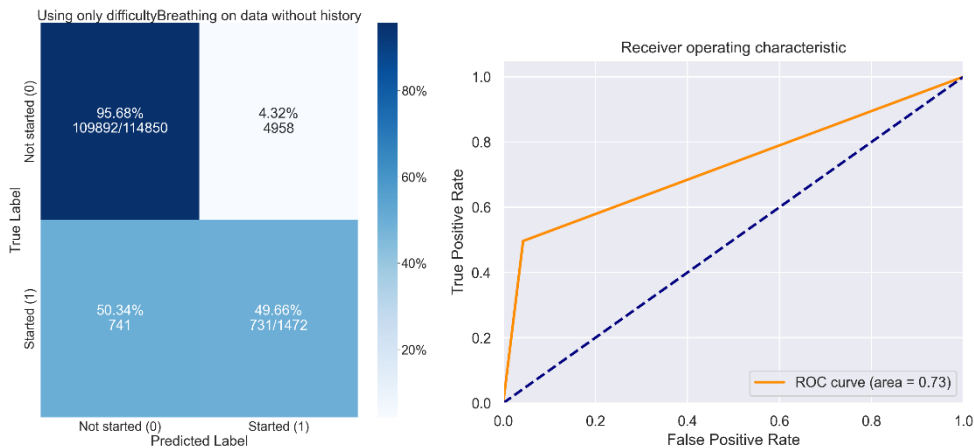
Logistisk regression hvor der er medtaget alle features undtagen `lungFunction` og `Habitualværdien` for `lungFunction`.



Her ses det, at resultaterne er næsten identiske med den fulde model. Det betyder, at man burde kunne undlade lungefunktionsmålingerne og opnå den samme forudsigelse som i en model med lungefunktionsmålinger.

Model: Åndenød. Kan `DifficultyBreathing` alene forklare igangsættelse?

Vi har undersøgt om `difficultyBreathing` kunne bruges til at forudsige igangsættelse, da denne er klart den bedste prediktor blandt de binære variable.



Her ses det overraskende, at denne variabel alene med en balanceret accuracy på 73% kan prædikere bedre end stoplyset, som har en accuracy på 71%. Andelen af sande positive på 49,66% er en smule højere end de 44,09% for stoplyset, og andelen af falske positive er med 4,32 også en del højere end stoplysets 3,02%.

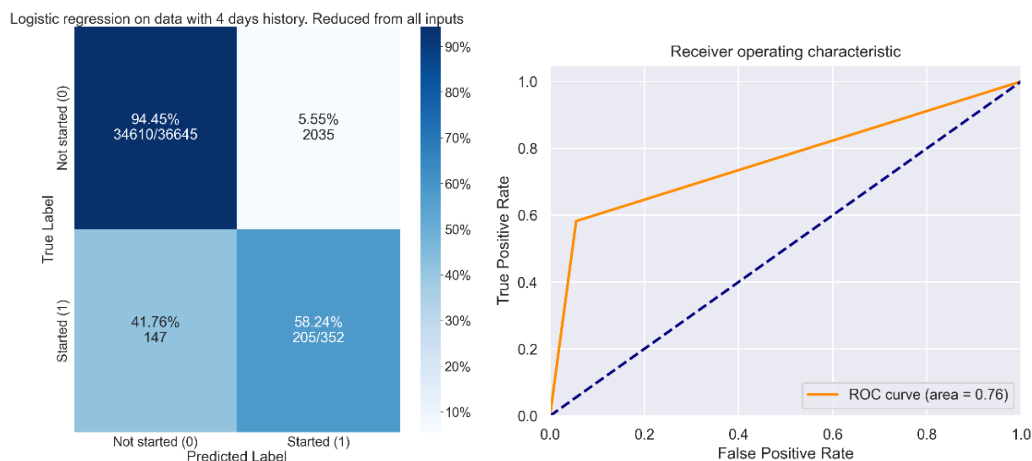
Fravælgelse af variabler på analysedatasæt "én række per dag per borger"

Model: Logistisk regression reduceret model

Logistisk regression med alle indgående features, inklusive habitualværdier og oplysninger om vejret samt data fra de foregående tre dage, men hvor der kun er medtaget features med signifikante parametre.

Her er medtaget variablerne increasedCough_0, coloredSputum_0, difficultyBreathing_0, weather_code, lungFunction_HabitualVaerdi, coloredSputum_1, difficultyBreathing_1, increasedCough_3 og difficultyBreathing_3. Disse variable er medtaget fordi de har statistisk signifikant ($p < 0.05$) effekt på modellens evne til at forudsige outcome.

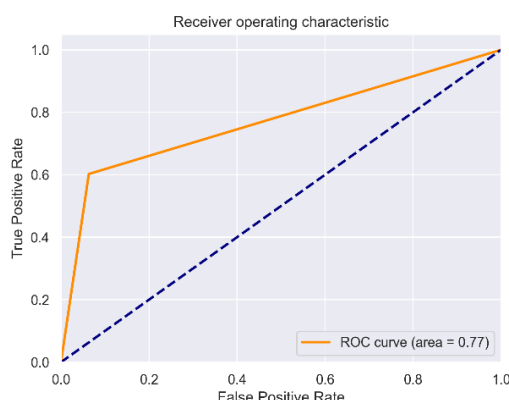
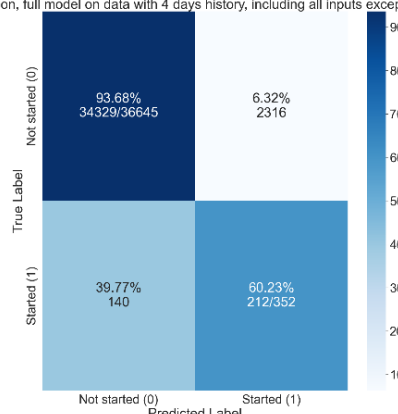
Evaluering af modellen viser at der er en accuracy på 76 pct., og andelen af sande positive er 41,76 pct., mens de falsk positive er 5,55 pct.



Model: Logistisk regression uden lungfunktion

Logistisk regression hvor der er medtaget alle features undtagen lungFunction og Habitualværdien for lungFunction, inklusive øvrige habitualværdier og oplysninger om vejret samt data fra de foregående tre dage.

Logistic regression, full model on data with 4 days history, including all inputs except lungFunction



Modellen har en accuracy på 77 pct., dvs. lidt lavere end modellen hvor lungefunktionen medtages. De sande positive er 39,77 pct., dvs. meget tæt på baseline, men de falske positive er 6,32 pct., hvilket er en del større end i baseline, hvor disse var 1,89 pct.

Opsummering af fravælgelse af variable

I de to nedenstående tabeller ses et overblik over performanceresultaterne fra de foregående afsnit. Som det ses, er der en del at hente ved brug af mere og mere komplekse modeller. Når vi har fjernet features med ikke-signifikante parametre, har det ikke givet anledning til væsentligt dårligere performance, ligesom det heller ikke har givet anledning til dårligere performance, når vi har fjernet lungefunktionsmålingerne.

En række per måling – alle modeller

Model	Balanceret accuracy	TPR	FPR
Baseline	71%	44,09%	3,02%
difficultyBreathing alene	73%	49,66%	4,32%
Logistisk regression fuld model	78%	62,76%	7,26%
Logistisk regression reduceret model	78%	61,38%	6,23%
Logistisk regression uden lungFunction	78%	62,76%	7,19%
Random Forest på alle variable	80%	68,97%	8,39%
XGBoost på alle variable	82%	74,14%	9,29%

En række per borger per dag

Model	Balanceret accuracy	TPR	FPR
Baseline	69%	38,91%	1,89%
Logistisk regression fuld model	77%	61,65%	8,30%
Logistisk regression reduceret model	76%	58,24%	5,55%
Logistisk regression uden lungFunction	77%	60,23%	6,32%

Diskussion

Evalueringsfokus

Gennem vores testning af modeller har vi valgt at fokusere på at optimere udfaldet af så lave falske negative værdier og høje sande positive værdier som muligt. Dog ser vi en tendens til at en optimering af de falske negative og sande positive værdier medfører en forværring med flere

sande negative og falske positive. Dette kompromis er noget der skal tages stilling til, hvis en endelig model skal idriftsættes. En drøftelse af balancen bør ske med det sundhedsfaglige personale i PreCare Klinikken, idet modellen vil være et dagligt værktøj for personalet og skal give mening for dem.

Fokus har været på at få så mange korrekte igangsættelser af borgere, hvor det til gengæld risikeres, at der også vil være flere borgere, der i dag ikke laver røde målinger, men som vil få røde målinger i den nye model, og som ikke giver anledning til igangsættelse af behandling. Dette vil bl.a. gøre at der skal foretages flere opkald til borgerne omkring deres målinger for at være sikker på deres tilstand ikke er kritisk, men en større andel af borgere med røde målinger vil få igangsat en behandling end i dag.

Konklusion

Vi kan generelt konkludere, at det er muligt at forbedre baseline algoritmerne. Vi har fokuseret på at optimere "balanced accuracy" metrikken, hvor der tydeligt ses en forbedring med identificering af sande positive og falske negative observationer. Som nævnt i diskussionen, vil evalueringsmetrikkerne være valgt alt efter, hvad der ønskes af udfald, dvs. hvis der ønskes færre falske negative eller falske positive end baselinen.

Resultaterne viser, at ved at bruge den balanceret accuracy som metrik, er det muligt at fjerne lungefunktionsmålingerne og stadig have en balanceret accuracy på 78%. Dog er det svært at sige, hvorvidt dette vil ændre sig med tiden alt efter mere data kommer i spil.

Resultaterne viser også, at der kan laves en solid model udelukkende med den kategoriske attribut: difficultyBreathing. Dette skyldes, at værdien er den første i beslutningstræet/baseline algoritmen og har derfor enorm stor magt i forhold til, hvorvidt stoplyset bliver sat i rød eller ej. Den tendens antages ikke som en kompleks beslutning og kan evt. ses om et problem i baseline algoritmen, da en beslutning kan være baseret 100% på baggrund af en variabel.

Vores analyse har også vist, at der ikke er en stor forskel på performance af de logiske regressio-ner vs. mere komplekse maskinlæringsmodeller. Det betyder, at hvis det ønskes, at modellen skal være transparent, vil den logistiske regression være en fin løsning. Dog kan der altid anvendes Explainable AI (se Future Work afsnit) for at håndtere black-box problemet med maskinlæringsmodellerne.

Future Work

Future Work er det arbejde vi ser som de næste skridt i projektet.

Beslutning om flere falske positiver og falske negativer

Der skal tages en beslutning om hvorvidt det ønskes at optimere de falske positiver eller de falske negativer, da der umiddelbart med de datasæt og modeller der er brugt, ikke er en stabil balance mellem de to. Balancen skal drøftes med det sundhedsfaglige personale i PreCare Klinikken, idet modellen vil være et værktøj, der påvirker sygeplejerskernes arbejde.

Flere tests på analysedatasættet "én række per dag per borger"

Et næste skridt vil være at anvende maskinlærings-modeller på analysedatasæt nr. 2, og derefter sammenligne med det første datasæt. Dette analysedatasæt vil også tillade, at nogle målinger ikke er angivet, hvor alle rækker uden værdier, blev slettet i det første datasæt.

Missing-værdierne i dette datasæt er imputeret ved brug af gennemsnittet af de omkringliggende variable. Standardmetoden til imputation er, at man imputerer flere gange med en varians, som tillader, at man taget højde for usikkerheden i imputationerne. I en uddybende analyse af datasættet med historik, vil det være relevant at benytte denne metode i stedet.

Dybere analyse og metoder til at håndtere ubalancen i dataene

For at håndtere ubalancen mellem sande negative og sande positive er der blevet oversamlet ved brug af SMOTE-metoden. Der er mange måder at håndtere ubalance i datasæt, og her vil der måske være mere at hente mhp. at forbedre balancen mellem falske positive og negative.

Mere evaluering mellem modellerne (krydsvalidering)

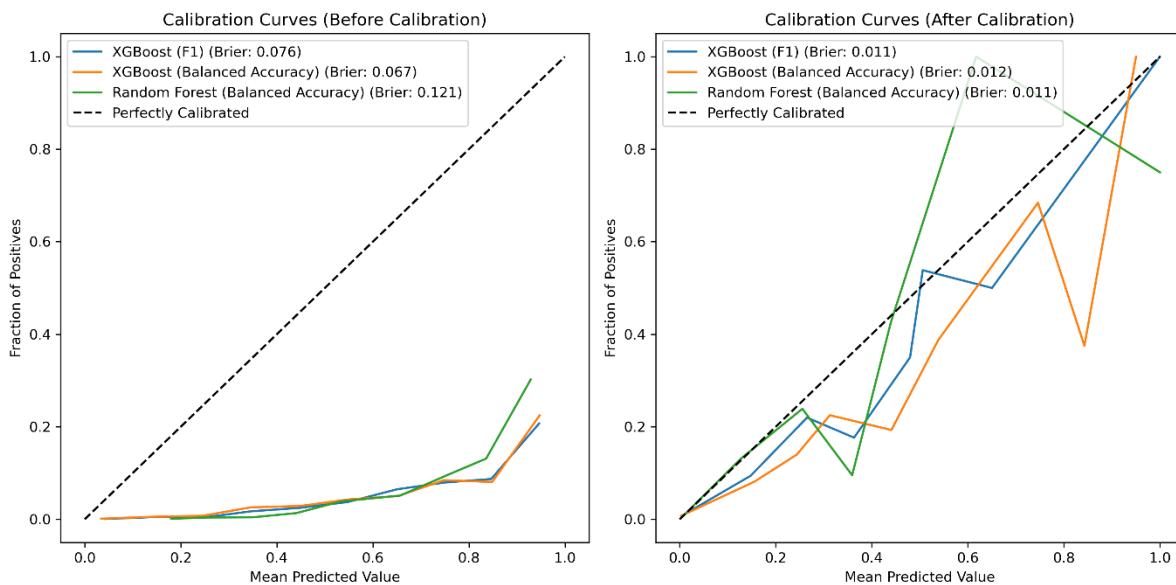
Da der er mange forskellige modeller, vil det være hensigtsmæssigt med højere sikkerhed at vurdere, hvilken model der bedst kan anbefales. Dette kan gøres ved at krydsvalidere modellerne gennem loop over ti fold, hvor der i hver fold laves en analyse/evaluering. Dermed opnås middelværdi og varians på estimerne af Balanced Accuracy (eller andre metrikker), True Positive Rate og False Positive Rate.

Modelkalibrering

Modelkalibrering refererer til processen med at justere forudsigelserne eller output-sandsynlighederne i en maskinlæringsmodel for at stemme overens med de sande sandsynligheder eller resultater i den virkelige verden. For eksempel, hvis en model forudsiger en sandsynlighed på 70% for en begivenhed, bør den ideelt set forekomme cirka 70% af tiden i virkeligheden. Men modellen kan konsekvent over- eller undervurdere de sande sandsynligheder.

Kalibrering er vigtig i forskellige anvendelser, især når output-sandsynlighederne bruges til beslutningstagning eller risikovurdering. Hvis en model er dårligt kalibreret, kan de forudsagte sandsynligheder muligvis ikke nøjagtigt afspejle sandsynligheden for, at en begivenhed faktisk forekommer, hvilket fører til forkerte beslutninger eller upålidelige risikovurdering.

I dette projekt har vi lavet en hurtig kalibrering af modellen, og derfor vil der skulle kigges dybere ind i dette. Der kan ses på billederne forneden, at jo tættere på den stiplede linje jo bedre er kalibrering af modellen. Vi ser at modellerne før kalibreringen ligger langt fra virkeligheden, og efter kalibreringen, derfor mener vi der er meget at hente ved en dybere analyse af kalibrering af modellen.



Mere omkring dette emne kan findes her: <https://scikit-learn.org/stable/modules/calibration.html>

Anvendelse af Explainable AI (XAI)

Explainable AI (XAI) er en del af maskinlæring, hvor komplekse modeller kan blive gjort transparente. Ved at anvende XAI kan der ses meget mere specifikt på, hvorfor modellerne er kommet til de forudsigelser, som de er. Det giver mulighed for at optimere både datasæt og model.

For borgerne specifikt, vil det være muligt at lave en "forklaring" i form af f.eks. "Knud er forudset til at blive igangsat indenfor de næste 4 dage, fordi Knud har haft farvet spyt, høj temperatur, og der har været høj fugtighed. Der er også mange andre måde XAI kan blive brugt på, fx at gruppere "observationer/borgere" og identificere bestemte trends for grupper af borgere så der kan optimeres ift. Behandlinger m.m.

Til at skabe forklaring anbefales at kigge på XAI metoden: [SHAP](#).

Bilag 6. Målerutiner

1. Formål

I PreCare-projektet bliver en gruppe udvalgte borgere bedt om at udføre egenmålinger, herunder lungefunktionsmålinger, måling af temperatur og iltmætning samt at besvare 3 ja/nej spørgsmål som en indikator på deres daglige symptomer på en sygdomsforværring. Borgerne anbefales at udføre målingerne en gang om dagen. Hver enkelt borgers egenmåling bliver triageret og indplaceret i en risikokategori (grøn, gul eller rød) ud fra en simpel algoritme (beslutningstræ), som sammen med årsag til evt. gul eller rød triagering, returneres til borgeren umiddelbart efter målingen.

Anbefalingen til borgerne om at måle sig dagligt bygger ikke på evidens, men på antagelsen om, at uden daglige målinger har klinikken ikke nok grundlag til at vurdere, om der er behov for behandling, og at de dermed potentielt ikke kan fange forværringer i tide. Anbefalingerne om daglige målinger tager eksempelvis ikke højde for eventuelle hændelser eller omstændigheder, der tilsiger behov for ændringer i borgernes målerutiner.

Formålet med denne analyse er at undersøge det reelle målebehov, og potentielt komme med et databaseret estimat af den optimale målefrekvens, der resulterer i mindst mulig intervention i borgerens hverdag uden tab af information, samt at undersøge om borgerne kan segmenteres i flere risikogrupper med hver deres optimale måleregime.

2. Beskrivelse af datagrundlaget

2.1 Datakilder

Data anvendt til analyser beskrevet i denne rapport kommer fra forskellige kilder. Borgernes egenmålinger, informationer om borgerne samt informationer om deres behandlingsforløb i PreCareKlinikken stammer fra Appinux. Vejrdata er hentet fra DMI, og pollendata er tilsendt fra Astma-Allergi Danmark.

Bilag A viser et overblik over datamodellen, der beskriver hvor diverse datavariabler er hentet samt hvordan data fra forskellige kilder er koblet sammen.

2.2 Deskriptiv statistisk analyse og datafuldstændighed

Datagrundlaget består af egenmålinger udført af 362 unikke borgere i perioden 05-10-2018 til 27-09-2023. Borgere med mindre end 90 dages inklusionstid fjernes fra datasættet og inkluderes derfor ikke i analyser. Efter filtrering af borgere på inklusionstid består datagrundlaget af 136.991 målinger fra 312 unikke borgere. Ud af de 136.991 egenmålinger har 1.817 af dem været ledsaget af en medicinsk igangsættelse, hvoraf 726 af igangsættelserne er sket efter en rød egenmåling og 757 efter en gul egenmåling. Datasættet indeholder i alt 24.307 gule målinger og 4.407 røde målinger.

Tabel 9 viser et overblik over borgernes inklusionstid, gennemsnitlige målrutiner, antallet af medicinske igangsættelser samt fordelingen af grønne, gule og røde egenmålinger i perioden for dataopsamlingen. Af tabellen fremgår det, at borgerne i gennemsnit måler sig 4,3 gange om ugen. Borgernes målerutiner er også illustreret i Figur 9, som viser, at tæt på 100 pct. af alle borgere måler sig mellem 1 og 2 gange om ugen gennemsnitligt, over 85 pct. af borgerne måler sig 2-3 gange om ugen gennemsnitligt, mens mere end 75 pct. af borgerne måler sig 3-4 gange om ugen gennem-

snitligt. Det ugentlige gennemsnitlige antal målinger er beregnet som antal målinger i alt i inklusionsperioden delt med antallet af inkluderede uger, og er derfor blot et estimat, der ikke indeholder information om, hvordan målingerne fordeler sig over de inkluderede uger.

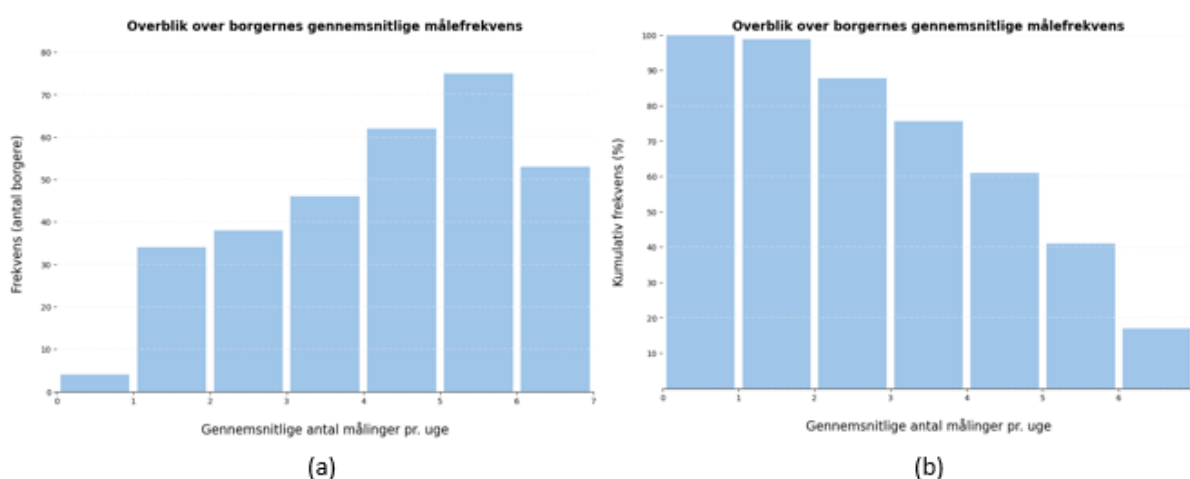
Tabel 9. Overblik over borgernes inklusionstid, målerutiner, medicinske igangsættelser samt fordeling af grønne, gule og røde målinger i perioden for dataopsamlingen

	Middel ($\pm$ std.)	Min.	25%-kvar-til	50%-kvar-til	75%-kvar-til	Max.
Inklusionstid (dage)	744 ($\pm$ 434)	90	368	689	1038	1771
Gennemsnitligt antal ugentlige målinger	4,3 ($\pm$ 1,6)	0,2	3,0	4,5	5,6	7,0
Antal igangsættelser i hele perioden	5,8 ($\pm$ 6,5)	0	1	4	8	41
Gennemsnitligt antal igangsættelser pr. inklusions-år*	3,1 ($\pm$ 2,7)	0,0	0,9	2,4	4,6	13,5
Antal gule målinger i hele perioden	77,9 ($\pm$ 73,5)	3	34	57	98	470
Gennemsnitligt antal gule målinger pr. inklusions-år*	44,0 ($\pm$ 35,4)	2,1	21,0	36,5	55,8	329,3
Antal røde målinger i hele perioden	14,1 ($\pm$ 15,7)	0	4	10	18	120
Gennemsnitligt antal røde målinger pr. inklusions-år*	9,1 ($\pm$ 10,7)	0,0	2,7	5,5	11,9	95,9

Note: * Gennemsnitligt antal målinger/igangsættelser pr. inklusionsår er beregnet som et mål for nemmere sammenligning af alle borgere, uanset varighed af inklusionstiden. For borgere, der har været inkluderet mindre end et år, vil dette derfor afspejle et estimat af antal målinger/igangsættelser pr. år med den antagelse, at den gennemsnitlige rate pr. år er uændret ift. raten i inklusionsperioden.

Kilde: PreCareKlinikken

Figur 9. Overblik over borgernes gennemsnitlige antal målinger pr. uge



Note: (a) viser den absolutte frekvens af borgere fordelt på antallet af gennemsnitlige målinger per uge, mens (b) viser den kumulative frekvens af borgere fordelt på antallet af gennemsnitlige målinger per uge.

Kilde: PreCareKlinikken

Tabel 10 viser overblikket over borgernes målerutiner, triageringer og antal igangsættelser for de fire risikogrupper hhv., som også er illustreret i Figur 10 for sammenligning af risikogrupperne. Af tabellen og figuren fremgår det, at borgernes gennemsnitlige målerutiner er sammenlignelige på tværs af risikogrupperne. Det fremgår også, at antallet af gule triageringer gennemsnitligt er større, des lavere risikogruppe, og samme tendens ses med antallet af røde triageringer samt antal medicinske igangsættelser. (Dog observeres stor spredning for borgernes målerutiner, triageringer samt antal igangsættelser).

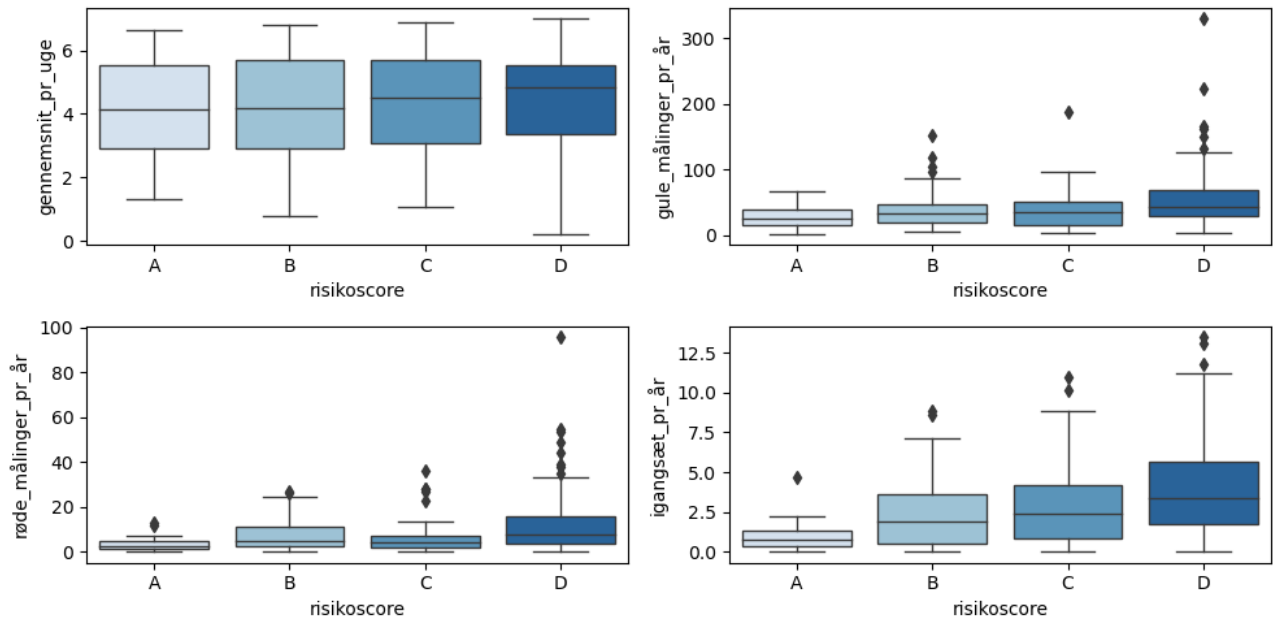
Tabel 10. Overblik over borgernes inklusionstid, målerutiner, medicinske igangsættelser samt fordeling af grønne, gule og røde målinger i perioden for dataopsamlingen grupperet på risikoscore ved inklusion

Risikogruppe	A	B	C	D
Antal borgere	19	85	70	134
Inklusionstid (dage)	849 ($\pm$ 398)	717 ($\pm$ 384)	971 ($\pm$ 424)	617 ($\pm$ 421)
Gennemsnitligt antal ugentlige målinger	4,1 ($\pm$ 1,6)	4,2 ($\pm$ 1,7)	4,3 ($\pm$ 1,6)	4,4 ($\pm$ 1,6)
Antal igangsættelser i hele perioden	2,4 ($\pm$ 2,2)	4,3 ($\pm$ 4,8)	7,4 ($\pm$ 6,8)	6,3 ($\pm$ 7,0)
Gennemsnitligt antal igangsættelser pr. inklusions-år*	1,0 ($\pm$ 1,1)	2,4 ($\pm$ 2,1)	3,0 ($\pm$ 2,7)	3,9 ($\pm$ 2,9)
Antal gule målinger i hele perioden	65,6 ($\pm$ 47,7)	65,2 ($\pm$ 48,9)	91,8 ($\pm$ 81,5)	81,1 ($\pm$ 84,0)
Gennemsnitligt antal gule målinger pr. inklusions-år*	29,6 ($\pm$ 17,7)	37,4 ($\pm$ 25,8)	38,8 ($\pm$ 29,9)	53,4 ($\pm$ 42,5)
Antal røde målinger i hele perioden	8,2 ($\pm$ 6,2)	13,0 ($\pm$ 14,4)	13,9 ($\pm$ 15,3)	15,6 ($\pm$ 17,5)
Gennemsnitligt antal røde målinger pr. inklusions-år*	3,7 ($\pm$ 3,6)	7,6 ($\pm$ 7,0)	6,3 ($\pm$ 7,1)	12,4 ($\pm$ 13,7)

Note: * Gennemsnitligt antal målinger/igangsættelser pr. inklusionsår er beregnet som et mål for nemmere sammenligning af alle borgere, uanset varighed af inklusionstiden. For borgere, der har været inkluderet mindre end et år, vil dette derfor afspejle et estimat af antal målinger/igangsættelser pr. år med den antagelse, at den gennemsnitlige rate pr. år er uændret ift. raten i inklusionsperioden.

Kilde: PreCareKlinikken

Figur 10. Boksplots over gennemsnitligt antal målinger pr. uge, gennemsnitligt antal gule målinger pr. år, gennemsnitligt antal røde målinger pr. år, samt gennemsnitligt antal igangsættelser pr. år grupperet på risikoscore ved inklusion



Kilde: PreCareKlinikken

Tabel 11. Deskriptiv statistisk overblik over egenmålingerne samt borgernes habituelværdier

Numeriske variable					
Variabel	Middel ($\pm$ std.)	Min.	Median	Max.	Manglende data
Lungefunktion	1,12 ($\pm$ 0,55)	0,03	1,01	9,07	4730 (3,46 %)
Puls	79,84 ($\pm$ 12,62)	20,00	79,00	245,00	2344 (1,71 %)
Temperatur	36,19 ($\pm$ 0,55)	34,90	36,20	41,90	2005 (1,47 %)
Iltmætning	94,04 ($\pm$ 4,54)	9,00	94,00	787,00	2340 (1,71 %)
Lungefunktion Habituelværdi	1,08 ($\pm$ 0,84)	0,21	1,00	155,00	351 (0,26 %)
Temperatur Habituelværdi	37,38 ($\pm$ 0,18)	35,00	37,40	37,50	0
Iltmætning Habituelværdi	89,30 ($\pm$ 6,18)	0,80	90,00	99,00	0
Puls Habituelværdi	98,79 ($\pm$ 6,71)	44,00	100,00	130,00	0
Lav puls Habituelværdi	53,93 ($\pm$ 0,46)	51,00	54,00	54,00	0
Kategoriske variable					
Variabel	% ja (antal)	% nej (antal)		Manglende data	
Farvet opspyt	0,7% (985)	99,1 (135.792)		0	
Øget åndenød	4,5% (6167)	95,3% (130.610)		0	
Øget hoste	2,4% (3334)	97,4% (133.443)		0	

Kilde: PreCareKlinikken

Borgernes egenmålinger er beskrevet i

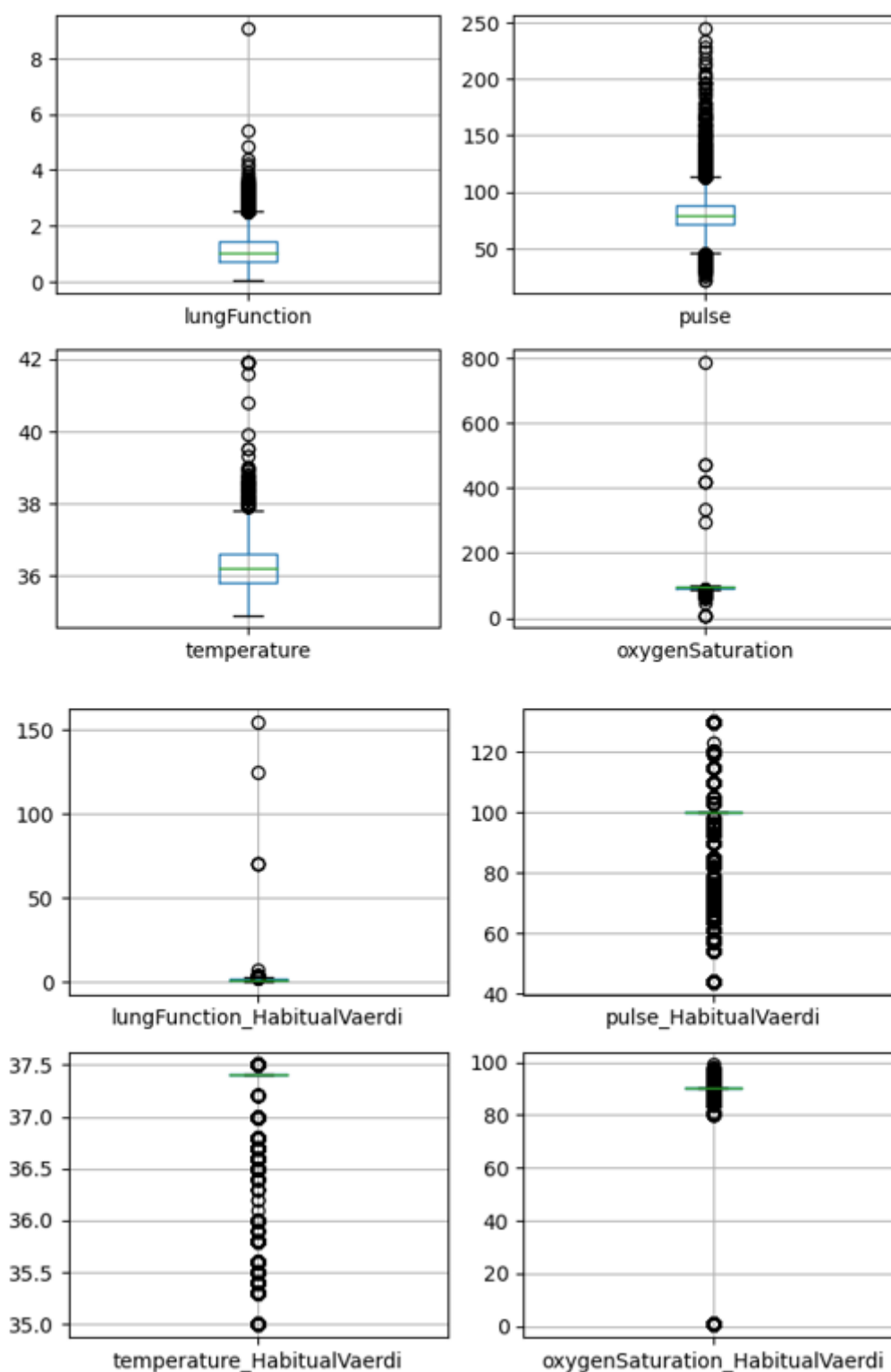
Tabel 11, der viser gennemsnit, minimum, median, maksimum samt antal manglende datapunkter

Numeriske variable					
Variabel	Middel ($\pm$ std.)	Min.	Median	Max.	Manglende data
Lungefunktion	1,12 ($\pm$ 0,55)	0,03	1,01	9,07	4730 (3,46 %)
Puls	79,84 ($\pm$ 12,62)	20,00	79,00	245,00	2344 (1,71 %)
Temperatur	36,19 ($\pm$ 0,55)	34,90	36,20	41,90	2005 (1,47 %)
Iltmætning	94,04 ($\pm$ 4,54)	9,00	94,00	787,00	2340 (1,71 %)
Lungefunktion Habituelværdi	1,08 ($\pm$ 0,84)	0,21	1,00	155,00	351 (0,26 %)
Temperatur Habituelværdi	37,38 ($\pm$ 0,18)	35,00	37,40	37,50	0
Iltmætning Habituelværdi	89,30 ($\pm$ 6,18)	0,80	90,00	99,00	0
Puls Habituelværdi	98,79 ($\pm$ 6,71)	44,00	100,00	130,00	0
Lav puls Habituelværdi	53,93 ($\pm$ 0,46)	51,00	54,00	54,00	0
Kategoriske variable					
Variabel	% ja (antal)	% nej (antal)		Manglende data	
Farvet opspyt	0,7% (985)	99,1 (135.792)		0	
Øget åndenød	4,5% (6167)	95,3% (130.610)		0	
Øget hoste	2,4% (3334)	97,4% (133.443)		0	

for hver af de 4 egenmålte variable samt de tilhørende habituelværdier og fordelingen i de tilstandsspørgsmål. Fordelingen i de egenmålte værdier og habituelværdierne er også vist i Figur 11. Fra tabellen og figuren kan vi se, at der er nogle enkelte ekstreme målepunkter i de egenmålte værdier samt de indtastede habituelværdier, og at disse er tydelige måle- eller tastefejl, eksempelvis lungefunktionsmålinger og lungefunktionshabituelværdier over 7, iltmætning over 100% eller under 70%. Ved nærmere undersøgelse blev det tydeligt, at de tydeligt fejlagtigt indtastede habituelværdier skyldtes tastefejl fra sygeplejerskerne, da disse hurtigt efterfølgende blev rettet til den korrekte værdi. For de egenmålinger med iltmætning over 100% manglede alle de andre egenmålte værdier, og ved nærmere undersøgelse blev det opdaget, at pulsoximetrene automatisk overfører ekstreme iltmålingsværdier, når de er ved at løbe tør for batteri.

Da triageringen af borgernes egenmålinger er direkte afledt af de indtastede habituelværdier og de målte/indtastede måleværdier, er det væsentligt at disse er korrekte, således at egenmålingen ikke resulterer i en falsk triagering, der potentielt kan være årsag til at en sygdomsforværring overses. For de egenmålinger med tydeligt fejlindtastede værdier (lungefunktion egenmåling samt habituelværdier og iltmætning egenmåling samt habituelværdier), er værdierne blevet rettet for de pågældende målinger for videre analyse af data. For pulsmålingerne er der også målinger, der formentlig er fejlmålinger, men disse er ikke rettet, da det er svært at skelne mellem de korrekte og forkerte målinger.

Figur 11: Boksplot over borgernes egenmålte lungefunktion, puls, temperatur og iltmætning



3. Analyse af triageringen i borgernes egenmålinger

3.1 Analyse af nuværende triageringsalgoritme (Appinux)

Da Appinux' triageringsalgoritme har ændret sig lidt løbende i projektets periode, implementeres et beslutningstræ svarende til den nuværende triageringsalgoritme. Alle egenmålingerne i datasættet er triageret på ny med denne algoritme (derfor vil triageringen ikke stemme fuldstændig overens med den triagering, der er registreret i Appinux).

Datagrundlaget indeholder i alt 22.195 gule målinger og 4.358 røde målinger fordelt på de 312 borgere, og langt hovedparten af disse målinger har ikke været ledsaget af en medicinsk igangsættelse. Tabel 12 viser triageringen af alle borgernes egenmålinger.

Tabel 12. Overblik over Appinux triagering af borgernes egenmålinger

Overblik over triagering af borgernes egen-målinger (Med nuværende Appinux triagerings-algoritme)			
Triagering	Antal i alt	Antal ledsaget af igangsættelse	Antal ikke ledsaget af igangsættelse
Grøn måling	110.224	348	109.876
Gul måling	22.195	631	21.564
Min. 3 på hinanden følgende gule*	1.819	268	1.548
Rød måling	4.358	624	3.734
I alt	136.777	1.603	135.174

Note: * Minimum 3 på hinanden følgende gule målinger er defineret som minimum 3 gule målinger, hvor den første måling ikke er foretaget mere end 14 dage før den sidste af de 3 målinger.

Kilde: PreCareKlinikken

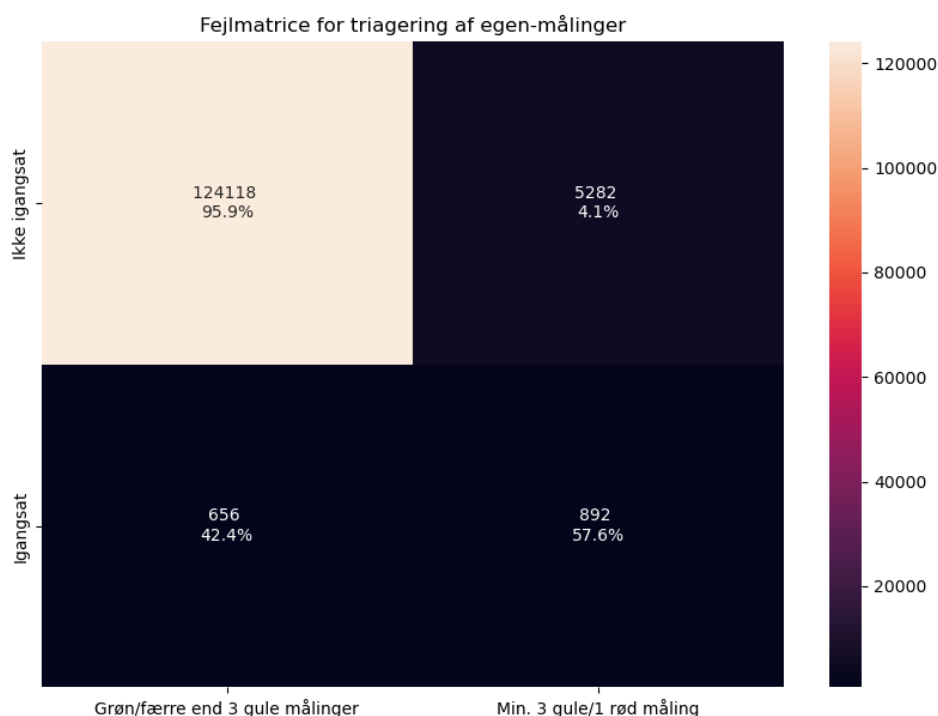
I tabellen ovenfor ses, at en stor andel af de røde eller minimum 3 på hinanden følgende gule målinger er falske positive, idet de ikke er ledsaget af en igangsættelse.

Figur 12 viser fejlmatrixen for Appinux-triageringen af borgernes egenmålinger, hvor sande positive er defineret som enten én rød egen-måling eller minimum tre på hinanden følgende gule målinger, der er ledsaget af en medicinsk igangsættelse, og sande negative er defineret som enten én grøn eller færre end tre på hinanden følgende gule egenmålinger, der ikke er ledsaget af en medicinsk igangsættelse.

Tabel 13

Tabel 13 viser et overblik over årsagen til gul og rød triagering af borgernes egenmålinger med og uden medicinsk igangsættelse. Ud af alle de gule og røde triageringer er lungefunktion årsag til triageringen for mere end 50 pct. af egenmålingerne. Af de 3.734 røde målinger, der ikke er ledsaget af en igangsættelse, er årsagen til den røde triagering fald i lungefunktion for 2.749 (74 pct.) af egenmålingerne, mens fald i lungefunktion er årsag til 10.938 (51 pct.) ud af de 21.564 gule egenmålinger, der ikke er ledsaget af en igangsættelse. Lungefunktion er altså den hyppigste årsag til gule og røde triageringer uden medicinsk igangsættelse (falske positive).

Figur 12. Fejlmatrixe for nuværende Appinux-triagering af borgernes egenmålinger



Kilde: PreCareKlinikken

Tabel 13. Oversigt over årsag til triagering af borgernes egenmålinger (med nuværende Appinux-algoritme)

Overblik over årsag til triagering af borgernes egenmålinger (Med nuværende Appinux triagerings-algoritme)					
Årsag til triagering	Gul triagering		Rød triagering		Alle
	Uden igang-sættelse	Med igang-sættelse	Uden igang-sættelse	Med igang-sættelse	
Lungefunktion	10.938	297	2.749	420	14.404
Puls	5.674	86	209	7	5.976
Opspyt har ændret farve	N/A	N/A	360	145	505
Iltindhold er for lavt	1.581	61	364	36	2.042
Temperatur er før høj	736	12	52	16	816
Øget åndenød eller hoste	668	28	N/A	N/A	696
I alt	21.564	631	3.734	624	26.553

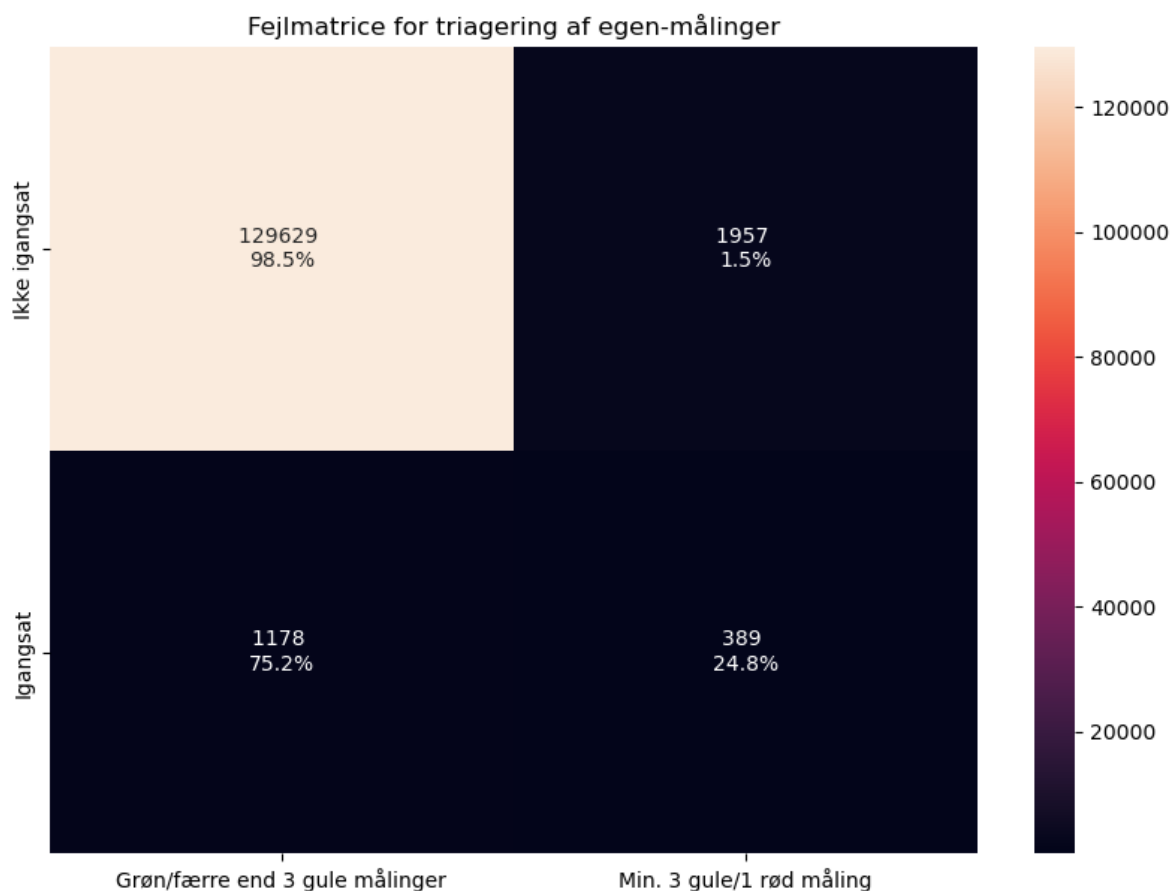
Kilde: PreCareKlinikken

3.2 Triageringsalgoritme uden lungefunktionsmåling

Den nuværende Appinux-triageringsalgoritme returnerer mange falske positive, hvor årsagen til triageringen er for lav lungefunktion i forhold til borgerens habituelle tilstand. Alle borgernes tilstandsmålinger triageres på ny af en ny triageringsalgoritme, der er tilsvarende Appinux-algoritmen, men hvor lungefunktionsmålingerne ikke er medtaget. Fejlmatrixen for triageringen uden lungefunktion er vist i Figur 13, hvor det fremgår, at andelen af falske positive er væsentligt mindre,

når lungefunktion ikke medtages (andelen af falske positive går fra ca. 4 pct. til ca. 1,5 pct.). På den anden side falder andelen af sande positive også fra 58 pct. til 25 pct., når lungefunktion ikke medtages, så på trods af at lungefunktionsmålingen resulterer i mange falske positive, tyder data også på, at denne er en god prædikator for en forværring, der kræver medicinsk behandling.

Figur 13. Fejlmatrixe for triagering af borgernes egenmålinger uden lungefunktionsmåling



Kilde: PreCareKlinikken

3.3 Triagering med udvidet datamodel

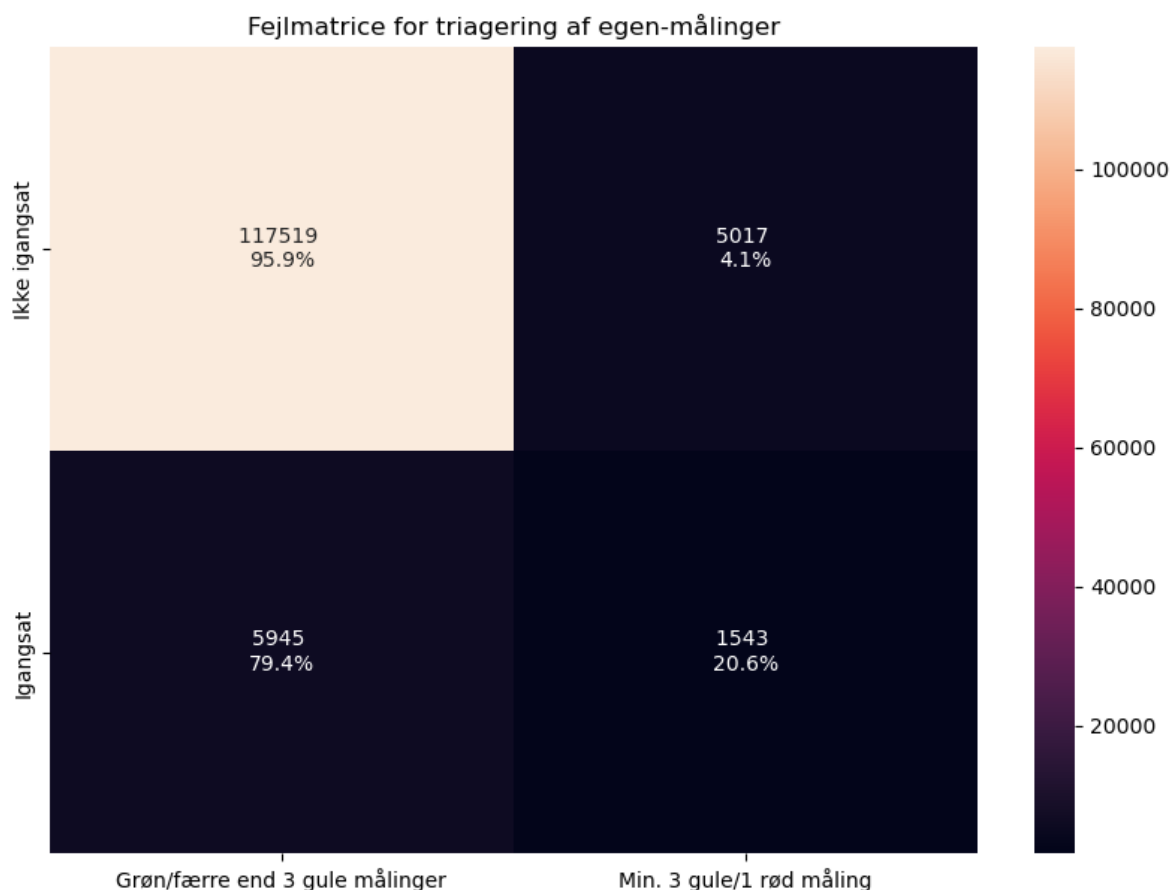
Med den nuværende datamodel, kobles et behandlingsforløb på den ene egenmåling, der er foretaget tættest på den medicinske igangsættelse, blandt de egenmålinger der er foretaget maksimalt en time efter igangsættelsen og er foretaget efter klokken 17:00 dagen forinden igangsættelsen. Dette betyder, at kun de målinger, der er foretaget indenfor et meget snævert tidsvindue relativt til den medicinske igangsættelse, får en positiv label (*igangsat*).

Baseret på antagelsen om, at der potentielt kan observeres en gradvis forværring i borgerens tilstand i dagene op til en medicinsk igangsættelse, udvides datamodellen, således at alle målinger, der er foretaget indenfor en uge forud for den medicinske igangsættelse, får tilknyttet en positiv label (*igangsat*). Figur 14 viser fejlmatrixen for triageringen med den udvidede datamodel, hvor alle målinger foretaget inden for en uge forud for igangsættelsen, får tilknyttet labelen *igangsat*. Fra fejlmatrixen fremgår det, at andelen af falske positive er uændret sammenlignet med triageringen ud fra den oprindelige datamodel (

Figur 12). Med den udvidede datamodel er andelen af falske positive ligeledes ca. 4 pct. (som med den oprindelige datamodel), mens andelen af sande positive er lavere med triageringen ud fra den udvidede datamodel.

Altså tyder det på, at de mange falske positive (gule og røde) målinger vi observerer, med den oprindelige datamodel ikke er centreret omkring den medicinske igangsættelse (og derved er et udtryk for den gradvise forværring af borgerens tilstand i dagene op til en medicinsk igangsættelse), men at de potentielt er fejlmålinger, der resulterer i en falsk triagering.

Figur 14. Fejlmatrix for triagering af borgernes egenmålinger med udvidet datamodel



Kilde: PreCareKlinikken

4. Analyse af variansen af borgernes egenmålinger

En forklaring på de mange falske positive triageringer er muligvis, at borgerne har svært ved at lave konsistente egenmålinger, der er retvisende for deres faktiske tilstand. For at undersøge dette, analyseres variansen i borgernes egenmålinger. Forud for variansanalysen fjernes alle egenmålinger, der ligger mindre end en uge før en medicinsk igangsættelse, da vi er interesseret i at se, hvor meget borgernes egenmålinger varierer, når de er i deres habituelle tilstand. Borgernes egenmålte puls, temperatur og iltmætning forventes at påtage en nogenlunde stabil middelværdi over hele inklusionsperioden (det forventes ikke, at nogen af de nævnte parametre forringes over tid). Mange borgerne har dog generelt en lungefunktion, der forringes over tid i inklusionsperioden, og standardafvigelsen beregnes derfor for perioder af én måned ad gangen (såfremt der er mere end 5 målinger den pågældende måned). Den samlede standardafvigelse for en specifik borger er

derefter beregnet som gennemsnittet af de månedlige standardafvigelser i hele inklusionsperioden. Denne beregningsmetode anvendes for at minimere effekten af en eventuelt nedadgående trend i lungefunktion i beregningen af den gennemsnitlige varians over tid.

Variansen for de fire egenmålte variable er opsummeret som gennemsnitlig standardafvigelse for alle inkluderede borgere i Tabel 14, der ligeledes viser den gennemsnitlige standardafvigelse for hver af de fire risikogrupper samt 5 aldersgrupper. Figur 15 viser fordelingen af standardafvigelsen for risikogrupperne for hver af de fire egenmålte parametre, hvor det fremgår, at der ikke er nogen signifikant forskel på standardafvigelserne for de 4 risikogrupper.

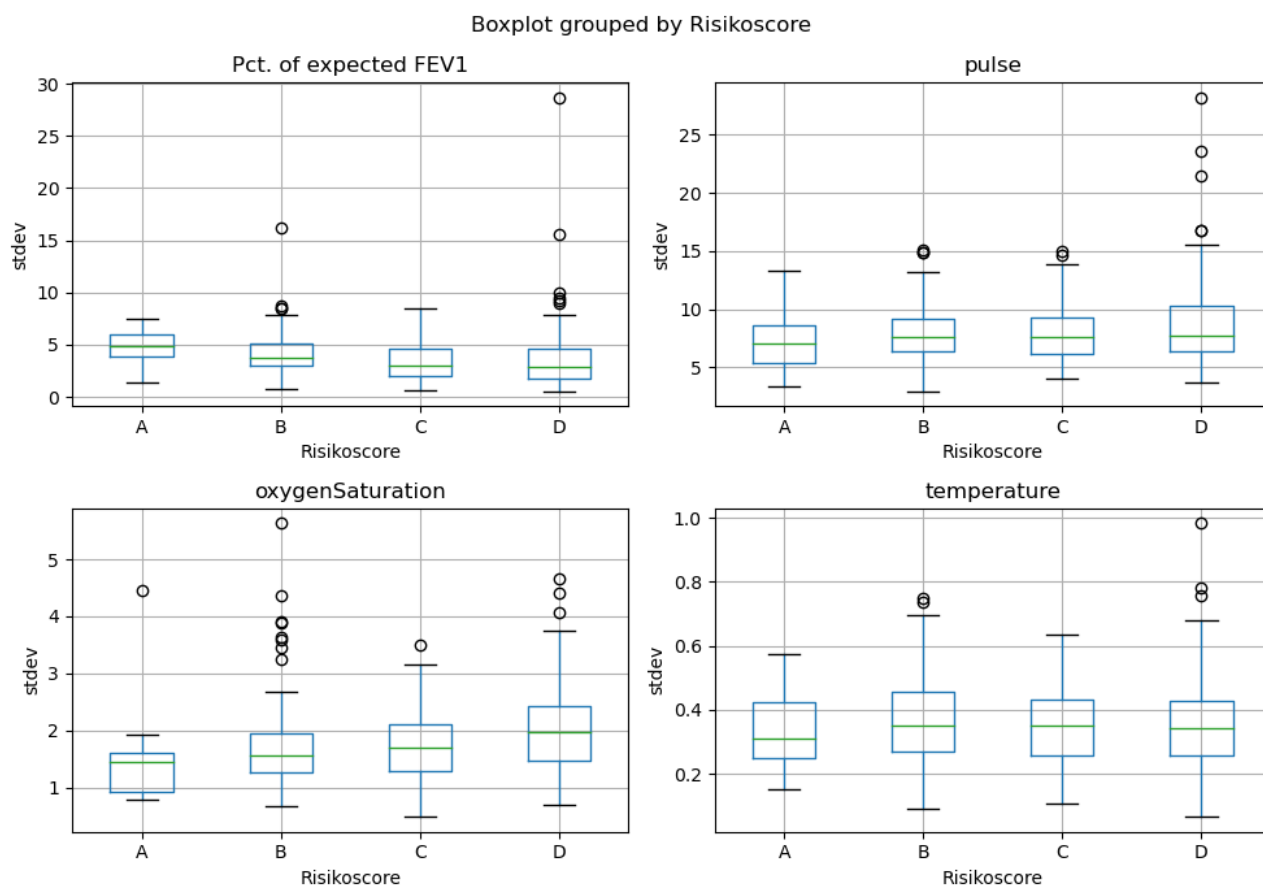
Tabel 14. Overblik over standardafvigelsen for borgernes egenmålinger fordelt på risikogrupper, aldersgrupper og for alle borgere samlet

Overblik over standardafvigelsen for borgernes egenmålinger fordelt på risikogrupper				
Risikogruppe	Lungefunktion (% af forv.)	Puls	Iltmætning	Temperatur
A	4,84	7,44	1,49	0,34
B	4,22	7,95	1,80	0,37
C	3,38	7,97	1,73	0,35
D	3,61	8,25	1,99	0,36
Overblik over standardafvigelsen for borgernes egen-målinger fordelt på aldersgrupper				
Aldersgruppe	Lungefunktion (% af forv.)	Puls	Iltmætning	Temperatur
<60 år	3,56	8,40	1,67	0,35
60 – 69 år	3,10	7,45	1,70	0,35
70 - 79 år	3,93	8,33	1,82	0,36
80 – 89 år	4,53	9,07	2,26	0,36
>90 år	5,17	10,41	2,56	0,36
Overblik over standardafvigelsen for alle borgernes egenmålinger				
Alle borgere	3,79	8,25	1,85	0,35

Note: Standardafvigelserne er angivet som den gennemsnitlige standardafvigelse for hele inklusionsperioden for alle borgere i den pågældende risikogruppe eller aldersgruppe ved inklusionstidspunktet.

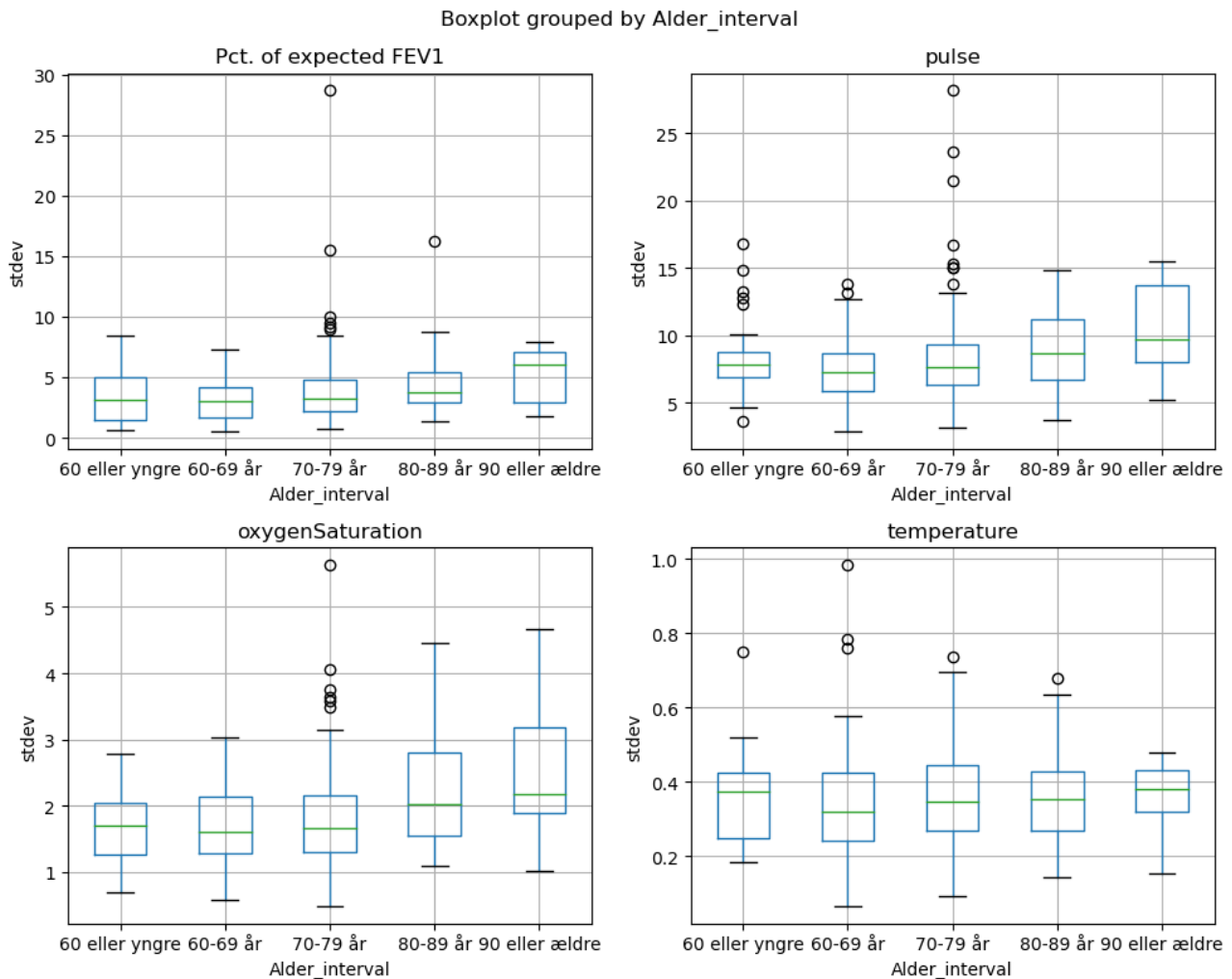
Kilde: PreCareKlinikken

Figur 15. Boxplot for standardafvigelsen af borgernes egenmålinger fordelt på risikogrupper



Kilde: PreCareKlinikken

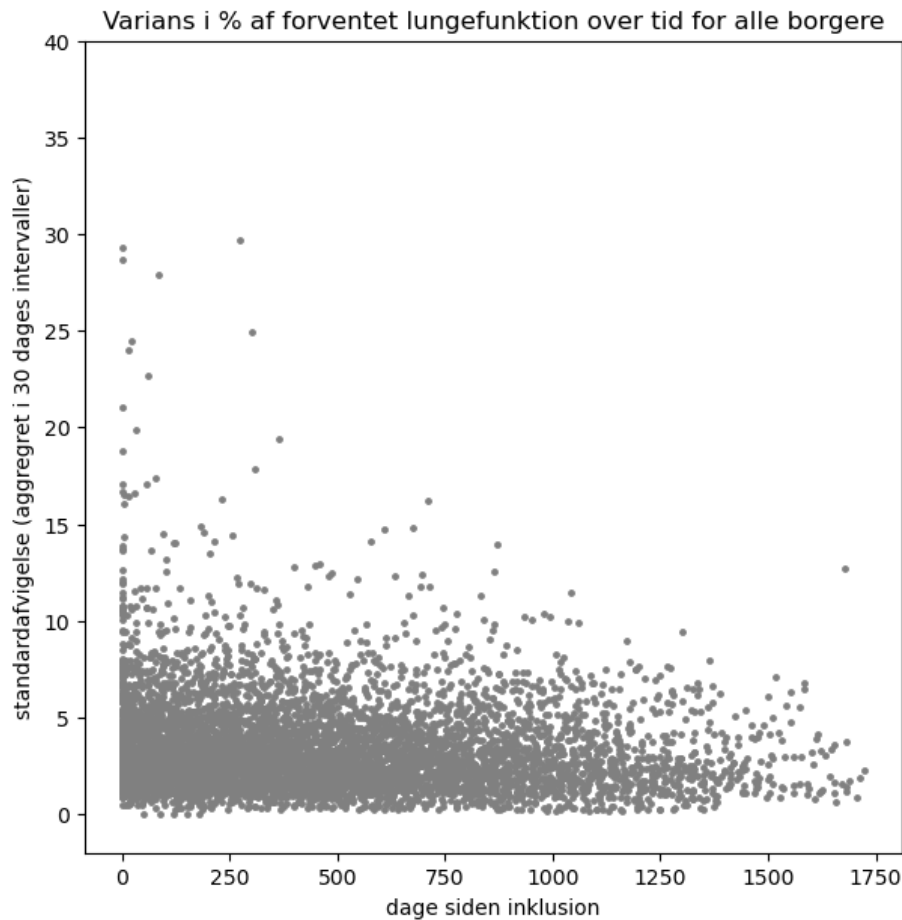
Figur 16. Boxplot for standardafvigelsen af borgernes egenmålinger fordelt på aldersgrupper



Kilde: PreCareKlinikken

Figur 16 viser fordelingen af standardafvigelsen for aldersgrupperne som boksplots for hver af de 4 egenmålte parametre, hvor det ligeledes fremgår, at der ikke er nogen signifikant forskel på standardafvigelserne for de 5 aldersgrupper.

Den observerede varians i borgernes egenmålinger kan muligvis skyldes, at borgerne har svært ved at anvende de udleverede måleapparater korrekt, hvilket resulterer i svingende målinger fra dag til dag, på trods af en uændret tilstand hos borgeren. Såfremt dette er tilfældet, kan det potentielt forventes at variansen falder over tid i takt med, at borgerne lærer at anvende apparaterne bedre og bedre. Figur 17 viser variansen for borgernes egenmålte lungefunktion (som % af forventet lungefunktion) over tid. Variansen er beregnet som en rullende standardafvigelse for 30-dages tidsperioder (såfremt tidsperioden har mere end 5 målinger), og plottet som funktion af dage siden inklusion for alle borgere individuelt. Fra plottet fremgår det, at variansen synes at være faldende over tid, hvilket kan tyde på, at borgerne bliver bedre til at måle deres lungefunktion med tiden. Analysen af varians over tid er kun lavet for lungefunktionsmålingerne, da dette skønnes at være den egenmåling, der er sværest for borgerne at foretage korrekt.

Figur 17. Standardafvigelsen for alle borgernes egenmålte lungefunktion over tid.

Note: Standardafvigelsen er beregnet som rullende standardafvigelse af 30 dages målinger for hele inklusionsperioden (Hvis der er færre end 5 målinger i en 30-dages tidsperiode er standardafvigelsen ikke vist i plottet).

Kilde: PreCareKlinikken

5. Modellering af data

For at kunne optimere på borgernes målerutiner er første skridt at sikre, at den triagering der sker på baggrund af egenmålingerne er pålidelig (fanger de faktiske forværringer i borgerens tilstand uden at resultere i for mange falske positive). Med den nuværende triageringsalgoritme resulterer den store varians i borgernes egenmålinger i mange falske positive og er derfor ikke et optimalt redskab til at opdage, hvornår borgeren faktisk oplever en forværret tilstand.

Den nuværende triageringsalgoritme er et simpelt beslutningstræ, der tager udgangspunkt i nogle habituelle værdier, som sygeplejerskerne har estimeret for hver enkelte borger. Til evaluering af om triageringsmetoden kan forbedres ved at anvende maskinlæring, er der foretaget forskellige afprøvninger af modellering af data. Der er afprøvet forskellige maskinlæringsmodeller på de egenmålte variable samt variable, der er afledt af disse, og desuden er forskellige eksterne variable medtaget i afprøvningen for at evaluere effekten af eksterne faktorer på borgernes tilstand. I alle

de afprøvede modeller beskrevet i de følgende afsnit er XGBoost¹⁵ anvendt som klassifikationsalgoritme uden yderligere hyperparameter-tuning.

5.1 Præparation af data

Forud for modellering er data præpareret og processeret med forskellige tiltag som beskrevet herunder.

For hver egenmåling i datasættet er % af forventet lungefunktion beregnet baseret på borgerens højde, alder og køn¹⁶. Den relative forskel mellem den pågældende måling (puls, temperatur, iltmætning og % af forventet FEV1) og den tilsvarende habituelværdi er beregnet¹⁷. Derudover er der afledt nogle nye variable: måned (kategorisk variabel svarende til den pågældende måned, målingen er foretaget), sæson (kategorisk variabel, der mapper den pågældende måling til en af de 4 sæsoner), samt lag for puls, temperatur, iltmætning og lungefunktion (svarende til den målte værdi for den pågældende variabel i den foregående måling). Eksterne faktorer såsom vejrdato (temperatur, luftfugtighed og lufttryk) samt pollental er ligeledes koblet på data for den pågældende dato for målingerne.

Datasættet er derefter splittet i et træningsdatasæt (80%) til at træne modellerne og et testdatasæt (20%) til at evaluere modellerne. I træningsdatasættet er de manglende værdier imputeret med MICE¹⁸, og de numeriske værdier er standardiseret¹⁹.

5.2 Evaluering af modelstørrelse

En række modeller med varierende antal inputvariable trænes og evalueres for at undersøge, om triageringen kan forbedres ved brug af maskinlæring, samt om triageringen kan forbedres ved at medtage eksterne faktorer såsom vejr, pollental og sæson. Først er der trænet en model på de rå egenmålinger alene (de 3 tilstandsspørgsmål, lungefunktion, puls, temperatur og iltmætning), og derefter er der trænet en række andre modeller med forskellige inputvariable. Et overblik over inputvariablene kan ses i Tabel 15, sammen med den balancerede nøjagtighed, sand-positivrate og falsk-positivrate for de forskellige modeller, når de er evalueret på testdatasættet.

Baseret på resultaterne i Tabel 15 tyder det på, at modellens evne til at kunne prædiktere igangsættelser ikke nødvendigvis bliver bedre med modelkompleksiteten. På trods af at de største modeller (Large og XLarge) har høj nøjagtighed skyldes dette, at antallet af falske positive er lavt, men antallet af sande positive er til gengæld også lavt. De største modeller er altså ikke ideelle til at fange igangsættelserne.

Den model med den højeste sand-positivrate er Medium-modellen, der er bygget på de 3 tilstandsspørgsmål samt FEV1 (% af forventet), temperatur, iltmætning og puls relativt til de respektive habituelværdier. Denne model er altså bygget på de samme inputvariable som Appinux-algoritmen, men opnår en væsentligt højere sand-positivrate på 73 pct. (sammenlignet med 58 pct. for Appinux-algoritmen). Dog ses der også en lidt højere falsk-positivrate på 10 pct. (sammenlignet med 4 pct. for Appinux-algoritmen).

¹⁵ [Python Package Introduction — xgboost 2.0.2 documentation](#)

¹⁶ Ved opslag i Appinux tabellen FEV1Reference.

¹⁷ Beregnet som måling – habituel-værdi/habituel-værdi.

¹⁸ Manglende værdier er imputeret ved brug af Multiple Imputation by Chained Equations (MICE) ([miceforest · PyPI](#)).

¹⁹ Hver af de numeriske variable er skaleret med MaxAbsScaler, der skalerer data således at den maksimale absolutte værdi for den pågældende variabel svarer til 1. Skaleringen er fittet på træningsdata og derefter brugt til at transformere testdatasættet ([sklearn.preprocessing.MaxAbsScaler — scikit-learn 1.3.2 documentation](#)).

Tabel 15 viser også at XXSmall-modellen, der alene er baseret på de 3 tilstandsspørgsmål, opnår en sand-positivrate på 61 pct., mens falsk-positivrate holdes nede på 5 pct. XSmall-modellen, der inkluderer de samme parametre som Medium-modellen undtagen lungfunktion resulterer i en sand-positivrate på 62 pct. og en falsk-positivrate på 9 pct.

Når de eksterne faktorer, såsom vejr og sæson, inkluderes i modellen, observeres et fald i sand-positivrate.

Tabel 15. Oversigt over inputvariable i de afprøvede maskinlæringsmodeller af forskellig kompleksitet samt den balancerede nøjagtighed (F1-score), sand-positivrate (TPR) og falsk-positivrate (FPR)

Model	Inkluderede parametre	Balanceret nøjagtighed	TPR	FPR
Baseline	Tilstandsspørgsmål Lungfunktion Temperatur Iltmætning Puls	94%	62,7%	9,6%
XXSmall	Tilstandsspørgsmål	96%	61,0%	5,0%
XSmall	Tilstandsspørgsmål Temperatur relativt til habituelværdi Iltmætning relativt til habituelværdi Puls relativt til habituelværdi	94%	62,0%	8,8%
Small	Tilstandsspørgsmål Vejrdata ²⁰ Pollendata ²¹ Måned Sæson	95%	52,6%	8,1%
Medium	Tilstandsspørgsmål Lungfunktion (% af forventet) relativt til habituelværdi Temperatur relativt til habituelværdi Iltmætning relativt til habituelværdi Puls relativt til habituelværdi	94%	73,2%	9,6%
Large	Tilstandsspørgsmål Lungfunktion (% af forventet) relativt til habituelværdi Temperatur relativt til habituelværdi Iltmætning relativt til habituelværdi Puls relativt til habituelværdi Euclidian distance ²² Lag 1	94%	68,8%	8,7%
XLarge	Tilstandsspørgsmål Lungfunktion (% af forventet) relativt til habituelværdi Temperatur relativt til habituelværdi Iltmætning relativt til habituelværdi	95%	62,0%	7,2%

²⁰ Vejrdata inkluderer dagens gennemsnitlige målte temperatur, luftfugtighed og lufttryk.

²¹ Pollendata inkluderer dagens målte pollental for El, Hassel, Elm, Birk, Græs, Alternaria, Cladosporium.

²² Euclidian distance er beregnet som summen af lungfunktion, temperatur, iltmætning og puls (relativt til de respektive habitue-værdier) kvadreret.

Puls relativt til habituelværdi Euclidian distance ²³ Lag 1 - 5 ²⁴
--

Kilde: PreCareKlinikken

5.3 Evaluering af prædiktionsvindue

I den nuværende datamodel er det kun den ene måling på dagen for en igangsættelse, der får tildelt en positiv label. Dette betyder, at modellen ikke er trænet til at genkende eventuelle mønstre i data for de målinger, der er foretaget i dagene op til en medicinsk igangsættelse. Ideelt set, vil vi gerne kunne prædiktere en sygdomsforværring i så god tid, at sygeplejerskerne eller borgerne selv kan initiere tiltag, der potentielt kan forhindre en medicinsk igangsættelse. Derfor testes forskellige prædiktionsvinduer (antal dage forud for en igangsættelse), for at undersøge om vi kan prædiktere igangsættelserne flere dage, før de forekommer. Prædiktionsvinduer på 0 til 7 dage er testet med XXSmall-modellen og Medium-modellen (prædiktionsvinduet på 0 dage svarer til modellerne beskrevet i det foregående afsnit). Prædiktionsvinduer er testet således, at alle målinger, der ligger inden for det specifikke prædiktionsvindue får en positiv label. Modellerne er derefter trænet på 80 pct. af data og evalueret på de resterende 20 pct. Resultaterne af modevalueringen for de 7 afprøvede prædiktionsvinduer er vist i Tabel 16.

For XXSmall-modellen kan vi se, at falsk-positivraten (FPR) forbliver stort set uændret uanset prædiktionsvinduet, mens sand-positivraten (TPR) falder, når prædiktionsvinduet udvides. Dette kan indikere, at det er svært at forudsige sygdomsforværringer flere dage i forvejen alene baseret på de 3 tilstandsspørgsmål (fejlmatrixerne er vist i bilag C). For Medium-modellen stiger sand-positivraten derimod, når prædiktionsvinduet udvides til en dag, hvorefter den falder en smule i takt med, at prædiktionsvinduet udvides yderligere. Falsk-positivraten stiger dog også i takt med, at prædiktionsvinduet udvides. Det tyder altså på, at mange af sygdomsforværringerne kan forudsiges flere dage forud for en igangsættelse på trods af, at modellen også returnerer en større andel af falske positive (Fejlmatrixerne er vist i bilag D).

Fordelingen af konfidensscoren for Medium-modellen med de 8 prædiktionsvinduer er vist i Figur 18, hvor de sande igangsættelser i testdatasættet er vist med orange søjler, og de sande ikke-igangsættelser er vist med blå søjler. Ved en perfekt model vil alle de sande igangsættelser have en konfidensscore på 1 og derved være placeret i højre side af søjlediagrammet fuldstændigt adskilt fra de sande ikke-igangsættelser, der vil være placeret yderst i venstre side med en konfidensscore omkring 0. Med et prædiktionsvindue på 0 dage observerer vi, at langt størstedelen af de sande ikke-igangsættelser får en lav konfidensscore, og mens konfidensscoren for de sande igangsættelser er spredt mere ud, dog med en stor del af dem mellem 0,8 og 1. Når prædiktionsvinduet udvides, overlapper de sande igangsættelser og ikke-igangsættelser i stigende grad. Trods den faldende grad af adskillelse, tildeles størstedelen af de sande ikke-igangsættelser en prædikeret sandsynlighed under 0,5, mens størstedelen af de sande igangsættelser tildeles en sandsynlighed over 0,5 med et prædiktionsvindue på 7 dage, hvilket indikerer, at vi til en vis grad kan forudsige igangsættelser, men med stigende grad af sikkerhed tættere på selve igangsættelsen.

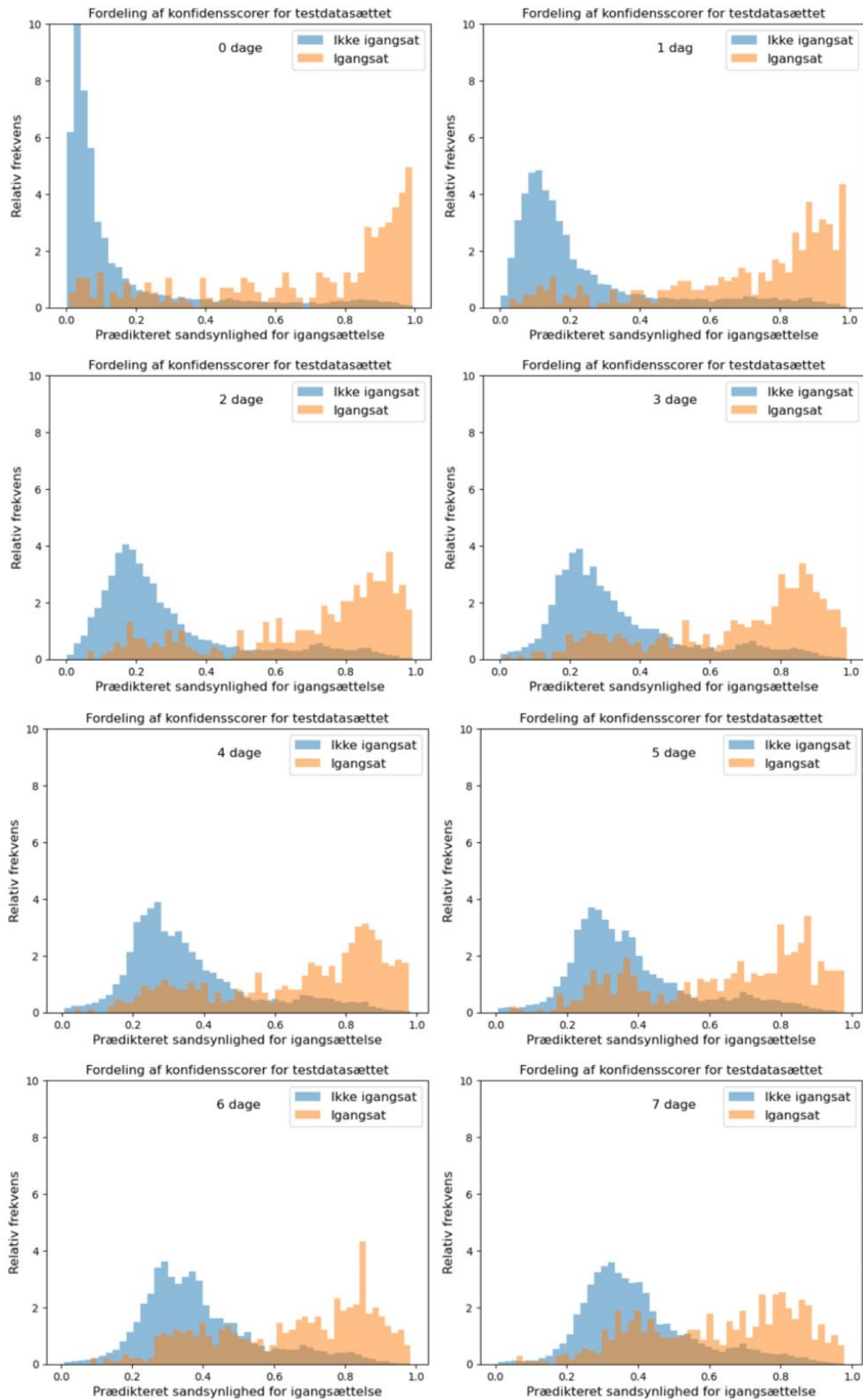
²³ Beregnet som summen af lungefunktion, temperatur, iltmætning og puls (relativt til habituelværdierne) kvadreret.

²⁴ Lag er medtaget for lag 1 til lag 5 (altså i alt 5 individuelle lag features)

Tabel 16. Overblik over resultater af evaluering af XXS og M modellerne med prædiktionsvinduer mellem 0 og 7 dage

Prædiktions- vindue	XXSsmall model			Medium model		
	Balanceret nøjagtighed	TPR	FPR	Balanceret nøjagtighed	TPR	FPR
0 dage	96%	61,0%	5,0%	94%	73,2%	9,6%
1 dag	96%	58,2%	5,0%	91%	79,4%	13,6%
2 dage	96%	55,0%	4,9%	90%	79,0%	15,3%
3 dage	96%	50,7%	4,9%	89%	76,0%	17,1%
4 dage	96%	46,7%	4,9%	88%	74,2%	17,9%
5 dage	95%	43,7%	4,9%	87%	71,5%	19,6%
6 dage	95%	41,2%	4,9%	86%	68,0%	20,7%
7 dage	95%	38,7%	4,9%	85%	68,0%	22,0%

Figur 18. Fordeling af konfidensscore for M-modellen med prædiktionsvinduer 0-7 dage.



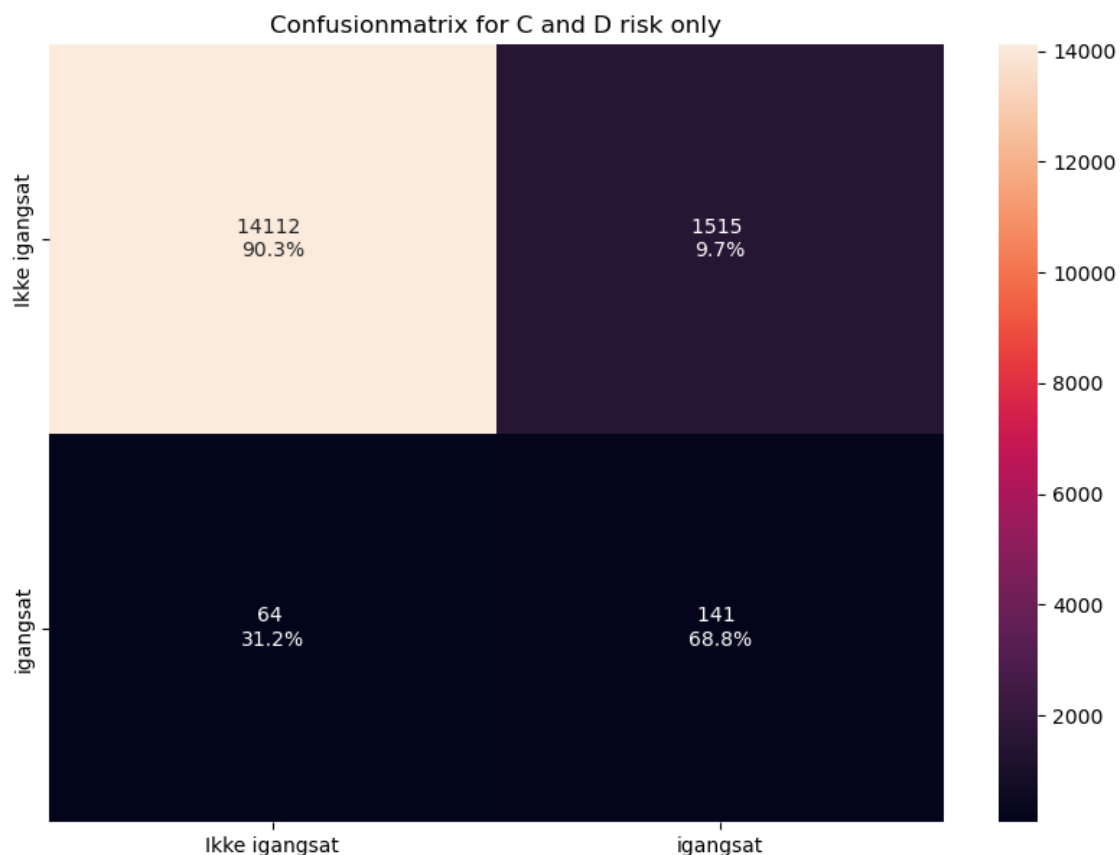
Note: Alle datapunkter i testdatasættet er repræsenteret i histogrammet, hvor x-aksen angiver den prædikterede sandsynlighed for en igangsættelse, mens y-aksen viser den relative frekvens. Alle sande ikke-igangsættelser er vist med blå søjler, mens de sande igangsættelser er vist med orange søjler.

Kilde: PreCareKlinikken

5.4 Evaluering af prædiktionsmålgruppe

I datasættet indgår KOL-borgere i alle risikogrupper (A-D), hvor der forventeligt kan være ret stor forskel i borgernes sygdomstilstand mellem de forskellige risikogrupper, hvor risikogruppe C og D er de borgere, der har den sværeste form for KOL, og derfor generelt oplever generelt flere sygdomsforværringer. Mønstrene i borgernes egenmålinger kan derfor også potentielt variere mellem risikogrupperne, hvilket kan resultere i støj i modeltræningen. For at undersøge dette er der trænet en ny prædiktionsmodel, hvor kun data fra borgere i risikogruppe C og D (ved inklusion) er medtaget i modeltræning og evaluering. Der er i alt 83.518 målinger fra borgere, der var i risikogruppe C og D ved inklusion i PreCareKlinikken. Alle egenmålinger fra disse borgere er splittet i et træningsdatasæt (80 pct.) og et testdatasæt (20 pct.), og en ny prædiktionsmodel er trænet på træningsdatasættet med de samme input variable som Medium-modellen beskrevet i Tabel 15 og evalueret på testdatasættet. Modellen opnår en vægtet F1 nøjagtighed på 94 pct. med en sand-positivrate på 69 pct. og en falsk-positivrate på 10 pct. (fejlmatrixen er vist i Figur 19). Modellen bliver altså tilsyneladende ikke bedre af at fjerne data fra de borgere, der var i risikogruppe A og B ved inklusion, hvilket kan skyldes, at datamængden i sin helhed bliver væsentligt mindre, hvilket gør det sværere for en model at lære de mønstre, der er karakteristiske for en sygdomsforværring.

Figur 19. Fejlmatrixe for Medium-prædiktionsmodellen trænet på borgere i risikogruppe C og D



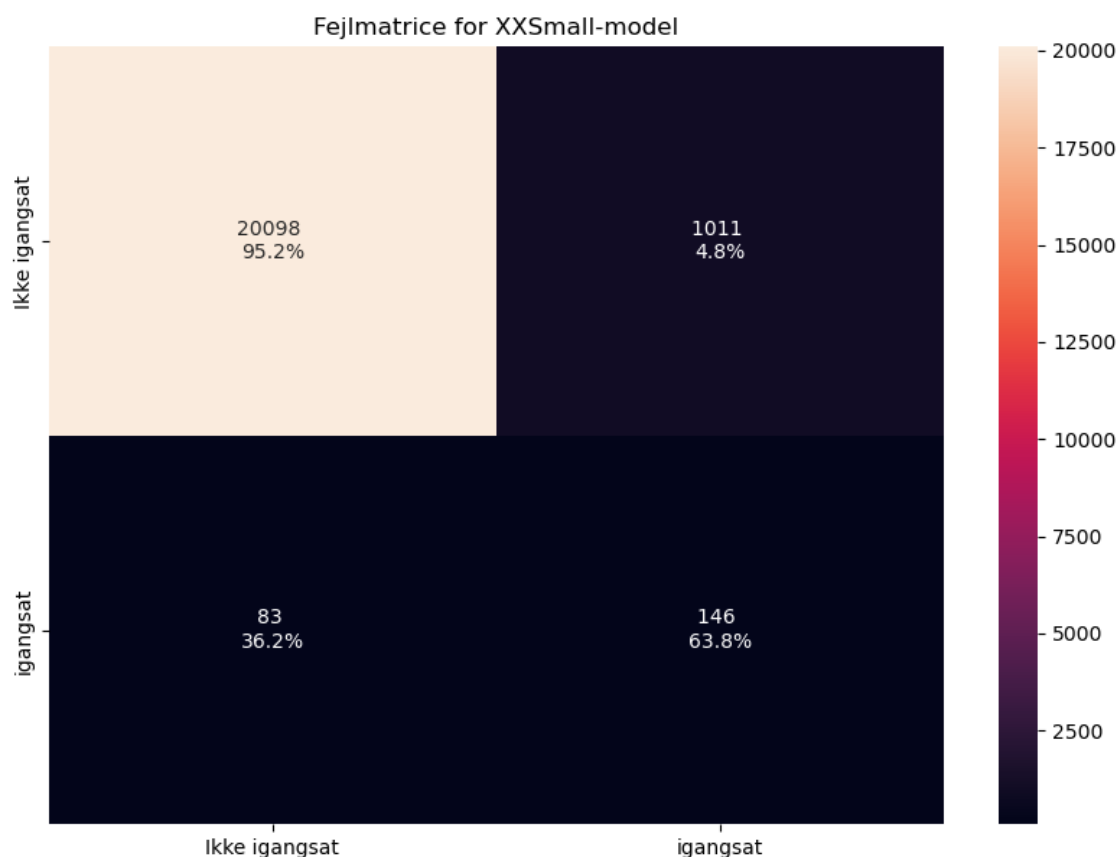
Kilde: PreCareKlinikken

5.5 Evaluering af generaliserbarheden for prædiktionsmodellen

Alle modellerne beskrevet i de foregående afsnit er trænet på 80 pct. af data og derefter evalueret på 20pct. af data ved tilfældigt at splitte alt data op i de specificerede proportioner. På den måde kan data fra én specifik borger forekomme både i træningsdatasættet og i testdatasættet, hvorfor det er svært at sige noget om, hvor god modellen er til at prædiktere forværringer hos nye borgere, hvis data ikke har været en del af træningsdatasættet. For at evaluere hvor generaliserbar modellen er (altså hvor godt den fungerer på helt nye borgere), testes XXSmall-modellen og Medium-modellen igen, men denne gang er data splittet således, at data fra de enkelte borgere kun forekommer i *enten* træningsdatasættet *eller* testdatasættet (data splittes stadig i 80 pct. træningsdata og 20 pct. testdata). Med denne metode til dataopdelingen opnår XXSmall modellen en vægtet F1 nøjagtighed på 97 pct. med en sand-positivrate på ca. 64 pct. og en falsk-positivrate på ca. 5 pct. (fejlmatricen er vist i Figur 20). Medium-modellen opnår en vægtet F1 nøjagtighed på 94 pct. med en sand-positivrate på ca. 72 pct. og en falsk-positivrate på ca. 9pct. (fejlmatricen er vist i Figur 21).

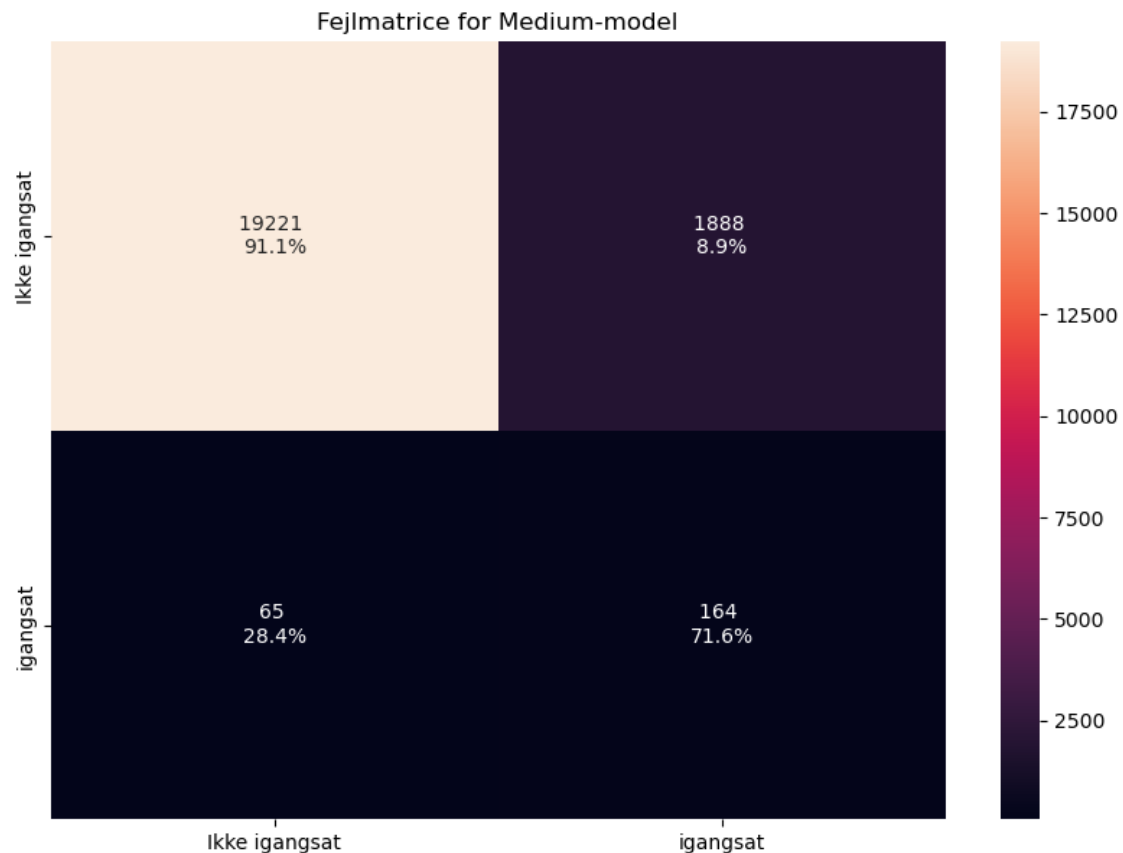
Evalueringen af de to afprøvede modeller tyder altså på, at modellerne er generaliserbare, når modellerne bruges til at prædiktere igangsættelser hos borgere, hvis data ikke indgår i træningsdatasættet, da resultaterne er sammenlignelige med modellerne, hvor data er splittet, således at data fra enkelte borgere kan forekomme både i trænings- og testdatasættet.

Figur 20. Fejlmatrice for XXSmall-prædiktionsmodellen når data fra individuelle borgere enten indgår i træningsdata eller i testdata



Kilde: PreCareKlinikken

Figur 21. Fejlmatrixe for Medium-prædiktionsmodellen når data fra individuelle borgere enten indgår i træningsdata eller i testdata



Kilde: PreCareKlinikken

6. Opsummering og diskussion

Formålet med analysen var at lave en dybdegående analyse af borgernes målerutiner, for at undersøge om det opsamlede data fra borgernes egenmålinger kan sige noget om en optimal målefrekvens. Borgerne er i perioden for dataopsamlingen blevet anbefalet at udføre egenmålingerne dagligt. Denne anbefaling bygger ikke på evidens og er ikke nødvendigvis den mest optimale målefrekvens. Optimalt set skal borgernes målerutiner være så lidt intervenserende i borgerens hverdag som muligt, men stadig kunne opsnappe forværringerne i borgerens tilstand før borgeren bliver indlæggelseskrævende. De følgende afsnit opsummerer de vigtigste fund og konklusioner i analysen af borgernes målerutiner.

6.1 Datagrundlaget

Analysen af datagrundlaget med de 136.991 egenmålinger opsamlet fra 312 borgere i perioden 05-10-2018 til 27-09-2023, viser at borgerne i gennemsnit måler sig 4,3 gange om ugen, og denne målefrekvens er sammenlignelig på tværs af risikogrupperne (Risikogruppe A måler sig i gennemsnit 4,1 gange om ugen, risikogruppe B 4,2 gange om ugen, risikogruppe C 4,3 gange om ugen, og risikogruppe D måler sig i gennemsnit 4,4 gange om ugen).

Ud af alle egenmålingerne er der kun ganske få, der mangler én eller flere af de målte variable. Lungefunktionsmålingen er den, der oftest mangler (i ca. 3,5% af målingerne), mens puls og iltmætning mangler i ca. 1,7% og temperatur i 1,5% af målingerne. De 3 tilstandsspørgsmål blev først obligatoriske at besvare i september 2022, og det er ikke muligt at se i data, om spørgsmålene er besvaret eller ej før denne dato (ubesvarede spørgsmål vil svare til at borgeren har svaret nej til de tre spørgsmål).

De 312 borgere har i gennemsnit 3,1 medicinske igangsættelser pr. inklusionsår, og antallet af igangsættelser er generelt højere des lavere risikogruppe borgeren var i ved inklusionstidspunktet.

Ved analyse af de 4 egenmålte variables fordeling blev det tydeligt, at der findes nogle måle- eller tastefejl for nogle af egenmålingerne og de registrerede habitueltværdier, eksempelvis lungefunktion over 7, iltmætning over 100 pct. eller pulsmålinger over 200. En korrekt triagering af borgernes egenmålinger beror på korrekte måleværdier og habitueltværdier, og fejlagtige værdier kan potentielt medføre, at en forværring i en borgers tilstand ikke afspejles i triageringen, eller at triageringen returnerer en forhøjet risiko for forværring hos en borger, der er i sin habituelle tilstand. Dette belyser et behov for en straks-validering af de målte og indtastede værdier, eksempelvis i form af absolutte grænseværdier, eller hvis den målte/indtastede værdi afviger signifikant fra den forrige værdi, kan det udløse en meddelelse, der spørger borgeren eller klinikerens om den målte/indtastede værdi er korrekt.

6.2 Triagering af borgernes egenmålinger

Analysen af triageringen af borgernes egenmålinger viser, at den nuværende Appinux-algoritme fanger ca. 58 pct. af igangsættelserne men også resulterer i en del falske positive (en falsk-positiv-rate på 4,1pct.), og at årsagen til de falske positive hovedsageligt skyldes lav lungefunktion (74 pct. af de røde egenmålinger, der ikke er ledsaget af en medicinsk igangsættelse skyldes fald i lungefunktion, mens fald i lungefunktion er årsag til 51 pct. af de gule målinger, der ikke er ledsaget af en medicinsk igangsættelse). Lungefunktion er altså den hyppigste årsag til gule og røde triageringer uden medicinsk igangsættelse (falske positive).

Triageringsalgoritmen blev afprøvet uden lungefunktionsmålingen, hvorved andelen af falske positive blev reduceret til 1,5 pct., men samtidigt faldt andelen af sande positive også til 25 pct. Dette indikerer, at store udsving i borgernes målte lungefunktion er årsag til mange falske positive, men at lungefunktionsmålingen stadigvæk har en stor kraft i forhold til at opdage forværringer, der leder til medicinsk igangsættelse hos borgerne med den nuværende Appinux triagerings-algoritme.

6.3 Variansanalyse

Variansen i borgernes egenmålinger er analyseret, hvor det blev særligt tydeligt, at borgerne har svært ved at udføre lungefunktionsmålingerne korrekt. Borgernes lungefunktionsmålinger (målt som % af forventet FEV1) har i gennemsnit en standardafvigelse på 3,8 pct. Det vil sige, at alle målingerne i gennemsnit afviger med 3,8 pct. fra middelværdien (og nogle målinger altså afviger mere, og andre målinger afviger mindre). For borgere med en generel dårlig lungefunktion, vil de observerede almindelige udsving i den målte lungefunktion hurtigt resultere i en gul eller endda rød triagering af egenmålingen på trods af borgeren ikke oplever en reel tilstandsforværring. De store udsving i lungefunktionsmålingerne gør det udfordrende at bruge disse målinger i en valid triagering af borgernes tilstand.

Analysen af variansen i lungefunktionsmålingerne indikerer desuden, at variansen bliver en smule mindre des flere dage de enkelte borgere har været inkluderet i PreCareKlinikken, hvilket kan tyde

på, at borgerne muligvis bliver bedre til at udføre korrekte egenmålinger med tiden. Generelt belyser analysen et behov for øget fokus på at undervise borgerne i at udføre korrekte lungefunktionsmålinger eller at assistere dem virtuelt, eksempelvis i form af en digital hjælper, der guider borgeren undervejs og beder borgeren gentage en måling, såfremt den virker urealistisk.

Der observeres også en smule varians i de resterende egenmålte variable, hvor puls i gennemsnit varierer med ca. 8 slag i minuttet, iltmætningen varierer gennemsnitligt med 1,85 pct, og temperaturen varierer gennemsnitligt med 0,35°C, men variationen i lungefunktionsmålingerne vurderes at være den mest kritiske i forhold til validiteten af egenmålingen.

6.4 Modellering af prædiktionsalgoritme

Maskinlæring er afprøvet for at vurdere om den nuværende Appinix triageringsalgoritme (der består af et simpelt beslutningstræ) kan forbedres. Én type maskinlæringsalgoritme (XGBoost) er afprøvet og ikke optimeret/tunet, da formålet er at vurdere muligheden for at anvende maskinlæring til triagering samt at evaluere vigtigheden af de forskellige variable ved forudsigelse af sygdomsforværring.

Effekten af modelkompleksiteten blev evalueret ved at bygge forskellige prædiktionsmodeller med varierende input variable, hvor den simpleste model er baseret på de 3 tilstandsspørgsmål alene, og de mere komplekse modeller består af de egenmålte variable, variable der er afledt af de målte variable samt eksterne variable såsom vejrdato, pollen data og sæson.

Baseret på afprøvningen tyder det på, at den bedste model opnås ved at inkludere de tre tilstandsspørgsmål, den målte puls (relativt til borgerens habituelværdi), den målte temperatur (relativt til borgerens habituelværdi), den målte iltmætning (relativt til borgerens habituelværdi), samt % af forventet FEV1 (relativt til borgerens habituelværdi), hvorved der opnås en estimeret sand-positivrate på 73,2 pct., og en estimeret falsk-positivrate på 9,6 pct. Modellen er altså bygget på de samme inputvariable som den nuværende Appinix triageringsalgoritme men er i stand til at detektere væsentligt flere igangsættelser, dog på bekostning af at der også returneres flere falske positive.

Den simpleste model, bestående af de 3 tilstandsspørgsmål alene, opnår en sand-positivrate på 61% med en falsk-positivrate på 5 pct. Dette indikerer at modellen er i stand til at detektere forværringer en smule bedre end den nuværende Appinix triageringsalgoritme uden at returnere signifikant flere falske positive, hvilket understreger den prædiktive kraft i de 3 tilstandsspørgsmål. Dette belyser samtidigt et potentiale i at erstatte nogle af de daglige egenmålinger med en række spørgsmål, der er nemme for borgerne at besvare, og som potentielt kan detektere forværringer i borgerens tilstand, hvorved borgeren kan promptes for at udføre en fuld egenmåling.

Forskellige prædiktionsvinduer blev testet, for at evaluere om det er muligt at forudsige forværringer i borgernes tilstand flere dage forud for en medicinsk igangsættelse. Prædiktionsvinduer på 0–7 dage blev testet, hvor 0 dage svarer til en prædiktionsvindue på dagen for igangsættelse, og 7 dage svarer til at kunne prædiktere en forværring op til 7 dage forud for igangsættelsen. De forskellige prædiktionsvinduer blev afprøvet for den simpleste model (med de 3 tilstandsspørgsmål som input variable), og den model der opnåede den højeste sand-positivrate. Afprøvningen af prædiktionsvinduerne viste, at sand-positivraten falder markant med den simpleste model, i takt med at prædiktionsvinduet udvides (med et prædiktionsvindue på 3 dage falder sand-positivraten til ca. 51 pct., og et prædiktionsvindue på 7 dage resulterer i en sand-positivrate på 39 pct., mens falsk-positivraten forbliver på ca. 5 pct. uanset prædiktionsvinduet).

For den mere komplekse model observeres derimod en umiddelbar stigning i sand-positivraten, når prædiktionsvinduet udvides. Med et prædiktionsvindue på 3 dage er sand-positivraten 76 pct.,

og et prædiktionsvindue på 7 dage resulterer i en sand-positivrate på 68 pct. Dog stiger falsk-positivraten også i takt med at prædiktionsvinduet udvides. Dette indikerer at målinger op imod 7 dage forud for en igangsættelse indeholder information, der kan bruges til at forudsige en medicinkrævende tilstandsforværring hos borgeren og underbygger hypotesen om, at det ikke er nødvendigt for borgerne at måle sig hver eneste dag for at kunne opdage forværringer i borgerens tilstand.

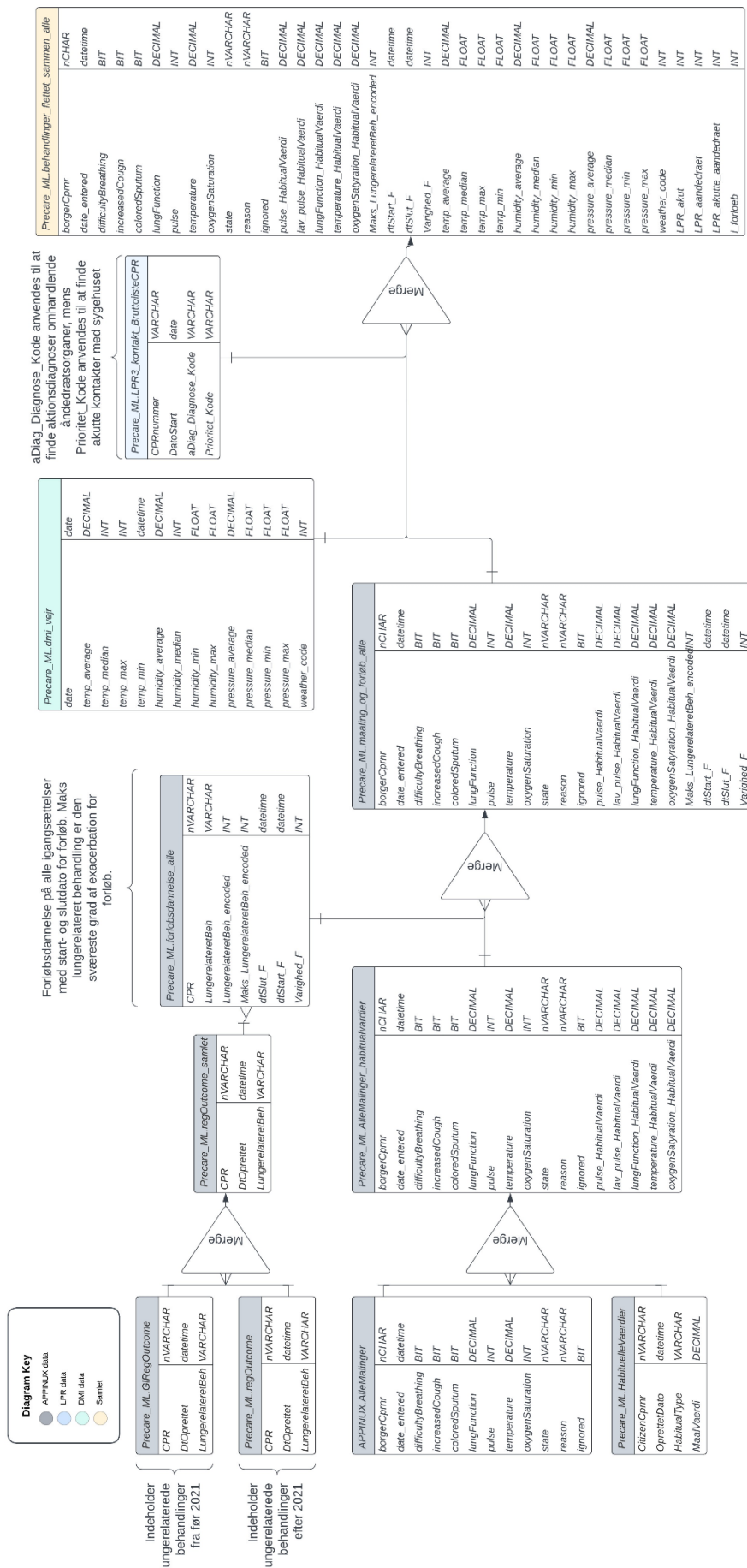
7. Konkluderende kommentarer

På baggrund af det nuværende datagrundlag er det svært at sige noget om den optimale målefrekvens for borgerne, da datakvaliteten generelt ikke er tilfredsstillende. Hvis datakvaliteten generelt forbedres (særligt hvis borgerne bliver bedre til at udføre valide lungefunktionsmålinger), vurderes det, at der er et stort potentiale for at udvikle en maskinlæringsmodel, der er i stand til at prædiktere forværringer hos borgere flere dage i forvejen. Derved kan målefrekvensen potentielt også sættes ned, således at borgere ikke behøver at udføre egenmålinger dagligt, men at borgeren derimod promptes for flere målinger, såfremt maskinlæringsmodellen returnerer en forhøjet sandsynlighed for en forværring inden for de nærmeste dage.

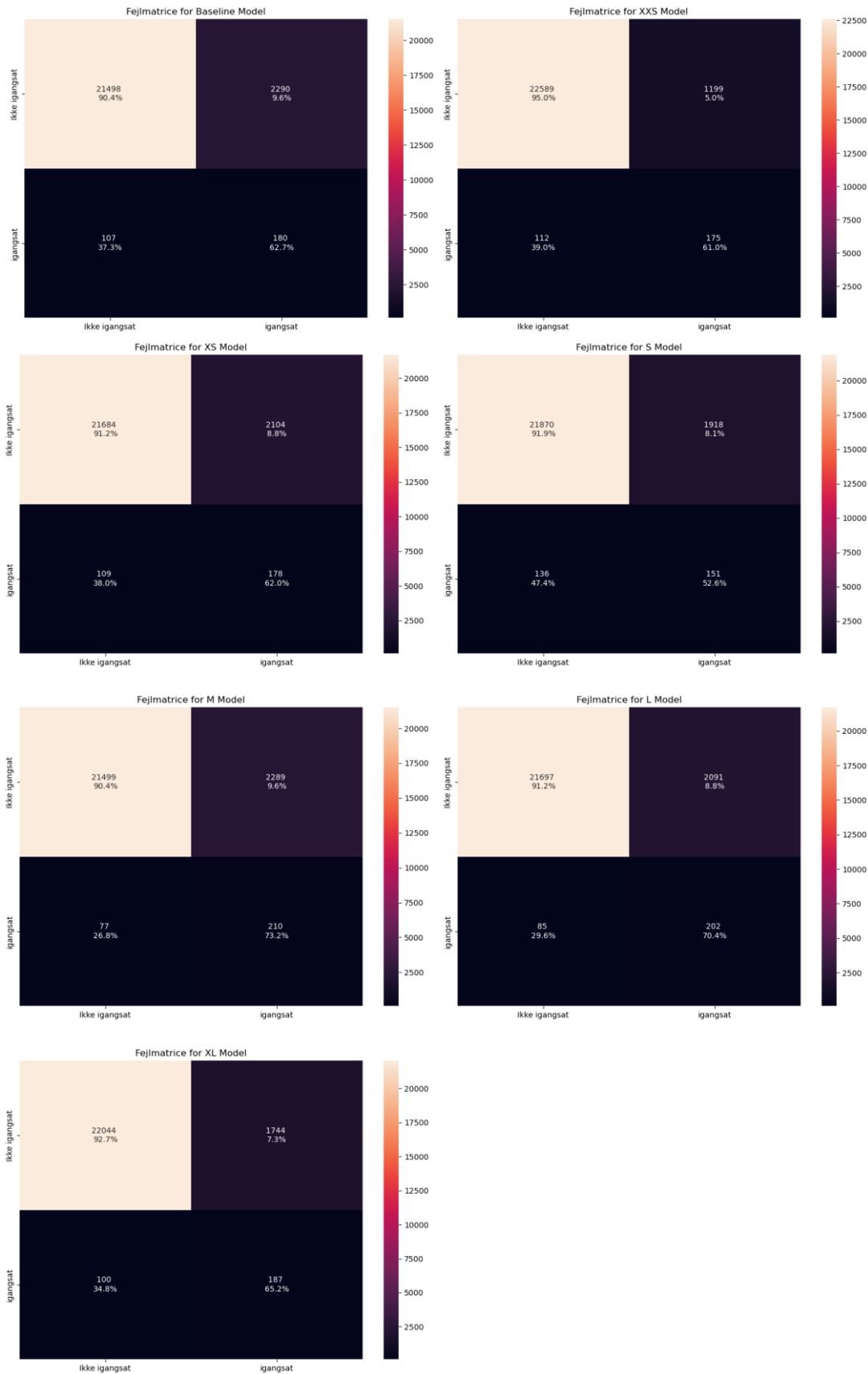
Analysen af borgernes egenmålinger har belyst et behov for straks-validering af indtastede og målte værdier, således at egenmålingerne bliver mere valide og antallet af forkerte triageringer mindskes. Derudover har analysen vist, at lungefunktionsmålingerne har en stor prædiktiv kraft, men samtidig er årsag til mange falske positive, da der er stor varians i borgernes egenmålte lungefunktion, hvilket belyser at der er et behov for øget fokus på at lære borgerne at udføre lungefunktionsmålingen korrekt eller eventuelt at gentage en måling, såfremt den afviger meget fra et normalområde eller den sidst udførte måling.

Analysen har ligeledes understreget den prædiktive kraft i de tre tilstandsspørgsmål, hvilket peger på et potentiale for at anvende disse til den daglige tilstandsevaluering, som kan danne grundlag for, at borgeren promptes for mere omfattende egenmålinger. Tilstandsspørgsmålene er mindre intervenerende end de nuværende egenmålinger, og den prædiktive kraft kan muligvis tilmed forbedres, hvis der inkluderes flere spørgsmål i den daglige tilstandsevaluering.

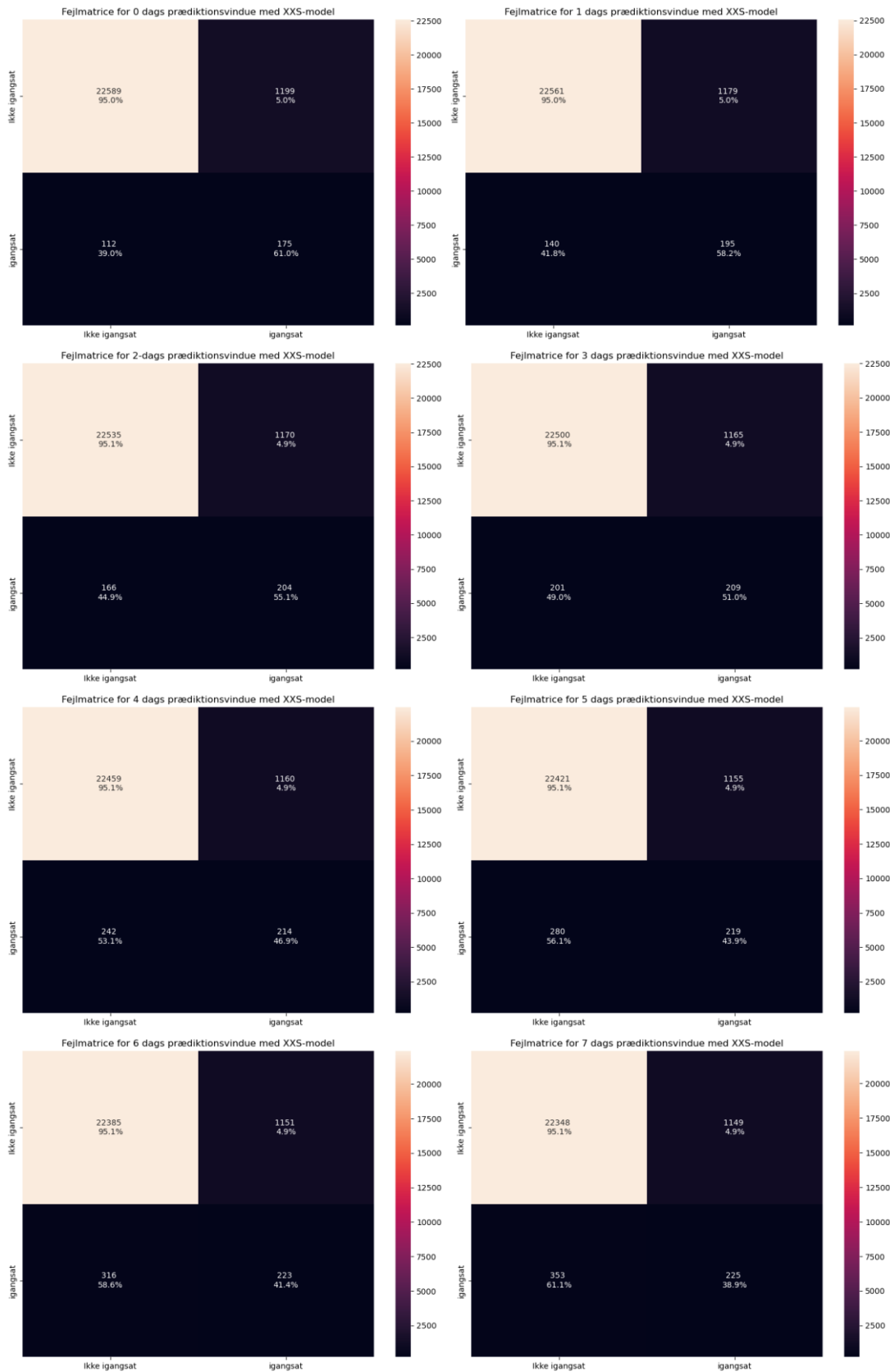
Bilag A: Overblik over datamodel



Bilag B: Fejlmatrixer for de afprøvede ML modeller med varierende input



Bilag C: Fejlmatrixer for de afprøvede prædiktionsvinduer XXS-Model



Bilag D: Fejlmatrixer for de afprøvede prædiktionsvinduer M-Model

