NORGES HANDELSHØYSKOLE
Bergen, Våren 2021

Fastlege vs Dr. Robot

En eksperimentell studie om bruk av robot til medisinsk konsultasjon

Kaja Witnes Schultz og Vilde Myrvold Thorud

Veileder: Helge Thorbjørnsen

Selvstendig masterutredning i

Økonomisk Styring og Strategi og Ledelse

NORGES HANDELSHØYSKOLE

Forord

Dette er en oppgave som er skrevet i forbindelse med masterstudiet ved NHH - Norges Handelshøyskole. Oppgaven utgjør 30 studiepoeng, og inngår som en del av vår mastergrad i Økonomi og Administrasjon. Utredningen er innen hovedprofilene Strategi og ledelse og Økonomisk Styring.

Vi var tidlig enige om at vi ønsket å skrive en oppgave som er rettet mot teknologisk utvikling i helsesektoren, etter å ha fått en økt interesse for dette i løpet av studietiden. Gjennom samtaler med vår veileder, Helge Thorbjørnsen, kom vi frem til at vi ønsket å gjennomføre et eksperiment innen et fagfelt som kan være relevant for flere bransjer - algoritmeaversjon.

Oppgavens tittel "Fastlege vs Dr. Robot" illustrerer at den teknologiske utviklingen har resultert i flere tilfeller der teknologi og menneske blir satt opp mot hverandre, som eksempelvis i helsesektoren. Det eksisterer en stor skepsis blant pasienter mot å ta i bruk teknologi – til tross for at teknologi ofte gjør det bedre enn mennesker. Dette synes vi er særlig interessant, og arbeidet med oppgaven har derfor vært utrolig lærerik. Det har spesielt vært spennende å opparbeide seg økt kunnskap om menneskets interaksjon med teknologi, i form av hvordan vi foretar beslutninger og evaluerer ulike alternativer opp mot hverandre. Vi håper funnene vil være relevant for akademia og næringslivet.

Vi ønsker å rette en stor takk til vår dyktige veileder, Helge Thorbjørnsen, som har vist stort engasjement og interesse for arbeidet vårt. Han har gitt oss gode råd og tilbakemeldinger på veien som vi har satt stor pris på.

Oslo, Juni 2021

Kaja Witnes Schultz og Vilde Myrvold Thorud

Sammendrag

Å benytte kunstig intelligens ("Artificial Intelligence", AI) i helsevesenet kan bidra til bedre tjenester og økt effektivisering. Det anses som nødvendig for å sikre en bærekraftig helsesektor i fremtiden. Imidlertid eksisterer det en irrasjonell motvilje tilknyttet å benytte AI-teknologi, også kalt algoritmeaversjon. Dette synes å gjelde til tross for at forskning viser at AI-teknologi presterer bedre enn mennesker. I den sammenheng har vi sett på om det foreligger algoritmeaversjon blant pasienter ved en medisinsk konsultasjon, og om graden av aversjon påvirkes når de opplyste symptomene varierer mellom lav og høy alvorlighetsgrad.

I oppgaven har vi benyttet kvantitativ metode, og på bakgrunn av eksisterende teori har vi utformet en modell, med tilhørende hypoteser. For å teste hypotesene har vi innhentet primærdata ved et eksperiment utformet som en spørreundersøkelse. Ved å benytte varians-, kjikvadrat- og medieringsanalyse har vi avdekket om det foreligger indikasjon på algoritmeaversjon ved i) valg av behandlingsmetode, ii) grad av tillit og oppfattet risiko tilknyttet diagnostisering og behandling, samt iii) evaluering av konsultasjonen, målt ved dimensjonene tilfredshet og forventet anger.

Hovedfunnene i oppgaven peker mot at det foreligger indikasjoner på algoritmeaversjon. Imidlertid kommer funnene kun til syne ved i) valg av behandlingsform og ii) grad av tillit til behandlingsformen. Graden av tillit deltakerne utviser til valgt behandlingsform påvirker også hvordan den evalueres. Det innebærer at dersom tilliten til behandlingsformen øker, vil det påvirke hvor god evalueringen av behandlingsformen er.

Videre kan vi ikke fastslå om det å variere alvorlighetsgrad (lav/høy) generelt påvirker graden av algoritmeaversjon. Imidlertid synes å være en sammenheng, der respondentene utviser lavere aversjon mot roboten ved symptomer av lav alvorlighetsgrad. Vi fant også at 80.8% av deltakerne ønsket å benytte fastlege med støtte fra en robot, da dette ble introdusert som et alternativ. Resultatet indikerer dermed at det kan være hensiktsmessig å tilby AI-teknologi som en behandlingsform dersom fastlegen fortsatt er endelig beslutningstaker.

Innholdsfortegnelse

Oversikt over tabeller

Oversikt over figurer

1 Innledning

Økt digitalisering gir omfattende endringer innen privat næringsvirksomhet, og dermed større forventning til at offentlig sektor også leverer kostnadseffektive tjenester som er av høy kvalitet. Kostnadseffektivitet er særlig gjeldende som følge av den demografiske utviklingen med en økt andel eldre, som utfordrer velferdsstatens bærekraft (Riekeles, 2018). Til tross for at Norge er verdensledende innen teknologimodenhet, har vi ikke lykkes med å overføre dette til offentlig sektor (Abelia, 2017). Dette støttes av Difis statusvurdering av digitalisering, der det hevdes at offentlig sektor ikke klarer å utnytte mulighetene som har oppstått fra digitaliseringen i samfunnet på en tilstrekkelig måte (Difi, 2016).

I rapporten *Vårt nye digitale Norge* har Accenture (2016), i samarbeid med World Economic Forum, kartlagt effekten av å digitalisere ulike sektorer, samt hvilken verdi dette vil ha for samfunnet. I rapporten fremkommer det blant annet at en økning i digitaliseringstakten i helsesektoren har et verdipotensiale på 25 milliarder kroner. Blant teknologiene som trekkes frem er økt bruk av kunstig intelligens ("Artificial Intelligence", AI), som muliggjør nye tjenester og økt innsikt gjennom selvlærende algoritmer.

Vi benytter EU sin definisjon av kunstig intelligens: "Kunstig intelligente systemer utfører handlinger, fysisk eller digitalt, basert på tolkning og behandling av strukturerte eller ustrukturerte data, i den hensikt å oppnå et gitt mål" (Regjeringen, 2020). Nasjonal helse- og sykehusplan (NHSP) 2020-2023 viser videre til at kunstig intelligens vil gjøre det "mulig å utnytte våre felles helsedata for å tilby en raskere og mer presis diagnostisering, bedre behandling og mer effektiv ressursbruk" (Regjeringen, 2019). Dette kan også resultere i at en kan avdekke flere "tidlig fase"-symptomer i forbindelse med alvorlige sykdommer (Anderssen, 2019).

Bedre utnyttelse av AI-teknologi kan derfor resultere i at pasientene får en mer nøyaktig og effektiv diagnostisering og behandling - uavhengig av hvilken lege en har og geografisk bosted (Anderssen, 2019). Funnene bekreftes av andre studier som finner at medisinsk kunstig intelligens blant annet kan gi bedre diagnostisering og behandling enn leger (Longoni et al., 2019). Dette gjelder blant annet identifisering av hjertesykdom (Hutson, 2017), kreftdiagnoser (Lohr, 2016; Haenssle et al, 2018) og øyesykdommer (Gulshan et al., 2016). AI- teknologi er

også bedre til prioritere korrekt akutt behandling, der robotens treffsikkerhet er på 90.2 %, mot legenes 77.5 % (Donnelly, 2017).

Til tross for større behov for kostnadseffektiv drift i helsevesenet, der økt benyttelse av kunstig intelligens kan bøte på utfordringene, foreligger det en motstand mot bruk av dette. Mer presist er det en systematisk psykologisk motstand mot bruk av statistiske algoritmer og automatiserte beslutningssystemer - også kalt algoritmeaversjon (Dietvorst et al., 2015). Det er flere studier som undersøker tilstedeværelsen av algoritmeaversjon, og forklarer hvorfor en slik aversjon finner sted. Blant forklaringene er overdreven tillit til menneskelige eksperter (Khaneman, 2011), ulik vekting av maskinelle feil og menneskelig feil (Dietvorst et al., 2015), sosiale behov (Baumeister & Leary, 1995; Deci & Ryan, 2008) og frykten for tapt individualitet (Longoni et al, 2019).

"Fastlegekrisen" er et begrep som jevnlig blir satt på dagsorden, og omhandler den økende mangelen på fastleger. En konsekvens av dette er at fastlegene må påta seg mer arbeid. Dette kommer til syne ved at fastlegene har en median på 52.5 arbeidstimer i uken, samt at 10% av fastlegene oppgir arbeidsuker på 75 timer (Dommerud, 2020; Storvik, 2018). Det eksisterer derfor en flaskehals i fastlegeordningen, og tiltak er nødvendig for å sikre en bærekraftig drift. AI-teknologi er i den sammenheng en løsning som kan bidra til å lempe på fastlegenes arbeidsbelastning, og legene selv har gradvis vist seg mer positive til å ta i bruk teknologien. Dette har særlig gjort seg gjeldende etter coronapandemien: I en spørreundersøkelse gjennomført av Intel (2020) oppga 84% av de forespurte aktørene i helsesektoren at de benytter, eller planlegger å benytte, kunstig intelligens relatert til helsediagnostikk, sammenlignet med 45% før pandemien i 2020.

Selv om helsepersonell synes å bli mer positive til å benytte kunstig intelligens, er en også avhengig av at pasientene selv stoler på teknologien, og ønsker å ta den i bruk. I henhold til litteraturen foreligger det en aversjon mot å få behandling av en robot. Dette er særlig relatert til helse, siden det er et svært personlig område, og kan innebære stor risiko om feil skulle forekomme (Asan et al., 2020). Det er derfor interessant at pasienter heller vil la seg behandle av en lege når forskningen er klar på at AI-teknologien presterer bedre. Det er derfor nødvendig med ytterligere kunnskap om algoritmeaversjon fra et pasientperspektiv, for å kunne dra bedre nytte av teknologiens mulighetsområder og verdiskapingspotensiale - både i fastlegeordningen og helsesektoren som sådan.

Fra litteraturen er det - så vidt vi kjenner til - ingen som har sett på algoritmeaversjon hos pasienter tilknyttet fastlegekonsultasjoner. Det er heller ikke forsket på om graden av algoritmeaversjon endres når alvorlighetsgraden på symptomer varierer. Funnene kan være relevant for implementering av AI-teknologi i sammenhenger der pasienter er direkte involvert, som eksempelvis i fastlegeordningen. Et eksempel her er at en kan tilby roboter som behandlingsmetode ved mindre alvorlige symptomer, dersom det er signifikante funn på at algoritmeaversjonen da er lavere. Når pasienter gradvis blir trygg på disse tjenestene, kan en gradvis utvide tjenestetilbudet. Dette leder frem til vårt forskningsspørsmål:

Er det algoritmeaversjon til stede hos pasienter i forbindelse med utførelse av en medisinsk konsultasjon, og er aversjonen svakere når symptomene er av lav alvorlighetsgrad?

2 Teoretisk bakteppe

I denne oppgaven anser vi det som hensiktsmessig å få en dypere forståelse av faktorene som påvirker menneskers forhold til AI-teknologi i helsesektoren. Først vil vi få en rikere innsikt i de psykologiske mekanismene som kan påvirke ens preferanser i medisinsk behandlingsmetode. I den sammenheng er tillit og risiko to sentrale faktorer som påvirker mennesker ved beslutningstaking. Vi vil så se på hvordan disse aspektene spiller inn i helsesektoren, og videre i møte med den stadig økende teknologiske utviklingen. Deretter tar vi for oss litteratur innen Human-Computer Interaction (HCI) og kunstig intelligens for å gi en forståelse av teknologiens utvikling, funksjoner og muligheter, samt menneskers interaksjon med denne. Til slutt vil vi ta for oss algoritmeaversjon. Dette danner grunnlaget for det vi ønsker å studere.

2.1.0 Tillit

En rekke definisjoner på tillit har blitt utarbeidet i litteraturen, der det synes å være en uenighet om omfanget av begrepet (Sekhon et al., 2014; Mayer et al., 1995; Korczynski, 2000; Tullberg, 2007). Kramer et al. (1996) hevder at mellommenneskelig tillit er basert på tidligere interaksjon med mennesker. Sekhon et al. (2014) mener imidlertid at tillit defineres som “hva en part tror om en annen part sine holdninger og oppførsel”. Mayer et al. (1995) har en lignende definisjon, der det hevdes at tilliten en part viser er basert på en overordnet vurdering av den andre partens kompetanse, velvilje og integritet. I forlengelse av denne

definisjonen vektlegger Korczynski (2000) at tillitsgiver bør foreta en vurdering av hvor sikker en er på at tillitshaver vil utføre tiltenkte handlinger. McAllister (1995) hevder videre at det eksisterer både kognitiv og affektiv tillit. Kognitiv tillit omfatter graden av kunnskap en har om den andre parts kompetanse og pålitelighet, som gjør at vi har "gode nok" grunner til å stole på vedkommende. Affektiv tillit handler om de emosjonelle og mellommenneskelige aspektene som eksisterer mellom partene.

På tross av uoverensstemmelse i litteraturen vedrørende definisjonens omfang, er villigheten til å være sårbar, og sikre forventningene en har til tillitshavers handlinger, sentralt for de fleste definisjonene. Vi velger derfor å benytte følgende definisjon for tillit: «En parts villighet til å være sårbar overfor en annen parts handlinger, basert på forventninger om at den andre vil utføre en bestemt handling som er viktig for tillitsgiver, uavhengig av muligheten til å observere eller kontrollere tillitshaver» (Mayer et al, 1995). Med dette følger det også en usikkerhet, ved at det er en risiko for at negative utfall kan forekomme (Jøsang & Presti, 2004; Rousseau et al. sitert i Sekhon et al., 2014; Alaszewiski, 2003; Giddens, 1991).

2.1.1 Tillit og risiko

Boon & Holmes (1991) viser til at risiko og tillit kan interagere med hverandre, siden tillit handler om "å få innsikt i andres motiver som påvirker en selv, i situasjoner som medfører risiko". Dette støttes av Giddens (1991) som i sin artikkel viser til at tillit er en form for å håndtere risiko som følge av uvitenhet og usikkerhet. Risikoaspektet øker spesielt når en transaksjon er involvert i tillitsvurderingen, for eksempel i form av en økonomisk gjenstand. På denne måten kan risikoaspektet betegnes som forventede konsekvenser dersom en feil skulle forekomme (Jøsang & Presti, 2004).

I litteraturen synes det å være få systemer og modeller som beskriver interaksjonen mellom tillit og risiko, som eksplisitt tar risikoaspektet i betraktning (Jøsang & Presti, 2004). I de fleste tilfeller må respondenten selv vurdere forholdet mellom risiko og tillit ved å kombinere de forskjellige faktorene i modellene (Grandson & Sloman, 2000; Falcone & Castelfranchi, 2001). Imidlertid kan det hevdes at at det eksisterer et inverst forhold mellom tillit og risiko. Med dette menes at transaksjoner som er av stor betydning, også krever høy grad av tillit og lav risiko. Transaksjoner som vurderes å være av liten betydning krever imidlertid lavere tillitsnivå, og kan innebære større risiko (Partick, 2002). Videre er det flere aspekter enn å

stole på tillitshaver som inngår når en skal utvise tillit. Det argumenteres blant annet for at viktigheten og betydningen av handlingen, kan være for stor til at tillitsgiver er villig til å vise tillit. Det kan dermed være en risiko som er for stor til at den kan tolereres (Falcone & Castelfranchi, 2001).

Andre forfattere i litteraturen har beskrevet forholdet mellom tillit og risiko i en matrise. Dimitrakos (2002) viser til at forholdet mellom tillit og risiko eksplisitt kan illustreres i en tillitsmatrise, der graden av tillit en viser i en gitt situasjon stammer fra en interaksjon mellom disse faktorene. Tilsvarende matrisemodell fremlegger Manchala (1998). Derimot, i stedet for å måle tilliten direkte, benyttes det her andre variabler som måler konseptet tillit - mer presist risikokostnaden og tidligere historie en har med tillitshaver. Dette resulterer i en risiko-tillits matrise som beslutter om en skal inngå en avtale med den andre parten, eller ikke.

Fra litteraturen synes det derfor å eksistere et forhold mellom tillit og risiko, og at dette forholdet kan være inverst. Tillit- og risikovurderinger er imidlertid særdeles situasjonsbetinget. Mennesket står overfor slike beslutningers hver dag, som både kan være bevisste og ubevisste, samt i forskjellige kontekster (Kahneman, 2011; Jøsang & Presti, 2004). Dette gjør at det er utfordrende å vise til forholdet mellom konseptene, og illustrere dette i konkrete modeller, eller systemer.

2.1.2 Oppfattet risiko

Grima et al. (2019) definerer oppfattet risiko som "en situasjon der en beslutningstaker har den nødvendige kunnskapen om konsekvensene av forskjellige alternativer, der sannsynlighetene for disse utfallene vurderes". På den måten kan oppfattet risiko også ses på som forventet tap. Av dette er det naturlig at en, ved en beslutning, vil forsøke å minimere, eller om mulig unngå tapet. Mitchell (1999) hevder at det særlig er viktig å legge til grunn at oppfattet risiko er en subjektiv vurdering - både av størrelsen på det potensielle tapet og sannsynligheten for den.

Den oppfattede risikoen kan være spesielt viktig i forbindelse med beslutningstaking, som eksempelvis relatert til valg mellom ulike tjenester og produkter. Tilknyttet dette består risikovurderingen av både ytre og indre faktorer. Ytre faktorer omfatter karakteristika ved produktet eller tjenesten, mens indre faktorer gjelder tidligere erfaringer forbrukeren har med

tjenesten eller produktet. Dersom tjenesten, eller produktet, vekker negative følelser hos konsumenten, vil dette øke den oppfattede risikoen. Individuelle personlighetstrekk, som eksempelvis risikovillighet, kan også påvirke grad av oppfattet risiko (Dowling sitert i Lacey et al., 2009).

Roselius (1971) argumenterer for at følgende fire kategorier påvirker den oppfattede risikoen: Tid, fare, ego og penger. Mohtar & Abbas (2015) har senere utvidet til følgende fem kategorier: Finansiell risiko, ytelsesrisiko, fysisk risiko, psykologisk risiko, sosial risiko og bekvemmelighetsrisiko. Finansiell risiko defineres som det finansielle tapet til beslutningtakeren. Ytelsesrisiko omfatter risikoen for at gjenstanden ikke fungerer som den skal, eller som forventet. Videre beskrives fysisk risiko som risikoen for at gjenstanden påfører fysisk skade på brukeren, mens psykologisk risiko omfatter de psykisk belastende følelsene risikoen kan vekke, som angst eller uro. Sosial risiko omhandler tap av sosial status, respekt, vennskap eller anerkjennelse. Til slutt defineres bekvemmelighetsrisiko som tap av innsats og tid som er lagt ned i forbindelse med å bli vant til produktet eller tjenesten.

Featherman & Pavlou (2003) kommer til lignende konklusjoner i sitt studie, som omhandler oppfattet risiko i forbindelse med å ta i bruk digitale tjenester. Imidlertid deler forfatterne inn oppfattet risiko i følgende hovedkategorier: i) prestasjonsrisiko og ii) psykologisk risiko. Førstnevnte kategori innehar følgende tre underkategorier: økonomi, mulighet/tid og innsats, mens psykologisk risiko har underkategoriene psykologi faktorer og sosiale forhold. Studiens resultater viser at viljen til å ta i bruk digitale tjenester reduseres som følge av prestasjonsrelaterte risikooppfatninger hos konsumentene. Opplevd brukervennlighet av de digitale tjenestene reduserte dermed konsumentenes oppfattede risiko (Featherman & Pavlou, 2003). Forfatterne viser at det er hensiktsmessig å analysere faktorene som påvirker konsumentenes oppfattede risiko tilknyttet bruk av en tjeneste. Resultatene kan bidra til å iverksette tiltak som kan redusere konsumentenes bekymring tilknyttet anvendelse av teknologien. I den sammenheng viser Lacey et al. (2009) til at oppfattet risiko hos konsumenter avtar ved bruk av produktet eller tjenesten, men at det også avhenger av den enkeltes personlighetstrekk.

2.1.3 Tillit og risiko i helsevesenet

Tillit og risiko spiller en sentral rolle i markeder der det er stor grad av asymmetrisk informasjon mellom tillitsgiver og tillitshaver. Helsesektoren, og forholdet mellom pasient og lege, er i den sammenheng et eksempel på dette (Hardin sitert i Dyer, 2016). Legen har incentiver til å ivareta et godt forhold til pasienten på bakgrunn av å være bundet av moralske og psykologiske forpliktelser. Av den grunn er det viktig å fremstå troverdig overfor tillitsgiver. Derfor hevdes det at pasienten, i en slik situasjon, antar at legen tar utgangspunkt i pasientens beste interesse, og av den grunn utviser tillit. Imidlertid avhenger tilliten pasienten viser av at en er mentalt tilgjengelig, som blant annet baseres på tidligere erfaringer (Dyer, 2016).

Alaszewiski (2003) argumenterer for at tilliten pasientenen har til profesjonelle aktører i helsevesenet er relativt situasjonsbetinget. Han fremmer at møtet med aktører i helsevesenet kan være preget av hurtige risiko- og tillitsvurderinger, dersom situasjonen er akutt. Derfor er tillitsgiver avhengig av å stole på tillitshaver som tar beslutninger på vegne av seg selv, til tross for at det innebærer en potensiell risiko om vedkommende tar feil avgjørelse. Tillitsgiver må dermed kunne stole på at tillitshaver har tilstrekkelig kunnskap og vilje til å benytte dette i pasientens beste interesse.

Graden av tillit påvirker også sannsynligheten for at pasienten vil følge legens anbefalinger. Zheng et al. (2017) finner i den sammenheng at når tilliten tillitsgiver har overfor legen er høy, så vil pasienten heller la tvilen komme legen til gode, fremfor å stille spørsmål ved legens intensjoner eller kompetanse. Dette støttes av Newcomer (1997) som hevder at tillit i medisinsk behandling hovedsakelig handler om i) legen er kompetent nok til å gi riktig diagnose og behandling, og om ii) vedkommende har den beste interessen for pasienten. Tilsvarende konklusjon, bare utledet med andre begreper, kommer Leisen & Hyman (2001) til. Forfatterne fremmer at tillit beror på i) velvilje og ii) en teknisk komponent. Velvilje omfatter blant annet å forstå pasientens individuelle behov, vise omsorg, ærlighet og respekt, mens den tekniske komponenten består av at legen klarer å foreta grundige evalueringer, samt å gi effektiv og riktig behandling.

2.1.4 Tillit i det digitale grensesnitt

Den økende digitale transformasjonen samfunnet gjennomgår, medfører at skillelinjene mellom det vi definerer som den digitale og analoge verden reduseres. Det digitale samfunnet synes å være både volatilt, usikkert, komplekst og ambisiøst (Deloitte, 2021). Tillit til nye teknologiske løsninger står derfor sentralt i kommersialiseringen av teknologien, slik at en kan dra nytte av den teknologiske verdiskapingen. I følge PWC (2017) er en avhengig av å bygge tillit til teknologiske løsninger for å kunne maksimere potensialet, samt håndtere medfølgende risiko. I tråd med dette står også konsumentene stadig overfor ulike tillitsvurderinger i møte med nye, ukjente teknologier og systemer. I den sammenheng er en avhengig av at konsumentene har tillit til at virksomheter, både offentlig og privat, leverer trygge tjenester, og behandler persondataene i tråd med lovreguleringer. Dette bekreftes av studier som viser at det er et positivt forhold mellom tillit og kundetilfredshet (Deepak et al., 2002; Anderson and Narus, 1991; Gummerus et al., 2004; Taylor and Hunter 2003). Derfor hevder PWC (2017) at organisasjoner og myndigheter må fokusere på implementering av tillit som en del av kjernevirksomheten.

2.2.0 HCI

Vi vil i det følgende gi en innføring i det tverrfaglige området Human-Computer Interaction (HCI). Dette for å gi en grunnleggende forståelse av interaksjonen mellom mennesker og maskiner. Mer presist gjelder dette den teknologiske utviklingen og viktigheten av å fokusere på brukerne. HCI defineres som vitenskapen som kombinerer informatikk, design, atferdsvitenskap og kunstig intelligens, og omhandler derfor interaksjonen som foregår mellom menneske og datamaskin (Ren & Bao, 2020). Noe av det viktigste i forbindelse med HCI er funksjonalitet og brukervennlighet (Karray et al., 2008; Fischer, 2001), samt å tilpasse løsningene etter brukernes bakgrunnskunnskaper og mål (Fischer, 2001).

2.2.1 Historisk utvikling

På 1970-tallet var fokuset innenfor HCI i stor grad på grafisk brukergrensesnitt (GUI) som bruk av ikoner, menyer, tastatur, datamus og datapenn. Formålet var å skape mer brukervennlige systemer (Fischer, 2001). I den videre utviklingen av HCI skjedde det et skifte: Fokuset gikk bort fra den fysiske interaksjonen, for eksempel å peke og klikke med en datamus, og mer mot hvordan brukeren forstår og interagerer med systemet (Harper et al.,

2008; Karray et al., 2003). Det var derfor en "kognitiv revolusjon" innenfor HCI på 1980- og 1990-tallet som reflekterte skiftet.

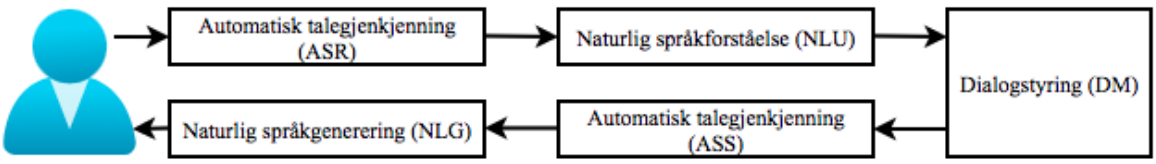
På 1990-tallet skiftet fokuset til kommunikasjonen mellom mennesker som var muliggjort av datamaskiner. Dette gjenspeiles i veksten av kommunikasjonsnettverk som koblet maskiner sammen. Utviklingen gikk derfor fra hvordan HCI kunne muliggjøre effektiv interaksjon med en maskin, til å fokusere på hvordan brukere kunne interagere med hverandre via en maskin. I senere tid har HCI blitt et tverrfaglig område som har gjennomgått enorme endringer (Harper et al., 2008). Økt teknologisk utvikling, og derav også maskinenes funksjonsområder, har resultert i et stadig større fokus på brukerne. Dette kommer til syne ved at brukerbehov er den sentrale driveren i utviklingen av maskiner og systemer (Hudlicka, 2003).

2.2.2 Menneskelige evner hos intelligente maskiner

I det følgende presenteres en kort gjennomgang av ulike evner som mennesker benytter ved kommunikasjon og interaksjon, henholdsvis lytte, tale, lese, skrive og syn. Dette er derfor også evner som intelligente maskiner må inneha (Ren & Bao, 2020).

Lytte og tale

Sansene som omhandler lyd blir brukt til gjensidig kommunikasjon mellom mennesker, og består av å lytte og prate. Maskiner må derfor ha slike evner for å kunne imitere et menneske i en interaksjonsprosess. Dette kan gjøres via et talesystem. Det består av automatisk talegjenkjenning ("Automatic Speech Recognition", ASR), naturlig språkforståelse ("Natural Language Understanding", NLU), dialogstyring ("Dialogue Management", DM), naturlig språkgenerering ("Natural Language Generation", NLG), og automatisk talegenerering (Automatic Speech Synthesis", ASS) (Ren & Bao, 2020). Se figur 1 nedenfor som illustrerer prosessen.



Figur 1: Prosess for talesystem, oversatt og hentet fra Ren & Bao (2020)

ASR skal transformere brukerens tale til stavelser og ord. Videre skal NLU analysere resultatet av talegjenkjenningen, og deretter benyttes syntaksanalyser og semantiske analyser til å transformere informasjonen. På denne måten blir informasjonen på en form som kan benyttes av DM. DM brukes så for å gjøre en omfattende analyse, blant annet basert på resultatet fra NLU, konteksten til dialogen, historisk informasjon om dialogen etc.. Dette benyttes til å vurdere intensjonen til brukeren. Deretter blir responsstrategien utviklet og gjennomført av systemet. NLG vil så organisere den passende responsen, og konvertere systemets respons til det naturlige språket brukeren kan forstå. Til slutt skal ASS generere teksten fra NLG til den endelige talen, som videre sendes tilbake til brukeren (Ren & Bao, 2020).

ASR'er utvikles både innenfor akademia og i industrien. Eksempler på sistnevnte er bedrifter som Google, Microsoft, IBM og Amazon, som har alle utviklet egne ASR'er. I tillegg til selve innholdet i talen, inneholder stemmen også følelser. Dermed er det ikke kun innholdet i talen som er viktig, men også hvilke følelser som er tilstedeværende. Derfor er det et stort fokus på følelsesmessig gjenkjenning, slik at denne informasjonen kan utnyttes videre i interaksjonen (Ren & Bao, 2020).

Lese og skrive

Å utstyre maskiner med evnene til å kunne lese og skrive er innenfor kategorien NLP. Her er målet å muliggjøre at maskinene skal kunne lese tekst generert av mennesker, forstå tankene og ideene det inneholder, samt å uttrykke tanker og ideer ved å generere spesifikke tegn og tekst (Ren & Bao, 2020).

Synet

Synet er den viktigste sansen, siden mer enn 80% av informasjonen som mottas fra den utvendige verden er via synet. I kommunikasjon med andre benytter mennesker synet til å gjenkjenne og vurdere ansiktsuttrykk, gester, fysisk atferd, følelser etc. Maskinsyn er vitenskapen om hvordan maskiner kan "se" som mennesker. Det vil si at en benytter et kamera til å erstatte menneskelige øyne for å produsere bilder. Tilsvarende benyttes en maskin til å erstatte den menneskelige hjernen for å prosessere bildene. En teknikk som benyttes er blant annet gjenkjenning av ansiktsuttrykk, som vil si å gjenkjenne tilstander for å kunne identifisere følelser hos objektet. Det finnes også teknikker for å oppdage og

gjenkjenne positurer, gester og øyebevegelser. Sistnevnte kan for eksempel benyttes til å oppdage om vedkommende er uoppmerksom (Ren & Bao, 2020).

2.2.3 Tilpasning av datasystemer

Tidligere har et fokusområde vært å ha tilstrekkelig informasjon tilgjengelig for brukerne. Den senere utviklingen med "big data" har resultert i at enorme mengder data genereres. Dette gjør at fokuset har gått fra å ha informasjon tilgjengelig, til å selektere ut hvilken informasjon som er relevant for den aktuelle oppgaven, gitt den antatte bakgrunnskunnskapen brukeren har (Fischer, 2001). En utfordring for designere av HCI-systemer er derfor å lage software til millioner av brukere, som samtidig er tilpasset hver enkelt bruker. For å møte denne utfordringen fremmer Fischer (2001) at en må utarbeide systemene slik at de klarer i) å si det rette, ii) å si ting til rett tid, og iii) å si ting på rett måte.

2.2.4 Identifisere affekt

Et viktig aspekt ved HCI er å identifisere affekt, som innebærer at intelligente maskiner har evnen til å forstå og uttrykke følelser. Dette ble først lansert i 1997 av Picard fra MIT. Det er godt kjent at følelser er en viktig faktor i kommunikasjon og interaksjon mellom mennesker. Av den grunn forventes det at intelligente maskiner også skal inneha evnen til å interagere på det emosjonelle planet (Ren & Bao, 2020). En av grunnleggerne av kunstig intelligens, omtalte i sin bok "The Society of mind", at spørsmålet ikke er om maskiner kan inneha følelser, men om maskiner kan være intelligente uten følelser (Minsky, 1988).

Å identifisere affekt kan gjøres ved å gjenkjenne brukerens affektive tilstand, for deretter å kunne tilpasse maskinens respons ved behov (Hudlicka, 2003). Dette kan gjøres på en rekke måter, som for eksempel ved måling av hjerterytme, ansiktsuttrykk, tonefall, kroppsbevegelser og gester, samt selvrapportering av for eksempel misnøye. Dette benyttes så for å analysere hva som kan være tilstedeværende affektive tilstander. Konteksten til oppgaven bør også inkluderes i analysen, eksempelvis oppgavens kompleksitet, lengde og tidspunkt på dagen den gjennomføres, samt den enkeltes historikk på for eksempel tidligere suksess og feiling. Følelsesmessig tilstand kan også identifiseres basert på personlighet, som eksempelvis grad av ekstroversjon og aggressivitet (Hudlicka, 2003). Flere metoder kombineres ofte for å identifisere affekt så presist som mulig.Når den affektive tilstanden er

identifisert avgjør systemet om det skal foretas tilpasninger ut mot brukeren, og eventuelt hvordan systemes skal tilpasses.

Det er en rekke situasjoner hvor menneskelige feil kan resultere i uhell og ulykker, og som kan reduseres ved at systemdesignet kan identifisere affekt. Dette gjelder særlig i situasjoner relatert til stress, usikkerhet, frustrasjon og kjedsomhet. Videre kan det også benyttes til å opprettholde positiv affekt i forbindelse med arbeidsoppgaver for at brukerne av systemet skal prestere optimalt.

En metode for å vurdere brukernes følelser er ved å analysere data som hentes inn fra HCI-enheter. Brukere interagerer med teknologi ved å benytte ulike enheter som datamus, touch-skjermer etc. Disse enhetene samler inn informasjon om brukeren ved at de fungerer som en sensor (Google, 2015). For eksempel kan en datamus gi presis informasjon om en persons motoriske bevegelser med millisekunders presisjon (Hehman et al., 2014). Denne informasjonen kan gi indikasjoner på brukerens emosjonelle og kognitive tilstand (Freeman et al., 2011; Kim & Choi, 2012).

Ved å analysere datamus-bevegelser kan en for eksempel oppdage hvilke deler av interaksjonen med systemet som fører til negative følelser. Denne informasjonen kan så benyttes til å få en bedre forståelse av hvor det kan gjennomføres systemforbedringer. Systemer kan derfor automatisk oppdage når det er sannsynlig at en bruker opplever negative følelsesmessige reaksjoner, for så å gripe inn. Dette kan eksempelvis være ved å gi brukerne mulighet til å uttrykke misnøye (Klein et al., 2002), komme med beklagende uttalelser (Tzeng, 2004), kompensasjon (Smith et al., 1999) eller forklaringer (Kuo et al., 2011). Resultatene fra en studie utført av Hibbeln et al. (2017) viser at negative følelser påvirker avstanden og farten ved bruk av datamus. Denne informasjonen kan dermed indikere om negative følelser er tilstedeværende, samt nivået på disse. Dette gir muligheter til å kunne designe systemer som kan identifisere, oppfatte og respondere på brukerens følelser.

2.2.5 HCI og tillit

Tillit kan i stor grad påvirke interaksjonen mellom menneske og maskin (Hancock et al., 2011; Lee & See, 2004; Parasuraman & Riley, 1997). Noen hevder til og med at dersom ikke tillit er tilstedeværende vil ikke maskinen brukes (Schaefer et al., 2016). Forskning viser også

at når brukeren har et valg mellom et menneske eller et automatisert system for å utføre en oppgave, vil vedkommende velge alternativet som han stoler mest på (Nickerson & Reilly, 2004).

Hancock et al. (2011) viser at tillit i forbindelse med HCI i stor grad er påvirket av prestasjoner. En studie viser at dersom roboten presterte bedre, resulterte det i høyere grad av tillit. Dette førte til at brukere valgte maskinen oftere, fremfor å gjøre oppgaven manuelt (Dzindolet et al., 2003). En annen studie viser videre at oppgavetype var viktigere enn prestasjonene til maskinen (Salem et al., 2015). Det er også viktig at brukeren har en riktig forståelse av maskinens faktiske egenskaper, slik at tilliten hverken blir for høy eller for lav, som henholdsvis kan føre til at maskinen brukes galt, eller ikke benyttes i det hele tatt (Parasuraman & Riley, 1997).

Sanders et al. (2019) viser at tillit også kan predikere brukervalg. I deres studie var det to ulike oppgaver som skulle gjennomføres, der deltakerne skulle velge mellom å benytte en robot eller et menneske. Analyser indikerte at brukernes valg signifikant ble predikert av deltakernes tillitsscore til roboten. Det var imidlertid kvalitative data som indikerte at det var andre faktorer som var viktigere i valget av robot enn tillit, siden deltakerne sjeldent nevnte tillit som en årsak til valget deres. Det bør også nevnes at tilliten til roboten var lavere enn til mennesket ved begge oppgaver. Dette til tross for at de presterte identisk. Dette indikerer dermed en høyere grad av tillit til mennesker, sammenlignet med roboter.

2.3.0 Kunstig intelligens

Kunstig intelligens (AI) kan forklares som menneskelig intelligens utøvet av maskiner. Menneskelig intelligens muliggjør at mennesker blant annet evner å lære, bruke logikk, resonnere, gjenkjenne mønstre, ta avgjørelser og løse problemer (Ren & Bao, 2020). Målet er derfor å bygge intelligente maskiner som kan tolke verden som mennesker, forstå språk og lære fra virkelige eksempler (Jones et al., 2018). Maskinene får stadig flere menneskelige egenskaper, som for eksempel evnen til å lytte, prate, lese, skrive, føle og inneha bevissthet (Ren & Bao, 2020). Videre er algoritmer viktig innenfor kunstig intelligens, og kan defineres som en rekke steg som en maskin følger for å utføre spesifikke oppgaver (Castelo et al., 2019).

2.3.1 Typer kunstig intelligens

Vi vil i det følgende gi en oversikt over ulike former for kunstig intelligens, der fokuset er innenfor medisinsk behandling og helse.

Maskinlæring

Maskinlæring (ML) analyserer strukturerte data. Denne type data kan for eksempel være fra røntgenbilder, genetiske data, resultater fra fysiske undersøkelser, elektroniske journaler, resultater fra laboratorietester og elektrofysiologiske data (Jiang et al., 2017). Når maskinlæring benyttes vil datasystemet først læres opp ved bruk av store datasett med relevante data, slik at det lærer å gjenkjenne mønstre, fremfor at det programmeres med spesifikke regler (Mesko et al., 2018; Jones et al., 2018). Den vanligste formen for bruk av ML i medisin er innenfor presisjonsmedisin, altså å predikere hvilken behandling som det er sannsynlig at vil fungere på en pasient (Davenport & Kalokata, 2019). I USA har for eksempel IBM Watson Health Cognitive computing system brukt ML-teknikker. Dette ved å benytte over en million forskningsartikler og historikken til et stort antall pasienter, for å skape et beslutningsstøttesystem for leger som behandler kreftpasienter. Her har målet vært å forbedre nøyaktigheten i diagnostisering av pasientene, samt å redusere kostnadene (Jones et al., 2018).

Nevrale nettverk og deep learning

Nevrale nettverk er en mer kompleks form for ML. Det er en samlebetegnelse for ulike datastrukturer med tilhørende algoritmer, som er inspirert av måten nervecellene i en hjerne er organisert på (Mesko et al., 2018; Dvergsdal, 2019).

Den mest komplekse formen for ML er deep learning, som er nevrale nettverk bestående av mange lag (Davenport & Kalokata, 2019; Jiang et al., 2017). Dermed kan mer komplekse ikke-lineære mønstre utforskes. Dette er nødvendig når dataene stadig øker i volum og kompleksitet. Bruken av deep learning i forskning ble nesten doblet i 2016. Majoriteten av deep learning blir benyttet i forbindelse med bildeanalyser, siden bilder av natur er komplekse og det eksisterer store volum (Mesko et al., 2018; Jiang et al., 2017). Deep learning-algoritmer gjør det mulig å bistå diagnostisering i forbindelse med kardiologi, dermatologi og onkologi (Mesko et al., 2018). Mest vanlig er det å benytte deep learning innenfor analyser av bilder i forbindelse med onkologi (Davenport & Kalokata, 2019).

Natural language processing

En form for deep learning er "natural language processing" (NLP). Det er er teknikker som henter ut informasjon fra ustrukturerte data, som for eksempel kliniske notater. Dette gjøres så om til strukturerte data som kan analyseres (Jiang et al., 2017). NLP består blant annet av stemmegjenkjenning, tekstanalyse og oversettelse, som er relatert til språk. Innen medisin brukes NLP i stor grad for å forstå og klassifisere klinisk dokumentasjon og publisert forskning (Davenport & Kalokata, 2019).

Typer av kunstig intelligens kombinert

Et AI-system må inneha den tradisjonelle ML-komponenten for å håndtere strukturerte data, samt en NLP-komponent for å håndtere ustrukturerte data. Deretter må systemet læres opp med helsedata før systemet kan assistere legen med eksempelvis diagnostisering og behandlingsalternativer. For eksempel har IBM Watson begge disse komponentene, som innenfor kreftforskning har gitt behandlingsanbefalinger som i 99% av tilfellene samsvarer med legens avgjørelser. I tillegg har IBM Watson for eksempel samarbeidet med Quest Diagnostics for å tilby genetiske analyser for diagnostisering, som blant annet i Japan har medført identifisering av en sjelden type leukemi (Jiang et al., 2017).

Maskiner gjør det bedre enn mennesker

Kunstig intelligens blir stadig mer utbredt innenfor ulike industrier, og i samfunnet ellers. Benyttelse av kunstig intelligens i helsesektoren omtales som revolusjonerende. Det er også ansett som den mest verdifulle teknologien, som har det største potensialet til å oppnå mange gjennombrudd (Ren & Bao, 2020). Maskiner gjør det i dag bedre enn mennesker på en rekke områder som kan bidra til gevinster som for eksempel bedre diagnostisering og økt effektivitet (Logg et al., 2019). De kan til og med utføre det som regnes for å være subjektive oppgaver, som å oppdage følelser i ansiktsuttrykk og tonefall (Castelo et al., 2019).

AI-teknologiens potensiale er stort på mange områder i forbindelse med helsetjenester (Davenport & Kalokata, 2019). Det er allerede en rekke bedrifter og organisasjoner som har vist at AI kan forbedre kvaliteten på helsetjenester og/eller redusere kostnadene. AI-baserte tjenester kan for eksempel gi mer nøyaktige diagnoser, samt benyttes i forbindelse med administrasjonsoppgaver og big data-analyser. AI-teknologi kan også assistere leger ved automatisk å innhente oppdatert medisinsk informasjon fra journaler, lærebøker og klinisk erfaring, for så å gi forslag til korrekt behandling. I tillegg kan AI-systemer hente ut nyttig

informasjon fra store databaser av pasienter, for å kunne gi prediksjoner av helseutfall. Det er generelt et stort fokus på å utvikle AI-teknologi som kan benyttes i forbindelse med onkologi, nevrologi, samt kardiologi. Dette er sykdommer som kan resultere i død, og dermed er tidlig diagnostisering avgjørende (Mesko et al., 2018; Jones et al., 2018; Davenport & Kalokata, 2019; Jiang et al., 2017; Jussupow et al., 2020).

Paul Mehl viste allerede i 1954 at statistiske prediksjoner gjør det bedre enn mennesker. Han gikk i gjennom resultater fra 20 studier på prognoser, på tvers av ulike områder, der statistiske prediksjoner presterte best (Mehl, 1954). Videre viste Dawes at menneskelige eksperter ikke gjorde det like godt som enkle lineære modeller, blant annet på klinisk diagnostisering (Dawes, 1979; Dawes et al., 1989). Videre er prognoser gjennomført av algoritmer mer nøyaktige enn de som er utført av mennesker. Dette er gjeldende innen for eksempel medisinsk diagnostisering (Beck et al., 2011; Grove et al., 2000).

Grove et al. (2000) gjennomførte en metaanalyse i forbindelse med diagnostisering, der de sammenlignet treffsikkerheten til kliniske og maskinelle prediksjoner innenfor medisin og psykologi. Kliniske prediksjoner er i dette tilfellet vurderinger gjennomført av psykologer og leger. Maskinelle prediksjoner innebærer blant annet statistiske og algoritmiske prediksjoner. I gjennomsnitt var maskinelle prediksjoner 10% mer nøyaktige enn de kliniske. De maskinelle prediksjonene utkonkurrerte kliniske i 33-47% av studiene som ble vurdert. I motsatt tilfelle var det kun ved 6-16% av studiene at kliniske prediksjoner var vesentlig mer nøyaktig. I omtrent halvparten av studiene var de noenlunde like nøyaktige. Dette var tilfellet uavhengig av domene (medisin/psykologi), type oppgave, hvem som foretok vurderingen (type stilling) og erfaring (nybegynner/ekspert).

Resultatene viser videre at de maskinelle prediksjonene i enda større grad var mer nøyaktige da psykologene og legene benyttet kliniske intervjuer. Årsaker kan være at mennesker er utsatt for ulike bias i kliniske vurderinger (Garb, 1998; Kahneman et al., 1982). Slike bias kan blant annet være å ikke tillegge optimal vekting til ulike symptomer, å anta at små utvalg er representative (representativitetsbias), og å legge for stor vekt på nyeste data (tilgjengelighetsbias). I tillegg får ikke mennesker tilbakemelding på hvor nøyaktige vurderingene deres er på samme måte som maskinene gjør, som igjen ikke gir de muligheten til å endre på uheldige vaner og feil gjort i vurderingene (Grove et al., 2000).

2.3.2 Utfordringer ved kunstig intelligens

Egenskapene ved deep learning-algoritmer gir typisk lite mening for mennesker. Dermed er forklaringene på modellenes resultater gjerne vanskelige, eller umulige å tolke (Davenport & Kalokata, 2019). Algoritmene er så komplekse at logikken bak dem også er ukjent selv for utviklerne. Denne mangelen på transparens kan dermed redusere troverdigheten (Vayeana et al., 2018). I den sammenheng kan det oppstå problemer hvis for eksempel en pasient har fått påvist diagnosen kreft basert på et røntgenbilde, og det ikke er mulig å forklare hvorfor vedkommende har fått det påvist (Davenport & Kalokata, 2019). Formidling til pasienten av detaljer om behandling er også viktig, og dermed må legen i det minste ha en kjennskap til det grunnleggende om hvordan algoritmene fungerer. Ved helautomatiserte medisinske avgjørelser kan graden av risiko forbundet med prosedyren, være avgjørende for hvor mye informasjon det er nødvendig å gi til pasienten om AI-teknologien som benyttes (Vayeana et al., 2018). Ved å kommunisere med pasientene om AI-teknologien som brukes, kan det også øke tilliten og aksepten (Vayeana et al., 2018).

Videre vil det også bli gjort feil av AI-systemer i forbindelse med diagnostisering og behandling. Det kan derfor være vanskelig å etablere troverdighet og fordele ansvar i forbindelse med dette. I tillegg kan det oppstå situasjoner der pasienter vil motta informasjon fra AI-systemet, der det heller hadde vært ønskelig å motta informasjonen fra en empatisk lege (Davenport & Kalokata, 2019). Det er altså ulike situasjoner hvor andre faktorer foretrekkes, som eksempelvis empati og personlig forhold, som ikke AI kan erstatte. Dermed kan ikke den menneskelige legen erstattes fullstendig (Mesanko et al., 2018).

2.3.3 Kunstig intelligens og tillit

Tillit til teknologi omfatter både tillit til selve teknologien og aktøren som leverer teknologien. Av den grunn hevder Siau & Wang (2018) at tilliten en har til teknologien og teknologileverandøren påvirker hverandre. I litteraturen synes det å være enighet om at det er essensielt å etablere en tillit til teknologien, for å få konsumenter til å ville benytte seg av den. Dette gjelder spesielt nye teknologiske løsninger som AI-teknologi (Lacey et al., 2009; PWC 2017)

I en rapport utarbeidet av Kantar (2017) fremmes det at tilliten til bruk av AI-teknologi og nye digitale tjenester varierer. Til tross for at benyttelse av kunstig intelligens har økt, er det

fremdeles en del som ikke viser like stor tillit til teknologien. Rapporten viser at da respondentene ble spurt om de stolte på at maskiner kunne hjelpe dem ved et problem, var 35% av respondentene negative til dette, og 38% var positive. Imidlertid synes det å være stor forskjell i svarene fordelt på aldersgrupper, der yngre viser større tillit enn eldre.

En annen studie som viser tilsvarende resultater så på forbrukeres vilje til å benytte seg av selvkjørende Uber. Omlag 44% av respondentene svarte at de ikke er villige til å benytte en selvkjørende bil (Gillath et al., 2020). Tilsvarende resultater støttes av andre studier, som hevder at 42% viser lav tillit til kunstig intelligens, og at 49% av respondentene ikke kunne nevne en AI-teknologi som de stolte på (Dujmovic, 2017). Den manglende tilliten til AI-teknologi kan særlig synes å omfatte situasjoner der risikoen for feil kan ha større konsekvenser, som eksempelvis i helsesektoren. På den måten kan aversjonen tilknyttet AI-teknologi i helsesektoren bli så høy at det ikke blir tatt i bruk i stor nok grad, slik at det blir vanskeligere å dra nytte av teknologiens gevinster (Asan et al., 2020).

I litteraturen er det forskjellige forklaringer på hvorfor mange ikke har tillit til AI-teknologi. Noen hevder det kan være fordi en ikke har forståelse for hvordan AI-teknologien fungerer, og derfor synes det er vanskelig å vise tillit til teknologien. En annen forklaring er at personlige karakteristika påvirker graden av tillit til teknologien (Gillath, 2020). Siau & Wang (2018) argumenterer for at følgende tre faktorer, med hver sine dimensjoner vist i parentes, påvirker graden av tillit til teknologi: Personlige karakteristikker (personlighet, evner), miljømessige forhold (kultur, type oppgave, institusjonelle faktorer) og teknologisk karakteristika (prestasjoner, prosesser, hensikt). Forfatterne mener at de teknologiske karakteristikkene har størst påvirkning på graden av tillit en viser til teknologiske produkter og tjenester. Eksempelvis vil det være enklere å lansere en teknologi som tilbys av en anerkjent bedrift som har store prestasjoner å vise til, fremfor en virksomhet som ikke har en like sterk merkevare.

Implementering av AI-teknologi i helsevesenet kan påvirke den mellommenneskelige tilliten pasientene utviser. Kerasidou et al. (2019) hevder at det er flere trekk ved AI-teknologien som påvirker tilliten pasientene har til helsevesenet, eksempelvis fravær av empati, medfølelse og tillit. Dette er trekk som spiller en vesentlig rolle i pasient-lege forholdet, blant annet for å legge til rette for høy grad av pasientautonomi. Det handler om medisinsk selvbestemmelse for pasienten, som anses å være særdeles viktig i dagens samfunn (Ursin, 2021). Derfor

hevder Kerasidou et al. (2019) at det er vesentlig å utvikle AI-systemer som klarer å ta hensyn til at pasienter har forskjellige verdier og prioriteringer relatert til sin helse. Formålet er dermed å å bevare pasientens selvbestemmelse, til tross for en økt teknologisk transformasjon i helsesektoren. I den forbindelse vurderes det som viktig gjøre noen refleksjoner rundt hvilken rolle AI-teknologien skal ha i helsevesenet, og hvordan en skal påse at en bevarer pasientautonomien.

Imidlertid reiser forfatterne også spørsmål ved hvorvidt det å overlate oppgaver til roboter vil kunne påvirke tillitsforholdet mellom legen og pasienten. Det begrunnes med at fordelene ved å benytte kunstig intelligens, som økt effektivitet, nøyaktighet og bedre personalisert behandling, potensielt kan utkonkurrere tillitsbehovet pasientene har (Zuboff sitert i Kerasidou et al., 2019). Det kan derfor tenkes at lege-pasient forholdet kan gå fra være et forhold som er basert på sårbarhet og tillit, til å bli et forhold der legene bistår pasientene ved behov, for ytterligere informasjon eller sparring (Kerasidou et al., 2019). Økt implementering av AI-teknologi i helsesektoren vil dermed kunne ha implisitte påvirkninger på mellommenneskelige forhold, som en også må ta til betraktning.

2.4.0 Algoritmeaversjon

AI-teknologi er nå raskere og mer nøyaktig enn mennesker på en rekke områder (Jussupow et al., 2020). Disse nye mulighetene resulterer i et valg mellom å benytte seg av en menneskelig ekspert eller algoritme til å utføre ulike oppgaver. Forskning viser at valgene ikke nødvendigvis baseres på objektive og rasjonelle kriterier. Mennesker foretar heller irrasjonelle beslutninger, og foretrekker mennesker fremfor algoritmer, selv når sistnevnte er bevist å være overlegne. Dette kan defineres som algoritmeaversjon (Dietvorst et al., 2015; Dietvorst et al., 2018; Castelo et al., 2019).

2.4.1 Indikatorer på algoritmeaversjon

Fra litteraturen synes det å være hovedsakelig tre måter å måle algoritmeaversjon, som Jussupow et al. (2020) illustrerer i sin metastudie. Den første formen omfatter valget mellom en algoritme eller et mennesket, tilknyttet råd eller utførelse av en oppgave. Aversjon er her tilstedeværende dersom mennesket velges fremfor algoritmen.

Den andre formen omfatter hvordan aktørenes (menneske/robot) vurderinger benyttes. Dette gjøres for eksempel ved at brukeren først utfører et eget estimat, for så å bli gitt menneskets og algoritmens vurdering. Deretter tar brukeren en endelig beslutning. Aversjon er tilstedeværende dersom de justerer estimatet sitt mindre mot algoritmen sitt estimat enn mot mennesket sitt. En annen måte å måle denne formen for aversjon på, er å undersøke hvor sterke preferansene er for de to ulike aktørene (Longoni et al., 2019). Høyere preferanser for mennesket kan derfor tolkes som aversjon.

Den siste formen er at mennesket og algoritmen blir evaluert av brukerne, både når det gjelder utfallet av oppgaven, samt en overordnet vurdering av aktørene. Dette kan for eksempel gjøres ved at brukerne vurderer tilliten de har til aktørene (Madhaven & Wiegmann, 2007; Önkal et al., 2009), samt hvor tilfredse de er (Palmeira & Spassova, 2015; Bigman & Gray, 2018). Dersom algoritmen blir rangert lavere enn menneske indikerer det aversjon.

2.4.2 Årsaker til algoritmeaversjon

Fra litteraturen fremkommer det ulike årsaker til aversjon tilknyttet bruk av AI-teknologi. Blant annet identifiserer Jussupow et al. (2020) følgende fire karakteristikker ved algoritmer, som påvirker aversjon: algoritme-autonomi, algoritmens prestasjoner, egenskaper ved algoritmen, og menneskelig involvering. Vi vil derfor i det følgende ta utgangspunkt i disse karakteristikkene når vi gjennomgår litteraturen.

Algoritme-autonomi

Algoritmer kan ha ulik grad av selvstyre. Spekteret går fra å være beslutningsstøtte, der brukeren tar den endelige beslutningen, til at algoritmen selvstendig utfører oppgaven, og at mennesket kun måler algoritmens prestasjoner. Førstnevnte kan kalles en "rådgivnings-algoritme", og sistnevnte kan kategoriseres som en "prestasjons-algoritme". I henhold til litteraturen synes brukere å vise indikasjon på aversjon mot begge typer algoritmer (Jussupow et al., 2020). Imidlertid viser studiens resultater at det er mindre grad av aversjon overfor "rådgivnings-algoritmer". Dette kan indikere at høyere grad av algoritme-autonomi fører til større grad av algoritmeaversjon. Årsaken til differansen kan blant annet forklares ved at brukeren opplever tap av kontroll når selve beslutningen overlates til algoritmen (Burton et al., 2019).

Algoritmens prestasjoner

Hvordan algoritmen presterer er avgjørende for hvordan brukere både interagerer med den, samt evaluerer den. Blant annet hevder Yeomans et al. (2019) at å vise hvordan algoritmen presterer reduserer algoritmeaversjonen. Dette støttes av Dzindolet et al. (2002), som viste at det å gi informasjon om hvordan algoritmen presterer, påvirker i hvilken grad beslutningstakerne stoler på algoritmene.

Videre finner Dietvorst et al. (2015) at valget mellom en algoritme eller menneske påvirkes av deres tidligere erfaringer med disse i den gitte situasjonen. Dersom en ikke har erfaring med algoritmen fra før er en villig til å stole på den. Imidlertid, når en ser algoritmen utføre den aktuelle oppgaven, og derfor også tidvis feiler, gjør dette at en stoler mindre på den. Som resultat kan bestlutningstakerens valg endres til å heller ville benytte et menneske. Derimot ble ikke mennesket valgt bort dersom vedkommende også gjorde feil, i likhet med algoritmen. Disse resultatene indikerer derfor at mennesker raskere forkaster en algoritme som feiler - også når mennesket gjør større feil enn algoritmen (Dietvorst et al., 2015). Resultatene viser også at algoritmeaversjonen synes å være sterkere når det er en selv som blir utsatt for trusselen om å bli erstattet av algoritmen, fremfor et annet menneske.

Tilsvarende funn viser også Logg et al. (2019) til. Forfatterne finner at mennesker har en tendens til å dømme feil utført av AI-teknologi hardere, enn om et menneske skulle gjøre samme feilen. Videre kom de frem til at eksperter heller stolte på egen dømmekraft enn algoritmen sine råd, til tross for at algoritmen gir bedre prediksjoner. Resultatene var konsistente uavhengig alder. Imidlertid så en at personer som var mer komfortabel med tall viste større tillit til algoritmen.

En forklaring på at algoritme og menneske vurderes ulikter teorien om "expectation-diconfirmation" (Bhattacherjee & Premkumar, 2004). Det er troen på at algoritmer er perfekte og ikke gjør feil (Dzindolet et al., 2002; Madhaven & Wiegmann, 2007; Goodyear et al., 2016). Når en innser at dette ikke er tilfellet, er sannsynligheten større for at en skylder på og "straffer" algoritmen, på bakgrunn av disse forventningene, enn hva de hadde gjort med mennesker.

Dermed er det klart fra litteraturen at algoritmer fortsatt gjør det bedre enn mennesket på flere oppgaver objektivt sett, men at en likevel evaluerer de ulikt. Noen oppgaver innebærer også større konsekvenser dersom de utføres dårlig, som diagnostisering eller behandling av sykdom. Brukere synes her å være mindre villig til å stole på algoritmer når risikoen er høyere (Castelo et al., 2019).

Egenskaper ved algoritmen

Castelo et al. (2019) viser at type oppgave, og de oppfattede egenskapene til algoritmen, er en viktig driver for aversjon. Funnene synes å være spesielt gjeldende ved subjektive oppgaver som innebærer moralske avgjørelser eller affektive aspekter. Dette begrunnes med at det da er nødvendig med menneskelige egenskaper som empati og intuisjon, noe det vurderes som at algoritmene ikke innehar. Imidlertid kan dette synes å ikke være en korrekt antakelse, siden algoritmer stadig blir bedre til å utføre subjektive oppgaver, eksempelvis ved at de klarer å analysere ansiktsuttrykk og følelser (Castelo et al., 2019).

Derimot ble algoritmene sett på som bedre til å kunne utføre objektive, kvantifiserbare oppgaver som innebærer egenskaper som logikk og rasjonalitet. Imidlertid kan oppfattelsen av hvor subjektiv oppgaven er påvirkes. Castelo et al. (2019) viser at hvis oppgaven blant annet ble omformulert til å være av mer objektiv karakter, ble algoritmeaversjonen redusert. I tillegg justerte de karakteristikkene ved algoritmene for å gjøre de mer menneskelige, som også gjorde at aversjonen ble redusert. Det kan derfor totalt sett synes at algoritmeaversjon reduseres, ved at brukerne i større grad oppfatter at algoritmene har de nødvendige egenskapene til å kunne gjennomføre oppgaven.

Tilsvarende viser Lee (2018) at graden av algoritmeaversjon særlig avhenger av om oppgaven er av mer subjektiv eller objektiv karakter. Førstnevnte krever ferdigheter som eksempelvis empati og intuisjon, mens sistnevnte blant annet krever logikk og rasjonalitet. I forbindelse med objektive oppgaver ble menneske og algoritme vurdert likt, blant annet på grad av tillit og hvor rettferdig avgjørelsene var. For subjektive oppgaver ble egenskapene rettferdighet og tillit trukket frem som de viktigste hos menneskelige beslutningstakere, mens egenskapene pålitelighet og fravær av bias var de viktigste hos algoritmer (Lee, 2018). Resultatene viste også at det oppstår negative følelser blant respondentene når algoritmer utfører subjektive oppgaver. I motsatt fall genereres det positive følelser når et menneske utfører oppgaven.

Dette fordi en kan oppnå sosial anerkjennelse om en benytter en menneskelig beslutningstaker.

Videre viser også enkelte studier at mennesker har en tendens til å utvise algorimeaversjon tilknyttet oppgaver der det oppleves at det er behov for å ta individuelle hensyn. Longoni et al. (2019) ser på tvers av ni separate studier, hvordan "unikhetsforsømmelse" hos pasienter påvirker bruk av AI-teknologi i helsesektoren. Unikhetforsømmelse defineres som "en bekymring for at AI-teknologi ikke evner, i like stor grad som mennesker, å ta hensyn til ens unike karakteristika og omstendigheter, som resulterer i en motstand mot benyttelse av medisinsk AI-teknologi" (Longoni et al., 2019).

I henhold til studiens resultater har forbrukerne mindre sannsynlighet, og lavere betalingsvilje, for å benytte kunstig intelligens enn menneskelige fagpersoner. Videre viser resultatene at en er mindre følsom for forskjeller i leverandørenes ytelse ved bruk av menneske. Det anses også som negativt at en leverandør tilbyr AI-teknologi fremfor menneskelige ressurser. Videre er mennesker som anser seg selv som spesielt unike mindre tilbøyelig til å benytte AI-teknologi. Imidlertid reduseres motstanden når det presiseres at behandlingen er tilpasset den enkelte pasienten. Longoni et al. (2019) viser også til at en kan unngå følelsen av unikhetsforsømmelse hos pasientene ved å øke graden av informasjon AI-teknologi får om pasienten, slik at teknologien kan personalisere behandlingen. Tilbøyeligheten til å benytte kunstig intelligens beror også på hvem som er beslutningstaker. Resultatene fra studien viser at motstanden mot å benytte teknologien elimineres hvis en benytter AI i medisin som en støttefunksjon, der lege er beslutningstaker (Longoni et al., 2019).

Imidlertid har Longoni et al. (2019) sine studier fått kritikk for forskningsdesignet og dets mulighet til å måle algoritmeaversjon. Pezzo & Beckstead (2020) fremmer i sin artikkel at resultatene fort kan tolkes feil. Forfatterne gir uttrykk for at studien ikke gir mulighet til å direkte sammenligne mennesker og AI-teknologi som behandlingsmetoder. Det henvises her blant annet til at utvalget ikke fikk velge mellom de forskjellige tjenestealternativene.

Menneskelig involvering

Mennesker kan påvirke algoritmer på to måter. Dette påvirker igjen hvilke egenskaper brukerne oppfatter at algoritmen innehar. Det første måten er at mennesker sikrer kvaliteten

på algoritmen, ved at de er involvert i utviklingen og opplæringen av den. I den forbindelse fremkommer det at brukere foretrekker algoritmer der det informeres om at teknologien er lært opp av mennesker (Jago, 2019). Den andre måten er at mennesker er involvert i bruken av algoritmen, ved at mennesket sitter med den endelige beslutningen (Palmeira & Spassova, 2015). På denne måten fremstår algoritmen som mer kontrollerbar (Dietvorst et al., 2018), og ble foretrukket fremfor kun å benytte algoritmen alene (Palmeira & Spassova, 2015; Longoni et al., 2019). Imidlertid er det uklart om resultatet blir bedre enn når en algoritme alene utfører oppgaven alene (Yeomans et al., 2019). Totalt sett fører menneskelig involvering til at algoritmen anses i større grad å inneha de passende og nødvendige egenskapene. På denne måten reduseres graden av autonomi hos algoritmen. Dette fører igjen til at aversjonen reduseres.

Dietvorst et al. (2018) fant i sine studier at mennesker er mer villig til å benytte seg av algoritmer når en har mulighet til å kunne foreta egne justeringer. I tillegg viser resultatene at en er lite sensitiv overfor graden en har mulighet til å justere algoritmen: Da deltakerne hadde mulighet til å gjennomføre små justeringer var resultatet omtrent det samme som for deltakerne som hadde mulighet til å foreta større justeringer. Da deltakerne kunne foreta justeringer, ble det i tillegg rapportert om høyere tilfredshet. Deltakerne vurderte det også som at algoritmen presterte relativt bedre enn dem selv, sammenlignet med deltakere som ikke fikk muligheten til å modifisere algoritmens prognoser. Det ble også rapportert om at deltakerne var like tilfredse - uavhengig av hvor mye de kunne justere algoritmens prognose.

Å la mennesker justere algoritmens prognoser gjør ofte resultatet mindre presist og kvaliteten dårligere (Carbone et al., 1983; Goodwin & Fildes, 1999; Hogarth & Makridakis, 1981; Lim & O'Connor 1995; Willemain, 1991). Imidlertid er fordelene ved å få flere til å benytte algoritmen større. Dersom en tillater mennesker å justere algoritmen i noen grad, kan det øke viljen til å benytte algoritmer vesentlig. På denne måten vil en få mer presise resultater og bedre prediksjoner, sammenlignet med om algoritmen ikke ble benyttet i det hele tatt.

3 Hypoteser og forskningsmodell

3.1.0 Valg av behandlingsform

Kunstig intelligens blir som nevnt i litteraturgjennomgangen stadig mer utbredt, også innen helsevesenet. Det er allerede en rekke bedrifter og organisasjoner som har vist at AI kan forbedre kvaliteten på helsetjenester og/eller redusere kostnadene (Davenport & Kalokata, 2019). Med den raske utviklingen innen kunstig intelligens, blir det stadig benyttet til flere oppgaver som tidligere kun var mulig å utføre av mennesker. Algoritmer er nå raskere og mer nøyaktig enn mennesker på en rekke områder (Jussupow et al., 2020). Dette kan blant annet sees i en metaanalyse innenfor medisin og psykologi, der prediksjonene gjort av maskiner i gjennomsnitt var 10% mer nøyaktige (Grove et al., 2000). Med stadig økt eksponering for teknologi øker kjennskapen, bruken og således villigheten til å benytte seg av kunstig intelligens (Davenport & Kalokata, 2019).

Til tross for at benyttelsen av AI-teknologi har økt, er det fremdeles en del som ikke viser like stor tillit til teknologien (Dujmovic, 2017). Forskning viser videre at det ofte eksisterer algoritmeaversjon (Dietvorst et al., 2015; Dietvorst et al., 2018; Castelo et al., 2019). I litteraturgjennomgangen foreslås det tre ulike former for algoritmeaversjon, der den ene er valget mellom algoritme eller menneske (Jussupow et al., 2020). Longoni et al., (2019) har i sin studie fått kritikk for at deltakerne som deltok ikke fikk et valg av behandlingsmetode (AI vs menneske) da algoritmeaversjon ble målt (Pezzo & Beckstead, 2020). Derfor har vi, i vårt studie, blant annet benyttet valg for å se om det kan gi en indikasjon på algoritmeaversjon. Vi tror, i tråd med litteraturgjennomgangen, at det vil være flere som vil benytte seg av et menneske enn en robot - også i konteksten med en medisinsk konsultasjon. Imidlertid, for å gjøre utgangspunktet for deltakernes valg av behandlingsform likt, har vi opplyst om at sannsynligheten for korrekt diagnostisering og behandling er lik. Vi har derfor følgende hypotese:

H1: En høyere andel av deltakerne vil velge fastlege enn robot for en medisinsk konsultasjon, når sannsynligheten for korrekt diagnostisering og behandling er den samme

Den manglende tilliten til AI-teknologi er særlig fremtredende der risikoen for feil medfører store konsekvenser (Asan et al., 2020). Noen oppgaver kan medføre store konsekvenser dersom det utføres dårlig, som diagnostisering eller behandling av sykdom. Av den grunn

synes brukere å være mindre villig til å stole på algoritmer når risikoen er høyere (Castelo et al., 2019). Imidlertid kan det tenkes at det er ulikheter også innad i helsevesenet, hva gjelder konsekvenser om en feil skulle forekomme. Dermed vil også konsekvensene ved ulike feil variere. På bakgrunn av litteraturgjennomgangen tror vi derfor at valget mellom robot og menneske kan påvirkes av den opplevde alvorlighetsgraden av egne symptomer, og dermed også hvor store konsekvensene av feil kan være. Dette gir følgende hypotese:

H1a: Deltakerne med symptomer av lav alvorlighetsgrad velger oftere robot, enn deltakerne med symptomer av høy alvorlighetsgrad

Dette innebærer at effekten postulert i H1 er moderert av alvorlighetsgrad av symptomene. Mer alvorlige symptomer vil føre til en høyere valgandel fastlege.

3.2.0 Grad av tillit utvist til behandlingsform

Som utledet i litteraturen har tillitsbegrepet flere definisjoner og favner bredt. Tilliten en viser i en gitt kontekst avhenger av flere faktorer, som blant annet tidligere personlige erfaringer (Kramer et al., 1996), og en overordnet vurdering av av den andre parts kompetanse og holdninger (Sekhon et al., 2014; Mayer et al., 1996; Korczynski, 2000; Tullberg, 2007). Litteraturen viser også at tillit er sterkt tilknyttet oppfattet risiko (Boon & Holmes, 1991; Giddens, 1991; Sekhon et al., 2014; Rousseau et al., 1998), og særlig i tilfeller der det er asymmetrisk informasjon mellom tillitsgiver og tillitshaver, som helsevesenet er et eksempel på (Dyer, 2016; Alaszewiski, 2003).

Fra litteraturen om tillit og helse, viser studier at tilliten mennesker viser overfor helsepersonell i stor grad baseres på i) aktørens velvilje til å hjelpe, og ii) kompetansen de besitter (Hyman, 2001; Newcomer, 1997). Tillit er også en måte å måle algoritmeaversjon på, ved at tilliten til aktørene måles og sammenlignes. Dersom algoritmen blir rangert lavere enn mennesket, indikerer det at det er en aversjon tilstede (Madhaven & Wiegmann, 2007; Önkal et al., 2009). Det er derfor interessant å undersøke om algoritmeaversjon, målt ved konseptet tillit, finner sted når respondentene selv har valgt behandlingsformen (fastlege/robot), og utfallet av konsultasjonen er den samme. Vi har derfor følgende hypotese:

H2: Deltakerne som har valgt robot utviser lavere grad av tillit, enn deltakerne som har valgt fastlege

Videre tror vi, basert på litteraturgjennomgangen, at høy alvorlighetsgrad på symptomene, som medfører at risikoen oppfattes som høyere, vil påvirke graden av tillit til roboten. Dette vil da komme til syne i form av at tilliten blir lavere, som videre kan indikere en høyere grad av algoritmeaversjon. Det gir følgende hypotese:

H2a: Deltakerne som har valgt robot utviser lavere grad av tillit når de har symptomer av høy alvorlighetsgrad, enn deltakerne med symptomer av lav alvorlighetsgrad

Dette innebærer at effekten postulert i H2 er moderert av alvorlighetsgrad av symptomene. Mer alvorlige symptomer vil føre til en lavere grad av tillit utvist til robot.

3.3.0 Grad av oppfattet risiko i forbindelse med behandlingsform

Studier viser at konseptet oppfattet risiko blant annet kan måles gjennom indikatorene finansiell risiko, ytelsesrisiko, fysisk risiko, psykologisk risiko, sosial risiko og bekvemmelighetsrisiko (Mohtar & Abbas, 2015; Featherman & Pavlou, 2003). Litteraturen viser videre at graden av tillit og risiko, som mennesket viser til teknologi, øker med gjentatt bruk av gjenstanden (Lacey et al., 2009; Featherman & Pavlou, 2003). Det er derfor rimelig å tenke at deltakerne som velger mennesket vil ha en lavere opplevd fysisk risiko tilknyttet diagnostisering og behandling, enn de som velger robot. Dette begrunnes med at deltakerne sannsynligvis har benyttet seg av en fastlegetjeneste før, og derfor har større tiltro til denne behandlingsmetoden, enn til en robot.

Tilsvarende er det også rimelig å anta at deltakerne som har valgt robot opplever en større psykologisk risiko. Dette fordi AI-teknologi ikke benyttes i norsk helsevesen i forbindelse med medisinske konsultasjoner i dag. Av den grunn har ikke deltakerne selv erfart behandlingsmetoden, men de har informasjonen de fikk innledningsvis om at begge behandlingsformene er like gode. Derfor er det naturlig å tenke at dette kan skape en usikkerhet som resulterer i en indre uro.

Ut fra litteraturen kan vi ikke se at oppfattet risiko har blitt brukt som en indikator på algoritmeaversjon. Imidlertid viser blant annet Asan et al. (2020) sin studie at manglende tillit til AI-teknologi særlig gjelder i situasjoner der den oppfattede risikoen for feil kan ha større konsekvenser, noe helsevesenet kan være et godt eksempel på. Til tross for at det er en del studier på forholdet mellom tillit og risiko, vurderes det som at det er mindre informasjon om forholdet mellom tillit og oppfattet risiko i forbindelse med algoritmeaversjon. Imidlertid sier teorien at tillit og risiko kan korrelere negativt med hverandre (Partick, 2002), og vi tror derfor at risiko også kan benyttes som en indikator på algoritmeaversjon. Dette begrunnes med at økt risiko tilknyttet vurderingen av en medisinsk konsultasjon, kan resultere i en høyere terskel for å utvise tillit til behandlingsformen. Det kan derfor tenkes at respondentene vil være mer kritisk til diagnostiseringen og behandlingen en får. Dette gir følgende hypotese:

H3: Deltakerne som har valgt robot utviser høyere grad av oppfattet risiko, enn deltakerne som har valgt fastlege

Basert på litteraturen kan det videre tenkes at den opplevde usikkerheten og risikoen tilknyttet diagnostisering og behandling av robot, kan være ulik på bakgrunn av alvorlighetsgrad på symptomer. Dette fordi det kan innebære ulik grad av konsekvenser dersom noe går galt (Asan et al., 2020). Derfor tror vi at høy alvorlighetsgrad, vil medføre at risikoen vil oppfattes som høyere. Dette kan dermed indikere høyere grad av algoritmeaversjon. Det gir følgende hypotese:

H3a: Deltakerne som har valgt robot utviser høyere grad av oppfattet risiko når de har symptomer av høy alvorlighetsgrad, enn deltakerne med symptomer av lav alvorlighetsgrad

Dette innebærer at effekten postulert i H3 er moderert av alvorlighetsgrad av symptomene. Mer alvorlige symptomer vil føre til en høyere grad av oppfattet risiko til robot.

3.4.0 Evaluering av behandlingsform

Litteraturgjennomgangen viser at et mål på algoritmeaversjon er at aktøren blir evaluert av brukerne, der det foreligger indikasjon på algoritmeaversjon hvis AI-teknologien får en dårligere evaluering enn menneske. Evalueringen omfatter både utfallet av oppgaven, samt en vurdering av selve aktøren (Madhaven & Wiegmann, 2007; Önkal et al., 2009; Palmeira &

Spassova, 2015; Bigman & Gray, 2018). I dette tilfelle er aktørene fastlegen og roboten. Dersom roboten blir rangert lavere enn fastlegen indikerer det aversjon.

Hvordan roboten presterer kan blant annet være avgjørende for hvordan deltakerne evaluerer den. I den sammenheng viser Dzindolet et al. (2002) til at å gi informasjon om hvordan algoritmen presterer påvirker i hvilken grad beslutningstakerne stoler på algoritmene. Derfor har vi informert om at begge aktører har en treffsikkerhet på 80% for diagnostiseringen. En rekke studier viser også at beslutningstakere reagerer på at algoritmen feiler. Når deltakerne innser at algoritmen ikke er perfekt, er sannsynligheten større for at de "straffer" den enn hva de hadde gjort med mennesker. Dermed kan algoritmen fortsatt gjøre det bedre enn mennesket objektivt sett, men de blir likevel evaluert ulikt (Dietvorst et al., 2015; Logg et al., 2009).

Fra litteraturen fremgår det at evalueringen som respondentene oppgir generelt er lavere for algoritme enn for menneske, og at dette indikerer algoritmeaversjon (Jussupow et al., 2020). I vårt tilfelle er evaluering en overordnet vurdering av behandlingsformen. Vi har tilpasset det vår kontekst ved å fokusere på dimensjonene tilfredshet og forventet anger, siden vi anser dette som de mest passende. I en kontekst med medisinsk konsultasjon tror vi at respondentenes tilfredshet med roboten vil være lavere, uavhengig av alvorlighetsgrad. På denne måten kan derfor evalueringen, i form av tilfredshet og forventet anger, gi indikasjon på algoritmeaversjon. Dette gir følgende hypotese:

H4: Deltakerne som har valgt robot utviser lavere grad av tilfredshet, enn deltakerne som har valgt fastlege

Vi tror videre at høyere alvorlighetsgrad vil påvirke evalueringen av roboten negativt. Mer presist tenker vi at respondentene med høy alvorlighetsgrad vil være mindre tilfredse, enn respondenter med lav alvorlighetsgrad. Dette gir derfor følgende hypotese:

H4a: Deltakerne som har valgt robot utviser lavere grad av tilfredshet når de har symptomer av høy alvorlighetsgrad, enn deltakerne med symptomer av lav alvorlighetsgrad.

Dette innebærer at effekten postulert i H4 er moderert av alvorlighetsgrad av symptomene. Mer alvorlige symptomer vil føre til en lavere grad av tilfredshet utvist til robot.

Vi får de samme hypotesene for forventet anger:

H5: Deltakerne som har valg robot utviser høyere grad av forventet anger, enn deltakerne som har valgt fastlege

H5a: Deltakerne som har valgt robot utviser høyere grad av forventet anger når de har symptomer av høy alvorlighetsgrad, enn deltakerne med symptomer av lav alvorlighetsgrad

Dette innebærer at effekten postulert i H5 er moderert av alvorlighetsgrad av symptomene. Mer alvorlige symptomer vil føre til en høyere grad av forventet anger utvist til robot.

3.5.0 Medierende effekter

Ettersom indikasjon på algoritmeaversjon kan måles på ulike måter, er en naturlig antakelse at de forskjellige indikatorene også kan påvirke hverandre. Litteraturen viser at det kan være et positivt forhold mellom tillit og kundetilfredshet (Deepak et al., 2002; Anderson and Narus, 1991; Gummerus et al., 2004; Taylor and Hunter 2003). Av den grunn tror vi at evalueringen deltakerne gjør av den valgte behandlingsformen kan påvirkes av graden av tillit de utviser. Det vil si at graden av tilfredshet og forventet anger delvis kan forklares av tilliten deltakerne har til den respektive behandlingsformen (fastlege/robot). Videre tror vi at det samme kan gjelde for oppfattet risiko, basert på at det kan være en negativ korrelasjon mellom tillit og risiko (Partick, 2002). Vi ønsker derfor å teste om tillit og oppfattet risiko kan fungere som mediatorer for evaluering av den valgte behandlingsformen. Dette gir følgende hypoteser:

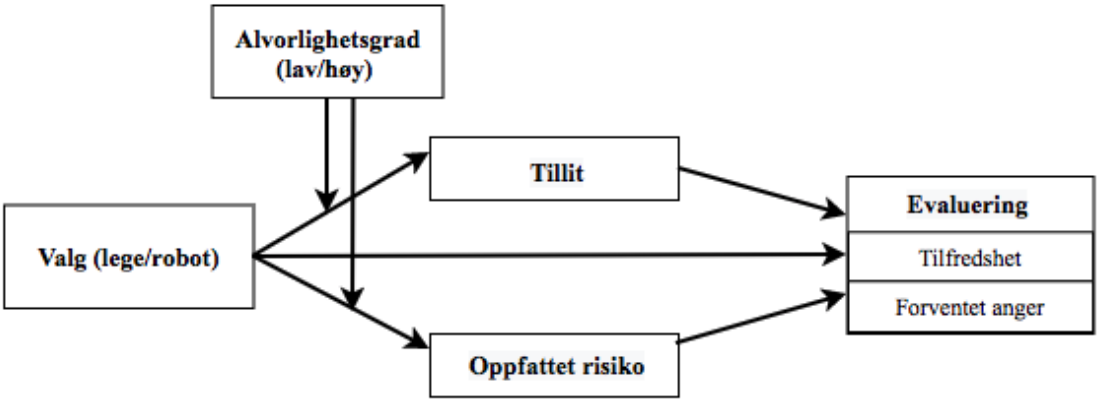
H6a: Effekten som valg har på tilfredshet, kan delvis forklares av tillit
H6b: Effekten som valg har på tilfredshet, kan delvis forklares av oppfattet risiko

H7a: Effekten som valg har på forventet anger, kan delvis forklares av tillit
H7b: Effekten som valg har på forventet anger, kan delvis forklares av oppfattet risiko

3.6.0 Forskningsmodell

Forskningsmodellen, som vist i figur 2, illustrerer våre hypoteser om årsak-virkningsforhold. Forskningsmodellen søker å se om det foreligger en indikasjon på algoritmeaversjon, representert ved våre hypoteser og forholdet mellom disse. Alvorlighetsgrad på symptomer

(lav/høy) er en uavhengig variabel da det er den som varieres. Imidlertid er det også en moderator, ettersom den påvirker relasjonen mellom variablene i modellen. Valg av behandlingsmetode (fastlege/robot) er en avhengig variabel for H1, og en uavhengig variabel for de resterende hypotesene. I modellen medieres effekten som valg av behandlingsform har på evaluering av tillit og oppfattet risiko. Valgt behandlingsform blir evaluert av deltakerne, ved å vurdere grad av tilfredshet og forventet anger. Dette er den avhengige variabelen.



Figur 2: Forskningsmodell

4 Metode

4.1.0 Forskningsdesign og tilnærming

Forskningsdesign defineres som en overordnet plan for hvordan forskningsspørsmålet skal besvares (Saunders et al., 2019), og omfatter design, tilnærming og forskningsmetode.

Et forklarende design er passende å benytte når en skal se på kausale sammenhenger mellom variabler (Saunders et al., 2019). I dette tilfellet vil vi se på om det foreligger algoritmeaversjon ved å se på ulike indikatorer, og hvordan aversjonen eventuelt endres når alvorlighetsgraden av symptomene endres. Vi benytter kvantitativ metode for å samle inn data, ved å gjennomføre et eksperiment. På denne måten får vi standardiserte data fra mange deltakere, som enkelt kan sammenlignes siden de er delt inn i ulike grupper. Av den grunn er det godt egnet for hypotesetesting (Grønmo, 2020). Studien har en en abduktiv fremgangsmåte. Det er passende når det er mye informasjon om et gitt tema, men betydelig mindre informasjon om det konkrete området en ønsker å studere, som det er i dette tilfellet (Achinstein, 2001). En slik fremgangsmåte tillater oss å ha en forskningsprosess der en kan

tilpasse og utvide eksisterende teori ut ifra funn. Oppgavens hypoteser er utledet på bakgrunn av eksisterende teori, som vil bli testet ved analyse av primærdata.

4.2.0 Forskningsstrategi

Forskningsstrategien omfatter den konkrete fremgangsmåten som benyttes for å besvare forskningsspørsmålet (Saunders et al., 2019). I vårt tilfelle er det benyttet en kvantitativ metode for datainnsamling ved bruk av en spørreundersøkelse. Ettersom vi ønsker å se på kausalitet der to variabler varieres, henholdsvis den oppfattede alvorlighetsgraden til symptomene og valg av behandlingsform, er det hensiktsmessig å gjennomføre et eksperiment. Siden det er delt inn i eksperimentgrupper er dataene enkle å sammenligne, og godt egnet for hypotesetesting. I tillegg har vi kontroll over omgivelsesfaktorene, slik at vi kan isolere faktorer som kan påvirke studiens interne validitet.

Eksperimentet er utformet med et 2*2 between-subject design. Videre er eksperimentet utført som en spørreundersøkelse. Manipulasjonen er de oppgitte symptomene. Deltakerne blir tilfeldig tilordnet en eksperimentgruppe med høy eller lav alvorlighetsgrad på symptomene. I tillegg gjør respondentene et valg av behandlingsform, slik at de på denne måten deles videre inn i grupper som ikke er tilfeldig. Dette gjør at vi ikke har full kontroll på eksperimentet. Valget begrunnes imidlertid med at studien blir mer reell, samt at det muliggjør en måte å måle indikasjon på algoritmeaversjon (Jussupow et al., 2020).

Valg av eksperiment med at en spørreundersøkelse begrunnes blant annet med at det er lett å kontrollere, samt at det er standardisert. Dette gir oss mulighet til å isolere effekter og samle inn store datamengder, som også er enkelt å sammenligne (Jacobsen, 2015). Imidlertid er det også noen svakheter ved å benytte spørreundersøkelser ved eksperiment. Spørreundersøkelser er utformet på forhånd, som gjør det vanskelig å foreta justeringer underveis, og det er heller ikke mulig å stille oppfølgingsspørsmål. En får heller ikke innsikt i hvorfor respondentene svarer som de gjør. For å imøtekomme noen av svakhetene bør en derfor gjennomføre en pre-test i forkant (Saunders et al., 2019). Vi anser det likevel som hensiktsmessig å benytte spørreundersøkelse siden det er en anerkjent metode ved eksperiment, samtidig som at standardiseringen tillater oss å samle inn store datamengder, som både er sammenlignbare og som kan benyttes til hypotesetesting.

4.3.0 Datainnsamling

4.3.1 Utvalg

Hensikten med å samle inn data gjennom kvantitativ metode er å danne et representativt bilde av populasjonen (Jacobsen, 2015). Populasjonen i vårt tilfellet er hele Norges befolkning etter fylte 18 år. Det er fordi alle har en fastlege, og har i varierende grad behov for helsetjenester. Vi har satt aldersgrensen på 18 år, slik at en er myndig og har mulighet til å ta selvstendige avgjørelser. I dette tilfellet er det ikke praktisk mulig å innhente data om hele populasjonen. Det må derfor benyttes et utvalg (Saunders et al., 2019).

Respondenter ble anskaffet ved bruk av frivillig deltakelse og selvutvelgelse. Vi oppfordret også respondenter til å videresende undersøkelsen, slik at det ble anskaffet respondenter ved en snøballeffekt. Metoden er fordelaktig å benytte, fordi det muliggjør rekruttering av mange respondenter, samt at det er gratis og raskt å gjennomføre (Jacobsen, 2015). Med utgangspunkt i at undersøkelsen ble sendt til vår omgangskrets, kan utvalget være en noe homogen gruppe. Ulempen er at det derfor kan være et systematisk skjevt utvalg, og dermed kan det være relevante grupper som ikke er inkludert. Av den grunn er det ingen garanti for at utvalget er representativt for populasjonen, noe det må tas hensyn til i en konklusjon. Det antas derfor at effektstørrelsen mellom de ulike gruppene kan være små. Effektstørrelsen defineres som et mål på størrelsesforskjellen mellom grupper. Dette er nyttig i tilfeller der målingene ikke har noen indre verdi, som ved bruk av Likert-skala (Sullivan & Feinn, 2012).

Ved beregning av størrelsen på utvalget har vi benyttet følgende formel:

$$n = \frac{2\left(Z_a + Z_{1-\beta}\right)^{2}\sigma^{2}}{\Delta^2}$$

Av formelen er n = antall deltakere per gruppe, $Z\alpha$ definerer oppgavens konfidensnivå og $Z1-\beta$ er studiens styrke. Videre er σ det estimerte standardavviket, samt at Δ omfatter den estimerte effektstørrelsen (Kadam & Bhalerao, 2010). For oppgaven benyttes et 95%-konfidensnivå ($Z\alpha$ = 1.96) og styrke på 5% ($Z1-\beta$ = 1,65). Vi estimerer et standardavvik på 1,2. Dette begrunnes med at utvalget anses som noe homogent, samt at målingen av responsen blir foretatt ved en Likert-skala på fem og syv punkter, noe som også begrenser spredningen naturlig. Siden vi forventer relativt små effekter settes effektstørrelsen til 1. Vi får n = 38 når

vi setter de oppgitte verdiene inn i formelen. Dette indikerer at vi bør ha minst 38 i deltakere i hver gruppe. Vi fikk til slutt totalt 239 respondenter.

4.3.2 Utforming av spørreskjema

Spørreundersøkelsen er utformet i Qualtrics. Det er totalt 22 spørsmål. Av disse er åtte tilknyttet kontrollvariabler. Alder, kjønn, arbeidsstatus og spørsmål om en snuser, kommer i begynnelsen av undersøkelsen. Spørsmål vedrørende nevrotisisme, risikovillighet, hvor tillitsfull en er, og teknologikompetanse er imidlertid lagt til slutt. Dette begrunnes med at spørsmålene ikke potensielt skal forstyrre svarene til respondentene. Etter de innledende spørsmålene er respondentene randomisert inn i to grupper ved bruk av randomiseringsfunksjonen til Qualtrics. De får da ulik informasjon om sykdomsbildet i form av at symptomene varieres mellom å være av lav og høy alvorlighetsgrad. Videre foretar de et valg om en robot eller en fastlege skal gjennomføre konsultasjonen. På denne måten inndeler de seg selv videre i grupper.

De påfølgende spørsmålene er like - uavhengig av symptomene de får og valget de foretar. Eneste forskjell mellom gruppene i undersøkelsen er at spørsmålene tilpasses etter valgte behandlingsform. Det er 10 spørsmål tilknyttet hypotesene, der hensikten er å måle en indikasjon på algoritmeaversjon. Dette måles ved valg av aktør, samt ved konseptene: Tillit, oppfattet risiko og evaluering. Sistnevnte er målt gjennom dimensjonene tilfredshet og forventet anger, som vi ønsker å se på hver for seg. Spørsmålene blir besvart ved hjelp av en sjutrinns eller femtrinns Likert-skala. Ved spørsmål der respondenten må ta et valg, er rekkefølgen på svaralternativene randomisert i Qualtrics, slik at det ikke oppstår en rekkefølgeeffekt.

Det er sentralt å påpeke at vi har begrenset antall spørsmål, både i forbindelse med kontrollvariablene og selve eksperimentet. Dette er gjort for å få nok deltakere til å svare på undersøkelsen, samt for å unngå “careless responding”. Dermed kan det være en fare for at vi ikke klarer å måle konseptene godt nok, som igjen kan påvirke resultatene våre.

4.3.3 Gjennomføring

Eksperimentet er gjennomført ved at respondenter besvarer undersøkelsen via PC, nettbrett eller mobil når og hvor de selv ønsker. Fordelene er at det er lav terskel for å delta siden det er

raskt og enkelt. Dette var viktig for å kunne få inn nok respondenter. På den andre siden kan respondentene bli påvirket av at undersøkelsen gjennomføres i ulike omgivelser og tidspunkt på døgnet. På den måten kan det være andre faktorer som påvirker respondentenes svar. Slike påvirkningsfaktorer kan være bråk og forstyrrelser i omgivelsene, samt om en er sulten, uopplagt eller lignende. Imidlertid er utvalget av såpass stor størrelse at det vurderes å ikke ha en særlig påvirkning på resultatet.

4.4.0 Skalaer og variabler

4.4.1 Skalaer

Variablene ble målt ved spørsmål eller påstander som ble stilt til deltakerne. Det ble benyttet syvpunkts og fempunkts Likert-skala, der svaralternativene gikk fra "i svært liten grad" til "i svært stor/høy grad". Grunnet skalaenes oddeetallutforming er det et naturlig midtpunkt med alternativet "4", eller "verken enig eller uenig" - avhengig av antall punkter på skalaen (Saunders et al., 2019). Begrunnelsen for å ha syvpunktsskala er at vi ønsket å fange opp variasjonen i responsene uten at det ble for mange alternativer. Vi besluttet å benytte fempunktsskala relatert til personlighetstrekk og påstander om deltakerne, siden dette er vanlig ved personlighetstester (Larsen et al., 2013). I tråd med Saunders et al. (2019) sin anbefaling om konsistens, hadde alle spørsmålene samme svaralternativer bortsett fra kontrollvariablene, og der deltakerne foretar valg av behandlingsform. Siden vi anså det som hensiktsmessig at deltakerne svarte på alle spørsmålene og påstandene, aktiverte vi tvungen respons i Qualtrics.

4.4.2 Uavhengig og modererende variabel: Opplevd alvorlighetsgrad av symptomer

Den uavhengige variabelen er den som systematisk endres for å måle effekten på den avhengige variabelen (Sanders et al., 2019). I vårt tilfelle er dette alvorlighetsgraden av symptomene. Denne variabelen blir også en modererende variabel, fordi den påvirker relasjonen til den avhengige variabelen. Respondentene har fått informasjon om symptomer av lav eller høy alvorlighetsgrad. Informasjonen de fikk om førstnevnte tilstand er: "Sår hals, feber, svelgevansker, hodepine, hovne og ømme lymfeknuter på halsen. På bakgrunn av symptomene dine foretar du noen søk på Google. Du er relativt sikker på at du har akutt halsbetennelse, og derfor må få antibiotika." Om sistnevnte tilstand er det opplyst følgende informasjon: "Muskel- og leddsmerter, slapphet, vondt i hode og dårlig søvn. På bakgrunn av

symptomene dine foretar du noen søk på Google, og er relativt usikker på hva det kan være - alt fra leddgikt og kreft, til utbrenthet."

Vi har benyttet Norsk Helseinformatikk, samt vært i samtale med en lege for å finne symptomer som er passende (personlig samtale med lege Jørgen Utvoll, 15.03.21). Det har vært viktig å gjøre informasjonen så reell og tydelig som mulig. Av den grunn er det også opplyst om at vedkommende har søkt opp symptomene selv, noe som i stor grad er normalt. Her har vi presisert hva et slikt nettsøk på symptomene kan vise. Tilfellet med lav alvorlighetsgrad viser tydelig diagnosen det mest sannsynlig er, og at dette er sykdom som innebærer lav risiko. Det motsatte gjelder ved høy alvorlighetsgrad. Her indikeres det at symptomene kan skyldes mye forskjellig, der det også kan være alvorlig. Ved å gjennomføre pre-tester undersøkte vi også hvordan disse symptomene ble oppfattet. Som følge av dette ble det foretatt små endringer i formuleringer for å tydeliggjøre symptomene ytterligere. I tillegg har vi inkludert kontrollspørsmål i selve undersøkelsen, der det stilles spørsmål om hvor alvorlig symptomene oppfattes å være.

4.4.3 Uavhengig og avhengig variabel: Valg mellom robot og fastlege

Når deltakerne foretar et valg av behandlingsmetode, er det en avhengig variabel, mens det er en uavhengig variabel i de resterende hypotesene. Valg av behandlingsmetode er mellom robot og fastlege, henholdsvis representert gjennom IBM Watson og Dr. Johansen. Vi besluttet å benytte navnet på en virkelig robot som eksisterer i dag for å gjøre det mer reelt og virkelighetsnært. Imidlertid kan kjennskapen og holdningene til selskapet IBM og produktet Watson påvirke svarene. Vi anser imidlertid sannsynligheten for at merkevareassosiasjoner i forbindelse med et såpass nytt fenomen ikke vil være problematisk. Hva gjelder fastlege, valgte vi å benytte et alminnelig navn. Det er kun etternavn som er opplyst, slik at ikke trekk ved fastlegen skal påvirke respondentenes svar.

Valget mellom robot og fastlege er en variabel som varieres for å se effekten på den avhengige variabelen. Deltakerne ble på bakgrunn av valget sitt delt inn i grupper, henholdsvis robot eller fastlege. Det ble derfor til slutt fire grupper basert på alvorlighetsgrad og valg.

4.4.4 Medierende variabler: Tillit og oppfattet risiko

Tillit og oppfattet risiko er våre medierende variabler. For å måle tillit er spørsmålene utarbeidet med utgangspunkt i følgende dimensjoner innenfor tillit: Velvilje og teknisk kompetanse. Dette er basert på tidligere studier innen tillit i medisinsk behandling, og er derfor de relevante dimensjonene i vårt tilfelle (Newcomer, 1997; Hyman, 2001). Oppfattet risiko blir målt ved spørsmål som er utarbeidet med utgangspunkt i følgende relevante dimensjoner innenfor oppfattet risiko: Ytelsesrisiko og psykologisk risiko (Mohtar & Abbas, 2015; Featherman & Pavlou, 2003).

For å sikre en sterk intern validitet har vi konstruert spørreundersøkelsen slik at hvert spørsmål er tilegnet et konsept, som vist i tabell 1.

Tillit	1. I hvilken grad mener du at (behandler) vil deg vel, og handler i din beste interesse?
	2. I hvilken grad mener du at (behandler) har stilt korrekt diagnose, og gitt deg rett behandling?
	3. I hvilken grad mener du at (behandler) har de rette egenskapene til å gjennomføre konsultasjonen?
	4. I hvilken grad mener du at (behandler) evner å vurdere din unike helsesituasjon?
Oppfattet risiko	1. I hvilken grad anser du konsekvensene for å være alvorlige dersom du får feil diagnose og behandling?
	2. I hvilken grad opplever du indre uro (engstelse, nervøsitet, stress) i forbindelse med diagnostisering og behandling?

Tabell 1: Spørsmål for tillit og oppfattet risiko

4.4.5 Avhengig variabel: Evaluering av konsultasjon - tilfredshet og forventet anger

Evaluering er vår andre avhengige variabel. Dette blir målt gjennom dimensjonene tilfredshet og forventet anger. Vi ser på disse individuelt fremfor en snittscore, fordi det er ønskelig å se effekten på disse dimensjonene hver for seg. Spørsmålene i forbindelse med tilfredshet er basert på spørsmål fra den anerkjente European Customer Satisfaction Index (Ciavolino & Dahlgaard, 2007), som styrker studiens interne validitet tilknyttet måling av dimensjonen tilfredshet (Ciavolino & Dahlgaard 2007). Spørsmålet for forventet anger er basert på spørsmål fra tidligere studier om dette, og da særlig Brewer et al. (2016) sitt metastudie som omhandler forventet anger og helseatferd (Brewer et al., 2016). I tabell 2 vises spørsmålene som er stilt for å måle hvert konsept.

Tilfredshet	1. I hvilken grad er du fornøyd med diagnostisering og behandling gjennomført av (behandler)?
	2. I hvilken grad vil du anbefale (behandler) til andre?
Forventet anger	1. I hvilken grad angrer du på valget ditt av (behandler)?

Tabell 2: Spørsmål for tilfredshet og forventet anger

4.4.6 Kontrollvariabler

Personlig informasjon

Vi har hentet inn informasjon om deltakernes kjønn, alder og arbeidsstatus for å kunne kontrollere resultatene mot disse variablene. Vedrørende tillit til AI-teknologi, viser Accenture sin rapport at differansen mellom kjønn varierer, der menn i noen tilfeller kan vise større tillit enn kvinner (McCabe, 2017). Det er også ulikheter mellom kjønn hva gjelder risiko, der menn er mer risikovillige enn kvinner (Charness & Gneezy, 2012; Dohmen et al., 2005). Til slutt viser Kantar (2017) at yngre mennesker har mer tillit til AI-teknologi enn eldre, i motsetning til Logg et al. (2019) som viser at det ikke er noen ulikheter i alder. Siden litteraturen er tvetydig, er dette noe vi ønsker å kontrollere resultatene mot.

Risikovillighet

Det kan forekomme store ulikheter i villigheten til å påta seg risiko (Dohmen et al., 2009). Det kan derfor påvirke hvordan deltakerne oppfatter risikoen rundt symptomene sine, samt om de velger robot eller fastlege. Derfor ønsker vi å kontrollere for dette. I utgangspunktet er det uenighet i litteraturen om direkte spørsmål som stilles til respondentene om risikovillighet faktisk viser vedkommendes risikovillighet. Dette begrunnes med at faktorer som blant annet at en overvurderer seg selv ("self-serving bias") eller at strategiske motiver kan gi feilaktige resultater (Camerer & Hogarth, 1999). For å unngå slike effekter kan en gjennomføre reelle eller hypotetiske eksperimenter med økonomiske incentiver, for å måle grad av risikovillighet. Et eksempel på dette er inntektslotteri (Schubert et al., 1999; Holt & Laury, 2002; Barr & Packard, 2002; Eckel et al., 2005; Eckel & Grossman, 2007).

I en studie gjennomført av Dohmen et al. (2009) stilte de både generelle spørsmål om risikovillighet, samt at det ble gjennomført et eksperiment som et inntektslotteri for de samme deltakerne. Resultatet var at generelle spørsmål om risikovillighet i stor grad ga de samme

resultatene som et inntektslotteriet, slik at førstnevnte derfor kan benyttes til å predikere risikovillighet. Av den grunn kan det derfor indikere at undersøkelser med direkte spørsmål kan gi gode målinger på den enkeltes generelle risikopreferanser. I det økonomiske domene synes det å være en generell enighet om at ens generelle risikopreferanser påvirker risikovilligheten i de fleste situasjoner. Imidlertid er det større uenighet om dette innenfor psykologi (Slovic, 1972a og 1972b; Weber et al., 2002). Dohmen et al. (2009) har derfor både stilt spørsmål om generell risiko, men også innenfor spesifikke kontekster som karriere, helse og finansielle avgjørelser. Resultatet tilsier at kontekstspesifikke spørsmål om risiko typisk er den beste måten å måle risikovillighet på, innenfor den respektive konteksten, men at det fungerer dårligere til å måle den generelle risikovilligheten.

For å måle risikovillighet innenfor helse så forfatterne på om deltakerne røyket (Feinberg, 1977). Resultatet var at røyking er sterkt assosiert med villigheten til å påta seg risiko innenfor helse. Dermed synes det å være en god indikator på risikovillighet innenfor helse. Imidlertid er det lite egnet for andre spesifikke kontekster som finansiell risiko, eller for å måle generell risikovillighet.

Vi valgte å benytte et generelt spørsmål om risikovillighet hentet fra Dohmen et al. (2009): "I hvilken grad er du generelt villig til å påta deg risiko?" Dette er basert på deres funn om at generelle spørsmål om risikovillighet faktisk kan måle dette på en god måte. Forfatterne benyttet i utgangspunktet en skala fra 0-10. Vi har valgt å benytte en skala fra 1-7, for å opprettholde konsistens med resten av spørreundersøkelsen. I vår undersøkelse har vi også inkludert spørsmålet "Snuser du fast?" for å måle risikovillighet tilknyttet personlig helse. Studien til Dohmen et al. (2009) ble gjennomført i Tyskland i 2009. Siden det er relativt få som røyker i Norge i dag (FHI, 2018), har vi besluttet å tilpasse spørsmålet til norske forhold. Generelt sett viser forskning at snus er mindre farlig enn røyk, men vi anser det likevel som en god måte å måle risikovillighet på. Det begrunnes med at det i stor grad eksisterer negative helsemessige konsekvenser ved snus som brukerne er villig til å utsette seg for (Stranden, 2017).

Nevrotisisme

Studier viser at personer som scorer høyt på nevrotisisme er mindre tilbøyelig til å benytte ny teknologi (Sharan & Ramano, 2020). Av den grunn er det også sentralt å kontrollere for denne effekten i vårt eksperiment. Vi har derfor inkludert følgende to påstander for å måle

nevrotisisme: i) "Jeg kjenner ofte på negative følelser (engstelse, sinne, tristhet)", og ii) "Jeg føler meg hjelpesløs i stressende situasjoner". Påstandene er hentet fra boken Personality psychology (Larsen et al., 2013), samt er tilpasset til vårt eksperiment i samråd med psykolog (samtale med psykolog Mariann Pettersen, 25.03.21).

Tilbøyelighet til tillit

Tilbøyeligheten til tillit er blant annet betinget av deltakerens personlighetstrekk og tidligere opplevelser (Kramer et al.,1996; Sekhon et al., 2014; Davis & Schoorman, 1995). Av den grunn er det relevant å kontrollere resultatene opp mot deltakernes egne vurderinger av deres tilbøyelighet til tillit. For å måle dette har vi tatt utgangspunkt i studien til Gleaser et al. (2000), og i samråd med psykolog har vi tilpasset spørsmålet til vårt eksperiment (samtale med psykolog Mariann Pettersen, 25.03.21). Dette resulterte i følgende påstand: "Jeg stoler på at de fleste andre mennesker vil meg vel, også personer jeg ikke kjenner så godt". Deltakerne skulle deretter rangere hvor enige de var på en fempunkts Likert-skala.

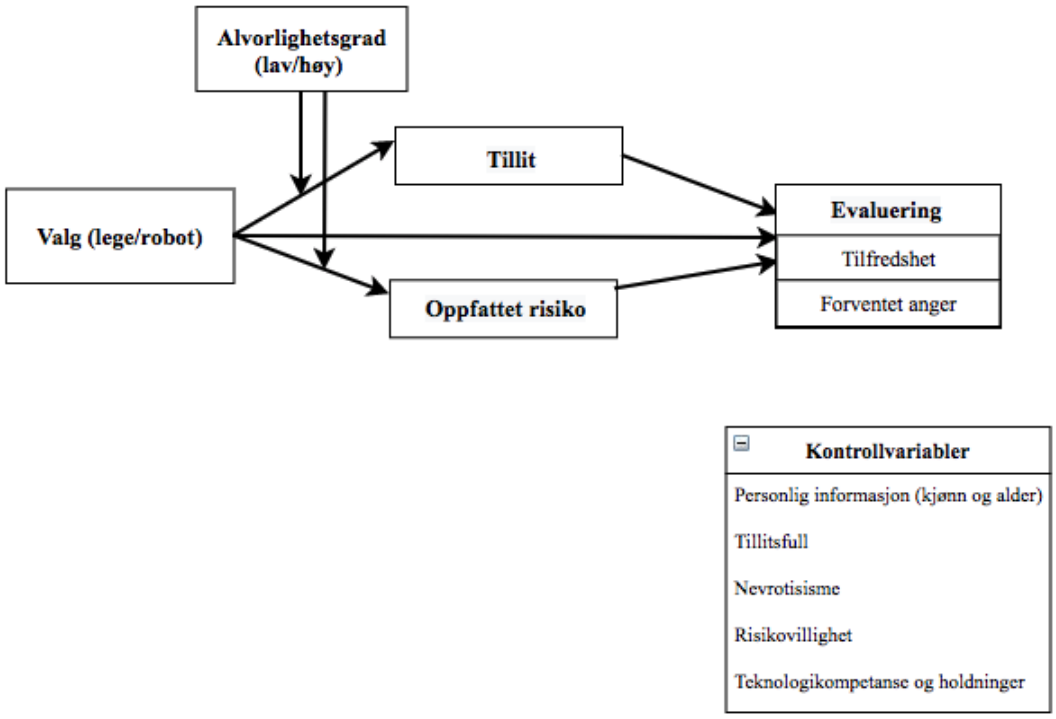
Teknologikompetanse og holdninger

Som nevnt i litteraturgjennomgangen kan personlige karakteristika, som holdninger og kompetanse om teknologi påvirke valg relatert til benyttelse av teknologiske løsninger (Gillath, 2020; Kantar, 2017). I den forbindelse er det hensiktsmessig å se om det eksisterer en sammenheng mellom behandlingsmetode og respondentens teknologikompetanse. Vi har derfor inkludert følgende to påstander i studien som måler dette blant deltakerne: i) "Jeg klarer å bruke nye teknologiske produkter og tjenester uten hjelp fra andre" og ii) "En bør være forsiktig med å erstatte mennesker som utfører viktige oppgaver med teknologi, fordi ny teknologi ikke er pålitelig". Deltakerne skulle så rangere hvor enig de var i påstandene på en fempunkts Likert-skala. Spørsmålene er hentet fra den anerkjente indeksen The Technology Readiness Index (TRI) (Parasuman & Colby, 2015), som styrker studiens reliabilitet.

I tabell 3 vises en oversikt over alle kontrollvariabler, med tilhørende spørsmål som måler konseptene. Videre, i figur 3, illustreres modellen og de forskjellige kontrollvariablene.

Risikovillighet	1. Snuser du?
	2. Til slutt, i hvilken grad er du generelt villig til å påta deg risiko?
Nevrotisisme	1. Jeg kjenner ofte på negative følelser (engstelse, sinne, tristhet)
	2. Jeg føler meg hjelpesløs i stressende situasjoner
Tillitsfull	1. Jeg stoler på at de fleste vil meg vel, også personer jeg ikke kjenner så godt
Teknologi	1. Jeg klarer å bruke nye teknologiske produkter og tjenester uten hjelp
	2. En bør være forsiktig med å erstatte mennesker som utfører viktige oppgaver med teknologi, fordi ny teknologi er ikke pålitelig

Tabell 3: Oversikt over kontrollvariabler med spørsmål



Figur 3: Forskningsmodell og kontrollvariabler

4.5.0 Dataanalyse

Dataanalyse innen kvantitativ forskning, omhandler bearbeiding av de innsamlede dataene, for deretter å benytte det til analyser og tolkning av resultater (Saunders et al., 2019). Vi vil i det følgende ta for oss hvordan vi har bearbeidet de innsamlede dataene, samt hvilke analyser vi har brukt. For dette formålet er statistikkprogrammet R benyttet.

4.5.1 Klargjøring av datasett

Datasettet ble først kontrollert for avvik. Vi hadde totalt 338 respondenter. Av disse var det 99 respondenter som ble utelatt fra analysen, fordi de kun hadde gjennomført deler av spørreundersøkelsen. Dette resulterte i totalt 239 respondenter i analysen. Etter at vi sendte ut spørreundersøkelsen, vurderte vi det ikke som nødvendig å inkludere arbeidsstatus. Arbeidsstatus ble på bakgrunn av dette ekskludert fra datasettet for videre analyser. Vi snudde også reverserte skalaer.

4.5.2 Deskriptiv statistikk

Deskriptiv statistikk betegnes som en numerisk beskrivelse av variablene. Det benyttes til å beskrive kontrollvariablene, henholdsvis respondentenes alder, kjønn, vurdering av egen risikovillighet og tillitsfullhet, teknologikompetanse, samt grad av nevrotiske karakteristika. Tilsvarende gjøres for de medierende og avhengige variablene. I henhold til Saunders et al. (2019) bør en inkludere variablenes sentraltendenser, noe vi har gjort. Det benyttes et 5% signifikansnivå i analysene, som vil si at vi aksepterer opptil fem prosent sannsynlighet for at vi forkaster en nullhypotese som egentlig er sann (Keller, 2011).

For å teste om kontrollvariablene er normalfordelt, benytter vi Shapiro-Wilk test. Testen forkaster nullhypotesen om normalfordeling når p-verdien er mindre enn eller lik 0.05. For de medierende og avhengige variablene vurderer vi om de er normalfordelt ved å undersøke deres skjevhet og spisshet. En skjevhet innenfor (-1,1) og en spisshet innen (-2,2) anses som akseptabelt. I så tilfelle anses variablene som akseptabelt normalfordelte (Saunders et al., 2019). I den sammenheng fremmes det også at en i vurderingen bør vise skjønn hva gjelder størrelsen på avviket og type variabel (Hair et al., 2010; Bryne, 2010).

For å avgjøre om det er statistisk signifikante forskjeller mellom eksperimentgruppene vedrørende de ulike kontrollvariablene, foretas en enveis ANOVA-test. Nullhypotesen (H0) om at gjennomsnittet til gruppene er like, forkastes dersom p-verdien er mindre enn eller lik 0.05. Dersom variablene ikke er normalfordelte gjennomføres det også en Kruskal Wallis H-test. Det er et ikke-parametrisk alternativ til enveis ANOVA-test, som benyttes om variablene ikke skulle være normalfordelt. Ved gjennomføring av testen forkaster vi nullhypotesen (H0), om at medianene er like, ved forklastningsgrensen på p= 0.05 (Agresti, 1990). I tilfeller der

Kruskal-Wallis testen forkastes, benyttes en Dunn`s test for å foreta en parvis sammenligning mellom gruppene. Testen forkaster nullhypotesen (H0), om likhet mellom gruppene, når p-verdien er mindre eller lik 0.05. (Dinno, 2015).

For å se på fordelingen av ordinale og nominale variabler i gruppene, ble det benyttet en kjikvadrat "goodness of fit"-test. Testen forkaster nullhypotesen (H0) om at fordelingen mellom kategoriene er lik ved forkasningsgrensen på p= 0.05. I den sammenheng må følgende tre forutsetninger være oppfylt: Det må være kategoriske variabler, observasjonene må være uavhengige, samt at den forventede frekvensen i hver gruppe må være minst 5 (Saunders et al., 2019; Keller, 2011).

4.5.3 Faktoranalyse

Faktoranalyse er en statistisk analysemetode (Harman, 1976; Rummel, 1970). En slik analyse søker å finne mønsteret av korrelasjoner blant de observerte variablene. Dersom det er høy korrelasjon mellom observerte variabler er det sannsynlig at de påvirkes av samme faktor. Hver faktor består derfor av observerte variabler som korrelerer i høy grad. Faktor kan derfor anses som underliggende dimensjoner som beskriver resultatene fra datamaterialet. En faktoranalyse benyttes til å redusere kompleksitet, i tillegg til å finne meningsfulle og forenklede mønstre (DeCoster, 1998; Kline, 1994). Eksplorerende faktoranalyse søker å utforske datamaterialet for å avdekke de underliggende dimensjonene, og er den mest brukte analysen (Kline, 1994).

Eksplorerende faktoranalyser benyttes blant annet i forbindelse med spørreundersøkelser. Dette for å undersøke om spørsmål eller påstander kan forklares ut fra underliggende dimensjoner. Analysen vil da vise hvilke spørsmål i en spørreundersøkelse som "henger sammen", og dermed i hvilken grad de ulike spørsmålene hører til de respektive faktorene (Svartdal, 2020b; Kim & Mueller, 1978b). Analysen er derfor hensiktsmessig å utføre i forbindelse med vår spørreundersøkelse. Den er utformet ved ulike spørsmål med utgangspunkt i anerkjente indekser og fra eksisterende teori, som er tilpasset vårt formål. Vi har to til fire spørsmål innenfor temaene tillit, oppfattet risiko, tilfredshet, nevrotisisme og teknologikompetanse. Det gjennomføres en eksplorerende faktoranalyse for å undersøke om spørsmålene innenfor de ulike konseptene korrelerer i høy grad. Slik påser vi at vi måler det vi ønsker å måle.

For å sjekke at utvalget er egnet for en faktoranalyse gjennomføres det en Kaiser-Meyer-Olkin (KMO) test og Bartlett's test of sphericity. KMO-testen måler andelen variasjon som kan være felles variasjon, og gir en indikasjon på om korrelasjonsmønsteret til de respektive spørsmålene muliggjør faktoranalyse (Mehmetoglu & Mittner, 2020). KMO-testen gir en verdi mellom 0 og 1. Jo større verdien er, desto mer egnet er utvalget for faktoranalyse. En verdi på 0.6 benyttes ofte som en nedre grense (Kaiser, 1974; Dzubian & Shirkey, 1974). Bartletts test sjekker om variablene er urelaterte, og dersom det er tilfellet er de uegnet for faktoranalyse. Signifikante resultater indikerer derfor at utvalget er egnet (Mehmetoglu & Mittner, 2020).

I forbindelse med å finne riktig antall faktorer benyttes Kaiser-kriteriet, som tilsier at faktorer som har en egenverdi over 1 skal beholdes. I tillegg benyttes Scree-testen, som sier at alle egenverdier skal plottes inn i fallende rekkefølge, der alle faktorer skal inkluderes frem til det fremkommer et stort fall i egenverdiene. Rotasjon av faktorene benyttes for å forbedre faktorløsningen og finne en endelig løsning. Oblique rotasjonsmetoder benyttes for faktorer som korrelerer med hverandre (Kim & Mueller, 1978a). Vi benytter derfor oblimin, fordi det gir mening at våre faktorer kan korrelere. Videre er faktorladning korrelasjonen av en observert variabel til en faktor. Det er vanlig å anse faktorladningene som høye dersom verdiene overstiger 0.6, og moderate hvis de er over 0.3 (Kline, 1994). Vi har derfor en nedre verdigrense på 0.3. Faktorladningene avgjør hvilke observerte variabler som hører til hvilke faktorer, samt hvilke observerte variabler som eventuelt tas ut av datasettet (Kline, 1994).

4.5.4 Korrelasjonsanalyse

Korrelasjonsanalyse benyttes for å finne relasjonen mellom to eller flere variabler. Analysen er basert på antakelsen om et lineært forhold mellom variablene. Relasjonen fremkommer som en korrelasjonskoeffisient med en verdi som strekker seg fra -1 til 1. En korrelasjonskoeffisient på 1 indikerer at de to variablene samvarierer perfekt positivt, og -1 indikerer at de to variablene samvarierer perfekt negativt. Dersom verdien er 0 er det ikke et lineært forhold mellom variablene (Gogtay & Thatte, 2017). Videre er en korrelasjonsmatrise et sett av korrelasjonskoeffisienter mellom et sett med variabler (Kline, 1994). Vi benytter en korrelasjonsmatrise for å undersøke forholdet mellom overraskende eller problematiske korrelasjoner før hypotesetestingen.

4.5.5 Statistiske tester

I det følgende presenteres de statistiske testene som er benyttet for å teste hypotesene våre. Det benyttes også her et 5% signifikansnivå i analysene, hvilket betyr at vi aksepterer opptil 5% sannsynlighet for at vi forkaster en nullhypotese som egentlig er sann (Keller, 2011). Det er sentralt å påpeke at om en ikke kan forkaste en nullhypotese, tilsier ikke det at en kan bevise nullhypotesen (Keller, 2011). Testene er vedlagt i appendix.

Kjikvadrat-test

Vi har benyttet en kjikvadrat-test for å se på om det foreligger en forskjell i valg av behandlingsform, samt for å undersøke om alvorlighetsgrad påvirker valget. Testen forkaster nullhypotesen (H0) om at det ikke foreligger en forskjell, ved forkasningsgrensen på p= 0.050. I den sammenheng må følgende tre forutsetninger være oppfylt: Det må være kategoriske variabler, observasjonene må være uavhengige, samt at den forventede frekvensen i hver gruppe må være minst 5 (Saunders et al., 2019; Keller, 2011).

Enveis variansanalyse

Vi har benyttet enveis ANOVA-test for å teste om det eksisterer forskjeller mellom gruppenes gjennomsnitt. I den sammenheng forutsettes det at variablene er normalfordelt og observasjonene er uavhengige, samt at gruppene har tilnærmet lik varians. Variablenes skjevhet og spisshet ble testet for se om de er innenfor akseptable avvik fra normalitet. Imidlertid synes det i litteraturen å være enighet om at ANOVA-analysen er robust mot avvik fra normalitet, om en har store utvalg (Saunders et al., 2019; Keller, 2011; Pallant, 2010). I vårt tilfelle vil avvik fra normalitet påpekes, men vi vurderer det som mindre problematisk på bakgrunn av utvalgets størrelse (n=239). Videre, for å teste for lik varians mellom gruppene, ble det gjennomført Levenes tester. Ved gjennomføring av denne testen, beholdes nullhypotesen om lik varians ved en p-verdi over 0.05 (Keller, 2011).

I tilfeller der forutsetningen om homoskedastisitet ikke ble oppfylt, er det også benyttet en ANOVA-test for ulik varians, også kalt Welch-test. For både ANOVA- og Welch-testen tilsier nullhypotesen (H0) at gjennomsnittet mellom gruppene er likt, mens alternativhypotesen (HA) viser til at gjennomsnittet til minst èn gruppe er forskjellig fra de andre. Hypotesene er gjensidig utelukkende.

$H0$: μHA robot = μHA fastlege = μLA robot = μLA fastlege

HA: minst én er forskjellig

Nullhypotesen (H0) og alternativhypotesen (HA) for ANOVA og Welch-testen illustreres ovenfor. Ved signifikante resultater viser ikke testen hvilke grupper det gjelder. Av den grunn ble det gjennomført planned contrast-tester. Dette for å se på om de signifikante resultatene fra analysene indikerer forskjeller som samsvarer med hypotesene våre. For signifikante resultater ble til slutt også gjennomført en ANCOVA-test. Dette for å foreta en robusthetssjekk av kovariatene opp mot resultatene (Saunders et al., 2019; Keller, 2019).

Paret t-test

Vi har benyttet en paret t-test for å teste for forskjeller i observasjoner som hører sammen i utvalget. For denne testen forutsettes det også normalfordeling. Derfor ble det gjennomført en Shapiro-Wilk test. Imidlertid er det viktig å påpeke at noe avvik fra forutsetningene også her kan aksepteres, som følge av utvalgets størrelse.

I vår oppgave er en paret t-test relevant å benytte når vi ønsker å se på om det er en forskjell i svaret til respondentene før og etter de får introdusert et tredje behandlingsalternativ. Dette er tilknyttet hypotesetestingen som gjøres i forbindelse med valg av behandlingsform. Siden vi ikke med sikkerhet kan kjenne til deltakernes preferanser hva gjelder behandlingsmetode, hverken før eller etter de får et nytt valgalternativ, er det passende å gjennomføre en to-halet t-test (Keller, 2011). Testens nullhypotese (H0) tilsier ingen differanse (m = mean difference) mellom observasjonene, mens alternativhypotesen (HA) viser til at det er en forskjell mellom observasjonene. Hvis t-verdien er over kritisk grense, og p-verdien er signifikant, forkastes nullhypotesen.

H0: m = 0

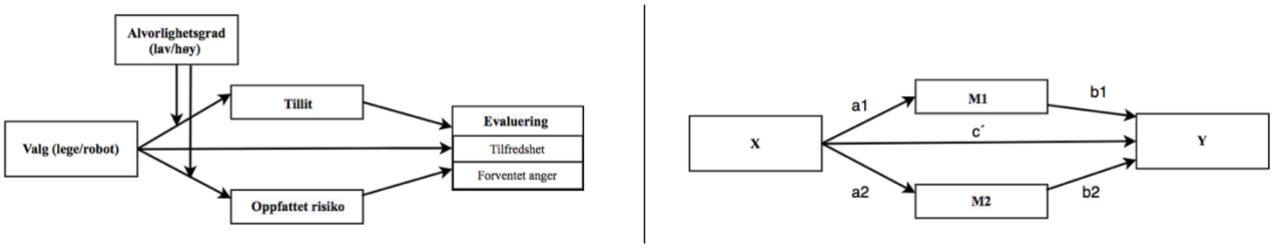
HA: m $\neq$ 0

Nullhypotesen (H0) og alternativhypotesen (HA) illustreres ovenfor. Om testen indikerer at det er en signifikant forskjell mellom observasjonene, vil vi se på hvilken variabel som har det høyeste snittet, for å kunne avgjøre hvilken vei ulikheten går. Vi vil også sammenligne andelene av de valgte behandlingsmetodene - både før og etter det nye alternativet ble innført.

Medieringsanalyse

Hvis en medierende effekt er til stede betyr at det eksister en tredje variabel: M. Her er den uavhengige variabelen (X) årsak til mediatoren (M), som igjen forårsaker den avhengige variabelen (Y). En mediator er altså en variabel som påvirker relasjonen til X og Y (MacKinnon et al., 2007). I vårt tilfelle har vi to mediatorer, henholdsvis tillit (M1) og oppfattet risiko (M2). Vi har også to uavhengige variabler, valg av behandlingsform og alvorlighetsgrad av symptomer. Vi undersøker først kun valg av behandlingsform, før også alvorlighetsgrad inkluderes. Sistnevnte resulterer i totalt følgende fire grupper: De med høy alvorlighetsgrad som har valgt fastlege (HA fastlege), de som har lav alvorlighetsgrad som har valgt fastlege (LA fastlege), samt de som har høy alvorlighetsgrad som har valgt robot (HA robot) og de som har lav alvorlighetsgrad som har valgt robot (LA robot). De uavhengige variablene vises, for enkelhets skyld, sammen som X i modellen under. Denne vises i figur 4. Videre er det to avhengige variabler, henholdsvis tilfredshet og forventet anger, som er to dimensjoner innenfor evalueringen. Disse vises også for enkelhets skyld sammen som Y i modellen. Imidlertid undersøkes de hver for seg med medieringsanalyse, fordi vi ønsker å se på effektene hver for seg

Vi benytter medieringsanalyse for å undersøke om de uavhengige variablene (X) påvirker de avhengige variablene (Y), gjennom variablene, tillit (M1) og oppfattet risiko (M2). Vi antar at det finnes flere mekanismer enn tillit og oppfattet risiko. Av den grunn vil det også eksistere en direkte effekt (c') mellom de uavhengige variablene og den avhengige variabelen, slik at vi har en delvis medierende effekt. Dersom c' derimot er null, forklares Y kun av mediatorene, og det eksisterer en fullstendig medierende effekt (Hayes, 2018). Se figur under for en oversikt over hvordan modellen blir for medieringsanalyse.



Figur 4: Generell modell for medieringsanalyse

Vi har benyttet Hayes (2021) sin PROCESS-makro versjon 3.5 i R. Modell 4 brukes for å gjennomføre medieringsanalyse for flere mediatorer enn én. Her får en både den totale effekten, direkte effekten, og indirekte effekten X har på Y. Den totale effekten er den X har på Y uten å inkludere mediatorene (c). Den direkte effekten (c') er den effekten X har på Y som ikke kan tilskrives mediatorene, henholdsvis M1 og M2. Videre har X en effekt på M1 (a1). I tillegg har M1 en effekt på Y (b1). Både a1 og b1 må være signifikante for at det eksisterer en medierende effekt med tillit. Det finnes da en indirekte effekt mellom X og Y. Hvis det er tilfellet vil den direkte effekten (c') være mindre enn den totale effekten som X har på Y (c). Det begrunnes med at noe av effekten tilskrives M1 når den inkluderes (Hayes, 2018). Forklaringen over gjelder også for oppfattet risiko, M2.

Vi har som nevnt fire grupper, og har derfor en kategorisk uavhengig variabel. Av den grunn er det ikke hensiktsmessig å se på hvordan én enhets økning i X påvirker Y, M1 og M2. Av den grunn utføres medieringsanalyse med multikategorisk X (Hayes & Preacher, 2014). Det vil si at effekten av å endre fra én gruppe til en annen undersøkes istedenfor. Dette settes opp slik at alle grupper settes opp mot hverandre (Hayes, 2016). Kategoriene må da kodes som dummyvariabler. For å kunne undersøke forskjellen mellom alle grupper er vi nødt til å gjennomføre testen tre ganger, med ulik koding av dummyvariablene. Dette svekker testen og kan øke sannsynligheten for type 1 feil (Hayes, 2018). Dette er derfor noe det må tas hensyn til i analysen.

For å dobbeltsjekke PROCESS-resultatene benyttes "bootstrappede"-konfidensintervall. Dersom vi får et konfidensintervall som ikke inneholder null er resultatet signifikant, siden det tilsier at effekten ikke kan være null. Vi genererte "bootstrappede" konfidensintervaller med 10 000 bootstrap-prøver. Dette ble gjort for alle analysene.

4.6.0 Evaluering av metode

I det videre vil vi gjøre en evaluering av metoden som benyttes i oppgaven, for å kunne vurdere studiens kvalitet. Til dette formål vil vi ta utgangspunkt i etablerte mål for validitet og reliabilitet.

4.6.1 Reliabilitet

Reliabilitet innebærer hvor pålitelig og konsistent studien er. Det vil si om en kan oppnå samme funn under andre omstendigheter eller med andre utvalg dersom studien replikeres. Det er derfor særdeles viktig å være grundig og kritisk, samt beskrive prosessen nøye der en begrunner valg og antakelser som tas. Dette betyr at en må være rigid når det gjelder forskningsmetoden, for å unngå trusler mot reliabilitet, som igjen påvirker funn og konklusjoner (Svartdal, 2020a).

Å legge til rette for høy grad av reliabilitet er ikke nødvendigvis enkelt. Det finnes en rekke trusler, i form av ulike feil og bias, som kan true reliabiliteten (Staff, 2015). I vårt studie er blant annet "feil" hos deltakerne en risiko. Spørreundersøkelsen ble gjennomført på nett, som betyr at deltakernes omstendigheter varierer - både hva gjelder tid på dagen og sted. Vi kan derfor ikke utelukke at respondentene har hatt ulik grad av forstyrrende elementer i omgivelsene når undersøkelsen ble gjennomført, eller at deltakerne kan ha vært sultne, trøtte, slitne eller lignende. Vi har ikke hatt mulighet til å observere deltakerne da de tok undersøkelsen, dermed er det vanskelig å kontrollere for deltakerfeil som kan påvirke respondentenes svar (Saunders et al., 2019). Dette kan påvirke reliabiliteten. Imidlertid ser vi at de fleste deltakerne fullførte på omtrent samme tid, hovedsakelig fem til syv minutter. Gjennomsnittlig tidsbruk er raskere enn ti minutter som vi forventet basert på pre-testen, men det viktigste er at deltakerne brukte noenlunde lik tid. Dette kan indikere at det ikke har forekommet vesentlige feil hos deltakerne.

Deltakerbias kan oppstå ved at deltakerne for eksempel er redd for at svarene skal kunne spores tilbake til seg selv, eller at de forsøker å svare det de tror forskerne er ute etter (Saunders et al., 2019). Derfor har vi unngått å etterspørre sensitiv informasjon, samt gjennomført undersøkelsen elektronisk. I tillegg har vi anonymisert besvarelsene, noe det også er informert om i begynnelsen av undersøkelsen.

Det kan også oppstå ulike "feil" og bias hos forskerne. Det kan forekomme feil i forbindelse med utformingen av undersøkelsen, der spørsmålene blant annet kan bli stilt feil, eller være misvisende (Saunders et al., 2019). Derfor har vi i stor grad basert spørsmålene på anerkjente undersøkelser. Vi har også tilpasset spørsmålene til studien ved behov, med utgangspunkt i teori, samt i samråd med fagpersoner. I tillegg gjennomførte vi en pre-test med 12 personer for å kunne få tilbakemeldinger, og rette på eventuelle feil og uklarheter. Forskerbias kan

blant annet oppstå ved at vi lar våre subjektive meninger og forutinntatthet påvirke analysen og tolkninger av resultatene (Saunders et al., 2019). Vi legger til rette for å unngå dette ved at vi begge gjennomførte analysen av dataene, samt at vi har dokumentert prosessen godt underveis og begrunnet alle valg. Videre er det viktig å sørge for å unngå "datafisking". Det vil si å tilpasse analysene til å få signifikante resultater på det vi ønsker å finne (Kvittingen & Torgersen, 2019). Vi har derfor planlagt i forkant hvilke analyser som skal gjennomføres, og holdt oss til disse uavhengig av hvilke resultater vi har fått.

4.6.2 Validitet

Intern validitet

Intern validitet knyttes til i hvilken utstrekning funnene i en studie kan tilskrives manipulasjonen som finner sted, fremfor eventuelle feil i forskningsdesignet (Pripp, 2018). Et laboratorieeksperiment, sammenlignet med et felteksperiment, foregår i kunstige omgivelser. Det er derfor lettere å utelukke andre variabler som kan påvirke resultatet. For å oppnå høy grad av validitet må en derfor kunne kontrollere variablene. På den måten kan en utelukke feilkilder, og kontrollere for eventuelle spuriøse effekter, som kan påvirke eksperimentets resultater (Keller, 2011). For å sikre høy grad av intern validitet har vi søkt å holde alle forhold konstante, unntatt de uavhengige variablene, som i vårt tilfelle er alvorlighetsgraden på de oppgitte symptomene, samt valg av behandlingsform. Vi vil i det følgende fremlegge potensielle trusler for den interne validiteten.

Respondentene ble tilfeldig fordelt inn i to eksperimentgrupper, henholdsvis høy eller lav alvorlighetsgrad av symptomer, samt at de videre valgte behandlingsform selv. Valg av behandlingsform er derfor en avhengig variabel for H1 og H1a, og videre en uavhengig variabel for de resterende hypotesene. Dette resulterte til slutt i totalt fire grupper. Tilfeldig inndeling av alvorlighetsgrad av symptomer, kan redusere eventuelle spuriøse effekter tilknyttet personlige karakteristika ved gruppenes deltakere (Saunders et al., 2019). Vi kunne også foretatt tilfeldig inndeling av behandlingsform. Imidlertid er valg en måte å måle indikasjon på algoritmeaversjon, og derfor var det ønskelig å ha den som en avhengig variabel først. Vi vurderte det som viktigere at deltakerne selv valgte mellom robot og fastlege, fremfor å redusere eventuelle spuriøse effekter.

Det kan også tenkes at noen av respondentene frykter at svarene de oppgir kan ha en reell påvirkning for dem i ettertid (Saunders et al., 2019). I vårt tilfelle kan respondentene blant annet frykte at funnene kan stimulere til økt bruk av AI-teknologi i helsevesenet. Hvis en misliker dette kan det tenkes at noen kan stille seg særlig negativ i undersøkelsen, som kan påvirke våre resultater. Imidlertid vurderer vi dette som lite sannsynlig, siden det opplyses om at studien er en masteroppgave, samt at vi ikke har en reell makt til å gjennomføre slike tiltak i helsevesenet.

Definisjonsvaliditet

Defininisjonsvaliditet omfatter validtiteten tilknyttet måleindikatorene, og ser på om de valgte indikatorene er passende, og faktisk måler det en ønsker å måle (Saunders et al., 2019). I vårt studie har vi følgende konsepter, som kan gi indikasjoner på algoritmeaversjon: Tillit, oppfattet risiko, samt evaluering. For sistnevnte ser vi på dimensjonene tilfredshet og forventet anger hver for seg. Det er benyttet spørsmål som dekker ulike områder innenfor de ulike konseptene. På den måten dekkes konseptet på en tilstrekkelig måte for vårt formål. Vi ønsker å måle hvorvidt det er ulik grad av algoritmeaversjon, formulert med syv hypoteser. Indikasjon på algoritmeaversjon måles ved å se på differansen i respondentenes svar i de forskjellige gruppene, målt ved de ovennevnte indikatorene. En effekt som kan påvirke indikatorene er den reelle opplevelsen av de oppgitte symptomene. Siden studien er fiktivt, er det en sannsynlighet for at deltakerne synes det er vanskelig å sette seg inn i sykdomsbildet, som ubevisst kan påvirke svarene deres. Eksperiementets design og resultat gir ingen indikasjon på om dette kan stemme. Dermed vurderes dette som en svakhet ved definisjonsvaliditeten.

Innholdsvaliditet søker å se på om spørsmålene i spørreundersøkelsen gir en tilstrekkelig dekning av de undersøkende konseptene (Saunders et al., 2019). I vårt tilfelle har vi gjennomgått litteraturen grundig, samt benyttet spørsmål fra anerkjente spørreundersøkelser, for å påse at måleindikatorene faktisk måler de respektive konseptene vi søker å studere. For å unngå "careless responding" er det generelt et begrenset antall spørsmål for hvert enkelt konsept. Dette påvirker innholdsvalidtiteten til de respektive konseptene.

Tillit og oppfattet risiko er store temaer i litteraturen, og det er derfor ulike måter å måle disse konseptene på. Vi har måttet begrense oss til spørsmål som er hentet fra to ulike dimensjoner for hvert tema, siden disse ble ansett som de mest relevante. Det kan svekke

innholdsvaliditeten på konseptene. For personlighetstrekkene nevrotisisme og tillisfullhet, har vi i samråd med psykolog kommet frem til spørsmål som i størst grad vil kunne dekke konseptene. Imidlertid, siden vi kun har stilt henholdsvis to og ett spørsmål om disse personlighetstrekkene i spørreundersøkelsen, vurderes det ikke som tilstrekkelig til å kunne fastslå deltakernes personlighetstrekk. Dette svekker derfor innholdsvaliditeten. Derimot vil det kunne gi en indikasjon på hvilke personlighetstrekk respondentene har på disse områdene.

Konstruktvaliditet referer til hvilken grad et sett med spørsmål faktisk måler tilstedeværelsen av innholdet som en har intensjon å måle (Saunders et al., 2019). I vårt tilfelle har vi gjennomført en standardisert spørreundersøkelse. I den forbindelse er det kritisk at eksperimentet er anonymt, slik at en unngår at deltakerne endrer adferd på bakgrunn av at svarene deres blir observert. Deltakerne ble derfor informert om at besvarelsene var anonyme i instruksen, for å legge til rette for at vi får ærlige svar.

En annen svakhet med eksperimenter er at det er en risiko for at deltakerne misforstår oppgaven. I så tilfelle vil ikke dataene være valide. Derfor er det særlig viktig at både respondentene og vi har en felles forståelse om spørsmålene som stilles. Sentinel misforståelse defineres som "misforståelser blant respondentene tilknyttet forståelsen av spørsmålene som stilles, og at dette avviker fra forskernes oppfatning" (Saunders et al., 2019). For å minimere sentinel misforståelse, ble det utarbeidet tydelige retningslinjer som ble gitt i begynnelsen av undersøkelsen. Det ble også gjennomført en pre-test av spørreundersøkelsen, som vi sendte ut til 12 personer for å få tilbakemeldinger. Av disse gikk vi gjennom spørreundersøkelsen nøye med tre personer, i forskjellig aldersgruppe, for å forsikre oss om at vi hadde en felles oppfatning av instruksene og spørsmålene. På bakgrunn av tilbakemeldingene foretok vi noen omformuleringer, slik at instruksjonene og spørsmålene ble tydeligere.

I alt gjennomførte ikke 99 av respondentene hele spørreundersøkelsen. En årsak kan være at de ikke forsto hva en skulle gjøre. Imidlertid besvarte 239 respondenter hele undersøkelsen, med omtrent samme tidsbruk, noe som kan peke i retning av at deltakerne forsto hvordan spørreundersøkelsen skulle gjennomføres. Andre årsaker til at de 99 respondentene ikke gjennomførte undersøkelsen kan være at en ikke hadde tid likevel, eller ønsket å ta den på et senere tidspunkt. Spørreundersøkelser på nett innebærer stor grad av fleksibilitet i

gjennomføringen av den. Det er derfor ikke uvanlig å påbegynne undersøkelsen, for deretter å fullføre den på et senere tidspunkt som passer bedre. Vi ser i retrospekt at vi kunne stilt et kontrollspørsmål om i hvilken grad instruksene var forståelige, som kunne gitt oss en bedre indikasjon på dette.

Ekstern validitet

Ekstern validitet tar for seg hvorvidt funnene fra en studie er generaliserbare (Dahlum, 2021). Det er i hovedsak om utvalget kan generaliseres til selve populasjonen, som i vårt tilfelle er hele den norske befolkning over 18 år. Ifølge Jacobsen (2015) er dette kun mulig dersom det er et sannsynlighetsutvalg. Vi har et ikke-sannsynlighetsutvalg, og vi kan derfor ikke være sikre på at det ikke finnes noen systematiske skjevheter i utvalget. For å rekruttere deltakere benyttet vi en kombinasjon av selvselektering og snøballeffekt, der det er tatt utgangspunkt i omgangskretsen til forskerne. Av den grunn kan derfor besvarelsene stamme fra en mer homogen gruppe enn hele den norske befolkning.

Studien er gjennomført som et eksperiment, og det har sine fordeler og svakheter. Det er hensiktsmessig, fordi det i større grad kan utelukke andre variabler som kan påvirke resultatet, og dermed styrke den interne validiteten. Imidlertid vil det derfor gå utover den eksterne validiteten, fordi det foregår i et kunstig miljø. Deltakerne kan påvirkes av dette ved at de oppfører seg annerledes enn de ville gjort i reelle omgivelser. Laboratorieeksperimenter er derfor kritisert for å være kunstige, noe som medfører at funnene kan være til mindre nytte utenfor laboratoriet (Levitt & List, 2009). I vårt eksperiment ble deltakerne eksponert for ikke-reelle stimuli i kliniske omgivelser. De forholdt seg til tenkte symptomer, der de skulle foreta et valg mellom en robot og fastlege. Deretter ønsket vi deltakernes hypotetiske vurderinger av dette. Derfor kan vi ikke med sikkerhet slå fast at roboten faktisk ville blitt benyttet, og at vurderingene av valgt behandlingsform faktisk ville blitt slik i en reell situasjon. Av den grunn er funnene mindre generaliserbare utenfor den eksperimentelle situasjonen.

Vi har benyttet aktøren IBM Watson, og generaliserer til hele bransjen basert på dette. En fare her er at merkevareassosiasjoner og bransjespesifikke assosiasjoner ikke nødvendigvis samsvarer fullt ut. Det kan påvirke responsen til deltakerne, som igjen kan svekke generaliserbarheten. Imidlertid er benyttelse av AI-teknologi fortsatt et relativt nytt fagfelt, særlig innenfor helsevesenet. Derfor er det naturlig å anta at det heller ikke er opparbeidet

spesielt sterke assosiasjoner til konkrete bedrifter. Vi anser derfor faren som liten for at merkevareassosiasjoner tilknyttet bruk av IBM Watson svekker generaliserbarheten. Videre kan studien vi har gjennomført være generaliserbart til lignende kontekster vedrørende bruk av AI-teknologi i helsevesenet. Imidlertid må studien da replikeres for å kunne etablere statistisk generaliserbarhet.

4.6.3 Oppsummering

Studier som har lav grad av reliabilitet, vil også ha lav validitet, siden feil og bias vil påvirke resultatene, og dermed også den videre tolkningen. Dersom studien har lav validitet vil dette påvirke troverdigheten til studiens resultater og konklusjoner. Derfor vil det også ha en uheldig påvirkning på den eksterne validiteten. Det begrunnes med at det da er høyst usannsynlig at en replikasjon av studien vil gi samme resultater og statistiske relasjoner. Tilknyttet dette er det sentralt å vurdere hvilke svakheter ved studien som kan aksepteres for at det fremdeles skal anses relevant.

Vi vurderer det som at studien innehar høy grad av reliabilitet, og derav også er pålitelig. Den interne validiteten vurderes også som høy nok, foruten at det kunne vært flere spørsmål tilknyttet konseptene for å måle dette mer nøyaktig. Imidlertid vurderes den eksterne validiteten som lav. Vi kan dermed konkludere med at vi i stor grad vet at vi måler det vi ønsker, men at vi ikke kan generalisere resultatene til hele populasjonen, eller i andre kontekster.

4.7.0 Etiske perspektiver

Vi vil i det følgende ta for oss de etiske aspektene ved studien. Etikk tar for seg normer og standarder for atferd som styrer moralske valg. Tilknyttet dette sier god forskningsetikk at en må ta hensyn til både deltakere av eksperimentet, og andre som kan påvirkes av studiens resultater (Saunders et al., 2019).

Deltakerne kan påvirkes negativt om vi ikke opprettholder en høy etisk standard. I den sammenheng er det blant annet sentralt å ha åpenhet rundt studiens hensikt. Respondentene ble derfor i forkant informert om at det var en spørreundersøkelse om ulik behandlingsmetodikk i helsevesenet. De fikk videre utdypende informasjon om studiens formål, samt manipulasjon av variabler, etter endt besvarelse. På denne måten sikres full

åpenhet, samtidig som at informasjonen som oppgis ikke blir ledende, og påvirker deltakernes besvarelser. Av tilsvarende hensikt, for å unngå deltaker bias, ble det innledningsvis også opplyst om at en deltar i en undersøkelse, og ikke i et eksperiment. Avslutningsvis ble deltakerne informert om at en kunne kontakte oss per e-post om en skulle ønske ytterligere informasjon om funn, samt behandling av personlig informasjon eller liknende. På denne måten har vi vært åpne om studiens hensikt uten at det har påvirket respondentenes besvarelser.

For å opprettholde en høy etisk standard er det også sentralt å skape forutsigbarhet, trygghet og sikkerhet for studiens respondenter (Saunders et al., 2019). For å sikre valgfrihet tilknyttet deltakelse i spørreundersøkelsen, ble deltakerne innledningsvis informert om at det var frivillig å delta, og at en når som helst kan trekke seg fra studien. For å gi deltakerne forutsigbarhet ble det også innledningsvis informert om at en ville få et case, etterfulgt av noen spørsmål som en skulle besvare etter beste evne. De ble også informert om at besvarelsene anonymiseres, og at det ikke var mulig å spore hvem som har svart hva. Videre stilte vi ingen spørsmål som kan betegnes som sensitive. Dette, kombinert med anonymisering av besvarelser, er tiltak som kan bidrar til trygghet og sikkerhet for respondenten, samtidig som det også kan bidra til å senke terskelen for å delta. Vi har på denne måten søkt å utvise passende adferd overfor deltakerne i hele prosessen.

I presentasjonen av dataene har vi søkt å være så objektive som mulig. I den sammenheng fremmer Jacobsen (2015) at en skal gjengi fullstendige resultater i riktig kontekst. Vi har fremlagt de metodiske valgene vi har gjort, samtidig som at vi også har vurdert studiens styrker og svakheter. Videre har vi også søkt å unngå "datafisking", ved på forhånd å beslutte hvilke dataanalyser som skal gjennomføres. På denne måten, unngår vi å tilpasse statistiske analyser til datagrunnlaget og hva vi ønsker å finne (Kvittingen & Torgersen, 2019). Videre har vi også lagt ved analysene i appendix, slik at resultatenes pålitelighet kan vurderes.

5 Analyse

Ved å benytte den metodiske tilnærmingen beskrevet i kapittel fire, ble data innsamlet til et datasett. Vi vil i det følgende først presentere deskriptiv statistikk, før vi gjennomfører ulike analyser. Til slutt tester vi hypotesene våre, ved å teste for forskjeller i gjennomsnitt, samt om det eksisterer medierende effekter.

5.1.0 Deskriptiv statistikk

Totalt 338 deltok i eksperimentet, der 239 ble inkludert i analysen. Vi benyttet Qualtrics sin randomiseringsfunksjon til å randomisere deltakerne inn i to ulike grupper. Deltakerne delte seg videre selv inn i grupper ved å velge behandlingsform (fastlege/robot). Av den grunn er det å forvente at det er noe skjevhet i gruppeinndelingene, men at fordelingen relatert til symptomer er forholdsvis lik. Imidlertid var det 99 personer som ikke fullførte spørreundersøkelsen. Derfor er det rimelig å anta at det kan være noe skjevhet i inndelingen av symptomer foretatt av Qualtrics.

Av analysen kan vi se følgende gruppeinndelinger: 123 personer fikk oppgitt mindre alvorlige symptomer, og 116 personer fikk oppgitt mer alvorlige symptomer, som er som forventet. Videre finner vi at det er 43 respondenter i gruppen LA robot, og 80 i gruppen LA fastlege. Videre er det henholdsvis 26 respondenter i gruppen HA robot, og 90 i gruppen HA fastlege.

5.1.1 Alder og kjønn

Siden vi rekrutterte deltakere ved en kombinasjon av selvselektering og snøballeffekt, er det å forvente lite variasjon i aldersfordelingen (Smith & Holcombe, 2008). Gjennomsnittlig alder på tvers av gruppene er 36.3. Videre er gjennomsnittlig alder for LA robot og LA fastlege - henholdsvis 37.93 og 36.62. Tilsvarende for HA robot er gjennomsnittlig alder 31.92, og 36.29 for HA fastlege. Resultatene for Shapiro-Wilk testen viser at aldersfordelingen for hver gruppe er signifikant forskjellig fra normalfordeling, med følgende verdier: LA fastlege (W= 0.844, p= 0.000), LA robot (W= 15.044, p= 0.000), LA fastlege (W= 0.820, p= 0.000) og HA fastlege (W= 0.720, p= 0,000). Et histogram av aldersfordelingen blant alle deltakerne illustreres i appendix 2.

Eksperimentets kjønnsfordeling har en overvekt av kvinner med 165 deltakere, sammenlignet med menn på 76 deltakere. Siden ingen av deltakerne definerte seg selv som “annet”, utgår denne kategorien i analysen. Fordelingen av kvinner var på henholdsvis 56 og 27 for LA fastlege og LA robot, samt 64 og 16 for HA fastlege og HA robot. Tilsvarende, for menn, er fordelingen på henholdsvis 25 og 16 for LA lege og LA robot, samt 27 og 10 for HA fastlege og HA robot. Kjikvadrat-testen $\chi 2(1)$ indikerer at det er en signifikant forskjell mellom kjønnene i to av gruppene: LA fastlege ($\chi 2$= 11.864, p= 0.000) og HA fastlege ($\chi 2$= 11.864, p= 0.000). Videre er det ikke signifikant forskjell i de andre to gruppene: LA robot ($\chi 2$=

2.814, p= 0.093) og HA robot ($\chi 2$= 1.385, p= 0.239). Fordelingen av kjønn per gruppe presenteres ved et histogram i appendix 2.

5.1.2 Tilbøyelighet til tillit

Variabelen tilbøyelighet til tillit, viser til deltakernes egen vurdering av hvor tillitsfull en er. Trekket måles ved å spørre om en tenker at andre vil en vel, også mennesker en ikke kjenner så godt. Dette måles ved en Likert-skala, der 1 er svært uenig og fem er svært enig. Analysen viser at deltakernes tilbøyelighet til tillit har et snitt på 3.83, en median på 4, samt et standardavvik på 0.77. Utvalget synes derfor å være relativt tillitsfulle. En oversikt over variabelens deskriptive statistikk kan sees i tabell 4. Shapiro-Wilk testen for på tvers av gruppene avviser normalfordeling (W = 0.816, p= 0.000).

Tillitsfull	Gjennomsnitt	Standardavvik	Median	S-W normalitetstest
LA fastlege	3,79	0,63	4	W=0.745, p = 0,000
LA robot	3,88	0,88	4	W=0.817, p = 0,000
HA fastlege	3,68	0,91	4	W=0.827, p = 0,000
HA robot	3,96	0,66	4	W=0.796, p = 0,000

Tabell 4: Selvrangering av tillit per gruppe

Siden dataen ikke er normalfordelt, er ikke forutsetningene for en enveis ANOVA-test møtt. Av den grunn blir det foretatt en Kruskal - Wallis H(3)-test. Testen viser at det ikke er signifikante forskjeller i vurderingen av hvor tillitsfulle de er, mellom gruppene ($\chi 2$ = 2.38, p= 0.401). Funnene bekreftes av en enveis ANOVA-test (F= 1.16, p= 0.326)

5.1.3 Nevrotisisme

Variabelen nevrotisisme, gir indikasjoner på respondentenes nevrotiske trekk, og omfatter «grad av negative følelser» og «følelse av hjelpeløshet». Besvarelsene ble foretatt på en 5-poengs Likert-skala, der 1 er svært uenig og 5 er svært enig. Resultatene, fordelt på grupper, vises i tabell 5 og 6. Gjennomsnittlig score på tvers av gruppene er på 2.41 (median 2) for «negative følelser», og på 2.24 (median 2) for «hjelpeløshet». Som vi kan se av tabellen avviser Shapiro-Wilk testen normalfordeling for alle grupper, for begge påstander. Av analysen kan vi se at deltakerne er relativt lite nevrotiske, og befinner seg i det nedre sjiktet av skalaen. Shapiro-Wilk test for både «negative følelser» og «hjelpeløshet» avviser også normalfordeling på tvers av gruppene (W= 0.881, p= 0.000; W= 0.853, p= 0.000).

Negative følelser	Gjennomsnitt	Standardavvik	Median	S-W normalitetstest
LA fastlege	2,46	1,14	2	W = 0.869, p = 0.000
LA robot	2,33	1,25	2	W = 0.865, p = 0.000
HA fastlege	2,47	1,10	2	W = 0.894, p = 0.000
HA robot	2,12	1,03	2	W = 0.853, p = 0.001

Tabell 5: Selvrangering negative følelser

Hjelpesløshet	Gjennomsnitt	Standardavvik	Median	S-W normalitetstest
LA fastlege	2,26	0,88	2	W = 0.853, p = 0.000
LA robot	2,19	1,10	2	W = 0.843, p = 0.000
HA fastlege	2,23	0,88	2	W = 0.829, p = 0.000
HA robot	2,38	1,02	2	W = 0.863, p = 0.003

Tabell 6: Selvrangering hjelpesløshet

En Kruskal - Wallis H(3)-test viser at det ikke er signifikante forskjeller mellom gruppene for hverken «negative følelser» ($\chi 2$= 2.73, p= 0.43) eller «hjelpeløshet» ($\chi 2$= 0.87, p= 0.830). Funnene støttes av en enveis ANOVA-test med følgende verdier: «negative følelser» (F= 0.814, p= 0.487) og «hjelpeløshet» (F= 0.26, p= 0.853).

5.1.4 Risikovillighet

For å måle deltakernes risikovillighet ble det stilt to spørsmål - et som omfatter deltakernes generelle risikovillighet, og et som søker å avdekke deres helserelaterte risikovillighet. Sistnevnte testes ved å spørre om respondentene snuser.

Generell risikovillighet

Resultatene for generell risikovillighet illustreres i tabell 7. Respondentene kunne svare på en 7-poengs Likert-skala. Av analysen kan vi se at respondentene er i overkant risikovillige, med et gjennomsnitt på tvers av gruppene er på 4.31 og en median på 4. Vi ser videre at Shapiro-Wilk testen avviser normalfordeling for alle grupper. Videre er det heller ikke normalfordeling på tvers av gruppene (W= 0.944, p= 0.000).

Generell risikovillighet	Gjennomsnitt	Standardavvik	Median	S-W normalitetstest
LA fastlege	4,17	1,22	4	W = 0.933, p = 0.000
LA robot	4,44	1,20	5	W = 0.919, p = 0.004
HA fastlege	4,18	1,42	4	W = 0.951, p = 0.001
HA robot	5,04	1,04	5	W = 0.918, p = 0.040

Tabell 7: Selvrangering generell risikovillighet

En Kruskal - Wallis H(3)-test indikerer at det er signifikante forskjeller i egen vurdering av generell risikovillighet, mellom gruppene ($\chi 2$= 11.23, p= 0.01). Resultatene støttes av en enveis ANOVA-test (F= 3.62, p= 0.014). En Dunn´s test viser at det er signifikante forskjeller i generell risikovillighet mellom HA robot og HA fastlege (Z= -3.10, p= 0.011), samt HA robot og LA fastlege (Z= 3.00, p= 0.016). Resultatene fra Dunn´s post-hoc test finnes i appendix 4. Som vi kan se av tabell 7, viser en sammenligning av scorene at HA robot har signifikant høyere oppfattet risikovillighet enn både LA fastlege og HA fastlege.

Helserelatert risiko

Vedrørende risikovillighet tilknyttet egen helse, oppga majoriteten at de ikke snuste (83%), mens henholdsvis 12% og 5% oppga at de snuste fast, eller en gang i blant. For HA fastlege viser fordelingen at 11 deltakere snuser, 75 snuser ikke, samt at fire snuser en gang iblant. For HA robot er fordelingen at seks snuser, 19 snuser ikke, mens én snuser en gang i blant. For LA fastlege er den på henholdsvis seks, 70 og fire. Til slutt, for LA robot, opplyses det at seks snuser, 45 snuser ikke, samt at to snuser kun en gang iblant. Kjikvadrat-testen $\chi 2(2)$ indikerer at det er en signifikant forskjell i alle gruppene vedrørende status på snusing: HA fastlege ($\chi 2$= 102, p= 0.000), HA robot ($\chi 2$= 19.92, p= 0.000), LA fastlege ($\chi 2$= 105.7, p= 0.000), LA robot ($\chi 2$= 45.25, p= 0.000). Fordelingen av snusing per gruppe presenteres ved et histogram i appendix 2.

5.1.5 Teknologikompetanse og holdninger

For å måle hvilke forhold deltakerne har til teknologi, ble det stilt to spørsmål - et vedrørende deres egen vurdering av teknologikompetanse, og et om deres holdninger til å erstatte menneskelige ressurser med teknologiske løsninger. Besvarelsene ble foretatt på en 5-poengs Likert-skala, der 1 var svært uenig og 5 svært enig for teknologikompetanse. Imidlertid er skalaen for teknologiske holdninger snudd. En oversikt over den deskriptive statistikken kan sees i tabell 8 og 9. Gjennomsnittlig vurdering av egen teknologikompetanse, på tvers av

gruppene, er på 4.37 (median 5). Vedrørende holdninger om å erstatte mennesker med teknologi er gjennomsnittet på tvers av gruppene 2.85 (median 3). Resultatene viser at respondentene anser seg selv som relativt teknologikompetente, og at de er relativt nøytrale i vurderingen om å erstatte mennesker med ny teknologi.

Shapiro-Wilk testene, viser at det ikke er en normalfordeling mellom gruppene - verken på spørsmålet som omfatter teknologikompetanse, eller holdninger til benyttelse av ny teknologi. Videre avviser også Shapiro-Wilk testen normalfordeling på tvers av gruppene (W= 0.734, p= 0.000; W= 0.910, p= 0.000).

Bruke ny teknologi	Gjennomsnitt	Standardavvik	Median	S-W normalitetstest
LA fastlege	4,27	0,80	4	W = 0.7832, P = 0.000
LA robot	4,72	0,50	5	W = 0.5773, P = 0.000
HA fastlege	4,23	0,84	4	W = 0.7400, P = 0.000
HA robot	4,62	0,57	5	W = 0.6607, P = 0.000

Tabell 8: Selvrangering benytte ny teknologi

Erstatte mennesker med teknologi	Gjennomsnitt	Standardavvik	Median	S-W normalitetstest
LA fastlege	2,44	0,91	2	W = 0.8780, P = 0.000
LA robot	3,53	1,08	3	W = 0.8960, P = 0.000
HA fastlege	2,59	0,95	2	W = 0.88802, P = 0.000
HA robot	3,88	0,71	4	W = 0.8124, P = 0.000

Tabell 9: Selvrangering teknologiske holdninger

En Kruskal - Wallis H(3)-test indikerer at det er signifikante forskjeller mellom gruppene i egen vurdering av teknologikompetanse ($\chi 2$= 17.18, p= 0.000), samt holdningene en har til teknologi ($\chi 2$= 56.64, p= 0.000). Funnene støttes av en enveis ANOVA-test, med verdier på henholdsvis (F=3.14, p= 0.001) for teknologikompetanse, og (F=22.44, p=0.000) for holdninger til teknologi.

For spørsmålet som omfatter teknologikompetanse viser Dunn´s test til signifikante forskjeller mellom gruppene for HA fastlege og LA robot (Z= -3.62, p= 0.001), samt LA fastlege og LA robot (Z= -3.26, p= 0.005). Som vi kan se av tabell 8 har LA robot signifikant høyere

snittscore for teknologikompetanse enn HA fastlege og LA fastlege. Resultatene fra Dunn´s test kan sees i appendix 4.

Når det gjelder påstanden om å erstatte mennesker med teknologi, viser Dunn´s testen til at det er signifikante forskjeller mellom alle gruppene (Z= -5.49, p= 0.000; Z= 0.82, p= 0.000; Z= 5.97, p= 0.000; Z= -4.55, p= 0.000; Z= 1.52, p= 0.000; Z= -5.12, p= 0.000). Resultatene fra Dunn´s test finnes i appendix 4. For å se på forskjellene mellom gruppene benyttes verdiene i tabell 9 til å se på snittscore. Resultatene indikerer at gruppene som har valgt robot er signifikant mer enig i påstanden om å erstatte mennesker med teknologi enn gruppene som har valgt fastlege.

5.1.6 Medierende og avhengige variabler

En oversikt over variablenes skjevhet, spisshet og sentraltendens vises i appendix 3.

Av analysen kan vi se at det er enkelte av de medierende og avhengige variablene for robot som ikke er innenfor de aksepterte grensene. For de medierende variablene er det særlig spørsmålene som omfatter om deltakerne har fått korrekt diagnose, og om roboten evner å vurdere ens unike helsesituasjon. Disse har høye skjevhetsverdier, på henholdsvis -1.48 og -1.46. Videre har spørsmålet om roboten har gitt korrekt diagnose også en relativt høy spisshetsverdi på 2.64.

Analysen viser også at enkelte av de avhengige variablene ikke er innenfor akseptable avvik fra normalitet. Dette gjelder spørsmålet tilknyttet robot om en vil anbefale behandlingsformen. Skjevhetsverdiene er her i overkant høye, på -1.27 og -1.12. Videre har også spørsmålet som omfatter hvor fornøyde deltakerne er med robotens behandling en spisshetsverdi på 2.06.

Til tross for at enkelte variabler ikke er innenfor de gitte akseptable grensene, vurderer vi det ikke som problematisk for analysen. Dette blant annet på grunn av størrelsen på utvalget og at avvikene ikke er spesielt store (Blanca et al., 2017; Pallant, 2010; Saunders et al., 2019).

5.2.0 Faktoranalyse

En faktoranalyse er nødvendig for å undersøke om konseptene som har flere enn ett spørsmål faktisk "henger sammen". Det vil si at spørsmålene innenfor samme konsept lader på én faktor (Svartdal, 2020a; Kim & Mueller, 1978b). Konseptene er henholdsvis tillit, oppfattet risiko, tilfredshet, nevrotisisme og teknologikompetanse.

5.2.1 Egnethet

For å undersøke om utvalget er egnet for faktoranalyse benytter vi Kaiser-Meyer-Olkin (KMO)-testen og Bartlett's test of sphericity. KMO-verdien var på 0.81, og viser derfor at utvalget er godt egnet til å kunne gjennomføre en faktoranalyse. Fra Bartletts test er resultatet signifikant ($\chi 2$= 1066,52, p= 0,000). Dermed er variablene ikke urelaterte, og utvalget er dermed egnet for faktoranalyse. Resultatet kan sees i tabell 10 under.

KMO	Bartlett
0,81	$\chi 2$= 1066,52, p= 0,000

Tabell 10: Utvalgets egnethet for faktoranalyse

Det ble derfor gjennomført en eksplorerende faktoranalyse med "oblimin" som rotasjonsmetode.

5.2.2 Initiell faktorløsning

Egenverdikriteriet på 1 resulterte i fire faktorer. Etter dette var det et stort fall i verdiene, slik at det ikke ble inkludert noen faktorer under 1. Faktorene forklarer kumulativt 54% av variasjonen. Se oversikt i tabell 11 under:

Komponenter	Egenverdi	% varians	Kumulativ %
1	2,21	18	18
2	1,31	11	29
3	1,19	10	39
4	1,91	16	54

Tabell 11: Initiell oversikt faktorer

Faktoranalysen ga derfor følgende mønstermatrise i tabell 12:

	Komponenter			
	1	2	3	4
I hvilken grad er du fornøyd med diagnostisering og behandling gjennomført av (behandler)?				0,84
I hvilken grad vil du anbefale (behandler) til andre?				0,81
I hvilken grad mener du at (behandler) vil deg vel, og handler i din beste interesse?	0,58			
I hvilken grad mener du at (behandler) har stilt korrekt diagnose, og gitt deg rett behandling?	0,38			0,50
I hvilken grad mener du at (behandler) har de rette egenskapene til å gjennomføre konsultasjonen?	0,86			
I hvilken grad mener du at (behandler) evner å vurdere din unike helsesituasjon?	0,90			
I hvilken grad anser du konsekvensene for å være alvorlige dersom du får feil diagnose og behandling?			0,36	
I hvilken grad opplever du indre uro (engstelse, nervøsitet, stress) i forbindelse med diagnostisering og behandling?			1,00	
Jeg klarer å bruke nye teknologiske produkter og tjenester uten hjelp				
En bør være forsiktig med å erstatte mennesker som utfører viktige oppgaver med teknologi, fordi ny teknologi er ikke pålitelig				0,32
Jeg kjenner ofte på negative følelser (engstelse, sinne, tristhet)		0,61		
Jeg føler meg hjelpesløs i stressende situasjoner		0,93		

Tabell 12: Initiell mønstermatrise

Faktorladningene er i stor grad konsistent med vårt teoretiske grunnlag. Spørsmålene om tilfredshet lader her på faktor 4. Tillit ser ut til å manifestere seg i faktor 1. Dette indikerer at alle spørsmålene måler tillit, men at vi ikke har klart å måle to ulike dimensjoner innenfor tillit. Vi har benyttet spørsmål fra to dimensjoner for å operasjonalisere tillitsbegrepet tilpasset vårt formål. Det er imidlertid ikke problematisk at disse dimensjonene ikke kan måles hver for seg, siden vi kun benytter tillit totalt sett i oppgaven. Imidlertid krysslader spørsmålet om "diagnose og behandling" med faktor 4, tilfredshet. Spørsmålene om oppfattet risiko lader på faktor 3.

Faktorladningene indikerer videre at spørsmålene innenfor nevrotisisme måler det samme. Imidlertid har vi ikke klart å måle teknologikompetanse, der det kun eksisterer en kryssladning med faktor 4, tilfredshet.

5.2.3 Endelig faktorløsning

Vi testet ulike løsninger, der styrker og svakheter ved disse er vurdert. Vi landet på løsningen med å ta ut spørsmålet om "diagnose og behandling" innenfor tillit. Dette på bakgrunn av den problematiske kryssladningen på faktor 1, tillit, og 4, tilfredshet. Det indikerer at spørsmålet ikke er egnet til å måle tillit fremfor tilfredshet, og motsatt. Ved å ta bort ett spørsmål svekker dette robustheten til denne faktoren. Dette må det derfor tas hensyn til når resultatene tolkes.

Videre tar vi ut begge spørsmålene om teknologikompetanse siden vi ikke har klart å måle dette, der kun det ene spørsmålet lader på faktor 4, tilfredshet.

Egenverdien på faktorene 1 og 4 har falt noe, men er fortsatt godt over 1. Den kumulative variansen faktorene forklarer har økt med 13 prosentpoeng. Se i tabell 13 under for oversikt.

Komponenter	Egenverdi	% varians	Kumulativ %
1	2,15	24	24
2	1,33	15	39
3	1,19	13	52
4	1,38	15	67

Tabell 13: Endelig oversikt faktorer

Som vi kan se av tabell 14, lader spørsmålene fortsatt på de samme faktorene, med det er nå ingen problematiske kryssladninger. Vi anser dette som beste løsning totalt sett, og benytter disse tilpasningene videre i analysene.

	Komponenter			
	1	2	3	4
I hvilken grad er du fornøyd med diagnostisering og behandling gjennomført av (behandler)?				0,47
I hvilken grad vil du anbefale (behandler) til andre?				1,01
I hvilken grad mener du at (behandler) vil deg vel, og handler i din beste interesse?	0,59			
I hvilken grad mener du at (behandler) har de rette egenskapene til å gjennomføre konsultasjonen?	0,94			
I hvilken grad mener du at (behandler) evner å vurdere din unike helsesituasjon?	0,84			
I hvilken grad anser du konsekvensene for å være alvorlige dersom du får feil diagnose og behandling?			0,37	
I hvilken grad opplever du indre uro (engstelse, nervøsitet, stress) i forbindelse med diagnostisering og behandling?			1,00	
Jeg kjenner ofte på negative følelser (engstelse, sinne, tristhet)		0,56		
Jeg føler meg hjelpesløs i stressende situasjoner		1,00		

Tabell 14: Endelig mønstermatrise

5.3.0 Korrelasjonsmatrise

Korrelasjonsmatrisen er presentert i tabell 15, og viser ingen overraskende eller problematiske korrelasjoner. Her fremkommer det blant annet at tilfredshet korrelerer signifikant med tillit og oppfattet risiko. Det viser at hypotesene våre om mediering kan stemme. Det samme gjelder forventet anger, som korrelerer signifikant negativt med oppfattet risiko. Imidlertid gjelder ikke det samme for tillit. Tilfredshet korrelerer også negativt signifikant med forventet

anger. Det kan gi en indikasjon på at tilfredshet og forventet anger henger sammen, siden de begge er dimensjoner innenfor evaluering. Tillit og oppfattet risiko korrelerer signifikant negativt, noe som derfor både er i tråd med litteraturen og våre antakelser.

Videre er det signifikant korrelasjon mellom kjønn og oppfattet risiko, samt kjønn og nevrotisisme. Dette indikerer derfor at kvinner både oppfatter risikoen som høyere, samt er mer nevrotiske. Dette er derfor i tråd med eksisterende litteratur (Charness & Gneezy, 2012; Dohmen et al., 2005). I tillegg er det en signifikant korrelasjon mellom oppfattet risiko og nevrotisisme, noe som derfor kan indikere at deltakere som har sterkere nevrotiske trekk oppfatter risikoen som høyere.

Korrelasjonsmatrisen kan ses i tabell under. Signifikante koeffisienter er markert med fet skrift og stjerne. En helsides versjon er vedlagt i appendix 5.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Kjønn	1,00													
2. Alder	0,07	1,00												
3. Snus	0,05	0,02	1,00											
4. Alvorlighetsgrad_Høy	0,03	0,09	**0,19***	1,00										
5. Valg_Høy	-0,09	-0,14	-0,11	**-0,19***	1,00									
6. Tilfredshet_Snitt	0,04	-0,12	-0,01	-0,11	**0,20***	1,00								
7. Forventet_Anger	0,07	-0,02	0,08	0,10	-0,06	**-0,13***	1,00							
8. Tillit_Snitt	0,02	-0,08	0,04	-0,08	-0,04	**0,63***	-0,05	1,00						
9. Risiko_Snitt	**0,21***	-0,04	0,06	**0,28***	-0,14	**-0,25***	**0,13***	**-0,16***	1,00					
10. Nevrotisisme_Snitt	**0,29***	**-0,17***	0,04	-0,03	-0,05	-0,03	0,00	-0,07	**0,22***	1,00				
11. Tillitsfull	-0,07	-0,05	0,00	0,02	0,14	-0,06	-0,05	0,05	-0,10	**-0,15***	1,00			
12. Risikovillighet	0,03	0,00	0,11	**0,20***	0,16	0,01	-0,02	0,00	0,09	-0,04	0,01	1,00		
13. Alvorlighetsgrad_Lav	0,05	0,08	0,07	NA	NA	0,12	0,05	0,07	**0,26***	0,15	-0,02	-0,05	1,00	
14. Valg_Lav	-0,07	0,04	-0,09	NA	NA	0,01	0,07	**-0,28***	-0,03	-0,05	0,06	-0,05	-0,16	1,00

Tabell 15: Korrelasjonsmatrise

5.4.0 Hypotesetesting

5.4.1 Valg av behandlingsform

Vi vil først teste om det totalt sett er en høyere andel som velger fastlege fremfor robot. Testene er lagt ved i appendix 6.

H1: En høyere andel av deltakerne vil velge fastlege enn robot for en medisinsk konsultasjon, når sannsynligheten for korrekt diagnostisering og behandling er den samme

En kjikvadrat-test viser at det foreligger signifikante forskjeller når det gjelder valg av behandlingsform. ($\chi 2$= 42.01, p= 0.000). Prosentvis er fordelingen 71% som velger fastlege og 29% som velger robot. Dermed støttes H1.

H1a: Deltakerne med symptomer av lav alvorlighetsgrad velger oftere robot, enn deltakerne med symptomer av høy alvorlighetsgrad

En kjikvadrats-test viser at forskjellen mellom gruppene er signifikant ($\chi 2$= 4.19 p= 0.041). Prosentvis er fordelingen ved høy alvorlighetsgrad 78% som velger fastlege og 22% som velger robot. For deltakerne med lav alvorlighetsgrad er det 65% som velger fastlege og 35% som velger robot. Dermed støttes også H1a.

Tredje behandlingsalternativ

Vi ønsker også å teste om valget av behandlingsmetode signifikant endrer seg når deltakerne får introdusert et tredje alternativ. Dette alternativet omfatter at Dr. Johansen kan benytte IBM Watson til å få informasjon og veiledning, men at fastlegen selv er ansvarlig for den endelige beslutningen. Her er fastlege kodet som 0, robot som 1, og de begge sammen som 2. Det er derfor gjennomført én test for hver gruppe. Her tester vi om det er signifikant forskjell mellom første og andre valg ved å se på differansen mellom disse verdiene.

Ingen av gruppene har normalfordeling med en Shapiro-Wilk test (W= 0.626, p= 0.000; W= 0.621, p= 0.000; W= 0.636, p= 0.000; W= 0.696, p= 0.000 for henholdsvis HA fastlege, LA fastlege, HA robot og LA robot). Grunnet utvalgenes størrelse benytter vi likevel en paret t-test som er to-halet. Alle gruppene hadde signifikante resultater (t= -19.374, p = 0.000; t= -16.350, p= 0.000; t= -25, p= 0.000; t= -9.545, p= 0,000 for henholdsvis HA fastlege, LA fastlege, HA robot og LA robot). Dette betyr derfor at for alle fire gruppene er valget signifikant forskjellig når de får et tredje alternativ. Totalt sett var det 80.8% av deltakerne som valgte IBM Watson sammen med Dr. Johansen. I de ulike gruppene HA fastlege, LA fastlege, HA robot og LA robot valgte henholdsvis 80%, 76%, 96% og 81% alternativet med begge to. Oppsummering av verdier kan sees i tabell 16.

Gruppe	Condition I-J	Mean	Mean difference	t	df	Sig.
HA fastlege	Valg_Etter - Valg_Før	1,611 0,000	**1,611***	19,374	89	0,000
LA fastlege	Valg_Etter - Valg_Før	1,538 0,000	**1,538***	16,350	79	0,000
HA robot	Valg_Etter - Valg_Før	1,962 1,000	**0,962***	25	25	0,000
LA robot	Valg_Etter - Valg_Før	1,786 1,000	**0,767***	9,545	42	0,000

Tabell 16: Resultater av paret t-test

5.4.2 Grad av tillit utvist til behandlingsform

Vi vil så gjennomføre tester tilknyttet tillit, der vi først vil undersøke om det foreligger forskjeller mellom gruppene. Levene-test gir signifikante resultater (F= 3.498, p= 0.016), og indikerer at det er ulik varians på tvers av gruppene, altså at det foreligger heteroskedastisitet. Det er derfor gjennomført en enveis variansanalyse som ikke krever at variansen er lik - mer presist en Welch enveis-test. Testen viser signifikante forskjeller mellom gruppene (F= 3.431, p= 0.021). Det er derfor videre gjennomført planned contrast-tester for å se på forskjellene mellom de ulike gruppene. Dette er gjort med utgangspunkt i å anta ulik varians. Se oversikt over verdier fra planned contrast-testene i tabell 17. Alle tester er lagt ved i appendix 7.

H2: Deltakerne som har valgt robot utviser lavere grad av tillit, enn deltakerne som har valgt fastlege

Resultatet fra å sammenligne gruppenes gjennomsnittsscore som har valgt fastlege og robot er signifikant (F= 7.941, p=0,005). Tillitsscoren er for de som har valgt fastlege 5.673, og 5,272 for de som har valgt robot. Den signifikante differansen er derfor 0.401 i favør av de som har valgt fastlege. Av den grunn støttes H2.

Vi tester videre om det også foreligger forskjeller mellom grupper som har valgt robot og fastlege, der alvorlighetsgraden varieres. Vi fikk signifikante forskjeller mellom LA fastlege og LA robot (F= 12.398, p= 0.001), og den signifikante differansen viser at gjennomsnittlig tillitsscore for de som har valgt fastlege er 0.712 høyere enn de som har valgt robot. Vi fikk også signifikant resultat for forskjell mellom LA fastlege og HA robot (F= 5.841, p= 0.016),

der den signifikante differansen viser at gjennomsnittlig tillitsscore til fastlege er 0.385 høyere enn til robot. Imidlertid er det ikke signifikant forskjell mellom HA fastlege og HA robot (F= 0.019, p= 0.892), eller mellom HA fastlege og LA robot (F= 2.088, p= 0.145).

Vi undersøker videre om det er forskjell mellom gruppene med ulik alvorlighetsgrad som har valgt robot.

H2a: Deltakerne som har valgt robot utviser lavere grad av tillit når de har symptomer av høy alvorlighetsgrad, enn deltakerne med symptomer av lav alvorlighetsgrad

Når det testes for forskjeller mellom gruppene LA robot og HA robot får vi signifikante forskjeller (F= 5.841, p= 0.016). Den signifikante differansen viser at gruppen HA robot har en gjennomsnittlig tillitsscore som er 0.327 høyere enn gruppen LA robot. Dette er derfor det motsatte av hypotesen, og H2a støttes ikke.

Videre fikk vi ikke signifikante resultater når vi undersøkte det samme for gruppene med ulik alvorlighetsgrad som har valgt fastlege (F= 2.868, p= 0.092).

Contrast Test

		Contrast	Sum of squares	Df	Mean Square	F	Sig.
Tillit	Does not assume equal variance	Fastlege_Robot	9,190	1	9,194	7,941	0,005
		HA Fastlege_HA Robot	0,020	1	0,021	0,019	0,892
		HA Fastlege_LA Robot	2,420	1	2,418	2,088	0,0150
		LA Fastlege_HA Robot	6,760	1	7,763	5,841	0,016
		LA Fastlege_LA Robot	14,350	1	14,355	12,398	0,001
		HA Robot_LA Robot	3,700	1	3,702	3,198	0,075
		HA Fastlege_LA Fastlege	3,320	1	3,321	2,868	0,092

Tabell 17: Resultater av planned contrast-tester

Robusthetssjekk

Kovariaten kjønn var også signifikant relatert til tillit for de signifikante hypotesene (F= 6.43, p= 0.011). Det var også en signifikant effekt mellom gruppene, etter å ha kontrollert for kjønn (F= 4.58, p= 0.003). De andre andre kovariatene - alder, informasjon om snusing, samt egen vurdering av nevrotisisme, tillitsfullhet og risikovillighet - hadde imidlertid ingen effekt på resultatene. Resultatene kan sees i tabell 18 under.

Tilfredshet	Egenv_risiko	Kjønn	Alder	Snus	Nevrotisisme	Tillitsfull
F-verdi	0.824	1.786	1.015	5.524	0.012	0.751
P-verdi	0.364	1.182	0.314	0.019	0.912	0.387

Tabell 18: Resultater av ANCOVA-analyse for tillit

5.4.3 Grad av oppfattet risiko i forbindelse med behandlingsform

Vi vil i det følgende presentere tester tilknyttet oppfattet risiko. For å teste om det var forskjeller i oppfattet risiko mellom gruppene, ble det foretatt en klassisk enveis variansanalyse. Dette er passende, siden resultatene til Levene-testen viste lik varians mellom gruppene (F= 0.49, p= 0.686). Vi fant signifikante forskjeller mellom gruppene tilknyttet oppfattet risiko (F= 5.63, p= 0.000). Det ble derfor gjennomført planned contrast-tester for å se på forskjellen mellom de ulike gruppene. Testresultatene er oppgitt i appendix 8.

H3: Deltakerne som har valgt robot utviser høyere grad av oppfattet risiko, enn deltakerne som har valgt fastlege

Resultatet fra å sammenligne gjennomsnittet til gruppene som har valgt fastlege og robot, er ikke signifikant (F= 2.73, p=0,099). Dermed støttes ikke H3.

Vi tester videre med planned contrast-tester om det foreligger forskjeller mellom grupper som har valgt robot og fastlege, der alvorlighetsgraden varieres. Vi fikk signifikant forskjell mellom gruppene HA fastlege og HA robot (F= 8.32, p= 0.004). Imidlertid har gruppen HA fastlege en score på 4.18, som er 0.4 høyere score enn gruppen HA robot sin score på 3.78. Funnene er derfor motsatt av forventet. Det er også signifikante resultater for gruppene HA fastlege og LA robot (F= 13.87, p= 0.000). Her har HA fastlege 0.79 høyere score på oppfattet risiko enn LA robot. Det er også motsatt av forventet. Det var ikke signifikante forskjeller mellom HA robot og LA fastlege (F= 3.56, p= 0.060), og LA fastlege og LA robot (F= 0.26, p= 0.610).

Vi undersøker videre om det er forskjell mellom gruppene med ulik alvorlighetsgrad som har valgt robot.

H3a: Deltakerne som har valgt robot utviser høyere grad av oppfattet risiko når de har symptomer av høy alvorlighetsgrad, enn deltakerne med symptomer av lav alvorlighetsgrad

Vi finner ikke signifikante forskjeller mellom gruppene LA robot og HA robot for oppfattet risiko (F=2.40, p=0.122). Dermed støttes ikke hypotese H3a.

Videre fikk vi signifikant forskjell da vi undersøkte gruppene HA fastlege og LA fastlege (F= 13.00, p= 0.000), der HA fastlege hadde en 0.70 høyere score for oppfattet risiko enn LA fastlege. Resultatet er derfor som forventet.

Robusthetsjekk

Kovariaten kjønn var også sterkt signifikant tilknyttet oppfattet risiko for hypotese H3 og H3a (F=11.51, P=0.008). Det var også en sterk signifikant effekt mellom gruppene, etter å ha kontrollert for kjønn (F=5.55, P=0,000). De andre andre kovariatene - alder, informasjon om snusing, samt egen vurdering av nevrotisisme, tillitsfullhet og risikovillighet - hadde imidlertid ingen effekt på resultatene. Resultatene fra ANCOVA-analysen kan sees i tabell 19.

Oppfattet risiko	Egenv_risiko	Kjønn	Alder	Snus	Nevrotisisme	Tillitsfull
F-verdi	1,860	11,510	0,200	1,200	0,063	1.413
P-verdi	0,172	0,008	0,653	0,270	0,792	0.235

Tabell 19: Resultater ANCOVA-analyse for oppfattet risiko

5.4.4 Evaluering av behandlingsform

Tilfredshet

Det gjennomføres så tester i forbindelse med tilfredshet, der vi først vil undersøke om det foreligger forskjeller mellom gruppene. Levene-test gir ikke signifikante resultater (F= 0.759, p= 0.518), og av den grunn indikerer det at det er lik varians på tvers av gruppene, altså at det foreligger homoskedastisitet. Det er derfor gjennomført en klassisk enveis variansanalyse. Resultatet er signifikant (F-verdi= 8.725, p= 0.001), og det er derfor videre gjennomført planned contrast-tester for å se på forskjellene mellom de fire ulike gruppene. Alle testresultater er lagt ved i appendix 9.

H4: Deltakerne som har valgt robot utviser lavere grad av tilfredshet, enn deltakerne som har valgt fastlege

Ved å se på forskjell mellom gruppene som har valgt robot og fastlege får vi akkurat ikke signifikante resultater (F= 3.782, p= 0.053), og H4 støttes ikke.

Vi tester videre om det kan foreligge forskjeller mellom grupper som har valgt robot og fastlege, der alvorlighetsgraden varieres. Resultatene fra disse testene varierer. Vi får signifikant forskjell mellom gruppene HA fastlege og HA robot (F=12.398, p= 0.001). Differansen er imidlertid motsatt av forventet, der gjennomsnittscore hos de som har valgt robot er 0.573 høyere enn de som har valgt fastlege. Videre er det signifikante forskjeller mellom HA fastlege og LA robot (F= 14.730, p= 0.001), der den signifikante differansen også her er motsatt av forventet. Her er gjennomsnittsscoren hos gruppen som har valgt robot 0.713 høyere enn fastlege. Vi får ikke signifikante forskjeller mellom LA fastlege og LA robot (F= 0.468, p= 0.495), og heller ikke mellom LA fastlege og HA robot (F= 2.271, p= 0.133).

Vi undersøker videre om det er forskjell mellom gruppene med ulik alvorlighetsgrad som har valgt robot.

H4a: Deltakerne som har valgt robot utviser lavere grad av tilfredshet når de har symptomer av høy alvorlighetsgrad, enn deltakerne med symptomer av lav alvorlighetsgrad.

Når det testes for forskjell mellom gruppene HA robot og LA robot, blir resultatet ikke signifikant (F= 0.762, p= 0.384), og H4a støttes ikke.

Det er også testet for forskjeller mellom gruppene HA fastlege og LA fastlege, noe som ga signifikante resultater (F= 14.303, p= 0.000), der scoren fra de som hadde lav alvorlighetsgrad var 0.92 høyere enn for de som hadde høy alvorlighetsgrad.

Robusthetssjekk

Kovariaten snus var også signifikant relatert til tilfredshet for de gruppene vi fant signifikante forskjeller (F= 5.52, p= 0.019). Det var også en signifikant effekt mellom gruppene, etter å ha kontrollert for snus (F= 6.19, p= 0,000). De andre andre kovariatene - alder, kjønn, samt egen

vurdering av nevrotisisme og risikovillighet - hadde imidlertid ingen effekt på resultatene. Resultatene kan sees i tabell 20.

Tilfredshet	Egenv_risiko	Kjønn	Alder	Snus	Nevrotisisme	Tillitsfull
F-verdi	0.824	1.786	1.015	5.524	0.012	0.751
P-verdi	0.364	1.182	0.314	0.019	0.912	0.387

Tabell 20: Resultater ANCOVA-analyse for tilfredshet

Forventet anger

I det videre presenteres resultatene relatert til hypotesetesting i forbindelse med forventet anger. Det ble foretatt en klassisk enveis variansanalyse for å teste om det var forskjeller i forventet anger mellom gruppene. Dette ble gjort ettersom Levene-testen viste lik varians mellom gruppene (F= 0.36, p= 0.836). Vi finner ingen signifikante resultater mellom gruppene for forventet anger (F =0.39, p =0.851). Testresultatene er oppgitt i appendix 10.

H5: Deltakerne som har valg robot utviser høyere grad av forventet anger, enn deltakerne som har valgt fastlege

H5a: Deltakerne som har valgt robot utviser høyere grad av forventet anger når de har symptomer av høy alvorlighetsgrad, enn deltakerne med symptomer av lav alvorlighetsgrad

Dermed støttes hverken H5 eller H5a.

5.4.5 Medierende effekter

Videre vil vi teste om tillit og risiko kan forklare noe av forskjellen mellom gruppene når det gjelder evaluering. Vi har sett på tilfredshet og forventet anger hver for seg, fordi det er ønskelig å se på begge dimensjoner. Den uavhengige variabelen må signifikant påvirke mediatoren (a), samtidig som mediatoren må signifikant påvirke den avhengige variabelen (b). Dette vil derfor resultere i at effekten som X har på Y reduseres (c'), fordi mediatoren inkluderes. Dersom c' blir null forekommer det fullstendig mediering, det vil si at effekten X har på Y kun er grunnet mediatoren. Dersom c' reduseres eksisterer det delvis mediering.

Tilfredshet

Vi vil både teste hypotesene vedrørende tillit og oppfattet risiko som mediatorer, samt undersøke om det foreligger forskjeller mellom grupper som har valgt robot og fastlege, der alvorlighetsgraden varieres. Fullstendige tabeller over resultater er vedlagt i appendix 11.

H6a: Effekten som valg har på tilfredshet, kan delvis forklares av tillit

H6b: Effekten som valg har på tilfredshet, kan delvis forklares av oppfattet risiko

Effekt a

Når gruppene fastlege og robot sammenlignes, får vi en signifikant effekt på tillit (p= 0.006, β= -0.394). Her er betakoeffisienten negativ, noe som betyr at effekten er sterkere for gruppen som har valgt fastlege. Bootstrap-prøvene bekrefter PROCESS-resultatene (CI= -0.771, -0.103), siden det ikke er null i konfidensintervallet. Videre var det ikke en signifikant effekt på oppfattet risiko (p= 0.099), noe også Bootstrap-resultatet viste.

Når alvorlighetsgrad varieres får vi for mediatoren tillit at det kun er gruppene LA fastlege vs LA robot som har signifikant resultat på effekt a (p= 0.001). De hadde videre en betakoeffisient på -0.649. Det vil si at effekten er sterkere for gruppen LA fastlege. Videre hadde i utgangspunktet HA fastlege vs LA robot signifikant resultat på effekt a (p= 0.038, β= -0.380). Med en negativ betakoeffisient er effekten sterkest for HA fastlege. Imidlertid viser Bootstrap-resultatene et konfidensintervall på -0.013, 0.591. Dermed er ikke effekten fra PROCESS-resultatene signifikant likevel, ettersom null er med i intervallet. Utover dette var det ingen signifikante resultater for effekt a.

For mediatoren oppfattet risiko, var effekten signifikant hos både gruppene HA fastlege vs LA fastlege (p= 0000, β= -0.552), og HA fastlege vs LA robot (p= 0.001, β= -0.608). Betakoeffisientene her var begge negative, og gruppen HA fastlege har i begge tilfeller sterkest effekt på oppfattet risiko. Utover dette var det ingen signifikante effekter.

Se tabell 21 under for oversikt over alle resultater for effekt a.

	Tilfredshet				
	β	t	p	Bootstrapping 95% CI	
				Lower	Upper
Effekt a, tillit					
Fastlege vs robot	**-0.394***	-2.802	0.006	-0.771	-0.103
HA fastlege vs LA fastlege	0.269	1.784	0.076	-0.013	0.591
HA fastlege vs HA robot	-0.082	-0.376	0.707	-0.498	0.319
HA fastlege vs LA robot	**-0.380***	-2.093	0.038	-0.911	0.046
LA fastlege vs HA robot	-0.351	-1.585	0.114	-0.797	0.032
LA fastlege vs LA robot	**-0.649***	-3.501	0.001	-1.213	-0.244
HA robot vs LA robot	-0.298	-1.225	0.222	-0.905	0.227
Effekt a, oppfattet risiko					
Fastlege vs robot	-0.236	-1.659	0.099	-0.672	0.045
HA fastlege vs LA fastlege	**-0.552***	-3.706	0.000	-1.119	-0.345
HA fastlege vs HA robot	-0.311	-1.440	0.151	-0.896	0.077
HA fastlege vs LA robot	**-0.608***	-3.380	0.001	-1.289	-0.322
LA fastlege vs HA robot	0.241	1.103	0.271	-0.180	0.813
LA fastlege vs LA robot	-0.055	-0.302	0.763	-0.560	0.427
HA robot vs LA robot	-0.297	-1.232	0.219	-0.966	0.186

Tabell 21: Resultater for effekt a

Effekt b

Når gruppene som har valgt fastlege og robot sammenlignes, har tillit en signifikant effekt på tilfredshet (p= 0.000, β= 0.654). Betakoeffisienten er effekten M har på Y, og kan tolkes som at når tilliten øker med ett punkt på Likert-skalaen, øker tilfredshet med 0.654. Oppfattet risiko har også en signifikant effekt på tilfredshet (p= 0.018, β= -0.116). Her vil tilfredshet reduseres med 0.116 dersom oppfattet risiko øker med ett punkt.

Når alvorlighetsgrad varieres, er effekten som tillit har på tilfredshet signifikant (p= 0.000, β= 0.650). Betakoeffisienten viser at når tillitsscoren øker med ett punkt på Likert-skalaen, øker tilfredshet med 0,650. Vedrørende effekten oppfattet risiko har på tilfredshet, er denne ikke signifikant (p= 0.118).

Se tabell 22 under for fullstendige resultater.

	Tilfredshet				
	β	t	p	Bootstrapping 95% CI	
				Lower	Upper
Effekt b, tillit					
Fastlege vs robot, Tillit → Tilfredshet	**0.653***	13.286	0.000	0.614	0.853
Valg og alvorlighetsgrad, Tillit → Tilfredshet	**0.650***	13.373	0.000	0.612	0.851
Effekt b, oppfattet risiko					
Fastlege vs robot, Oppfattet risiko → Tilfredshet	**-0.116***	-2.381	0.018	-0.196	-0.027
Valg og alvorlighetsgrad, Oppfattet risiko → Tilfredshet	-0.077	-1.569	0.118	-0.163	0.016

Tabell 22: Resultater for effekt b

c'-effekt

Videre er effekt c´, som er effekten X har på Y som ikke kan tilskrives mediatorene, signifikant for gruppene som har valgt fastlege og robot (p= 0.000, β= 0.500). Når alvorlighetsgrad varieres er resultatene signifikante for gruppene HA fastlege vs LA fastlege (p= 0.003, β= 0.338), HA fastlege vs HA robot (p= 0.003, β= 0.488), LA fastlege vs LA robot (p= 0.002, β= 0.436), og HA fastlege vs LA robot (p= 0.000, β= 0.774). Bootstrap-resultatene viser også det samme. Se tabell 23 under for fullstendig oversikt.

	Tilfredshet				
	β	t	p	Bootstrapping 95% CI	
				Lower	Upper
Effekt c'					
Fastlege vs robot	**0.500***	4.648	0.000	0.379	0.873
HA fastlege vs LA fastlege	**0.338***	2.960	0.003	0.135	0.715
HA fastlege vs HA robot	**0.488***	3.032	0.003	0.224	1.016
HA fastlege vs LA robot	**0.774***	5.583	0.000	0.649	1.284
LA fastlege vs HA robot	0.151	0.921	0.358	-0.207	0.595
LA fastlege vs LA robot	**0.436***	3.120	0.002	0.234	0.854
HA robot vs LA robot	0.286	1.586	0.114	-0.082	0.778

Tabell 23: Resultater for effekt c´

c-effekt

Vi ser videre at effekten c, altså effekten som X har på Y uten å ta med mediatorene, er signifikante for gruppene HA fastlege vs LA fastlege (p= 0.000, β= 0.555), HA fastlege vs HA robot (p= 0.035, β= 0.459), og HA fastlege vs LA robot (p= 0.002, β= 0.574). Se tabell 24 for oversikt.

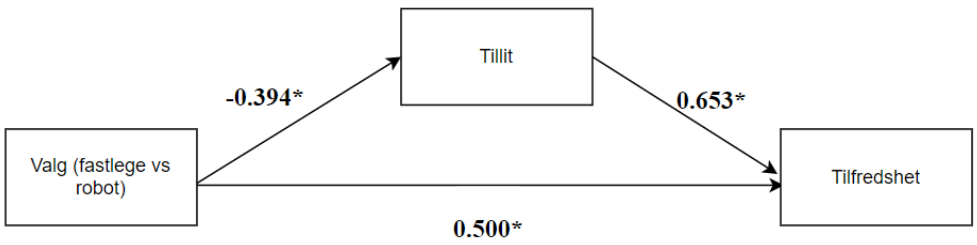
Tilfredshet

	β	t	p	Bootstrapping 95% CI	
				Lower	Upper
Effekt c					
Fastlege vs robot	0.269	1.897	0.059		
HA fastlege vs LA fastlege	**0.555***	3.722	0.000		
HA fastlege vs HA robot	**0.459***	2.125	0.035		
HA fastlege vs LA robot	**0.574***	3.190	0.002		
LA fastlege vs HA robot	-0.096	-0.438	0.662		
LA fastlege vs LA robot	0.019	0.102	0.919		
HA robot vs LA robot	0.115	0.476	0.635		

Tabell 24: Resultater for effekt c

Oppsummering

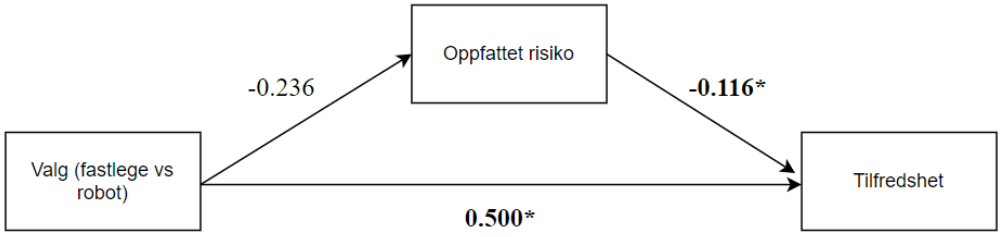
Når gruppene som har valgt fastlege og robot sammenlignes er det en signifikant effekt på tillit (a), samt at tillit har en signifikant effekt på tilfredshet (b). Effekten som valg har på tilfredshet, uten å inkludere mediatorene (c), er ikke signifikant (p= 0.059). Imidlertid er det ikke nødvendig for at det skal foreligge en medierende effekt (Shrout & Bolger, 2002). Den direkte effekten er signifikant (c'), men denne er ikke null. Det betyr at det ikke er kun tillit som forklarer effekten som valg har på tilfredshet. Av den grunn eksisterer det en delvis medierende effekt, og H6a støttes. Se figur 5 for en oversikt over effektene. Signifikante resultater er markert i fet skrift og med stjerne.



Figur 5: Oversikt over effekter for mediatoren tillit på tilfredshet

Figur 6 gir videre en oppsummering av resultatene for mediatoren oppfattet risiko, der vi ser på gruppene fastlege og robot. Oppfattet risiko har en signifikant effekt på forventet anger (b),

samt at det er en signifikant direkte effekt (c'). Imidlertid har ikke gruppene en signifikant effekt på oppfattet risiko (a). Av den grunn eksisterer det ikke en medierende effekt for oppfattet risiko, og H6b støttes ikke.



Figur 6: Oversikt over effekter for mediatoren oppfattet risiko på tilfredshet

Når alvorlighetsgrad varieres er det gjennomgående for alle grupper at tillit påvirker tilfredshet signifikant (b). Imidlertid er det kun gruppene LA fastlege vs LA robot som har en signifikant effekt på tillit (a). De har også en signifikant total effekt (c), som vil si at LA fastlege vs LA robot påvirker tilfredshet signifikant når mediatorene ikke inkluderes. I tillegg har LA fastlege vs LA robot en signifikant direkte effekt på tilfredshet, som ikke kan tilskrives mediatoren (c'). Denne effekten er ikke null, og av den grunn eksisterer det en delvis medierende effekt. Totalt sett er dette de eneste gruppene som har en medierende effekt når alvorlighetsgrad varierer. Se figur i appendix 11 for en figur med oversikt over effektene.

Oppfattet risiko (M2) påvirker ikke tilfredshet signifikant (b2), og det kan derfor heller ikke foreligge en medierende effekt.

Forventet anger

Vi vil både teste hypotesene om tillit og oppfattet risiko som mediatorer, samt undersøke om det foreligger forskjeller mellom grupper som har valgt robot og fastlege, der alvorlighetsgraden varieres. Se appendix 11 for tabeller av analyseresultater.

H7a: Effekten som valg har på forventet anger, kan delvis forklares av tillit

H7b: Effekten som valg har på forventet anger, kan delvis forklares av oppfattet risiko

Effekt a

Gruppene som har valgt fastlege og robot har en signifikant effekt på mediatoren tillit (p= 0.006, β= -0.394). Denne effekten er sterkere for fastlege ettersom betakoeffisienten er

negativ. Bootstrap-prøvene viser det samme (CI= -0.771, -0.103). Imidlertid finner vi ikke at det er en signifikant effekt for fastlege vs robot på oppfattet risiko (p= 0.099).

Når vi ser på forskjeller mellom gruppene der alvorlighetsgrad inkluderes, finner vi at gruppene HA fastlege vs LA robot har signifikant effekt på tillit (p= 0.038, β= -0.380). Imidlertid viser bootstrap-prøvene det motsatte (CI= -0.911, 0.046), og av den grunn er resultatet fra PROCESS likevel ikke signifikant. Videre har gruppene LA fastlege vs LA robot en signifikant effekt på tillit (p= 0.001, β= -0.649). Utover dette er det ingen signifikante resultater for effekten som gruppene har på tillit.

Både gruppene HA fastlege vs LA fastlege, og HA fastlege vs LA robot, påvirker oppfattet risiko signifikant (p= 0.000, β= -0.552; p= 0,001, β= -0.608). Utover dette er det ingen signifikante resultater for effekt a. Det samme viser bootstrap-prøvene.

Effekt b

For gruppene fastlege og robot har tillit en signifikant effekt på forventet anger (p= 0.000, β= -0.423). Ved å øke tillitsscoren med ett punkt på Likert-skalaen, vil det redusere forventet anger med 0.423. Videre har også oppfattet risiko en signifikant effekt på forventet anger (p= 0.001, β= 0.188). Når oppfattet risiko øker med ett punkt, øker forventet anger med 0.188. Bootstrap-prøvene bekrefter PROCESS-resultatene. Når alvorlighetsgrad varieres har tillit en signifikant effekt på forventet anger (p= 0,000, β= -0.433). Videre har oppfattet risiko en signifikant effekt på forventet anger (p= 0.001, β= 0.196).

Effekt c'

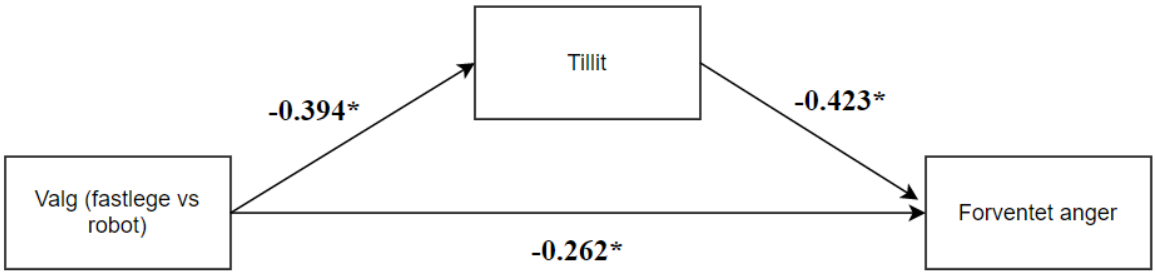
Fastlege vs robot har en signifikant direkte effekt på forventet anger (p= 0.042, β= -0.262), som ikke kan tilskrives mediatorene (c'). Bootstrap-prøvene bekrefter PROCESS-resultatet (CI= -0.625, -0.089). Når alvorlighetsgrad inkluderes har gruppene LA fastlege vs LA robot en signifikant direkte effekt på forventet anger (p= 0.014, β= -0.424), som ikke kan tilskrives mediatorene (c'). Videre viste i utgangspunktet PROCESS-resultatene ingen andre signifikante resultater. Imidlertid viste bootstrap-prøvene signifikant resultat for gruppene HA fastlege vs LA robot (CI= -0.731, -0.024), slik at de har en signifikant direkte effekt på forventet anger likevel.

Effekt c

Hverken PROCESS eller bootstrap-prøvene viste signifikante resultater på total direkte effekt som gruppene har på forventet anger. Dette gjelder både for gruppene som har valgt fastlege og robot, samt når alvorlighetsgrad inkluderes.

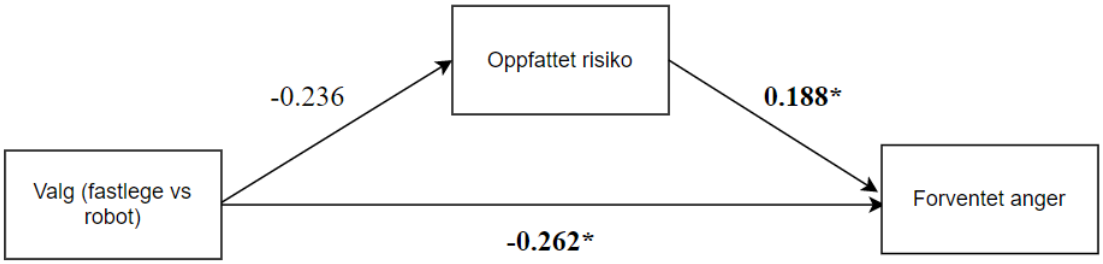
Oppsummering

Figur 7 under gir en oppsummering av resultatene for mediatoren tillit, der vi ser på gruppene fastlege og robot. Signifikante resultater er markert med fet skrift og stjerne. Gruppene har en signifikant effekt på tillit (a), samt at tillit har en signifikant effekt på forventet anger (b). I tillegg er det en signifikant direkte effekt på forventet anger (c'). Denne effekten er imidlertid ikke null, noe som betyr at det ikke er kun tillit som forklarer effekten som valg har på forventet anger. Av den grunn er det en delvis medierende effekt, og H7a støttes. Det er imidlertid ikke en signifikant total direkte effekt (c),men det er heller ikke nødvendig for at det skal foreligge en medierende effekt (Shrout og Bolger, 2002).



Figur 7: Oversikt over effekter for mediatoren tillit på forventet anger

Figur 8 gir en oppsummering av resultatene for mediatoren oppfattet risiko, der vi ser på gruppene fastlege og robot. Oppfattet risiko har en signifikant effekt på forventet anger (b), samt at det er en signifikant direkte effekt (c'). Imidlertid har ikke gruppene en signifikant effekt på oppfattet risiko (a). Av den grunn eksisterer det ikke en medierende effekt for oppfattet risiko, og H7b støttes ikke.



Figur 8: Oversikt over effekter på mediatoren oppfattet risiko på forventet anger

Når vi inkluderer alvorlighetsgrad får vi i større grad varierende resultater. Se appendix 11 for figurer for gruppene som hadde signifikante indirekte effekter.

Mediatoren tillit har gjennomgående i alle grupper en signifikant effekt på forventet anger (b). Videre har gruppene LA fastlege vs LA robot en signifikant effekt på tillit (a). I tillegg har de en signifikant direkte effekt på forventet anger, som ikke kan tilskrives mediatoren (c'). Det er ingen signifikante totale direkte effekter (c), men det er heller ikke ikke nødvendig (Shrout og Bolger, 2002). Ettersom c' ikke er null, er det ikke kun tillit som forklarer effekten som valg og alvorlighetsgrad har på forventet anger. Av den grunn eksisterer det en delvis medierende effekt for gruppene LA fastlege og LA robot.

Oppfattet risiko har også gjennomgående i alle grupper en signifikant effekt på forventet anger (b). Videre påvirket gruppene HA fastlege vs LA fastlege, og HA fastlege vs LA robot, oppfattet risiko signifikant (a). Ingen har en signifikant direkte effekt (c') på forventet anger, men det er ikke nødvendig for at det skal være en medierende effekt til stede (Shrout & Bolger, 2002). Det eksisterer derfor delvis medierende effekter i alle nevnte grupper, siden c' ikke er null.

5.5.0 Oppsummering av resultater

Vi vil nå gi en kort oppsummering av resultatene i analysen. En oppsummering er presentert i tabell 25.

	Hypotese	Resultat
H1	En høyere andel av deltakerne vil velge fastlege enn robot for en medisinsk konsultasjon, når sannsynligheten for korrekt diagnostisering og behandling er den samme	Støttes
	H1a: Deltakerne med symptomer av lav alvorlighetsgrad velger oftere robot, enn deltakerne med symptomer av høy alvorlighetsgrad	Støttes
H2	Deltakerne som har valgt robot utviser lavere grad av tillit, enn deltakerne som har valgt fastlege	Støttes
	H2a: Deltakerne som har valgt robot utviser lavere grad av tillit når de har symptomer av høy alvorlighetsgrad, enn deltakerne med symptomer av lav alvorlighetsgrad	Støttes ikke
H3	Deltakerne som har valgt robot utviser høyere grad av oppfattet risiko, enn deltakerne som har valgt fastlege	Støttes ikke
	H3a: Deltakerne som har valgt robot utviser høyere grad av oppfattet risiko når de har symptomer av høy alvorlighetsgrad, enn deltakerne med symptomer av lav alvorlighetsgrad	Støttes ikke
H4	Deltakerne som har valgt robot utviser lavere grad av tilfredshet, enn deltakerne som har valgt fastlege	Støttes ikke
	H4a: Deltakerne som har valgt robot utviser lavere grad av tilfredshet når de har symptomer av høy alvorlighetsgrad, enn deltakerne med symptomer av lav alvorlighetsgrad.	Støttes ikke
H5	Deltakerne som har valg robot utviser høyere grad av forventet anger, enn deltakerne som har valgt fastlege	Støttes ikke
	H5a: Deltakerne som har valgt robot utviser høyere grad av forventet anger når de har symptomer av høy alvorlighetsgrad, enn deltakerne med symptomer av lav alvorlighetsgrad	Støttes ikke
H6	H6a: Effekten som valg har på tilfredshet, kan delvis forklares av tillit	Støttes
	H6b: Effekten som valg har på tilfredshet, kan delvis forklares av oppfattet risiko	Støttes ikke
H7	H7a: Effekten som valg har på forventet anger, kan delvis forklares av tillit	Støttes
	H7b: Effekten som valg har på forventet anger, kan delvis forklares av oppfattet risiko	Støttes ikke

Tabell 25: Oppsummering av resultater fra hypotesetesting

5.5.1 Valg av behandlingsform

Vi ønsket å undersøke om det er en forskjell i andel som valgte robot og fastlege. I den sammenheng fikk vi signifikante resultater på at det er forskjell mellom gruppene, slik at H1 støttes. Deretter ville vi se på om alvorlighetsgrad av symptomer kan påvirke valget av behandlingsmetode. Her fikk vi signifikante resultater på at det er flere deltakere med lav alvorlighetsgrad som velger robot, dermed støttes også H1a. Til slutt fikk vi signifikante resultater når respondentene får introdusert et tredje alternativ, der 80.8% ville blitt behandlet av en fastlege med støtte av en robot.

5.5.2 Grad av tillit utvist til behandlingsform

Vi ønsket å undersøke om de som valgte fastlege hadde en høyere tillitsscore enn de som valgte robot. H2 støttes siden det er signifikante forskjeller mellom gruppene som har valgt fastlege og robot. Videre ville vi se på om ulik alvorlighetsgrad på symptomene påvirket tillitsscoren til respondentene, og om det er ulikheter mellom respondentene som har valgt robot og fastlege. Vi fikk signifikante forskjeller mellom LA fastlege og LA robot, samt LA fastlege og HA robot. Mer presist utviser de som har valgt fastlege høyere tillit. Vedrørende

H2a er det signifikante resultater for forskjell mellom LA robot og HA robot, imidlertid er differansen motsatt av forventet. Av den grunn støttes ikke H2a.

5.5.3 Grad av oppfattet risiko i forbindelse med behandlingsform

Deretter ble det testet om oppfattet risiko er høyere hos respondentene som har valgt robot enn fastlege. H3 støttes ikke, ettersom vi ikke fikk signifikante resultater. I tillegg ville vi se på om ulik alvorlighetsgrad på symptomene påvirket oppfattet risiko hos respondentene, og om det er ulikheter mellom de som har valgt robot og de som har valgt fastlege. Vi fikk signifikante forskjeller mellom HA fastlege og HA robot, samt HA fastlege og LA robot. Her var differansen motsatt av forventet for begge. Resultatene indikerer at det kan eksistere en forskjell basert på alvorlighetsgrad, og at de som har valgt robot har en lavere oppfattet risiko relatert til diagnostisering og behandling. Vi fikk også signifikante forskjeller mellom HA fastlege og LA fastlege. Her var differansen som forventet ved at scoren er høyest ved høy alvorlighetsgrad. Videre er det ikke signifikante resultater for forskjell mellom LA robot og HA robot, slik at H3a ikke støttes.

5.5.4 Evaluering av behandlingsform

Tilfredshet

Det ble så testet for om det foreligger en forskjell i vurdering av tilfredshet mellom deltakerne som har valgt fastlege og robot. Vi fikk akkurat ikke signifikant forskjell mellom gruppene, slik at H4 ikke støttes. Videre ville vi undersøke om ulik alvorlighetsgrad på symptomene påvirker tilfredshetsscoren til respondentene, og om det er ulikheter mellom respondentene som har valgt robot og fastlege. I den forbindelse fikk vi signifikante forskjeller mellom HA fastlege og HA robot, samt HA fastlege og LA robot. Her var differansen motsatt av forventet for begge. Vi fikk også signifikante resultater for HA fastlege og LA fastlege, der differansen var som forventet ved at den var høyest for lav alvorlighetsgrad. Det er ikke signifikante resultater for forskjell mellom LA robot og HA robot, slik at H4a ikke støttes.

Vedrørende medierende effekter for gruppene som har valgt fastlege og robot, eksisterer det for tillit en delvis medierende effekt på tilfredshet, slik at H6a støttes. Imidlertid er resultatene mer varierende når alvorlighetsgrad inkluderes. Her var det kun for LA fastlege og LA robot at tillit delvis medierte tilfredshet. Det var ingen medierende effekter for oppfattet risiko, slik at H6b ikke støttes. Tilsvarende resultater fikk vi da alvorlighetsgrad ble inkludert.

Forventet anger

Vi fikk ikke signifikante resultater på om det foreligger forskjell mellom noen av gruppene. Av den grunn støttes ikke H5, som omhandler om det foreligger en forskjell i forventet anger mellom deltakerne som har valgt fastlege og robot. Av samme grunn er det heller ingen forskjell mellom gruppene HA robot og LA robot, slik at H5a støttes heller ikke,

For gruppene som valgte fastlege og robot eksisterer det for tillit en delvis medierende effekt på forventet anger, og H7a støttes. Da alvorlighetsgrad ble inkludert eksisterte det kun en delvis medierende effekt på forventet anger for LA fastlege og LA robot. Videre hadde ikke oppfattet risiko en medierende effekt på forventet anger, for gruppene som har valgt fastlege og robot. Dermed støttes ikke H7b. Til slutt eksisterte det en delvis medierende effekt på forventet anger for gruppene HA fastlege, LA fastlege, HA fastlege og LA robot da alvorlighetsgrad ble inkludert.

6 Diskusjon

I det videre vil vi diskutere resultatene fra analysen i forrige kapittel. Oppgavens formål har vært å besvare følgende forskningsspørsmål:

Er det algoritmeaversjon til stede hos pasienter i forbindelse med utførelse av en medisinsk konsultasjon, og er aversjonen svakere når symptomene er av lav alvorlighetsgrad?

Bakgrunnen for oppgaven og forskningsspørsmålet, er utfordringene som knyttes til implementering av AI-teknologi i helsesektoren. For å sikre en bærekraftig helsesektor i Norge er en avhengig av optimalisering og effektivisering av driften, noe økt bruk av AI-teknologi bidrar til (Riekeles, 2018). I tillegg viser en rekke studier at AI-teknologi gjør det bedre enn mennesker (Logg et al., 2019; Castelo et al., 2019; Beck et al., 2011; Dawes et al., 1989; Grove et al., 2000). Til tross for dette viser mennesket en motvilje til å benytte seg av teknologien på flere områder - også relatert til helse. Dette kalles algoritmeaversjon. Vi ønsket å undersøke om det foreligger indikasjoner på algoritmeaversjon i forbindelse med en medisinsk konsultasjon. Videre ville vi se på om aversjonen endres når alvorlighetsgraden på de oppgitte symptomene er forskjellig - variert mellom lav eller høy.

For å måle indikasjoner på algoritmeaversjon, har vi utformet ulike hypoteser. Disse er utformet på bakgrunn av litteratur om algoritmeaversjon, der noen av de tidligere har blitt benyttet for å måle aversjon mot å benytte AI-teknologi. Tidligere er ikke oppfattet risiko før blitt benyttet som indikator for å måle fenomenet. Imidlertid, siden litteraturen gir indikasjoner på at risiko og tillit kan korrelere negativt (Partick, 2002), kan det argumenteres for at dette også kan være en måte å måle aversjon på.

Vi vil i det følgende ta for oss hypotesene, og drøfte de respektive resultatene opp mot funn fra tidligere forskning. Deretter vil vi foreta en helhetlig diskusjon av om det foreligger indikasjon på algoritmeaversjon, hensyntatt alle hypotesene.

6.1.0 Valg av behandlingsform

Litteraturen viser at valg av behandlingsform er en velfungerende måte å måle indikasjon på algoritmeaversjon (Jussupow et al., 2020). Det har også foreligget kritikk mot studier som ikke har benyttet dette når algoritmeaversjon er blitt studert (Pezzo & Beckstead, 2020). Våre funn er i tråd med tidligere studier, ved at vi fikk signifikante resultater på valg av behandlingsform som var i disfavør av roboten, der kun 29% valgte denne. En årsak kan være at deltakerne har manglende tillit til AI-teknologi på bakgrunn av at konsekvensene ved feil kan anses som store, siden det omhandler egen helse (Asan et al., 2020; Castelo et al., 2019).

Vi fikk også signifikante resultater ved valg av behandlingsform, da vi testet for forskjeller mellom lav og høy alvorlighetsgrad: For høy alvorlighetsgrad var det 22% som valgte robot, og ved lav alvorlighetsgrad 35%. Med 13 prosentpoeng forskjell er dette over en 50% økning i valg av robot. Funnene indikerer derfor at graden av algoritmeaversjon reduseres når alvorlighetsgraden av symptomene er mindre alvorlige.

Funnet om at alvorlighetsgrad kan påvirke valg av behandlingsform kan blant annet forklares ved at konsekvensene av potensielle feil vurderes ulikt. Ved et sykdomsbilde der det er åpenbart og tydelig hva som er galt, og pasienten ikke anser dette som spesielt alvorlig, kan derfor også konsekvensene ved feil anses som mindre. Av den grunn kan det tenkes at pasienten kan være mer villig til å benytte seg av en robot. I motsatt fall, dersom sykdomsbildet er mer komplekst og diffust, og konsekvensene følgelig kan anses som store om noe skulle gå galt, kan det tenkes at en ikke er villig til å ta risikoen om å benytte en robot.

Resultatene er derfor i tråd med antakelsene våre om at oppfattet risiko innenfor egen helse kan være ulik, og at dette kan påvirke villigheten til å benytte seg av en robot (Asan et al., 2020).

Når alvorlighetsgraden er høy kan pasienten også føle et større behov for å møte et empatisk menneske. Samtidig kan det også foreligge et større behov for å få informasjon, og ha en å sparre med. Antakelsene er i så fall i tråd med litteraturen, som viser til at dette anses som manglende egenskaper ved AI-teknologien (Kerasidou et al., 2019; Lee, 2018; Castelo et al., 2019). Det kan derfor tenkes at disse behovene ikke er like sterke når alvorlighetsgraden på symptomer oppfattes som lav.

En annen forklaring for funnene kan være at pasientene foretar ulik vekting i vurderingen av sannsynligheten for å få feil diagnose og behandling av robot og fastlege. Det er innledningsvis i spørreundersøkelsen opplyst om at treffsikkerheten for begge behandlingsformene er på 80%. Imidlertid viser tidligere studier at AI-teknologi "straffes" i mye større grad enn menneske hvis feil gjøres (Dietvorst et al, 2015; Logg et al., 2019). Den oppgitte prosentandelen for feil, som er på 20%, kan dermed bli ilagt ulik vekting for robot og menneske. Det kan derfor tenkes at vektingen forsterkes når alvorlighetsgraden er høy, siden konsekvensene av feil er mer alvorlig.

Det kan også tenkes at respondentene som har valgt robot gjør andre antakelser, som også påvirker valget deres, og som en kvantitativ studie ikke klarer å fange opp. En slik antakelse kan være kortere ventetid på legekontoret om en velger robot som behandlingsform. Dette begrunnes med at det er en kjent sak at det ofte kan være forsinkelser hos fastlegen. Videre, kan en annen antakelse være at robot er et rimeligere alternativ enn fastlege. Et annet moment kan være at ønsket om en rask og effektiv konsultasjon veier tungt, og at det derfor ikke er ønskelig å måtte snakke med en fastlege. Dette kan særlig synes gjeldene ved lav alvorlighetsgrad, siden behovet for en sparringspartner da ikke er så stort.

Litteraturen viser også at mennesker i større grad er villig til å benytte seg av en robot dersom de kan foreta små justeringer, eller at et menneske tar den endelige avgjørelsen (Dietvorst et al., 2018). Tilsvarende funn finnes også innen medisinsk behandling: Longoni et al. (2019) viser at pasientene foretrekker å benytte AI-teknologi, dersom det benyttes som en støttefunksjon og legen er endelig beslutningstaker. Vi finner tilsvarende funn i vårt studie

ved at det er en signifikant forskjell fra deltakernes initielle valg: 80.8% valgte alternativet om å benytte Dr. Johansen sammen med IBM Watson da de fikk mulighet til det. I de ulike gruppene strakk prosentandelen seg fra en minsteverdi på 76% hos LA fastlege, til 96% hos HA robot. Resultatene indikerer at respondentene er villig til å benytte en robot, men at det fortsatt er ønskelig at et menneske er involvert, og gjør de endelige vurderingene. På den måten får en den menneskelige kontakten, som kan være ønskelig i forbindelse med en konsultasjon, samtidig som at risikoen for menneskelige feil minimeres ved å benytte en robot (Dietvorst et al., 2018). Av de som først valgte robot var det også mange som heller ønsket seg dette alternativet. Funnene kan indikere at det i stor grad er ønskelig å bevare den menneskelige kontakten innen helsetjenester.

Konklusjonen er at valg av behandlingsform gir en indikasjon på algoritmeaversjon. Det å endre på alvorlighetsgraden på symptomer påvirker også graden av algoritmeaversjon. Videre foreligger det ikke indikasjoner på algoritmeaversjon når en får et tredje alternativ, der deltakerne også kan velge å få behandling fra robot og menneske sammen.

6.2.0 Grad av tillit utvist til behandlingsform

Å måle tillit er i henhold til litteraturen, også en god måte å måle algoritmeaversjon (Jussupow et al., 2020). Det eksisterer signifikante forskjeller i tillitsscoren mellom gruppene som har valgt fastlege og robot. Den gjennomsnittlig tillitsscoren til fastlege er signifikant høyere enn til robot, der differansen er 0.401. Funnene gir derfor indikasjoner på algoritmeaversjon. Det er også signifikante resultater ved lav alvorlighetsgrad, der robot og fastlege sammenlignes. Her er differansen 0.712 i favør fastlegen, som også kan indikere algoritmeaversjon ved lav alvorlighetsgrad.

Årsaker til forskjellen i tillitsscore kan være fravær av menneskelige egenskaper hos roboten, som eksempelvis empati, intuisjon og medfølelse. Egenskapene kan anses å være av stor viktighet når det gjelder noe så viktig og personlig som ens egen helse. Funnene er dermed i tråd med litteraturen som viser til at fravær av slike menneskelige egenskaper ved AI-teknologi påvirker tilliten pasientene utviser (Kerasidou et al., 2019; Lee, 2018; Castelo et al., 2019).

En annen årsaksforklaring kan være at deltakerne tror at roboten ikke klarer å ta hensyn til individuelle vurderinger. Mer presist kan det være at respondentene tenker at roboten kan være like dyktig som fastlegen, siden treffsikkerheten er 80% for begge, men at den derimot ikke er i stand til å utføre individuelle vurderinger som er spesielt ved ens eget tilfelle. På den måten kan det være at pasientene tror at det foreligger en større sannsynlighet for at en selv kan være blant de 20% som ikke får riktig diagnose og behandling. I så tilfelle faller dette inn under begrepet "unikhetsforsømmelse" fra litteraturen. Det omhandler at algoritmer ikke evner å ta hensyn til pasientens unike karakteristika og forhold (Longoni et al., 2019).

Dersom de overnevnte utfordringene er gjeldende kan dette bidra til å svekke tillitskomponenten "velvilje", som finnes innenfor medisinsk behandling. Dette innebærer blant annet å forstå pasientens individuelle behov, vise omsorg, ærlighet og respekt, og at tillitshaver handler i beste interesse for pasienten (Newcomer ,1997; Leisen og Hyman, 2001).

Når vi inkluderer alvorlighetsgrad viser funnene imidlertid at det ikke er signifikante resultater mellom gruppene HA fastlege og HA robot. Vi får heller ikke signifikante resultater når HA fastlege og LA robot sammenlignes, samt LA fastlege og HA robot. Det er derfor ikke mulig å si noe om alvorlighetsgrad generelt påvirker tillitsscoren. Vi kan dermed heller ikke si noe om algoritmeaversjonen, målt ved tillit, kan være lavere ved lav alvorlighetsgrad. En forklaring kan være at forskjellene på gruppenes opplevde alvorlighetsgrad på symptomene (lav/høy) ikke er store nok, siden vi ikke fikk signifikante forskjeller på det.

Det som er spesielt interessant er at vi får signifikante resultater på at HA robot har 0.327 høyere snittscore på tillit til behandlingsformen, enn LA robot. Det er motsatt av hva vi trodde. Snittscoren for robot ved høy alvorlighetsgrad er fortsatt lavere enn for fastlege. Funnene er derfor fortsatt i tråd med tillitsutfordringene en har til AI-teknologi, slik at det fortsatt kan indikere algoritmeaversjon. En forklaring for funnet kan være at respondentene selv har valgt behandlingsform. I den sammenheng kan det tenkes at de som er i gruppen HA robot kan ha en særlig tiltro til AI-teknologien, ettersom de har valgt robot på tross av at konsekvensene ved feil er store.

Konklusjonen er at tillitsscore gir en indikasjon på algoritmeaversjon. Imidlertid er det usikkert om alvorlighetsgrad generelt påvirker graden av algoritmeaversjon. Den kan i så fall

synes å være motsatt for gruppene som har valgt robot, i form av at når alvorlighetsgraden er høy, viser deltakerne av ulike grunner en lavere grad av algoritmeaversjon.

6.3.0 Grad av oppfattet risiko i forbindelse med behandlingsform

Ettersom litteraturen sier at det er en sentral sammenheng mellom tillit og risiko (Boon & Holmes, 1991; Giddens, 1991; Jøsang & Presti, 2004), vurderte vi det som hensiktsmessig å inkludere oppfattet risiko som en indikasjon på graden av aversjon. Spesielt siden vi ville se på forholdet mellom høy og lav alvorlighetsgrad. Imidlertid fikk vi ikke signifikante forskjeller mellom gruppenes score vedrørende oppfattet risiko, når en sammenligner gruppene som har valgt fastlege og robot. Funnene kan indikere at det ikke foreligger indikasjon på algoritmeaversjon, at oppfattet risiko ikke er egnet som måleindikator, eller at vi ikke har klart å måle oppfattet risiko godt nok.

Når vi inkluderer alvorlighetsgrad, viser imidlertid funnene at den oppfattede risikoen er høyere for respondenter som er i gruppen HA fastlege enn respondentene i gruppen HA robot. Vi fant også at den oppfattede risikoen er høyere for gruppen HA fastlege, enn gruppen LA robot. Begge resultatene strider mot hypotesene våre, og tidligere funn i litteraturen (Mohtar & Abbas, 2015; Lacey et al., 2019).

Årsakene bak funnene er trolig sammensatt. En forklaring kan være at deltakerne med høy alvorlighetsgrad er mer usikre på diagnostiseringen og behandlingen enn de med mindre alvorlige symptomer. Dermed har de også en høyere oppfattet risiko, siden konsekvensene er større for feil behandling. Funnene stemmer da overens med teorien (Partick, 2002). I den sammenheng kan det derfor synes at symptomenes alvorlighetsgrad kan veie tyngre enn det faktum at en robot blir benyttet.

En annen forklaring kan være at de som valgte robot har hatt dårlige opplevelser med fastleger tidligere, og derfor kan ha en underliggende tanke om at “alle andre alternativer er mindre risikable”. Det kan også tenkes at respondentene som valgte robot har såpass stor tiltro til AI-teknologien sine egenskaper, at de anså den oppfattede risikoen tilknyttet diagnostisering og behandling som lav. Forklaringen kan støttes av at deltakerne tok et aktivt valg om å ha robot som behandlingsmetode. Av den grunn kan det også tenkes at scoren på oppfattet risiko hadde vært høyere om deltakerne ikke hadde fått muligheten til å velge selv.

I henhold til litteraturen kan oppfattet risiko deles inn i usikkerhet og konsekvenser (Fischhoff et al., 1985). Her kan det gjøres noen betraktninger relatert til at HA fastlege har høyere oppfattet risiko enn HA robot. Siden gruppene som fikk oppgitt høy alvorlighetsgrad hadde lik informasjon hva gjelder symptomer, bør vurderingen, rent objektivt, være lik på tvers av gruppene. Imidlertid hevdes det i litteraturen at oppfattet risiko er en subjektiv opplevelse (Mitchell, 1999; Grima et al, 2019). Derfor kan det tenkes at enkelte deltakere har en høyere oppfattet risiko enn andre, og at dette kommer til syne i valget av behandlingsmetode. Mer presist, kan det synes at deltakerne som har valgt fastlege generelt har en høyere subjektiv oppfattet risiko enn de som har valgt robot.

Konklusjonen er at oppfattet risiko ikke gir en indikasjon på algoritmeaversjon. Videre kan vi ikke generelt si om alvorlighetsgrad påvirker oppfattet risiko. Imidlertid påvirkes oppfattet risiko signifikant ved høy alvorlighetsgrad, der vurderingen var i favør av robot med lavere oppfattet risiko. Graden av oppfattet risiko var også i favør robot når gruppen hadde lav alvorlighetsgrad, mot gruppen HA fastlege.

6.4.0 Evaluering av behandlingsform

Å benytte evaluering som en indikator på algoritmeaversjon er også i tråd med litteraturen. Dette kan gjøres ved å se på evalueringen av de ulike behandlingsformene. Dersom menneske får en bedre evaluering enn robot foreligger det aversjon (Jussupow et al., 2020). Vi har i vårt tilfelle delt evalueringen opp i tilfredshet og forventet anger, og vi vil i det følgende diskutere begge.

Tilfredshet

Vi fikk ikke signifikante forskjeller mellom gruppene som valgte fastlege og robot. Resultatene kan indikere at det enten ikke foreligger algoritmeaversjon når det måles ved tilfredshet, at vi ikke har klart å måle tilfredshet godt nok, eller at det ikke egner seg som måleindikator. Det er motstridende med tidligere studier. En årsak til funnene kan være at vi kun har inkludert to spørsmål på tilfredshet, slik at vi ikke har klart å fange opp alle aspekter ved dette. I den forbindelse bør det presiseres at resultatet akkurat ikke er signifikant, med en p-verdi som var 0.053.

Da alvorlighetsgrad ble inkludert fikk vi imidlertid signifikante forskjeller mellom gruppene HA fastlege og LA robot, som var motsatt av forventet. Det betyr at LA robot var mer tilfreds enn HA fastlege. En årsak til funnene kan være at alvorlighetsgrad veier tyngre enn det faktum at det er en robot som benyttes, slik at det er dette som blir utslagsgivende for hvor tilfreds en er. Å benytte en robot kan derfor i utgangspunktet gi en lavere tilfredshet, men ved lav alvorlighetsgrad blir den ikke lav nok til at den er dårligere enn for HA fastlege.

Vi fikk også signifikante forskjeller på fastlege og robot der respondentene hadde høy alvorlighetsgrad. Imidlertid er resultatet motsatt av forventet, slik at det ikke er i tråd med hypotesene våre. Tilfredshet omfavner hele tjenesteforløpet, og vurderingen kan derfor innebære alle spektre ved konsultasjonen. En årsak til at tilfredsheten er høyest for robot kan derfor være at deltakerne, som har valgt robot, ikke vurderer det som negativt at roboten mangler menneskelige egenskaper som empati og tillit. Resultatene er i den sammenheng motsatt av tidligere funn (Lee, 2018; Castelo et al., 2019). En annen årsak kan være at de som har valgt robot ikke anser disse egenskapene som viktige for sin bruk av tjenesten.

En årsaksforklaring kan også være respondentene har hatt dårlige opplevelser med fastleger tidligere, slik at robot uansett anses som et bedre alternativ. Det kan også være at respondentene anser roboten som best egnet til å utføre oppgaven. Det kan være fordi de som har valgt robot har større tiltro til at teknologien kan gjøre en bedre jobb enn en fastlege. Det kan begrunnes med at en er godt kjent med teknologi og dets muligheter. Argumentet kan spesielt synes gjeldende for deltakere som har valgt robot ved høy alvorlighetsgrad, siden konsekvensene da er større. Funnene kan derfor indikere det motsatte av algoritmeaversjon, siden snittscoren er høyere for robot enn menneske. Resultatene er derfor i tråd med litteraturen som handler om algoritmeverdsettelse, som defineres som fenomenet der mennesker kan ha preferanser for AI-teknologi over mennesker (Logg et al., 2019).

Analysen viser videre at tillit hadde en medierende effekt på tilfredshet for gruppene som har valgt fastlege og robot. Dette er derfor i tråd med antakelsene våre om at indikatorene kan henge sammen, ettersom de måler samme fenomen. Av den grunn er det viktig at tillit til behandlingsformen må være tilstedeværende for at en også skal kunne være tilfreds med konsultasjonen. Når alvorlighetsgrad inkluderes har tillit kun en medierende effekt på tilfredshet ved lav alvorlighetsgrad. Imidlertid er det gjennomgående for alle grupper at tillit

påvirker tilfredshet signifikant. Det betyr at ved å øke tilliten til behandlingsformen, vil dette kunne øke tilfredsheten.

Videre har ikke oppfattet risiko en medierende effekt på tilfredshet. Dette samsvarer derfor med resultatene vi har ellers ved at det ikke foreligger indikasjon på algoritmeaversjon for oppfattet risiko.

Forventet anger

Vi fikk ikke signifikante forskjeller mellom gruppene som valgte fastlege og robot. Av den grunn har vi ikke klart å få en indikasjon på algoritmeaversjon ved å benytte forventet anger. Funnene kan derfor indikere at det enten ikke foreligger algoritmeaversjon når det måles ved forventet anger, at vi ikke har klart å måle forventet anger godt nok, eller at det ikke egner seg som måleindikator. Det at deltakerne fikk opplyst at begge behandlingsformene hadde like stor sannsynlighet for korrekt diagnostisering og behandling, kan bidra til at graden av anger reduseres. Derfor kan det tenkes at vi hadde fått andre resultater om vi ikke hadde opplyst om dette innledningsvis.

Som ved tilfredshet har tillit en medierende effekt på forventet anger for gruppene som har valgt fastlege og robot. Dette er derfor også i tråd med antakelsene våre om at indikatorene kan henge sammen, ettersom de måler samme fenomen. Av den grunn er det viktig at tillit til behandlingsformen må være tilstedeværende for at en ikke skal føle anger i forbindelse med konsultasjonen. Når alvorlighetsgrad inkluderes er det også her en medierende effekt med tillit ved lav alvorlighetsgrad. Imidlertid er det også gjennomgående for alle grupper at tillit påvirker forventet anger signifikant. Det betyr at den forventede angeren kan reduseres, ved å øke tilliten til behandlingsformen.

For mediatoren oppfattet risiko var det en medierende effekt for gruppene HA fastlege, LA fastlege og LA robot. I tillegg var det gjennomgående for alle grupper at oppfattet risiko hadde en signifikant effekt på forventet anger. Dermed kan det å påvirke hvordan pasientene oppfatter risikoen, i forbindelse behandlingsformen, være nyttig for å redusere forventet anger.

Andre psykologiske mekanismer som kan påvirke evaluering

Gjennomgående ser vi indikasjon på algoritmeaversjon ved valg av behandlingsmetode og grad av tillit som utvises til behandlingsmetoden, men derimot ikke når en foretar en overordnet evaluering av behandlingen. Funnene i forbindelse med tilfredshet og forventet anger, kan derfor også forklares ved forskjellige psykologiske mekanismer som påvirker menneske ved beslutningstaking og evaluering.

Det kan blant annet tenkes at de som har valgt robot post-rasjonaliserer valget sitt, som anses å være en psykologisk forsvarsmekanisme for å rettferdiggjøre valg (Aslaksen, 2019). I dette tilfellet kan det være at deltakerne rettferdiggjør valget sitt om å velge robot ved evalueringen de foretar. Post-rasjonalisering forekommer oftest når en blir bedt om en begrunnelse for et valg eller en handling, som er tilfellet her.

Tilsvarende kan det også tenkes at deltakerne har blitt påvirket av "self perception theory". Teorien sier at når en ikke har tidligere erfaring med en gitt situasjon, kan holdninger dannes basert på adferden en har utvist i den konkrete situasjonen (Bem, 1972). I dette tilfellet kan det tenkes at gruppen som har valgt robot dermed danner tydelige positive holdninger til valget, på bakgrunn av at dem selv faktisk foretok det. Det er naturlig å anse seg selv som en fornuftig person som foretar gode beslutninger. Holdningene kan derfor komme til syne ved at evalueringen av robot som behandlingsform kan være for positiv. På denne måten underbygger en for en selv at det har blitt tatt en velbegrunnet og riktig beslutning.

En annen forklaring kan være at deltakerne blir påvirket av en "sunk cost fallacy". Mer presist kan det være at deltakernes "investering" i valget, i form av mental energi og tid, har vært av såpass stor størrelse at det påvirker evalueringen av behandlingsformen (Arkes & Blumer, 1985). Ettersom det er mer usikkerhet involvert i å velge robot, kan det tenkes at deres "investering" er større enn deltakerne som har valgt fastlege. Det kan også tenkes at "investeringen" er større ved høy alvorlighetsgrad, siden risikoen for feil da innehar større konsekvenser. Dette kan resultere i at en i større grad står fast ved valget om at robot var riktig, og at en dermed setter en høyere score.

Konklusjon evaluering

Konklusjonen er at tilfredshet ikke gir indikasjon på at det foreligger algoritmeaversjon. Videre er det usikkert om alvorlighetsgrad generelt påvirker graden av

algoritmeaversjon. Imidlertid utgjør alvorlighetsgrad en forskjell, i form av at snittscoren er signifikant høyere i favør av HA robot sammenlignet med HA fastlege. Tilsvarende gjelder for LA robot som har signifikant høyere snittscore enn HA fastlege. Forventet anger gir heller ikke indikasjon på at det foreligger algoritmeaversjon. Tillit fungerte som en mediator på både tilfredshet og forventet anger, for gruppene som valgte fastlege og robot.

6.5.0 Indikasjoner på algoritmeaversjon

Ved å benytte ulike måter å måle algoritmeaversjon, kan vi se om det totalt sett kan foreligge indikasjoner på algoritmeaversjon i forbindelse med bruk av robot til en medisinsk konsultasjon. Vi kan også se om alvorlighetsgrad utgjør en forskjell på graden av algoritmeaversjon. En oppsummering av om det foreligger indikasjoner på algoritmeaversjon vises i tabell 26.

For å oppsummere ga valg av behandlingsform indikasjoner på at det foreligger algoritmeaversjon, der også alvorlighetsgrad signifikant påvirker graden av aversjon. Videre ga måling av tillit også indikasjoner på at det foreligger algoritmeaversjon. Hva gjelder alvorlighetsgrad kan vi ikke fastslå at det generelt påvirker algoritmeaversjon. Imidlertid påvirket alvorlighetsgrad de som har valgt robot, der tilliten var høyest hos de som hadde høy alvorlighetsgrad. Dette er motsatt av forventet. Oppfattet risiko ga ikke indikasjoner på algoritmeaversjon. Heller ikke her kan vi fastslå om alvorlighetsgrad generelt påvirker algoritmeaversjon, siden resultatene varierer. Hverken tilfredshet eller forventet anger ga indikasjon på algoritmeaversjon. Også her varierte resultatene når alvorlighetsgrad ble inkludert.

Hva gjelder medierende effekter var dette tilstedeværende for tillit på både tilfredshet og forventet anger, da vi så på gruppene som hadde valgt fastlege og robot. Imidlertid medierte ikke oppfattet risiko hverken tilfredshet eller forventet anger. Videre var resultatene varierende da alvorlighetsgrad ble inkludert.

	Algoritmeaversjon	Resultat
Valg	Valg av behandlingsform gir en indikasjon på algoritmeaversjon	Ja
	Alvorlighetsgrad påvirker graden av algoritmeaversjon	Ja
Tillit	Grad av tillit utvist til behandlingsform gir en indikasjon på algoritmeaversjon	Ja
	Alvorlighetsgrad påvirker graden av algoritmeaversjon	Varierende resulutater
Oppfattet risiko	Grad av oppfattet risiko i forbindelse med behandlingsform gir en indikasjon på algoritmeaversjon	Nei
	Alvorlighetsgrad påvirker graden av algoritmeaversjon	Varierende resultater
Tilfredshet	Grad av tilfredshet utvist til behandlingsform gir en indikasjon på algoritmeaversjon	Nei
	Alvorlighetsgrad påvirker graden av algoritmeaversjon	Varierende resultater
	Tillit har en medierende effekt på tilfredshet	Ja
	Oppfattet risiko har en medierende effekt på tilfredshet	Nei
Forventet anger	Grad av forventet anger utvist til behandlingsform gir en indikasjon på algoritmeaversjon	Nei
	Alvorlighetsgrad påvirker graden av algoritmeaversjon	Varierende resultater
	Tillit har en delvis medierende effekt på forventet anger	Ja
	Oppfatter risiko har en medierende effekt på forventet anger	Nei

Tabell 26: Oppsummering av indikasjoner på algoritmeaversjon

7 Konklusjon

Formålet med det siste kapittelet er å besvare oppgavens forskningsspørsmål. Deretter vil vi se på oppgavens begrensninger. Til slutt vil vi diskutere teoretiske og praktiske implikasjoner, før vi kommer med anbefalinger til videre forskning.

Hensikten med denne oppgaven har vært å undersøke om det foreligger indikasjoner på algoritmeaversjon for pasienter i forbindelse med en medisinsk konsultasjon, samt om aversjonen er svakere når alvorlighetsgraden oppleves som lav. For å svare på dette har vi presentert teori om de psykologiske mekanismene tillit og risiko som påvirker menneske ved beslutningstaking. Vi har også presentert teori innen Human-Computer Interaction (HCI), kunstig intelligens og algoritmeaversjon, for å få en rikere innsikt i teknologiens utvikling, funksjoner og muligheter, samt menneskets interaksjon med denne. På bakgrunn av den presenterte teorien, ble det utviklet en modell med tilhørende hypoteser. For å svare på hypotesene, ble det samlet inn data kvantitativt, ved et eksperiment som var utformet som en spørreundersøkelse. Respondentene ble rekruttert ved en kombinasjon av selvselektering og snøballeffekt. Vi har dermed et ikke-sannsynlighetsutvalg, og av den grunn svekkes generaliserbarheten fra utvalget til populasjonen.

Funnene i oppgaven peker mot at det foreligger indikasjon på algoritmeaversjon. Imidlertid kommer funnene kun til syne ved valg av behandlingsform og grad av tillit til behandlingsformen, som fra litteraturen også har blitt benyttet som måleindikator på algoritmeaversjon. De andre hypotesene om oppfattet risiko, tilfredshet og forventet anger ga ikke signifikante resultater. Det er derfor ikke mulig å si noe om det foreligger indikasjoner på algoritmeaversjon basert på dette.

Videre finner vi at når alvorlighetsgraden inkluderes, har det varierende effekt på graden av algoritmeaversjon som utvises. Ved valg av behandlingsform finner vi signifikante forskjeller. Det indikerer at algoritmeaversjonen kan være sterkere når alvorlighetsgraden av symptomene oppfattes som høy. I vurderingen av tillit fikk vi varierende resultater. For de som valgte robot var resultatene motsatt av forventet, der deltakerne som hadde høy alvorlighetsgrad utviste høyest tillit. For oppfattet risiko og tilfredshet er også resultatene varierende. Forventet anger hadde ikke signifikante forskjeller mellom noen grupper, og av den grunn påvirket heller ikke alvorlighetsgrad graden av algoritmeaversjon. Totalt sett kan vi ikke fastslå generelt at alvorlighetsgrad påvirker graden av algoritmeaversjon. Imidlertid synes det å være en sammenheng, og av den grunn vurderes det som nødvendig med videre studier for å kunne fatte entydige konklusjoner.

Vedrørende medierende effekter finner vi at tillit har en delvis medierende effekt på både tilfredshet og forventet anger, da vi så på gruppene som hadde valgt fastlege og robot.
Av den grunn er det viktig at tillit til behandlingsformen må være tilstedeværende for at en også skal kunne være tilfreds med konsultasjonen og ikke føle på anger. Imidlertid fungerte ikke oppfattet risiko som en mediator. Da alvorlighetsgrad ble inkludert var resultatene mer varierende, og det kan derfor ikke konkluderes med at alvorlighetsgrad generelt påvirker de medierende effektene.

Det kan derfor synes at det foreligger indikasjoner på algoritmeaversjon blant pasienter ved gjennomføring av en medisinsk konsultasjon, der en robot benyttes i forbindelse med diagnostisering og behandling. Dette kommer til syne ved valg av behandlingsform og grad av tillit som utvises. Imidlertid er det ikke indikasjoner på algoritmeaversjon ved å se på oppfattet risiko, tilfredshet og forventet anger. Videre får vi varierende resultater vedrørende om alvorlighetsgraden av symptomene påvirker graden av algoritmeaversjon, der det kun er signifikant i forbindelse med valg av behandlingsform.

7.1.0 Begrensninger

Denne oppgaven har noen begrensninger som kan være av betydning for utredningen. Vi fikk ikke signifikante resultater tilknyttet kontrollspørsmålet om respondentenes vurdering av alvorlighetsgraden på symptomene (lav/høy). Signifikante forskjeller som vi har fått mellom gruppene kan fremdeles tilskrives denne manipulasjonen, men det er en svakhet i form av at vi ikke nødvendigvis har klart å gjøre forskjellen mellom høy og lav alvorlighetsgrad stor nok. Dersom vi hadde klart dette er det mulig at vi kunne fått flere signifikante resultater, og eventuelt større differanser mellom gruppene.

En annen begrensning gjelder antall spørsmål som benyttes for å måle konseptene. For å få nok deltakere, samt unngå "careless responding", er det hovedsakelig stilt to til fire spørsmål per konsept. Det anses som en svakhet siden vi antakelig ikke har klart å måle konseptene godt nok. Dette kan derfor påvirke resultatene vi har fått. I ettertid ser vi at vi kunne inkludert flere spørsmål i undersøkelsen, siden respondentene brukte kortere tid enn det som var forventet basert på pre-testen.

En tilsvarende svakhet er at vi ikke har tilstrekkelig antall spørsmål til å konstatere at de respektive personlighetstrekkene foreligger hos deltakerne. Det samme gjelder for teknologikompetanse og holdninger til teknologi. Dermed er det en sannsynlighet for at det kan foreligge sammenhenger mellom kontrollvariablene og konseptene, som vi ikke har klart å fange opp. Dette gjelder spesielt teknologikompetanse, som vi ikke klarte å måle, men som det kunne vært interessant å kontrollere for.

En annen begrensning relateres til at studien var fiktiv. En kan ikke med sikkerhet vite at deltakerne klarte å forestille seg symptomene en hadde fått tildelt, og dermed gi en reell vurdering av behandlingsformen. Det kan godt være at deltakerne hadde foretatt andre valg om de faktisk hadde hatt de oppgitte symptomene. I en reell situasjon, kan det blant annet tenkes at en ikke hadde ønsket å benytte robot likevel. Det begrunnes med at tanken om å benytte ny teknologi i utgangspunktet kan virke både nytt og spennende, men at en likevel ikke er villig til å ta valget når det blir et virkelig scenario. Tilsvarende gjelder for vurdering av behandlingsform, da det trolig oppleves annerledes å gjennomgå en medisinsk konsultasjon med en fastlege eller robot, fremfor å visualisere det.

Relatert til dette er det også en begrensning at vi har gjennomført eksperimentet på nett ved en spørreundersøkelse. Det gjør at studien kan oppfattes som kunstig. For å gjøre det mer reelt kunne eksperimentet blitt gjennomført på et legekontor. I den sammenheng kan det benyttes en fastlege og et datasystem, eksempelvis IBM Watson, som behandlere. Det lot seg dessverre ikke gjøre for denne oppgaven, og det er vanskelig å si om resultatene ville vært annerledes i et mer virkelighetsnært scenario.

Utvalget som gjennomførte eksperimentet utgjør også en begrensing. Siden vi rekrutterte deltakerne ved en kombinasjon av selvselektering og en snøballeffekt, bør en være forsiktig med å generalisere funnene til hele populasjonen. Det vurderes som at utvalget er relativt homogent, som kommer til syne ved testene i den deskriptive statistikken. Imidlertid anses utvalget som tilstrekkelig gitt omfanget av oppgaven. Videre er en svakhet at vi ikke har fått 38 deltakere i gruppen HA robot, som er ønsket minsteverdi for deltakere per gruppe. Årsaken er at vi ønsket at deltakerne skulle velge behandlingsform selv, ettersom det er en viktig indikator på algoritmeaversjon. Vi var imidlertid forberedt på at vi derfor kunne få enn ønsket i gruppene HA robot og LA robot.

Til slutt er tidsrommet for oppgaven en begrensning, siden den ble skrevet over ett semester. Det samme er de begrensede økonomiske ressursene vi hadde til rådighet. Hvis vi hadde hatt et lengre tidsrom og økonomiske midler, kunne vi gjennomført et mer omfattende eksperiment i en reell kontekst. Dersom vi fortsatt skulle gjennomført en spørreundersøkelse, kunne vi med økonomiske midler fått et mer representativt utvalg, ved bruk av et selskap som gjennomfører datainnsamling, for eksempel Norstat.

7.2.0 Implikasjoner og forslag til videre forskning

7.2.1 Teoretiske implikasjoner

I denne oppgaven har vi sett på om pasienter viser indikasjon på algoritmeaversjon i forbindelse med en medisinsk konsultasjon, og om aversjonen er svakere ved lav alvorlighetsgrad på symptomene. Våre funn gir derfor ny innsikt i litteraturen, siden det har gitt mer kunnskap om situasjoner der det kan foreligge algoritmeaversjon innen helse. Vi har ellers ikke funnet at det eksisterer forskning på algoritmeaversjon ved varierende alvorlighetsgrad innenfor helse. Ved å undersøke dette får vi varierende resultater. Dermed kan vi ikke generelt fastslå at alvorlighetsgrad påvirker graden av algoritmeaversjon.

Imidlertid synes det å være en sammenheng. Derfor er dette også bidrag til ny innsikt i litteraturen, men at ytterligere studier er nødvendig. Funnene våre er noe fremtidige oppgaver og prosjekter kan dra nytte av for videre forskning.

Vi har tatt utgangspunkt i studier som har benyttet valg som indikasjon på algoritmeaversjon, samt kritikken Longoni et al. (2019) sitt studie fikk av Pezzo og Beckstead (2020), der valg ikke ble benyttet. Av den grunn kunne deltakerne velge behandlingsmetode selv. Her fikk vi signifikante resultater. Dermed styrker funnene våre tidligere funn i litteraturen. Mer presist at det foreligger indikasjoner på algoritmeaversjon ved valg av behandlingsmetode mellom et menneske og AI-teknologi. Derfor er dette en velfungerende måleindikator på algoritmeaversjon (Jussupow et al., 2020).

Tilsvarende styrker også funnene våre tidligere funn i litteraturen, som viser at det foreligger indikasjon på algoritmeaversjon når tilliten en har til menneske og robot sammenlignes (Madhaven & Wiegmann, 2007; Önkal et al., 2009). Vedrørende tilfredshet var funnene så vidt ikke signifikant hva gjelder forskjell mellom fastlege og robot, noe som dermed kan innebære at tilfredshet i utgangspunktet kan fungere godt som en indikator på algoritmeaversjon (Jussupow et al., 2020). Imidlertid krever dette videre forskning.

7.2.2 Praktiske implikasjoner

Vi har funnet at det foreligger indikasjon på algoritmeaversjon tilknyttet valg av behandlingsmetode ved en medisinsk konsultasjon, der deltakerne kan velge mellom fastlege og robot. I tillegg viser resultatene at alvorlighetsgrad påvirker graden av algoritmeaversjon, ved at flere velger robot når alvorlighetsgraden er lav. Informasjonen kan være relevant ved implementering av AI-teknologi i fastlegeordningen, og helsesektoren som sådan. Ved en eventuell implementering av robot, kan en først introdusere løsningen ved sykdom som er av lav alvorlighetsgrad, siden det er flere som vil benytte denne tjenesten. Dette er viktig fordi en må ha pasienter som er villig til å benytte seg av tjenesten for å kunne tilby det.

Videre er det viktig at opplevelsen med roboten er god, i form av at behandlingsformen scorer høyt nok på andre konsepter som måler algoritmeaversjon, som for eksempel tillit og tilfredshet. Det at tillit har en medierende effekt på tilfredshet kan derfor være en faktor å ta med i vurderingen i forbindelse med bruk av teknologiske tjenester. Derfor kan det være

aktuelt å ta hensyn til dette ved å øke tilliten pasientene har til tjenesten, dersom en skal innføre AI-teknologi i medisinsk konsultasjon. Ved å øke tilfredsheten kan en derfor også øke bruken av tjenesten. Det er fordi en er avhengig av at pasienter er villig til å benytte seg av roboten fremfor menneske gjentatte ganger.

Dersom en klarer å få pasienter til å benytte seg av roboten flere ganger vil de på denne måten bli vant med teknologien, samt bli trygge på at det er et minst like godt alternativ som en menneskelig behandler. På sikt kan dette føre til lavere algoritmeaversjon, og eventuelt at algoritmeaversjon til slutt ikke er tilstedeværende. Hvis en lykkes med dette kan en etterhvert tilby tjenester i tilfeller med høyere alvorlighetsgrad.

Tilsvarende strategi kan benyttes relatert til andre områder innenfor helse, eksempelvis i forbindelse med legevakttjenester, samt i forbindelse med diagnostisering og behandling av psykiske lidelser. Sistnevnte vil i første omgang være best egnet for lettere psykiske lidelser. Her kan en robot foreta diagnostisering, for deretter å henvise videre til et passende digitalt selvhjelpsprogram. Slike digitale hjelpemidler er noe som eksisterer per i dag og utvikles kontinuerlig (DigPsyk, 2021; iFightDepression, 2021; Heradstveit, 2020). For tyngre psykiske lidelser vil roboten trolig kunne fungere som et støtteverktøy i tillegg til psykolog, både hva gjelder diagnostisering og behandling.

7.2.3 Anbefalinger til videre forskning

For videre forskning anbefaler vi først og fremst å gjennomføre vårt eksperiment i en større skala. På denne måten kan en generalisere og spesifisere til en større andel av befolkningen. I den sammenheng anbefales det også å ha et mer variert utvalg, hva gjelder kjønn og alder, der det også kontrolleres for teknologikompetanse, grad av nevrotisisme, risikovillighet og tillitsfullhet. Dermed kan en finne resultater tilknyttet grupper av disse variablene, samt generalisere på tvers av dem. Vi foreslår også at det inkluderes flere spørsmål for å kunne måle hvert konsept bedre. I tillegg bør det utarbeides beskrivelse av symptomer som gir signifikant resultat på forskjell mellom hvordan høy og lav alvorlighetsgrad oppfattes.

Det anbefales også å gjennomføre eksperimentet i en mer reell situasjon, der deltakerne møter opp og får bistand fra en fastlege eller en robot. Dette kan bidra til å redusere eventuelle «bias» ved at deltakerne ikke klarer å forestille seg den konkrete situasjonen. Det er også

interessant å se om en får tilsvarende funn ved andre helserelaterte situasjoner, der pasienten er direkte involvert. Dette kan eksempelvis være tilknyttet oppfølgingssamtaler.

Videre anbefales det å gjennomføre en kvalitativt studie for å undersøke hvordan pasientene tenker i forbindelse med beslutningstaking og vurdering av behandlingsformen. På denne måten kan en få en rikere innsikt i tiltak som kan foretas for å gjøre pasienter mer villig til å benytte seg av en robot. Fra litteraturen fremmes det flere momenter som kan inkluderes for å redusere algoritmeaversjon. Blant tiltakene er at en kan inkludere informasjon som viser at roboten klarer å vurdere din unike helsesituasjon, for å unngå unikhetsforsømmelse (Longoni et al, 2019). Et annet moment er at en kan gjøre fastlegen til endelig beslutningstaker, som stemmer overens med våre funn i litteraturen (Longoni et al., 2019). Videre kan en også for eksempel sørge for at oppgaven roboten utfører fremstår som mer objektiv. Dette begrunnes med at AI-teknologi er ansett for å være dyktig til å utføre denne type oppgaver (Castelo et al., 2019). På bakgrunn av dette kan det derfor være nyttig å få innsikt i hvilke konkrete tiltak som kan redusere algoritmeaversjonen, og dermed øke villigheten til å benytte AI-teknologi.

8 Litteraturliste

Abelia. (2017). Teknologi og digitalisering. Hentet fra: https://www.abelia.no/politikk/omstillingsbarometeret/teknologi-ogdigitalisering/#digitalisering-av-offentlige-tjenester

Accenture. (2016). *Vårt nye digitale Norge.* Hentet fra: https://www.accenture.com/t20170920t205302z__w__/no-en/_acnmedia/pdf-59/accenture-digitale-norge-2.pdf

Achinstein, P. (2001). *The Book of Evidence*. Oxford: Oxford University Press.

Alaszewski, A. (2003). Risk, Trust and Health. *Health, Risk & Society, 5:3*, 235-239. 10.1080/1369857031000160694l

Anderson, J. C. & Narus J. A (1991). Partnering as a Focused Market Strategy. *California Management Review. 33 (Spring)*, 95 - 113

Anderssen, H. (2019). AI vil revolusjonere helsevesenet - men Norge ligger langt bak. *Healthtalk.* Hentet fra: https://www.healthtalk.no/alle-artikler/ai-vil-revolusjonere-helsevesenet/

Arkes, H. R., & Blumer, C. (1985). The psychology of sunk costs. *Organizational Behavior and Human Decision Processes, 35*, 124-140.

Asan, O., Bayrak, A. E. & Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: focus on clinicians. *J Med Internet Res., 2020;22:e15154.* 10.2196/15154

Aslaksen, P. (2019). Rasjonalisering (psykologi). I *Store Norske Leksikon*. Hentet fra: https://snl.no/rasjonalisering_-_psykologi

Barr, A. & Packard, T. (2002). Revealed Preference and Self-Insurance: Can We Learn from the Self-Employed in Chile? *World Bank Policy Research Working Paper No. 2754.*

Baumeister, R. F. & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117(3)*, 497.

Beck, H. P., McKinney, J. B., Dzindolet, M. T. & Pierce, L. G. (2009). Effects of human-machine competition on intent errors in a target detection task. *Human Factors, 51(4)*, 477–486.

Bem, D. J. (1972). Self-perception theory. *In L. Berkowitz (Ed.), Advances in Experimental Social Psychology*, 1-62. New York: Academic Press.

Bhattacherjee, A. & Premkumar, G. (2004). Understanding Changes in Belief and Attitude toward Information Technology Usage: A Theoretical Model and Longitudinal Test. *MIS Quarterly, 28(2)*, 229–254

Bigman, Y. E. & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21–34.

Boon, S. & Holmes, J. (1991). The Dynamics of Interpersonal Trust: Resolving Uncertainity in the Face of Risk. *Cooperation and Prosocial Behaviour, Cambridge University Press,* 190-211.

Brewer, N. T., DeFrank, J. T. & Gilkey, M. B. (2016). Anticipated Regret and Health Behaviour: A Meta-Analysis. *Health Psychology, 2016, Vol. 35, No. 11,* 1264 –1275. http://dx.doi.org/10.1037/hea0000294

Burton, J. W., Stein, M. & Jensen, T. B. (2019). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 1–20.

Camerer, C. F. & Hogarth, R. M. (1999). The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. *Journal of Risk and Uncertainty, 19(1),* 7–42.

Carbone, R., Andersen, A., Corriveau, Y. & Corson, P. P. (1983). Comparing for different time series methods the value of technical expertise individualized analysis, and judgmental adjustment. *Management Sci. 29(5),* 559–566.

Castelo, N., Bos, M. W. & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, *56*(5), 809–825.

Charness, G. & Gneezy, U. (2012). Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization, 83(1),* 50–58. https://doi.org/10.1016/j.jebo.2011.06.007

Ciavolino, E. & Dahlgaard, J. J. (2007). ECSI - Customer Satisfaction Modelling and Analysis: A Case Study. *Total Quality Management and Business Excellence.* 10.1080/14783360701240337

Dahlum, S. (2018). Validitet. I *Store Norske Leksikon*. Hentet fra: https://snl.no/validitet

Davenport, T. & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal, Vol 6, No 2*, 94–98.

Dawes, R. M. (1979). The Robust Beauty of Improper Linear Models in Decision Making. *American Psychologist, Vol. 34, No. 7,* 571-582

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668-1674.

Deci, E. L., & Ryan, R. M. (2008). Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian Psychology, 49(3),* 182.

DeCoster, J. (1998). Overview of Factor Analysis. Hentet fra: http://www.stat-help.com/notes.html

Deepak, S., Jagdip Singh, J. & Sabol, B. (2002). Consumer Trust, Value, and Loyalty in Relational Exchanges. *Journal of Marketing, Vol. 66*, 15-37.

Deloitte. (2021). *Deloitte Tech Trends 2021*. Hentet fra: https://www2.deloitte.com/no/no/pages/technology-media-and-telecommunications/articles/Deloitte-tech-trends-2021.html

Dietvorst, B., Simmons, J. P. & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126.

Dietvorst, B., Simmons, J. P. & Massey, C. (2018). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science*, *64*(3), 1155–1170.

Difi. (2016). *Status digitalisering i offentlig sektor.* Hentet fra: https://www.difi.no/rapporter-og-statistikk/undersokelser/kunnskapsgrunnlag-om-digitalisering/status-digitalisering-i-offentlig-sektor

DigPsyk. (2021). *Om DigPsyk.* Hentet fra: https://digpsyk.no/om-digpsyk

Dimitrakos, T. (2002). A Service-Oriented Trust Management Framework. In R. *Trust, Reputation, and Security: Theories and Practice*, 53–72. Springer.

Dinno, A. (2015). Nonparametric Pairwise Multiple Comparisons in Independent Groups using Dunn's Test. *The Stata Journal*, *15(1)*, 292–300. https://doi.org/10.1177/1536867X1501500117

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J. & Wagner, G. G. (2009). Individual Risk Attitudes: Measurement, Determinants and Behavioral Consequences. *ROA Research Memorandum, 2009/6.*

Dommerud (2020). Slik skal fastlegekrisen løses. *Aftenposten.* Hentet fra: https://www.aftenposten.no/norge/i/zG2zJO/slik-skal-fastlegekrisen-loeses

Donnelly, L. (2017). Forget your GP, robots will soon be able to diagnose more accurately than almost any doctor. *The Telegraph.*

Dujmovic, J. (2017). Opinion: What's holding back artificial intelligence? Americans don't trust it. Hentet fra: https://www.marketwatch.com/story/whats-holding-back-artificial-intelligence-americans-dont-trust-it-2017-03-30

Dvergsdal, H. (2019). Nevralt nettverk. I *Store Norske Leksikon.* Hentet fra: https://snl.no/nevralt_nettverk

Dyer, T., Owens, J. & Robinson, P. (2016) The acceptability of healthcare: from satisfaction to trust. *Community. Dent Health*, *33*,1–10.

Dzindolet, M. T., Pierce, L. G., Beck, H. P. & Dawe, L. A. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors*, *44*(1), 79–94.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58,* 697–718.

Dzubian, C. D. & Shirkey, E. C. (1974). When is a correlation matrix appropriate for factor analysis? *Psychological Bulletin, 81(6),* 358–361

Eckel, C., & Grossman, P. (2007). Men, Women and Risk Aversion: Experimental Evidence. *Handbook of Experimental Results.* Elsevier Science.

Eckel, C., Johnson, C. & Montmarquette, C. (2005). Saving Decisions of the Working Poor: Short- and Long-Term Horizons. *Research in Experimental Economics, Vol. 10,* 219–260. Elsevier Science.

Falcone, F. & Castelfranchi, C. (2001) Trust and deception in virtual societies, 55-90..

Featherman, M. S., & Pavlou, P. A. (2003). Predicting e-services adoption: A perceived risk facets perspective.

Feinberg, R. M. (1977). Risk-aversion, Risk and the Duration of Unemployment. *Review of Economics and Statistics, 59(3),* 264–271.

Fischer, G. (2001). User Modeling in Human-Computer Interaction. *User Modeling and User-Adapted Interaction 11,* 65-86.

Folkehelseinstituttet (FHI). (2018). *Utbredelse av røyking i Norge*. Hentet fra: https://www.fhi.no/nettpub/tobakkinorge/bruk-av-tobakk/utbredelse-av-royking-i-norge/

Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in Motion Reveals Mind in Motion. *Frontiers in Psychology (2:59),* 1-6.

Garb, H. N. (1998). Studying the clinician. *American Psychological Association.*

Giddens, A. (1990) The Consequences of Modernity. Cambridge: Polity 1991. Modernity and Self-Identity: Self and Society in the Late Modern Age.

Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B. & Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Comput. Hum. Behav. 2021,* 115.

Goodwin P. & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? *J. Behav. Decision Making 12(1),* 37–53.

Goodyear, K., Parasuraman, R., Chernyak, S., Madhavan, P., Deshpande, G. & F. Krueger. (2016). Advice Taking from Humans and Machines: An fMRI and Effective Connectivity Study. *Frontiers in Human Neuroscience, 10 (542).*

Google. (2015). *Compatibility Definition Android 6.0.* Google Inc.

Grandison, T. & Sloman, M. A. (2000) Survey of Trust in Internet Applications. *IEEE Communications Surveys, 3(4),* 2–16.

Grima, S., özen, E., Boz, H., Spteri, J. & Thalassinos, E. (2019). Contemporary Issues In Behavioral Finance. *Contemporary issues in economic and financial analysis, Vol. 101.*

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E. & Nelson, C. (2000). Clinical Versus Mechanical Prediction: A Meta-Analysis. *Psychological Assessment, Vol. 12, No. 1,* 19-30. 10.1037//1040-3590.12.1.19

Gulshan, V., Peng, L., & Coram, M. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Journal of American Medical Association, 316,* 2402–2410.

Gummerus, J., Liljander, V., Pura, M. &Van Riel, A. (2004). Customer Loyalty to Content-Based Web Sites: The case of an Online Health-Care Service. *Journal of Services Marketing. 18(3),* 175-186.

Grønmo, S. (2020). Kvantitativ metode. *I Store norske leksikon.* Hentet fra: https://snl.no/kvantitativ_metode

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F. T, Buhl, F. T., Blum, A., (…) & Uhlmann, L. (2018). Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology, 29 (8),* 1836–1842.

Hancock, P. A., Billings, D. R., & Schaefer, K. E. (2011). Can you trust your robot? *Ergonomics in Design, 19(3),* 24–29.

Harper, R., Rodden, T., Rogers, Y. & Sellen, A. (2008). *Being Human: Human-Computer Interaction in the year 2020.* Microsoft Corporation.

Hayes, A. F. (2016). *Mediation, moderation, and conditional process analysis.* Statistical Horizons.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach (Methodology in the social sciences).* Guilford Press.

Hayes, A. F. & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology, Vol. 67, Issue 3,* 451-470. https://doi.org/10.1111/bmsp.12028

Hehman, E., Stolier, R. M. & Freeman, J. B. (2014). Advanced Mouse-Tracking Analytic Techniques for Enhancing Psychological Science. *Psychological Science (20:10),* 1183-1188.

Heradstveit, O. (2021). *Youwell: Denne digitale plattformen gir nye muligheter for behandling innen psykisk helse.* Hjelptilhjelp.no. Hentet fra: https://www.hjelptilhjelp.no/Hjelpeapparatet/denne-digitale-plattformen-gir-nye-muligheter-for-behandling-innen-psykisk-helse

Hibbeln, M., Jenkins, J. L., Schneider, C., Valacich, J. S. & Weinmann, M. (2007). How is your user feeling? Inferring emotion through Human-Computer Interaction devices. *MIS Quarterly, Vol. 41, No. 1,* 1-21.

Hogarth, R. M. & Makridakis, S. (1981). Forecasting and planning: An evaluation. *Management Sci. 27(2),* 115–138.

Holt, C. A. & Laury, S. K. (2002). Risk Aversion and Incentive Effects. *American Economic Review, 92(5),* 1644–1655.

Hudlicka, E. (2003). To feel or not to feel: The role of affect in human–computer interaction. *Int. J. Human-Computer Studies 59 (2003),* 1–32. Elsevier Science Ltd.

Hutson, M. (2017). Self-taught artificial intelligence beats doctors at predicting heart attacks. Science. Hentet fra https://www.sciencemag.org/news/2017/04/self-taught-artificial-intelligence-beats-doctors-predicting-heart-attacks

iFightDepression. (2021). *iFightDepression: selvhjelpsverktøy mot depresjon.* Hentet fra: https://ifightdepression.com/no/

Intel. (2020). *Advancing the adoption of AI.* Hentet fra: https://newsroom.intel.com/wp-content/uploads/sites/11/2020/11/AdvancingAI-Infographic.pdf

Jacobsen, D. I. (2015). *Hvordan gjennomføre undersøkelser? Innføring i samfunnsvitenskapelig metode (3rd ed.)*. Cappelen Damm AS.

Jago, A. S. (2019). Algorithms and Authenticity. *Academy of Management Discoveries, 5(1),* 38–56

Jiang, F., Jiang, Y, Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q, Shen, H &

Jones, L. D., Golan, D., Hanna, S. A. & Ramachandran M. (2018). Artificial intelligence, machine learning and the evolution of healthcare. A bright future or cause for concern? *Bone Joint Res, Vol. 7,* 223–225. 10.1302/2046-3758.73.BJR2017-0147.R1

Jussupow, E., Benbasat, I. & Heinzl, A. (2020) Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. *Research Papers 168.*

Jøsang, A. & Prest, S. (2004). Analysing the Relationship between Risk and Trust. *International Conference on Trust Management, iTrust 2004: Trust Management,*135-145

Kadam, P. & Bhalerao, S. (2010). Sample size calculation. *US National Library of Medicine National Institutes of Health.* Hentet fra: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2876926/

Kahneman, D., Slovic, P., & Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases. Cambridge University Press.

Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39,* 31–36.

Kantar (2017). *Svekket tillit i en digital verden.* Hentet fra: https://kantar.no/kantar-tns-innsikt/svekket-tillit-i-en-digital-verden/

Karray, F., Alemzadeh, M., Saleh, J. A. & Arab, M. N. (2008). Human-Computer Interaction: Overview on State of the Art. *International Journal on smart sensing and intelligent systems, Vol. 1, No. 1.*

Keller, G. (2011). *Managerial Statistics (9th ed).* CENGAGE Learning Custom Publishing.

Kerasidou, A. (2019). Empathy, compassion and trust balancing artificial intelligence in health care. The Ethox Centre, Nuffield Department of Population Health, University of Oxford.

Kim, H. J. & Choi, Y. S. (2012). Exploring Emotional Preference for Smartphone Applications. *Proceedings of the Consumer Communications and Networking Conference,* 245-249

Kim, J. O., & Mueller, C. W. (1978a). Factor analysis: Statistical methods and practical issues. *Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-014.* Newbury Park, CA: Sage.

Kim, J. O., & Mueller, C. W. (1978b). Introduction to factor analysis: What it is and how to do it. *Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-013.* Newbury Park, CA: Sage.

Klein, J., Moon, Y., & Picard, R. W. (2002). This Computer Responds to User Frustration: Theory, Design, and Results. *Interacting with Computers (14:2),* 119-140.

Kline, P. (1994). *An easy guide to factor analysis*. Routeledge.

Korczynski, M. (2000). The political economy of trust. *Journal of Management Studies, 37(1),* 1–21. doi:10.1111/1467-6486.00170

Kramer, R. M., & Tyler, T. R. (Eds.). (1996). *Trust in organizations: Frontiers of theory and research.* Sage Publications Inc.

Kuo, Y. F., Yen, S. T., & Chen, L. H. (2011). Online Auction Service Failures in Taiwan: Typologies and Recovery Strategies. *Electronic Commerce Research and Applications (10:2),* 183-193.

Kvittingen, I. & Torgersen, E. (2019). *P-hacking: Slik fisker forskere etter resultater.* Forskning.no. Hentet fra: https://forskning.no/forskningsetikk-om-forskning-statistikk/p-hacking-slik-fisker-forskere-etter-resultater/1322270

Lacey, S., Bruwer, J., & Li, E. (2009). The role of perceived risk in wine purchase decisions in restaurants.

Larsen, R. Buss, D. & Wismeijer A. (2013). *Personality Psychology. Domains of Knowledge about Human Nature.* New York: McGraw-Hill.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*. https://doi.org/10.1177/2053951718756684

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46,* 50–80.

Leisen, B. & Hyman, M. R. (2001). An Improved Scale for Assessing Patients' Trust in Their Physician. *Health Marketing Quarterly, 19:1,* 23-42. 10.1300/J026v19n01_03

Levitt, S. D., & List, J. A. (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review, 53(1),* 1–18. https://doi.org/10.1016/j.euroecorev.2008.12.001

Lim, J. S. & O'Connor, M. (1995). Judgemental adjustment of initial forecasts: Its effectiveness and biases. *J. Behav. Decision Making 8(3),* 149–168.

Logg, J. M., Minson, J. A. & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151,* 90–103.

Lohr, S. (2016). IBM is counting on its bet on Watson, and paying big money for it. *The New York Times.* Hentet fra: https://www.nytimes.com/2016/10/17/technology/ibm-is-counting-on-its-bet-on-watson-and-paying-big-money-for-it.html

Longoni, C., Bonezzi, A. & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research.*

MacKinnon, D. P., Fairchild, A. J. & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology.* 10.1146/annurev.psych.58.110405.085542

Madhaven, P. & D. A. Wiegmann. (2007). Effects of Information Source, Pedigree, and Reliability on Operator Interaction With Decision Support Systems. *Human Factors, 49(5),* 773–785.

Manchala, D. W. (1998). Trust Metrics, Models and Protocols for Electronic Commerce Transactions. *In Proc. of the 18 International Conference on Distributed Computing Systems,* 312–321. IEEE Computer Society

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust, *Academy of Management Review, 20(3),* 709–734.

McCabe, B. (2017). Which Gender Is More Likely To Trust Artificial Intelligence. Hentet fra: https://www.linkedin.com/pulse/which-gender-more-likely-trust-artificial-bill-

mccabe/?lipi=urn%3Ali%3Apage%3Ad_flagship3_profile_view_base_post_details%3BlQzy2K9FTleafH5MBYgOCw%3D%3D

McKnight, D., Harrison, C. & Norman, L. (2000). What is Trust? A Conceptual Analysis and an Interdisciplinary Model.

Mehl, P. (1954). *Clinical Versus Statistical Prediction: A Theoretical Analysis and Review of the Evidence.* Echo Point Books & Media.

Mehmetoglu, M. & Mittner, M. (2020). *Innføring i R for statistiske dataanalyser.* Universitetsforlaget.

Mesko, B., Hetényi, G. & Győrffy, Z. (2018). Will artificial intelligence solve the human resource crisis in healthcare? *BMC Health Services Research, 18:545.* https://doi.org/10.1186/s12913-018-3359-4

Minsky, M. (1988). *The Society of Mind.* Simon & Schuster.

Mitchell, V. W. (1999). Consumer perceived risk. Conceptualisations and models. *European Journal of Marketing, Vol. 33 Nos. 1-2,* 163-95.

Mohtar, S. B., & Abbas, M. A. (2015). Consumer resistance to innovation due to perceived risk: Relationship between perceived risk and consumer resistance to innovation. *Journal of Technology and Operation Management, 10(1),* 1–13.

Newcomer, L. N. (1997). Measures of Trust in Health Care. *Health Affairs, 16 January/February,* 50–51.

Nickerson, J. V., & Reilly, R. R. (2004). A model for investigating the effects of machine autonomy on human behavior. *In Proceedings of the 37th Annual Hawaii International Conference on System Sciences,* 1–10.

Pallant, J. (2010). *SPSS Survival Manual (4th ed.).* Berkshire: McGraw-Hill Education.

Palmeira, M. & Spassova, G. (2015). Consumer reactions to professionals who use decision aids. *European Journal of Marketing*, 302–326.

Parasuraman, A., & Colby, C. L. (2015). An Updated and Streamlined Technology Readiness Index: TRI 2.0. *Journal of Service Research, 18*(1), 59–74. https://doi.org/10.1177/1094670514539730

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39,* 230–253.

Partick, A. (2002) Building Trustworthy Software Agents. *IEEE Internet Computing, 6(6),* 46–53.

Pezzo, M. V., & Beckstead, J. W. (2020). Patients prefer AI to humans, so long as the AI is better than the humans: A commentary on Longoni, Bonezzi, and Morewedge. *Judgment and Decision Making, 15 (3),* 449–451.

Pripp, A. H. (2018). *Validitet*. Tidsskriftet Den Norske Legeforening. Hentet fra: https://tidsskriftet.no/2018/09/medisin-og-tall/validitet

PWC. (2017). *Consumers trust your tech even less than you think.* Hentet fra: https://www.pwc.com/us/en/services/consulting/library/consumer-intelligence-series/trusted-tech.html

Regjeringen. (2020). *Nasjonal strategi for kunstig intelligens.* Hentet fra: https://www.regjeringen.no/no/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/?ch=3

Regjeringen. (2019) *Nasjonal helse- og sykehusplan.* Hentet fra: https://www.regjeringen.no/no/dokumenter/nasjonal-helse--og-sykehusplan-2020-2023/id2679013/

Ren, F. & Bao, Y. (2020). A Review on Human-Computer Interaction and Intelligent Robots. *International Journal of Information Technology & Decision Making, Vol. 19, No. 1,* 5–47. 10.1142/S0219622019300052

Riekeles, H. (2018). *Ta tak i sykelønnsordningen, Solberg!*. Civita. Hentet fra: https://www.civita.no/velferd-og-den-nordiske-modellen/ta-tak-i-sykelonnsordningen-solberg

Roselius, T., (1971). Consumer rankings of risk reduction methods. *The Journal of Marketing. 35(1),* 56-61.

Rummel, R. J. (1970). *Applied Factor Analysis.* Northwestern University Press.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. *In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction,* 141–148.

Sanders, T., Kaplan, A., Koch, R. & Schwartz, M. (2019). The Relationship Between Trust and Use Choice in Human-Robot Interaction. *HUMAN FACTORS, Vol. 61, No. 4, June 2019,* 614–626. 10.1177/0018720818816838

Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research Methods for Business Students (8th ed.).* Pearson Education Limited.

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors, 58,* 377–400.

Schubert, R., Brown, M., Gysler, M. & Brachinger, H. (1999). Financial Decision-Making: Are Women Really More Risk-Averse? *American Economic Review Papers and Proceedings, 89(2),* 381–385.

Sekhon, H., Ennew, C., Kharouf, H. & Devlin, J. (2014). Trustworthiness and trust: influences and implications. *Journal of Marketing Management, 30/(3-4),* 409-430.

Sharan, N. N., & Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon*, *6*(8). https://doi.org/10.1016/j.heliyon.2020.e04572

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological Methods, 7,* 422-445.

Siau, K., & Wang, W. (2018). Building Trust in Artificial Intelligence, Machine Learning, and Robotics. *Cutter business technology journal, 31(2),* 47-53. Cutter Consortium

Slovic, P. (1972a). Psychological Study of Human Judgment: Implications for Investment Decision Making. *Journal of Finance, 27,* 777–799.

Slovic, P. (1972b). Information Procession, Situation Specificity, and the Generality of Risk-Taking Behavior. *Journal of Personality and Social Psychology, 22,* 128–134.

Smith, A. K., Bolton, R. N., & Wagner, J. (1999). A Model of Customer Satisfaction with Service Encounters Involving Failure and Recovery. *Journal of Marketing Research (36:3),* 356-372.

Smith, E. J., & Holcombe, W. N. (2008). Age Distribution for Students Enrolled in College Preparatory Courses. Tallahassee, FL.

Staff, A. (2015). *Bias*. De nasjonale forskningsetiske komiteene. Hentet fra: https://www.forskningsetikk.no/ressurser/fbib/uavhengighet/bias/

Storvik, A. G. (2018). Kartlegging: Fastlegene jobber i snitt 55,6 timer i uken. *Dagens magasin.* Hentet fra: https://www.dagensmedisin.no/artikler/2018/03/06/kartlegging-fastlegene-jobber-i-snitt-556-timer-i-uken/

Stranden, A. L. (2017). *Snus er langt mindre farlig enn antatt.* Forskning.no. Hentet fra: https://forskning.no/sykdommer-royking-kreft/snus-er-langt-mindre-farlig-enn-antatt/311292

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size-or Why the P Value Is Not Enough. *Journal of Graduate Medical Education, 4(3),* 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Svartdal, F. (2020a). Faktoranalyse. I *Store Norske Leksikon.* Hentet fra: https://snl.no/faktoranalyse

Svartdal, F. (2020b). Reliabilitet. *I Store Norske Leksikon.* Hentet fra: https://snl.no/reliabilitet

Taylor, S. A. & Hunter, G. (2003). An exploratory investigation into the antecedents of satisfaction, brand attitude, and loyalty within the (B2B) eCRM industry. *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behaviour, 16, 19.*

Tullberg, J. (2007). Trust - The importance of trustfulness versus trustworthiness. 37(5), 2059-2071.

Tzeng, J. Y. (2004). Toward a More Civilized Design: Studying the Effects of Computers that Apologize. *International Journal of Human–Computer Studies (61:3),* 319-345.

Ursin, L. (2021.) Pasientautonomi. I *Store Medisinske Leksikon*. Hentet fra: https://sml.snl.no/pasientautonomi

Vayena, E., Blasimme A. & Cohen, G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Med 15(11).* https://doi.org/10.1371/journal.pmed.1002689

Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology.* 10.1136/svn-2017-000101

Weber, E. U., Blais, A. R. & Betz, N. E. (2002). A Domain-Specific RiskAttitude Scale: Measuring Risk Perceptions and Risk Behaviors. *Journal of Behavioral Decision Making, 15,* 263–290.

Willemain, T. R. (1991). The effect of graphical adjustment on forecast accuracy. *Internat. J. Forecasting 7(2),* 151–154.

Yeomans, M., Shah, A., Mullainathan, S. & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414.

Zheng, S., Hui, S. F. & Yang, Z. (2017) Hospital trust or doctor trust? A fuzzy analysis of trust in the health care setting. *Journal of Business Research, 78,* 217–25. https://doi.org/10.1016/j.jbusres.2016.12.017.

Önkal, D., Goodwin, P., Thomson, M., Gönül, S. & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, *22*(4), 390–409.

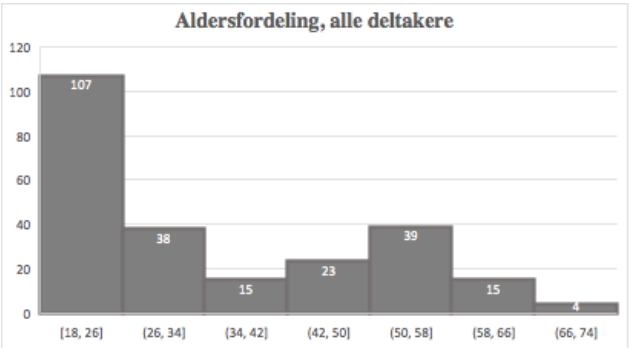
9 Appendix

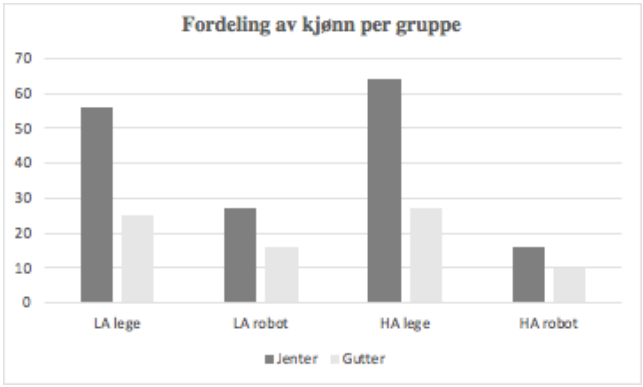
Appendix 1: Oversikt kontrollvariabler

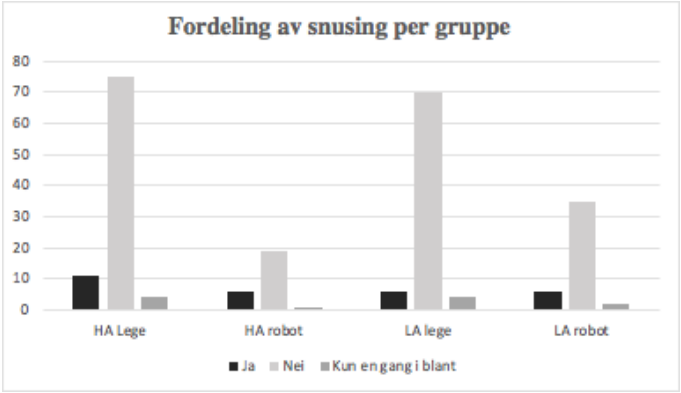
Område	Variabel	Målenivå	Verdier
Personlig informasjon	Alder	Nominal	18 ≤
	Kjønn	Nominal	Mann, Kvinne, Annet
Tilbøylighet til tillit	Egenvudering tillitsfull	Ordinal	Fempunkts Likert-skala*
Anlegg for nevrotisisme	Egenvurdering negative følelser	Ordinal	Fempunkts Likert-skala*
	Egenvurdering hjelpesløshet	Ordinal	Fempunkts Likert-skala*
Risikovillighet	Egenvurdering risikovillighet	Ordinal	Fempunkts Likert-skala*
	Helserelatert risiko	Nominal	Ja, nei, kun en gang iblant
Teknologikompetanse	Egenvurdering teknologikompetanse	Ordinal	Fempunkts Likert-skala*
	Holdning til benyttelse av ny teknologi	Ordinal	Fempunkts Likert-skala*

*verdier: Svært uenig - Svært enig

Appendix 2: Histogram med fordeling av alder, kjønn og snus







Appendix 3: Deskriptiv statistikk medierende og avhengige variabler

	N	Mean	SD	Skewness	SE of Skewness	Kurtosis	SE of Kurtosis
Medierende variabler							
Robot_vil deg vel	239	5,27	1,64	-0,81	0.157	0,08	0.312
Robot_korrekt diagnose	239	5,55	1,39	-1,48	0.157	2,64	0.312
Robot_rette egenskaper	239	5,42	1,38	-1,02	0.157	1,07	0.312
Robot_anger behandlingsform	239	5,42	1,51	-1,11	0.157	0,98	0.312
Robot_unik helsesituasjon	239	5,05	1,38	-1,46	0.157	-0,27	0.312
Robot_konsekvenser feil diagnose	239	4,36	1,58	-0,33	0.157	-0,78	0.312
Robot_uro ifm diagnostisering	239	2,72	1,43	0,51	0.157	-0,31	0.312
Lege_vil deg vel	239	5,73	1,13	-0,64	0.157	-0,37	0.312
Lege_korrekt diagnose	239	5,31	1,21	-0,46	0.157	-0,33	0.312
Lege_rette egenskaper	239	5,79	1,14	-0,99	0.157	0,5	0.312
Lege_anger behandlingsform	239	5,76	1,21	-0,96	0.157	0,42	0.312
Lege_unike helsesituasjon	239	5,47	1,13	-0,65	0.157	0,08	0.312
Lege_konsekvenser feil diagnose	239	4,59	1,69	-0,26	0.157	-0,92	0.312
Lege_uro ifm diagnostisering	239	3,12	1,55	0,35	0.157	-0,9	0.312
Avhengige variabler							
Valg_Fastlege_Robot	239	0,28	0,45	0.92	0.157	-1,16	0.312
Lege_fornøyd med behandling	239	5,36	1,30	-0,56	0.157	-0,28	0.312
Robot_fornøyd med behandling	239	5,68	1,29	-1,27	0.157	2,06	0.312
Lege_anbefaling behandlingsform	239	5,06	1,36	-0,48	0.157	-0,14	0.312
Robot_anbefaling behandlingsform	239	5,42	1,74	-1,12	0.157	0,99	0.312
Lege_ anger behandlingsform	239	2,06	1,41	1,42	0.157	1,33	0.312
Robot_anger behandlingsform	239	1,96	1,34	0,89	0.157	-1,03	0.312

Appendix 4: Dunn´s test

Dunn´s post-hoc test - generell risikovillighet

	Comparison	Z	P.unadj	P.adj
1	HA Lege - HA robot	-3.09504382	0.001967838	0.01180703
2	HA Lege - LA lege	-0.06654755	0.946941899	1.00000000
3	HA robot - LA lege	3.00834252	0.002626769	0.01576061
4	HA Lege - LA robot	-1.25592496	0.209143177	1.00000000
5	HA robot - LA robot	1.83491023	0.066518991	0.39911395
6	LA lege - LA robot	-1.17788628	0.238841944	1.00000000

Dunn´s post-hoc test - teknologikompetanse

	Comparison	Z	P.unadj	P.adj
1	HA Lege - HA robot	-2.2586718	0.0244063340	0.146438004
2	HA Lege - LA lege	-0.3632645	0.7164073192	1.000000000
3	HA robot - LA lege	1.9742346	0.0483550712	0.290130427
4	HA Lege - LA robot	-3.6229533	0.0002912584	0.001747551
5	HA robot - LA robot	-0.6840878	0.4939197264	1.000000000
6	LA lege - LA robot	-3.2591563	0.0011174410	0.006704646

Dunn´s post-hoc test - påstand om å erstatte mennesker med teknologi

	Comparison	Z	P.unadj	P.adj
1	HA Lege - HA robot	-5.4902148	4.014451e-08	2.408671e-07
2	HA Lege - LA lege	0.8202138	4.120942e-01	1.000000e+00
3	HA robot - LA lege	5.9722782	2.339631e-09	1.403778e-08
4	HA Lege - LA robot	-4.5492135	5.384680e-06	3.230808e-05
5	HA robot - LA robot	1.5257317	1.270767e-01	7.624601e-01
6	LA lege - LA robot	-5.1257327	2.963829e-07	1.778297e-06

Appendix 5: Korrelasjonsmatrise

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Kjønn	1,00													
2. Alder	0,07	1.00												
3. Snus	0,05	0,02	1,00											
4. Alvorlighetsgrad_Høy	0,03	0,09	**0,19***	1,00										
5. Valg_Høy	-0,09	-0,14	-0,11	**-0,19***	1,00									
6. Tilfredshet_Snitt	0,04	-0,12	-0,01	-0,11	**0,20***	1,00								
7. Forventet_Anger	0,07	-0,02	0,08	0,10	-0,06	**-0,13***	1,00							
8. Tillit_Snitt	0,02	-0,08	0,04	-0,08	-0,04	**0,63***	-0,05	1,00						
9. Risiko_Snitt	**0,21***	-0,04	0,06	**0,28***	-0,14	**-0,25***	**0,13***	**-0,16***	1,00					
10. Nevrotisisme_Snitt	**0,29***	**-0,17***	0,04	-0,03	-0,05	-0,03	0,00	-0,07	**0,22***	1,00				
11. Tillitsfull	-0,07	-0,05	0,00	0,02	0,14	-0,06	-0,05	0,05	-0,10	**-0,15***	1,00			
12. Risikovillighet	0,03	0,00	0,11	**0,20***	0,16	0,01	-0,02	0,00	0,09	-0,04	0,01	1,00		
13. Alvorlighetsgrad_Lav	0,05	0,08	0,07	NA	NA	0,12	0,05	0,07	**0,26***	0,15	-0,02	-0,05	1,00	
14. Valg_Lav	-0,07	0,04	-0,09	NA	NA	0,01	0,07	**-0,28***	-0,03	-0,05	0,06	-0,05	-0,16	1,00

Appendix 6: Analyser av kjikvadrat-test for valg av behandlingsform, H1 og H1a

Alvorlighetsgrad * Hvem vil du skal gjennomføre konsultasjonen? Crosstabulation

Count

		Alvorlighetsgrad		
		Høy alvorlighetsgrad	Lav alvorlighetsgrad	Total
Hvem vil du skal gjennomføre konsultasjonen?	Fastlege	90	79	169
	Robot	26	43	69
Total		116	122	238

Chi-Square Test - Robot

	Value	df	Sig.
Chi-Square	4.188	1	0.041
N of Valid Cases	239		

Chi-Square Test - Valg

	Value	df	Sig.
Chi-Square	42.017	1	0.000
N of Valid Cases	239		

Appendix 7: ANOVA-analyser tillit, H2 og H2a

Levene's Test for Homogenity of Variance

	Df	F	Sig.
Tillit	3	3,498	0,016

Welch's ANOVA

		F	df1	df2	Sig.
Tillit	Welch	3,431	3	85,504	0,021

ANOVA

		Sum of squares	Df	Mean Square	F	Sig.
Tillit	Between groups	14,610	3	4,871	4,207	0,006
	Within groups	272,090	235	1,158		
	Total	286,700	238			

Contrast Coefficients

	Condition			
Contrast	HA fastlege	HA robot	LA fastlege	LA robot
1	1	-1	1	-1
2	1	-1	0	0
3	1	0	0	-1
4	0	1	-1	0
5	0	0	1	-1
6	0	1	0	-1
7	1	0	-1	0

Contrast Test

		Contrast	Sum of squares	Df	Mean Square	F	Sig.
Tillit	Does not assume equal variance	1	9,190	1	9,194	7,941	0,005
		2	0,020	1	0,021	0,019	0,892
		3	2,420	1	2,418	2,088	0,0150
		4	6,760	1	7,763	5,841	0,016
		5	14,350	1	14,355	12,398	0,001
		6	3,700	1	3,702	3,198	0,075
		7	3,320	1	3,321	2,868	0,092

Appendix 8: ANOVA-analyser oppfattet risiko, H3 og H3a

Levene's Test for Homogenity of Variance

	Df	F	Sig.
Oppfattet risiko	3	0.493	0,686

ANOVA

		Sum of squares	Df	Mean Square	F	Sig.
Oppfattet risiko	Between groups	27,900	3	9,315	5,628	0,000
	Within groups	389,000	235	1,655		
	Total	416,900	238			

Contrast Coefficients

Contrast	Condition			
	HA fastlege	HA robot	LA fastlege	LA robot
1	1	-1	1	-1
2	1	-1	0	0
3	1	0	0	-1
4	0	1	-1	0
5	0	0	1	-1
6	0	1	0	-1
7	-1	0	1	0

Contrast Test

	Contrast	Sum of squares	Df	Mean Square	F	Sig.
Oppfattet risiko Assume equal variance	1	4,500	1	4,523	2,733	0,099
	2	13,800	1	13,771	8,320	0,004
	3	23,000	1	22,962	13,873	0,000
	4	5,900	1	5,892	3,560	0,060
	5	0,400	1	0,430	0,260	0,610
	6	0,400	1	3,972	2,400	0,122
	7	21,500	1	21,517	13,000	0,000

Appendix 9: ANOVA-analyser tilfredshet, H4 og H4a

Levene's Test for Homogenity of Variance

	Df	F	Sig.
Tilfredshet	3	0,759	0,518

ANOVA

		Sum of squares	Df	Mean Square	F	Sig.
Tilfredshet	Between groups	26,200	3	8,725	5,954	0,001
	Within groups	344,400	235	1,465		
	Total	370,600	238			

Contrast Coefficients

Contrast	Condition: HA fastlege	HA robot	LA fastlege	LA robot
1	1	-1	1	-1
2	1	-1	0	0
3	1	0	0	-1
4	0	1	-1	0
5	0	0	1	-1
6	0	1	0	-1
7	1	0	-1	0

Contrast Test

	Contrast	Sum of squares	Df	Mean Square	F	Sig.
Tilfredshet Assume equal variance	1	5,500	1	5,541	3,782	0,053
	2	17,800	1	17,800	12,148	0,001
	3	20,700	1	20,730	14,147	0,000
	4	3,300	1	3,328	2,271	0,133
	5	0,700	1	0,686	0,468	0,495
	6	1,100	1	1,117	0,762	0,384
	7	21,000	1	20,959	14,303	0,000

Appendix 10: ANOVA-analyser forventet anger, H5 og H5a

Levene's Test for Homogenity of Variance

	Df	F	Sig.
Forventet anger	4	0.361	0.836

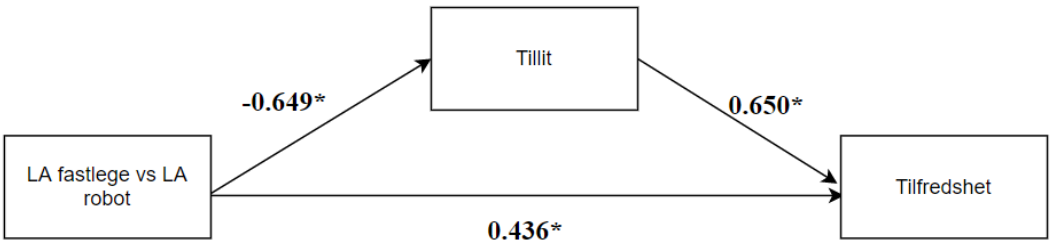
ANOVA

		Sum of squares	Df	Mean Square	F	Sig.
Forventet anger	Between groups	2,500	4	0,628	0,339	0,851
	Within groups	0,000	234	0,000		
	Total	2,5	238			

Appendix 11: PROCESS-resultater tilfredshet og forventet anger

Tilfredshet

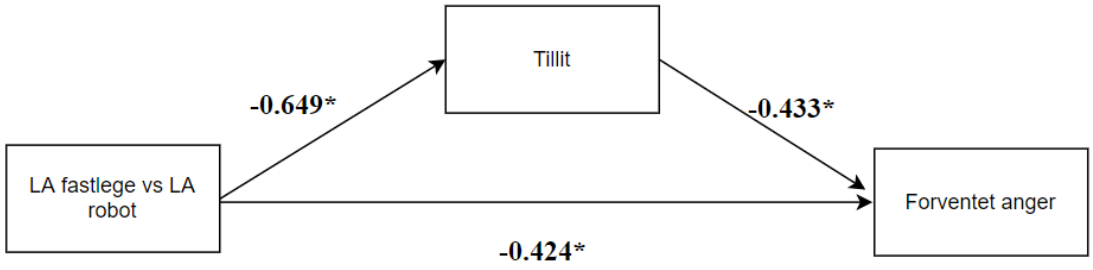
	β	t	p	Bootstrapping 95% CI	
				Lower	Upper
Effekt c					
Fastlege vs robot	0.269	1.897	0.059		
HA fastlege vs LA fastlege	**0.555***	3.722	0.000		
HA fastlege vs HA robot	**0.459***	2.125	0.035		
HA fastlege vs LA robot	**0.574***	3.190	0.002		
LA fastlege vs HA robot	-0.096	-0.438	0.662		
LA fastlege vs LA robot	0.019	0.102	0.919		
HA robot vs LA robot	0.115	0.476	0.635		
Effekt a					
Fastlege vs robot	**-0.394***	-2.802	0.006	-0.771	-0.103
HA fastlege vs LA fastlege	0.269	1.784	0.076	-0.013	0.591
HA fastlege vs HA robot	-0.082	-0.376	0.707	-0.498	0.319
HA fastlege vs LA robot	**-0.380***	-2.093	0.038	-0.911	0.046
LA fastlege vs HA robot	-0.351	-1.585	0.114	-0.797	0.032
LA fastlege vs LA robot	**-0.649***	-3.501	0.001	-1.213	-0.244
HA robot vs LA robot	-0.298	-1.225	0.222	-0.905	0.227
Effekt b					
Fastlege vs robot, Tillit → Tilfredshet	**0.653***	13.286	0.000	0.614	0.853
Valg og alvorlighetsgrad, Tillit → Tilfredshet	**0.650***	13.373	0.000	0.612	0.851
Effekt c'					
Fastlege vs robot	**0.500***	4.648	0.000	0.379	0.873
HA fastlege vs LA fastlege	**0.338***	2.960	0.003	0.135	0.715
HA fastlege vs HA robot	**0.488***	3.032	0.003	0.224	1.016
HA fastlege vs LA robot	**0.774***	5.583	0.000	0.649	1.284
LA fastlege vs HA robot	0.151	0.921	0.358	-0.207	0.595
LA fastlege vs LA robot	**0.436***	3.120	0.002	0.234	0.854
HA robot vs LA robot	0.286	1.586	0.114	-0.082	0.778



	Tilfredshet				
	β	t	p	Bootstrapping 95% CI	
				Lower	Upper
Effekt c					
Fastlege vs robot	0.269	1.897	0.059		
HA fastlege vs LA fastlege	**0.555***	3.722	0.000		
HA fastlege vs HA robot	**0.459***	2.125	0.035		
HA fastlege vs LA robot	**0.574***	3.190	0.002		
LA fastlege vs HA robot	-0.096	-0.438	0.662		
LA fastlege vs LA robot	0.019	0.102	0.919		
HA robot vs LA robot	0.115	0.476	0.635		
Effekt a					
Fastlege vs robot	-0.236	-1.659	0.099	-0.672	0.045
HA fastlege vs LA fastlege	**-0.552***	-3.706	0.000	-1.119	-0.345
HA fastlege vs HA robot	-0.311	-1.440	0.151	-0.896	0.077
HA fastlege vs LA robot	**-0.608***	-3.380	0.001	-1.289	-0.322
LA fastlege vs HA robot	0.241	1.103	0.271	-0.180	0.813
LA fastlege vs LA robot	-0.055	-0.302	0.763	-0.560	0.427
HA robot vs LA robot	-0.297	-1.232	0.219	-0.966	0.186
Effekt b					
Fastlege vs robot, Oppfattet risiko → Tilfredshet	**-0.116***	-2.381	0.018	-0.196	-0.027
Valg og alvorlighetsgrad, Oppfattet risiko → Tilfredshet	-0.077	-1.569	0.118	-0.163	0.016
Effekt c'					
Fastlege vs robot	**0.500***	4.648	0.000	0.379	0.873
HA fastlege vs LA fastlege	**0.338***	2.960	0.003	0.135	0.715
HA fastlege vs HA robot	**0.488***	3.032	0.003	0.224	1.016
HA fastlege vs LA robot	**0.774***	5.583	0.000	0.649	1.284
LA fastlege vs HA robot	0.151	0.921	0.358	-0.207	0.595
LA fastlege vs LA robot	**0.436***	3.120	0.002	0.234	0.854
HA robot vs LA robot	0.286	1.586	0.114	-0.082	0.778

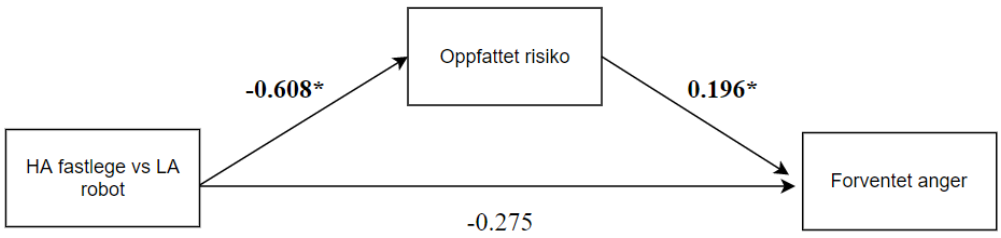
Forventet anger

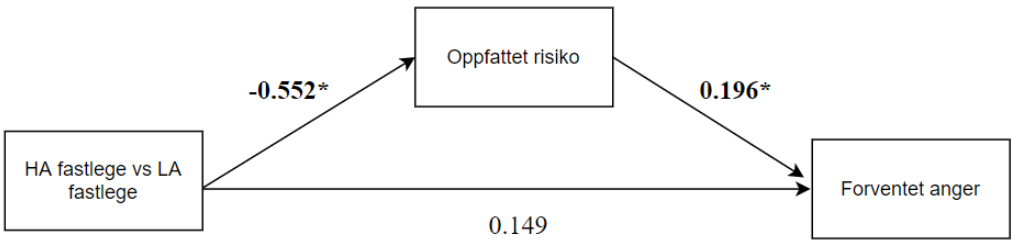
	β	t	p	Bootstrapping 95% CI	
				Lower	Upper
Effekt c					
Fastlege vs robot	-0.140	-0.978	0.329		
HA fastlege vs LA fastlege	-0.076	-0.492	0.623		
HA fastlege vs HA robot	-0.086	-0.385	0.701		
HA fastlege vs LA robot	-0.229	-1.233	0.219		
LA fastlege vs HA robot	-0.010	-0.044	0.965		
LA fastlege vs LA robot	-0.153	-0.809	0.420		
HA robot vs LA robot	-0.143	-0.576	0.566		
Effekt a					
Fastlege vs robot	**-0.394***	-2.802	0.006	-0.771	-0.103
HA fastlege vs LA fastlege	0.269	1.784	0.076	-0.013	0.591
HA fastlege vs HA robot	-0.082	-0.376	0.707	-0.498	0.319
HA fastlege vs LA robot	**-0.380***	-2.093	0.038	-0.911	0.046
LA fastlege vs HA robot	-0.351	-1.585	0.114	-0.797	0.032
LA fastlege vs LA robot	**0.649***	-3.501	0.001	-1.213	-0.244
HA robot vs LA robot	-0.298	-1.225	0.222	-0.905	0.227
Effekt b					
Fastlege vs robot, Tillit → Forventet anger	**-0.423***	-7.200	0.000	-0.690	-0.361
Valg og alvorlighetsgrad, Tillit → Forventet anger	**-0.433***	-7.323	0.000	-0.703	-0.375
Effekt c'					
Fastlege vs robot	**-0.262***	-2.040	0.042	-0.625	-0.089
HA fastlege vs LA fastlege	0.149	1.071	0.285	-0.207	0.609
HA fastlege vs HA robot	-0.060	-0.308	0.758	-0.474	0.328
HA fastlege vs LA robot	-0.275	-1.628	0.105	-0.731	-0.024
LA fastlege vs HA robot	-0.209	-1.051	0.295	-0.738	0.177
LA fastlege vs LA robot	**-0.424***	-2.489	0.014	-0.976	-0.191
HA robot vs LA robot	-0.214	-0.977	0.329	-0.718	0.123



Forventet anger

	β	t	p	Bootstrapping 95% CI	
				Lower	Upper
Effekt c					
Fastlege vs robot	-0.140	-0.978	0.329		
HA fastlege vs LA fastlege	-0.076	-0.492	0.623		
HA fastlege vs HA robot	-0.086	-0.385	0.701		
HA fastlege vs LA robot	-0.229	-1.233	0.219		
LA fastlege vs HA robot	-0.010	-0.044	0.965		
LA fastlege vs LA robot	-0.153	-0.809	0.420		
HA robot vs LA robot	-0.143	-0.576	0.566		
Effekt a					
Fastlege vs robot	-0.236	-1.659	0.099	-0.672	0.045
HA fastlege vs LA fastlege	**-0.552***	-3.706	0.000	-1.119	-0.345
HA fastlege vs HA robot	-0.311	-1.440	0.151	-0.896	0.077
HA fastlege vs LA robot	**-0.608***	-3.380	0.001	-1.289	-0.322
LA fastlege vs HA robot	0.241	1.103	0.271	-0.180	0.813
LA fastlege vs LA robot	-0.055	-0.302	0.763	-0.560	0.427
HA robot vs LA robot	-0.297	-1.232	0.219	-0.966	0.186
Effekt b					
Fastlege vs robot, Oppfattet risiko → Forventet anger	**0.188***	3.227	0.001	0.046	0.336
Valg og alvorlighetsgrad, Oppfattet risiko → Forventet anger	**0.196***	3.279	0.001	0.050	0.349
Effekt c'					
Fastlege vs robot	**-0.262***	-2.040	0.042	-0.625	-0.089
HA fastlege vs LA fastlege	0.149	1.071	0.285	-0.207	0.609
HA fastlege vs HA robot	-0.060	-0.308	0.758	-0.474	0.328
HA fastlege vs LA robot	-0.275	-1.628	0.105	-0.731	-0.024
LA fastlege vs HA robot	-0.209	-1.051	0.295	-0.738	0.177
LA fastlege vs LA robot	**-0.424***	-2.489	0.014	-0.976	-0.191
HA robot vs LA robot	-0.214	-0.977	0.329	-0.718	0.123





Appendix 12: Undersøkelse

I denne versjonen er Dr. Johansen og IBM Watson byttet ut med «behandler» der det er forskjell mellom de ulike versjonene. En forhåndsvisning av spørreundersøkelsen vises her: https://nhh.eu.qualtrics.com/jfe/preview/SV_6nTiASf6wbY3TVk?Q_CHL=preview&Q_SurveyVersionID=current

Intro

Tusen takk for at du ønsker å delta i en undersøkelse utført av studenter ved Norges Handelshøyskole (NHH). Undersøkelsen utgjør datagrunnlaget for en masteroppgave, som omhandler ulike behandlingsmetoder i helsevesenet. Du må være 18 år for å delta.
I undersøkelsen vil du få informasjon om et case, og spørsmål du skal ta stilling til. Det er viktig at du leser teksten nøye og besvarer spørsmålene etter beste evne. Når du har klikket deg videre i undersøkelsen, er det ikke mulig å gå tilbake.
Undersøkelsen tar omtrent 10 minutter å gjennomføre. Det er frivillig å delta, og du kan når som helst trekke deg uten å oppgi noen grunn. Alle data som samles inn anonymiseres, og det er ingen mulighet til å identifisere hvem som har svart hva i ettertid.

- ○ Jeg samtykker
- ○ Jeg samtykker ikke

Deltakerinformasjon

Kjønn

- ○ Mann
- ○ Kvinne
- ○ Annet

Hvor gammel er du?

Hva er din arbeidsstatus?

- ○ Student
- ○ Yrkesaktiv
- ○ Annet

Snuser du fast?

- Ja
- Nei
- Kun en gang iblant

Høy alvorlighetsgrad valg

Forestill deg at du kjenner på følgende symptomer tilknyttet helsen din, og av den grunn har behov for en medisinsk konsultasjon:

Muskel- og leddsmerter, slapphet, hjertebank, hodepine og søvnløshet.
På bakgrunn av symptomene dine foretar du noen søk på Google, og er relativt usikker over hva det kan være - alt fra leddgikt og kreft, til utbrenthet.

Du har valget mellom fastlegen Dr. Johansen og roboten IBM Watson. Begge kan gjennomføre konsultasjonen. IBM Watson er en programvare på datamaskin basert på kunstig intelligens. IBM Watson kan i likhet med Dr. Johansen innhente all nødvendig informasjon, og gjennomføre medisinske tester. Selve diagnostiseringen har en treffsikkerhet på 80% uavhengig av hvem av dem du velger, som vil si at det er 80% sannsynlighet for at du får korrekt diagnose.

Lav alvorlighetsgrad valg

Forestill deg at du kjenner på følgende symptomer tilknyttet helsen din, og av den grunn har behov for en medisinsk konsultasjon:

Sår hals, feber, svelgevansker, hodepine, hovne og ømme lymfeknuter på halsen.
På bakgrunn av symptomene dine foretar du noen søk på Google. Du er relativt sikker på at du har akutt halsbetennelse, og ønsker derfor antibiotika.

Du har valget mellom fastlegen Dr. Johansen og roboten IBM Watson. Begge kan gjennomføre konsultasjonen. IBM Watson er en programvare på datamaskin basert på kunstig intelligens. IBM Watson kan i likhet med Dr. Johansen innhente all nødvendig informasjon, og gjennomføre medisinske tester. Selve diagnostiseringen har en treffsikkerhet på 80% uavhengig av hvem av dem du velger, som vil si at det er 80% sannsynlighet for at du får korrekt diagnose.

Hvor alvorlig opplever du at symptomene dine er? (1 = svært lite alvorlig, 7 = svært alvorlig)

Hvem vil du skal gjennomføre konsultasjonen?

- Dr. Johansen
- IBM Watson

Høy alvorlighetsgrad

Behandler stiller deg spørsmål som for eksempel hva slags plager du har, hvor lenge det har vedvart, og hvordan dette påvirker deg i hverdagen. Videre innhenter legen informasjon fra journalen din for å se på tidligere sykdomshistorie. Du blir også sendt til legekontorets laboratorium for blodprøve, samt måling av blodtrykk og kroppstemperatur. Det hentes inn statistikk på tilfeller med samme type symptomer og sykehistorikk etter behov.

Etter en grundig evaluering konkluderer Behandler med at du har symptomer som tyder på at du er utbrent. Du blir sykemeldt og oppfordret til å redusere tempoet i hverdagen. I tillegg bør du holde deg i fysisk aktivitet. Dersom du ønsker det får du også en henvisning til psykolog. Det er satt opp en kontrolltime om fire uker, men du skal gi beskjed dersom symptomene forverrer seg.

Lav alvorlighetsgrad

Behandler stiller deg spørsmål som for eksempel hva slags plager du har, hvor lenge det har vedvart, og hvordan dette påvirker deg i hverdagen. Videre innhenter roboten informasjon fra journalen din for å se på tidligere sykdomshistorie. Du blir også sendt til legekontorets laboratorium for blodprøve, samt måling av blodtrykk og kroppstemperatur. Det hentes inn statistikk på tilfeller med samme type symptomer og sykehistorikk etter behov.

Etter en grundig evaluering konkluderer behandler med at du har en halsbetennelse forårsaket av en bakterieinfeksjon. Du får derfor resept på en antibiotikakur som må følges. Det anses ikke som nødvendig å sette opp en kontrolltime. Du bes om å ta kontakt dersom symptomene forverres, siden det kan bli nødvendig med ytterlige undersøkelser.

Tillit til behandlingsform

I hvilken grad mener du at behandler vil deg vel, og handler i din beste interesse? (1 = i svært liten grad, 7 = i svært stor grad)

I hvilken grad mener du at behandler har stilt korrekt diagnose, og gitt deg rett behandling? (1 = svært liten grad, 7 = svært høy grad)

I hvilken grad mener du at behandler har de rette egenskapene til å gjennomføre konsultasjonen? (1 = svært liten grad, 7 = svært høy grad)

I hvilken grad mener du at behandler evner å vurdere din unike helsesituasjon? (1 = svært liten grad, 7 = svært høy grad)

Oppfattet risiko i forbindelse med behandlingsform

I hvilken grad anser du konsekvensene for å være alvorlige dersom du får feil diagnose og behandling? (1 = svært liten grad, 7 = svært stor grad)

I hvilken grad opplever du indre uro (engstelse, nervøsitet, stress) i forbindelse med diagnostisering og behandling? (1 = i svært liten grad, 7 = i svært stor grad)

Evaluering - Tilfredshet og forventet anger

Husk å forsøke å sette deg inn i situasjonen som pasient når du svarer på spørsmålene. Du vil nå få spørsmål om vurderingsformen, der du skal foreta en evaluering på en skala fra 1-7.

I hvilken grad er du fornøyd med diagnostisering og behandling gjennomført av behandler? (1 = svært misfornøyd, 7 = svært fornøyd)

I hvilken grad angrer du på valget ditt av behandler? (1 = svært liten grad, 7 = svært høy grad)

I hvilken grad vil du anbefale behandler til andre? (1 = svært liten grad, 7 = svært høy grad)

Tredje alternativ

Du får nå et tredje alternativ for hvordan konsultasjonen kunne blitt gjennomført. Dr. Johansen kan benytte seg av IBM Watson for å få informasjon og veiledning, men er selv ansvarlig for den endelige beslutningen.

Hvem ville du valgt?

- ○ Dr. Johansen
- ○ IBM Watson
- ○ Dr. Johansen med hjelp fra IBM Watson

Spørsmål relatert til personlighet og ferdigheter (neste side)

Du vil nå få noen påstander, der du skal avgi ditt svar på en skala fra svært uenig til svært enig.
Hvor godt kjenner du deg igjen i disse påstandene?

	Svært uenig	Uenig	Verken enig eller uenig	Enig	Svært enig
Jeg klarer å bruke nye teknologiske produkter og tjenester uten hjelp	○	○	○	○	○
En bør være forsiktig med å erstatte mennesker som utfører viktige oppgaver med teknologi, fordi ny teknologi er ikke pålitelig	○	○	○	○	○

Hvor godt kjenner du deg igjen i disse påstandene?

	Svært uenig	Uenig	Verken enig eller uenig	Enig	Svært enig
Jeg kjenner ofte på negative følelser (engstelse, sinne, tristhet)	○	○	○	○	○
Jeg føler meg hjelpesløs i stressende situasjoner	○	○	○	○	○
Jeg stoler på at de fleste andre mennesker vil meg vel, også personer jeg ikke kjenner så godt	○	○	○	○	○

Generell risikovillighet

Til slutt, i hvilken grad er du generelt villig til å påta deg risiko? (1 = i svært liten grad, 7 = i svært stor grad)

Slutt

Du har nå deltatt i et eksperiment som undersøker viljen til å benytte kunstig intelligens relativt til fastleger. Dette er gjort i en kontekst der alvorlighetsgraden av symptomene som deltakerne opplyses om varierer, slik at den oppfattede risikoen er ulik. Vi ønsker å se på om viljen til å benytte kunstig intelligens er sterkere når den oppfattede risikoen er lav. Caset du leste om var fiktivt, men benyttelse av kunstig intelligens i helsevesenet øker stadig.

Skulle du ha noen ytterligere spørsmål om eksperimentets gjennomføring og resultater, behandling av personlig informasjon eller lignende, kan vi kontaktes på følgende e-postadresser:

Kaja.Schultz@student.nhh.no
Vilde.Thorud@student.nhh.no

Takk for din deltakelse!