Høgskulen på Vestlandet

MOØ300 Masteroppgave

MOØ300-O-2022-VÅR-FLOWassign

Predefinert informasjon

Startdato:	09-05-2022 00:00	**Termin:**	2022 VÅR
Sluttdato:	23-05-2022 14:00	**Vurderingsform:**	Norsk 6-trinns skala (A-F)
Eksamensform:	Masteroppgave		
Flowkode:	203 MOØ300 1 O 2022 VÅR		
Intern sensor:	Jill Merethe Loga		

Deltaker

Navn:	Iselin Pedersen Kristiansen
Kandidatnr.:	423
HVL-id:	081276@hvl.no

Informasjon fra deltaker

Egenerklæring *:	Ja
Jeg bekrefter at jeg har registrert oppgavetittelen på norsk og engelsk i StudentWeb og vet at denne vil stå på vitnemålet mitt *:	Ja

Gruppe

Gruppenavn:	7
Gruppenummer:	15
Andre medlemmer i gruppen:	Karoline Fløysand Vollan, Live Petrea Brunsvik Haraldsen

Jeg godkjenner avtalen om publisering av masteroppgaven min *

Ja

Er masteroppgaven skrevet som del av et større forskningsprosjekt ved HVL? *

Nei

Er masteroppgaven skrevet ved bedrift/virksomhet i næringsliv eller offentlig sektor? *

Nei

Høgskulen på Vestlandet

MASTEROPPGAVE

Betydningen av syntetisk data i helsesekotren og helseforskning

The Importance of Synthetic Data in the Healthcare Sector and in Health Research

Live Haraldsen, Iselin Pedersen Kristiansen & Karoline Fløysand Vollan

Master i Innovasjon og ledelse
Institutt for økonomi og administrasjon
Veileder: Kjersti Berg Danilova
23.05.2022

Jeg bekrefter at arbeidet er selvstendig utarbeidet, og at referanser/kildehenvisninger til alle kilder som er brukt i arbeidet er oppgitt, *jf. Forskrift om studium og eksamen ved Høgskulen på Vestlandet, § 12-1.*

Sammendrag

Formålet med studien er å belyse hvilket mulighetsrom syntetisk data kan gi den norske helsesektoren og hvilke utfordringer dette medfører. Med økende bruk av data i samfunnet stilles det særlige krav til hvordan personopplysninger behandles. Flere sektorer har nylig begynt å sette søkelys på hvordan syntetiske data kan benyttes fremfor reelle data, for å omgå utfordringer knyttet til personvern. Vi har studert hvilken betydning syntetiske data kan ha for helsesektoren sett i lys av behandling, deling og lagring av data. Syntetisk data er foreløpig ikke implementert i helsesektoren og dermed fremlegger studien de identifiserte teknologiske, innovative og juridiske dimensjonene av fenomenet.

Studien har en kvalitativ forskningsmetode med et eksplorativt design. Det begrunnes med at fenomenet som studeres er forholdsvis nytt i Norge og det foreligger lite forskning på feltet. Kvalitativ metode kan gi oss en dypere forståelse og innsikt i forskjellige synspunkt som kan belyse problemstillingen. I den forbindelse har vi benyttet oss av semistrukturerte dybdeintervju for å samle inn data.

Våre funn indikerer at syntetiske data kan ha en positiv konsekvens i henhold til *lovverk, helseforskning*, *innovasjon* og *kompetanse* innen helsesektoren. Syntetisk data kan være en løsning på personvernsproblematikken og samtidig tilgjengeliggjøre ytterligere data for deling mellom helseforetak. Sett i lys av *helseforskning* og *innovasjon* kan syntetisk data brukes til *kompetansehevende* tiltak, i tillegg til å fremskynde tidkrevende prosesser. Som følge av mulighetene ble det avdekket utfordringer knyttet til *lovverk*, *helseforskning*, *kompetanse* og *tillit*. Våre funn viser at det foreligger et fravær av *rettslige* avklaringer knyttet til syntetiske data. Bakgrunnen for det gjenspeiles i at fenomenet er lite utbredt i Norge, som kan ses i lys av funn som viser begrenset *tillit* og *kompetanse* knyttet til syntetisk data. Fra et *forskerperspektiv* avdekker funn en antakelse om skepsis til bruk av syntetiske data, relatert til datakvalitet og etiske implikasjoner.

Studien bidrar med ny og verdifull innsikt for bruk av syntetisk data innen helsesektoren. Kompleksiteten ved fenomenet syntetisk data presenteres og det identifiseres faktorer som kan tilrettelegge for mulighetsskapende bruk av syntetisk data. Avslutningsvis bidrar studien med identifisering av tema for videre forskning.

Summary

The purpose of the study is to shed light on what opportunities synthetic data can provide the Norwegian healthcare sector and what challenges this entails. With the increasing use of data in society, specific demands are set in regards to how personal data is processed. Several sectors have recently begun to focus on how synthetic data can be used rather than real data, to circumvent challenges related to privacy. We have studied the significance of synthetic data for the healthcare sector in terms of the processing, sharing and storage of data. Synthetic data has not yet been implemented in the healthcare sector and thus the study presents the identified technological, innovative and legal dimensions of the phenomenon. The study has a qualitative research method with an exploratory design. The reason is that the phenomenon being studied is relatively new in Norway, as such there is limited existing research in the field. Qualitative methods can offer us a deeper understanding and insight through different points of view that can shed light on the problem. In this regard, we have used semi-structured in-depth interviews to collect data.

Our findings indicate that synthetic data can have a positive consequence in accordance with legislation, health research, innovation and competence in the healthcare sector. Synthetic data can be a solution to the privacy issues and at the same time make additional data available for sharing between health trusts. In terms of health research and innovation, synthetic data can be used for competence-enhancing measures, as well as to speed up time-consuming processes. As a result of the opportunities, challenges related to legislation, health research, competence and trust were identified. Our findings show that there is an absence of legal clarifications related to synthetic data. The background for this is reflected in the fact that the phenomenon is not widespread in Norway. This can be seen in the light of findings that show limited trust and competence related to synthetic data. From a researcher's perspective, the findings reveal an assumption of skepticism about the use of synthetic data, related to data quality and ethical implications.
The study provides new and valuable insight for the use of synthetic data in the healthcare sector. The complexity of the phenomenon of synthetic data is presented, and factors are identified that can facilitate the utilization of the possibilities created through the use of synthetic data. Finally, the study contributes to the identification of topics for further research.

Forord

Denne masteroppgaven er et avsluttende prosjekt på et toårig masterstudium i innovasjon og ledelse ved Høgskulen på Vestlandet. Studien tar for seg hvilken betydning bruk av syntetisk data kan ha for helsesektoren. Vi håper at oppgaven kan bidra til å belyse hvilke muligheter syntetisk data kan gi helsesektoren og i helseforskning, samt synliggjøre potensielle utfordringer tilknyttet bruk. Videre mener vi at studien er viktig i lys av at syntetiserte helsedata kan tilgjengeliggjøre ytterligere data i helsesektoren og tilrettelegge for deling av forskningsresultater. Det aktuelle fenomenet er ikke tidligere forsket på innenfor helsesektoren i Norge og vi er følgelig de første som belyser temaet.

Vi vil takke Synnøve Olset som har vært en betydningsfull støttespiller og motivator gjennom hele prosjektet. Fra introduksjon til tematikken, motiverende samtaler og innspill til informanter. Vi ønsker å rette en stor takk til informantene som tok seg tid til å bli intervjuet, dere har bidratt med verdifull innsikt og perspektiver innenfor fenomenet som studien søker å belyse. Videre vil vi takke familie og venner som har bistått med gjennomlesning og korrektur. Vi ønsker også å takke vår veileder Kjersti Berg Danilova for et solid samarbeid, gode råd og tilbakemeldinger. Vi setter stor pris på deres innsikt, tid og interesse for tematikken. Oppgaven hadde ikke vært det den er uten dere. Avslutningsvis vil vi takke hverandre for et konstruktivt og givende samarbeid gjennom masterstudiet og i masteroppgaven.

Bergen, mai 2022
Live Haraldsen, Iselin Pedersen Kristiansen og Karoline Fløysand Vollan

Innholdsfortegnelse

Figurliste

Tabelliste

1.0 Innledning

1.1 Bakgrunn og relevans

Helsesektoren er i et digitalt skifte og den teknologibaserte helsenæringen utvikler seg raskt (Direktoratet for e-helse, 2020; Tennøe & Prabhu, 2017). Samtidig viser en OECD-rapport (2019) at helsesektoren ligger mellom 10 og 15 år bak andre sektorer når det gjelder digital transformasjon. Bruken av kunstig intelligens (KI) kan imidlertid bidra til forbedring av tunge prosesser, eller deler av slike prosesser, i tillegg til å fremme innovasjon og samhandling (Meld. St. 7, (2019-2020)). Ved å benytte KI i helsesektoren kan det forbedre dagens helsetilbud gjennom mer presis diagnostisering, bedre oppfølging og behandling (Hokstad, 2019).

Stordata ses som et sentralt begrep for tilgjengelig data, samtidig som kvalitet i data fremheves som en av de viktigste kriteriene for at den skal gi verdi (Andreu-Perez et al., 2015). Data i helsesektoren inneholder ofte sensitive personopplysninger og sektoren er derfor pålagt å opptre i samsvar med reguleringer og retningslinjer, ved forvaltning og deling av helsedata (Personopplysningsloven, 2018). Samtidig øker behovet for å produsere og behandle data. Dette aktualiserer spørsmål knyttet til hvordan helsesektoren kan produsere tilstrekkelig mengder av verdiskapende data som kan lagres, deles og analyseres, samtidig som personvern ivaretas.

Syntetiske data har blitt introdusert som en potensiell løsning når det gjelder personvernutfordringer ved databehandling og bedre utnyttelse av KI-modeller (Ramos & Subramanyam, 2021). Dette er data som genereres av en maskinlæringsmodell og gjenspeiler de originale dataene uten personopplysninger (Nowok et al., 2016, 1). Syntetiske data har fått mye oppmerksomhet de siste årene og som følge av dette har flere sektorer begynt å se på mulighetene bruken av denne teknologien kan innebære. I 2021 startet Helse Vest IKT prosjektet SYNDATA, for å evaluere nytten av å bruke syntetiske helsedata innen forskning og innovasjonsprosjekter i helsesektoren (Helse Stavanger, u.å.).

1.2 Tema og problemstilling

Ifølge Ramos og Subramanyam (2021) forventes det en betydelig økning i behovet for data som kreves for å kunne benytte KI fra 2020 og mot 2030. Behovet kan vise seg å være så markant at reelle eller ekte data, fra direkte målinger tilknyttet ressurskrevende prosesser og personvern, ikke lenger vil ha muligheten til å kunne dekke det. Her presenteres syntetiske data som en løsning for å kunne berike reelle data (Ramos & Subramanyam, 2021).

Til tross for at bruk av syntetisk data er et tilsynelatende nytt fenomen i Norge, har likevel enkelte aktører gjort bruk av slike data, der i blant Visma, Skatteetaten og Nav. Skatteetaten ferdigstilte i 2019 et prosjekt om bruk av syntetiske data for å utvikle et syntetisk folkeregister (Birkeland, 2019). Dette prosjektet var det første i sitt slag og ble igangsatt for at folkeregisteret skulle samsvarere med den oppdaterte folkeregisterloven som trådte i kraft i 2017.Utvikling av et syntetisk folkeregister for testing vil hjelpe Skatteetaten å ivareta sikkerhet og personvern samtidig som utvikling kan finne sted (Birkeland, 2019). Etterhvert som behandling, lagring og deling av persondata har blitt mer utbredt, har kravene som settes for dataforvaltere blitt strengere. Dette har skjedd både nasjonalt og internasjonalt, eksempelvis gjennom General Data Protection Regulation, bedre kjent som GDPR (Sørebø et al., 2020).

Ønsket for denne masteroppgaven er å forske på et fenomen som tidligere har blitt lite belyst, som samtidig kan skape verdi for mange aktører i en valgt bransje. Med inspirasjon fra SYNDATA ønsker vi gjennom denne studien å se nærmere på bruk av syntetisk data i helsesektoren og helseforskning. Det anses som nødvendig å avdekke mulighetsrommet syntetisk data gir, for å forstå de nåværende utfordringene tilknyttet databehandling i helsesektoren og for helseforskning. Formålet med studien er å se på hvilke muligheter syntetiske data kan gi den norske helsesektoren, i tillegg til hvilke utfordringer dette medfører. Vi har utarbeidet en overordnet problemstilling for studien:

"Hvilken betydning kan bruk av syntetisk data ha for helsesektoren?"

For å sikre en konstruktiv ramme og sørge for at problemstillingen besvares, er det videre utarbeidet to forskningsspørsmål:

"Hvordan kan bruk av syntetisk data skape muligheter for helsesektoren?"

"Hvilke utfordringer bør helsesektoren være bevisst på tilknyttet bruk av syntetisk data?"

1.3 Metode og avgrensninger

Tematikken i studien er kompleks og sammensatt, noe som innebærer et behov for en tydelig avgrensning. Vi skal studere bruk av syntetiske persondata i norsk helsesektor og helseforskning på et overordnet nivå, hvor avanserte og tekniske aspekter ikke tas med. Dette gjøres gjennom dybdeintervju med 11 fagpersoner og forskere som har tilknytning til helsesektoren eller IKT. Studien vil presentere relevante terminologier og teori for å gi en innføring og skape en forståelse av temaet.

1.4 Studiens bidrag

Studien vil bidra til å belyse den teknologiske utviklingen som helsesektoren i Norge står overfor. Særlig med fokus på produksjon, forvaltning og deling av stordata ved hjelp av syntetiske data, samtidig som personvern ivaretas. Det er en begrenset mengde med eksisterende forskning på syntetisk data. Gjennom å besvare den overordnede problemstillingen og forskningsspørsmålene tilfører denne studien relevant forskning på feltet. Studien samler teori og funn på en konstruktiv måte som er ment å skape verdi for leseren, samt identifisere muligheter for videre forskning.

1.5 Masteroppgavens oppbygging

Masteroppgaven er presentert i seks kapitler. Kapittel to tar for seg relevant teori som skal danne det teoretiske rammeverket for studien. Her er formålet å sikre en stødig teoretisk grunnmur og forankring. Kapittel tre forklarer studiens forskningsdesign, -tilnærming og -metode. Vi tar også for oss datamaterialets kvalitet i form av validitet og reliabilitet, så vel som etiske hensyn og personvern. I kapittel fire presenteres relevante sitater fra informantene med beskrivelse av hva disse forteller oss, i tillegg til en tabell som oppsummerer empiriske hovedfunn. Bakgrunnen for dette er å danne oversikt og tilrettelegge for besvarelse av forskningsspørsmålene. I kapittel fem presenteres og diskuteres de mest vesentlige empiriske

funn, sett i sammenheng av teori og annen forskning. Kapittel seks er den siste og avsluttende delen. Her presenteres en konklusjon på studien, hvor problemstillingen forsøkes å besvares gjennom forskningsspørsmålene. Videre uttrykkes praktiske og teoretiske implikasjoner, før kapittelet avrundes med metodiske begrensninger og forslag til fremtidig forskning.

2.0 Teoretisk rammeverk

Kapittelet er delt inn i seks deler, hvorav noen av dem har underkategorier. Den første delen presenterer teknologi i helsesektoren som skal gi innsikt i feltet for studien vår. I den andre delen søker vi å skape forståelse for kompleksiteten ved bruk av teknologi og syntetiske data. Den tredje delen introduserer og belyser personvern og GDPR, som er en viktig regulerende faktor ved bruk og deling av persondata. Del fire introduserer og belyser aktuelle typer innovasjon og innovasjonsprosesser. Del fem trekker frem helseforskning, og nåværende utfordringer tilknyttet dette feltet. Kapittelet avsluttes med sentral teori om kompetanse og tillit.

2.1 Teknologi i helsesektoren

Teknologi som kunstig intelligens og stordataanalyser legger grunnlaget for verdiøkende aktiviteter i samfunnet (Meld. St. 22, (2020-2021)). Dette ser vi i hvordan den teknologiske utviklingen har ført til banebrytende endringer i blant annet produksjon, tjenesteytelser og digital samhandling (Tenneø & Prabhu, 2017). Ved å benytte teknolgi som kunstig intelligens, maskinlæring og stordata, får man mer ut av dataene enn det som tidligere har vært mulig (Pettersen, 2019). Bruk av kunstig intelligens muliggjør man bedre utnyttelse av felles helsedata, slik at helsesektoren kan tilby raskere og mer presis diagnostisering, samt bedre oppfølging og behandling (Pettersen, 2019). Det foreligger imidlertid et behov for mer data dersom slike teknologier skal utvikles og benyttes (Tennøe & Prabhu, 2017). I den forbindelse er det avgjørende med tilgang til datasett med høy grad av kvalitet (Andreu-Perez et al., 2015). Som et resultat av dette skal det tilrettelegges for deling av data, både i offentlig og privat sektor og mellom dem (Kommunal- og moderniseringsdepartementet, 2020). Økt deling og bedre utnyttelse av dataene vil på den måten bidra til økt verdiskaping. Helsesektoren håndterer funksjonelle og profesjonelle siloer, som kan ses som strukturer av fragmentert omsorg og profesjonell praksis (Van Rossum et al., 2016). Dette kan skape

barrierer for å optimalisere arbeidsprosesser og strukturer (Van Rossum et al., 2016). Data deles i begrenset grad internt i en virksomhet, med eksterne aktører og mellom offentlige virksomheter som har et felles behov (Holstad, 2014). Videre hevdes det at store deler av dagens dataproduksjon i byråkratiet, produseres til eget formål i den enkelte virksomhet eller fagområde. Dataene utvikles uten en oppfatning om at andre aktører har behov for de samme dataene, både råvarene og produktet (Holstad, 2014).
Basert på dagens utvikling innen teknologi åpnes det opp for nye muligheter ved anvendelse av helsedata, samtidig som teknologien åpner opp for nye utfordringer angående personvern, informasjonssikkerhet og risiko (Tennøe & Prabhu, 2017). Utviklingen bidrar til å endre måten helsedata genereres og anvendes, og utfordrer hvordan den har vært organisert, forvaltet og regulert (Helsedatautvalget, 2017). Som følge av dette vil vi se nærmere på teori om data og de relevante teknologiske verktøyene for utvikling av syntetiske data.

2.1.1 Data

Data er et ord som mange har et forhold til, men som kan være vanskelig å definere. For å forstå data i kontekst med forskning i helsesektoren, er det nødvendig å nærmere definere begrepet. *"Data er informasjon, spesielt fakta eller tall, samlet inn for å bli undersøkt og vurdert og brukt for å hjelpe beslutningstaking, eller informasjon i elektronisk form som kan lagres og brukes av en datamaskin"* (Cambridge Dictionary, u.å). Vi anser definisjonen som relevant for vår studie på bakgrunn av dens omfang. Som definisjonen forklarer representeres data i flere ulike format, både strukturerte og ustrukturerte. Strukturerte data finnes ofte i organisert og formatert med en satt struktur. Telefonnummer og navnelister er eksempler på strukturerte data. Ustrukturert data er data som ikke er organisert, hvor det ikke kan identifiseres en gjenkjennelig struktur, eksempelvis bilder og lydfiler (Heggernes, 2020).

2.1.2 Datakvalitet

Verdien av data kan anses å gradvis øke gjennom gjenbruk og deling. For at data skal kunne brukes til å fatte beslutninger, kreves det en vesentlig grad av kvalitet (Andreu-Perez et al., 2015). Datakvalitet er et stort begrep hvor det er vanskelig å enes om en spesifikk definisjon og tolkning (Oliveira et al., 2005). Det begrunnes med at data ikke innehar fysiske egenskaper som et produkt har, og dermed gjør det vanskelig å vurdere kvaliteten (Veregin, 1999). Kvaliteten på dataene karakteriseres imidlertid som immaterielle egenskaper, slik som

fullstendighet og konsistens (Veregin, 1999). Oliveira et al. (2005) hevder at datakvalitet kan ses i to ulike perspektiv, henholdsvis database og ledelse. Datakvalitet i databaser er et teknisk synspunkt, mens i et lederperspektiv inngår aspekter som tilgjengelighet, troverdighet, relevans, tolkbarhet og objektivitet (Oliveira et al., 2005). Ifølge forskning kan datakvalitet bestå av flere dimensjoner, hvor de deles inn i to kategorier (Hazen et al., 2014). Den første kategorien er iboende, som omhandler attributter som er objektive og innebygd i dataene. I den iboende kategorien inngår faktorene nøyaktighet, aktualitet, konsistens og fullstendighet. Den andre kategorien er kontekstuelle, som refererer til attributter som er kontekstavhengige ut ifra hva dataene blir observert eller brukt til. Den kontekstuelle dimensjonen innebærer faktorer som relevans, kvantitet, verdiøkning, troverdighet, tilgjengelighet og omdømme til dataene (Hazen et al., 2014).

Ifølge Bray og Parkin (2008) er det fem dimensjoner innenfor datakvalitet i forskning, herunder kompletthet, validitet, reliabilitet, sammenlignbarhet og aktualitet. I bruk av helsedata innebærer kompletthet eksempelvis i hvilken grad den gitte helsetilstanden som forskes på dekkes av befolkningen og hvorvidt dataene inkluderer alle tilfellene i en spesifikk målpopulasjonen. Validitet handler om i hvilken grad dataene reflekterer virkeligheten, og i hvor stor grad dette er nøyaktig (Bray & Parkin, 2008). Reliabilitet tar for seg i hvilken grad det er mulig å reprodusere innholdet, mens sammenlignbarhet omhandler i hvilken grad data kan gi et sammenligningsgrunnlag på tvers av tid, geografi og ulike datakilder. Aktualitet handler om hastighet, og innebærer tid fra en hendelse har skjedd til informasjon er tilgjengelig for brukere av data (Bray & Parkin, 2008; Nasjonalt servicemiljø for medisinske kvalitetsregistre, u.å).

2.1.3 Stordata (Big Data)

Helsesektoren forvalter store mengder helsedata, og for å danne et bilde på kompleksiteten som påvirker fenomenet som studeres, er det hensiktsmessig å presentere begrepene stordata og stordataanalyse. I denne studien har vi valgt å bruke Laneys definisjon av stordata, på bakgrunn av det den forklarer begrepet på en forståelig måte. *"Stordata er store mengder omfattende variert data som genereres, fanges opp og behandles med høy hastighet"* (Laney 2001 referert i Günther et al., 2017, s. 191). Günther et al. (2017) hevder at verdien til stordata handler om hvorvidt selskaper klarer å omsette sosiale eller økonomiske verdier gjennom å benytte stordata. Verdien ligger i selve datamaterialet, som kan skape fremtidige

bruksmuligheter, i form av ny innsikt og kapital sett i et forretningsperspektiv. Zaslavsky et al. (2013) trekker frem at stordata kan gi økt inntekt, hjelpe å predikere fremtidige hendelser og redusere risiko.

Heggernes (2020) presenterer tre kategorier for inndeling av stordata. Fenomenet kan inndeles i menneskegenererte, systemgenererte og maskingenererte data. All type data er i prinsippet generert av mennesker, men med menneskegenererte data menes data som skapes og deles av mennesker. Det forekommer ved deling av data på sosiale medier. Systemgenererte data er transaksjonsdata som produseres gjennom IT-baserte systemer og benyttes for å støtte opp under forretningsdriften til en virksomhet. Dette er en upersonlig form for data og presenteres gjennom statistikk. Maskingenererte data genereres fra maskiner med en kontinuerlig produksjon av store mengder data. I maskingenererte data, som i KI-modeller, kan man registrere mønstre i atferd og tilpasse virksomheten etter dette. Videre hevder Heggernes (2020) at det er mest vekst innen maskingenererte data, på bakgrunn av at det i hovedsak består av strukturerte data som enklere kan analyseres.

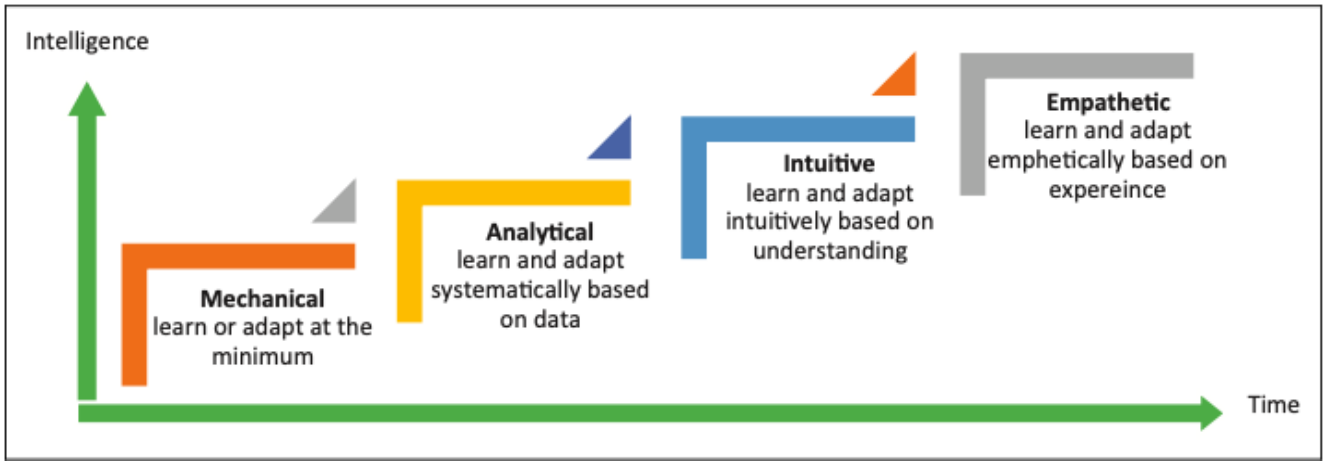
Benke og Benke (2018) hevder at stordata kan assosieres med ressursene som er tilknyttet beregninger, som er nødvendig for å håndtere det komplekse og økende volumet av data fra mange kilder. Dette kan ses i sammenheng med at det foreligger tre karakteristikker om hva stordata innebærer, herunder hastighet, variasjon og volum (Zaslavsky et al, 2013). Hastighet går ut på hvor ofte data blir generert, variasjon omfatter de ulike typene dataene som blir samlet inn og volum beskriver mengden data som samles inn. For at stordata skal kunne skape verdi må den analyseres og tolkes (Zaslavsky et al, 2013). Stordataanalyse kan forklares som en analyse som håndterer store volum med data fra flere ulike datakilder, som har til hensikt å oppdage nye sammenhenger og innsikter (Lanestedt, 2016).

2.1.4 Kunstig intelligens

Kunstig intelligens (KI) er et samlebegrep for emner relatert til data og anses som sentralt i produksjon av syntetiske data. På bakgrunn av dette gis det her en kort innføring i begrepet og noen aktuelle underkategorier. EU-kommisjonen sin ekspertgruppe definerer kunstig intelligens som:

"*Systemer som utfører handlinger, fysisk eller digitalt, basert på tolkning og behandling av strukturerte eller ustrukturerte data, i den hensikt å oppnå et gitt mål"* (EU-kommisjonen, 2018, s.1).

Huang og Rust (2018) presenterer fire underkategorier for kunstig intelligens som er mekanisk, analytisk, intuitiv og empatisk.



(Figur 1: *De fire intelligensene*. Huang & Rust, 2018, s. 158)

Mekanisk intelligens defineres som "*evnen til å automatisk utføre rutinemessige, gjentatte oppgaver*" (Huang & Rust, 2018, s.158). Det kan forstås som å lære seg eller tilpasse seg et minimumskrav. Sternberg (1997) hevder at vi har en begrenset bevissthet rundt mekaniske prosesser, da det er prosesser som har blitt utført gjentatte ganger. Mekanisk KI har som hensikt å etterligne menneskelig automatisering, som gjøres gjennom begrenset læring og tilpasningsevne, for å sikre konsistens (Sternberg, 1997).

Analytisk intelligens defineres som "*evnen til å behandle informasjon for problemløsning og lære av den*" (Sternberg 2005 referert i Huang & Rust, 2018, s.158). Dette omhandler læring og tilpasning av systematisk data, i tillegg til hvordan behandling av informasjon, logiske resonnementer og matematiske ferdigheter foregår (Sternberg, 1999). Slike ferdigheter kan oppnås gjennom spesialisering, ekspertise og øving i kognitiv tenking (Huang & Rust, 2018). Dataanalyse og maskinlæring fremheves av Huang og Rust (2018) som de mest sentrale analytiske KI-applikasjonene.

Intuitiv intelligens defineres som "*evnen til å tenke kreativt og effektivt tilpasse seg nye situasjoner*" (Sternberg 2005, referert i Huang & Rust, 2018, s.159). Definisjonen beskriver

hvordan det kan forstås som visdom fra helhetlig og erfaringsbasert tenking (Sternberg, 2005, referert i Huang & Rust, 2018, s.159). En slik KI-intelligens handler om å lære og tilpasse seg intuitivt basert på forståelse. Intuitiv intelligens består av faglige ferdigheter og evner, som ofte kan kreve innsikt og kreativ problemløsning. Hovedforskjellen på analytisk KI og intuitiv KI er dermed forståelse.

Empatisk intelligens defineres som "*evnen til å gjenkjenne og forstå andre mennesker følelser, reagere passende følelsesmessig og påvirke andres følelser*" (Goleman 1996, referert i Huang & Rust, 2018, s.159). I empatisk intelligens er formålet å lære og tilpasse seg situasjonen på en empatisk måte basert på opplevelser. Ifølge Johnson (2014) inngår faktorer som mellommenneskelige og sosiale ferdigheter, som har til hensikt å hjelpe mennesker å være følsom overfor andres følelser for å klare og samarbeide godt.

Hybrid intelligens søker å bruke komplimenterende styrker ved menneskelig og maskinell intelligens for å utvide det menneskelige intellekt. En slik intelligens skjer gjennom å utføre komplekse oppgaver hvor en oppnår bedre resultater som hver av intelligensene kunne gjort alene (Dellermann et al., 2019, s. 276). Det stilles spørsmål om hvorvidt det er mulig å bytte ut menneskelig intelligens med teknologi. Foreløpig kan ikke kunstig intelligens erstatte menneskelige evner, som empati, sympati, resonnering og abstrakt tenking, men heller bidra til frigjøring av tid i repetitive oppgaver (Østbye, 2020). I skjæringspunktet mellom intelligente maskiner og menneskelig arbeidskraft, identifiseres det et mulig potensial (Davenport & Kirby, 2016). Utnyttelse av potensialet kan forekomme ved å bruke KI-teknologi som støtte for, og effektivisering av menneskelige ressurser (Kolbjørnsrud, 2017). Det poengteres at KI vil være en sentral del av helsesektoren i fremtiden, men at det er opp til mennesker å bestemme i hvilken grad det skal være komplimenterende (Hokstad, 2019). På bakgrunn av dette understrekes viktigheten av å følge med på den teknologiske og digitale utviklingen, for å sikre at det er mennesker som leder an og ikke maskiner (Hokstad, 2019).

2.1.5 Maskinlæring

For at syntetiske data skal kunne benyttes må de produseres. Dette skjer gjennom KI-systemer og mer spesifikt gjennom maskinlæring. Temaet introduseres derfor kort for å danne et bilde på kompleksiteten ved produksjon av syntetiske data. Maskinlæring kan omtales som en underkategori av KI, som omfatter alle funksjoner som gjør det mulig for

maskiner å lære fra data uten å være spesifikt programmert til det (Jakhar & Kaur, 2020). Maskinlæring kan anvendes på ulike mengder data, til tross for dette vil bruk av stordata øke sannsynligheten for et mer nøyaktig resultat (Mendling et al., 2017). Maskinen har evnen til å lære av seg selv og gradvis forbedre nøyaktigheten for hver gang analyser genereres. En slik evne kan ses som et læringsaspekt (Jakhar & Kaur, 2020). Denne iterative tilnærmingen til læring uten menneskelig innblanding, gir mulighet til å få innsikter som algoritmene ikke er spesifikt programmet til å finne, såkalte uventede funn (Jakhar & Kaur, 2020). Et eksempel på et uventet funn kan være at det avdekkes en uforventet trend fra modellen, eller at tallene fra analysen viser noe annet enn tallene alene. Den økte tilgjengeligheten på stordata hevdes å ha ført til et bredere anvendelsesområde, så vel som identifisering av behovet for kvaliteten på dataene (Mendling et al., 2017).

2.1.6 Syntetisk data

Syntetisk data kan defineres som *"data som etterligner de originale observerte dataene og bevarer relasjonene mellom variabler, men inneholder ingen avslørende informasjon*" (Nowok et al., 2016, s. 1). Grunntanken med syntetiske data er å erstatte noen eller alle av de observerte verdiene fra reelle data, ved å benytte sannsynlighetsfordelinger, slik at de essensielle statistiske egenskapene til det opprinnelige datasettet bevares (Nowok et al., 2016). Syntetiske data er derfor kunstig data som gjenspeiler virkelige data, matematisk eller statistisk. De genereres av en maskinlæringsmodell som er trent på personopplysninger, og innehar de samme egenskapene som reelle data, men uten personopplysninger (Meld. St. 22, (2020-2021)). Hensikten med syntetisk data er likevel at det skal være realistisk og validering vil dermed være en forutsetning, for å teste hvorvidt syntetiske data er tilsvarende reelle data (Chen et al., 2019).

I dag blir syntetisk data hovedsakelig brukt som testdata i systemutviklingsprosjekter. Samtidig pekes det på andre mulige bruksområder, slik som forskning og utvikling (Walonoski et al., 2020), men for å bruke syntetiske data på disse områdene må syntetiske datasett gjøres allment tilgjengelig. Dette vil være mulig å oppnå ettersom dataene ikke inneholder personopplysninger (Meld. St. 22, (2020-2021)).

2.1.6.1 Muligheter ved bruk av syntetisk data

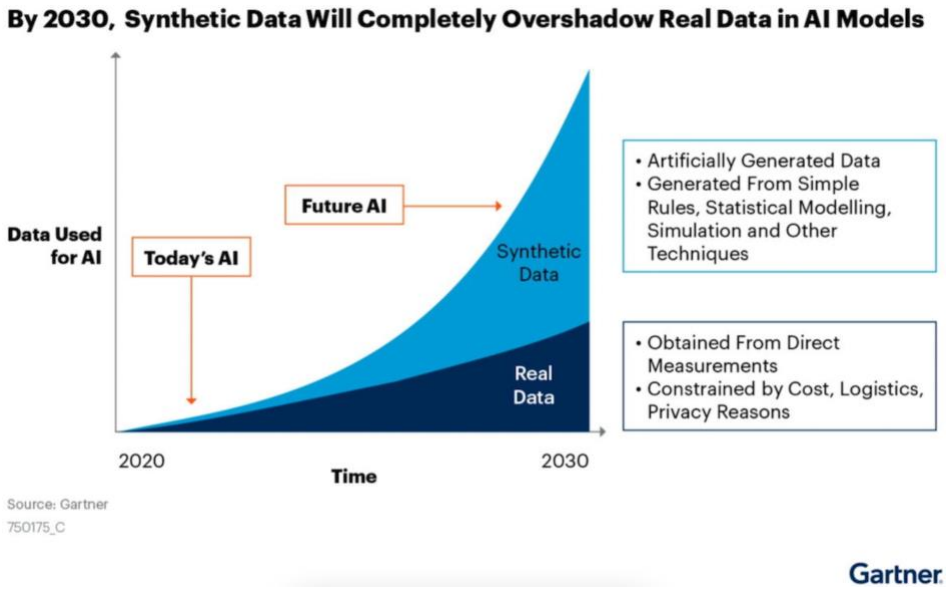
Syntetisk data muliggjør generering av realistiske data for helseforskning, systemimplementering og trening av KI-modeller (Chen et al., 2019). Videre belyses det at

tilgangen på reelle data kan være kostbart, strengt regulert og innebærer ofte etiske utfordringer, slik som personvernhensyn. Helsejournaler inneholder i stor grad sensitiv informasjon og det er derfor spesielt utfordrende for helsesektoren å behandle reelle data. Chen et al. (2019) påpeker i den forbindelse at selskaper kan omgå personvernutfordringer ved å benytte syntetiske data, på bakgrunn av at det beskytter pasientens personopplysninger og identitet. Syntetiske helsedata kan derfor behandles, lagres og deles uten å gå på bekostning av personvern og sikkerhet (Walonoski et al., 2020). På den måten kan syntetisk data være et nyttig verktøy for situasjoner der tilgangen til reelle data er krevende eller unødvendig (Chen et al., 2019). Et identifisert bruksområde for syntetiske data er forskning, hvor Walonoski et al. (2020) hevder at det kan muliggjøre forskere å frigi meningsfulle data, koder og resultater. Videre påpekes det at syntetiske data kan gi samme resultater som reelle data, men at det forutsetter at de er riktig konstruert og validert.

Ramos og Subramanyam (2021) fremhevet i sin analyse at syntetisk data er fremtiden innenfor kunstig intelligens. For at KI-modeller eller systemer skal kunne generere til sitt fulle potensial kreves det syntetiske data (Ramos & Subramanyam, 2021). I artikkelen fokuseres det på mulighetene syntetiske data gir for KI-utvikling og trening av slike modeller. Reelle data mangler ofte avgjørende kunnskap og ved bruk av syntetiske data kan man tillegge domenekunnskap inn i treningen av KI-modeller. Det kan føre til forbedring av kvaliteten på prediksjonene. Domenekunnskap er ifølge Ramos og Subramanyam (2021) kunnskap som man vet er relevant for KI-modellen og syntetiske data kan derfor bidra til å forbedre slike modeller.

Videre i analysen belyses det hvordan reell data i KI-modeller inneholder kun noen få potensielle scenarioer, mens ved å benytte syntetiske data vil det kunne fullstendiggjøre datasettene betraktelig (Ramos & Subramanyam, 2021). Det kan forklares ved at syntetiske data kan generere utallige sjeldne eller ukjente hendelser. Som følge av dette får man mer fullstendige datasett. Videre fremheves det at syntetiske data kan effektivisere testing av komplekse KI-modeller, gjennom å gi innsikt til ukjente hendelser som ellers ikke vil bli oppdaget med reell data. Dette gjør at syntetiske data i større grad er mer robuste enn reell data innenfor kunstig intelligens. Det ble også avdekket at syntetiske data kan bevare multivariate sammenhenger mellom variabler, som reelle data kan ha vanskeligheter med å finne. Det betyr at ved å benytte syntetiske data muliggjør dette at KI-modeller kan analysere flere variabler som opptrer samtidig (Ramos & Subramanyam, 2021).

På bakgrunn av funnene har Ramos og Subramanyam (2021) utarbeidet en modell som belyser fremtidens behov for syntetisk data. Her tydeliggjøres det at behovet for data som brukes til kunstig intelligens står overfor et drastisk skifte mot 2030. Syntetiske data vil i større grad innta rollen for benyttede data i kunstig intelligens i årene fremover. Siden reelle data i KI-modeller innehar flere begrensninger, slik som kostnader, logistikk og personvernhensyn, kan syntetiske data forstås som løsningen på disse begrensningene. En viktig årsak til dette er å sikre at tilgangen på data øker i takt med det gradvis økende behovet for store mengder data. I Norge har arbeidet begynt for å møte det fremtidige behovet for større mengder data, som kan ses i sammenheng med arbeidet til Norsk Helsenett. Selskapet har et prosjekt som arbeider med generering av syntetiske data, for å berike eksisterende testdata (Direktoratet for e-helse, 2022).



(Figur 2: *By 2030, Synthetic Data Will Completely Overshadow Real Data in AI Models.* Ramos & Subramanyam ved Gartner, 2021).

Andrews (2021) forklarer at bruk av syntetisk data gir generelt mer nøyaktige KI-modeller, som underbygger teorien til Ramos og Subramanyam (2021). Andrews (2021) viser videre til utfordringen med å samle og merke datasett i reelle data, som kan inneholde flere millioner elementer. På bakgrunn av dette hevder han at bruk av reelle data i KI er tidkrevende og fører ofte til høye kostnader. Andrews (2021) påpeker at ved å heller benytte syntetiske data kan det redusere kostnader, samt forebygge bias. Panch et al. (2019) hevder at algoritmisk bias påvirker ulikheter i helsesystemer negativt. Algoritmisk bias er tilfeller der bruken av en

algoritme forsterker eksisterende ulikheter i samfunnet, eksempelvis etnisk bakgrunn, kjønn og seksuell legning (Panch et al., 2019). Bias i forskning kan føre til at resultatene som presenteres ikke er en refleksjon av virkeligheten (Staff, 2015). Syntetiske data kan ifølge Andrews (2021) bidra til å sikre mangfold i data, slik at den virkelige verden i større grad kan representeres i KI-modeller.

2.1.6.2 Utfordringer ved bruk av syntetisk data

Som ved alle typer data er det ikke bare fordeler, og bevisstgjøring rundt mulige utfordringer og begrensninger er derfor viktig. James et al. (2021) trekker frem flere hensyn så må tas ved bruk av syntetisk data. Her blir utfordringene knyttet til produksjon og anvendelse, nye tekniske og organisatoriske tiltak, kompetanse samt ulike former for aksept.
Ifølge James et al. (2021) er en stor utfordring hvordan de syntetiske dataene produseres og hvordan anvendelsen av metoden måles, for å sikre at egenskapene mellom det originale og det syntetiske datasettet er sammenlignbare. Videre påpekes det at anvendelsen av metoden for å produsere syntetiske data kan være tidkrevende. Det betyr at dersom en aktør, slik som helsesektoren, skal generere syntetisk data så vil det mest sannsynlig kreve nye tekniske og organisatoriske tiltak. For å gjennomføre slike tiltak, vil det være behov for nye forretningsprosesser og kompetente personer til å produsere syntetiske data. Kompetente personer anses som viktig, på bakgrunn av behovet for å unngå over- eller undertilpasning av modellene som brukes til å syntetisere data (James et al., 2021). Det hevdes at syntetiske data og dens kvalitet er avhengig av god kvalitet på input dataene og data som maskinlæringsmodellen skal basere seg på (Dilmegani, 2022).

James et al. (2021) trekker også frem utfordringen med å sikre at en unngår reidentifisering. Her hevdes det at det er en risiko for reidentifisering, spesielt i tilfeller der en pasient har verdier som anses som "uteliggere". Uteliggere kan forklares som ekstremverdier og defineres som "*en observasjon (eller undergruppe av observasjoner) som ser ut til å være inkonsistent med resten av datasettet"* (Barnett & Lewis, 1994, referert i Wilkinson, 2018, s. 256). James et al. (2021) eksemplifiserer uteliggere i dette tilfellet som en sjelden sykdom eller demografi. Arora og Arora (2022) beskriver at det ikke finnes robuste og objektive metoder for å avgjøre om et syntetisk datasett er tilstrekkelig forskjellig fra det opprinnelige reelle datasettet. Dermed stilles det spørsmål rundt hvorvidt syntetiske data kan klassifiseres som uidentifiserbar data.

Det er uklart i hvilken grad konklusjoner basert på syntetiske data vil bli akseptert av vitenskapelige og medisinske miljøer (James et al., 2021). På bakgrunn av dette trekkes det frem at tillit til syntetisk data har en stor betydning for hvorvidt syntetiske data vil bli benyttet. Her hevdes det at inntil større tillit til syntetiske data har blitt etablert, vil kliniske miljøer forvente å se resultater direkte demonstrert på virkelige emner.
Aksept innen ulike bransjer går ut på hvorvidt syntetisk databruk vil bli ansett som egnet for et gitt formål. Syntetisk data er et område i vekst og som følge av dette trengs det mer arbeid for å forstå hvordan metodene kan brukes systematisk, og hvordan det kan skreddersys rundt hvert bruksområde (James et al., 2021). Regulatorisk aksept omhandler hvorvidt lover og regelverk vil akseptere deling av syntetiske data i fremtiden (James et al., 2021). Arora og Arora (2022) påpeker at fraværet av lovgivning angående syntetisk data utgjør potensielle risikoer for forbrukerne. I Norge betyr det at til tross for at teknologiselskaper er bundet av personopplysningsloven. når de håndterer kundedata for målrettet annonsering, finnes det ingen åpenbare begrensninger for å spre syntetiske representasjoner av slike sensitive data.

2.2 Personvern

I en stadig mer digitalisert verden hvor lagring, behandling og deling av data står i fokus, har behovet for kontroll og regulering av persondata aldri vært viktigere. Personopplysninger er et begrep som brukes i økende grad i diskusjoner om data og personvern. Datatilsynet (2019a) definerer personopplysninger som alle opplysninger og vurderinger som kan knyttes til deg som enkeltperson. Personvern, sett i lys av data, kan forklares som retten til å ha kontroll over egne personopplysninger, vite hva som er lagret av slike opplysninger, og hvordan opplysningene brukes (Datatilsynet, 2019b). Innsamling, bruk og forvaltning av persondata reguleres strengt, både nasjonalt og internasjonalt. I Norge reguleres persondata i henhold til Personopplysningsloven, som legger føringer for hvordan aktører og forskere skal behandle personopplysninger (Personopplysningsloven, 2018). Loven er således viktig for å fremme deling av data på en ansvarlig og tillitsskapende måte. Vi finner det som konstruktivt å presentere de mest relevante av de grunnleggende personvernprinsippene som regulerer atferden ved behandling av personopplysninger (Datatilsynet, 2019b):

Formålsbegrensning
"Personopplysninger skal samles inn for spesifikke, uttrykkelig angitte og berettigede formål

og ikke videre behandles på en måte som er uforenlig med disse formålene" (Personopplysningsloven, 2018). For å sikre at personopplysninger bare behandles for og til legitime, angitte og spesifikke formål, skal hvert formål identifiseres og beskrives. Formålene skal presenteres på en måte som gir alle involverte en samkjørt forståelse over hva personopplysningene skal brukes til. Formålet må har rettslig grunnlag og samsvare med etiske og rettslige samfunnsnormer. Opplysningene kan ikke gjengis til andre formål enn det opprinnelige (Datatilsynet, 2019b).

Dataminimering

"Personopplysninger skal være adekvate, relevante og begrenset til det som er nødvendig for formålene de behandles for" (Personopplysningsloven, 2018). Prinsippet utfordrer datasamler til å være kritisk til datamengde og kvalitet, som samles inn for å nå formålet med innsamlingen. Mengden av innsamlet data skal ikke overskride det som anses nødvendig for å oppfylle formålet (Datatilsynet, 2019b).

Integritet og konfidensialitet

"Personopplysninger skal behandles på en måte som sikrer tilstrekkelig sikkerhet for personopplysningene, herunder vern mot uautorisert eller ulovlig behandling og mot utilsiktet tap, ødeleggelse eller skade, ved bruk av egnede tekniske eller organisatoriske tiltak" (Personopplysningsloven, 2018). Databehandler plikter å sørge for at innsamlet data ikke går tapt, ødelegges eller skades. Videre plikter datasamler at behandlingen av opplysningene skal foregå slik at integritet, konfidensialitet og tilgjengelighet beskyttes (Datatilsynet, 2019b).

Bergsjø og Bergsjø, (2019) presenterer Datatilsynets ti sentrale personvernutfordringer forbundet med store datasett. Dette er forhold som innebærer en motstrid med de ulike prinsippene for personvern, og det kan derfor være hensiktsmessig å gjøres bevisst på disse utfordringene. Som de relevante utfordringene kan følgende trekkes frem; bruk av data til nye formål, risiko for reidentifisering og feil i faktagrunnlag. Bruk av data til nye formål kan innebære en utfordring i møte med personvernprinsippet om formålsbegrensning, da en sentral del av stordata er gjenbruk av data (Bergsjø & Bergsjø, 2019). Risiko for reidentifisering øker som resultat av stordataanalyser med data fra flere kilder, der enkeltindivider kan identifiseres til tross for anonyme datasett (Bergsjø & Bergsjø, 2019). Her er lagring av helsedata sentralt hvor man ofte ønsker å se på analyse av data på tvers av

flere enheter. Feil faktagrunnlag siktes til at beslutninger som påvirker den enkelte skal baseres på korrekte opplysninger (Bergsjø & Bergsjø, 2019). Stordataanalyse har en svakhet med hensyn til kontekst som kan føre til at beslutninger tas på grunnlag av opplysninger som var ment for et annet formål. Analysens beslutninger gjenspeiler dermed ikke nødvendigvis den aktuelle situasjonen (Bergsjø & Bergsjø, 2019). Dette kan også føre til følgefeil gjennom blant annet videre beslutninger eller at beriking av data skjer i ukorrekt kontekst.

2.2.1 GDPR

For å imøtegå det digitale skiftet, utarbeidet EU en ny personvernforordning, General Data Protection Regulation (GDPR), som skal gi enkeltpersoner mulighet til å få oversikt og kontroll på hvilke data både offentlige og private aktører besitter (Sørebø et al., 2020). GDPR trådte i kraft 25. mai 2018 i EU og gjaldt som norsk rett gjennom personopplysningsloven fra 1. juli samme år. I praksis skal personvernforordningen sørge for at enkeltpersoner må gi samtykke til at selskaper innheter, lagrer og behandler deres personopplysninger. Dette gir enkeltindividet kontroll og eierskap over egne opplysninger (Sørebø et al., 2020). Rammene satt av GDPR, personopplysningsloven og lignende reguleringer påvirker hvordan forskere forholder seg til persondata. Utover dette er det nasjonale forskningsetiske komiteer og utvalg som skal påse at forskning både i offentlig og privat format, skjer i tråd med etiske normer (De nasjonale forskningsetiske komiteene, u.å.). Dette skal sikres gjennom forebyggende arbeid, vedtak, rådgivning og granskning av enkeltsaker.

2.3 Innovasjon

Syntetisk data kan bidra til å endre det vi i dag ser på som data og innovere måten helsedata blir produsert og levert på. På bakgrunn av dette anser vi innovasjon som et vesentlig begrep å inkludere i denne studien.

Innovasjon er et komplekst begrep som kan være vanskelig å definere (Vigar et al., 2020). Begrepet innovasjon har flere forskjellige definisjoner, og kan forstås fra akademiske, økonomiske eller kommersielle perspektiver. I denne studien vil vi se nærmere på innovasjon sett fra et akademisk perspektiv. Joseph. A. Schumpeter, en pionér innenfor feltet, beskrev innovasjon som *"nye produkter, nye produksjonsmetoder, nye forsyningskilder, utforskning av nye markeder og nye måter å organisere virksomheten på"* (Fagerberg et al., 2005, s. 6).

Rapporten "Oslo Manual", som er utarbeidet i samarbeid med Organisation for Economic Co-operation and Development og Eurostat (OECD) (2005), bygger i stor grad på Schumpeter sin definisjon av innovasjon. Her vises det til føringer gjennom et minimumskrav for hva som kan regnes som en innovasjon. Her må produktet eller prosessen være ny, eller betydelig forbedret for at det kan kalles en innovasjon. Denne studien tar utgangspunkt i OECDs definisjon av innovasjon:

> *"En innovasjon er implementering av et nytt eller vesentlig forbedret produkt (vare eller tjeneste), eller prosess, en ny markedsføringsmetode, eller en ny organisasjonsmetode i forretningspraksis, arbeidsorganisasjon eller eksterne relasjoner"* (OECD & Eurostat, 2005, s. 46)

Gjennom å se på OECDs brede definisjon, ser en at begrepet innovasjon kan anvendes på flere mulige innovasjoner. Eksempelvis tjenesteinnovasjoner og prosessinnovasjoner. Videre viser definisjonen at innovasjon er en iterativ prosess der denne initieres av persepsjonen av et nytt marked og/eller en ny mulighet innen tjeneste. En slik mulighet åpner opp for utvikling, produksjon og markedsføringstiltak med mål om kommersiell suksess (Garcia & Calantone, 2002).

2.3.1 Ulike typer innovasjon

Det finnes en rekke forskjellige typer innovasjoner, slik som produkt-, prosess-, markedsførings- og organisatorisk innovasjon (OECD & Eurostat, 2005). Det er interessant for oss å se nærmere på innovasjonstyper, på bakgrunn av at syntetiske data kan muliggjøre innovasjon på flere områder. Samtidig har vi valgt å knytte syntetisk data opp mot tjenesteinnovasjon og prosessinnovasjon, siden vi anser disse innovasjonstypene som særlig relevante sett i lys av problemstillingen. Syntetiske data produseres og analyseres ved hjelp av teknologiske komponenter og som følge av dette anser vi det som hensiktsmessig å introdusere begrepene digital- og datadrevet innovasjon.

Innenfor produktinnovasjon finner vi både vare- og tjenesteinnovasjon, hvor dette defineres som "e*n vare eller tjeneste som er ny eller vesentlig forbedret med hensyn til dens egenskaper eller tiltenkte bruksområder*" (OECD & Eurostat, 2005, s. 48). Syntetisk data

kan knyttes til tjenesteinnovasjon, på bakgrunn av at en slik innovasjon kan utnytte ny kunnskap og teknologi, være en kombinasjon av de nevnte eller være basert på nye bruksområder. Videre beskriver rapporten "Oslo Manual" at produktinnovasjon gjerne innebærer betydelige forbedringer av tekniske spesifikasjoner, komponenter og materialer, integrert programvare, brukervennlighet eller andre funksjonelle egenskaper (OECD & Eurostat, 2005).

En prosessinnovasjon kan defineres som *"implementering av en ny eller betydelig forbedret produksjon eller leveringsmetode"* (OECD & Eurostat, 2005, s. 49). Prosessinnovasjon tar for seg endringer av hvordan tjenester blir laget eller levert (Tidd & Bessant, 2018). Ifølge Karlsson og Tavassoli (2015) innebærer prosessinnovasjon betydelige endringer i teknikker, utstyr og/eller programvare. Videre påpekes det at prosessinnovasjon også kan forstås som ny eller forbedret støtteaktivitet for varer eller tjenester, som for eksempel vedlikeholdssystemer, regnskap eller databehandling. Prosessinnovasjoner kan utformes for å redusere enhetskostnadene ved produksjon eller levering, og for å øke eller forbedre produkt- og leveringskvalitet. Prosessinnovasjoner kan videre føre til økt ytelse ved å redusere kostnader, økt produktivitet og tilgang til ressurser (Hervas-Oliver et al., 2014).

Kombinasjon mellom økende teknologisk endring og rask adopsjon av digitale produkter og tjenester, fører til et større fokus på effektiv utnyttelse av digital innovasjon (Iden et al., 2020). Digital innovasjon kan defineres som "*gjennomføring av nye kombinasjoner av digitale og fysiske komponenter for å produsere nye produkter*" (Yoo et al. 2010, s. 725). Basert på denne definisjonen, forstås digital innovasjon som bruk av digital teknologi under en innovasjonsprosess, og kan brukes til å beskrive resultatet til en innovasjon (Nambisan, 2017). Ifølge Osmundsen et al. (2018) kan digital innovasjon ses i to ulike perspektiv, henholdsvis resultat og prosess. Et resultatorientert perspektiv kan defineres som "*et nytt produkt eller tjeneste som skaper ny verdi for adoptanter, utviklet ved å kombinere digital teknologi på nye måter eller med fysiske komponenter* (Osmundsen et al., 2018, s. 7). Prosessorientertperspektivet dreier seg om *"å kombinere digital teknologi på nye måter eller med fysiske produkter, for å utvikle et nytt produkt eller tjeneste som skaper ny verdi for adoptanter"* (Osmundsen et al., 2018, s. 7). Basert på forskning kan det forstås som at det foreligger en gjensidig avhengighet mellom digital innovasjon som prosess og som resultat (Osmundsen et al., 2018).

Datadrevet innovasjon kan ses som en del av digital innovasjon, på bakgrunn av at begrepene er basert på teknologi og innovasjon. Fenomenet datadrevet innovasjon referer til bruk av integrasjon av dataanalyse og innovasjon (Rizk et al., 2020). OECD definerer datadrevet innovasjon som "*Bruk av data eller analyser for å forbedre eller fremme nye produkter, prosesser, organisasjonsmetoder eller markeder*" (OECD, 2015, s. 21). Sett i lys av definisjonen forutsetter datadrevet innovasjon tilgang på kompetanse, store datavolum og en digital infrastruktur for lagring, bearbeiding og analyse av data (St. meld 22, (2020-2021)).

2.3.2 Innovasjonsprosess

I Norge benyttes syntetiske data i testsituasjoner, og det er interessant for oss å undersøke hvorvidt bruksområdet er overførbart til helsesektoren. På bakgrunn av dette vil vi undersøke om syntetiske data kan gi verdi i en innovasjonsprosess, og i det følgende presenteres teori.

Veien fra en idé oppstår til et ferdig produkt er utviklet, kalles gjerne for en innovasjonsprosess (Aasen & Amundsen, 2011). Slike innovasjonsprosesser blir definert forskjellig, ofte basert på hvilken næring eller bransje den tar utgangspunkt i. Tidd & Bessant (2018) beskriver innovasjonsprosesser som en prosess som handler om å konvertere ideer til virkelighet og å skape verdi fra dem. Den innovative atferden til bedrifter er også preget av ulike institusjoner, slik som lover, regler, normer og rutiner, hvor disse utgjør insentiver og hindringer for innovasjon (Fagerberg et al., 2005).

OECD og Eurostat (2005) trekker frem fire steg i en innovasjonsprosess, herunder ressurser (input), aktiviteter, utganger (output) og resultater. Tidd og Bessant (2018) beskriver det første steget som en søkeprosess, her begynner en bedrift å se etter signaler fra omgivelsene som involverer et potensial for endring. Slike signaler kan komme fra blant annet teknologiske muligheter, behov og trender, konkurrenters atferd og reguleringer som skaper et press på organisasjoner eller endringer i markedet (Tidd & Bessant, 2018). Det neste steget dreier seg om utvalg, hvor hensikten er å løse innspillene fra søkeprosessen til et konsept. I denne fasen poengterer Tidd og Bessant (2018) viktigheten av at utvalget er i tråd med bedriftens strategi og at den potensielle innovasjonen bygger på etablerte teknologiske- og kompetanseområder.

I det tredje steget skal innovasjonen implementeres. Bedriften må ha innsikt i om det er et behov for innovasjonen gjennom å foreta ulike undersøkelser. Mens innovasjonen utvikles gjennomføres det ulike tester for å finne feil og løse disse underveis (Tidd & Bessant, 2018). Ifølge Tidd og Bessant (2018) kan en utfordring i denne fasen være å håndtere et økende press på ressurser i form av tid, penger og å anskaffe kunnskap gjennom FoU-tiltak. Det siste steget tar for seg verdiskaping, som kan forstås som resultatet av innovasjonen (OECD & Eurostat, 2005).

2.4 Helseforskning

Et planlagt anvendelsesområde for syntetiske data i Norge er helseforskning. Det er derfor hensiktsmessig for oss å inkludere litteratur om helseforskning, siden det er av særlig relevans for denne studien.

I Norge har myndighetene en klar forventning til helse- og omsorgstjenestene, hvor disse skal være tilpasset brukerens behov, modernisert og orientert mot forskning (Alstveit et al., 2016). Helseforskning er vitenskapelig forskning som har som formål å frembringe "*ny kunnskap om helse og sykdom*" (Helseforskningsloven, 2009). Innen helseforskning stilles det særlige krav til forsvarlighet, aktsomhet og etisk bevissthet. Spesielt i forskning der mennesker, humant biologisk materiale eller helseopplysninger involveres (NOU 2005:1, 2005). På bakgrunn av dette er helseforskning preget av lange og uforutsigbare søknadsprosesser (Larsen, 2017). Et eksempel på en slik søknad er godkjenning fra De regionale komiteer for medisinsk og helsefaglig forskningsetikk (REK). I Norge kreves det en forhåndsgodkjenning fra REK for å starte og gjennomføre all medisinsk og helsefaglig forskning (Forskningsetikk, 2014). REK har til hensikt å være en kontrollinstans for å forhindre at det blir gjennomført uetiske forskningsprosjekt.

For å løfte forskning innen helse ble det utarbeidet et prosjekt kalt “Helseanalyseplattformen” (HAP), som skal være en nasjonal infrastruktur og plattform for helseanalyse (Åm et al., 2021). Plattformen skal bidra til bedre helseforskning, styrke grunnlaget for kunnskapsbaserte helse- og omsorgstjenester, og samtidig stimulere til innovasjon (Meld. St. 22, (2020-2021)). Her skal helseregistre og andre relaterte datakilder kobles sammen, slik at forskere får en felles plattform, samt redusere tiden forskere bruker på innhenting av data. Plattformen har til

hensikt å benytte syntetiske data levert av Norsk helsenett (NHN), som gjør at løsningen kan håndtere helsedata på en måte som ivaretar personvern (Direktoratet for e-helse, 2019). HAP vil forenkle tilgangen til helsedata for forskere og tilrettelegge avanserte analyser på tvers av ulike datakilder, herunder helseregistre, pasientjournaler, grunndata og andre informasjonskilder (Åm et al., 2021).

Deler av arbeidet med Helseanalyseplattformen er imidlertid satt på vent grunnet Schrems II-dommen. Her avgjorde EU-domstolen at dersom personopplysninger skal overføres til land utenfor EU/EØS, må det være et overføringsgrunnlag i henhold til personopplysningsloven (Digdir, u.å). Deler av arbeidet som er stoppet er knyttet til dataanalyse, mens arbeidet i forbindelse med søknad og saksbehandling foregår fortsatt.

2.5 Kompetanse

På bakgrunn av at syntetisk data er et nytt felt, trengs det kompetanse. Dette kan vi knytte opp til hvordan helsesektoren har bygget et nasjonalt nettverk for kunstig intelligens, "Kunstig intelligens i norsk helsetjeneste", forkortet som KIN (Helsedirektoratet, 2022). Hensikten med dette nettverket er å utvikle, bygge og dele kompetanse på feltet. Med økt bruk av data kan samfunnet spares for kostnader, men det trengs her kompetanse for å møte fremtidens behov i helsesektoren (Østbye, 2020). På bakgrunn av at studien tar for seg det tekniske begrepet, syntetisk data, er det derfor hensiktsmessig å trekke inn digital kompetanse og definere dette.

> *"Digital kompetanse innebærer trygg og kritisk bruk av digitale verktøy og medier til arbeid, fritid og kommunikasjon. Den er underbygget av grunnleggende ferdigheter innen IKT: bruk av datamaskiner for å hente, vurdere, lagre, produsere, presentere og utveksle informasjon, og å kommunisere og delta i samarbeidsverktøy via internett"* (EU-kommisjonen, 2006, s.15-16).

Digital kompetanse krever god forståelse og kunnskap vedrørende de teknologiske applikasjoner som databaser, informasjonsprosessering, lagring samt forståelse for muligheter og risiko knyttet til disse (EU-kommisjonen, 2006). Digital kompetanse er vanskelig å måle siden det rommer mer enn bare tekniske ferdigheter (Aspøy & Andersen 2015; Slettemås 2014). Begge definisjonene anses som hensiktsmessige å introdusere, fordi de samlet gjenspeiler kompleksiteten ved kompetanse.

Dersom Norge skal utnytte mulighetene som KI og andre muliggjørende teknologier gir, vil det blant annet kreve kobling av forskning, teknologi og data på tvers av helseforetak og institusjoner (Pettersen, 2019). Her har blant annet kompetanse en sentral rolle, fordi en trenger kompetanse for å implementere og anvende nye teknologier. Som følge av dette har det blitt et større fokus på de menneskelige ressursene i organisasjoner og deres kompetansenivå (Johansen & Sæterdal, 2017).

I helsesektoren kan maskiner fungere som besluttningsstøte for helsepersonell, eksempelvis maskinlæring brukes til å tolke store mengder data på kort tid og luke ut menneskelige feil (Hokstad, 2019). Videre belyses behovet for at bruk av KI bør kombineres med menneskelige egenskaper som fagkunnskap og etiske betraktninger (Hokstad, 2019). Som følge av dette vil helsesektoren være avhengig av kompetent helsepersonell for å validere algoritmiske utfall fra kunstig intelligens systemer og kontekstualisere disse for pasienter (Stanfill & Marc, 2019).

2.5 Kompetanse

På bakgrunn av at syntetisk data er et nytt felt, trengs det kompetanse. Dette kan vi knytte opp til hvordan helsesektoren har bygget et nasjonalt nettverk for kunstig intelligens, "Kunstig intelligens i norsk helsetjeneste", forkortet som KIN (Helsedirektoratet, 2022). Hensikten med dette nettverket er å utvikle, bygge og dele kompetanse på feltet. Med økt bruk av data kan samfunnet spares for kostnader, men det trengs her kompetanse for å møte fremtidens behov i helsesektoren (Østbye, 2020). På bakgrunn av at studien tar for seg det tekniske begrepet, syntetisk data, er det derfor hensiktsmessig å trekke inn digital kompetanse og definere dette.

> *"Digital kompetanse innebærer trygg og kritisk bruk av digitale verktøy og medier til arbeid, fritid og kommunikasjon. Den er underbygget av grunnleggende ferdigheter innen IKT: bruk av datamaskiner for å hente, vurdere, lagre, produsere, presentere og utveksle informasjon, og å kommunisere og delta i samarbeidsverktøy via internett"* (EU-kommisjonen, 2006, s.15-16).

Digital kompetanse krever god forståelse og kunnskap vedrørende de teknologiske applikasjoner som databaser, informasjonsprosessering, lagring samt forståelse for muligheter og risiko knyttet til disse (EU-kommisjonen, 2006). Digital kompetanse er vanskelig å måle siden det rommer mer enn bare tekniske ferdigheter (Aspøy & Andersen 2015; Slettemås 2014). Begge definisjonene anses som hensiktsmessige å introdusere, fordi de samlet gjenspeiler kompleksiteten ved kompetanse.

Dersom Norge skal utnytte mulighetene som KI og andre muliggjørende teknologier gir, vil det blant annet kreve kobling av forskning, teknologi og data på tvers av helseforetak og institusjoner (Pettersen, 2019). Her har blant annet kompetanse en sentral rolle, fordi en trenger kompetanse for å implementere og anvende nye teknologier. Som følge av dette har det blitt et større fokus på de menneskelige ressursene i organisasjoner og deres kompetansenivå (Johansen & Sæterdal, 2017).

I helsesektoren kan maskiner fungere som besluttningsstøte for helsepersonell, eksempelvis maskinlæring brukes til å tolke store mengder data på kort tid og luke ut menneskelige feil (Hokstad, 2019). Videre belyses behovet for at bruk av KI bør kombineres med menneskelige egenskaper som fagkunnskap og etiske betraktninger (Hokstad, 2019). Som følge av dette vil helsesektoren være avhengig av kompetent helsepersonell for å validere algoritmiske utfall fra kunstig intelligens systemer og kontekstualisere disse for pasienter (Stanfill & Marc, 2019).

2.6 Tillit

Når vi skal undersøke syntetisk data i helsesektoren vil tillit være en av flere faktorer som kan påvirke villigheten til anvendelse. Det finnes lite forskning på tillit til syntetiske data, og som følge av dette er det aktuelt å se til nærliggende teori hvor det finnes likhetstrekk. På bakgrunn av at syntetisk data hovedsakelig består av KI-komponenter anser vi teori om tillit knyttet til kunstig intelligens som sammenlignbart.

Begrepet tillit er kjent for å være vanskelig å definere eller måle (Rousseau et al., 1998). Gambetta (1988) hevder at det er et vagt begrep som er vanskelig å beskrive og forstå. Den britiske filosofen Onora O'Neill, mener det er tre essensielle elementer som inngår i

tillit, herunder kompetanse, ærlighet, og pålitelighet. Kompetanse handler om hva vi gjør, ærlighet handler om hva vi sier, mens pålitelighet dreier seg om at vi utfører oppgavene som vi forplikter oss (Ingierd, 2017). Videre knytter O´Neill kompetanse opp mot troverdighet, hvor hun påstår at troverdighet kommer før tillit. Johnson-George og Swap (1982) hevder at vilje til å ta risiko ses på som en av de få egenskapene som er felles for alle situasjoner som innebærer tillit. Basert på arbeidet til Gambetta (1988) og Johnson- George og Swap (1982) utarbeidet Mayer et al. (1995, s. 712) en forenklet definisjon av tillit, hvor dette defineres som *"villigheten til å ta risiko"*. Videre forklarer de at denne definisjonen inneholder en vesentlig faktor for tillit, nemlig sårbarhet.

Ifølge Gillath et al. (2021) er mangel på tillit en av de største hindringene som står i veien for å dra full nytte av fordelene kunstig intelligens har å tilby. En årsak til manglende tillit til kunstig intelligens kan være at en ikke stoler på dens beslutninger. Ofte handler dette om at kunstig intelligens ikke kan forklare hvorfor eller hvordan det har kommet frem til en beslutning. Dette kalles for black box problematikken (Gille et al., 2020). Gillath et al. (2021) forklarer videre at å skape tillit til kunstig intelligens handler innledningsvis om å redusere oppfatningen av risiko knyttet til bruk.

2.7 Konklusjon

I dette kapittelet har vi presentert aktuelle begreper og teorier for å gi en innsikt i fenomenet syntetisk data, sett i kontekst av helsesektoren og helseforskning. De siste årene har helsesektoren tatt store steg i bruk av KI-teknologi. Definisjonen av data viser omfanget av hvor mye som innbefattes i begrepet. Gjennom KI, maskinlæring og stordata er utnyttelsen av data på et høyere nivå enn det som tidligere har vært oppnåelig. For å at analysene gir verdi og er av best mulig kvalitet, er datakvalitet avgjørende.
Personvern og lovverk er to viktige faktorer som i stor grad preger produksjon, behandling og deling av helsedata. Syntetiske data er data som ikke inneholder personopplysninger og kan benyttes for å sikre tilstrekkelige mengder av helsedata, så vel som å berike eksisterende reelle data. Teori om innovasjon og helseforskning ble introdusert og fremlagt som potensielle områder for bruk av syntetiske data. Bruk av syntetiske data krever kompetanse og at brukerne har tillit til både teknologien og dataene. Gjennomgangen av litteratur har vist at det er relativt lite forskning på bruk av syntetiske data, og vi vil med denne oppgaven bidra med kunnskap relatert til utnyttelse og utfordringer.

Dette kapittelet har belyst kompleksiteten som angår syntetiske data og hvilke implikasjoner bruken av en slik teknologi kan medføre. Tematikken som er presentert danner det teoretiske grunnlaget som skal ses sammen med innsamlet data. I det kommende kapittelet presentere vi vår metodiske tilnærming til studien.

3.0 Metode

I dette kapittelet skal vi ta for oss den metodiske tilnærmingen som ligger til grunn for å besvare problemstillingen og forskningsspørsmålene som studien reiser. Først presenterer vi metoden, tilnærmingen og designet for forskningen. Videre fremlegges bakgrunn for datagrunnlag og analyse av datamaterialets kvalitet. Avslutningsvis tar metodedelen for seg etiske betraktninger og hensyn til personvern.

3.1 Forskningsdesign

Forskningsdesign er ment som en overordnet plan for studien, som har til hensikt å forklare hvordan problemstilling skal belyses og besvares (Easterby-Smith et al., 2018). Hensikten med studien vår er å undersøke hvilken betydning bruk av syntetisk data kan ha for helsesektoren. Det belyses ved å se på hvilke muligheter som kan skapes, samt hvilke utfordringer en bør være bevisst. På bakgrunn av at det foreligger lite forskning på feltet fra tidligere, har vi valgt å anvende et eksplorativt forskningsdesign. I eksplorativt design har forskeren lite kunnskap om forskningsområdet fra tidligere, men ønsker å få en dypere innsikt i temaet for å forstå og tolke fenomenet som skal forskes på (Gripsrud et al., 2010). Det samsvar med våre forutsetninger for denne studien. Videre foreligger det to hovedteknikker for datainnsamling i eksplorativt design, henholdsvis fokusgrupper og individuelle dybdeintervju, vi har valgt å benytte oss av sistnevnte.

3.2 Forskningstilnærming

Forskningstilnærming er forbeholdt i hvilken grad en benytter seg av allerede etablert teori (Jacobsen, 2015). Det skilles mellom to hovedtilnærminger; induktiv og deduktiv tilnærming. Tilnærmingen man velger å benytte seg av, avhenger av i hvilken grad en anvender allerede etablert teori. En induktiv tilnærming handler om at forskeren ønsker å samle inn data for å

kunne etablere egne teorier (Jacobsen, 2015). En deduktiv tilnærming dreier seg om å bruke eksisterende teori som grunnlag og teste den gjennom et forskningsprosjekt En abduktiv tilnærming er en kombinasjon av de to nevnte (Jacobsen, 2015). Her kan man starte med å ta utgangspunkt i empiri og anvende teori underveis i forskningsprosessen for å forstå empirien (Easterby-Smith et al., 2018). Vi anså det som hensiktsmessig å velge abduktiv tilnærming for vårt forskningsprosjekt. Grunnlaget for det er at vi ønsket å ha en åpen tilnærming som tilrettelegger for å innhente ny informasjon for å besvare problemstillingen, samtidig som det ses i lys av teori (Jacobsen, 2015).

3.3 Kvalitativ forskningsmetode

Når man skal samle inn data for et forskningsprosjekt skilles det mellom kvalitativ og kvantitativ metode i forskningslitteraturen (Easterby-Smith et al., 2015). Kvantitativ metode kan ses statistisk med datagrunnlag fra et større utvalg, hvor dataene blir representert gjennom tallverdier. I kvalitativ metode foreligger som regel datamaterialet i tekstform og er basert på et mindre utvalg (Askheim & Grenness, 2008). Vår studie er basert på en kvalitativ forskningsmetode, for å få et dypere innblikk og en større forståelse av fenomenet syntetisk data (Yin, 2018). Videre har vi valgt kvalitativ metode på bakgrunn av at vi ønsker å gjennomføre dybdeintervju for å få innsikt i informantenes erfaringer, tolkninger og meninger om tematikken. På den måten vil vi i vesentlig grad åpne opp for at datamaterialet kan føre til at den overordnede problemstillingen og forskningsspørsmålene blir besvart.

3.4 Datagrunnlag

For å kunne besvare forskningsspørsmålene i denne oppgaven er vi avhengig av datainnsamling. Basert på det er det derfor vesentlig å på forhånd ha en plan om hvordan vi ønsker å hente inn data, hvordan vi skal analysere data og hvordan vi har til hensikt å bruke data (Yin, 2014). Primærdata går ut på at forsker selv må samle inn de nødvendige dataene (Easterby-Smith et al., 2018). I vårt forskningsprosjekt har vi tatt utgangspunkt i kvalitativ primær datainnsamling, som blir supplert med litteratur. Det begrunnes med at det foreligger lite forskning på feltet og at vi ønsket å anvende data som var direkte knyttet til vår problemstilling. Vi har foretatt datainnsamling gjennom intervjuer i form av individuelle dybdeintervju, hvor dette anses som studiens primærdata (Halvorsen, 2014). Våre informanter vil dermed ha en sentral rolle i denne oppgaven. Primærdata har følgelig blitt

supplert og sett i sammenheng med etablert teori. Søk etter litteratur har i all hovedsak foregått gjennom søkemotorene Oria og Google Scholar. Vi har benyttet oss av publiserte artikler, relevante fagtekster, rapporter og fagbøker. I tillegg til informasjon fra informantene våre, har vi gjennomført desktop undersøkelser. På bakgrunn av at vår studie er inspirert av innovasjonsprosjektet Pilot SYNDATA, har de bistått oss med verdifull innsikt fra deres prosjekt. De har delt dokumenter, fagartikler, møter og annen betydningsfull informasjon som har vært nyttig for vårt forskningsprosjekt.

3.4.1 Kvalitative primærdata

Innsamling av primærdata har vært avgjørende for studien vår. Det kan være ressurskrevende å samle inn data, på bakgrunn av at det krever tid til forberedelser, gjennomførelse og transkribering av intervju (Easterby-Smith et al., 2018). Det er flere måter å samle inn primærdata på og for vår studie så vi det som hensiktsmessig å benytte oss av semistrukturerte dybdeintervju. I alt har vi foretatt oss 10 intervju, med informasjon fra 11 informanter. Forutenom et intervju, ble samtlige intervju gjennomført som individuelle dybdeintervju. I kvalitativt intervju er målet å få en forståelse av informantens perspektiv (Easterby-Smith et al., 2018). Det inkluderer å få innsikt i individets synspunkter, samt forståelse for hvorfor de har det bestemte synspunktet. Før vi startet datainnsamlingen ønsket vi å gjennomføre intervjuene ansikt til ansikt, siden det er med å etablere tillit og åpenhet, samt flyt og sammenheng i intervjuet. På grunn av covid-19 og andre praktiske årsaker, som avstander, valgte vi likevel å gjennomføre intervjuene digitalt. Det begrunnes med at enkelte informanter var utilgjengelige for fysisk gjennomføring og for å sikre at planlagte intervju kunne gjennomføres i tråd med anbefalinger og restriksjoner i samfunnet på daværende tidspunkt.

3.4.2 Valg av informanter

I henhold til vår problemstilling ønsket vi å rekruttere informanter som innehar dyptgående kunnskap og erfaring innenfor syntetiske data, innovasjon og helseforskning. Vi anså det som relevant å snakke med mennesker med ulike perspektiver og tilnærminger til tematikken, på bakgrunn av at det kan generere forskjellige innfallsvinkler og synspunkt. Ved å ha benytte oss av et strategisk utvalg, ønsket vi å øke sannsynligheten for å innhente relevant data (Johannessen et al., 2016). Det finnes ulike måter å utforme et strategisk utvalg på og vi fant

det hensiktsmessig å benytte oss av kriteriebasert utvalg, i tillegg til snøballutvelgelse. Kriteriene som ble satt for vår studie var at informantene skulle ha kunnskap og kompetanse innen syntetiske data, innovasjon og helseforskning (Johannessen et al., 2011). Innovasjonsprosjektet SYNDATA hjalp oss med å knytte kontakt med potensielle informanter. Dette var svært behjelpelig for oss i søken om å finne aktuelle personer med relevant kompetanse innenfor det feltet vi ønsket å undersøke. Vi fikk tilsendt en interessentliste med en rekke navn. Det startet en søkeprosess fra vår side, hvor vi undersøkte hvem vi ønsket å sende en henvendelse om å delta i vårt forskningsprosjekt. Dette var kandidater som ble ansett til å kunne bidra med faglig kompetanse og innsikt i henhold til tematikken. For å rekruttere de aktuelle kandidatene tok vi direkte kontakt via e-post, hvor resten av kommunikasjonen opptil intervjuet foregikk. I løpet av intervjuprosessen benyttet vi oss av snøballutvelgelsen, hvor vi forhørte oss med informantene våre om noen hadde andre aktuelle kandidater som vi burde kontakte (Easterby-Smith et al., 2021). I de tilfellene hvor vi fikk konkrete forslag, undersøkte vi muligheten og nytten av å kontakte disse.

3.4.3 Semistrukturerte dybdeintervju

Vi har tatt utgangspunkt i å bruke semistrukturerte dybdeintervjuer som kvalitativ datainnsamlingsmetode i vårt forskningsprosjekt. Det er en struktur som tilrettelegger for at en skal kunne være fleksibel og dynamisk, ved å stille oppfølgingsspørsmål i løpet av intervjuet (Easterby-Smith et al., 2018). Samtidig er det en delvis styrt struktur, med utgangspunkt i en intervjuguide som er utarbeidet på forhånd. Den inneholder spørsmål vi ønsket å få avdekket underveis i intervjuene. For å sikre at intervjuguiden inkluderte spørsmål som ville innhente relevant data for våre forskningsspørsmål, så vi det som hensiktsmessig å til tidligere akademisk arbeid relatert til vårt. Det gjorde vi ved å benytte vi Oria (Høgskulen på Vestlandets kunnskapsdatabase) for å søke opp akademiske tekster og tidligere forskning på syntetiske data. Vi leste gjennom flere tekster og identifiserte hvilke tema som ble belyst der. Etter denne prosessen satt vi igjen med temaene; syntetisk data, innovasjon, helseforskning og data. Dette var temaer vi ønsket å stille spørsmål om, eksempelvis:

- Hva anser du som de største utfordringene med innsamling og behandling av data?
- Hva vet du om syntetiske data?
- Hva er dine tanker om bruk av syntetiske data i forskning?
- Hvordan ser du på forskning i relasjon til innovasjon?

- Hvilke tanker har du rundt innovasjon og forskning?

Videre har vi stilt oppfølgingsspørsmål som "*hva legger du i det begrepet?*" eller "*når du sier ..., hva mener du da?*". Som følge av at vi stilte oppfølgingsspørsmål i løpet av intervjuene, ga informantene oss informasjon som nødvendigvis ikke var i tråd med intervjuguiden. Det viste seg likevel å gi verdifull innsikt for å belyse problemstillingen. Vi anså det som fordelaktig at alle informanter i samme kategori ble stilt de samme spørsmålene fra samme intervjuguide. Det begrunnes med å være konsistent ved innsamlingen av data, samt at datainnsamlingen er relevant for problemstillingen. Det er essensielt at spørsmålene vi har stilt ble oppfattet og forstått av informantene, da misforståelser mellom intervjuer og intervjuobjekt kan skade resultatene i undersøkelsen. For å unngå dette har vi sendt ut intervjuguiden på forhånd, slik at informanten kan sette seg inn i den og eventuelt henvende seg til oss dersom noe er uklart.

3.4.4 Gjennomføring av intervju

Intervjuene ble gjennomført digitalt gjennom digitale kommunikasjonsplattformer som Google Meet og Microsoft Teams. Til tross for enkelte barrierer et digitalt intervju kan medføre, som eksempelvis tolkning av kroppsspråk og sittestilling, anså vi det som hensiktsmessig grunnet covid-19 pandemi (Skorstad, 2018). Det begrunnes ved at vi fikk en effektiv og fleksibel intervjuprosess til tross for sykdom, avstand og andre faktorer som kunne ha påvirket gjennomføringen av fysiske intervju (Easterby-Smith et al., 2018).
Vi har gjennomført 10 intervjuer, hvor vi anser gjennomførelsen av intervjuene som innsiktsfulle. I flertallet av intervjuene har alle tre prosjektmedlemmer vært tilstede, og i forkant ble det fordelt roller og ansvarsområder. Det var to som styrte intervjuet, hvor en hadde primæransvar for å følge intervjuguide, mens den andre skulle være observatør og komme med eventuelle oppfølgingsspørsmål. Den tredje personen har vært skribent og har hatt ansvar for å notere ned fortløpende.

I invitasjon til intervjuet har vi opplyst informanten om at intervjuet er beregnet å vare i cirka 60 minutter, og det er denne tiden vi har satt av til hver informant. De fleste intervjuene varte mellom 40-50 minutter. Erfaringen dette ga oss var at vi fikk god tid til å avrunde med intervjuobjektet, hvor vedkommende kunne stille spørsmål tilbake til oss dersom noe har vært

uklart. I tillegg til det nevnte, ga det også informanten tid til å legge til informasjon som ikke er kommet frem underveis i intervjuet, dersom dette var ønskelig.

Utvalg av informanter som ble intervjuet

Tabellen nedenfor presenterer en oversikt over informantene i studien. Vi har valgt å kategorisere bransjen vedkommende jobber i, samt deres stillingstittel. I de to siste kolonnene vises intervjuets varighet, oppgitt i minutter, samt deres informantnummer.

Bransje	**Stillingstittel**	**Tid på dybdeintervju**	**Informant**
Helse og IKT	Data Scientist	41:22	1
Helse og IKT	Spesialkonsulent	43:28	2
Offentlig og IKT	Virksomhetsarkitekt	37:35	3
Forskning og IKT	Leder	36:33	4
Forskning, IKT og utdanning	Produktområdeleder	48:08	5
Helse og omsorg	Klinikksjef	35:35	6
Helse og utdanning	Forskningsrådgiver	47:37	7
Offentlig og IKT	Avdelingsdirektør	59:11	8
Helse og IKT	Leder	32:47	9
Offentlig og IKT	Testleder	44:38	10
Offentlig og IKT	Testansvarlig	44:38	11

Tabell 1: Tabell som viser utvalget av informanter som ble intervjuet

Det ble gjennomført 10 dybdeintervjuer med elleve informanter, som alle hadde en tilknytning til helse eller IKT og som har ulike stillinger i de respektive selskapene (Tabell 1). Lengden på intervjuene varierte fra 32:47-59:11, hvor informantens erfaring og innsikt i syntetiske data var en faktor for lengden. Videre kan valg av semistrukturerte dybdeintervju også ha påvirket lengden på intervjuene. Etter gjennomføring av intervjuene, ble all innsamlet data analysert.

3.5 Analyse av data

Etter at datamaterialet er samlet inn, er det hensiktsmessig å tydeliggjøre en fremgangsmåte for hvordan materialet skal analyseres. Ifølge Easterby-Smith et al (2018, s. 235) bør datamaterialet kategoriseres og organiseres før en skal analysere det. Det kan være krevende da det ikke foreligger en tydelig fremgangsmåte eller prosedyre for hvordan en slik analyse skal utføres. Likevel er det vanlig å undersøke, kategorisere, sette opp i tabell, teste eller på en annen måte sammenslå empiriske funn (Yin, 2018). Siden vi gjennomførte 10 dybdeintervju på 30-60 minutter, var det en stor mengde data som skulle gjennomgås. Vi bestemte derfor at det ville være nyttig å transkribere samtlige intervjuer. Dette ville samtidig gi oss en mer oversikt over de fysiske data.

For å få en oversikt startet vi analysen med kategorisering og organisering av det transkriberte datamaterialet. Vi satt opp funnene våre i en tabell og sorterte disse i underkategorier som samsvarte med det teoretiske rammeverket for studien. Disse er datakvalitet, lovverk, tillit, innovasjon, forskning og kompetanse, og som anses som temaene i studien. Videre valgte vi å skille mellom hovedfunn og andre funn. På den måten kunne vi hente frem vesentlig informasjon og sitater som er blitt gjengitt i kapittel fire. Ved en slik analyse fikk vi en helhetlig oversikt over informantenes perspektiv sett i lys av tematikkene, samtidig som vi fikk se de i sammenheng med hverandre. Hovedfunn er funn som kan direkte knyttes opp mot problemstillingen for studien, mens andre funn er informasjon som kan sette hovedfunn i en større kontekst. Disse funnene kan også bidra til å belyse andre faktorer som ikke denne studien fokuserer på, men som kan være relevant for videre forskning. Funnene blir presentert etter muligheter og utfordringer, hvor de blir diskutert opp mot relevante teori.

3.6 Vurdering av datamaterialets kvalitet

Underveis i gjennomføringen av en kvalitativ studie er det hensiktsmessig å vurdere kvaliteten på det innsamlede datamaterialet som brukes i forskningsarbeidet. Studiets kvalitet er avhengig av hvordan forskerne tilnærmer seg forskningsprosessen fra utarbeidelse av hypoteser og forskningsspørsmål, til forskningsprosjektet blir publisert (Easterby-Smith, 2018). I kvalitativ forskning er det viktig å vurdere datamaterialets styrker og svakheter, samt valg av litteratur. Ved innhenting av litteratur er det vesentlig å være bevisst på hva en leter etter, samtidig som det er viktig å være kildekritisk og vurdere kildens relevans sett i lys av egen studie. Ifølge Yin (2018) kan kvaliteten av et studiedesign vurderes ut i fra studiens validitet og reliabilitet. I utførelsen av undersøkelser bør en alltid forsøke å minimere vanskeligheter tilknyttet validitet og reliabilitet (Jacobsen, 2018). Andre faktorer som kan være med på å vurdere forskningens datakvalitet er overførbarhet og troverdighet (Shenton, 2004). Videre kan det være aktuelt å kartlegge hvorvidt funn gjort i datainnsamlingen samsvarer med etiske betraktninger og personvern (Yin, 2018).

3.6.1 Validitet

Validitet handler om studiens gyldighet og pålitelighet sett i lys av resultatene (Gripsrud et al., 2016). Begrepet kan betegnes som hvor godt en klarer å måle det en har til hensikt å måle (Easterby-Smith et al., 2015). Man skiller mellom intern og ekstern validitet (Easterby-Smith et al., 2015).

Intern validitet

Intern validitet er ment som en forsikring om at resultatene i forskningen er sanne, og at konklusjonene som trekkes er korrekt ved å eliminere systematiske kilder til mulig skjevhet (Easterby-Smith et al., 2018). Kvalitative studier har gjerne sterk intern validitet, men i liten eller ingen grad ekstern validitet (Johannessen et al., 2016). Dette kan forklares ved at i en kvalitativ studie er forskeren interessert i informantenes subjektive meninger, erfaringer eller tanker. Det er derfor ikke gitt at disse er representative. Fokuset på å intervjue informanter av relevans, som har gitt oss verdifull innsikt om temaet, har bidratt til å sikre høy grad av intern validitet. Det begrunnes med kriteriebasert utvalg av informanter og snøballutvelgelsen. Vi anser vårt utvalg som valid, samtidig som at vi ser at det kan være en svakhet at vi kun har 11 informanter. Det er derfor ikke gitt at denne undersøkelsen er representativ. Videre har vi

gjennom vår studie forsøkt å ha fokus på høy grad av intern validitet ved å presentere dataene korrekt, samt vise til holdbarhet i de slutningene vi har tatt basert på funnene.

Ekstern validitet

Ekstern validitet beskriver i hvilken grad resultatene av forskningen kan generaliseres til andre setting eller kontekster (Easterby-Smith et al., 2018). Forskningsprosjektet har fokus på tematikker som innovasjon og forskning, som kan ses å være overførbart til andre sektorer og bransjer. På bakgrunn av det kan man stadfeste noe ekstern validitet i studien. Likevel ser vi som følge av at studien ikke har et stort utvalg av informanter, at det gjør generalisering av funn utfordrende. Til tross for dette mener vi at studiet bidrar til å belyse og gi en større innsikt for et tema som ikke er veletablert i norsk helsesektor eller helseforskning. Vi antar likevel at fenomenet syntetisk data vil være sentralt i forbindelse med den teknologiske utviklingen vi ser i samfunnet.

3.6.2 Reliabilitet

Reliabilitet omhandler hvor pålitelige og troverdige forskningsresultatene er (Easterby-Smith et al., 2018). En enkel forklaring av reliabilitet er hvor godt vi måler det vi har til hensikt å måle (Gripsrud et al., 2016). En forutsetning for å oppnå reliabilitet er at forskningen samsvarer med virkeligheten, og at det er stabilitet i målingene (Easterby-Smith et al., 2018). Det innebærer at det skal være mulig for en forsker å gjennomføre en identisk studie og avdekke de samme funnene og trekke en tilsvarende konklusjon. For å imøtekomme dette er man avhengig av hvilke data som benyttes, hvordan datamaterialet samles inn og hvordan det blir bearbeidet i etterkant (Johannessen et al., 2016). Det kan imidlertid være vanskelig å etterprøve kvalitativ forskning, på bakgrunn av at forskernes personlige egenskaper og kunnskaper kan påvirke forskningsprosessen. Målet er å minimere bias og menneskelige feil. Transparens i forskningen og detaljerte beskrivelser kan imidlertid bidra til å styrke reliabiliteten. Det begrunnes med at det gir et bedre utgangspunkt for å etterprøve en kvalitativ studie (Easterby-Smith et al., 2018). Det å gjennomføre semistrukturerte dybdeintervju, kan ha hatt en negativ effekt i henhold til reliabilitet i studien. Det begrunnes med at semistrukturerte intervju tilrettelegger for at hver samtale er dynamisk og endres i takt med informantens svar, dermed kan det være vanskelig å få en tilnærmet lik gjenskaping. For å styrke reliabiliteten i den grad det er mulig i vår studie, har vi derfor valgt å beskrive vår forskningsprosess i detalj i metodedelen. Det vil si valgene for vi har foretatt oss i henhold til

datainnsamling, bearbeiding av data og hvordan vi kom frem til funnene i studien. I tillegg har vi valgt å vedlegge intervjuguide og et transkribert intervju for mest mulig transparens. Med det håper vi at leseren kan få en forståelse for de valg og prosedyrer vi har fulgt, som kan bidra positivt for studiens reliabilitet.

3.6.3 Utfordringer og kritisk vurdering av data

Underveis i datainnsamlingen observerte vi at det var problematisk å få konkrete svar og eksempler fra enkelte av våre informanter. Vi ble da nødt til å se over intervjuguiden vår og reflektere over måten vi stilte oppfølgingsspørsmål på. Etter at vi ble bevisst på dette, ble vi bedre på å stille direkte oppfølgingsspørsmål. Vi ba da intervjuobjektet om å komme med konkrete eksempler på positive eller negative virkninger ved bruk av syntetiske data inn mot helseforskning.

Videre ble alle intervjuene gjennomført via digitale plattformer, noe som kan ha påvirket kvaliteten av intervjuet. Det forklares med at det kan oppstå barrierer hvor det vanskeligere å tolke og observere kroppsspråk, sittestilling og blikkontakt. En annen begrensning ved studien er at utvalget ikke kan regnes som representativt. Likevel vil vi påpeke at vi har et tilstrekkelig datagrunnlag for å belyse studiens problemstilling og forskningsspørsmål.

3.7 Etiske betraktninger og personvern

Et forskningsprosjekt må forholde og underordne seg etter etiske prinsipper og juridiske retningslinjer (Johannessen et al., 2011). Easterby-Smith et al. (2018) fremlegger 10 etiske retningslinjer for forskning. Prinsippene kan ses i to deler, og omhandler hvem eller hva som skal beskyttes. I den første delen er målet å beskytte informantens interesser. Forskeren må påse at en *ikke skader* informanten, samt ivareta og respektere informantens *verdighet*. Videre skal forskeren sørge for å motta et *fullt informert samtykke* fra alle informanter og beskytte *personvernet* deres. Forskeren må overholde *taushetsplikten* og sikre *konfidensialitet* i datamaterialet. Den andre delen i prinsippene retter søkelys mot å ivareta integritet opp mot forskningsmiljøet. Det er mulig ved å være *transparent* i forskningsprosessen.

Gjennom hele studiens forløp har vi hatt fokus på å ta etiske hensyn overfor de involverte partene. Før vi startet datainnsamlingen meldte vi prosjektets plan og formål til Norsk senter for forskningsdata (NSD). NSD er en instans som må godkjenne søknader om forskning, før det er aktuelt å starte med og samle inn data. Før man setter i gang med datainnsamlingen er det viktig at informanten har lest, forstått og samtykket til å delta i forskningsprosjektet, og gjort seg kjent med forskningens formål og fremstilling (Easterby-Smith, et al., 2018). Vi utarbeidet en samtykkeerklæring, som informantene fikk tilsendt på forhånd av intervjuene og ble bedt om å sende tilbake med signatur. Samtykkeerklæringen skal ivareta informantens interesser. Erklæringen presiserer studiens formål, informantens rett til innsyn, taushetsplikten vi som forskere er ansvarlig for, plan for oppbevaring av datamaterialet, samt hvor lenge vi hadde til hensikt å ha materialet lagret. Videre har vi forholdt oss til gjeldende regler og retningslinjer for innhenting, lagring og oppbevaring av data for å sikre informantenes personopplysninger. Det inkluderer følgelig at vi har fulgt Høgskulen på Vestlandet sine retningslinjer for behandling av personvernopplysninger og helseforskningsdata.

Vi ønsket som nevnt å ivareta informantene i hele studiens forløp. Det har vært gjeldende for vår datainnsamling, hvordan vi har analysert og tolket datamaterialet, i tillegg til presentasjon av funnene. Formålet var å sikre at vi har tolket informasjonen fra informanten korrekt. Videre har vi forsøkt å opptre respektfulle overfor enhver person som har vært delaktig eller hjulpet oss med å gjennomføre denne studien. Vi har hatt fokus på å være åpen og ærlig underveis i forskningsprosjektet, både i intervjusetting, med veileder og med andre fagfolk vi har vært i kontakt med. I det legger vi åpenhet om vår kunnskap og kompetanse på feltet, samt hvordan vi ligger an i prosessen. Som en del av forskningsetikken vil vi påpeke at vårt forskningsprosjekt er uavhengig forskning.

4.0 Analyse

I følgende kapittel presenteres empiriske funn gjort i studien. På bakgrunn av at vi har valgt å anonymisere alle informantene våre, vil de bli henvist ved informant og et spesifikt nummer for bruk av sitater. For å presentere funnene på en ryddig og oversiktlig måte har vi valgt å dele de inn i de to hovedtematikkene for studien. Henholdsvis muligheter og utfordringer, med flere underkategorier for hvert emne. Avslutningsvis oppsummerer vi hovedfunnene fra datainnsamlingen i en tabell.

4.1 Muligheter

For å belyse mulighetsrommet for syntetisk data i helsesektoren, ønsket vi å høre informantenes synspunkter og meninger knyttet til fenomenet. I den forbindelse skal vi presentere funn knyttet til lovverk, forskning, innovasjon, kompetanse og data.

4.1.1 Lovverk

Et gjennomgående funn omhandlet hvordan syntetisk data kan redusere eller potensielt fjerne barrierer knyttet til GDPR. *"Syntetiske data har et veldig stort potensial, siden hvordan vi behandler data er veldig regulert og som følge av dette kan vi bare jobbe på et begrenset område"* (Informant 1). Som følge av dette fremkom det at syntetisk data kan muliggjøre bredere behandling av data. Flere informanter sammenlignet syntetisk data opp mot reelle data, som er data virksomheter og forskere bruker i dag. Reelle data bringer med seg mange regulative hensyn både ved bruk, databehandling og oppbevaring. Dette gjør det vanskeligere å utnytte potensialet til dataene maksimalt. Det ble hevdet at en står mye friere ved å benytte syntetisk data, fordi man kan gjøre flere analyser uten å gå på bekostning av regelverket. Det kom tydelig frem fra intervjuene at nøkkelfunksjonen til syntetisk data er å unngå de regulatoriske barrierene i form av personvern.

Et annet funn omhandler datalagring, hvor for eksempel lagring av data på amerikanske skytjenester er ulovlig, som følge av Schrems II-dommen. Norge må derfor forholde seg til norske eller europeiske skyløsninger, til tross for at markedet er dominert av amerikanske teknologigiganter.

> *"Ved bruk av persondata må vi vite hvor servere som skal lagre data er lokalisert, dersom det er utenfor EU, byr dette på store utfordringer. Bruk av syntetiske data, fjerner denne barrieren"* (Informant 7).

Det fremkom at syntetisk data imidlertid gjør det mulig å benytte skyløsninger for datalagring utenfor Europa. I EU/EØS er det pålagt å opptre i samsvar med GDPR, mens utenforstående land opererer med andre reguleringer knyttet til data. Lagring av data med amerikanske retningslinjer er derfor problematisk sett i lys av reguleringer. Her kan bruk av syntetiske data ha en verdi. Informantene knyttet dette til Schrems II-dommen og hvordan det har resultert i at arbeidet med dataanalyse i Helseanalyseplattformen (HAP) er satt på pause. På bakgrunn av dette har det vært nødvendig for HAP å se på alternative løsninger for å kunne realisere data- og analysetjenestene som skal leveres. Flere av informantene trakk frem hvordan syntetisk data var den åpenbare løsningen, og at HAP har begynt arbeidet med å syntetisere data, slik at det kan tilbys som en erstatning for reelle data.

Det ble avdekket flere fordeler med syntetisk data, spesielt tilknyttet redusert risiko. I uttalelsen som presenteres, ble syntetiske data sammenlignet med anonymiserte data og hvordan bruk av syntetisk data muliggjør deling på tvers av avdelinger og sektorer. Det kom også frem at ved å bruke syntetisk data kan en få et mer helhetlig bilde på den syntetiske personen.

> *"Syntetisk data har ikke de begrensningene som anonymiserte data har. Det kan brukes mye friere, det er ikke noe bekymring eller fare for lekkasje, personvern og den type ting. Det åpner en helt annen mulighet for å både dele data og snakke om data. Fordi vi kan ha hele bildet på de syntetiske personene, istedenfor å bare bruke deler av en anonymisert person"* (Informant 10).

Funn viste at risiko for lekkasje blir fjernet ved bruk av syntetisk data. Dette kan knyttes til testing, som flere informanter poengterte at er et etablert bruksområde for syntetisk data. Syntetisk data har vært på radaren til flere av informantene, fordi det ville gjort testing og utvikling enklere. I den forbindelse ble det løftet frem hvordan syntetisk data kan være en mulig løsning på GDPR problematikk i en testfase på helseløsninger.

> *"EU har store ambisjoner og det som skjer nå i digital helse, er at en god del av helseløsningene får mer KI komponenter i seg. Da må man teste de løsningene både før og under utvikling. Under en slik utvikling kan det være ganske verdifullt å ha tilgang til syntetisk data sånn at man kan teste i et GDPR fritt landskap, altså at man ikke trenger å ta hensyn til det*" (Informant 11).

Ved å benytte reelle data til testaktiviteter kan dette medførte store konsekvenser. På bakgrunn av dette ble syntetisk data fremhevet av informantene som en løsning for å unngå negative utfall knyttet til testing.

> "*Ved bruk av syntetisk data vil det redusere risiko for at testaktivitet skjer med reelle personnummer og at konsekvenser av dette treffer journalen og historikken til den faktiske personen"* (Informant 8).

Informanten fortalte om en spesifikk hendelse der det ble foretatt flere testaktiviteter uten faglig kvalifisering. Et fastlegekontor tok i bruk en testaktør med et reelt fødselsnummer og gjorde en rekke testaktiviteter. Som følge av dette arvet den faktiske personen historikken til testaktøren. Denne hendelsen forfølger den aktuelle personens helsejournal fortsatt. Dette eksempelet kan ses i lys av behovet for syntetisk data. *"Det er et godt eksempel på hvorfor vi trenger syntetiske testdata også med unike identifikatorer sånn som fødselsnummer"* (Informant 8).

Som vi ser har flere av informantene trukket frem at personopplysningsloven legger føringer for databehandling, deling av data, testing og lagring. Syntetisk data kan anses som en mulig løsning på de nevnte utfordringene. Analysen viser videre at bruk av syntetisk data muliggjør å utnytte og optimalisere dataene ytterligere, slik at ny verdi og innsikt kan oppnås.

4.1.2 Forskning og innovasjon

Vi har valgt å sammenslå kategoriene forskning og innovasjon, på bakgrunn av at samtlige informanter nevner disse to i sammenheng med hverandre.

Det fremkom at det er flere barrierer knyttet til et forskningsprosjekt. Informantene påpekte at etiske implikasjoner kan være begrensninger for forskning. I den forbindelse ble

godkjenninger, samtykke fra pasient og foretak som eier dataene, trukket frem som eksempler. For å gjennomføre et forskningsprosjekt kreves det innhenting av data, og informantene nevnte i den sammenheng at syntetisk data kan ha en verdi for forskere i denne prosessen. For å samle inn helsedata er det nødvendig å gjennomgå en søknadsprosess, hvor det er krav om samtykke for å tilgang på data. Syntetisk data kan i slike tilfeller være en løsning, siden det ikke inneholder personopplysninger. Funn belyste hvordan syntetisk data kan gi mer tilgjengelige data og på den måten redusere tidsløpet til et forskningsprosjekt. *"Forskning tar lang tid og er komplisert slik som det er nå. Hvis vi får produsert syntetiske data, vil det åpne opp for mer tilgjengelighet og forskning"* (Informant 2).

Dette ble videre støttet med *"plattformer med syntetisk data skal det gå fra 17 måneder til 17 sekunder. (...). Her kan du få en god start mens du venter på reelle data"* (Informant 3). I denne konteksten ble HAP trukket frem som et eksempel på en plattform som skal tilby syntetiske data. Det ble videre beskrevet mer detaljert hvordan syntetiske data kan benyttes i startfasen av et forskningsprosjekt og hvilke muligheter dette kan gi.

> *"Man kan bruke syntetiske data i påvente av reelle data. Det betyr at kartleggingsarbeidet kan starte tidligere. Hvis en kan tilgjengeliggjøre syntetiske data til en som har behov, kan man starte tidligere og få innsikt i detaljer over hvilke data som finnes, så vil også mest sannsynlig søknaden bli mer presis. Du får tilgang på en representasjon på ekte data. Søknaden blir derfor mer presis og kan gjøre saksbehandlingen raskere"* (Informant 3).

Funnet identifiserte en mulighet knyttet til hvordan syntetisk data kan være nyttig i startfasen av et forskningsprosjekt, siden det kan føre til en mer presis søknad. Vi ser i denne sammenheng at syntetisk data kan bli brukt som en støttefunksjon for datainnsamling, ved at søknaden blir mer spesifikk, som kan føre til redusert saksbehandlingstid. Videre kom det frem at syntetisk data også kan bistå i hypotesetesting. *"Det å få tilgjengeliggjort syntetiske data basert på reelle data, gjør at man kan se på hva som ligger i de ulike dataene. Dette gjør at man kan forme, danne grunnlag og teste hypoteser"* (Informant 7). Funnet viser at syntetiske data gjør det mulig å danne et datagrunnlag og teste ulike hypoteser tidlig i forskningsprosessen. Ved å kartlegge forskningsområdet, kan det hjelpe forskere med å få en større forståelse og se sammenhenger.

> *"Når man er i en eksplorativ fase innen forskning, da leter man etter sammenhenger, her kan man bruke syntetiske data. Man kan ikke lagre data sånn vilkårlig i forhold til GDPR. Så sånn sett kan det være hensiktsmessig med syntetisk data"* (Informant 9).

Videre fremkommer det av funn at syntetisk data kan gi mulighet til å lagre data uavhengig av GDPR. Som følge av dette kan det gi forskere mulighetsrom til å drive med innsiktsarbeid og lagre syntetiske data. Sett i sammenheng med dette fremkom det at ved bruk av syntetisk data kan det tilrettelegges for raskere, bedre og mer helseforskning.

Funn avdekket at bruk av syntetisk data gjør det mulig å etterprøve forskningsdata, på grunn av tilgjengelige datasett uten personopplysninger. *"Det er definitivt mye lettere å validere forskningsdata, fordi syntetisk data er 100% anonymisert"* (Informant 2). Dette ses i lys av dagens retningslinjer for forskning, som gjør det vanskelig å validere et datasett. Det inngår i GDPR som regulerer deling av data med personopplysninger. En annen informant poengtere hvordan syntetiske data kan bidra til å skape validitet i forskningen.

> *"Den eneste måten å sammenligne er å bruke samme datasett. Hvis noen andre prøver å gjenskape den forskningen du har gjort, så må de bruke det samme datasettet for å finne ut av det samme. Hvis man bruker ekte data, så har man ikke lov til å oppgi det samme datasettet"* (Informant 1).

Her beskrives det hvordan en forsker ikke kan gjenskape et tidligere forskningsprosjekt, siden en ikke får tilgang til samme datasett i henhold til GDPR. Som følge av dette vil det være nærmest umulig å etterprøve forskningsfunn. Bruk av syntetisk data kan dermed forenkle kvalitetssikring i forskning.

Flere av informantene ga uttrykk for at syntetisk data kan være innovasjonsfremmende i en forskningssammenheng. Det forklares ved at syntetisk data kan effektivisere søknadsprosesser og fjerne regulatoriske hindringer. En positiv konsekvens av dette er at en kan behandle og oppbevare data friere. Samtidig viser funn at syntetisk data kan ha flereanvendelsesområder, som for eksempel i innovasjonsprosjekter. Her kan syntetiske data være ressurseffektivt, både i form av tid og kostnader. Dersom man klarer å utvikle store og realistiske datasett, ble det hevdet at det ville gitt en oppblomstring av innovasjon og kreativitet i form av analyseverktøy, teknikker og miljøer.

Det fremkom av samtlige informanter at syntetiske data kan gi verdi i et innovasjonsprosjekt, på bakgrunn av at en i større grad kan utnytte eksisterende data. I en innovasjonsprosess ble også tidsbesparelse ved bruk av syntetisk data nevnt som en mulig positiv konsekvens. Bruk av syntetisk data kan fremme innovasjon i helsesektoren, fordi det optimaliserer datagrunnlaget. På bakgrunn av dette vil de eksisterende dataene få en større nytteverdi. *"Det er en stor mengde data hos oss som ikke brukes til forskning og innovasjon. Syntetisk data kan gjøre dette tilgjengelig. (..) Med syntetisk data så åpner man opp et stort område av innovasjon, fordi data blir tilgjengelig uten like mye hindringer"* (Informant 2).

I en innovasjonsprosess er det flere faser, hvor man innledningsvis ved en produktidé har behov for data for å teste et produkt. I en slik prosess bruker bedrifter gjerne reelle testdata. Det krever samtykke, hvor personopplysningsloven preger denne prosessen med flere regulatoriske krav og vilkår. Ved å benytte syntetisk data som en støttefunksjon kan denne fasen forenkles, ved å se på hvorvidt produktidèen er mulig å gjennomføre i praksis. Her trekkes også tidsbesparelse i en tidlig innovasjonsfase inn som en mulig effekt av syntetisk data. En ringvirkning av dette kan være reduserte kostnader knyttet til testing av produktet.

> *"Syntetisk data kan brukes som en støttefunksjon i en innovasjonsprosess. Ved produktidé er man ute etter å se om produktet eller tjenesten lar seg gjøre. For å sjekke ut dette trenger man samtykke. Hvis man bruker syntetiske data kan man gjøre dette uten å måtte ta samtykke (..). Man kan teste i et GDPR fritt miljø, som kan spare mye tid tidlig i innovasjonsløpet"* (Informant 9).

Som tidligere funn har avdekket, kan syntetisk data fungere som et innovasjonsfremmende verktøy i flere settinger. Helsesektoren er en offentlig sektor og som følge av dette preget av byråkrati og regelverk. Det er viktig for å sikre rettferdig og sikker pasientbehandling. Det legger imidlertid noen føringer for helsesektoren når det gjelder innovasjon. Som følge av dette har helsesektoren en inkrementell tilnærming til innovasjon.

> *"Innovasjon i helsesektoren handler om å bruke eksisterende løsninger, og bruke de på en ny og smartere måte, og ikke det å være revolusjonerende og komme opp med nye løsninger. Det er ikke der innovasjon oppstår i helsesektoren. Innovasjon oppstår i det daglige arbeidet i pasientbehandlingen ved at man tar i bruk eksisterende*

> *løsninger på en smartere måte og kanskje kobler sammen nye og eksisterende løsninger på en ny måte"* (Informant 8).

Her fremkommer det hvordan innovasjon i helsesektoren eksempelvis handler om å implementere nye elementer inn i en eksisterende løsning. En klinikksjef understøttet dette ved å komme med et konkret eksempel på hvordan syntetisk data kan styrke den nåværende helsetjenesten *"Syntetiske data kan brukes til beslutningsstøtte, ved å foreslå tilleggsdiagnoser eller spørsmål som bør stilles til pasient"* (Informant 6). Her beskrives det hvordan man kan benytte et utviklet register bestående av syntetisk data for å kunne predikere diagnoser. Syntetisk data kan derfor fungere som en potensiell støttefunksjon for fagpersoner i kliniske miljøer. Videre funn viste hvordan syntetiske data kan anvendes som en beslutningsstøtte for leger og fagpersoner, dersom det foreligger nok datavolum.

> *"Om man har en stor mengde data så kan man identifisere et mønster og en hel rekke elementer, som kan hjelpe leger med å diagnostisere eksempelvis kreft, fordi data peker i den retningen. (...). Algoritmer kan hjelpe å identifisere disse tingene, ved at legen kan registrere noe om en pasient. Da kan det eventuelt komme opp et rødt flagg i systemet som sier noe om sannsynligheten for at det er en eller annen sykdom. Her kan syntetiske data bli benyttet."* (Informant 4).

Trente maskinlæringsmodeller kan avdekke mønstre, som et menneske nødvendigvis ikke ville klart. Avdekking av slike mønstre kan bistå kliniske fagpersoner til å sette en diagnose og på den måten kan man få stilt en raskere og mer helhetlig diagnose. Det er mulig som følge av at maskinlæringsmodeller kan predikere mulige utfall, basert på store datavolum. Videre ble det forklarte hvordan syntetisk data kan anvendes som beslutningsstøtte for diagnostisering og hvordan det på samme måte kan brukes inn mot medisinering.

> *"Dersom man skal medisinere en diagnose hvor det er et spekter av medisiner, så kan syntetiske data bidra til å forutsi hvilken medisin man burde prøve først. Klarer man å få støtte på sånne ting, slik at pasienten slipper å gå igjennom testfase av 3-4 ulike medisiner, så er det veldig interessant. Hvis man får til det, så er det bare fantasien som begrenser mulighetsrommet for dette"* (Informant 6).

Her presenteres syntetisk data som et potensielt nyttig verktøy når det gjelder å unngå feilmedisinering. Det bygger på at man sparer tid på å korte ned veien til riktig medisin, og som følge av det reduseres den medisinsk utprøvning. En positiv konsekvens av dette er at pasienten slipper unødvendig belastning. Relatert til det identifiseres et stort potensial som ligger til grunn dersom man klarer å utvikle troverdige syntetiske datasett.

Et funn viste at delte berikede syntetiske datasett kan bidra til økt samarbeid på tvers av sektorer. Åpne og delte berikede syntetiske datasett kan derfor virke innovasjonsfremmende. Berikede data er en kombinasjon av syntetiske og reelle data, som har til hensikt å fullstendiggjøre et datasett. Her kan berikede syntetiske data benyttes i en hel verdikjede. Både ut mot grensesnitt, og mot alle tilhørende tilstøtende systemer, som tidligere har vært vanskelig. På den måten kan ulike problemstillinger testes i en hel verdikjede, fordi systemene kommuniserer sammen.

> *"Et annet bruksområde på syntetisk data, som kan fremme innovasjon er dersom vi lager berikede syntetiske data, så kan vi også tilrettelegge syntetiske data for bruk i en hel verdikjede. Det har vært en problemstilling frem til nå, fordi en verdikjede består av integrasjon mellom mange systemer. (...). Det kan vi enkelt løse med syntetiske data. Det mener jeg er innovasjonsfremmende, for da har man plutselig mulighet til å teste forskjellige problemstillinger i en hel verdikjede, fremfor å bare teste i et lokalt system"* (Informant 8).

Basert på analysen ser vi at syntetisk data kan ha en verdi i forskning. Det kan gi tidsbesparelse i et forskningsprosjekt, muliggjøre etterprøving og videreutvikling av forskningsresultater. Videre viser analysen hvordan syntetisk data kan benyttes som støttefunksjon i helsesektoren og på den måten fremme innovasjon. Andre funn peker på syntetisk data som innovasjonsfremmende gjennom redusering av siloer, avdekking av mønstre, økt produktivitet og tilgang på ressurser.

4.1.3 Kompetanse

Kompetanseheving ble identifisert som en mulighet ved bruk av syntetisk data. Her påpekte en informant hvordan anvendelse av syntetisk data kan være relevant for utdanning og opplæring. Bruk av syntetisk data vil i den sammenheng bidra med tilrettelegging for økt

kompetanse på en harmløs måte, som følgelig ikke vil påvirke den daglige driften. Dette ble understøttet av en informant til.

> *"Vi bruker de syntetiske testdataene ikke bare til testformål, men også opplæringsformål. Vet også at flere offentlige instanser snakker mye om at de ikke ønsker å snakke om syntetiske testdata lenger, men syntetiske data. De tenker også at de som driver med opplæring og innføring ikke bør jobbe med produksjonsdata, for det trenger de ikke"* (Informant 10).

I det overnevnte sitatet bruker informanten eksempler fra andre i bransjen som ser verdien i å benytte syntetiske data til opplæring og innføring. Det startet med at de benyttet syntetiske testdata, men etter hvert som bruken har blitt mer omfattende og erfaringen har økt, har bruksområdet utvidet seg.

Videre belyste enkelte informanter hvordan bruk av syntetiske data kan forenkle samhandling mellom ulike systemer. En informant trakk frem et konkret prosjekt, for å gi en beskrivelse av hvordan syntetisk data kan være en løsning for å unngå siloer. Her måtte dataene lages i alle systemene de brukte for å kunne sikre at systemene hadde samme informasjon og dermed kunne kommunisere sammen. *"Vi startet med å måtte lage dataene i alle systemene slik at man kan få snakket sammen. Det er noe av det problemet vi ønsker å løse med syntetisk data"* (Informant 11).

Funnet viser hvordan økt kompetanse på syntetisk data kan skape verdier som kan være med på å sikre kommunikasjon mellom systemer. Videre kan det bidra til å redusere manuell inntasting av data i systemer. Slik inntasting kan føre til menneskelig feil og bidra til at eksisterende siloer bevares. Siden det i stor grad er personvern som begrenser deling av data mellom systemer, kan syntetisk data være en bro for å unngå problematikken og tilrettelegge for deling.

Videre ble det poengtert at kunstig intelligens og syntetisk data er et fokusområde i Europa og at Norge er med på satsningen for å bygge kompetanse knytte til dette.

> *"Det største prosjektet vi er med på nå er en EU-satsning, Norge har satt inn 1,3 milliarder (…) vi ser på muligheter for kunstig intelligens og syntetisk data. Det går*

på å tilgjengeliggjøre de rette volumene og de rette type syntetiske data for trening av kunstig intelligens modeller innenfor EU. Alle aktører innenfor EU kan komme til oss og gjøre bruk av det vi har tilgjengelig av syntetiske data og en plattform for å trene sin kunstige intelligens modeller" (Informant 8).

Sitatet refererer til en satsning på tvers av EU, som ser på om bruk av KI og syntetisk data kan være med å tilgjengeliggjøre tilstrekkelig mengde og riktig type data for å kunne trene KI-modeller. Her er Norge et av landene som har satset og har en forventning om at slike samarbeid vil gi tilgang til riktig og nok data, samtidig øke kompetanse innenfor feltet.

4.1.4 Data

Funn viste at bruk av KI kan bidra til å redusere siloer, og informantene trakk frem at ved å utvikle KI-modeller på syntetiske data muliggjør det deling av data. Det forklares ved at det ikke foreligger identifiserbare opplysninger i datasettene.

"Det som er viktig med kunstig intelligens er at det ikke er noen siloer tilstede, så jo større bredde det er, jo mer kan man bruke det. Man øker kvaliteten ved å bryte ned siloer, fordi man kan finne mønstre når siloer ikke er tilstede" (Informant 4).

I uttalelsen blir kunstig intelligens sett i sammenheng med syntetiske data, som kan gi flere fordeler. Ved å bryte ned siloer kan det følgelig føre til at en klarer å avdekke ulike mønstre på tvers av enheter, avdelinger og bransjer. Bakgrunnen for dette er at èn faktor i et bestemt ledd, kan påvirke andre faktorer i andre ledd. Klarer man å identifisere slike hendelser, gjør det mulig å skape mer bredde og mangfold i dataene. På bakgrunn av dette ble det fremhevet av flere informanter at syntetisk data kan ha en positiv virkning gjennom å redusere siloer.

"Ved bruk av syntetiske data vil man i større grad sikre og tilrettelegge for videre forskning, da det ikke er like mange siloer på grunn av personvern. Men med syntetiske data så kan man si "dette har jeg jobbet med, bruk det videre". Det vil absolutt føre til mer mangfold. Mer realistisk industristandard, spesielt innen medisin. Mange siloer man kan bruke og knytte sammen" (Informant 1).

Sitatet beskriver hvordan syntetisk data kan redusere siloer, og på den måten kan det enklere tilrettelegges for videre forskning. Som følge av at barrieren knyttet til personvern fjernes, kan forskningsresultater i større grad deles på tvers. På den måten kan en øke tilgjengeligheten av informasjon. En annen informant understøttet dette og mente at en av fordelene ved å benytte syntetisk data, fremfor reelle data, er tilgjengeligheten.

> *"Syntetiske data lar oss tilgjengeliggjøre de dataene man har behov for, der man trenger dem, når man trenger dem. Vi klarer å imøtekomme det behovet ved hjelp av syntetiske data, fremfor bruk av produksjonsdata. Det går på både hastighet og volum"* (Informant 8).

Her ble det forklart hvordan syntetisk data er mer tilstrekkelig fremfor bruk av produksjonsdata sett i sammenheng med hastighet og volum, som er vesentlig i KI-modeller. Ved å få tilstrekkelig mangfold i syntetiske data vil det i større grad være mulig å bruke syntetisk data som en erstatter for, eller beriker til reelle data.

En av informantene arbeider med å produsere syntetiske helsedata og fremhevet flere muligheter tilknyttet dette. Vedkommende gjorde rede for hvordan syntetiske data gjør det mulig å redusere bias, som kan bidra til å styrke datasettets kvalitet. Bakgrunnen for dette er at i syntetiske datasett kan en gå inn i dataene og benytte kunnskap til å produsere data som gjenspeiler realiteten. Det medfører at en enklere kan arbeide aktivt for å redusere bias i helsedata. I reelle data er det flere tilfeller hvor dataene ikke innehar gitte ekstremverdier, til tross for at vi vet at det finnes. Slike ekstremverdier kan være en sjelden sykdom eller en uvanlig hendelse. I en slik situasjon kan en skape data som dekker egenskapene i en reell befolkning, og på den måten bidra til å øke kvaliteten på syntetiske datasett.

> *"Helsedata er inkomplett og med masse bias. I syntetiske data kan man benytte kunnskap til å generere nye datasett eller fjerne bias. Det gjør at man kan jobbe med blant annet bias litt mer aktivt. Noen ganger har vi ikke konkrete scenarioer i et reelt datasett, men det kan man generere i et datasett med syntetiske data. (...). I syntetiske datasett kan man produsere data og øke kvaliteten, med spesielle scenarier. Det er ikke lov med reelle data, men i en syntetisk verden er det lovlig"* (Informant 2).

Et annet aspekt som ble avdekket under datainnsamlingen, var hvorvidt en kan klare å utvikle tilstrekkelige maskinlæringsmodeller. Her menes maskinlæringsmodeller som er kompetente nok til å produsere kvalitetssikker syntetisk data, på en slik måte at det syntetiske folkeregisteret kan klare å være dynamisk. Dersom det er mulig har ikke datasettet behov for å bli kontinuerlig kontrollert av mennesker. På den måten vil det syntetiske folkeregisteret i stor grad videreutvikle seg selv, likt som en reell befolkning hvor innbyggere blir født, får sykdommer, gifter seg og dør. *"Det å ha gode nok maskinlæringsmodeller og statistiske modeller er viktig, slik at datasettet kan utvikle seg i takt med det norske samfunnet."* (Informant 11).

En av informantene beskrev mulighetsrommet til syntetisk data som en tjeneste og hvilken verdi dette kan gi, på tvers av bransjer og sektorer.

> *"Målet er å etablere syntetisk data as a service. Jeg ser for meg at vi kan ha en hylle med x-antall algoritmer som er basert på veldefinerte brukerbehov (...) vi har satt et scope (...) og definert brukerhistoriene (...) og etablert algoritmene (...) som du kan ta i bruk til å produsere de syntetiske dataene som du har behov for. Da har vi plutselig etablert syntetiske data as a service og det er en hyllevare, disse algoritmene. (...) Så slipper man hele utviklingsprosessen, da kan vi tilgjengeliggjøre data (...) når du trenger dem og der du trenger dem"* (Informant 8).

Her ble det beskrevet at man i nærliggende fremtid kan tilby syntetisk data som en utviklet programvare, som en tjeneste til konsumenter. Det belager seg på at man klarer å utvikle gode algoritmer, som produserer syntetisk data i henhold til det behovet brukerne ønsker å dekke. Da kan man se for seg at syntetisk data vil fungere som en hyllevare, hvor en kan plukke ut de ulike momentene man trenger der og da. På den måten kan man nå målet om å kunne tilby riktig data, til riktig tid og sted. Videre fremkom det at en slik tankegang med syntetisk data "as a service" vil gjøre det enkelt og tilgjengelig for innovasjon og næringsfremmende virksomheter. Samtidig påpekte vedkommende at veien frem til produksjon av syntetiske data ved bruk av maskinlæringsmodeller kan være en lang og vanskelig prosess, fordi det er kostnadskrevende.

4.2 Utfordringer

Ved å avdekke hvilke muligheter syntetisk data kan gi, fremkom det fra datagrunnlaget flere utfordringer som helsesektoren bør være bevisst på tilknyttet bruk. I denne delen av analysen vil vi presentere utfordringene i de ulike kategoriene.

4.2.1 Lovverk

Gjennom datainnsamlingen ble det avdekket utfordringer som fremkommer grunnet regulatoriske hindringer ved produksjon av syntetiske data. Et funn beskrev hvordan Personopplysningsloven legger føringer på tilgang og bruk av reelle data. Her ble arbeidet med Helseanalyseplattformen trukket frem som et eksempel. Gjennom bruk av syntetisk data skal HAP møte behovet for mer helsedata. For å produsere syntetisk data er det imidlertid nødvendig å innhente og bruke reelle data. En forutsetning for å få tilgang til reelle data, er at det må oppgis et formål. Dette kan anses som problematisk, siden syntetiske datasett i Helseanalyseplattformen skal være tilgjengelig for helseforskere og kan derfor bli benyttet til ulike formål og forskningsprosjekter. På bakgrunn av dette kan det være vanskeligere å søke om tilgang på reelle data for syntetisering, da det ikke foreligger et spesifikt formål.

> *"Det å tilgjengeliggjøre syntetiske data er ikke noe problem fra et rettslig perspektiv, men for å få utlevert dataene for å syntetisere de, trenger man et formål. Det er der det ligger. Både det å få tilgang på de dataene som man ønsker å syntetisere og det å utføre det"* (Informant 3).

Et annet funn viser paradoksalt at fraværet av juridiske retningslinjer på syntetiske data kan være en utfordring. Funnet belyste kompleksiteten ved produksjon av berikede syntetiske data og hvordan manglende reguleringer setter begrensninger for dette. For å produsere berikede syntetiske data forutsetter det at en kan knytte ulike datakilder sammen. Imidlertid er disse datakildene plassert i forskjellige registre, hvor disse har ulike registereiere.

> *"Så fort vi skal skape disse berikede syntetiske dataene og koble oss til forskjellige registre hvor det sitter forskjellige registereiere så er det masse juridiske hindringer,*

> *det er en lang søknadsprosess, behandlingsprosess og godkjenningsprosess. Det er GDPR, data eier, ansvar dataeier har overfor sine registrerer, det er altså registerloven. Om vi ønsker å opprette og lagre det syntetiske uttrekket en forsker lager, så er det plutselig opprettelse av et nytt register, og da er det registerloven som kommer inn og setter krav og forventninger og de har ikke juridisk prøvd ut det godt nok enda om det også gjelder selv om dataen er syntetiske"* (Informant 8).

Siden produksjon av syntetisk data er i startfasen er det usikkert hvordan lover, slik som registerloven, vil være gjeldende for syntetisk data. Som følge av dette kan det anses som at foreligger et fravær av rettslige avklaringer knyttet til syntetiske data. Videre ble det trukket frem at en konsekvens av GDPR er langvarige søknads- og behandlingsprosesser når de forsøker å produsere berikede syntetiske data.

4.2.2 Forskning

For at syntetisk data skal bli tatt i bruk i forskning, viste funn at datasettene må være plausible, som betyr at de må være sannsynlige og troverdige. En informant stilte spørsmål til hvorvidt dette er mulig å oppnå med syntetiske data *"Vanskeligere å bruke syntetiske data til forskning, fordi de syntetiske dataene må være plausible. De må representere befolkningen, men likevel ikke"* (Informant 5). Her ble det hevdet at det vil være vanskeligere å bruke syntetisk data i forskning. Årsaken til dette er at syntetiske data skal representere en befolkning uten å gå på bekostning av personidentifiserende opplysninger. På bakgrunn av dette kan det medføre at de syntetiske dataene ikke stemmer overens med de reelle dataene. Videre funn viste at syntetisk data også må være reliabel for å kunne tilføre forskere en verdi ved bruk.

Noen av informantene trakk frem reguleringer som må være til stede for å sikre etisk forskning og at dette kan være en utfordring ved bruk av syntetisk data. En informant fremhevet utfordringen knyttet til åpne og tilgjengelige syntetiske data.

> *"De etiske reglene på forskningsprosjekt, godkjenning og samtykke fra pasient og foretak som eier dataene er en begrensing fordi man er avhengig av det. Men det er*

> *veldig viktig at vi har dem, for man kan ikke tillate all forskning, selv om det er en bremsing. Hvis vi har åpne syntetiske data er det en fare etisk sett"* (Informant 1).

Her ble det belyst at dersom syntetisk data fjerner behovet for REK-godkjenning, kan det være risikabelt sett i lys av etiske betraktninger. Bakgrunnen for dette er at dersom det ikke finnes en kontrollinstans ved bruk av syntetiske data til forskning, kan etisk grenser overskrides. En kontrollinstans skal forhindre at forskere gjennomfører forskningsprosjekter som kan skade enkeltindivider eller sårbare grupper i samfunnet. Videre ble det poengtert at søknadsprosesser og samtykke kan være en barriere for forskning, men at det er viktig for å sikre at etiske hensyn blir ivaretatt.

4.2.3 Kompetanse

Et gjennomgående funn var at kompetanse på syntetisk data ikke er tilstede i helsesektoren. Noen av informantene påpekte derimot at kompetanse på syntetisk data vil bli bedre over tid, og forklarte at det er behov for erfaringer på feltet for å kunne bygge kompetanse. Funn viste videre til den avgjørende rollen eksperter har for kvalitet knyttet til produksjon av syntetiske helsedata. Dette var spesielt relatert til hvordan domeneeksperter i større grad kan ta gode valg og på den måten unngå negative konsekvenser.

> *"Du trenger domeneeksperter som sitter tett på den ekte løsningen, for å være med å gjøre gode valg rundt data som skal støtte opp om den samme løsningen. For helsedata f.eks. da bør det være helseforetakene, i samarbeid med Norsk Helsenett eventuelt" (Informant 10).*

Organisering og tilrettelegging for implementering av syntetisk data bør derfor gjøres med hensyn til domeneeksperter, på bakgrunn av deres unike kompetanse. Sett i lys av dette viste et annet gjennomgående funn at kompetanse er avgjørende for å produsere syntetiske data med kvalitet, fordi det er komplekst. *"Kvaliteten av dataene er helt avhengig av hvem som produserer den"* (Informant 2). I syntetiske datasett trakk informantene frem at volum og variasjon er viktig for å sikre kvalitet. Her foreligger det et behov for kompetanse, for å kunne sikre at egenskapene til dataene er på et ønsket nivå. I den forbindelse ble det avdekket at dette er en stor utfordring.

> *"Noe av den største utfordringen med syntetiske data er jo ikke det å få samlet de inn, eller laget de, men dekket den samme variasjonen og spredningen på egenskaper i den syntetiske data som vi har i den reelle norske befolkningen" (Informant 11).*

Videre ble det belyst av enkelte informanter at kompetanse vil være nødvendig i flere ledd ved bruk av syntetisk data. Her ble det forklart at dersom det skal benyttes i helsesektoren, vil det kreve endringer i både tekniske og organisatoriske prosesser. En av informantene hevdet at den største utfordringen tilknyttet bruk av syntetisk data vil være de organisatoriske endringene. Det vil føre til et behov for opplæring av ansatte som skal bruke syntetiske data.

> *"Tekniske utfordringer kan man løse, men største utfordringen som vil kreve mer er både prosesser, organisatoriske, og menneskene som bruker dataene. Det krever opplæring til å benytte den type data" (Informant 2).*

I uttalelsen ble det belyst hvordan organisatoriske endringer som påvirker prosessene, menneskene og kompetansen, må til for å sikre at brukerne kan benytte syntetiske data til det gitte formålet. En annen informant støttet funnet og stadfestet at bruken av syntetisk data må ha en konkret funksjon som skaper verdi. Det ble forklart ved at det ikke er et poeng å innføre noe som ikke gjør en prosess smidigere eller som kan berike eksisterende data.

4.2.4 Tillit

Gjennom intervjuene ble det ikke identifisert eller belyst noen muligheter sett i lys av tillit ved bruk av syntetisk data. I denne delen av kapittelet vil vi imidlertid presentere de ulike utfordringene som informantene trakk frem.
Blant flere av informantene fremkom det at foreligger en viss grad av skepsis til syntetisk data. Når det gjelder tillit til teknologi og det kunstige er det mange aspekter som må tas i betraktning. Her ble det gjort et forsøk på å sammenligne bruk av syntetisk data med en annen teknologi som er mer kjent.

> *"Sammenlignbart med førerløse biler, teknologien er der, men du stoler ikke på at bilen tar de etiske beslutningene. Sånn er det med de problemstillingene vi kommer inn på her når det gjelder syntetisk data" (Informant 6).*

På spørsmål om informanten selv ville benyttet syntetisk data i sitt kliniske arbeid uttrykket vedkommende en mangel på tillit til teknologien. Informanten mente at teknologien, her KI, ikke kan ta like veloverveide og etiske beslutninger som et menneske kunne tatt.
Det ble avdekket at flere av informantene uttrykte en lignende form for skepsis ved bruk av syntetisk data. Det ble identifisert at flere mente at tillit til syntetisk data, spesielt blant forskere, vil være vanskelig og at det vil kreve arbeid for å overbevise dem. Et gjennomgående funn viser at det er liten grad av tillit knyttet til syntetisk data. Det ble beskrevet som en av de største barrierene for hvorvidt syntetisk data vil bli benyttet i fremtiden. Her ble vanen med å bruke reelle data trukket frem som en mulig forklaring.

> *"Den største utfordringen er tillit til dataene i klinisk miljø. I det kliniske miljøet, vil de bruke syntetisk data? Vil de bruke dette? Dette er helt ukjent for dem. De er vant til å bruke reelle data" (Informant 2).*

Dette funnet ble støttet av en annen informant som mente at det vil kreve overbevisning for å få forskere til å benytte seg av syntetisk data. *"Mange forskere som er opptatt av statistikk vil være skeptisk. Så det kan være tøft å tilby forskere syntetiske data, du må overbevise de" (Informant 5).*

Videre funn viste at tillit til syntetisk data må skje over tid, både blant forskere og konsumenter. Det ble identifisert at store deler av mistilliten kan relateres til kvaliteten på syntetiske data. Kvaliteten ses her i forskningssammenheng og blir knyttet til validitet, reliabilitet og plausibilitet. Flere av informantene sammenlignet syntetiske data med reelle data, og hevdet at det vil kreve tillitsskapende arbeid dersom syntetiske data skal benyttes på lik linje som reelle data.

> *"Det handler om å skape tillit ut til konsumentene til syntetiske data. En forsker må jo ha tillit til de syntetiske dataene, å forstå og bli overbevist om at dette er gode nok data sammenlignet med produksjonsdata" (Informant 8).*

Et annet funn relatert til tillit var konkurransen blant forskere og de høye forventningene som de kan stille til seg selv. *"Forskere er tøffe, smarte og har knivskarp konkurranse. Den type tilnærming. Dette må benchmarkes litt" (Informant 5).* Flere av informantene belyste dette, og påpekte at bruk av en ny tilnærming til datainnsamling i forskning må bearbeides. En av informantene forklarte at for å kunne overbevise forskere til å benytte syntetisk data må det skapes et referansepunkt. Det kan gjøre det enklere for forskere å sammenligne syntetiske data med reelle data. Tillit til syntetisk data, dens reliabilitet og mulighet for anvendelse er vesentlige utfordringer som ble belyst i flere intervjuer.

> *"Hvor reliable er disse dataene? Det er sånn at hvis feilmarginene blir for stor så slutter man å bruke disse, og da er det ikke verdt noen ting. Det bør være en beslutningsstøtte som dette dataverket gir oss" (Informant 6).*

For at syntetisk data skal benyttes forventes det at de er reliable slik at brukerne har tillit til å støtte seg på dataene ved beslutningstaking. Dersom det viser seg å være store feilmarginer og brukere slutter å benytte seg av syntetisk data på grunn av det, så vil investeringen i syntetiske data være verdiløs. Dette funnet ble understøttet av flere og videre ble krav om fullstendig anonymisering presentert.

> *"Det som er viktig når vi snakker om syntetiske data er den anonymiseringen, det skal ikke være mulig å finne tilbake til en gitt identitet i datagrunnlaget, det er viktig. Hvis det skal det brukes til forskning er det viktig at det er statistisk riktig. Det må være på plass hvis vi skal kunne si at vi kan bruke syntetisk data" (Informant 3).*

Her ble viktigheten med anonymisering av identitet og korrekt datagrunnlag trukket frem. Dette ble fremlagt som grunnleggende for å skape tillit. Videre viser funnet at det er vesentlig at syntetiske datasett er statistisk riktig, som vi kan se i lys av andre funn.

4.2.5 Data

Samtlige informanter belyste utfordringer knyttet til det å kunne sikre kvalitet i produksjon av syntetiske data. Det fremkom gjennom flere av våre informanter at datakvalitet er en stor barriere for bruk av syntetisk data i dag. Informantene beskrev flere faktorer som må være

tilstede for å sikre datakvalitet ved en produksjon av syntetiske data. Funn viste likevel at det foreligger en usikker på hvordan en kan oppnå dette i praksis, og at det krever ressurser for å kunne produsere syntetiske data med tilstrekkelig kvalitet.

> *"Vi tror på at vi kan løse veldig mye med kun syntetiske data. Det store spørsmålet er hvordan skal vi få produsert de? Altså hvor tungt skal vi satse på dette og hvordan skal vi få det til i praksis for å sikre gode nok syntetiske data?"* (Informant 3).

I forbindelse med ressurser var det flere informanter som trakk frem at det er krevende og kostbart å produsere syntetisk data. Dette kan også ses i lys av et annet funn som viste til at utgiften tilknyttet produksjon av syntetisk kan være innovasjonshemmende, spesielt for mindre selkaper. Et gjennomgående funn viste at produksjon av et syntetisk datasett er en komplisert prosess og at hvem som utvikler dataene vil være en vesentlig faktor for å sikre kvalitet. Videre ble det reist flere spørsmål hvorvidt syntetiske data faktisk kan gjenspeile virkeligheten.

> *"Utfordringene ligger i spørsmålet om hvor godt gjenspeiler syntetisk data realiteten? Hvordan kan du skape data som du vet gjenspeiler ekte data og realiteten? (...) en feil som ligger i datasettet, legger man ikke merke til det, så produserer man feile data, da kan ideene som er bygget opp av dette også være feil. Det største spørsmålet er hvordan syntetiske data produseres. Det er en utfordring"* (Informant 1).

For at man skal lage troverdige og gode syntetiske datasett, må man kunne validere at datasettet gir en riktig gjenspeiling av den reelle verden. Det ble beskrevet hvordan en relativt liten feil i et syntetisk datasett, som for eksempel en gravid mann, kan få store følgefeil dersom det ikke blir oppdaget. Det blir da skapt livshistorier som ikke er overførbare til virkeligheten, og som følge av dette vil ikke de syntetiske datasettene lenger realistiske.

Gjennom andre funn fremkom det at volum, bredde og variasjon er viktige faktorer for å sikre kvalitet i et syntetisk datasett. Viktigheten med å aggregere opp nok bredde på dataene for å sikre mengder med informasjon er avgjørende. Dette ble støttet og utdypet i et annet intervju. *"Vi må produsere store nok, men også liten nok mengde data, slik at man har kontroll*" (Informant 10). Her ble en reell problemstilling vedkommende hadde hatt i sin

arbeidssituasjon nevnt. Et testdatasett som bestod av flere ID-profiler, ble utviklet og produsert ved bruk av syntetisk data. Formålet var at det skulle etterligne en reell norsk befolkning. Problemet som etter hvert ble lagt merke til, var at flere av ID-profilene levde liv som ikke var overførbare til det norske samfunnet. Eksempelvis ble det ikke oppdaget at flere ekteskap ikke var gjensidig fra begge parter.

Et annet funn påpekte en mulig fallgruve ved produksjon av syntetiske data, som gjaldt maskinlæringsmodell algoritmene. Man kan klare å utvikle gode maskinlæringsmodeller med spesifikke algoritmer for syntetiske data. Til tross for dette er det vanskelig å sikre at samme algoritmer ikke er benyttet i produksjon av reelle data. Det kan være problematisk dersom en kan avdekke likheter mellom ekte og syntetiske personer, fordi da vil ikke syntetisk data lenger være uten personidentifiserende informasjon. Dette ses i samsvar med funnet hvor et legekontor brukte en testaktør med et reelt fødselsnummer.

> *"Syntetiske data kan jo fortsatt oppleve en del problemstillinger, som man ser på og er bevisste på. Man kan jo fort, uten å være klar over det, sitte å lage de samme algoritmene som f. eks Skatteetaten lager for å produsere sine reelle fødselsnummer, og dermed reelle data. Selv om man har kunstig skapt de algoritmene for å lage syntetiske data. Så det kan være så sammenfallende at du ser likhetstrekk mellom syntetiske personer og reelle personer."* (Informant 8).

Dette funnet ble understøttet i et annet intervju fra en informant som hadde et statistisk ståsted *"Man bygger de syntetiske dataene på en sånn måtene at de skal gi mer eller mindre de samme statistiske resultatene som reell data. Paradoksalt nok kan de bli identifisert likevel, selv om de er syntetiske"* (Informant 5). Her ble det presentert et konkret eksempel, som tok for seg statistikk om kommunen Modalen, som er et lite lokalsamfunn. Dersom en bruker en analyse med syntetiske data, og det blir avdekket at det betales barnebidrag dersom man er prest, så er det sannsynlig at dette stemmer overens med virkeligheten. På den måten kan det være mulig å identifisere personlige opplysninger om et enkeltindivid, selv ved bruk av syntetisk data.

4.3 Hovedfunn

Nedenfor fremstilles hovedfunnene fra datainnsamlingen i studien presenteres inndelt i de aktuelle temaene identifisert fra dybdeintervjuene, som er videre kategorisert i muligheter og utfordringer.

	Muligheter	Utfordringer
Lovverk	▪ Kan være en løsning rundt lovverk for personvern ▪ Muliggjør lagring i skyløsninger utenfor Europa ▪ Kan redusere risiko for tap eller lekkasje av data	▪ Produksjon av berikede syntetiske data er ikke juridisk utprøvd enda ▪ Syntetiske data omfattes ikke av eksisterende lovverk ▪ Krever reelle data for produksjon som har regulative begrensninger
Forskning	▪ Kan brukes i påvente av reelle data i en forskningsprosess ▪ Tilrettelegger for effektivisering av forskningsprosessen ▪ Deling av forskningsdata på tvers kan gi bedre forskning og rom for etterprøving av resultater	▪ Kvaliteten avhenger av volum, variasjon og bredde ▪ Usikkert om forskere vil ta det i bruk grunnet kvaliteten og fordi det er lite etablert
Innovasjon	▪ Øker produktivitet og tilgang til ressurser ▪ Ses som innovasjon både som prosess og resultat ▪ Kan anvendes i prosessinnovasjon for at produkt eller tjeneste kan nå markedet raskere	▪ Kostbart og krevende for mindre selskap å produsere

Kompetanse	▪ Kan brukes til opplæring og kompetanseheving ▪ Fellesprosjekt i EU gi ny erfaring og kunnskap	▪ Kompetanse for kvalitet i dataproduksjon ▪ Balanse mellom datamengde og kontroll ▪ Mulig utfordringer med følgefeil i datasett ▪ Kompetanse for å tas i bruk likestilt med reelle data ▪ Kan kreve organisatoriske endringer i prosesser ▪ Kan være risiko for re-identifisering
Tillit		▪ Tar tid og krever kompetanse for å skape tillit ▪ Begrenset grad av tillit til KI som etisk beslutningstaker ▪ Generell skepsis til bruk særlig i forskningssammenheng

Tabell 2: Hovedfunn fra individuelle dybdeintervju

5.0 Diskusjon

I denne studien søker vi svar på hvilken betydning syntetisk data kan ha for helsesektoren og helseforskning. For å besvare problemstillingen kobler vi funn fra analysen med litteratur fra det teoretiske rammeverket for studien. Vi vil gi et dypere innblikk i hvilke muligheter som bruk av syntetiske data kan gi helsesektoren, samt hvilke utfordringer som bør bevisstgjøres tilknyttet bruk. Tema som diskuteres er henholdsvis hovedfunn innen lovverk, helseforskning, innovasjon, datakvalitet, kompetanse og tillit.

5.1 Muligheter og utfordringer relatert til lovverk

Hvordan persondata blir behandlet er strengt regulert gjennom personopplysningsloven. Data i helsesektoren inneholder i stor grad personopplysninger og på den måten må helseforetak opptre i samsvar med de lovfestede personvernprinsippene ved bruk av helsedata (Personopplysningsloven, 2018). Vi fant at syntetisk data er en mulig løsning knyttet til personvernutfordringer, som kan ses i lys av Ramos og Subramanyam (2021) og Chen et al. (2019). Teknologisk utvikling, økende databruk og reguleringer som GDPR, kan forstås som motiverende faktorer for bruk av syntetiske data. Det ses i sammenheng med utviklingen og bruk av syntetiske testdata i Norge, eksempelvis Tenor testdatasøk og Helsetanken populasjon.

Videre viser funn at Schrems II dommen legger føringer knyttet til lagring og overføring av personopplysninger til land utenfor EU/EØS (Digdir, u.å). Dommen gjør at det ikke lengre er tillatt å lagre persondata i skyløsninger utenfor Europa. Funn fra studien viser imidlertid at ved å benytte syntetiske data kan det muliggjøre lagring utenfor EU/EØS, ettersom dataene ikke inneholder personopplysninger.
Vårt funn om at syntetiske data kan redusere risiko for tap eller lekkasje av dataene, ses i tråd med personvernprinsippet om integritet og konfidensialitet. Funnet indikerer at det vil spesielt være knyttet til utilsiktet tap, ødeleggelse eller skade på dataene (Personopplysningsloven, 2018). En slik redusert risiko anses som fordelaktig for helsesektoren, siden det kan gi større mulighetsrom for å utnytte datasett friere. Funn fra denne studien viser at flere assosierer syntetiske data med testsituasjoner, og at dette er et etablert bruksområde for syntetisk data i Norge. Videre fremlegger analysen at syntetiske data er hensiktsmessig å benytte ved test- og opplæringsformål, siden det anses som tryggere å bruke fremfor reelle data.

I nyere tid arbeides det med å etablere offentlige plattformer hvor syntetiske data kan benyttes i helseforskning. I den forbindelse er det ulike utfordringer knyttet til produksjon og lagring av syntetiske helsedata. For å produsere syntetiske data er det en forutsetning å få tilgang på reelle data. Som følge av dette viser funn noen utfordringer sett i lys av personvernprinsippene, spesielt rundt krav om formålsbegrensing og dataminimering. Hensikten med en slik plattform er at den skal være tilgjengelig for helseforskere i ulike felt.

Det er derfor ikke gitt at de reelle dataene vil ha et enkelt formål, men at de syntetiserte dataene skal brukes i ulike kontekster og forskningsprosjekter.

Vårt funn om produksjon av berikede syntetiske data viser at manglende regulering er en utfordring. Ved en slik produksjon er det flere lover som inntrer, på bakgrunn av at berikede syntetiske data er en kombinasjon av syntetiske data og reelle data. Foreløpig er det ikke juridisk utredet, og retningslinjene for produksjon er derfor uklare. Som følge av dette er det uvisst hvordan en slik produksjonsprosess vil bli regulert i fremtiden. Denne utfordringen kan knyttes til Arora og Arora (2022), som problematiserer fraværet av lovgivning når det gjelder syntetiske data. Dette kan videre ses i lys av James et al. (2021), som hevder at en mulig utfordring ved bruk av syntetisk data kan være regulatorisk aksept, spesielt knyttet til deling av syntetiske data. Det er et interessant funn, siden det belyser hvordan syntetisk data er løsningen på de nåværende regulatoriske hindringene, slik som personvern. Imidlertid er det nærliggende å tro at det kan oppstå reguleringer knyttet til syntetiske data i fremtiden.

5.2 Muligheter og utfordringer relatert til forskning

Studien viser at syntetisk data kan ha funksjon inn mot forskning og utvikling, på bakgrunn av at syntetiske datasett kan tilgjengeliggjøre et større omfang av data enn det som reelle data kan gi. Tendenser fra analysen viser at helsesektoren har behov for mer, bedre, raskere tilgjengelig helsedata, og at det er oppnåelig ved bruk av syntetiske data, slik som Ramos og Subramanyam (2021) tematiserer. Det fremkommer av analysen at syntetiske data klarer å imøtekomme behovet for tilgjengelige data når man trenger dem, som ses i sammenheng med volum og hastighet. Dette underbygges av Bray og Parkin (2008) som hevder at datakvalitet er knyttet til aktualitet, som baserer seg på hastighet. Det innebærer tid fra en hendelse inntreffer til informasjonen er tilgjengelig for brukeren av data.

Våre funn viser at syntetiske data kan benyttes i en forskningsprosess, hvor tilgjengelige syntetiske datasett kan brukes i påvente av reelle data. De disponible dataene muliggjør for forsker å kartlegge, forme hypoteser og danne grunnlaget for forskningen tidligere i prosessen. Det kan resultere i en mer nøyaktig utarbeidet søknad som sendes til vurdering for godkjenning av forskningsprosjektet. Forskere kan på den måten utforme en mer presis

søknad for sitt forskningsprosjekt, hvor en positiv konsekvens kan være redusert behandlingstid. Det ses i lys av at det på nåværende tidspunkt er ressurs- og tidkrevende å få tilgang til helsedata (Larsen, 2017). Som følge av dette har det blitt rettet søkelys på hvordan det kan tilrettelegges for bedre tilgang for helsedata i Norge. I våre funn nevnes Helseanalyseplattformen, hvor formålet er å distribuere tilgjengelige syntetiske data for forskere. Våre funn om at en slik plattform vil gjøre enklere for forskere å oppsøke kvalitetssikre datasett og øke muligheten for gjenbruk av dataene, kan ses i lys av Åm et al. (2021).

Som følge av åpne og tilgjengelige syntetiske datasett innen medisinsk forskning, belyser et interessant funn fra studien en utfordring knyttet til etiske implikasjoner. Funnet viser at dersom forskere benytter seg av syntetiske data som datagrunnlag i et forskningsprosjekt, kan en REK-godkjenning omgås. REK-godkjenning ses i tråd med NOU 2005:1 hvor det stilles særlige krav til forsvarlighet og etisk bevissthet i helseforskning. Det ses som problematisk på bakgrunn av at en ikke kan tillate all type forskning etisk sett. Likevel kan vi forstå en slik godkjenning kan være en barriere for forskningsprosessen, sett i lys av lang behandlingstid. Som følge av dette kan funnet også tolkes som en mulighet, på bakgrunn av at syntetiske data kan gjøre det oppnåelig å sette i gang med forskningsprosjekt uten en REK-godkjenning.

Videre viser funn fra studien at tilgjengelige syntetiske datasett tillater etterprøving av forskningsfunn. Etterprøving av forskning er vanskelig grunnet personvernhensyn. Imidlertid vil anvendelsen av syntetiske data forenkle etterprøving og videreutvikling av datasett brukt i forskning, som vi ser i samsvar med funn fra studien. Denne muligheten kan knyttes til Walonoski et al. (2020), som hevder at syntetiske data gjør det mulig å behandle, lagre og dele datasett. I en forskningssammenheng kan dette bidra til å tilrettelegge for videre forskning ved deling av datasett, som på nåværende tidspunkt er begrenset grunnet personvernhensyn.

Studien indikerer at det er uvisst hvorvidt forskere vil ta i bruk syntetiske data i sitt arbeid. Årsaken forklares med en generell usikkerhet rundt kvaliteten og plausibiliteten tilknyttet syntetiske data. Innenfor forskning er datakvalitet ofte knyttet til blant annet kompletthet, validitet og aktualitet (Bray & Parkin, 2008). Samtidig viser funn fra studien at kvalitet på syntetiske data også innebærer volum, variasjon og bredde. Disse egenskapene fremheves, siden det kan påvirke i hvilken grad syntetiske data gjenspeiler virkeligheten. Funn viser at en

utfordring med syntetiske data er å dekke den samme variasjonen og spredningen på egenskaper som finnes i den reelle norske befolkningen. Dersom syntetiske datasett innfrir disse faktorene, kan det trolig øke tillit til datakvalitet.

5.3 Muligheter og utfordringer relatert til innovasjon

Funn fra denne studien viser at syntetiske data kan gi mer tilgjengelig data, som åpner opp for bedre og enklere tilgang til datasett. På bakgrunn av at syntetiske data består av stordata og KI-komponenter kan det anses som en datadrevet innovasjon (OECD, 2015; Yoo et al., 2020). Osmundsen et al. (2018) presenterer digital innovasjon og hvordan det kan ses i to ulike perspektiv, herunder som en prosess og et resultat. Funn fra studien viser hvordan syntetisk data kan være en innovasjon definert ut ifra begge perspektivene. I et resultatorientert perspektiv kan syntetiske data skape verdi som en støttefunksjon til både klinisk fagpersonell og forskere i helsesektoren. Ved å se på syntetisk data som en prosess, kan syntetiske data berike reelle data og på den måten kombineres data på nye måter. Som følge av dette tilrettelegges det for bruk av syntetiske data i en hel verdikjede.

Videre funn fra studien viser at målet med syntetiske data er at det skal etableres "as a service". Det betyr at syntetiske datasett skal være tilgjengelig, som en tjeneste, der brukere kan hente ut syntetiske data. Her trekkes det frem at syntetiske data kan skape hyllevare-algoritmer, som kan gjenbrukes av mange konsumenter innenfor forskjellige områder. Funnet om syntetiske data "as a service" viser at det kan bidra til innovasjon og næringsfremmende virksomhet. Bakgrunnen for dette er at det kan være kostbart og krevende å produsere syntetiske data, spesielt for mindre selskaper. Dette funnet kan dermed ses å fravike teori på feltet (Andrews, 2021; Ramos & Subramanyam, 2021). Dersom syntetiske datasett er tilgjengelige "as a service" vil dette tilrettelegge for at bedrifter får data som de har behov for, som kan bidra til økt innovasjonsaktivitet. Videre kan syntetiske data "as a service" knyttes til tjenesteinnovasjon, på bakgrunn av at syntetisk data kan ses som en ny teknologi, som kan anvendes på nye bruksområder og gi betydelige forbedringer i brukervennlighet (OECD & Eurostat, 2005).

Videre fremlegges det at syntetiske data kan benyttes i en innovasjonsprosess. Bakgrunnen for dette er at ved en produkt- eller en tjenesteidé vil det være nødvendig å foreta ulike tester for å avdekke mulige feil og løse dem fortløpende (Tidd og Bessant, 2018). Funn viser at

syntetiske data kan muliggjøre testing både før og under utvikling med tilgjengelige og realistiske data, som samsvarer med Walonoski et al. (2020). Videre belyser funn fra denne studien at en positiv konsekvens kan være tidsbesparelse, siden det ikke vil være nødvendig å foreta personvernshensyn og samtykke. På bakgrunn av dette kan bruk av syntetiske data føre til at et produkt eller en tjeneste raskere når markedet.

Vårt funn om at syntetisk data kan gi en forbedret støtteaktivitet for databehandling, ses i lys av at syntetiske data kan bidra til å redusere siloer og optimalisere arbeidsprosesser (Van Rossum et al., 2016). Ved bruk av syntetiske data kan deling av data forenkles, som kan være en innvirkende faktor for å bryte ned siloer i helsesektoren (Holstad, 2014). Funn viser at ved å ha åpne og delte syntetiske datasett kan siloer reduseres, noe som kan bidra til å avdekke mønstre i helsesektoren. Videre fremkommer det at fjerning av siloer kan føre til økt samarbeid på tvers av sektorer. Det samsvarer med Holstad (2014), som poengterer at å redusere siloer vil bidra til økt verdiskaping gjennom mer deling og bedre utnyttelse av dataene mellom offentlige og private virksomheter. Syntetiske data kan dermed forstås som innovasjonsfremmende.

Studien indikerer at syntetisk data kan virke innovasjonsfremmende i helsesektoren på flere områder. Funn viser at helsetjenester får flere KI-komponenter, og som følge av dette er det fordelaktig å benytte syntetiske data ved testing i en slik innovasjonsprosess. Bakgrunnen for dette er at helsetjenester med KI-komponenter trenes ofte på reelle data og ved å benytte syntetiske data kan testing foregå i et GDPR fritt område. Videre fremkommer det fra analysen at syntetisk data kan brukes som en støttefunksjon, i form av beslutningsstøtte ved å predikere tilleggsdiagnoser i det kliniske miljøet. I den forbindelse kan bruk av syntetiske data forstås som en hybrid intelligens, hvor formålet er å støtte, forbedre og akselerere menneskelige ressurser (Kolbjørnsrud, 2017). I det nevnte tilfellet kan syntetiske data ses som en prosessinnovasjon og dermed forstås som en ny støtteaktivitet for en helsetjeneste (Karlsson & Tavassoli, 2015). Syntetisk data som støttefunksjon kan optimalisere datasettene i det kliniske miljøet og følgelig lykkes med å avdekke hendelser hvor flere datavariabler opptrer samtidig (Ramos & Subramanyam, 2021). For at KI-modeller skal klare å oppdage sammenhenger hvor flere datavariabler opptrer samtidig, forutsetter det nok volum og variasjon (Zaslavsky et al., 2013). Videre viser funn at syntetiske data øker produktivitet og tilgang på ressurser i en forskningssammenheng, som kjennetegner prosessinnovasjoner (Hervas-Oliver et al., 2014).

Analysen fremlegger at syntetiske data muliggjør aktivt redusering av bias. Funnet antyder hvordan man kan bruke kunnskap for å fremme kvaliteten i datasettet. Gjennom bruk av syntetiske data vil det være mulig å tillegge domenekunnskap i KI-modeller, slik at kvaliteten på modellens prediksjoner forbedres (Ramos & Subramanyam, 2021). Videre tematiserer funnet hvordan de reelle datasettene i helsesektoren ikke er komplette, fordi det mangler konkrete hendelser. Syntetiske data vil imidlertid gjøre det mulig å fullstendiggjøre datasettene ved å generere flere sjeldne og ukjente hendelser (Ramos & Subramanyam, 2021). På den måten kan bruk av syntetiske data bedre speile virkeligheten, fordi man kan berike med data som ikke eksisterer i reelle data (Andrews, 2021).

5.4 Muligheter og utfordringer relatert til datakvalitet og kompetanse

Norge har behov for spesialister som kan utvikle og ta i bruk muliggjørende teknologier, spesielt innenfor kunstig intelligens (Meld. St. 22 (2020-2021)). Funn fra studien viser at det foreligger et behov for kompetanse og domeneeksperter innenfor syntetiske data. I helsesektoren anbefales det derfor et samarbeid mellom helseforetakene og domeneeksperter som sitter tett på løsningen med syntetiske data. Et hovedfunn trekker frem at kvaliteten på syntetiske data avhenger av hvem som produserer dem. Basert på funnet vil kompetanse være en forutsetning ved produksjon av syntetiske data. Dette underbygger James et al. (2021), som trekker frem viktigheten av å ha kompetente individer til å produsere syntetiske data. For å øke kompetansen på feltet, viser funn at det foregår en europeisk satsning på kunstig intelligens og syntetiske data, hvor Norge er delaktig i initiativet.

Funn fra denne studien vektlegger balansen mellom datamengde og kontroll. Ved syntetisering av data foreligger det et behov for volum, variasjon og bredde i datasettene for å sikre validitet (Zaslavsky et al, 2013). Samtidig viser funn at datamengde ikke kan gå på bekostning av kontroll. Funn viser at det kan være vanskelig å identifisere feil ved kvaliteten og å ha oversikt, dersom omfanget på dataene er for stor. For at det skal være mulig å avdekke feil i datasettene, vil graden av kontroll derfor være avgjørende. Variasjon i syntetiske datasett er derimot vesentlig for å dekke spredning på egenskaper som finnes i den reelle norske befolkningen. For å opprettholde kontroll og kvalitet, kan det tolkes som at ved produksjon av syntetiske data kan en stå i en spagat mellom volum og variasjon.

Videre fremkommer det at kvalitet i syntetiske data er også knyttet til hvorvidt disse er tilstrekkelig realistiske og dermed plausible. Dette samsvarer med Chen et al. (2019), som tematiserer viktigheten av realistiske syntetiske datasett. Likevel er det noen utfordringer knyttet til *for* realistiske syntetiske data, spesielt med tanke på reidentifisering (Arora & Arora, 2022; James et al., 2021). Funn fra studien som belyser denne utfordringen viser til hvordan ekstremverdier i syntetiske datasett kan avsløre personer på bakgrunn av en spesifikk geografi, yrke eller helsetilstand. En slik utfordring kan knyttes til Bergsjø og Bergsjø (2019), som problematiserer reidentifisering ved bruk av flere datakilder. Samtidig viser andre funn fra denne studien at ved å utelukke ekstremverdier kan det føre til bias i datasett, som kan ses i lys av Panch et al. (2019) og Staff (2015). Som følge av dette kan det tolkes som at det ikke finnes objektive metoder for å evaluere hvorvidt syntetiske data er tilstrekkelig forskjellig fra reell data og samtidig representative, slik som Arora og Arora (2022) problematiserer. Relatert til risiko for reidentifisering kan vi forstå dette som en interessant problemstilling ved syntetisk data, på bakgrunn av at det foreligger motstridende forskning på feltet. I den forbindelse kan vi trekke paralleller til James et al. (2021) som hevder at for å produsere syntetiske data er kompetanse vesentlig for å kunne sikre personvern i syntetiske datasett.

En utfordring ved syntetiske data som trekkes frem er mulighet for følgefeil. I syntetiske datasett er det derfor avgjørende med god kvalitet på input dataene som maskinlæringsmodellen skal basere seg på. Troverdigheten vil svekkes dersom det blir produsert syntetiske datasett med feil verdier, siden de ikke vil være realistiske (Bergsjø & Bergsjø, 2019). Basert på teori kan vi forstå utfordringen med følgefeil som spesielt knyttet til produksjon av berikede syntetiske data. På bakgrunn av at dataene som blir benyttet til berikelse ikke nødvendigvis er ment for samme formål og på den måten gjenspeiles ikke den aktuelle situasjonen (Bergsjø & Bergsjø, 2019). Funn viser at kompetansen er avgjørende dersom syntetiske data skal tas i bruk på lik måte som produksjonsdata i helsesektoren. Syntetisk data er komplekst og som følge av dette er det flere faktorer innen kompetanse som kan være vesentlige. Andre funn fremhever kompetanse i forhold til planlegging av produksjon, utnytting og deling av syntetiske data, som samsvarer med utfordringene James et al. (2021) beskriver.

Når det gjelder utfordringer knyttet til produksjon av syntetiske data er funnene todelt. Her ser vi at noen av informantene hevdet at selve produksjonen er komplisert og vanskelig, mens

andre mente at utfordringene vil være organisatoriske endringer. Det kan tolkes som at informantene har ulike perspektiver og oppfatninger på bakgrunn av deres kompetanse og erfaringer på feltet. Funn fra studien tilsier videre at kompetanse og datakvalitet henger sammen. Datakvalitet er essensielt for at syntetiske data skal gi en verdi, men for å sikre kvalitet kreves det kompetanse. Funn viser at kompetanse på feltet vil gradvis øke i takt med produksjon av syntetiske data. Som følge av dette er flere av informantene usikre på hvor tungt syntetisk data skal satses på i helsesektoren. Funn viser hvordan bruk av syntetiske data kan føre til både tekniske og organisatoriske endringer, som samsvarer med James et al. (2021). Slike endringer kan forstås som nye tekniske prosesser og behov for opplæring knyttet til bruk av syntetiske data.

I denne studien anser vi tekniske og organisatoriske endringer som en utfordring, i tillegg til behovet for riktig kompetanse til å gjennomføre de tekniske endringene knyttet til syntetisk data. På bakgrunn av disse funnene kan det tenkes at dersom syntetisk data skal anvendes i helsesektoren vil det kreve økt bevissthet rundt menneskene som tar i bruk syntetiske data og deres kompetansenivå (Johansen & Sæterdal, 2017).

5.5 Utfordringer og anbefalinger relatert til tillit

Ifølge Gillath et al. (2021) er mangel på tillit en av de største hindringene som står i veien for å dra full nytte av fordelene kunstig intelligens har å tilby. Sett i sammenheng med bruk av syntetisk data i helsesektoren, viser funn fra denne studien at det er en utfordring å skape tillit til syntetiske data. James et al. (2021) problematiserer dette og hevder at det er uvisst hvorvidt konklusjoner basert på syntetiske data vil bli akseptert av vitenskapelige og medisinske miljøer. Funn viser at manglende tillit kan stamme fra blackbox-problematikken (Gille et al. 2020), som knyttes til bruk av syntetisk data som beslutningstøtte. Årsaken til dette skyldes tvil om kunstig intelligens kan ta etiske beslutninger og hensyn i slike situasjoner. På bakgrunn av dette kan det forstås som at foreligger en form for sårbarhet og villighet til å ta risiko ved å benytte syntetiske data (Mayer et al., 1995).

Et annet funn viser at tillit til syntetisk data skapes over tid, som kan bety at når kompetanse på området øker kan dette påvirke tilliten til syntetisk data positivt. Dette samsvarer med O´Neill, som belyser hvordan kompetanse og pålitelighet er essensielle elementer som inngår

i tillit (Ingierd, 2017). Det kan derfor forstås i denne sammenheng at dersom forskere og klinisk fagpersonell vurderer syntetisk data som pålitelig, kan dette bidra til å skape tillit til syntetisk data. For å oppnå tillit kan vi dermed tolke datakvalitet som en vesentlig faktor for hvorvidt syntetiske datasett blir ansett som pålitelige. Syntetisk data er ikke implementert i helsesektoren, og som følge av dette viser funn at syntetisk data i stor grad er ukjent for både forskere og klinisk fagpersonell. På bakgrunn av dette anser vi omdømme som et relevant attributt ved datakvalitet (Hazen et al., 2014), siden det kan forstås som å ha en innvirkning på tilliten til syntetisk data i det kliniske miljøet.

Som funn viser foreligger det en generell skepsis til syntetisk data, spesielt i forskningssammenheng. Tidligere forskning (Gillath et al., 2021; Gille et al., 2020), viser at det er mulig å skape tillit til kunstig intelligens og vi anser det derfor som overførbart til syntetisk data. Gillath et al. (2021) hevder at ved å redusere oppfatningen av risiko knyttet til bruk kan dette skape tillit. Sett i lys av dette kan det anbefales å gjøre forskere kjent med Helseanalyseplattformen, som skal tilby syntetiske datasett. Bruk av Helseanalyseplattformen kan føre til at forskere får egne erfaringer med syntetisk data i en skjermet arena. James et al. (2021) tematiserer dette og hevder at inntil større tillit har blitt etablert, vil kliniske miljøer forvente å se resultater direkte demonstrert på virkelige emner. Det kan derfor tenkes at erfaringer gjennom en slik plattform kan innledningsvis bidra til å redusere oppfatning av risiko tilknyttet syntetisk data.

6.0 Konkluderende avslutning

6.1 Konklusjon

Den teknologibaserte helsenæringen utvikler seg raskt, og som følge av dette oppstår nye muligheter og utfordringer. Teknologi og innovasjon kan være vesentlige faktorer for forbedring av tunge prosesser i helsesektoren, samtidig foreligger det ulike regulatoriske hindringer knyttet til dette. Kunstig intelligens, i form av syntetisk data, kan muliggjøre bedre utnyttelse av helsedata og deling på bakgrunn av innebygd personvern. I denne studien har vi valgt å undersøke hvilken betydning syntetisk data kan ha for helsesektoren og i helseforskning. For å besvare den overordnede problemstillingen ble det utformet to forskningsspørsmål:

"Hvordan kan bruk av syntetisk data skape muligheter for helsesektoren?"

"Hvilke utfordringer bør helsesektoren være bevisst på tilknyttet bruk av syntetisk data?"

Relatert til det første forskningsspørsmålet har vi avdekket at syntetisk data kan være en løsning for personvernutfordringer knyttet til data. Bruk av syntetiske data tilgjengeliggjør store mengder med data og gjør det mulig å avdekke mønstre i et datasett, som kan være tidkrevende og potensielt uoppnåelig med reelle data. Faktorer som redusering av siloer og deling av data internt mellom avdelinger og eksternt mellom aktører, vil på den måten virke innovasjonsfremmende, samtidig som personvern ivaretas. Studien har funnet at syntetisk data kan bistå i ulike kliniske utredninger som en beslutningsstøtte i helsesektoren. Videre kan syntetiske data i større grad muliggjøre etterprøving av forskningsresultater, samt gi verdi gjennom ytterligere tilgjengelige data som kan effektivisere en forskningsprosess.

Det andre forskningsspørsmålet bygger videre på det første ved å belyse noen av utfordringene helsesektoren må ta stilling til ved anvendelse av syntetiske data. Det er nødvendig med tilstrekkelig kompetanse for å sikre kvalitet i produksjon av syntetiske data. Her ses datakvalitet i tråd med faktorer som volum, variasjon, bredde og kontroll. For at helsesektoren og forskere skal ta i bruk syntetiske data, må det foreligge en tillit til at dataene er plausible og gjenspeiler den virkelige verden. Imidlertid er det nødvendig å ta hensyn til mulig reidentifisering av enkeltindivider. Fravær av lovgiving og retningslinjer tilknyttet syntetisk data ses per i dag også som en utfordring.

6.2 Implikasjoner

Gjennom studien har vi avdekket noen teoretiske og praktiske implikasjoner, så vel som metodiske begrensninger, som presenteres i de følgende delkapitlene.

6.2.1 Teoretiske implikasjoner

Studien har identifisert hvilken betydning bruk av syntetisk data kan ha på helsesektoren og helseforskning. På bakgrunn av at det foreligger lite eksisterende forskning på syntetiske data og at det ikke er implementert i helsesektoren, har det vært utfordrende å trekke slutninger.

Vi er blant de første som ser nærmere på fenomenet og studien gir dermed et vesentlig bidrag til litteraturen om syntetisk data. Formålet med studien er å gi leseren økt kunnskap og forståelse på et nytt felt i Norge, som bærer preg av lite modenhet. I studien har vi avdekket et behov for syntetiske data og fremhevet en teknologi som anses å bli viktig i fremtiden, på bakgrunn av regulatoriske krav til personvern og sikkerhet ved databehandling. Samtidig belyses utfordringer knyttet til kompetanse knyttet til produksjon og kvalitet av syntetiske data. Ved å fremlegge mulighetsrommet som syntetisk data kan gi og samtidig belyse hvilke utfordringer som bør bevisstgjøres ved bruk, kan det tilrettelegge for optimal utnyttelse av syntetiske data. Vi anser studien som relevant for å forstå kompleksiteten ved syntetisk data.

6.2.2 Praktiske implikasjoner

Vår studie fremhever hvordan syntetisk data kan gi verdi for helsesektoren og for helseforskning, i form av mer og bedre datagrunnlag. Vi har kastet lys over flere muligheter og utfordringer som vil være aktuelle ved implementering av syntetiske data. Ved å være bevisst på de nevnte utfordringene kan man identifisere kritiske faktorer for å unngå mulige fallgruver. Dersom syntetiske data skal tas i bruk på lik linje som reelle data, vil kompetanse og datakvalitet være en forutsetning. Samtidig er det viktig å påpeke at syntetisk data vil skape et stort mulighetsrom for innovasjon og forskning. For å sikre at syntetiske data blir tatt i bruk i praksis, bør det være fokus på tillitsskapende tiltak.

Studien er relevant for aktører som tar i bruk, eller som ønsker å benytte syntetiske data. Eksempelvis Norsk Helsenett, Helse Vest IKT, Helseanalyseplattformen, Skatteetaten og NAV. Flere av resultatene i studien vurderes som overførbare til andre sektorer, på bakgrunn av generell økende bruk av stordata og kunstig intelligens i næringslivet. Studien har belyst hvordan reguleringer legger føringer for behandling av persondata i helsesektoren, som kan ses gjeldende for andre bransjer. Resultatene er spesielt aktuelle for selskaper som behandler og analyserer store mengder data, men som ikke har tilgang på fullstendige datasett eller som trenger supplerende data.

6.3 Begrensninger og forslag til videre forskning

6.3.1 Begrensninger ved studien

I starten av arbeidet med masteroppgaven utforsket vi ulike tematikker, og utarbeidingen av vår problemstilling skjedde først etter kontakt med SYNDATA-prosjektet litt ut i semesteret. Vi valgte å forske på betydningen bruk av syntetisk data kan ha for helsesektoren. På bakgrunn av at vi har tatt utgangspunkt i en omfattende sektor, kan det være sannsynlig at vi har gått glipp av nyttig informasjon ved at vi har avgrenset oss til å intervjue 11 fagpersoner tilknyttet feltet. Vi har reflektert over antall informanter, hvor vi ser at flere intervju med forskere, kunne gitt en dypere og mer spesifikk innsikt i fenomenet og dets påvirkning. Samtidig har vi totalt intervjuet 11 personer som kunne belyse temaet fra ulike perspektiver.

En annen begrensning ved studien er gjennomføring av intervju. Samtlige dybdeintervju ble gjennomført via digitale kommunikasjonsplattformer på grunn av koronarestriksjoner og logistikk. Det kan ha påvirket sannsynligheten for eventuelle misforståelser underveis og bearbeiding av informasjon. Gjennomføring av digitale semistrukturerte dybdeintervju kan hatt en innvirkning på tolkning av kroppsspråk, samt grad av åpenhet og tillit. Samtidig er ikke temaet i utgangspunktet sensitivt og vi oppfattet ikke at informanter opplevde det slik.

Videre var vår faglige kunnskap og våre erfaringer på feltet minimale ved oppstart. Dette sett opp mot fenomenets kompleksitet, gjorde det vanskelig å se hele bildet og samtidig ta hensiktsmessige begrensninger. Vår manglende kunnskap på tidspunktet forstår vi som en mulig faktor for tematikkens bredde, hvor det kunne vært fordelaktig med ytterligere avgrensninger. Til tross for dette har læringskurven vært bratt og kunnskapen har økt parallelt med utarbeidingen av masteroppgaven.

Avslutningsvis kan det trekkes frem at første halvdel av studien ble utført med en annen problemstilling. Den bar preg av idealisme, hvor vi ønsket å se på hvordan bruk av syntetiske data i helsesektoren tilrettela for innovasjon i forskning. På bakgrunn av at syntetiske data foreløpig ikke er etablert som en tjeneste i helsesektoren, ble dette for abstrakt og utfordrende å undersøke. Den overordnede problemstillingen og forskningsspørsmålene ble derfor endret underveis. Vi valgte til slutt å undersøke hvilke muligheter bruk av syntetisk data kan gi helsesektoren, samt hvilke utfordringer som bør bevisstgjøres som følge av dette. Gjennom

presentasjon av relevant teori og datainnsamling, mener vi at studien besvarer problemstillingen og forskningsspørsmålene.

6.3.2 Forslag til videre forskning

Fenomenet syntetisk data kan skape et mulighetsrom som gir stor verdi for helsesektoren og helseforskning. Vår studie har fokusert på en sektor som er svært kompleks og bærer preg av strengt regulert personvern. For videre forskning kunne det vært interessant å sett på implementering av syntetiske data i helsesektoren og hvordan det kan skje på en konstruktiv måte. Her kunne fokus på forvaltning og deling av syntetiske data vært en spennende vinkling. På bakgrunn av at studien har belyst eksempler på bruk av syntetiske data gjennom Skatteetaten og Helseanalyseplattformen, kunne en casestudie på en eller begge av dem vist seg interessant. Herunder kunne bruk av syntetiske testdata i en innovasjonsprosess gitt verdifull innsikt for tematikken. Fremtidig forskning på syntetisk data i andre sektorer og bransjer kunne blitt studert for å identifisere forskjeller og likheter i funn fra helsesektoren, og trukket paralleller på tvers av disse. Avslutningsvis anser vi studien som en grobunn på feltet, som identifiserer en rekke muligheter for videre forskning. Vi håper at innsikten vi har gitt vil inspirere og gi verdi for fremtidig forskning på syntetisk data.

7.0 Referanseliste

Aasen, T. M., & Amundsen, O. (2011). *Innovasjon som kollektiv prestasjon*. Gyldendal Akademisk.

Alstveit, M., Halvorsen, A., Willumsen, E., & Ødegård, A. (2016). Lederen som innovatør og balansekunstner: en kvalitativ studie av lederes erfaringer fra forskningssamarbeid mellom helse- og velferdstjenestene og høyere utdanning. *Nordisk Tidsskrift for Helseforskning*, *12*(2). https://doi.org/10.7557/14.4051

Andreu-Perez, J., Poon, C., Merrifield, R., Wong, S., & Yang, G. (2015). Big Data for Health. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1193-1208. https://doi.org/10.1109/JBHI.2015.2450362

Andrews, G. (2021, 8 juni). *What is Synthetic data?* Nvidia. Hentet 20. mars 2022 fra https://blogs.nvidia.com/blog/2021/06/08/what-is-synthetic-data/

Arora, A. & Arora, A. (2022). Synthetic patient data in health care: a widening legal loophole. *The Lancet, 399(10335)*, 1601-1602. https://doi.org/10.1016/S0140-6736(22)00232-X

Askheim, O., & Grenness, T. (2008). *Kvalitative metoder for markedsføring og organisasjonsfag*. Universitetsforl.

Aspøy, T. M., & Andersen, R. K. (2015). *Digital kompetanse i arbeidslivet*. Fafo-rapport 2015:28.

Benke, K., & Benke, G. (2018). Artificial Intelligence and Big Data in Public Health. *International Journal of Environmental Research and Public Health*, *15*(12), 2796. https://doi.org/10.3390/ijerph15122796

Bergsjø, L., & Bergsjø, Håkon. (2019). *Digital etikk: Big data, algoritmer og kunstig intelligens*. Universitetsforlaget.

Birkeland, J. (2019, 21. februar). *Syntetisk folkeregister.* Computerworld. Hentet 4. mai 2022 fra https://www.cw.no/offentlig-sektor-skatteetaten-test/syntetisk-folkeregister/776216

Bray, F. & Parkin, D. M. (2008). Evaluation of data quality in the cancer registry: Principles and methods. Part I: Comparability, validity and timeliness. *European Journal of Cancer*, *45*(5), 747–755. https://doi.org/10.1016/j.ejca.2008.11.032

Cambridge Dictionary (u.å). Data. *Cambridge Dictionary.* Hentet 16. mai 2022 fra https://dictionary.cambridge.org/dictionary/english/data

Chen, J., Chun, D., Patel, M., Chiang, E., & James, J. (2019). The validity of synthetic clinical data: A validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Medical Informatics and Decision Making*, *19*(1), 44–44. https://doi.org/10.1186/s12911-019-0793-0

Datatilsynet (2019a, 17. juni). *Hva er en personopplysning?* Datatilsynet. Hentet 11. mars 2022 fra https://www.datatilsynet.no/rettigheter-og-plikter/personopplysninger/

Datatilsynet. (2019b, 16. juli). *Personvernprinsippene.* Datatilsynet. Hentet 18. april 2022 fra https://www.datatilsynet.no/rettigheter-og-plikter/personvernprinsippene/

Davenport, T.H., & Kirby, J. (2016). *Only humans need apply: Winners and losers in the age of smart machines*. Harper Business.

De nasjonale forskningsetiske komiteene. (u.å.). *Ny lovgivning om personopplysinger - hva betyr det for forskning?*. De nasjonale forskingsetiske komiteene. Hentet 18. februar 2022 fra https://www.forskningsetikk.no/ressurser/gdpr/

De nasjonale forskningsetiske komiteene. (2014, 10. oktober). *Regionale komiteer for medisinks og helsefaglig forskningsetikk (REK)*. De nasjonale forskingsetiske komiteene. Hentet 18. mai 2022 fra https://www.forskningsetikk.no/ressurser/gdpr/

Dellermann, D., Calma. A., Lipusch, N., Weber, T., Weigel, S., Ebel, P. (2019). *The future of human-ai collaboration: a taxonomy of design knowledge for hybrid intelligence systems*. I: Hawaii international conference on system sciences (HICSS). Hawaii, USA https://doi.org/10.24251/HICSS.2019.034

Digdir. (u.å.). *Hva er Schrems II-dommen*. Digitaliseringsdirektoratet. Hentet 22. april 2022 fra https://www.digdir.no/handlingsplanen/hva-er-schrems-ii-dommen/2581

Dilmegani, C. (2022, 14. februar). *The Ultimate Guide to Synthetic Data: Uses, Benefits & Tools.* AI Multiple. Hentet 25. februar 2022 fra https://research.aimultiple.com/synthetic-data/

Direktoratet for e-helse. (2019). *Forprosjekt - Utredning om bruk av kunstig intelligens i helsesektoren*, IE-1058. Direktoratet for e-helse. https://www.ehelse.no/publikasjoner/utredning-om-bruk-av-kunstig-intelligens-i-helsesektoren

Direktoratet for e-helse. (2020). *Utviklingstrekk 2020,* IE-1055. Direktoratet for e-helse. https://www.ehelse.no/publikasjoner/rapport-utviklingstrekk-2020

Direktoratet for e-helse. (2022, februar). *Innsiktsrapport - Behov for data til kunstig intelligens i helsetjenesten*, IE-1096. Direktoratet for e-helse. https://www.ehelse.no/publikasjoner/behov-for-data-til-kunstig-intelligens-i-helsetjenesten

Easterby-Smith, M., Thorpe, R., Jackson, P. R., & Jaspersen, L. J. (2015). *Management and business research* (5, utg.). SAGE.

Easterby-Smith, M., Thorpe, R., Jackson, P. R., & Jaspersen, L. J. (2018). *Management and business research* (6, utg.). SAGE.

Easterby-Smith, M., Thorpe, R., Jaspersen, L. J & Valizade D. (2021). *Management and business research* (7. utg.). SAGE.

EU-kommisjonen. (2006). *Key Competences for Lifelong Learning: European Reference Framework*. EU-kommisjonen https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32006H0962#:~:text=The Reference Framework sets out,7) Sense of initiative and

EU-kommisjonen. (2018, 24. april). *Artificial Intelligence in Europe*. EU-kommisjonen https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=EN

Fagerberg, J., Mowery, D. C., Nelson, R. R., Asheim, B. T., Bruland, K., & Grodal, S. (2005). *The Oxford handbook of innovation*. Oxford University Press.

Gambetta, D. G. (Ed.). 1988. Can we trust trust? In D. G. Gambetta (Ed.), Trust: 213-237. *Basil Blackwell.*

Garcia, R., & Calantone, R. (2002). A critical look at technological innovation typology and innovativeness terminology: a literature review. *The Journal of Product Innovation Management*, *19*(2), 110–132. https://doi.org/10.1111/1540-5885.1920110

Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., & Spaulding, R. (2021, februar). Attachment and trust in artificial intelligence. *Computers in Human Behavior*, *115*, 106607. https://doi.org/10.1016/j.chb.2020.106607

Gille, F., Jobin, A., & Ienca, M. (2020). What we talk about when we talk about trust: Theory of trust for AI in healthcare. *Intelligence-Based Medicine*, *1-2*, 100001. https://doi.org/10.1016/j.ibmed.2020.100001

Goleman, D. (1996). *Emotional Intelligence: Why It Can Matter More than IQ*. Bloomsbury Publishing.

Gripsrud, G., Olsson, U.H. og Silkoset, R. (2010). *Metode og dataanalyse: beslutningsstøtte for bedrifter ved bruk av JMP*. (2. utg.) Høyskoleforlaget.

Günther, W. A., Rezazade Mehrizi, M. H., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems*, *26*(3), 191–209. https://doi.org/10.1016/j.jsis.2017.07.003

Halvorsen, K. (2014). *Å forske på samfunnet - en innføring i samfunnsvitenskapelig metod*e (5. utg.). Cappelens Forlag AS.

Hazen, B., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, *154*, 72–80. Hentet 6. mai 2022 fra https://doi.org/10.1016/j.ijpe.2014.04.018

Heggernes. T. A., (2020). *Digital forretningsforståelse : fra store data til små biter* (3. utg.). Fagbokforlaget.

Helsedatautvalget. (2017). *Et nytt system for enklere og sikrere tilgang til helsedata: Rapport fra Helsedatautvalget 2016-2017*. Regjeringen. https://www.regjeringen.no/contentassets/1fe9cf37e64344e1a3b3c62f950b100b/170630_helsedatalovutvalget.pdf

Helseforskningsloven. (2009). Lov om medisinsk og helsefaglig forskning. (LOV-2020-12-04-133). Lov data. Hentet 20. mai 2022 fra https://lovdata.no/dokument/NL/lov/2008-06-20-44

Helse Stavanger (2022). *PILOT SYNDATA.* Helse Stavanger Stavanger Universitetssykehus. Hentet 27.04.2022 fra https://helse-stavanger.no/pilot-syndata

Hervas-Oliver, J.L., Sempere-Ripoll F. & Boronat-Moll, C. (2014). Process innovation strategy in SMEs, organizational innovation and performance: a misleading debate? *Small Business Economics*, *43*(4), 873–886. https://doi.org/10.1007/s11187-014-9567-3

Hokstad, I. (2019). Kunstig intelligens krever sunn fornuft. *Tidsskriftet for den Norske Legeforening, 139*(13). https://doi.org/10.4045/tidsskr.19.0435

Holstad, B. (2014). Mer deling og bedre informasjonsforvaltning kan stanse tidstyver. *Stat & styring, 2, 32-34*. https://doi-org.galanga.hvl.no/10.18261/ISSN0809-750X-2014-02-14

Huang, M-H., & Rust, R. T. (2018). Artificial Intelligence in Service. *Journal of Service Research : JSR*, *21*(2), 155–172. https://doi.org/10.1177/1094670517752459

Iden, J., Eikebrokk, T.R. & Marrone, M. (2020). Process reference frameworks as institutional arrangements for digital service. *International Journal of Information Management, 54*, 102150. https://doi.org/10.1016/j.ijinfomgt.2020.102150

Jacobsen, D. (2015). *Hvordan gjennomføre undersøkelser? : Innføring i samfunnsvitenskapelig metode* (3. utg.). Cappelen Damm akademisk.

Jakhar, D., & Kaur, I. (2020). Artificial intelligence, machine learning and deep learning: Definitions and differences. *Clinical and Experimental Dermatology*, *45*(1), 131-132. https://doi.org/10.1111/ced.14029

James, S., Harbron, C., Branson, J., & Sundler, M. (2021). Synthetic data use: exploring use cases to optimise data utility. *Discover Artificial Intelligence*, *1*(1), 1–13.

Ingierd, H. (2017). Troverdighet før tillit. *Nytt norsk tidsskrift*, *3*, 317–323. https://doi.org/10.18261/issn.1504-3053-2017-03-09

Johannessen, A., Christoffersen, L., & Tufte, P. A. (2011). *Forskningsmetode for økonomisk-administrative fag* (3. utg., p. 490). Abstrakt forl.

Johannessen, A., Christoffersen, L., & Tufte, P. (2016). *Introduksjon til samfunnsvitenskapelig metode* (5. utg.). Abstrakt.

Johansen, O., & Sætersdal, H. I. (2017). *HR og personalledelse*. Fagbokforlaget.

Johnson-George. C, & Swap, W. (1982). Measurement of specific interpersonal trust: Construction and validation of a scale to assess trust in a specific other. *Journal of Personality and Social Psychology*. 43(6), 1306-1317. https://doi/10.1037/0022-3514.43.6.1306

Johnson, H. (2014, 22. februar). 6 Soft Skills Every Professional Needs. *Business 2 Community*. Hentet 24. januar 2022 fra https://www.business2community.com/human-resources/6-soft-skills-every-professional-needs-0788441

Karlsson, C. & Tavassoli, S. (2015). Innovation strategies of firms: What strategies and why? *The Journal of Technology Transfer*, *41*(6), 1483–1506. https://doi.org/10.1007/s10961-015-9453-4

Kolbjørnsrud, V. (2017). Kunstig intelligens og lederens nye jobb. *Magma*, s. 33-42. Hentet 26. februar 2022 fra https://biopen.bi.no/bixmlui/bitstream/handle/11250/2460933/Kunstig%20intelligens%202017.pdf?sequence=1&isAllowed=y

Kommunal- og moderniseringsdepartementet. (2020). *Nasjonal strategi for kunstig intelligens*. Regjeringen. https://www.regjeringen.no/contentassets/1febbbb2c4fd4b7d92c67ddd353b6ae8/no/pdfs/ki-strategi.pdf

Lanestedt, G. (2016). Stordata og kunnskapsbasert forvaltning. *Stat & Styring*, (2), 52-54.

Larsen, E. (2017). Informasjon om helsen din fra helseregistre og biobanker kan bidra til livreddende forskning. Men hvor mye skal forskerne få tilgang til? *Tidsskriftet GENialt. Bioteknologirådet.* Hentet 13 mai 2022 fra https://www.bioteknologiradet.no/2017/01/norske-helseregistre-en-skjult-gullgruve/

Mayer, R., Davis, J., & Schoorman, F. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792

Meld. St. 7 (2019-2020). *Nasjonal helse- og sykehusplan 2020-2023.* Helse- og omsorgsdepartementet. https://www.regjeringen.no/contentassets/95eec808f0434acf942fca449ca35386/no/pdfs/stm201920200007000dddpdfs.pdf

Meld. St. 22 (2020-2021). *Data som ressurs: Datadrevet økonomi og innovasjon.* Kommunal- og moderniseringsdepartementet. https://www.regjeringen.no/contentassets/4f357e18bd314dc08c8e1b447b71b700/no/pdfs/stm202020210022000dddpdfs.pdf

Mendling, J., Decker, G., Reijers, H. A., Hull, R. (2018). How do machine learning, robotic process automation, and blockchains affect the human factor in business process management? *Communications of the Association for Information Systems*, 43, 297-320. https://doi.org/10.17705/1CAIS.04319

Nambisan, S. (2017). Digital Entrepreneurship: Toward a Digital Technology Perspective of Entrepreneurship. *Entrepreneurship Theory and Practice*, *41*(6), 1029–1055. https://doi.org/10.1111/etap.12254

Nasjonalt servicemiljø for medisinske kvalitetsregistre. (u.å). *Datakvalitet.* Nasjonalt servicemiljø for medisinske kvalitetsregistre. Hentet 8. mai 2022 fra https://www.kvalitetsregistre.no/datakvalitet

NOU 2005: 1. (2005). *God forskning: bedre helse*. Helse- og omsorgsdepartementet. https://www.regjeringen.no/contentassets/848476c900bb455abdca39ccef4733af/no/pdfs/nou2005200500010000dddpdfs.pdf

Nowok, B., Raab, G.M. & Dibben, C. (2016). Bespoke Creation of Synthetic Data in R. (11. utg). *Journal of Statistical Software*. 10.18637/jss.v074.i11

OECD & Eurostat. (2005). *Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data* (3. utg.). OECD Publications. https://ec.europa.eu/eurostat/documents/3859598/5889925/OSLO-EN.PDF

OECD. (2015). *Data-driven innovation: Big data for growth and Well-Being.* OCED Publications. https://doi.org/10.1787/9789264229358-en

Oliveira, P., Rodrigues, F., & Henriques, P. R. (2005, November). A formal definition of data quality problems. I *ICIQ*.

Osmundsen, K., Iden, J. & Bygstad, B. (2018). *Hva er digitalisering, digital innovasjon og digital transformasjon*. NOKOBIT, 26(1), 1-15.

Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: implications for health systems. *Journal of Global Health*, *9*(2), 010318–010318. https://doi.org/10.7189/jogh.09.020318

Personopplysningsloven. (2018). *Lov om behandling av personopplysninger* (LOV-2018-06-15-38). Lovdata. Hentet 20. april 2022 fra https://lovdata.no/dokument/NL/lov/2018-06-15-38?q=personopplysningsloven

Pettersen, K. H. (2019). Kunstig intelligens vil endre helsetjenesten. *Tidsskrift for den norske Legeforening*, *139*(14). https://doi.org/10.4045/tidsskr.19.0479

Ramos, L & Subramanyam, J. (2021). *Forget About Your Real Data: Synthetic Data Is the Future of AI*. Gartner. Hentet 15 mars 2022 fra https://www.gartner.com/document/4002912

Rizk, A., Ståhlbröst, A., & Elragal, A. (2020). Data-driven innovation processes within federated networks. *European Journal of Innovation Management*, *25*(6), 498–526. https://doi.org/10.1108/EJIM-05-2020-0190

Rousseau, D., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *The Academy of Management Review*, *23*(3), 393–404. https://doi.org/10.5465/AMR.1998.926617

Shenton, A. (2004). Strategies for ensuring trustworthiness in qualitative research projects. *Education for Information*, *22*(2), 63–75. https://doi.org/10.3233/EFI-2004-22201

Skorstad, E., Schulze, O. & Nilsen, D. Øyvind. E. (2008). Rett person på rett plass: psykologiske metode i rekruttering og lederutvikling (s. 267). Gyldendal Akademisk.

Staff, A. (2015, 23. juni). *Bias.* De nasjonale forskningsetiske komiteene. Hentet 08. mai 2022 fra https://www.forskningsetikk.no/ressurser/fbib/uavhengighet/bias/

Stanfill, M., & Marc, D. T. (2019). Health Information Management: Implications of Artificial Intelligence on Healthcare Data and Information Management. *Yearbook of Medical Informatics*, *28*(1), 056–064. https://doi.org/10.1055/s-0039-1677913

Sternberg, R. J. (1997). A Triarchic View of Giftedness: Theory and Practice. *Handbook of Gifted Education*, N. Coleangelo and G. A. Davis, eds. Allyn and Bacon, 43-53.

Sternberg, R. J. (1999). The Theory of Successful Intelligence. *Review of General Psychology*, *3*(4), 292-316. https://doi.org/10.1037/1089-2680.3.4.292

Sternberg, R. J. (2005). The Theory of Successful Intelligence. *Interamerican Journal of Psychology*, 39(2), 189-202.

Sørebø, Ø., Fredriksen, J., Simnica, F. & Mollestad, H. (2020,1. oktober). EUs personvernforordning (GDPR). *Praktisk økonomi og finans*, *3*, 240–256. https://doi.org/10.18261/issn.1504-2871-2020-03-07

Tennøe, T. & Prabhu, R. (2017). Kunstig intelligens og norsk politikk. *Nytt norsk tidsskrift, 2, 205–216. https://doi.org/10.18261/issn.1891-1781-2017-02-09*

Tidd, J., & Bessant, J. (2018). *Managing innovation: integrating technological, market and organizational change* (6. utg.). Wiley.

Van Rossum, L., Aij, K. H., Simons, F. E., van der Eng, N., & ten Have, W. D. (2016). Lean healthcare from a change management perspective. *Journal of Health Organization and Management*, *30*(3), 475–493. https://doi.org/10.1108/JHOM-06-2014-0090

Veregin, H. (1999). Data quality parameters. *Geographical information systems*, *1*, 177-189. https://www.geos.ed.ac.uk/~gisteac/gis_book_abridged/files/ch12.pdf

Vigar, G., Cowie, P., & Healey, P. (2020). Innovation in planning: creating and securing public value. *European Planning Studies*, *28*(3), 521–540. https://doi.org/10.1080/09654313.2019.1639400

Walonoski, J., Klaus, S., Granger, E., Hall, D., Gregorowicz, A., Neyarapally, G., Watson, A., & Eastman, J. (2020) Synthea™ Novel coronavirus (COVID-19) model and synthetic data set. *Intelligence-Based Medicine*, *1-2*, 100007–100007. https://doi.org/10.1016/j.ibmed.2020.100007

Wilkinson, L. (2018). Visualizing Big Data Outliers Through Distributed Aggregation. *IEEE Transactions on Visualization and Computer Graphics*, *24*(1), 256–266. https://doi.org/10.1109/TVCG.2017.2744685

Yin, R.K. (2014). *Case Study Research: Design and Methods* (5. utg.). SAGE.

Yin, R. K. (2018). *Case Study Research: Design and Methods* (6. utg.). SAGE.

Yoo Y, Henfridsson O, Lyytinen K (2010) Research commentary— t*he new organizing logic of digital innovation: an agenda for information systems research*. Inf Syst Res 21(4):724–735

Zaslavsky, A., Perera, C., & Georgakopoulos, D., (2013). Sensing as a Service and Big Data. *Cornell University*. https://doi.org/10.48550/ar.Xiv.1301.0159

Østbye, T. (2020). Digitalisering som varmer. *Praktisk økonomi & finans*, *1*, 39–46. https://doi.org/10.18261/issn.1504-2871-2020-01-06

Åm, H. Frøyhaug, M., & Tøndel, G. (2021. Helsedata som gullgruve? *Nytt norsk tidsskrift*, *1-02*, 86–98. https://doi.org/10.18261/issn.1504-3053-2021-01-02-08

8.0 Vedlegg

8.1 Vedlegg 1 – Godkjenning fra NSD

Meldeskjema / Masteroppgave - Syntetisk data / Vurdering

Vurdering

Referansenummer
980681

Prosjekttittel
Masteroppgave - Syntetisk data

Behandlingsansvarlig institusjon
Høgskulen på Vestlandet / Fakultet for økonomi og samfunnsvitskap / Institutt for økonomi og administrasjon

Prosjektperiode
23.03.2022 - 01.07.2022

Meldeskjema

Dato
10.03.2022

Type
Standard

Kommentar
OM VURDERINGEN
Personverntjenester har en avtale med institusjonen du forsker eller studerer ved. Denne avtalen innebærer at vi skal gi deg råd slik at behandlingen av personopplysninger i prosjektet ditt er lovlig etter personvernregelverket.

Personverntjenester har nå vurdert den planlagte behandlingen av personopplysninger. Vår vurdering er at behandlingen er lovlig, hvis den gjennomføres slik den er beskrevet i meldeskjemaet med dialog og vedlegg.

DEL PROSJEKTET MED PROSJEKTANSVARLIG
For studenter er det obligatorisk å dele prosjektet med prosjektansvarlig (veileder). Del ved å trykke på knappen «Del prosjekt» i menylinjen øverst i meldeskjemaet. Prosjektansvarlig bes akseptere invitasjonen innen en uke. Om invitasjonen utløper, må han/hun inviteres på nytt.

TYPE OPPLYSNINGER OG VARIGHET
Prosjektet vil behandle alminnelige kategorier av personopplysninger frem til den datoen som er oppgitt i meldeskjemaet.

LOVLIG GRUNNLAG
Prosjektet vil innhente samtykke fra de registrerte til behandlingen av personopplysninger. Vår vurdering er at prosjektet legger opp til et samtykke i samsvar med kravene i art. 4 og 7, ved at det er en frivillig, spesifikk, informert og utvetydig bekreftelse som kan dokumenteres, og som den registrerte kan trekke tilbake.

Lovlig grunnlag for behandlingen vil dermed være den registrertes samtykke, jf. personvernforordningen art. 6 nr. 1 bokstav a.

PERSONVERNPRINSIPPER
Personverntjenester vurderer at den planlagte behandlingen av personopplysninger vil følge prinsippene i personvernforordningen om:
• lovlighet, rettferdighet og åpenhet (art. 5.1 a), ved at de registrerte får tilfredsstillende informasjon om og samtykker til behandlingen
• formålsbegrensning (art. 5.1 b), ved at personopplysninger samles inn for spesifikke, uttrykkelig angitte og berettigede formål, og ikke behandles til nye, uforenlige formål
• dataminimering (art. 5.1 c), ved at det kun behandles opplysninger som er adekvate, relevante og nødvendige for formålet med prosjektet
• lagringsbegrensning (art. 5.1 e), ved at personopplysningene ikke lagres lengre enn nødvendig for å oppfylle formålet

DE REGISTRERTES RETTIGHETER
Så lenge de registrerte kan identifiseres i datamaterialet vil de ha følgende rettigheter: innsyn (art. 15), retting (art. 16), sletting (art. 17), begrensning (art. 18), og dataportabilitet (art. 20).
Personverntjenester vurderer at informasjonen om behandlingen som de registrerte vil motta oppfyller lovens krav til form og innhold, jf. art. 12.1 og art. 13.

Vi minner om at hvis en registrert tar kontakt om sine rettigheter, har behandlingsansvarlig institusjon plikt til å svare innen en måned

FØLG DIN INSTITUSJONS RETNINGSLINJER
Personverntjenester legger til grunn at behandlingen oppfyller kravene i personvernforordningen om riktighet (art. 5.1 d), integritet og konfidensialitet (art. 5.1. f) og sikkerhet (art. 32).
Ved bruk av databehandler (spørreskjemaleverandør, skylagring eller videosamtale) må behandlingen oppfylle kravene til bruk av databehandler, jf. art 28 og 29. Bruk leverandører som din institusjon har avtale med.

For å forsikre dere om at kravene oppfylles, må dere følge interne retningslinjer og/eller rådføre dere med behandlingsansvarlig institusjon.

MELD VESENTLIGE ENDRINGER
Dersom det skjer vesentlige endringer i behandlingen av personopplysninger, kan det være nødvendig å melde dette til oss ved å oppdatere meldeskjemaet. Før du melder inn en endring, oppfordrer vi deg til å lese om hvilken type endringer det er nødvendig å melde: https://www.nsd.no/personverntjenester/fylle-ut-meldeskjema-for-personopplysninger/melde-endringer-i-meldeskjema

Du må vente på svar fra oss før endringen gjennomføres.

OPPFØLGING AV PROSJEKTET
Personverntjenester vil følge opp ved planlagt avslutning for å avklare om behandlingen av personopplysningene er avsluttet.

Lykke til med prosjektet!

8.2 Vedlegg - Samtykkeerklæring

NSD skjema

Vil du delta i forskningsprosjektet *"Bruk av syntetisk data for å tilrettelegge for innovasjon innen helseforskning"*?

Dette er et spørsmål til deg om å delta i et forskningsprosjekt hvor formålet er å *undersøke bruk av syntetisk data som innovasjon innen helseforskning*. I dette skrivet gir vi deg informasjon om målene for prosjektet og hva deltakelse vil innebære for deg.

Formål
Formålet med prosjektet er å undersøke hvordan syntetisk data som innovasjon kan påvirke helseforskning i Norge. Syntetisk data handler om å fjerne personlige opplysninger fra data, slik at syntetisk data kan benyttes i forskning på en mer effektiv og hensiktsmessig måte. Bruken av syntetisk data er relativt nytt i Norge og derfor et spennende felt å undersøke på.

Vi vil gjennomføre kvalitative undersøkelser for å undersøke forskningsspørsmålet, samt bruke relevante forskningsartikler, bøker, mm. Under kvalitative undersøkelser, vil vi gjennomføre dybdeintervjuer med relevante informanter tilknyttet temaet vi undersøker.

Problemstilling: «*Hvordan kan bruk av syntetiske data tilrettelegge for innovasjon innen helseforskning i Norge?*»

Forskningsspørsmål:

- På hvilken måte kan tilgjengeliggjøring av data effektivisere forskning?
- På hvilken måte kan effektivisering av forskning bidra til mer innovasjon?

Forskningsprosjektet er en masteroppgave som gjennomføres ved Høgskulen på Vestlandet, avdeling Kronstad, ved linjen innovasjon og ledelse.

Hvem er ansvarlig for forskningsprosjektet?
Høgskulen på Vestlandet er ansvarlig for prosjektet. Forskningsprosjektet utføres i samarbeid med prosjektet SYNDATA hos Helse Vest IKT som vil bistå med interessentliste.

Hvorfor får du spørsmål om å delta?
I dette forskningsprosjektet ønsker vi å intervjue mennesker med spesifikk kompetanse, erfaring eller innsikt innen syntetisk data i Norge. Utvalget vil derfor bestå av mennesker som besitter relevant informasjon for forskningsprosjektet. Vi ønsker å gjennomføre 12-15 intervjuer med aktuelle objekter.

Kontaktopplysninger av intervjuobjekter er innhentet gjennom egen research eller har blitt oppgitt gjennom samarbeid med SYNDATA.

Hva innebærer det for deg å delta?
Hvis du velger å delta i prosjektet, innebærer det at du deltar i et dybdeintervju. Det vil ta cirka 60 minutter. Intervjuguiden inneholder spørsmål om blant annet syntetisk data, innovasjon, dagens prosesser rundt data og forskning. Dine svar fra intervjuet vil bli lagret elektronisk ved passordbeskyttede notater.

Det er frivillig å delta
Det er frivillig å delta i prosjektet. Hvis du velger å delta, kan du når som helst trekke samtykket tilbake uten å oppgi noen grunn. Alle dine personopplysninger vil da bli slettet. Det vil ikke ha noen negative konsekvenser for deg hvis du ikke vil delta eller senere velger å trekke deg.

Ditt personvern – hvordan vi oppbevarer og bruker dine opplysninger
Vi vil kun bruke opplysningene om deg til formålene vi har fortalt om i dette skrivet. Vi behandler opplysningene konfidensielt og i samsvar med personvernregelverket.

- *De som vil ha tilgang til dine opplysninger er studentgruppen fra Høgskulen i Vestlandet som gjennomfører forskningsprosjektet.*
- *Vi vil erstatte navn og kontaktopplysningene dine med en kode som lagres på en egen navneliste adskilt fra øvrige data.*
- *Vi vil lagre datamaterialet på en passordbeskyttet mappe.*

- *Vi vil anonymisere deg og dine opplysninger, slik at du ikke vil gjenkjennes ved en eventuell publikasjon.*

Hva skjer med opplysningene dine når vi avslutter forskningsprosjektet?
Opplysningene anonymiseres frem til prosjektet avsluttes/oppgaven er godkjent, noe som etter planen er 01.07.2022. Etter denne tidsfristen vil vi slette all data og personopplysninger vi har om deg.

Dine rettigheter
Så lenge du kan identifiseres i datamaterialet, har du rett til:
- innsyn i hvilke personopplysninger som er registrert om deg, og å få utlevert en kopi av opplysningene,
- å få rettet personopplysninger om deg,
- å få slettet personopplysninger om deg, og
- å sende klage til Datatilsynet om behandlingen av dine personopplysninger.

Hva gir oss rett til å behandle personopplysninger om deg?
Vi behandler opplysninger om deg basert på ditt samtykke.

På oppdrag fra Høgskulen på Vestlandet har NSD – Norsk senter for forskningsdata AS vurdert at behandlingen av personopplysninger i dette prosjektet er i samsvar med personvernregelverket.

Hvor kan jeg finne ut mer?
Hvis du har spørsmål til studien, eller ønsker å benytte deg av dine rettigheter, ta kontakt med:

- *Høgskulen på Vestlandet* ved Kjersti Berg Danilova (veileder) ved kjersti.danilova@nhh.no, Live Haraldsen (student) ved 590629@stud.hvl.no, Karoline Fløysand Vollan (student) ved 594968@stud.hvl.no eller Iselin Pedersen Kristiansen (student) ved 081276@stud.hvl.no.
- Vårt personvernombud: Trine Anikken Larsen ved trine.anikken.larsen@hvl.no
- Hvis du har spørsmål knyttet til NSD sin vurdering av prosjektet, kan du ta kontakt med:
  NSD – Norsk senter for forskningsdata AS på epost (personverntjenester@nsd.no) eller på telefon: 55 58 21 17.

Med vennlig hilsen

Live Haraldsen, Karoline Fløysand Vollan og Iselin Pedersen Kristiansen
(Studenter)

Kjersti Berg Danilova
Prosjektansvarlig (Veileder)

Samtykkeerklæring

Jeg har mottatt og forstått informasjon om prosjektet *"Bruk av syntetisk data for å tilrettelegge for innovasjon innen helseforskning"*, og har fått anledning til å stille spørsmål. Jeg samtykker til:

- å delta i intervju.
- at innsamlet informasjon jeg har oppgitt vil kunne brukes inn i forskningsprosjektet.

Jeg samtykker til at mine opplysninger behandles frem til prosjektet er avsluttet.

_____

(Signert av prosjektdeltaker, dato)

8.3 Vedlegg – Intervjuguide til fagperson

*Spørsmål med * stilles hvis de jobber med data i hverdagen*

Tusen takk for at du ønsker å stille til intervju med oss, det setter vi stor pris på. Som nevnt tidligere, så skal vi skrive en masteroppgave om hvordan bruk av syntetisk data kan tilrettelegge for innovasjon innen helseforskning. Vår oppgave bygger på samarbeid med SYNDATA, som du også er en del av. Som du vet er jo dette et ganske uopptråkket område, så det er veldig spennende for oss å skrive master om dette. Informasjonen som blir gitt gjennom intervjuet vil bli anonymisert, og du kan trekke samtykke i etterkant dersom du ønsker det ved å sende oss en mail. Intervjuet vil vare i ca. 1 time.

Arbeidsgiver & stillingstittel:

Kan du beskrive din arbeidshverdag og dine arbeidsoppgaver?

Hva jobber du med for øyeblikket?

Hva anser du som de største utfordringene med innsamling og behandling av data?

Må du forholde deg til data i din arbeidshverdag?
- hvilken type data er det snakk om?

Hvordan ser du på forskning i relasjon til innovasjon?

Hvilke tanker har du rundt innovasjon og forskning?

Er du kjent med begrepet syntetisk data?

Hva vet du om syntetisk data?

Hva ser du på som barrierer for syntetisk data?
-Hva betyr dette i praksis? Kan du komme med et praktisk eksempel der GDPR har trenert eller påvirket en prosess.

-Hvordan forholdt dere til det?

Hvilke fordeler kan oppstå ved å ta i bruk syntetisk data?

Hva er dine tanker rundt bruk av syntetisk data i helseforskning?
-Hvilken rolle tror du syntetisk data kan ha i fremtiden innen helseforskning?

Hva er dine tanker om bruk av syntetisk data sett opp imot helseforskning?

Hvilken type data tror du utgjør majoriteten om 10 år?

Tror du bruk av syntetisk data kan påvirke innovasjon i helseforskning?
-Hvis "ja", beskriv hvordan og på hvilken måte?

*Hvis syntetisk data var tilgjengelig for deg, ville du benyttet det da?

- Hvordan ville dette påvirket måten du jobber på (prosessen) og eventuelt resultat?

Tror du at bruk av syntetisk data kan gi verdi for helseforskning?

- På hvilken måte?

Hvordan skal den syntetiske dataen forvaltes? Hvem skal ha eie andeler? Eventuelt hvordan skal disse fordeles?

Støttespørsmål intervjuguide

Hvilket forhold har du data: innsamling og håndtering?

*Hvilken tilnærming har du til innsamling av data?

- Bruker du ulike metoder for dette?

*Hva skjer med innsamlet data når et forskningsprosjekt er avsluttet?

*Hvordan er prosedyrene for gjenbruk av tidligere innsamlet data?

*Hva er viktig å tenke på i en datainnsamlingsprosess?

*Hvilke behov har du som når du skal samle inn data?

*Kan du beskrive en ideell prosess for innhenting av data?

*Hvis du kunne endret/fjernet et steg i prosessen ved dagens datainnsamling i forskning, hva ville det vært?

8.4 Vedlegg – Intervjuguide til forsker

Intervjuguide til forsker

Arbeidsgiver & stillingstittel:

Kan du beskrive din arbeidshverdag og dine arbeidsoppgaver?

Hva forsker du på for øyeblikket?

Hvilke formål har du med din forskning og hva ønsker du å oppnå?

Kan du ta oss igjennom en forskningsprosess og forløp?

Hvilken tilnærming har du til innsamling av data?
- Bruker du ulike metoder for dette?

Hva skjer med innsamlet data når forskningsprosjektet er avsluttet?

Hvordan er prosedyrene for gjenbruk av tidligere innsamlet data?

Hva mener du er barrierene med data for deg som forsker?

Hva er viktig å tenke på i en datainnsamlingsprosess?

Hvilke behov har du som forsker når du skal samle inn data?

Kan du beskrive en ideell prosess for innhenting av data?

Hvis du kunne endret/fjernet et steg i prosessen ved dagens datainnsamling, hva ville det vært?

Hvordan ser du på forskning i relasjon til innovasjon?

Hvilke tanker har du rundt innovasjon og forskning?

Er du kjent med begrepet syntetisk data?

Hva vet du om syntetisk data?

Hva er dine tanker om bruk av syntetisk data i forskning?

Hvis syntetisk data var tilgjengelig for deg, ville du benyttet det da?
- Hvordan ville dette påvirket måten du forsker på (prosessen) og eventuelt resultat?

Tror du at bruk av syntetisk data kan gi verdi for forskere?
- På hvilken måte?

Hvordan skal den syntetiske dataen forvaltes? Hvem skal ha eie andeler? Eventuelt hvordan skal disse fordeles?

8.5 Vedlegg – Utvalgt transkribert intervju

Utvalgt transkribert intervju

Arbeidsgiver & stillingstittel:
Bachelor i IT og master i chain og supply management. Jobber med styringsdata i ***. Jobbet med innovasjon innenfor teknologi, ser på nye trender innenfor helse. Hovedsakelig innenfor maskinlæring og innovasjonsprosjekt. Vi har ikke stillingsbeskrivelse, men jeg jobber som spesialkonsulent innenfor maskinlæring og data.

Kan du beskrive din arbeidshverdag og dine arbeidsoppgaver?
Teknisk sett, eller folkelig? Vi jobber i seksjon som er smidig. Vi bruker smidig metodikk for prioritere hvilke oppgaver vi skal utføre. Vi jobber slik at vi har lister prioritert, som prioriteres av produkteier. Vi kjører sprint på 2-3 uker hvor vi utfører oppgavene, avhengig av størrelsen på oppgavene. Vi har forskjellige områder, klinisk, HR, osv. Dette er siloer, vi er eksperter på ulike områder. På slutten av sprinten har vi en demo hvor vi viser resultater til produkteier, hvor vi tester osv. Kunden tester, også får vi tilbakemelding på det. Scrum, da jeg begynte var det ikke snakk om dette. Vi implementerte Scrum og ble inspirert av Statoil, nå er det litt mer utbredt.

Må du forholde deg til data i din arbeidshverdag?

- hvilke type data er det snakk om?

For min del er det helsedata, kliniske data. Vårt største datasystem er DIPS, det er aktiviteten (data) fra sykehuset, men vi bruker også fagsystemer. Vi beriker data fra systemer vi har. Men hovedkilden er NPA. Det er agnostisk og hvilken som produserer dem. Ut fra NPA får vi pasientaktivitet fra de fire sykehusene som er i ***.

Er dette anonymiserte eller maskert data?
Vi har ikke anonymisert datasettet, vi avidentifisert (vi har ikke fødselsnummer, men pasient ID fra DIPS). Men på grunn av GDPR måtte vi skille mellom sensitivt, aggregert og

anonymisert. Ordet anonymisert gir forvirring, så vi bruker av-anonymisert. Vi har et dataregister som er av-anonymisert og en som er delvis (gjerne aggregering). Det er ikke 100 %, man kan ikke bli fullstendig anonymisert.

Hva er din rolle i *?**

Grunnen til at jeg sa ja er flere ting, men det første som berører meg (arbeidsgiver) er personvern. For at vi benytter data som pasientdata som ligger hos oss så må det være en del som er på plass. Dataen skal være til et bestemt forhold. Sykehusene eier dataen, med syntetisk data så åpner man opp et stort område av innovasjon, fordi data blir tilgjengelig uten like mye hindringer. Reelle data som syntetiske data. Det kan brukes innenfor forskning og innovasjon. Forskning tar lang tid og komplisert slik som det er nå. Hvis vi får produsert syntetisk data, vil det åpne opp for mer forskning og tilgjengelighet. Internt og eksternt. Dette opplever vi hver gang vi går inn i et innovasjonsprosjekt, personvern. Det er tunge prosesser. Innovasjonsprosjekter som kunne vært utført på 3 måneder tar 3 år.
Helseanalyseplattformen ble stoppet på grunn av dette her. De ser nå på syntetiske data. Dersom de kan få til å lage syntetisk data, så blir det en Helseanalyseplattform. De prøver å gjøre dette. Hvis de får det til blir det stort. Vi prøver å få det til i vår lille skala med NTNU, men ideen er den samme.

Tror du bruk av syntetisk data kan påvirke innovasjon i helseforskning?

- **Hvis "ja", beskriv hvordan og på hvilken måte?**

Fra forskning til produkt blir veien mye kortere. Du kan forske så mye du kan og publisere det. Syntetisk data kan du gå videre og lage et produkt. Muligheter for å implementere produkter. Du fjerner delen med personvern og det er en omfattende del av helsedata.

Tror du at bruk av syntetisk data kan gi verdi for helseforskning?

- **På hvilken måte?**

Det er en stor mengde data hos oss som ikke brukes til forskning og innovasjon. Syntetisk data kan gjøre dette tilgjengelig. Men mange andre. Helsedata er inkomplett og med masse bias i det. I syntetiske data kan vi benytte kunnskap og genere nye datasett eller fjerne bias. Vår helsedata er veldig ubalansert. Eks antall pasienter som ikke møter til sine kliniske timer, mange frafall. Det er få i forhold til populasjon. Veldig lite av populasjon ønsker å bli studert og få har den egenskapen du ønsker å forske. I syntetisk data kan du genere datasett som gjør at du kan jobbe med bias litt mer aktivt. Noen ganger har vi ikke scenarier i et reelt datasett,

men det kan man generere i datasettet med syntetisk data (scenarioer som er i klinikk, men som ikke kommer frem). Kvaliteten av dataene er helt avhengig av hvem som produserer den. i syntetiske datasett data kan man produsere data og øke kvaliteten, spesielle scenarier. Det er ikke lov med reelle data. I en syntetisk verden er det lovlig.

Hvis man får en gravid mann så får man en data som ikke er realistiske på det datasettet. Det må være etablerte prosesser, for det kan bli misbrukt. Du må registrere det på en viss måte for at det skal være kvalitet. Største problemet vi har nå er at leger og sykepleiere skriver feil koder når de registrerer diagnoser eller behandlinger. Det samme vil det være i en syntetisk verden. Det krever en modellering. Men man har kontroll på dette, sannsynligheten for at man lager noe som ikke gir mening er liten. Dette er maskinstyrt. Når det er menneske styrt så blir sannsynligheten større. Det viktigste er at man modellerer riktig.

Om du generer syntetisk data fra reelle data og man allerede har problemer med reelle data kan det overføres til syntetisk data. MEN her kan man kunstig kompensere for dette. Antall pasienter som blir behandlet med en viss diagnose er ikke representativ på grunn av bruk av feil kode. Her kan man aktivt ta klinisk kunnskap og generer det man mener er mer realistisk. Både dette med bias og balansere data. Man har full fleksibilitet i den syntetisk verden, ulikt fra reelle data.

Utfordringer med syntetisk data, hvilke barrierer?
Den største utfordringen er tillit til dataen i klinisk miljø. I det kliniske miljøet, vil de bruke syntetisk data? Vil de bruke dette? Dette er helt ukjent for dem. De er vant til å bruke reelle data. Det krever arbeid i miljøene for å bygge tillit til dataene. Organisatorisk arbeid som kreves. Det finnes nok av utfordringer. Tekniske utfordringer kan man løse, men største utfordringen som vil kreve mer er både prosesser, organisatoriske, og menneskene som bruker dataene. Det krever opplæring til å benytte den type data. Prosesser innenfor sykehusene.

Vil det være lettere å validere forskningsdata? For eksterne?
Ja, definitivt mye lettere å validere forskningsdata, fordi syntetisk data er 100% anonymisert, da er det fritt tilgjengelig for alle innenfor helse. Jeg snakker ikke om det kommersielle. Det er fremdeles sykehusene som eier dataene i grunnen. Snakker om fritt - da snakker jeg innenfor helse og det offentlige. Det er private aktører som vil benytte denne type data til

innovasjon og utvikle nye produkter. Men i USA er det samarbeid mellom privat og offentlig om dette.

Hvordan fungerer det dersom det er Volvat som har avlastet Haukeland?
Fortsatt Haukeland som eier dataene. De får tilgang til å løse oppgaven, men sekundær dataen får de ikke. Det er Bergen som eier. Du som eier dine helsedata, men sykehusene kan bruke de som primær bruk for å behandle deg og som sekundærbruk for å forske. Det er forskjell mellom primær og sekundær bruk. Syntetisk data er sekundær bruk. Man skal ikke bruke dette for å behandle pasienter. Man kan lage en algoritme som kan brukes for å behandle i teorien, men det er kun et scenario, jeg har ikke hørt om noen som har gjort det. Hva har innovasjon med dette å gjøre? Det kommer sjeldent produkter ut fra helseforskning. Problemet med prosjektet med NTNU er personvern.

8.6 Stikkordregister

Analytisk intelligens: *"Analytisk intelligens er evnen til å behandle informasjon for problemløsning og lære av den"* (Sternberg 2005 referert i Huang & Rust, 2018, s.158).

Data: *"Data er informasjon, spesielt fakta eller tall, samlet inn for å bli undersøkt og vurdert og brukt for å hjelpe beslutningstaking, eller informasjon i elektronisk form som kan lagres og brukes av en datamaskin"* (Cambridge Dictionary, u.å).

Datadrevet innovasjon: *"Datadrevet innovasjon er bruk av data eller analyser for å forbedre eller fremme nye produkter, prosesser, organisasjonsmetoder eller markeder*" (OECD, 2015, s. 21).

Datakvalitet: Datakvalitet karakteriseres som immaterielle egenskaper, som fullstendighet og konsistens (Veregin, 1999).

Digital innovasjon: *"Digital innovasjon er gjennomføring av nye kombinasjoner av digitale og fysiske komponenter for å produsere nye produkter"* (Yoo et al. 2010, s. 725).

Digital innovasjon som prosess: "*et nytt produkt eller tjeneste som skaper ny verdi for adoptanter, utviklet ved å kombinere digital teknologi på nye måter eller med fysiske komponenter* (Osmundsen et al., 2018, s. 7)

Digital innovasjon som resultat: *"å kombinere digital teknologi på nye måter eller med fysiske produkter, for å utvikle et nytt produkt eller tjeneste som skaper ny verdi for adoptanter"* (Osmundsen et al., 2018, s. 7)

Digital kompetanse: *"Digital kompetanse innebærer trygg og kritisk bruk av digitale verktøy og medier til arbeid, fritid og kommunikasjon. Den er underbygget av grunnleggende ferdigheter innen IKT: bruk av datamaskiner for å hente, vurdere, lagre, produsere, presentere og utveksle informasjon, og å kommunisere og delta i samarbeidsverktøy via internett"* (EU-kommisjonen, 2006, s.15-16).

Empatisk intelligens: "*Empatisk intelligens er evnen til å gjenkjenne og forstå andre mennesker følelser, reagere passende følelsesmessig og påvirke andres følelser*" (Goleman 1996, referert i Huang & Rust, 2018, s.159).

GDPR: General Data Protection Regulation som skal gi enkeltpersoner mulighet til å få oversikt og kontroll på hvilke data både offentlige og private aktører har (Sørebø et al., 2020).

Hendelsesdata: *"Hendelsesdata er hendelser som inntreffer enten i en virksomhet eller innbyggers liv, eller som følge av at data endres. Basert på hendelser vil man ha behov for tjenester"* Digdir (u.å)

Hybrid intelligens: Hybrid intelligens søker å bruke komplimenterende styrker ved menneskelig og maskinell intelligens for å utvide det menneskelige intellekt. En slik intelligens skjer gjennom å utføre komplekse oppgaver hvor en oppnår bedre resultater som hver av intelligensene kunne gjort alene (Dellerman et al., 2019, s. 276).

Innovasjon: *"En innovasjon er implementering av et nytt eller vesentlig forbedret produkt (vare eller tjeneste), eller prosess, en ny markedsføringsmetode, eller en ny organisasjonsmetode i forretningspraksis, arbeidsorganisasjon eller eksterne relasjoner"* (OECD & Eurostat, 2005, s. 46)

Innovasjonprosess: Veien fra en idé oppstår til et ferdig produkt er utviklet (Aasen & Amundsen, 2011).

Intuitiv intelligens: "*Intuitiv intelligens er evnen til å tenke kreativt og effektivt tilpasse seg nye situasjoner*" (Sternberg 2005, referert i Huang & Rust, 2018, s.159).

Kunstig intelligens: "*Systemer som utfører handlinger, fysisk eller digitalt, basert på tolkning og behandling av strukturerte eller ustrukturerte data, i den hensikt å oppnå et gitt mål"* (EU-kommisjonen, 2018, s.1).

Maskinlæring: Maskinlæring kan omtales som en underkategori av KI, som omfatter alle funksjoner som gjør det mulig for maskiner å lære fra data uten å være spesifikt programmert til det (Jakhar, 2020, s. 131).

Mekanisk intelligens: "*Mekanisk intelligens er evnen til å automatisk utføre rutinemessige, gjentatte oppgaver*" (Huang & Rust, 2018, s.158).

Personopplysninger: Personopplysninger som alle opplysninger og vurderinger som kan knyttes til deg som enkeltperson (Datatilsynet, 2019).

Personvern: Personvern kan forklares som retten til å ha kontroll over egne personopplysninger, vite hva som er lagret av slike opplysninger, og hvordan opplysningene brukes (Datatilsynet, 2019).

Prosessinnovasjon: *"Prosessinnovasjon er implementering av en ny eller betydelig forbedret produksjon eller leveringsmetode"* (OECD & Eurostat, 2005, s. 49).

Schrems II-dommen: Her avgjorde EU-domstolen at dersom personopplysninger skal overføres til land utenfor EU/EØS, må det være et overføringsgrunnlag i henhold til personopplysningsloven (Digdir, u.å.).

Stordata: *"Stordata er store mengder omfattende variert data som genereres, fanges opp og behandles med høy hastighet"* (Laney 2001 referert i Günther et al., 2017, s. 191).

Stordataanalyse: *"Stordataanalyse kan forklares som en analyse som håndterer store volum med data fra flere ulike datakilder, som har til hensikt å oppdage nye sammenhenger og innsikter"* (Lanestedt, 2016).

Syntetisk data: *"Syntetisk data er data som etterligner de originale observerte dataene og bevarer relasjonene mellom variabler, men inneholder ingen avslørende informasjon*" (Nowok et al., 2016, s. 1).

Tillit: *"Tillit er villigheten til å ta risiko"* (Mayer et al., 1995, s. 712).

Uteligger: "*Uteligger er en observasjon (eller undergruppe av observasjoner) som ser ut til å være inkonsistent med resten av datasettet"* (Barnett & Lewis, 1994, referert i Wilkinson, 2018, s. 256)

Vare- og tjenesteinnovasjon: "*Vare- og tjenesteinnovasjon er* en *vare eller tjeneste som er ny eller vesentlig forbedret med hensyn til dens egenskaper eller tiltenkte bruksområder*" (OECD & Eurostat, 2005, s. 48).