Rapport
Syntetisk data inom intensivvård
AI Sweden & Region Västerbotten
Fas 1

Till denna rapport hör följande bilagor:

2_Rapport från Syndata AB (på svenska)
    Attachment 1a: Recovered dataset evaluation report (in English)
    Attachment 1b: Discharged dataset evaluation report (in English)
Attachment 2: Results from AI Sweden (in English)
Attachment 3: Results from RVB Region Västerbotten (in English)

Table of Contents

Bakgrund

En ständigt återkommande utmaning i arbetet med AI-modeller inom sjukvården är behovet av mer data. Varje region kommer ha problem med för liten mängd data eftersom de bara har sin egen regions data att tillgå. Att kunna dela data mellan regioner så att man kan träna sina AI-modeller på större datamängder skulle därför kunna bidra positivt till AI-utvecklingen inom sjukvården. Eftersom det rör sig om data av mycket känslig karaktär är dock sådan datadelning svår att göra. En möjlig lösning skulle vara att arbeta med syntetisk data som inte innehåller data från verkliga patienter och därför troligen kan delas mellan regioner.

Region Västerbotten arbetar med ca 2000 patienter i de AI-modeller som nu tas fram, men för att få ännu bättre modeller skulle den mängden behöva vara betydligt större. Källsystemet från vilket data hämtas heter MetaVision, vilket är ett kliniskt informationssystem som används för elektronisk journalföring och som beslutsstöd inom perioperativ, neonatal- och intensivvård inom Region Västerbotten. Systemet samlar in relevant data från medicinteknisk utrustning och ger tillsammans med sjukvårdspersonalens utlåtande och integrationer till andra system en samlad bild av patientens anestesi och intensivvårdsförlopp. Detaljrikedomen per patient gör att det finns väldigt många parametrar kopplat till varje patient.

Det finns inte mycket information om hur utfallet blir om man använder syntetisk data i AI-modeller, så det skulle vara värdefullt att utvärdera. En samverkan mellan regioner behövs för att få tillgång till mer data och då skulle delning av syntetiserade data vara en tänkbar lösning. Det är då viktigt att utvärdera både utfall av AI-modellens träffsäkerhet och det juridiska kring att dela syntetisk data mellan regioner i sjukvården. Man bör även utreda säkerhetsaspekter som exempelvis möjlighet till återidentifikation av patienter.

Region Västerbotten har ett pågående AI-projekt: AMHOS – AI/ML i Hälso- och sjukvård som är delvis finansierat av EU-medel genom Tillväxtverket. I projektet har det tagits fram två AI-baserade modeller för prediktion av vårdtid på intensivvården:

- Prediktion av tid till utskrivning pga tillfrisknande
- Prediktion av tid till utskrivning (alla orsaker)

Dessa två prediktionsmodeller tränas med två separata dataset, vilka vi i denna rapport refererar till som dataset:

- “recovered” (tid till utskrivning pga tillfrisknande)
- “discharged” (tid till utskrivning pga alla orsaker)

Det är även påbörjat ett arbete med prediktioner inom ett nytt område: prediktion av vårdtid på post-op. Därefter planeras prediktion av hypotension vid operation och prediktion av smärta på post-op.

Mål och avgränsning

AI Sweden bidrar till att accelerera användningen av AI i Sverige och vill stötta en nationell användning av syntetiska data för utveckling och träning av verksamhetsnyttiga AI-modeller inom sjukvården.

Projektets Fas 1 har fokus på praktiskt användning av syntetisk data i AI-modeller för att närmare utforska hur användbar syntetisk data är när man tar fram AI-modeller inom sjukvård. Baseret på den vårddata som Region Västerbotten har samlat in genom arbetet med att ta fram prediktionsmodeller för prediktion av vårdtid på intensivvården, så utreder vi hur väl syntetisk data - som skapas med olika metoder - förhåller sig till originaldata och hur träffsäkerheten för prediktionsmodeller tränade på syntetisk data skiljer sig åt jämfört med modeller tränade på originaldata.

Projektets Fas 1 är avgränsat till att utreda förutsättningar och möjligheter kring användning av syntetisk data inom sjukvården. Utvärderingen av skapad syntetisk data är avgränsad till kvantitativ jämförelse mellan genererad syntetisk data och originaldata, samt en utvärdering av hur syntetisk data förhåller sig till originaldata vid träning och validering av befintliga prediktionsmodeller hos Region Västerbotten. Inom ramarna för projektets Fas 1 så kommer dock inte alternativa modeller för prediktion av vårdtid att utvärderas.

Konstellation

Region Västerbotten (RVB)
Per Ericson, Specialist Digitalisering.
Projektledning, infrastruktur, utvärdering och resultatspridning.

Robert Wiksten, Förvaltningsledare teknik Vårdstöd Special.
Infrastruktur och utvärdering.

Sara Lundsten, Förvaltningsledare verksamhet, anestesi-SSK.
Återkoppling på utfall från kliniskt perspektiv.

Frans Vincent, klinisk ledare VSS, narkosläkare.
Återkoppling på utfall från kliniskt perspektiv.

Petter Lindgren, Senior Data Scientist (konsult, Sogeti).
Testa AI-modeller med olika former av syntetisk data.

Mattias Andersson, Data Engineer (konsult, Sogeti).
Datauttag och tester av modellerna.

Syndata AB
Mattias Ripoll, CEO, Project Manager
Roxana Buzatoiu, Data Scientist
Keshav Padiyar, Data Scientist

Douglas Garcia, Data Scientist
Guillermo Padres, Tech Lead

AI Sweden / Örebro Universitet (ÖU)
Andreas Persson, Data Scientist / Forskare

AI Sweden
Henrik Ahlén, AI Change Agent Healthcare
Övergripande projektledning

Förutsättningar och möjligheter med syntetisk data

Syntetisk data är en "konstgjord" version av riktig data som efterliknar originaldatan vad gäller egenskaper såsom statistiska relationer och korrelationer, men som saknar kopplingar till identifierbara individer från originaldatan. Genom introduktionen av generativa maskininlärningsmetoder[1] så har det även introducerats oändliga möjligheter till att använda maskininlärning till att skapa verklighetstrogen syntetisk data. Inom sjukvården så har detta även skapat förutsättningar för att generera och dela syntetisk vårddata som saknar kopplingar till identifierbara individuella patienter.

En av de främsta anledningen till att använda syntetisk data är att det skapar förutsättningar för att kunna dela data. När det kommer till medicinska tillämpningar så skapar syntetisk data dessutom nya förutsättningar för att använda maskininlärning och träna AI-modeller, t.ex. för att predikterar vårdtyngd. Genom delad syntetisk data så kan man utöka antalet observationer i träningsdatan, vilket leder till bättre och mer robusta AI-modeller. Det finns därför stor potential för en rad olika typer (modaliteter) av syntetisk data inom medicinska tillämpningar, exempelvis: *tabulärdata* bestående av formulärdata, texter, medicinska/tekniska mätvärden, etc.; *bilddata* såsom röntenbilder, foton av hudsjukdomar, etc.; *ljuddata* såsom ultraljudinspelningar, etc.

Inom detta projekt har vi valt att fokusera på syntetisering av *tabulärdata* då det är denna typ av data som används hos Region Västerbotten för att träna de prediktionsmodeller som används för prediktion av vårdtid på intensivvården

Originaldata

Den data som arbetats med i projektet kommer från Region Västerbottens PDMS (patient Data Management System) som används på intensivvårdsavdelningen i Umeå. Det innehåller data från 2017 till 2021 och omfattar ca 1700 observationer. Det görs detaljerad journalföring och systemet har koppling till medicinteknisk utrustning vilket gör att det finns många värden per observation. Det kan vara upp till 15 000 olika variabler, men det är sällan som samtliga används. I processen att skapa prediktionsmodeller utifrån datan så har de 100 viktigaste variablerna valts ut. I den processen är data med direkt personkoppling inte med, dock ålder och kön eftersom det är relevanta variabler för prediktionsmodellen. Den typiska patienten är en man i åldern 60-70 år, det finns dock patienter i alla åldrar ända ner till ettåriga barn. En stor del av patienterna har andningssvårigheter eller allvarliga infektioner.

[1] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).

Det finns tre huvudsakliga kategorier av data:

- Formulärdata, som är inskrivet av en människa (sjuksköterska eller läkare) vid specifika tider, typiskt sett en inmatning per arbetsskiftet.
- Medicinsk/teknisk data, som kommer från medicinteknisk utrustning, som ventilatorer, patientmonitoreringsutrustning, etc. MT-utrustningen ger i regel ett värde per minut till PDMS.
- Numerisk data, som exempelvis värden från ventilatorn (t ex syresaturation) men också beräknade värden, t ex hur många antibiotiska läkemedel som är beställda för en patient.

Server och verktyg för åtkomst av avidentifierad originaldata

Det var viktigt att inte datan skulle lämna Region Västerbotten eftersom man då skulle frångå en av grundtankarna med att använda syntetisk data. För avidentifiering så togs uppgifter med direkt koppling till person bort, såsom personnummer, namn och adress. För att komma åt den avidentifierade datan användes VPN-uppkopplingar för åtkomst mot en server inom Region Västerbotten. I samband med godkännande av VPN-konto krävs att man tecknar sekretess- och ansvarsförbindelse.

Det gjordes installation av alla verktyg som behövdes på servern för att kunna göra generera syntetisk data. I detta fall handlade det om Python, Git och Jupyter Notebook. För att flera användare samtidigt ska kunna jobba mot samma server så behöver även RDS (Remote Desktop Services) installeras på servern. Eftersom användarna satt på olika plattformar så behövdes olika lösningar för anslutningen. Användare på Windows använde fjärrskrivbordet medan de på Linux använde Remmina.

Metoder för att skapa syntetiska data

När det kommer till generativa maskininlärningsmetoder för att allmänt skapa syntetisk data så är det företrädesvis två metoder som används: GAN (Generative Adversarial Network)[2], samt VAE (Variational Auto Encoder)[3]. Båda dessa metoder har sedermera anpassats specifikt för att generera syntetisk tabulärdata:

- CTGAN (Conditional Tabular GAN): en tävlingsinriktad modell bestående av två nätverk - en *generator* och en *diskriminator* - där generatorn tränas för att generera trovärdig data som är så pass lik originaldatan som möjligt, medans diskriminatorn tränas till att försöka urskilja originaldatan från data skapad av generatorn.
- TVAE (Tabular VAE): en tvåstegsmodell bestående av en *kodare* och en *avkodare.* Kodaren tränas till att transformera originaldatan till en lågdimensionell latent representation, medans avkodaren omvänt tränas till att avkoda och transformerar

[2] Kingma, Diederik P., and Welling, Max. “Auto-encoding variational bayes.” *International Conference on Learning Representations* (2013).

[3] Xu, Lei , et al. “Modeling Tabular data using Conditional GAN.” *Advances in Neural Information Processing Systems (NeurIPS)* (2019).

tillbaka den lågdimensionella representation till data som återspeglar originaldatan.

Anledningen till dessa anpassningar, specifikt för tabulärdata, är att just tabulärdata ofta består av en blandning av olika typer av värden (vilket även beskrivs ovan under rubrik **Originaldata**). Både CTGAN och TVAE modeller tränas utifrån de konditionala förutsättningarna att träningsdata kan bestå av attribut med olika värden vilket delas in i två grupper av attribut:

- Kategoriska attribut: består av diskreta värden, t.ex. formulärdata, journaldata.
- Numeriska attribut: består av kontinuerliga värden, både heltal och flyttal.

Alternativt till att dela in attribut i grupper om antingen kategoriska eller numeriska attribut, så kan en kopula-fördelningsfunktion[4] användas för att istället skapa en enhetlig kumulativ hyperdimensionell fördelning av samtliga attributvärden. Implementationer av både CTGAN och TVAE, såväl som stöd för modeller baserade på kopula-fördelningsfunktion, finns öppet tillgängliga genom SDV (Synthetic Data Vault)[5]. SDV inkluderar även olika verktyg och metoder för att utvärdera genererad syntetisk data (jämfört med originaldata). För att statistiskt uppskatta ett mått på kvaliteten av den syntetiserade datan så används främst två stycken fördelningstester:

- CS Test: använder Chi-Squared-fördelningstest för att uppskatta fördelningen mellan två attribut bestående av diskreta värden.
- KS Test: använder Kolmogorov-Smirnov-fördelningstest för att uppskatta fördelningen mellan två attribut bestående av kontinuerliga värden.

Dessa tester utförs i regel per attribut genom att jämföra varje attribut från originaldatan med motsvarande attribut från den syntetiserade datan. Det slutliga kombinerade måttet aggregeras sedan som ett medelvärde av alla individuella tester för alla kategoriska respektive numeriska attribut.

För resultat av denna studie angående mått på kvaliteten, se Bilaga 2, Tabell 1.

För- och efterbehandling av originaldata / syntetiserad data

Även ifall generativa modeller, specifikt anpassade för syntetisering av tabulärdata (t.ex. CTGAN), generellt kan hantera att träningsdata kan bestå av olika datatyper, så krävs ändå en viss form av förbehandling av originaldata, samt en viss form av efterbehandling av den genererade syntetiska datan. För den originaldata som användes för denna studie så förbehandlas originaldata enligt följande:

1. Ta bort alla redundanta attribut, t.ex. alla attribut som är fullständigt funktionellt beroende av ett annat attribut.
2. Säkerställ att alla värden för ett attribut är av samma datatyp. Detta inkluderar även att hantera värden som är N/A (Not Available), vilka bör inkluderas, men som kan

[4] Schmidt, Thorsten. "Coping with copulas." *Copulas-From theory to application in finance* 3 (2007): 34.

[5] https://sdv.dev/

behöva omvandlas, t.ex. för kontinuerliga värden omvandlas till ett numeriska motsvarande NaN (Not a Number).

3. Identifiera vilka attribut som räknas som kategoriska respektive numeriska attribut.

Motsvarande så efterbehandlades även den syntetiserade datan enligt följande:

1. Avrunda alla numeriska attribut som förväntas vara av typen heltal.
2. Återskapa redundanta attribut.
3. Säkerställ att alla attributvärden ligger inom gränserna för vad som är rimligt, t.ex. säkerställ att attributvärdena ligger inom extremgränserna för originaldatan.

Patientintegritet utifrån syntetisk data

Ett av huvudsyftena med att använda generativa maskininlärningsmetoder, för att syntetisera data, är att anonymisera originaldatan. Genererad syntetisk data måste dock utvärderas för att säkerställa att den syntetiserade datan är tillräckligt olik originaldatan så att det inte går att återidentifiera en enskild individ utifrån den syntetiserade datan.

SDV inkluderar även metoder för att uppskatta ett mått på integritet (privacy) genom att estimera sannolikheten att enskilda individer, från originaldatan, ska kunna identifieras baserat på den syntetiserade datan. Omvänt så bygger dessa metoder på att en modell tränas med den syntetiserade datan och med avseende på nyckelattribut (dvs. attribut som kan tänkas kunna identifiera en individ), samt utifrån vissa attribut som räknas som känsliga attribut (dvs. attribut som kan tänkas kunna härleda nyckelattribut). Originaldata används sedan för att prediktera och jämföra utfallen med motsvarande nyckelattribut från originaldata. Detta innebär att för optimal integritet så ska inga predikterade utfall överensstämma med motsvarande värden för nyckelattribut från originaldata. Likt statistiska mått på kvalitet så måste även olika mått på integritet uppskattas utifrån kategoriska respektive numeriska attribut. För denna studie så har vi dock enbart utgått utifrån kategoriska attribut och uppskattat mått på integritet utifrån:

- Kategorisk generaliserad CAP (correct attribution probability): utgår ifrån att individ-par med minsta Hamming-avstånd för motsvarande känsliga attribut från den syntetiserade datan och originaldatan identifieras. Måttet på integritet ackumuleras sedan som sannolikheten att motsvarande nyckelattribut för varje identifierat individ-par inte stämmer överens.

- Kategorisk RF (random forest): bygger på att en RF klassificerare tränas baserat på den syntetiserade datan, samt med avseende på identifierade nyckelattribut. Denna klassificerare används sedan för att uppskatta ett mått på integritet genom den ackumulerade sannolikheten för felprediktion vid validering utifrån originaldatan.

När det handlar om tabulärdata så är det även korrelationen mellan attributen som avgör integriteten. Målet med att träna en generativ modell för att syntetisera data är att försöka lära modellen att bevara dessa korrelationer mellan attributen. Ett problem är dock att en modell kan lära sig att identifierar dessa korrelationsmöster för väl så att den individuella integritet äventyras. För att uppnå en högre nivå av personlig integritet så kan därför generativa modeller tränas med metoder för att förstärka den personliga integriteten.

Genom biblioteket SmartNoise (som utgör en del av Open Differential Privacy)[6], så finns det stöd för att träna två varianter av CTGAN modeller som dessutom tillämpar metoder[7,8,] för att förstärka den personliga integriteten (och som vi även har utvärderat i denna studie):

- DP-CTGAN (Differential Privacy): denna metod bygger på att integriteten förstärks genom att en viss form av statistiskt brus appliceras på modellens gradienter under träningsprocessen. Detta brus gör så att datan från en individuell patient gör mindre inverkan på den tränade modellens parametrar.
- PATE-CTGAN (Private Aggregation of Teacher Ensembles): denna metod delar upp den totala träningsdatan i delmängder och tränar separata modeller (sk. teachers) för varje delmängd. Den personliga integriteten skyddas sedermera genom att ett predikterat utfall aggregeras som majoriteten av utfallen av denna ensemble av teachers.

För resultaten av denna studie angående mått på patientintegritet, se Bilaga 2.

Syntetisk data vs originaldata

Inom ramarna för detta projekts Fas 1 så har totalt sju stycken generativa modeller tränats och använts. Av dessa sju modeller är fyra stycken öppna och allmänt tillgängliga modeller, dvs. de två vedertagna modellerna CTGAN och TVAE (vilka beskrivs under rubrik **Metoder för att skapa syntetiska data**), samt de två varianter av CTGAN som dessutom tillämpar metoder för att förstärka den personliga integriteten (vilka beskrivs under föregående rubrik **Patientintegritet utifrån syntetisk data**). Utöver de öppna modellera så har även Syndata utvecklat tre stycken modeller (Modell 1-3), där Modell 1 och 3 är baserade på CTGAN, medans Modell 2 är en modell baserat på en Gaussisk kopula-fördelningsfunktion.

För en detaljerad rapport angående modellerna som Syndata har tagit fram, se Bilaga 1.

För var och en av de syntetiska datamängderna tränades motsvarande modeller för att prediktera vårdtid. Dessa modeller tränades med samma parametrar som för modellerna tränade utifrån originaldata. På det sättet är indata det enda som skiljer träningen av modeller baserade på syntetiska data med träningen av modeller baserade på originaldata. Utvärdering gjordes sedan utifrån dessa kriterier:

- Error rate - hur bra rangordnar modellen patienter utifrån tid kvar till utskrivning?
- Utskrivningshastighet - har modellen liknande generella utskrivningshastighet som motsvarar verkligheten?
- Robusthet - ger modellen liknande resultat för olika dataset genererade från samma generativa modell?

---

[6] https://opendp.org/

[7] Abadi, Martin, et al. "Deep Learning with Differential Privacy." *Proc. of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (2016).

[8] Papernot, Nicolas, et al. “Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data.” International Conference on Learning Representations (ICLR) (2017).

- Fördelning av responsvariabler - har responsvariablerna liknande fördelning som originaldatan?

De flesta prediktionsmodeller tränade på Syndatas syntetiska dataset gav låg error rate, nästan i paritet med originalmodellerna. Utskrivningshastigheten, däremot, predikterades ofta till att vara snabbare än i verkligheten. För att åtgärda detta så tog Syndata fram en förbättrad modell (Modell 3), där fördelningen för responsvariablerna optimerades. Utskrivningshastigheten blev då betydligt bättre, men modellens error rate ökade något.

De testade öppna modellerna gav likvärdiga resultat när det gäller utskrivningshastighet och fördelning av responsvariabler som Syndatas modeller, men hade betydligt högre error rate. Robustheten var generellt hög för alla modeller.

Generellt sett ser det lovande ut att kunna använda syntetiskt medicinsk data för att träna prediktionsmodeller för att prediktera vårdtid. Speciellt imponerande var det att Syndatas modeller hade så pass låg error rate.

För en mer detaljerad rapport på valideringen, se Attachment 3.

Sammanfattning av insikter från Fas 1

Här sammanfattar vi våra insikter från projektets Fas 1. Insikter som kan vara nyttiga för andra aktörer som sitter i inledningsfasen av liknande projekt. I efterföljande stycke så presenterar vi en diskussion och frågeställning kring fortsatt arbete under Fas 2 av detta projekt. Från projektets Fas 1 så kan vi summera våra insikter enligt följande:

- **Åtkomst av originaldata**
  Genom att använda VPN så går det att ge åtkomst till känslig data utan att datan lämnar servern där den är lagrad. I samband med godkännande av VPN-konton fick utvecklarna även godkänna särskild ansvars- och sekretessförbindelse.

  Det kan dock vara en utdragen process - både juridiskt och tekniskt - att ge åtkomst till alla aktörer som är inblandade i ett projekt. Det är därför viktigt att arbetet med åtkomst initieras redan från början av ett projekt av denna typ.

  När flera aktörer använder samma data och samma server så kan det dessutom uppstå nya tekniska problem, exempelvis att servern blir överbelastad. Det kan därför vara bra att även ta höjd för eventuella nya tekniska utmaningar.

- **Analys av originaldata**
  Det är viktigt att förklara datasetet och dess tillhörande attribut tydligt. Generativa modeller för att skapa syntetisk tabulärdata förutsätter (generellt) att träningsdatan delas in i kategoriska respektive numeriska attribut (vilket ytterligare beskrivs under stycke **Metoder för att skapa syntetiska data**). Det är därför viktigt att kommunicera vilka attribut som en den generativ modell ska användas som numeriska attribut och vilka som ska användas som kategoriska attribut.

- **Förbehandling av träningsdata**
  Det underlättar både för- och efterbehandling (såväl som träning av generativa modeller), ifall tabulärdata uppfyller normalformerna enligt teorin bakom relationsdatabaser[9]. Den originaldata som användes för denna studie bestod av vissa redundanta attribut som var fullständigt funktionellt beroende av ett annat attribut, dvs. värdena för ett attribut kan helt härledas baserat på värdena av ett annat attribut. För denna studie så togs helt enkelt fullständigt funktionellt beroende attribut bort genom förbehandling av träningsdatan och återskapades sedan genom efterbehandling av den syntetiserade datan.

- **Träning av generativa modeller**
  Traditionella förlustfunktioner är inte alltid det bästa måttet på inlärning vid träning av generativa modeller. Eftersom GAN modeller är uppbyggda av två nätverk som tävlar mot varandra så kommer måttet på förlust att fluktuera i takt med att det ena eller det andra nätverket blir bättre genom inlärning. För att träna en optimal GAN modell (samt att undvika överträning) så syntetiserade vi ett valideringsdataset efter varje träningsepok och använde istället kvalitetsmått på den syntetiserad valideringsdata (Bilaga 2, Tabell 1) för att styra träningsprocessen.

- **Validering av syntetiserad data**
  Det är viktigt att samma teckenkodning används för alla dataset - speciellt ifall det finns attribut bestående av textsträngar med specialtecken, t.ex. svenska å, ä och ö. Måtten på fördelning för diskreta attribut gav initialt väldigt låga resultat. Detta visade sig enbart vara ett resultat av att olika teckenkodning används för originaldata och den syntetiserade datan. För att kringgå detta problem så omvandlas alla teckensträngar till rent ASCII-format innan datan validerades.

- **Efterbehandling av syntetiserad data**
  Det kan vara fördelaktigt att efterbehandlat den syntetiserade datan och kontrollera en form av “rimlighet” när det gäller värdena för den syntetiserade datan. För numeriska attribut så visade det sig snabbt att den syntetiserade datan kunde innehålla värden som ligger utanför vad som kan tänkas vara rimligt för vissa attribut, t.ex. negativa värden för attribut som egentligen bara kan bestå av positiva värden (så som en individs ålder). För att kringgå detta problem så filtrerade vi de syntetiserade dataseten och tog bort alla individer med ett attributvärde utanför extremgränserna för motsvarande attribut från originaldatan. Resultatet av denna filtrering blev dock att den syntetiserade datan inte innehöll några individer som sticker ut ifrån mängden av individer från originaldatan, dvs. det sytetiserades inga etremindivder (outliers), vilka kan vara av intresse när det handlar om medicinsk data. Utan vidare inblandning av medicinsk expertis så är det dock svårt att utröna vad som kan räknas som rimligt eller ej för respektive attribut. En möjlighet för fortsatt arbete skulle därför kunna vara att låta medicinsk expertis granska den syntetiserade datan och avgöra ifall de genererade värdena verkligen är rimliga.

[9] Padron-McCarthy, Thomas, and Risch, Tore. *Databasteknik*. Studentlitteratur, 2005.

- **Träning och validering av befintliga prediktionsmodeller**
  Att skapa prediktionsmodeller baserat på medicinsk syntetisk data ser överlag lovande ut. Modellerna är bara några få procent sämre i träffsäkerhet än originalmodellerna. Detta skapar goda förutsättningar för att generera än bättre prediktionsmodeller baserade på en större sammanslagen datamängd syntetiska data från flera regioner.

Underlag för Fas 2 fortsättningsprojekt

Under Fas 1 så har vi rent objektivt utgått från den syntetiserade datan och jämfört den med originaldatan. Både kvantitativt och hur väl den fungerar vid träning och validering av befintliga prediktionsmodeller hos Region Västerbotten. Projektets Fas 1 har varit avgränsat till att initialt utvärdera förutsättningar och möjligheter kring användning av syntetisk data. För Fas 2 fortsättningsprojektet så finns det därför ett par primära frågor som återstår att utredas:

- **På vilket sätt kan de generativa modellerna förbättras?**
  Det kortsiktiga målet för Fas 2 är att finjustera parametrarna för de utvärderade generativa modeller och på så vis även förbättra den syntetiska datan. Det originaldataset som har använts under projektets Fas 1 är generellt sett ett relativt litet dataset när det kommer till att träna maskininlärningsmodeller (då det endast omfattar ca 1700 observationer). För att ytterligare förbättra de generativa modellerna så skulle det dock behövas mera originaldata (företrädesvis motsvarande dataset från andra regioner). Ett alternativ till att dela originaldata mellan regioner skulle då kunna vara att gemensamt träna generativa modeller, exempelvis genom federerad maskininlärning[10].

- **Kan prediktionsmodeller förbättras utifrån en större sammanslagna datamängd syntetiska data?**
  En större sammanslagen datamängd syntetiska data skapar även andra förutsättningar för att träna, validera och förbättra olika prediktionsmodeller. Syntetisk data kommer dock aldrig att bli bättre än den originaldata som används från första början för att träna generativa modeller, dvs. syntetisk data har ingen möjlighet att frambringa information som inte redan återfinns i originaldatan. För att förbättra de prediktionsmodeller som har använts inom detta projekt så skulle det därför behövas en större sammanslagen datamängd syntetisk data (företrädesvis då en sammanslagen datamängd med syntetiska data från olika regioner).

Utöver ovanstående primära frågor så finns det även ett antal mera subjektiva frågor som återstår att besvara i Fas 2 fortsättningsprojektet, exempelvis:

- **Vilka är de lämpligaste tillämpningsområdena för syntetisk data?**
  När modellerna skulle kunna bli bättre med mer observationer, men data är av känslig karaktär och är svår att dela mellan dataägare. Vår tanke är att det också kan användas till att reducera bias i modellerna. Det ska utredas i nästa fas av projektet.

[10] Zhao, Zilong, et al. "Fed-TGAN: Federated Learning Framework for Synthesizing Tabular Data." *arXiv preprint arXiv:2108.07927* (2021).

- **När är det olämpligt att använda syntetisk data?**
  När data är av icke känslig art och det redan finns tillräckligt med data för att träna och utvärdera olika modeller, då är syntetisering överflödig. Det går heller inte att rädda dåliga modeller genom att syntetisera mera data utifrån få observationer originaldata. Åter igen, syntetisk data har ingen möjlighet att frambringa information som inte redan återfinns i originaldatan

- **Går det att använda syntetisk data för balansering av dataset i syfte att minska bias?**
  Går det att använda syntetisk data för balansera ett dataset med bias, t.ex. går det att öka andelen kvinnor i ett dataset som har en större andel män i originaldatan? Hur påverkar det modellens kvalitet? Detta kommer att undersökas närmare i Fas 2 av detta projektet. I praktiken är det möjligt att balansera ett dataset genom att fylla på den underrepresenterade klassen med syntetisk data. Resultatet av en sådan balansering blir oftast att den underrepresenterade klassen får en bättre precision medan den överrepresenterade klassen får sämre precision. Från fall till fall så måste man därför avgöra vad som är viktigast: att minska bias eller att ha en så bra generell metod som möjligt trots bias.

- **Vad finns det för juridiska frågor kring användandet av syntetisk data?**
  Går det att obehindrat dela ett syntetiskt dataset? Vad finns det för juridiska hinder och förutsättningar kring att dela syntetisk data? Allt detta är frågor som vi har kvar att besvara under Fas 2 av detta projekt.