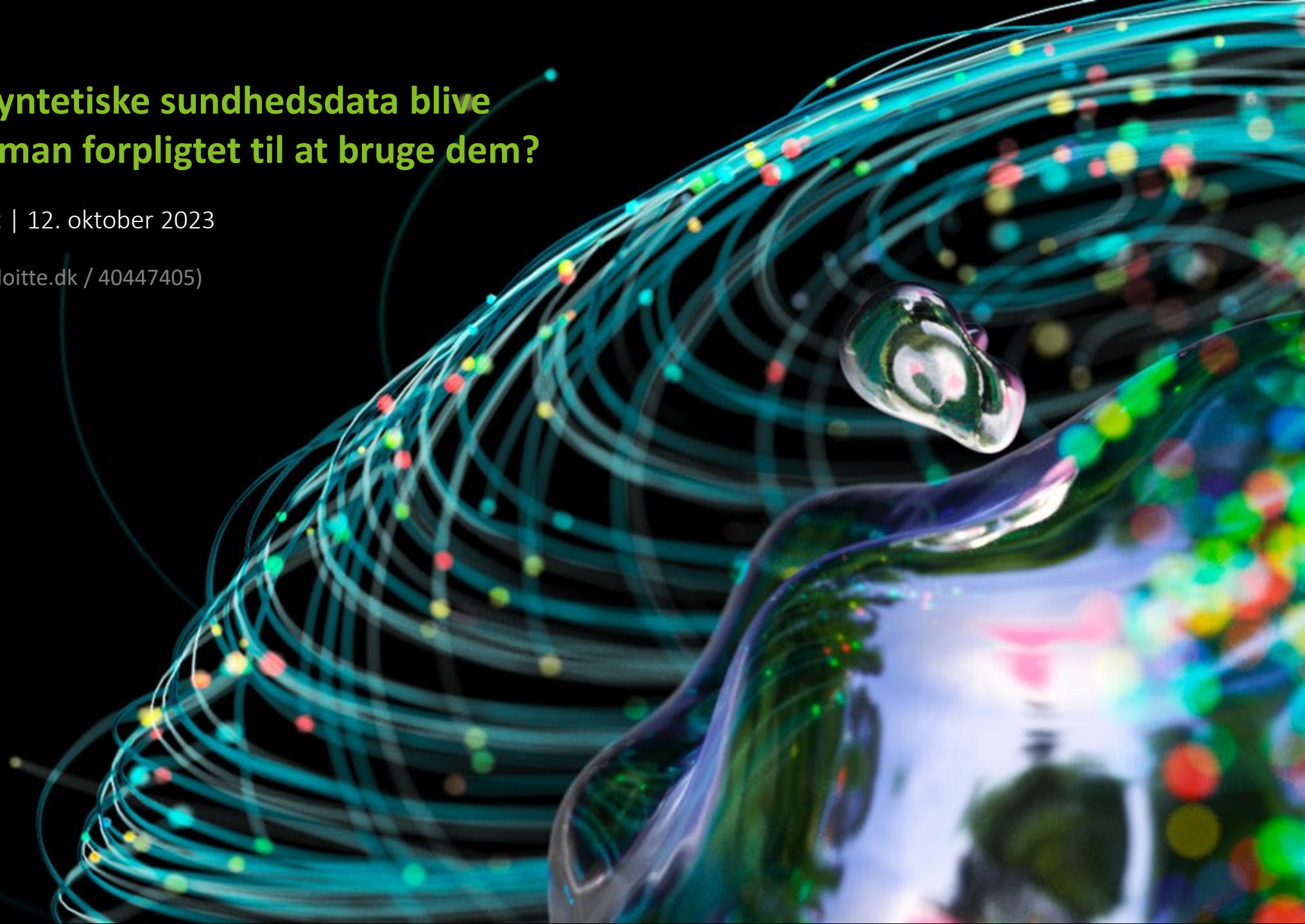
Hvor gode kan syntetiske sundhedsdata blive – og hvor bliver man forpligtet til at bruge dem?

E-Sundhedsobservatoriet | 12. oktober 2023

Thor Hvidbak (thvidbak@deloitte.dk / 40447405)



Deloitte.

Hvor gode kan syntetiske sundhedsdata blive – og hvor bliver man forpligtet til at bruge dem?

Spørgsmål 1: Hvor gode kan syntetiske sundhedsdata blive?

Det korte svar: Ret gode.

- De kan på mange typer data blive endog **rigtig gode** – hvilket vi skal se eksempler på
- Vi skal også se eksempler på, at syntetiske data kan supplere ægte data og samlet set give **bedre performance og robusthed** i udviklingen af AI-modeller – eller bruges målrettet til at reducere bias

Overordnet er der **to typer af begrænsende faktorer** for datakvaliteten:

- **Den ikke-tekniske:** Ofte er det ikke værktøjerne, som er den begrænsende faktor, men derimod at man kan gøre de syntetiske data *for gode* – og derfor af privacy-mæssige grunde *vælger* at reducere kvaliteten

- **Den tekniske:** Den nuværende modenhed er højest på tabulære data og billeddata, mens der for andre typer af datasæt stadig er begrænsninger. Endvidere stiller komplekse datamodeller særlige krav

Såvel datakvalitet som reidentifikationsrisiko skal altid varedeklareres på en overskuelig og ensartet vis

Spørgsmål 2: Og hvor bliver man forpligtet til at bruge syntetiske data?

Det korte svar: Når det er ”good enough”. Men det er dog mere end blot et sikrere alternativ…

- Som minimum på **testdata** (data til test og udvikling af it-systemer)
  - Dette har fx i mange år været linjen fra det norske datatilsyn. Og EU presser i tiltagende grad på for, at medlemsstaterne mere systematisk fremmer brugen af privatlivsbevarende teknikker (PET’er)

- Samt i en lang række situationer tillige som et bedre alternativ til konventionel anonymisering
  – hvor de syntetiske data både kan sikre **langt bedre beskyttelse og samtidig ofte en bedre anvendelighed**

- Samt derudover generelt på områder, hvor man kan opnå lige så gode resultater som med ægte data
  – og dermed iht. GDPR’s **dataminimeringskrav** er forpligtet til at bruge den metode, som har mindst risiko

- Endelig skal vi se eksempler på, at man af kvalitets- eller effektivitetsmæssige grunde *vælger* metoden til

Såvel datakvalitet som reidentifikationsrisiko skal altid varedeklareres på en overskuelig og ensartet vis

Kontekst | Mulighederne accelererer med udviklingen inden for generativ AI

Generative AI Capabilities					
Text tagging	**Programming Logic**				
Text summarization	**Code generation**				
Text comparison	**Code documentation**	Language Translation			
Composition	**Text-to-code**	**Voice Generation**	**Image Generation**	**Video Generation**	
Document creation	Web Applications	**Voice Synthesis**	Design	Video Editing	3D Models
Text	**Code**	**Speech**	**Image**	**Video**	**3D**
WRITING	***LOGIC***	***LANGUAGE***	***CREATIVITY***	***CREATIVITY***	***CREATIVITY, LOGIC***

- Længe for ChatGPT & Venner fik deres folkelige gennembrud har man dog brugt generativ AI (nærmere bestemt **Generative Adversarial Networks** – de såkaldte GAN-modeller) til at lave syntetiske data
- Vi kommer fx til at vise **scanningsbilleder**, som vi genererede i 2020, og **syntetisk fritekst (anamnesen** i henvisninger fra almen praksis) genereret i foråret 2022 – inkl. en standardmetode til at risikovurdere data
- Erfaringerne med at ”kontrollere” GAN-modeller kan tages med ind i arbejdet med **trustworthy generativ AI**

Med generativ AI kan vi også skabe simplere syntetiske datasæt alene ud fra datamodel

Videoen viser et demo-modul, der ”on demand” kan generere testdata vha. ChatGPT4. Man indlæser datamodel (tabeller med relationer) og kan herefter bestille test-borgere. Løsningen holder styr på relationerne og datamodellen. Man kan også let hurtigt tilføje eller ændre i test-populationen. Tilsvarende interfaces har vi lavet i løsninger baseret på egentlige syntetiske data. Mange myndigheder er i dag nødt til at trække på deres it-udviklere for at få nye testborgere.

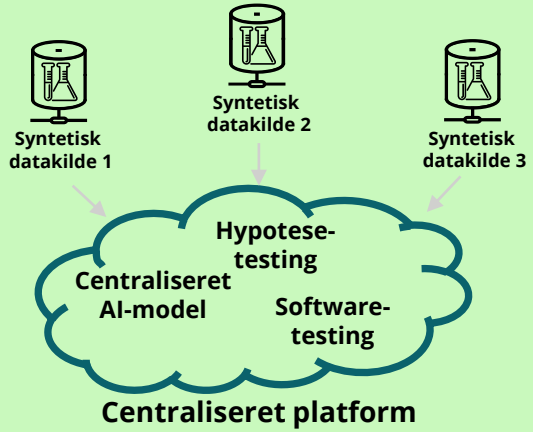
Kontekst | Syntetiske data er dog kun ét af flere greb i privacy-værktøjskassen

Avancerede platforme vil kombinere flere privacy-teknikker. Og syntetiske data kan *ikke* erstatte validering på ægte data.

Syntetiske data

hypotese → udviklingsfase

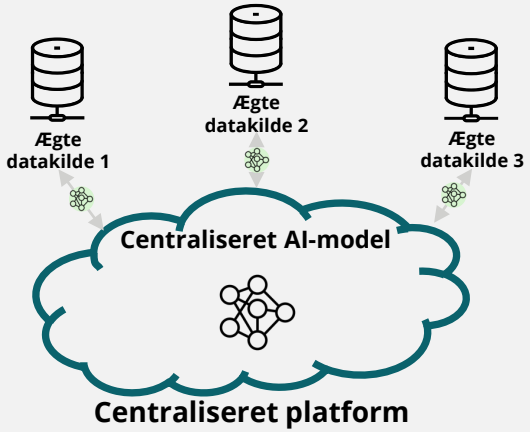
- ✓ Centralisering for datadiversitet
- ✓ Hypoteseudvikling og -test
- ✓ Balancering af datasæt
- ✓ Softwaretest og -udvikling
- ✓ Er modsat de andre irreversibel



Fødereret læring

udviklingsfase → drift

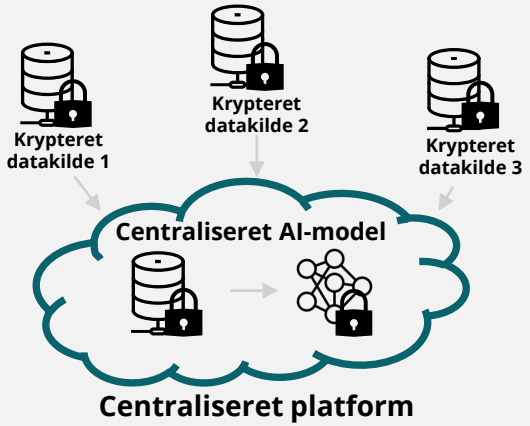
- ✓ Datadiversitet igennem føderation (model/analyse kommer til data)
- ✓ Ægte data (kræver hjemmel)
- ✓ Fordele ressourceforbrug
- ✓ Dataejerne har stadig kontrol



Krypterede beregninger

udviklingsfase → drift

- ✓ Datadiversitet igennem krypteret centralisering – *en slags meget avanceret pseudonymisering*
- ✓ Ægte data (kræver hjemmel)
- ✓ Samle ressourceforbrug



Et yderligere lag af sikkerhed kan blive tilføjet til alle tre teknologier gennem **differential privacy**, der utydeliggør de underliggende ægte data ved at tilføje tilfældig støj på en måde, så "tabet" af privat information gøres matematisk målbart.

Læs mere: https://radar.dk/holdning/danmark-halser-efter-i-privatlivsbeskyttelse-og-moderne-metoder-til-sikker-datadeling

Syntetiske data er på godt og ondt en god efterligning

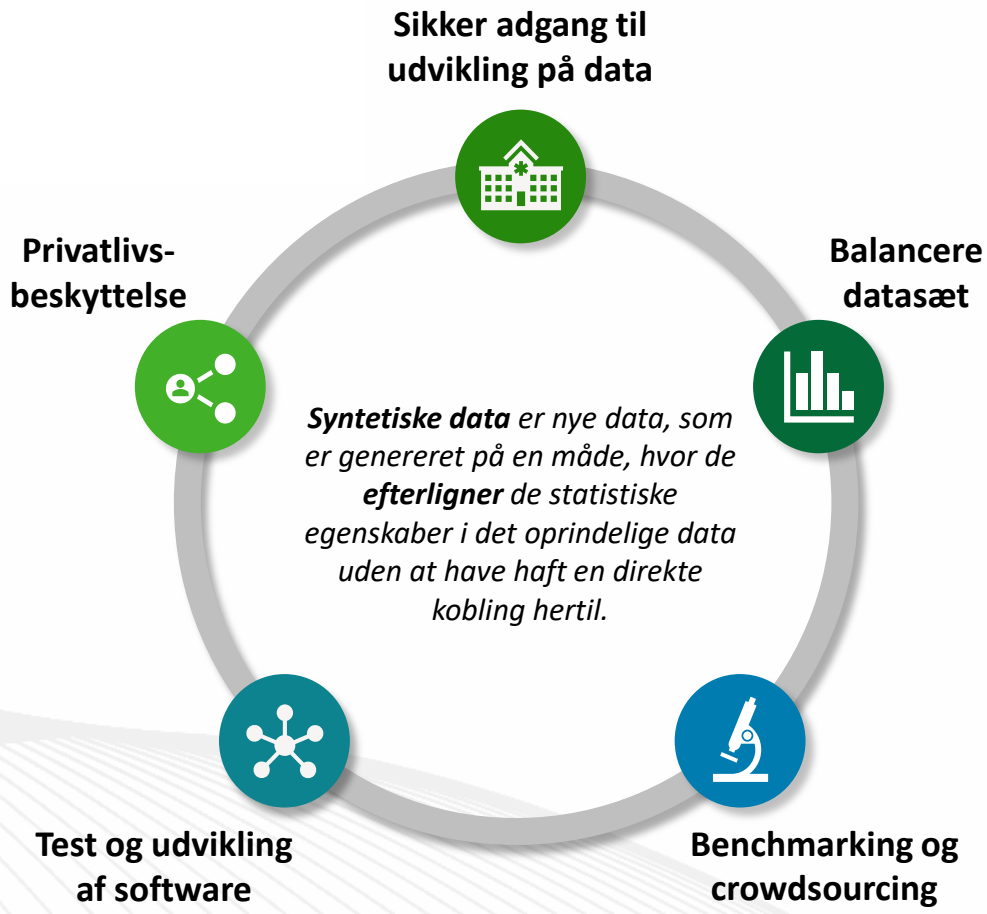
Syntetiseringen bryder koblingen til de originale data mere radikalt end anden anonymisering og har på godt og ondt et element af tilfældighed. Men med de rette metoder kan der genereres syntetiske data af høj kvalitet, som til mange formål er *både* mere anvendelige *og* langt mere sikre end konventionelt anonymiserede data.

The "good":

- ✓ En meget god efterligning = **højere anvendelighed** end konventionelt anonymiserede data
- ✓ **Meget højere privatlivsbeskyttelse** (koblingen til individer i det oprindelige datasæt brydes)
- ✓ Nemmere at **operationalisere** og gøre målbar end eksempelvis differential privacy

The "bad":

- ➢ Stadig kun en god efterligning
- ➢ Bedst til generelle sammenhænge
- ➢ Privatlivsrisiko er ikke nødvendigvis elimineret (modsat hvad nogle producenter hævder…)
- ➢ Svært at generere nogle typer komplekse datasæt

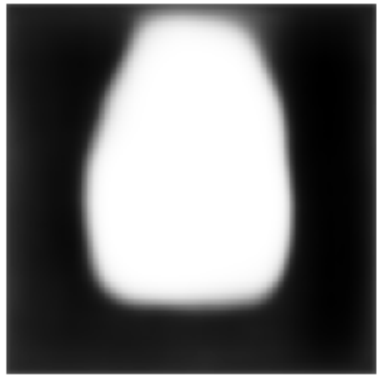


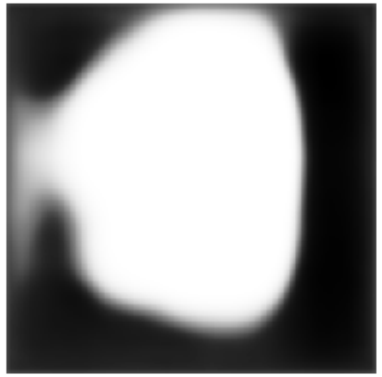
Hvordan? – Forestil jer at spille ”tampen brænder” 1 mio. gange…

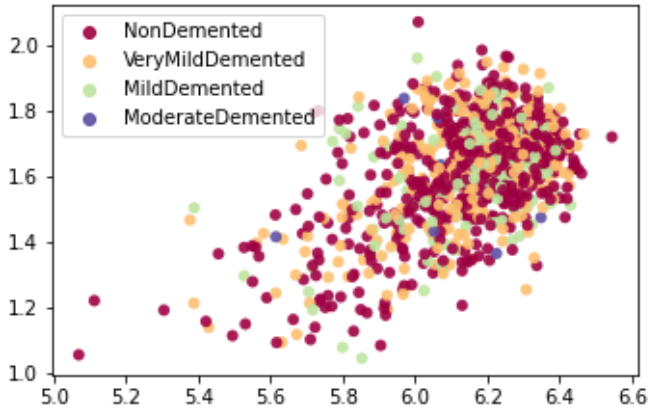
Generative Adversarial AI

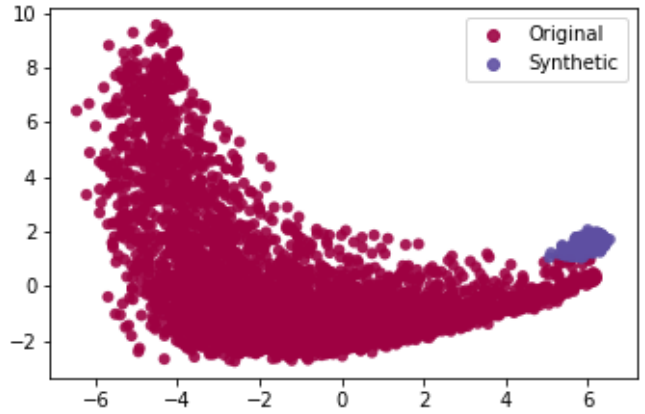
To konkurrerende AI-modeller, hvor modellen, som skaber de syntetiske data, aldrig får adgang til de rigtige data

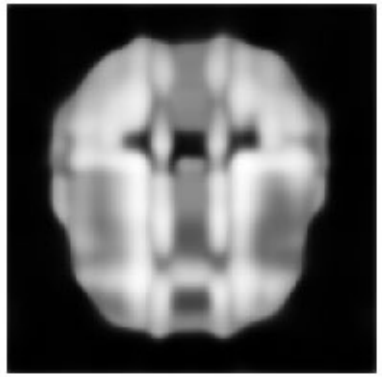
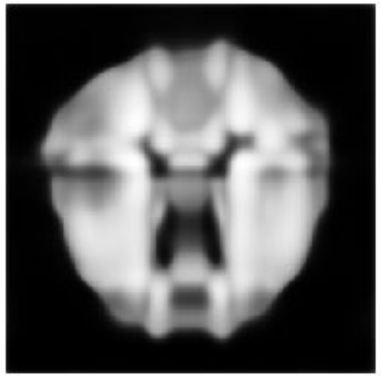
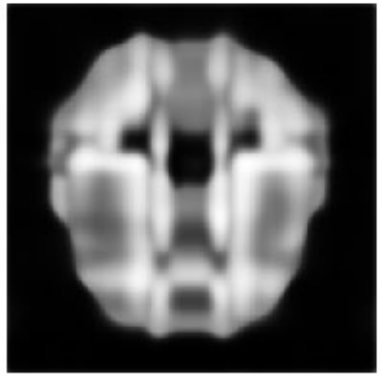
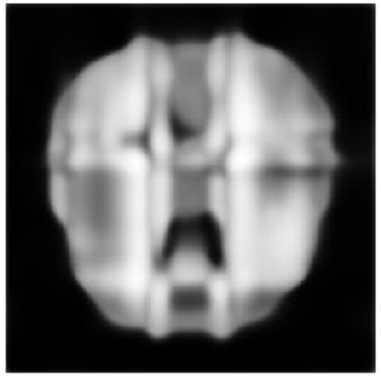
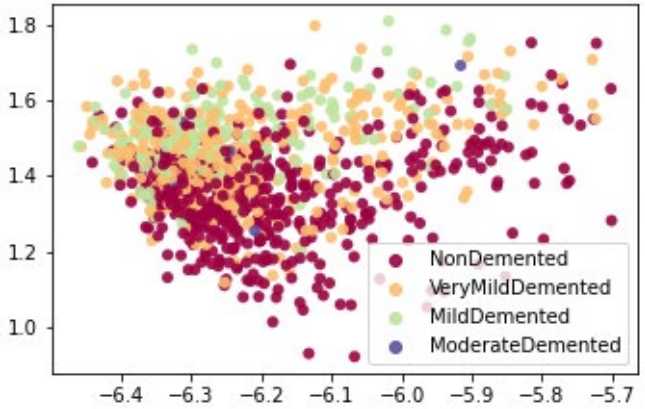
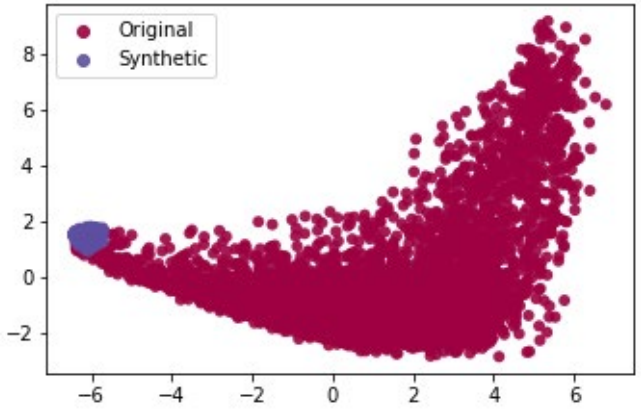
Den generative model starter med tilfældig støj og prøver sig frem – og kommer gradvist tættere på

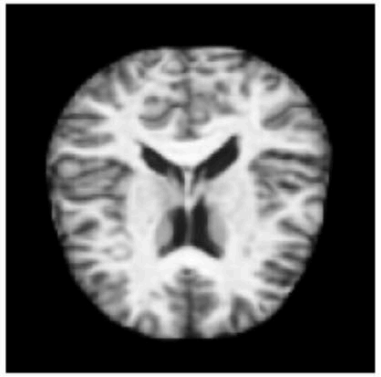


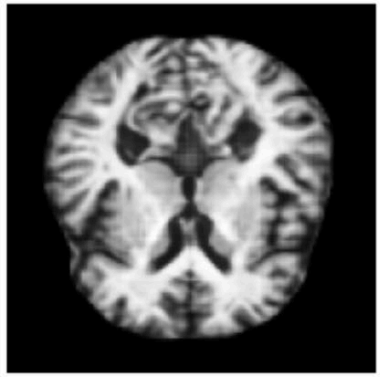


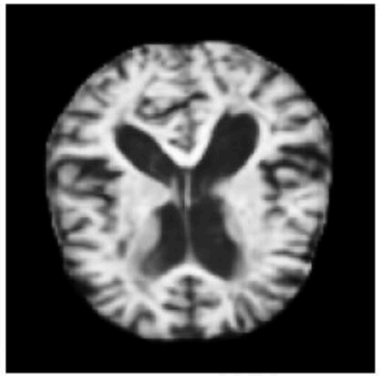

NonDemented
VeryMildDemented
MildDemented
ModerateDemented
2.0
1.8
1.6
1.4
1.2
1.0
5.0
5.2
5.4
5.6
5.8
6.0
6.2
6.4
6.6

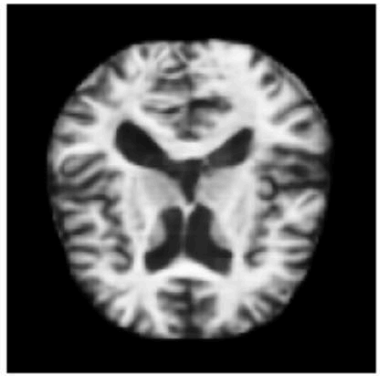

Original
Synthetic
10
8
6
4
2
0
−2
−6
−4
−2
0
2
4
6






1.8
1.6
1.4
1.2
1.0
−6.4
−6.3
−6.2
−6.1
−6.0
−5.9
−5.8
−5.7
NonDemented
VeryMildDemented
MildDemented
ModerateDemented

Original
Synthetic
8
6
4
2
0
−2
−6
−4
−2
0
2
4
6



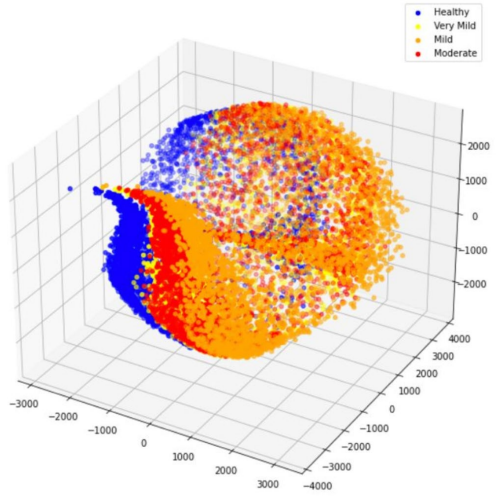


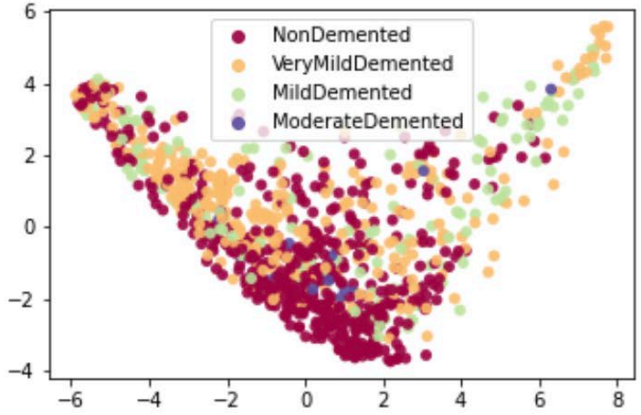


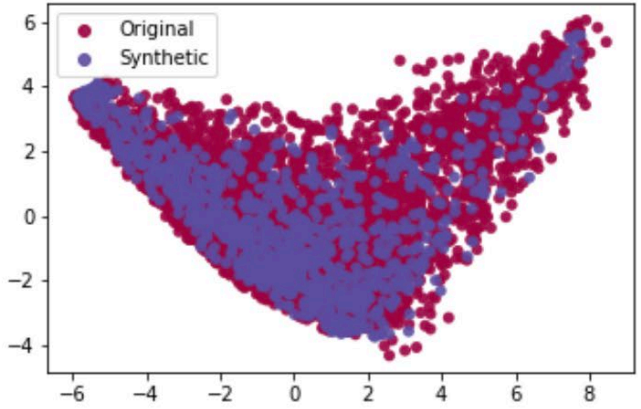


"It was interesting to see that synthetic data could be used to supplement the real data to improve AI predictive performance."

– Henning Langberg (Chief Innovation Officer at Rigshospitalet), 2020





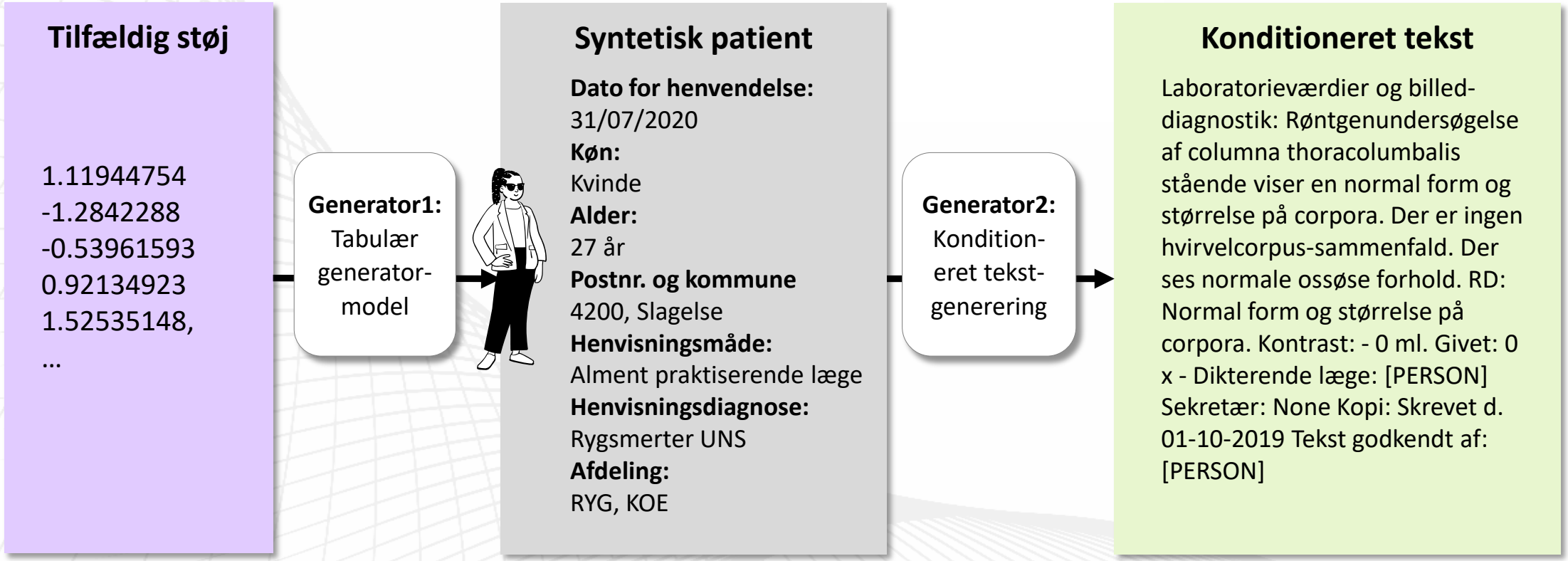


novo nordisk fonden

REGION H Rigshospitalet

Generering af fritekst (længe før LLM-modellerne fik deres folkelige gennembrud)

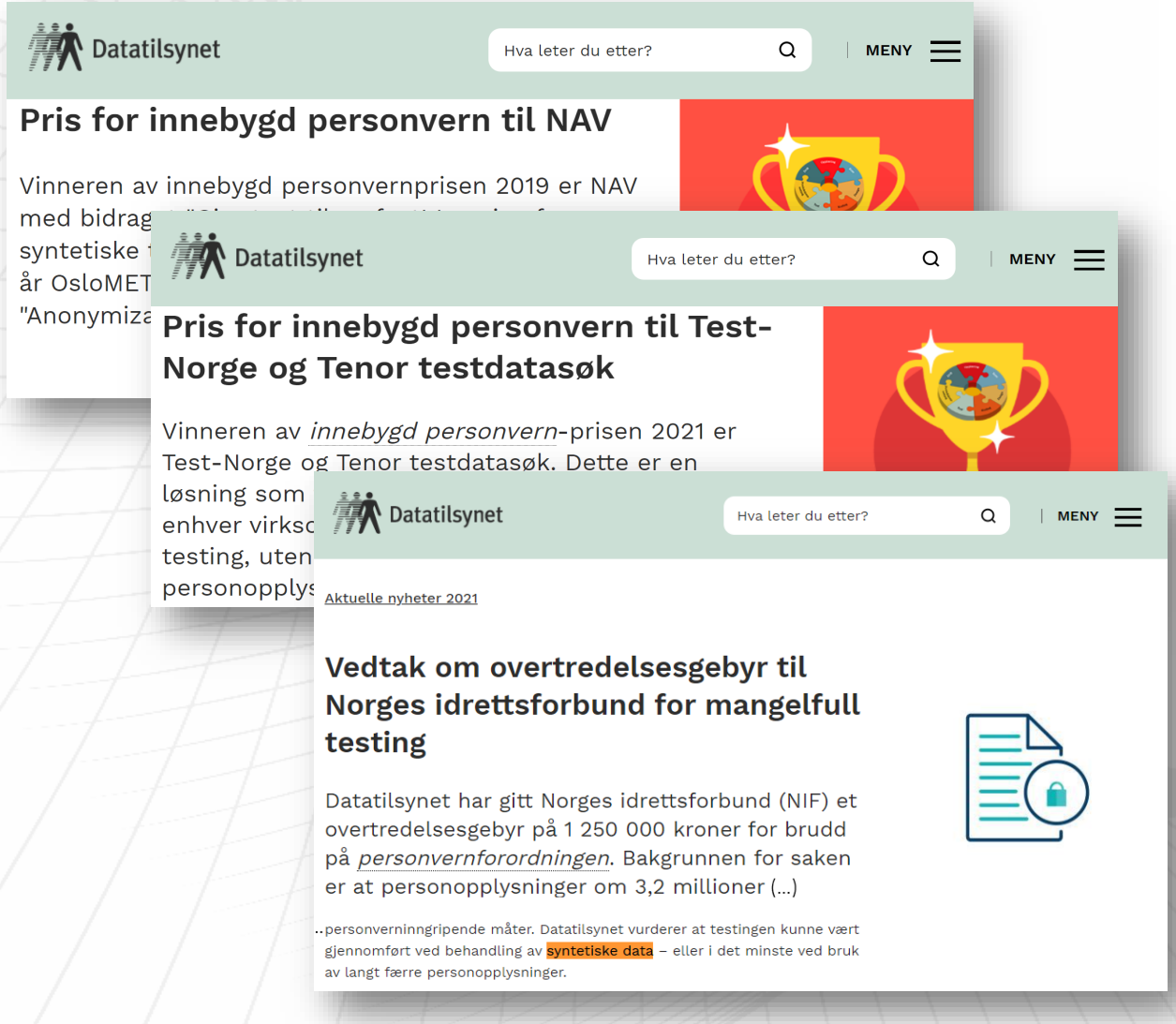
Generering af syntetiske patienthenvisninger i en løsning, der i to trin bruger to typer syntetisering. Data bruges til træning af en AI-model til automatiseret henvisningstriage. Der er udviklet en standardmetrik for privacy-risikoen, også på tekst.



REGION SJÆLLAND

Test | Anbefalet af det norske datatilsyn som et dataminimerende alternativ siden 2018

Og siden 2021 har datatilsynet udskrevet bøder til organisationer, som fortsat bruger person-/produktionsdata til test.



Test | Syntetiske data ”on demand”: Dataminimering og mere effektive testprocesser

Mulighed for at lægge udviklingsmiljøet i skyen uden borgerdata. Og mindre manuelt arbejde med at finde/lave testdata. Kompleks datamodel fordelt på ca. 1.500 tabeller – så meget af arbejdet handlede om at dokumentere data og relationer.

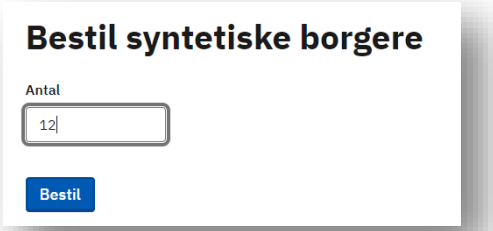
A. Syntetisering af testdata | B. Webapplikation | C. Integration og test

Brugerrejse

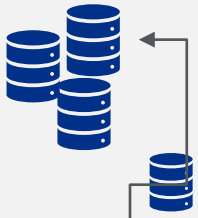
1. Generér testdata
Test og domæneekspert starter et testdata projekt via web interfacen og påbegynder syntetisering.

2. Publicér data
Efter at have undersøgt dataet og evt. manuelt tilføjet ydertilfælde publiceres data til skyen.

3. Test af modul
Det syntetiske test data kan nu tilgås af test modulet i Azure DevOps.



STAR on-prem



Indhentning af ønsket data on prem

Generator-modul

Træning af AI model og generering af syntetisk testdata

Web interface
Valgte dataområder sendes til syntetisering

Publicér til Azure

Automatiseret kvalitetssikring før der publiceres

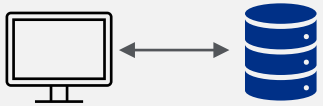
STAR Azure platform

Azure

Syntetisk testdata sendes op i Azure blob storage

Container læser syntetisk testdata og tilgås via testmodul

Test-modul



- Målrettet testere
- Er automatisk genereret ud fra ægte data
- Kan bruges til at generere gode realistiske arketyper af test individer
- Planlagt at følge samme genererings-struktur som konstrueret data

✓ **Gode muligheder for at bruge i cloud – Stærk dataminimering og beskyttelse**

- ☑ HasCV
- ☑ IsEnrolled
- ☑ WithContactGroup
- ☑ HasAbsence
- ☐ WithPension
- ☐ WithProfession
- ☐ WithEducation
- ☐ HasInterviews
- ☐ ..

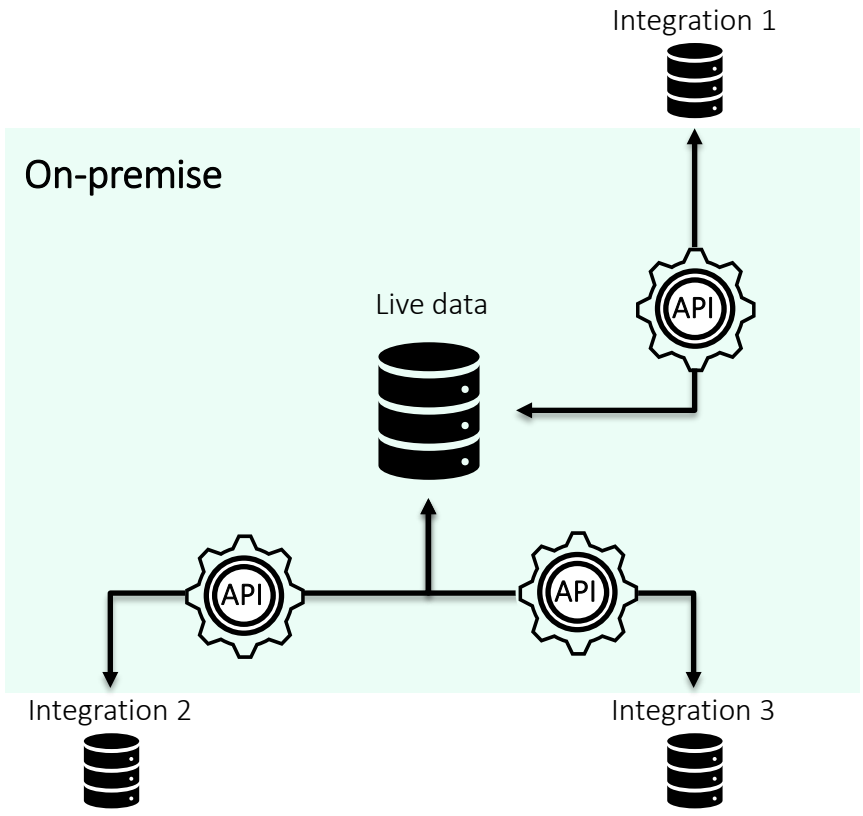
CPR: 310857-0000
Navn: Iver Syntesia Bech
Fødselsdag: 1957-08-31
Adresse: Tagensvej 56 4th, 2200
Enrolled: Tilmeldt
Kontaktgruppe: Seniorjob
Fravær: Sygemeldt

Test | Syntetiske API'er kan simulere integrationer, så man kan holde test i eget miljø

Dataminimerende og tidsbesparende.

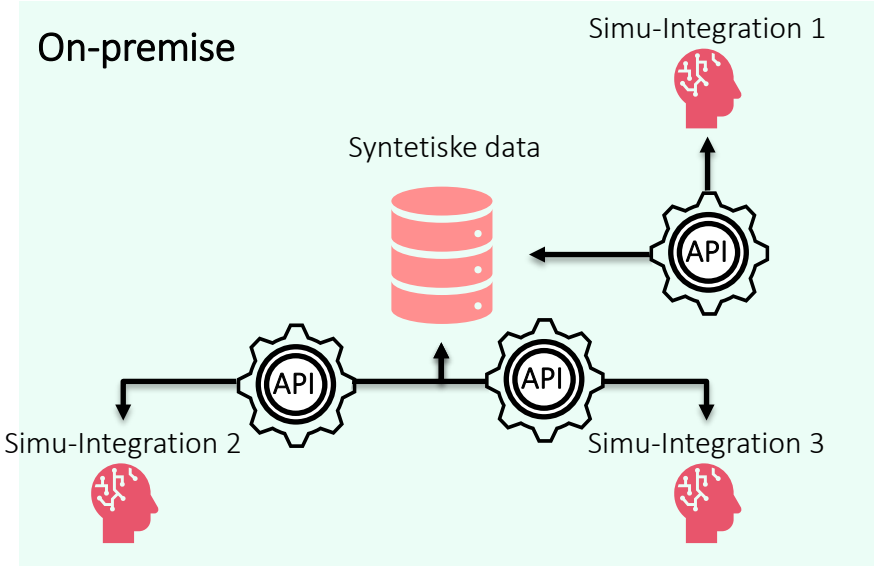
Produktionsmiljø

Produktionsscenarie med integrationer



Testmiljø med simulerede integrationer

Testmiljø med syntetiske testdata og simulerende integrationer



Borgen vil have Danmark med på vognen

Åben høring om anvendelse af sundhedsdata

Udvalget for Digitalisering og It og Sundhedsudvalget

Onsdag den 12. april 2023 kl. 10.00-12.00, vær. 1-133

Formål

Formålet med høringen er at få belyst potentialer samt etiske og retssikkerhedsmæssige problemstillinger ved indsamling og deling af sundhedsdata.

Udvalgene ønsker i den forbindelse at få belyst, hvilke initiativer der skal iværksættes for *at* sikre borgernes ret til at modtage den bedst mulige behandling og for, *at* behandlerne har de samme oplysninger til rådighed på tværs af sektorer.

Udvalgene ønsker også konkrete forslag til, hvordan der reelt sikres anonymitet, herunder om syntetisk data kan være vejen frem på en række områder.

Andre spørgsmål, som ønskes belyst, er, hvordan borgerne skal forholde sig til deres rettigheder i forhold til datasikkerhed, samtykke og indsigt. I den forbindelse ønskes konkrete forslag til modernisering af lovgivningen på området, som både tager højde for den teknologiske udvikling, nye forsknings- og udviklingsmuligheder og dataetikken.

EU presser på – men vi skal dog til de angelsaksiske lande for at finde guidelines

USA har for nylig vedtaget lovgivning om PET (privacy-teknikker). UK's ICO har i juni 2023 udgivet en opdateret vejledning.

- "Personal data shall be (…) adequate, relevant and **limited to what is necessary** in relation to the purposes for which they are processed ('**data minimisation**')." – GDPR (2016), artikel 5

- "There are techniques enabling analyses on databases that contain personal data, such as (…) the use of **synthetic data or similar methods** and other state-of-the-art privacy-preserving methods that could contribute to a more privacy-friendly processing of data. **Member States should provide support to public sector bodies to make optimal use of such techniques**." – Data Governance Act (2022), præambel, stk. 7

- "The health data access bodies (…) **should apply tested techniques that ensure electronic health data is processed in a manner that preserves the privacy** of the information contained in the data." – European Health Data Space (2022, *forhandles stadig*), præambel, stk. 43

- "In the AI regulatory sandbox personal data lawfully collected for other purposes shall be processed for the purposes of developing and testing certain innovative AI systems (…) where those requirements cannot be effectively fulfilled by processing **anonymised, synthetic or other non-personal data**." – AI Act (2021, *forhandles stadig*)

- "This includes **recommendations about technology for preserving privacy**." – EU/DIGST: Trustworthy AI for the Danish Local and Regional Authorities (2023), udbudstekst (Deloitte har for nylig vundet opgaven)

Ikke ”kun” compliance og privacy, men også en forudsætning for skalering af AI

Skalerbar AI-udvikling kræver diversitet i træningsdata

Gartner

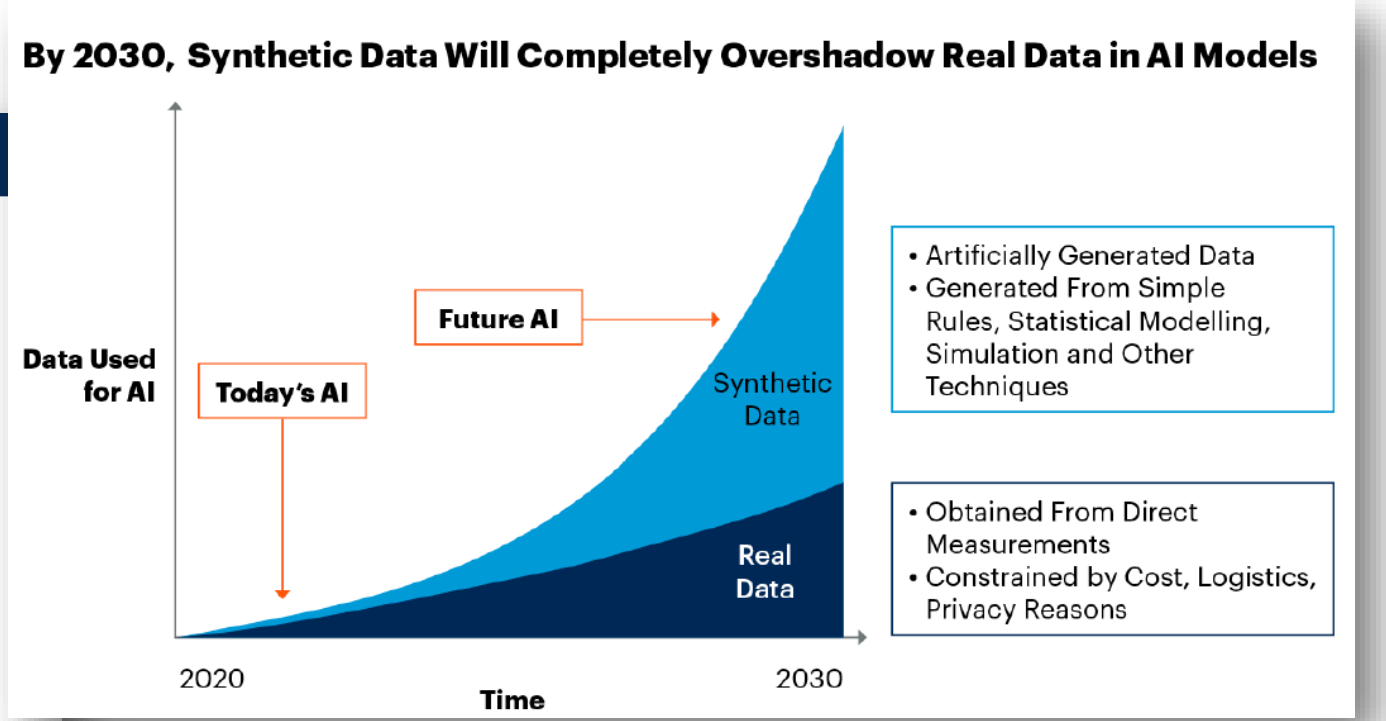
Maverick* Research: Forget About Your Real Data — Synthetic Data Is the Future of AI

Published 24 June 2021 - ID G00750175 - 22 min read

By Analyst(s): Leinar Ramos, Jitendra Subramanyam

Initiatives: Artificial Intelligence

> Synthetic data is often seen as a lower-quality substitute, useful only when real data is inconvenient to get, expensive or constrained by regulation. This misses the true potential of synthetic data. The fact is you won't be able to build high-quality, high-value AI models without synthetic data.



"We predict **massive increase in adoption** as synthetic data:

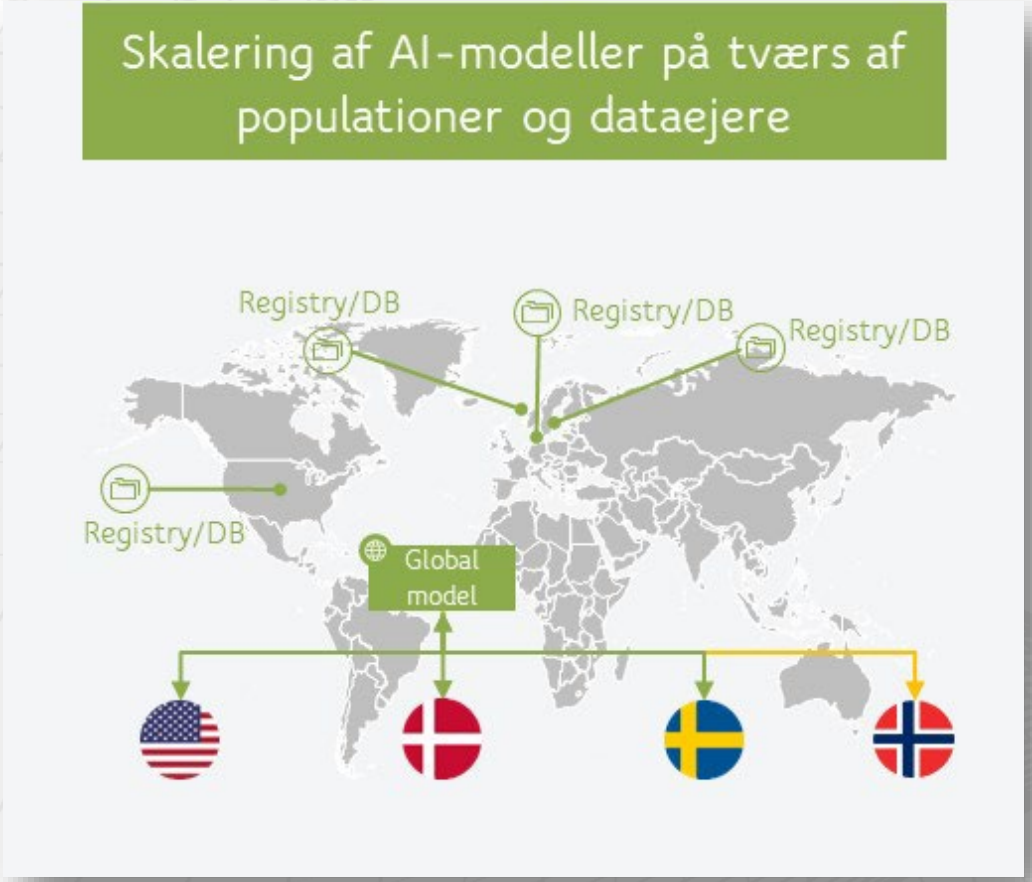
- **Avoids using PII [persondata] when training** machine learning (ML) models via synthetic variations of original data or synthetic replacement of parts of data.
- **Reduces cost and saves time** in ML development as it is cheaper and faster.
- **Improves ML performance** as more training data leads to better training outcomes."

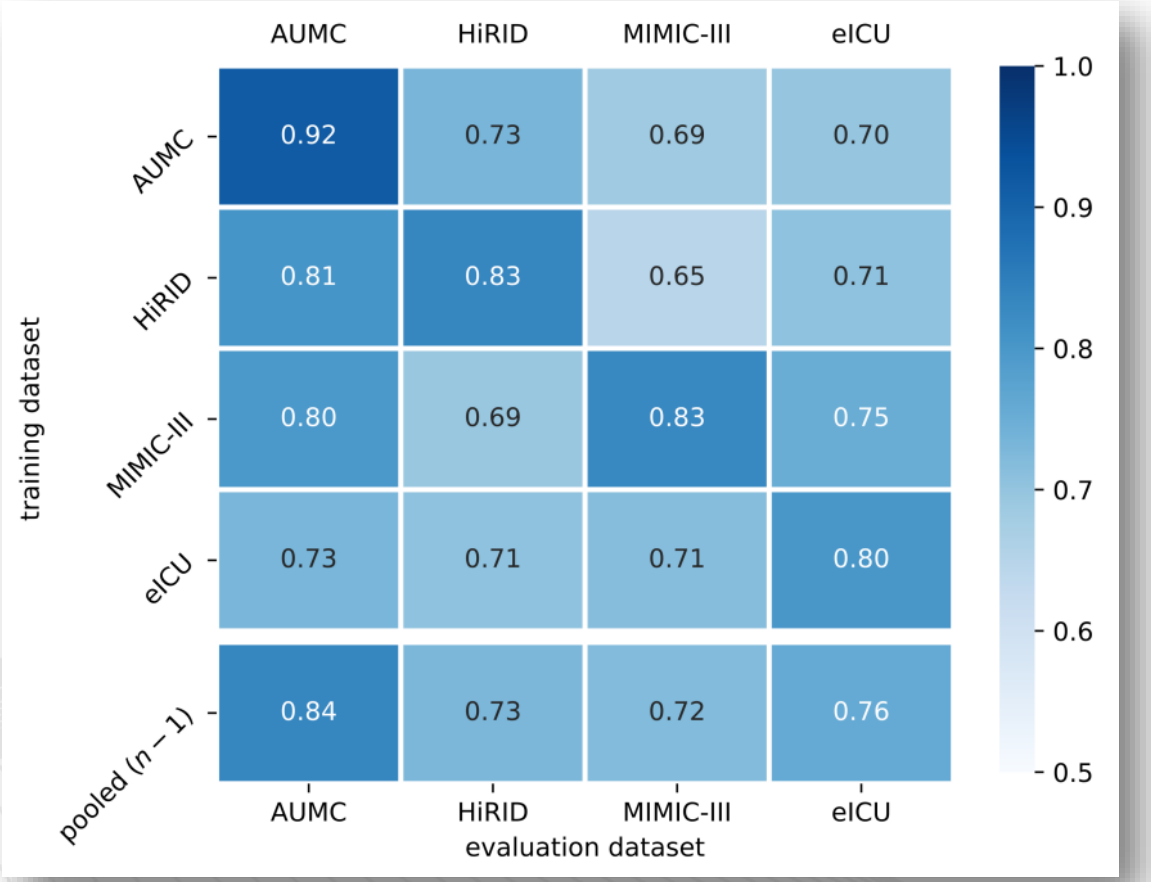
(Gartner, august 2022)

Vi skal kunne benchmarke (sammenligne) AI-modeller

Skalering af AI lider under manglende robusthed på tværs af populationer. Man kan ikke på forhånd vide, hvor godt en model rejser fra én setting (fx land, region) til en anden. Men man ved, at performance stiger med mere alsidige data.

Syntetiske benchmark-datasæt vil kunne muliggøre sammenligning af AI-modeller og accelerere ibrugtagningen.





Moor et al. (2021): *Predicting sepsis in multi-site, multi-national intensive care cohorts using deep learning*

… det gør man bl.a. i Australien

Her med eksempler på plots, der viser distributioner i syntetiske sepsis-datasæt sammenlignet med de ægte data.

scientific data

OPEN

DATA DESCRIPTOR

The Health Gym: synthetic health-related datasets for the development of reinforcement learning algorithms

Nicholas I-Hsien Kuo[1]✉, Mark N. Polizzotto[2], Simon
Anders Sönnerborg[7], Maurizio Zazzi[8], Michael Böhm
Sebastiano Barbieri[1]

In recent years, the machine learning research community
availability of openly accessible benchmark datasets. Clini
to their confidential nature. This has hampered the develo
machine learning applications in health care. Here we intr
of highly realistic synthetic medical datasets that can be fr
and compare machine learning algorithms, with a specific
three synthetic datasets described in this paper present pa
and sepsis in the intensive care unit, and people with huma
antiretroviral therapy. The datasets were created using a r
The distributions of variables, and correlations between va
the synthetic datasets mirror those in the real datasets. Fu
disclosure associated with the public distribution of the sy

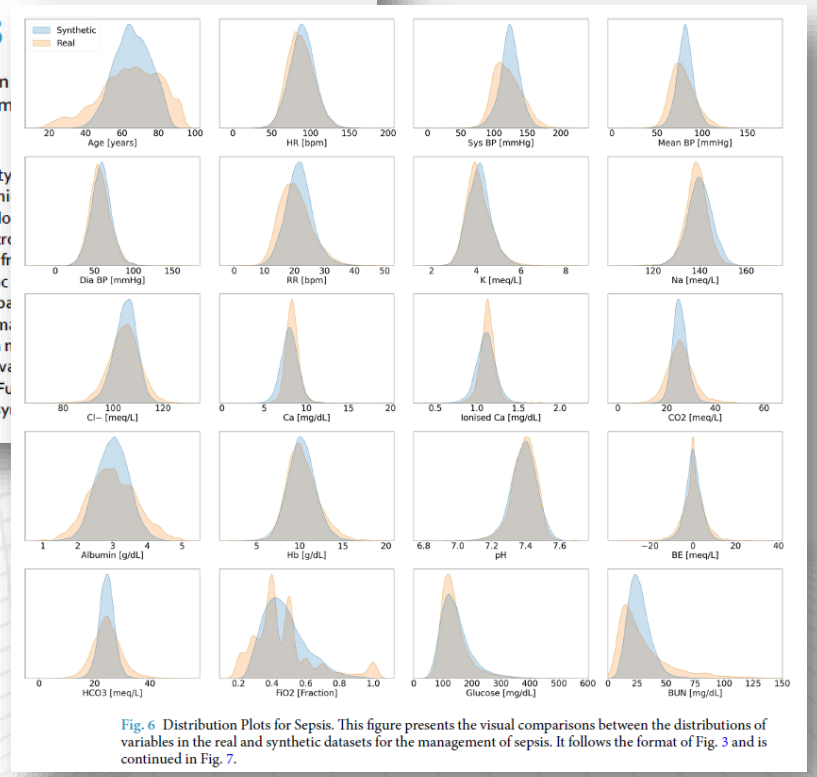


Fig. 6 Distribution Plots for Sepsis. This figure presents the visual comparisons between the distributions of variables in the real and synthetic datasets for the management of sepsis. It follows the format of Fig. 3 and is continued in Fig. 7.

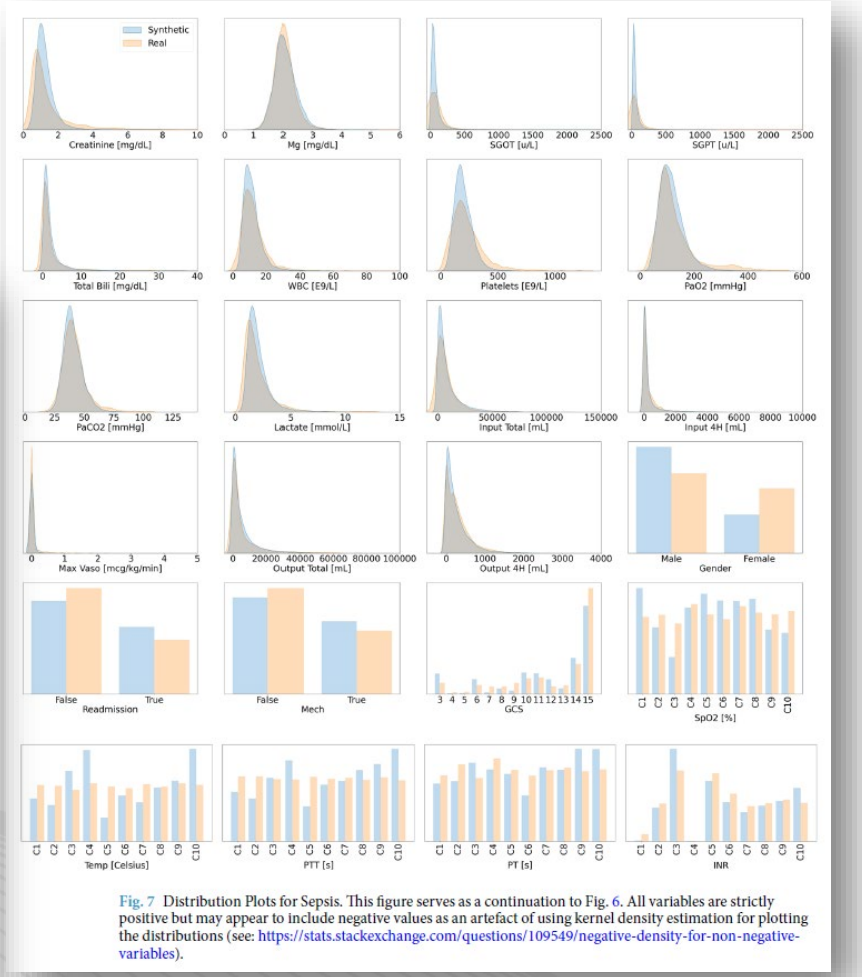


Fig. 7 Distribution Plots for Sepsis. This figure serves as a continuation to Fig. 6. All variables are strictly positive but may appear to include negative values as an artefact of using kernel density estimation for plotting the distributions (see: https://stats.stackexchange.com/questions/109549/negative-density-for-non-negative-variables).

Hvor gode data kan blive – en reel mulighed for at træne modeller på anonyme data

Muligheden for at lave robust anonymiserede data, der kan bruges til accelereret træning og udvikling af AI-modeller.

npj | digital medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Check for updates

EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records

Jinsung Yoon[1]✉, Michel Mizrahi[1], Nahid Farhady Ghalaty[1], Thomas Jarvinen[1], Ashwin S. Ravi[1], Peter Brune[1], Fanyu Kong[1], Dave Anderson[1], George Lee[1], Arie Meir[2], Farhana Bandukwala[1], Elli Kanal[2], Sercan Ö. Arık[1] and Tomas Pfister[1]

Privacy concerns often arise as the key bottleneck for the sharing of data between consumers and data holders, particularly for sensitive data such as Electronic Health Records (EHR). This impedes the a
with tremendous potential. One promising approach for such privacy conc
generative modeling framework, EHR-Safe, for generating highly realistic a
based on a two-stage model that consists of sequential encoder-decoder
innovations focus on the key challenging aspects of real-world EHR data:
categorical features with distinct characteristics, and time-varying features
evaluations, we demonstrate that the fidelity of EHR-Safe is almost-identical
trained on them) while yielding almost-ideal performance in practical priv

npj Digital Medicine (2023)6:141 ; https://doi.org/10.1038/s41746-023-00888

”Under numerous evaluations, we demonstrate that the fidelity of EHR-Safe is **almost-identical with real data** (<3% accuracy difference for the models trained on them) while yielding **almost-ideal performance in practical privacy metrics**.”

Table 2. Fidelity results with utility metrics.

Utility with all features

Target	Models	Metrics	MIMIC-III		eICU	
			Train on Real	Train on Synth	Train on Real	Train on Synth
Mortality	GBDT	AUC	**0.762**	**0.736**	0.943	0.938
		AP	**0.304**	**0.261**	0.600	0.534
	RF	AUC	0.723	0.710	**0.954**	**0.945**
		AP	0.276	0.251	**0.600**	**0.580**
	GRU	AUC	0.728	0.667	0.937	0.938
		AP	0.278	0.193	0.567	0.528
	LR	AUC	0.712	0.680	0.872	0.818
		AP	0.233	0.207	0.323	0.260
	Average	AUC	0.731	0.689	0.926	0.909
		AP	0.272	0.228	0.522	0.475

Et helt andet trade-off mellem datakvalitet og re-identifikationsrisiko

Hurtig sammenligning uden optimering af modellen. Alligevel er det muligt at få et helt andet forhold mellem datakvalitet og risiko end med de konventionelle anonymiseringsmetoder (der endvidere ofte ikke er egnet til AI-udvikling).

Datasæt	Risiko for at lære noget nyt om en person	Nøjagtighed
Diabetes (ægte)	76,2%	97%
GAN-syntetiseret (uden optimering)	5.3%	92%
K3-anonymiseret	76,2%	95%
K5-anonymiseret	76,2%	96%

gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Male	47.0	0	0	ever	48.20	5.7	100	0
Female	40.0	0	0	never	27.63	6.5	126	0
Male	19.0	0	0	never	20.80	4.0	100	0
Female	59.0	0	1	never	60.26	8.8	145	1
Male	47.0	0	0	No Info	27.32	6.1	126	0

gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Female	47.4	0	0	never	34.87	7.3	283	1
Female	61.6	0	1	not current	27.32	6.2	144	0
Male	61.4	0	0	No Info	17.66	5.0	131	0
Female	3.9	0	0	never	28.04	5.0	127	0
Female	10.5	0	0	No Info	16.78	4.0	73	0

gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Male	43.0-47.0	0	0	never,ever	44.7-48.8	5.7	100	0
Female	40.0	0	0	never	27.6-29.9	6.5	126	0
Male	18.0-19.0	0	0	never	20.5-21.1	4.0	100	0
Female	53.0-69.0	0	1	never,No Info	57.8-62.0	8.8	145	1
Male	47.0	0	0	No Info	27.3-32.1	6.1	126	0

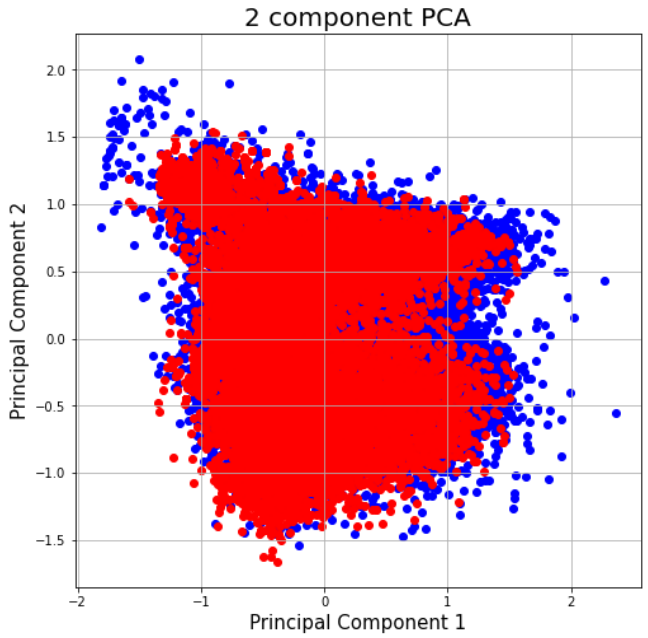
gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Male	43.0-51.0	0	0	never,ever	44.7-49.0	5.7	100	0
Female	40.0	0	0	never	27.4-29.9	6.5	126	0
Male	18.0-19.0	0	0	never	20.5-21.1	4.0	100	0
Female	53.0-69.0	0	1	never,No Info	57.8-62.0	8.8	145	1
Male	47.0	0	0	No Info	27.3-32.1	6.1	126	0

Hvordan varedeklareres datakvalitet?

Sammenligning af fordelinger og performance af modeller trænet på syntetiske vs. ægte data.

Sammenligning af fordelinger

Beskrivelse:
Ved brug af en præ-trænet AI model kan vi komprimere teksterne ned til tal og sammenligne de syntetisk genererede anamneser (rød) med de ægte (blå)



Ægte model mod ægte data

Beskrivelse:
For at lave en baseline træner vi en model simpel AI model til at klassificere (visitere) hvilken afdeling, anamnesen hører til (ortopædkir, gastro mm.).

Træningsdata: Alle ægte anamneser før 2021
Testdata: Alle ægte anamneser fra start 2021

Mest hyppige afdelinger	Sensitivitet	Præcision	F1 score
ORTOPÆDKIR	85%	98%	91%
KARDIOLOGISK	76%	92%	83%
GYNÆKOLOGISK	85%	92%	88%
KIR	64%	86%	73%
PÆDIATRISK	84%	70%	76%
MAMMARADIOLOGI	100%	82%	90%
Vægtet gennemsnit over alle (inkl. de afdelinger udestående fra denne tabel)	76%	79%	76%

Syntetisk model mod ægte data

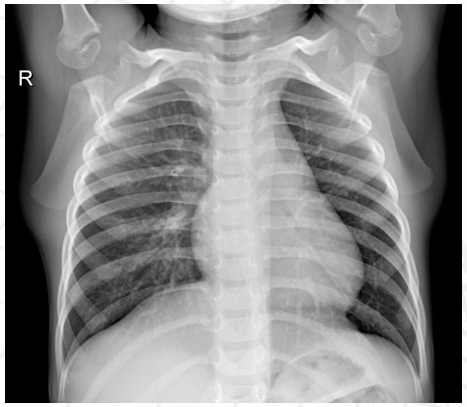
Beskrivelse:
Vi sammenligner hvor meget den syntetiske model taber i ydeevne fra den ægte model, når det evalueres på testdatasættet.

Træningsdata: Alle syntetiske anamneser
Testdata: Alle ægte anamneser fra start 2021
Epochs: 50

Mest hyppige afdelinger	Sensitivitet	Præcision	F1 score
ORTOPÆDKIR	83%	86%	84%
KARDIOLOGISK	73%	74%	73%
GYNÆKOLOGISK	80%	63%	70%
KIR	30%	78%	43%
PÆDIATRISK	45%	72%	56%
MAMMARADIOLOGI	0%	0%	0%
Vægtet gennemsnit over alle (inkl. de afdelinger udestående fra denne tabel)	56%	56%	52%

Hvordan varedeklareres datakvalitet?

Billeddata.



Healthy

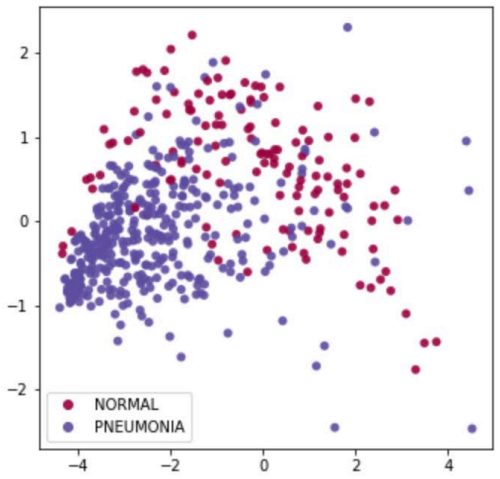
Pneumonia

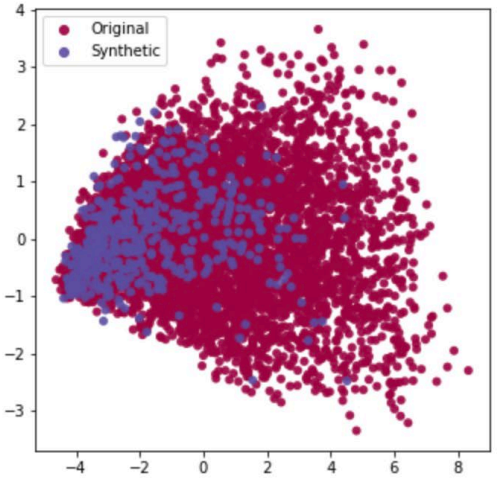
Pneumonia dataset

- Balanced data
- ~5000 datapoints. Varying sizes, most > 1000px
- A lot of kid images (different bone structure)
- Resized to 512x512

Classification model

- Encode images using inception embeddings
- Multiclass logistic regression with embeddings as input
- Accuracy on real data
  - Training data: 85.4%
  - Test data: 85%



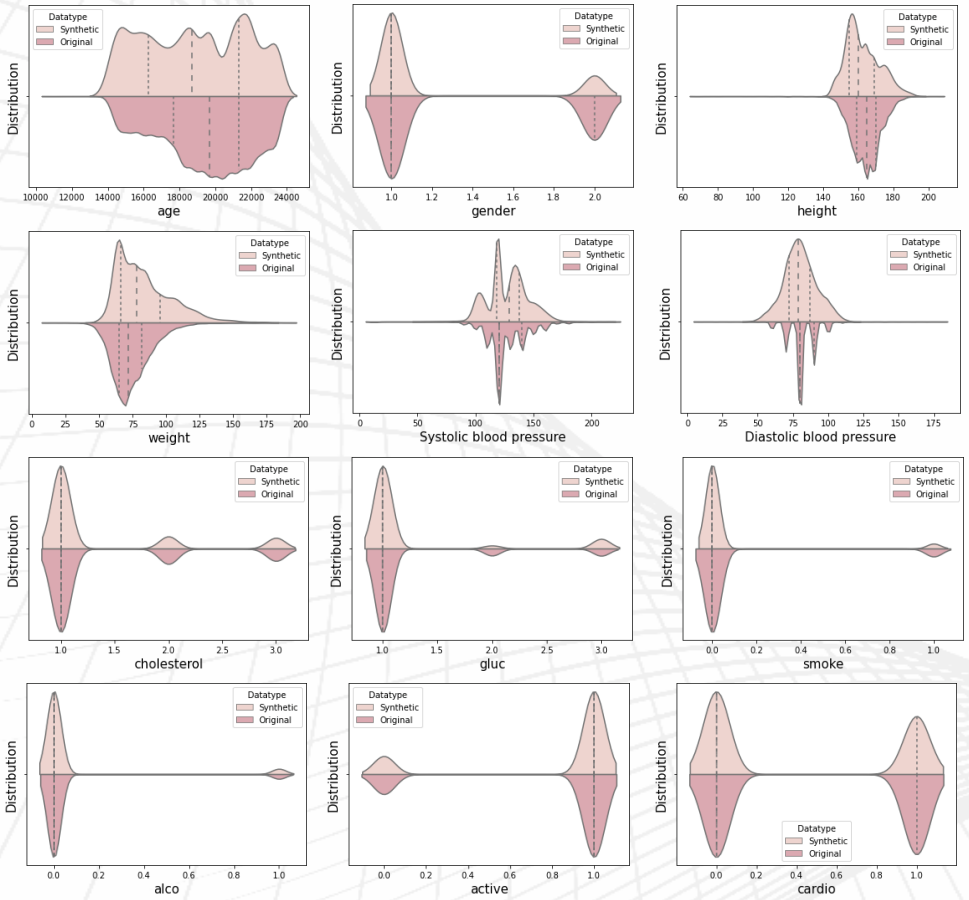


kaggle

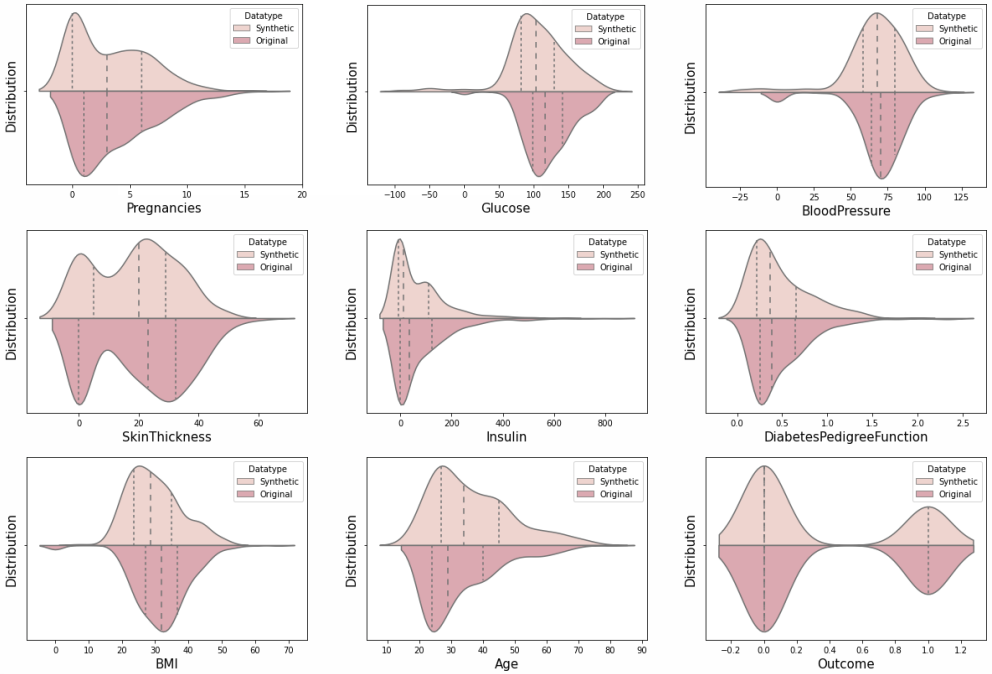
Hvordan varedeklareres datakvalitet?

Univariat sammenligning.

Cardiovascular Variables



Diabetes Variables

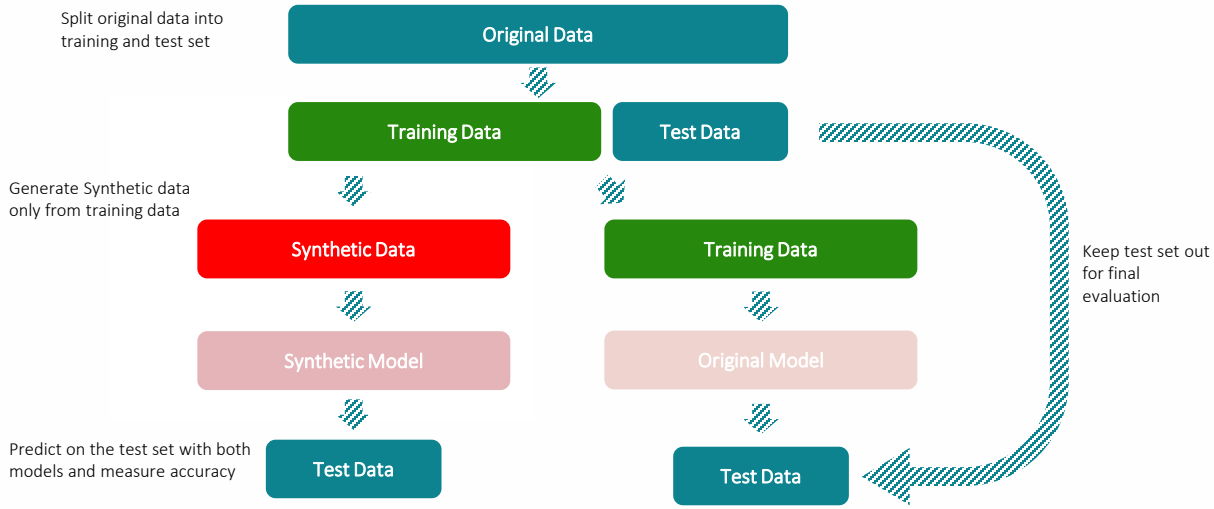


Hvordan varedeklareres datakvalitet?

Multivariat sammenligning.

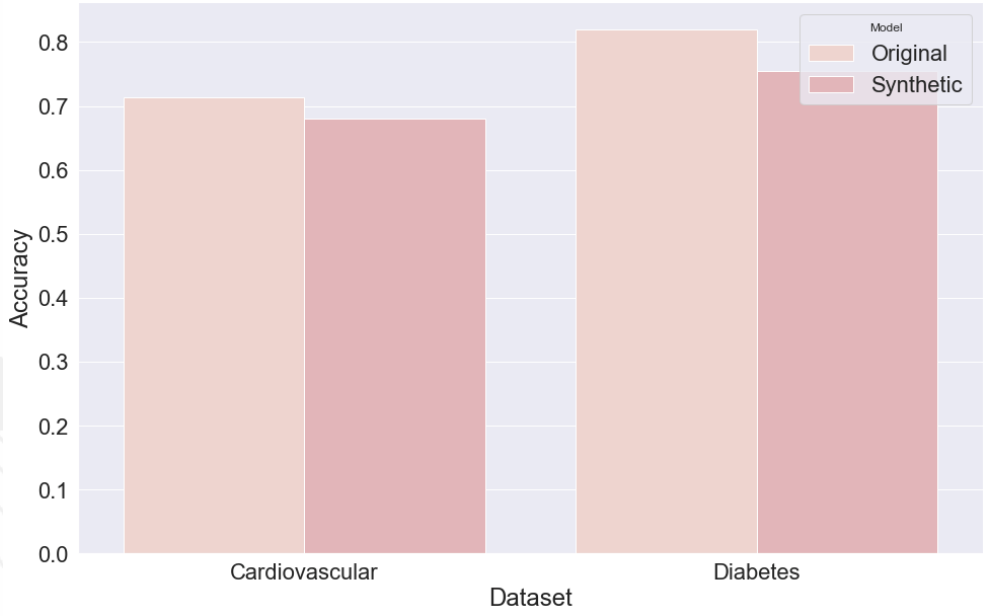
Random Forest Regression model

- Sample configurations for model trained with and without synthetic data
  - 500 trees
  - Square root splits



Result

- Model developed solely on synthetic data **retains high accuracy on real data** coinciding with results from other studies



… samlet i standardiserede rapporter for hhv. datakvalitet og reidentifikationsrisiko

Due Diligence Checklist
Reasoning of the business case for the project, the five safes framework (ethical considerations, data protection etc.) and governance structure.

Privacy Assurance Report
Unique matches and too close to original data removed and group inference reported. Bias is reported inform of plots and measurements of attributes with discrimination risks.

Utility Assessment Report
Evaluation of the synthetic data's utility by comparisons to original data using univariate, bivariate, multivariate analyses, bias and distinguishability.

Summary Report
Settings used in the pipeline, name and email of the user/data scientist and general outcomes of privacy assurance and utility assessment reported.

Privacy Report

Summary

This report analyses the effectiveness of an anonymisation of a dataset and the residual risks that follows.The anonymised dataset aimed for sharing in this report consists of 20000 anonymised individuals. This anonymised dataset origins from a population of 395933 where 277153 has been sampled for anonymising and sharing.

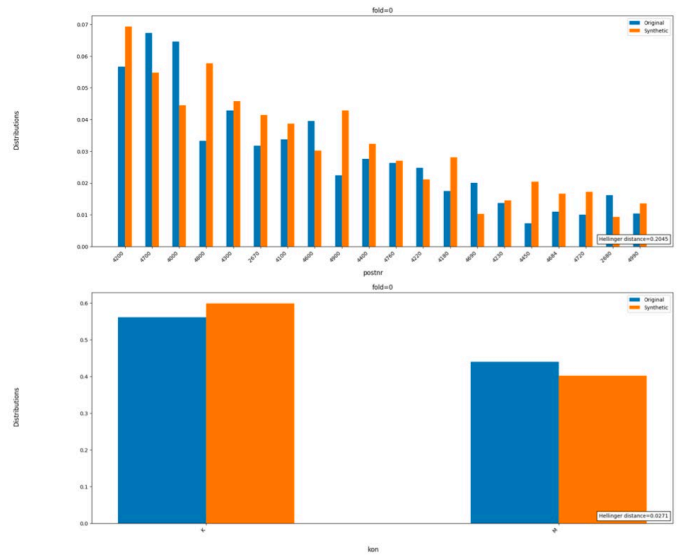
Identical matches: 0.0%
Combinations of attacks evaluated: 2751
Re-identification risk: 2.5%

Effectiveness of anonymisation: Effective

For in-depth details regarding the anonymisation, please see the following chapter.

Utility Report

Univariate comparision of discrete values



Vi har udviklet en metode til en standardiseret måling af reidentifikationsrisiko

Deloittes kvantitative risikovurdering operationaliserer og udbygger den nyeste forskning i et standardiseret rapportformat. Modsat konventionel anonymisering kan man i syntetiske data ikke triangulere sig tilbage til de ægte individer – men der er stadig en risiko for, at man med en vis statistisk sandsynlighed kan udlede noget. Derfor skal data altid kontrolleres.

Privacy Report

Summary

This report analyses the effectiveness of an anonymisation of a dataset and the residual risks that follows.The anonymised dataset aimed for sharing in this report consists of 20000 anonymised individuals. This anonymised dataset origins from a population of 395933 where 277153 has been sampled for anonymising and sharing.

Identical matches: 0.0%
Combinations of attacks evaluated: 2751
Re-identification risk: 2.5%

Effectiveness of anonymisation: Effective

For in-depth details regarding the anonymisation, please see the following chapter.

Reidentification risks

1. Identical matches

The number of synthetic individuals fully replicated from the origin population including quasi-identifiers and sensitive attributes.

The synthetic population contains 0 datapoints fully replicated datapoints from the origin population.

2. Indirect reidentification risk

The risk of learning sensitive information from a synthetic population from quasi-identifier matches in the origin population.

The synthetic population has an estimated reidentification risk of 0.01 passing the safety threshold of 0.09.

Near-identical match risks

Near-identical matches can also impose a reidentification risk, however, with significantly higher uncertainty of learning new information due to the increased plausible deniability from the adversary cannot match on all quasi-identifiers and must create assumptions. Albeit outliers can be hidden in near-identical matches and through inference an adversary can potentially learn sensitive information about a uniquely outlying individual. Therefore, additional analyses are be conducted to support the decision in sharing data using measures of distance between all individuals' attributes to identify near-identical matches.

Critical replicated synthetic outlier evaluation

Evaluation of potential critical synthetic outliers being generated too close to a unique individual in the origin population.

The synthetic dataset contains 10 synthetic datapoints critically close to the unique outliers in the origin population.

Figure 2 Critical synthetic outlier evaluation showing NNDR on x-axis and DCR on y-axis to the left. Synthetic individuals (green cross) in the critical region (red) should be evaluated. The right plot shows synthetic population (green cross) compared with origin population (blue dots).

Rapporten simulerer flere tusinde angreb på de anonymiserede data og giver en samlet, sammenlignelig metrik for risikoen. Eksemplet til højre viser en reidentifikationsrisiko på 2,5% – forstået som risikoen for, at en angriber kan udlede noget om et individ i datasættet. Vel at mærke under meget konservative antagelser: Det antages jf. GDPR, at angriberen allerede har adgang til de ægte data og kan bruge den viden til at vurdere, hvilke af de mange angreb, der har størst chance for at afsløre information. Til sammenligning er risikoen med K=5-anonymisering hele 76% i eksemplet vist ovenfor. EMA har publiceret en guideline vedr. deling af data fra kliniske forsøg med en tærskelværdi på 9%.

Hvor gode kan syntetiske data blive – og hvor bliver man forpligtet til at bruge dem?

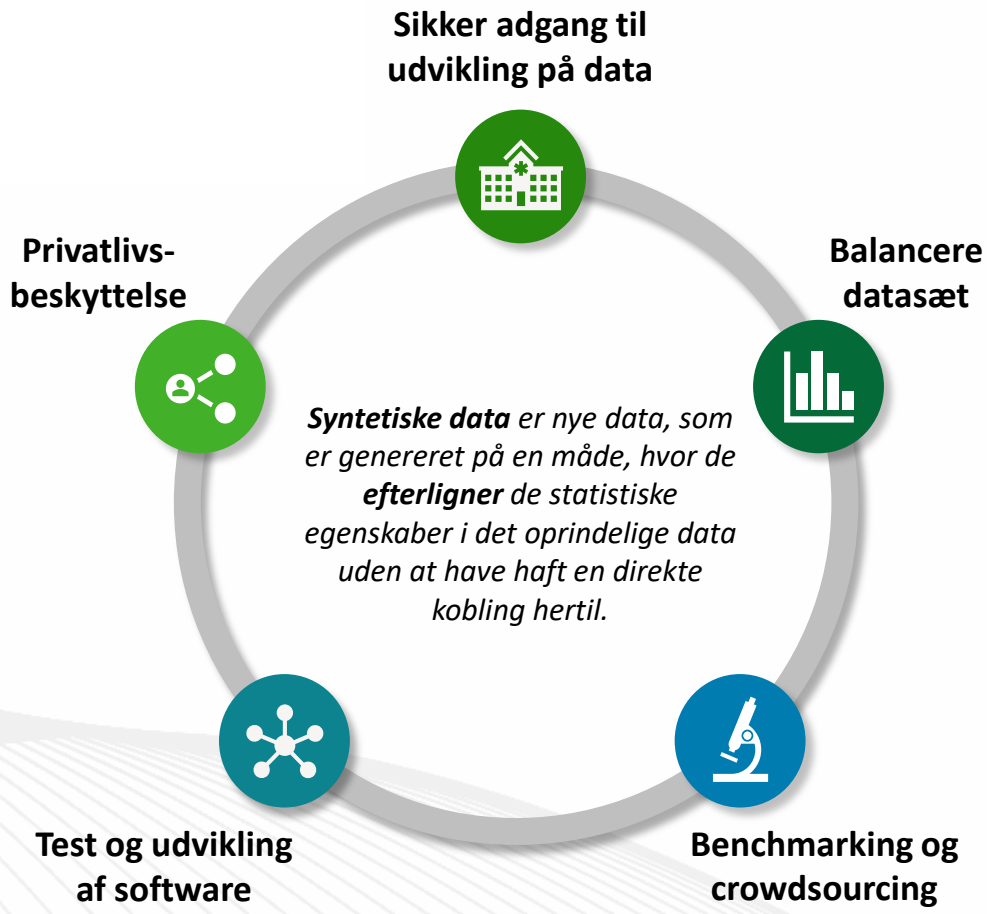
De kan blive ret gode. Og man bliver forpligtet til at bruge dem, hvor dataminimeringskravet tilsiger det. Herudover kan der være en række andre grunde til at arbejde med syntetiske data.

Hvorfor: De tre D'er

- Dataminimering
- Datadeling
- Datadiversitet

Hvor: Hovedtyper

- Test og udvikling (privacy og mere effektive tests)
- Sikker adgang til udvikling på følsomme data
- Booste datadiversitet og balancere datasæt
- Muliggøre benchmarking og crowdsourcing



Hvor gode kan syntetiske sundhedsdata blive – og hvor bliver man forpligtet til at bruge dem?

Spørgsmål?