ELAN-VIPP

Het *ELAN Virtuele Patiënten en Populatie* project

Onderweg naar een digitale tweeling mét ELAN?

Health Campus Den Haag

Nederlandstalig Platform Survey Onderzoek (NPSO) bijeenkomst Synthetische data, 12/10/2023

Prof. Dr. Marco Spruit (LUMC/LIACS)

Over… Marco Spruit

Pretpakket VWO
Propedeuse Nederlands
Propedeuse Muziekwetenschap
Bovenbouwstudie α-Informatica

Als Ontwikkelaar

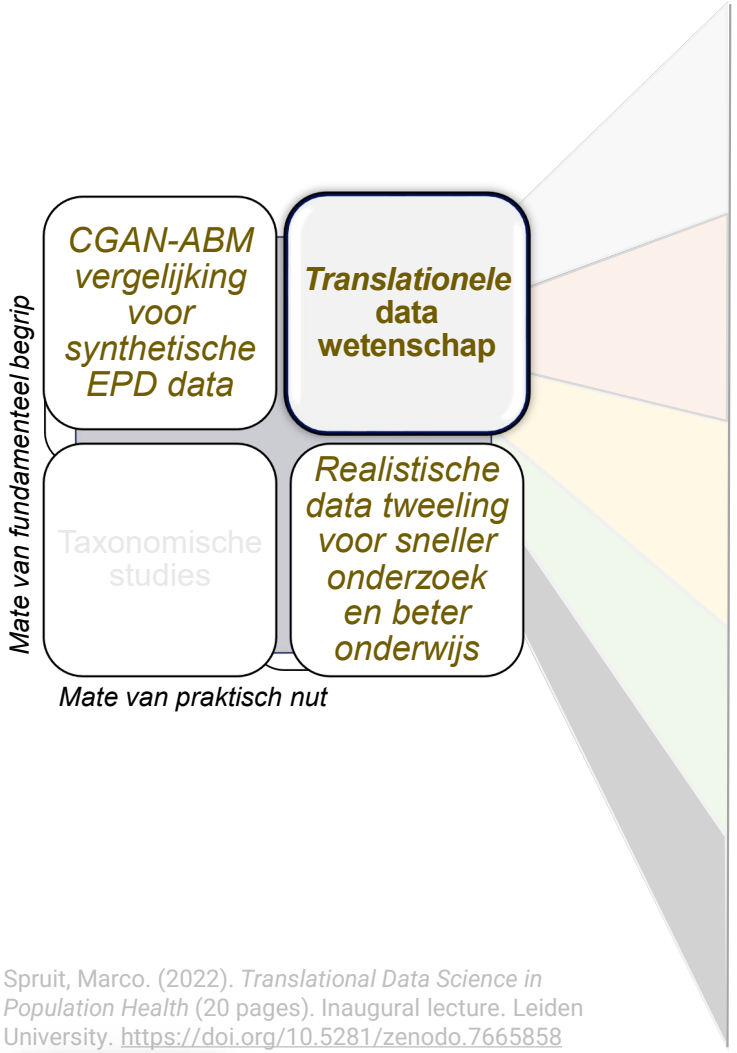
- 1993
  - *Information Retrieval* programmeur
    - ZyLAB Europe BV
- 1995
  - *Big Data* systeemontwikkelaar
    - Koninklijke Marine
- 1997
  - Productsoftware ontwikkelaar/ondernemer
    - Insertable Objects, Wizzer BV

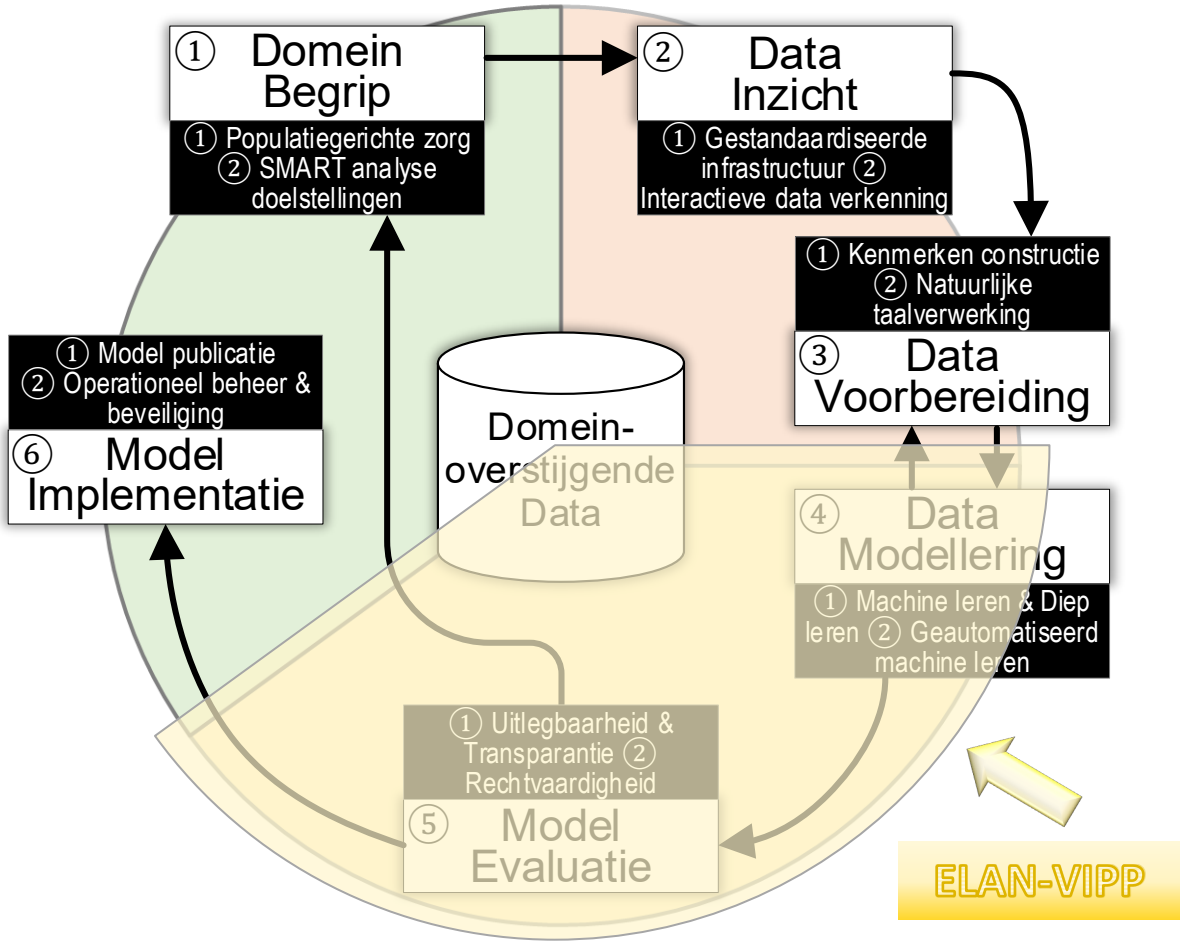
Als Wetenschapper

- 2003
  - OiO in Computationele Linguïstiek
    - Universiteit van Amsterdam (UvA)
- 2007
  - Universitair (hoofd)docent Informatiekunde
    - Universiteit Utrecht >> ADS Lab (2015)
- 2020
  - **Hoogleraar *Advanced Data Science in Population Health***
    - **LUMC/LIACS** @ Universiteit Leiden (HCDH)
    - PH Living Lab, CAIRELab, TDS Lab, SIG HDS

Health Campus Den Haag

Over… *Translationele* Datawetenschap





Spruit, Marco. (2022). *Translational Data Science in Population Health* (20 pages). Inaugural lecture. Leiden University. https://doi.org/10.5281/zenodo.7665858

Data Techniek | Data Analyse | e-Health Implementatie

Health Campus Den Haag

Over… Health Campus Den Haag

1. Verkleinen gezondheidsverschillen
2. Duurzame benadering
3. Breed gezondheidsperspectief

Leids Universitair Medisch Centrum

Populatiegericht zorgonderzoek

DE HAAGSE HOGESCHOOL

Zorg/Data-onderzoek

HagaZiekenhuis

Ziekenhuis data

Reinier de Graaf

Ziekenhuis data

parnassia groep
Specialist in geestelijke gezondheid

Mentale gezondheid

Universiteit Leiden

Data science & AI

H+ MC

Ziekenhuis data

VREDE EN RECHT
Den Haag

oa WMO data

+

hadoks

Huisarts data

GGD Haaglanden

oa Vaccinatie data

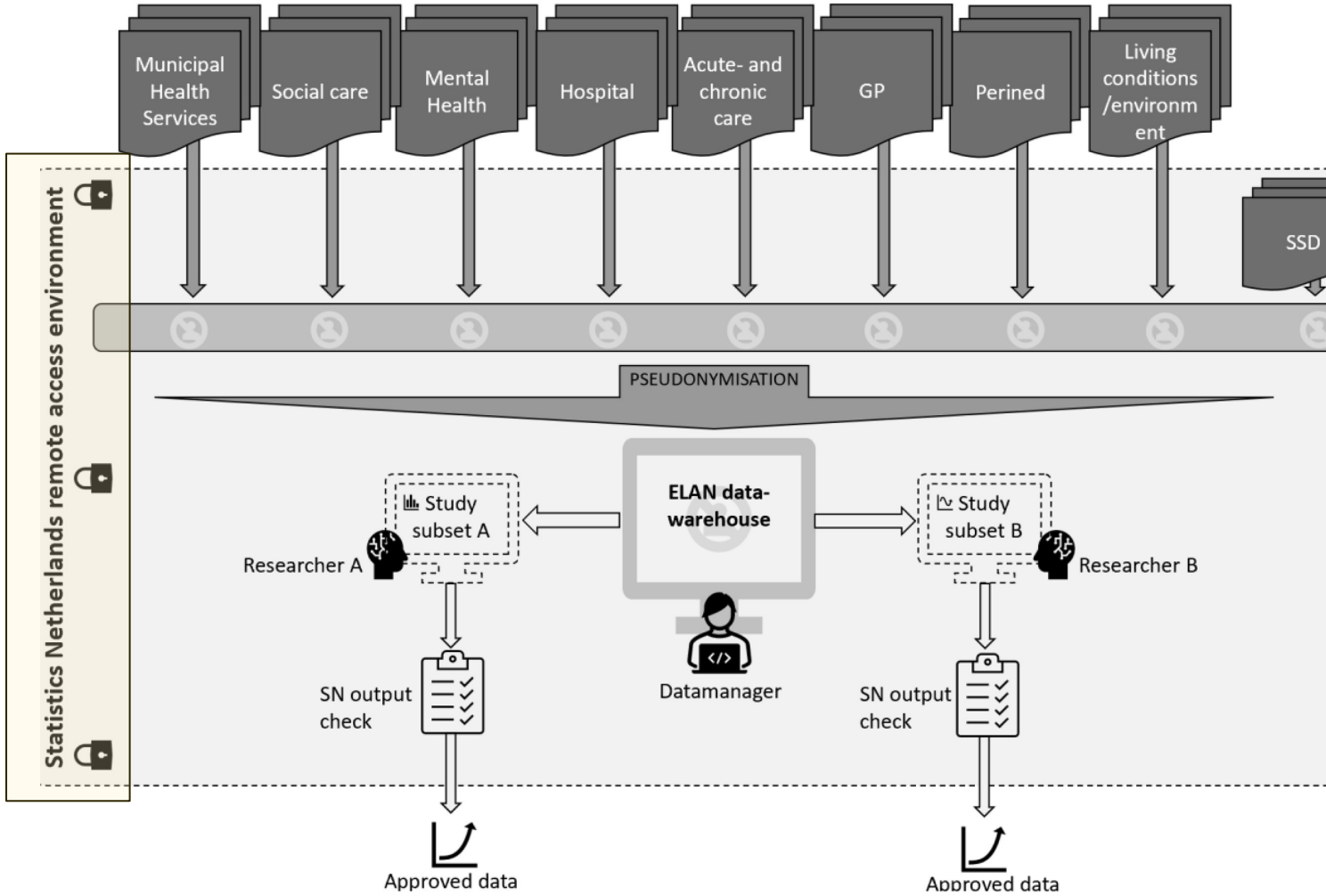
Health Campus Den Haag

Over… ELAN

- Het *"Extramurale LUMC Academische Netwerk"* (ELAN): CBS-RA

1. trans-disciplinair
2. domein-overstijgend
3. regionale data-infrastructuur

- inclusief bijbehorende beleidsstructuur, dat gezondheidsbeleid en onderzoek ondersteunt
- Totaal aantal patiënten:
  - ~1,2 miljoen records



Health Campus Den Haag

Ardesch,F., Meulendijk,M., Kist,J., Vos,R., Vos,H., Kiefte-de Jong,J., Spruit,M., Bruijnzeels,M., Bussemaker,J., Numans,M., & Struijs,J. (2023). A data-driven population health management approach: The Extramural LUMC Academic Network data infrastructure. *Health Policy, 132*, 104769. [online]

Over… Synthetische data → Waarvoor een oplossing?

Nu… ELAN via CBS-RA

1. Tijdrovend proces, vanwege
2. Zeer strenge beveiliging, met
3. Veel gebruiksbeperkingen, inclusief
4. Beperkingen tbv CDSS in praktijk;
5. ‘Enkel’ tabulaire, *small* data
6. Niet voor onderwijs doeleinden

USP

- Unieke rijkheid aan populatiedata op individueel niveau, dus volop kansen voor nieuwe wetenschappelijke en beleidsmatige inzichten

Toekomst… synthetische ELAN ‘twin’

1. Factor 10(?) sneller onderzoek → innovatiekracht
2. Semi-automatische ‘bewijsvoering’ via evaluatiemetrieken?
3. Open sourcing/access mogelijk? → Innovatiekracht
4. Vrije modelontwikkeling mogelijk
5. Multimodale data → Rijker dus realistischer en waarheidsgetrouwer
6. Onderwijs verbetert en verdiept aanzienlijk → Aantrekkelijker en leerzamer voor studenten

Health Campus Den Haag

Health Campus

Den Haag

Synthetische data → Onderwijs innovatie

LUMC Project Raamplan Implementatie Artsopleiding (PRIMA) 2020

NFU opdracht: *"Ontwikkelingen in technologie, waaronder informatie- en communicatietechnologie en kunstmatige intelligentie, leiden tot nieuwe mogelijkheden in preventie, diagnostiek en behandeling"*

Kans! AI-Tech met één revisie op constructieve en coherente wijze grondig integreren

VIPP

Marco Spruit Karoly Szuhai

Synthetische data: *PRIMA 2020* Onderwijs advies

- VIPP!
- Vergroot het enthousiasme van zowel studenten als docenten voor Technologie, AI & Big Data
- Door een EPD-afgeleide dataset te ontwikkelen van ‘zo-goed-als-echte’ patiënten en lokale bevolking

- 2021-: VIPP in PHM startvak
- 2023-2024: AI week in j1 GNK

Sterktes

- Geeft Impliciet een **unieke Leidse smaak** aan de GNK programma's
- Maakt **zo-goed-als-echte** data-analyses mogelijk (*i.e.* medisch betekenisvolle uitkomsten)
- Gemeenschappelijk referentie-kader voor casuïstiek om **alle cursussen** te helpen verbinden

Zwaktes

- Eenmalige **ontwikkeling** is resource-intensief, vergt maatwerk (*i.e.* promovendus)
- Reeds in gebruik zijnde open datasets (*e.g.* GWAS) vereisen handmatig **koppelen** aan VIPP
- Beleid tbv draagvlak behoud & medewerking van **data providers** (LUMC, GGDH)

Kansen

- Realistische data-analyses in onderwijs/onderzoek kunnen worden **getest** zonder METC
- De grote behoefte aan zo-goed-als-echte zorgdatasets zal VIPP tot een nationale **gouden standaard** maken.
- De master **PHM** wil VIPP graag verweven in haar curriculum.

Dreigingen

- Vereist formeel **bewijs** dat echte patiënten nooit (opnieuw) kunnen worden geïdentificeerd
- AVG-gerelateerd **vertrouwen** garanderen, gegeven lastig uit te leggen bewijzen (XAI)
- Hoeveel inspanning is er nodig voorbij "**plausibele** ontkenning" (LaPlace-ruis)?

Health Campus Den Haag

Health Campus Den Haag

Synthetische data → Technieken

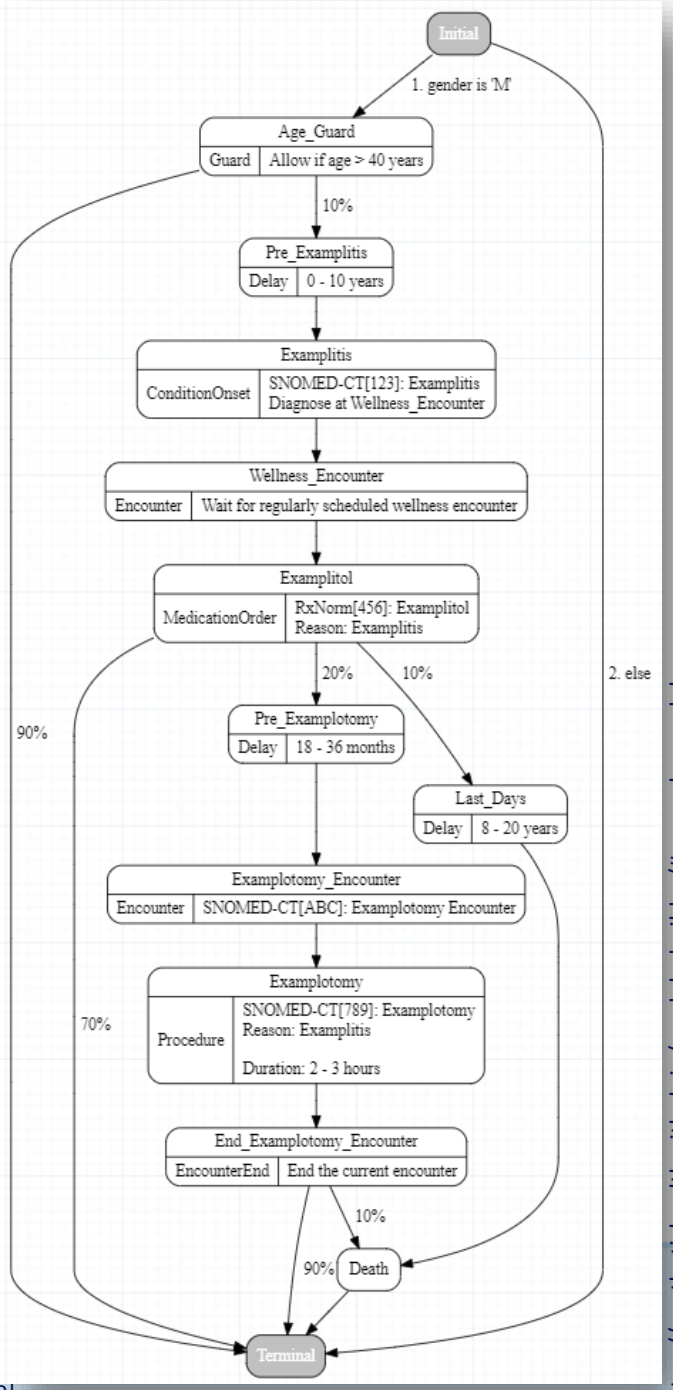
Drie benaderingen voor onderzoek en toepassingen

A. **Regel-gebaseerd → Agent-Based Modelling (ABM)**
B. Neuraal netwerk-gebaseerd → Generative Adversial Networks (GAN)
C. Off-the-shelf → Syntho

Ammar Faiq Jim Achterberg Marcel Haas Hugo Álvarez

A. Agent-Based Modelling (ABM)

- *Synthea:* "an open-source Synthetic Patient Population Simulator"
  - **Doel:** *"[…] simulate the lifespans of synthetic patients, modeling the 10 most frequent reasons for primary care encounters and the 10 chronic conditions with the highest morbidity* ***in the United States****"*
  - ***Benadering:*** Synthetische data (bewust) **NIET** obv originele EPD data (onwenselijk, onmogelijk), maar obv publiekelijk beschikbare data



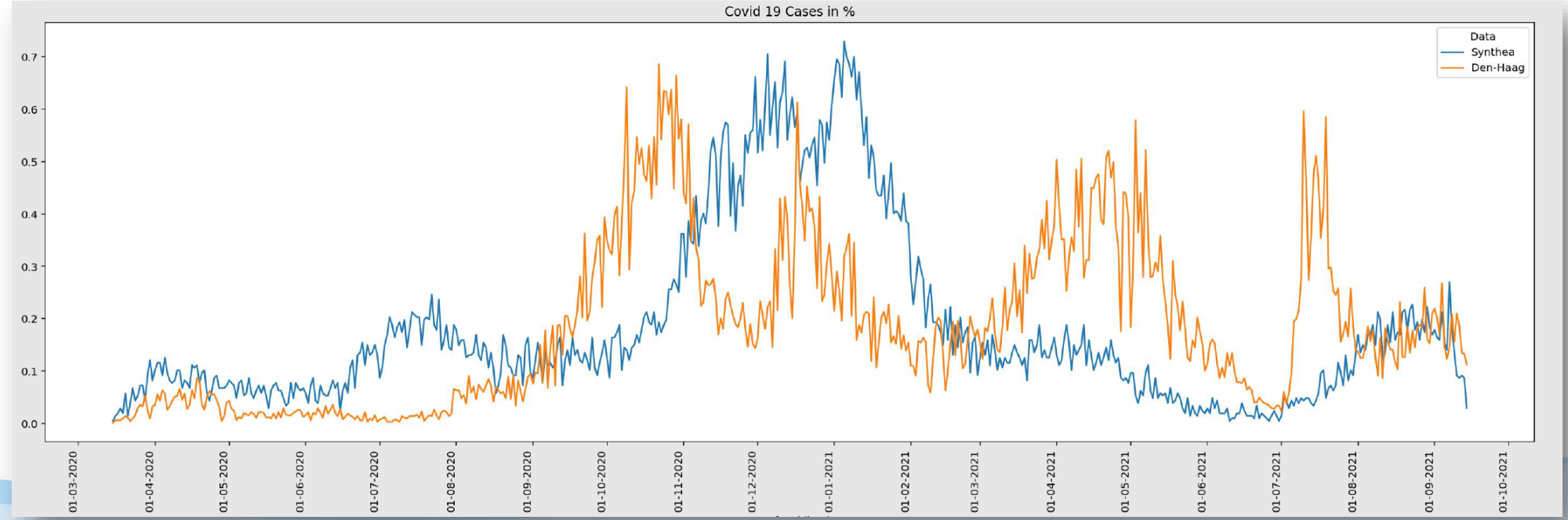
https://synthetichealth.github.io/module-builder/#example_module

Health Campus Den Haag

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., ... & McLachlan, S. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association, 25(*3), 230-238. [online]

Voorbeeld: Synthea voor PHM

- MSc *Fundamentals of Population Health Management* vakdoel
  - Praktijkopdracht data-analyse COVID19 prevalentie voorspelmodel
  - Vereist integrale analyse van medische, sociale, regionale, etc data

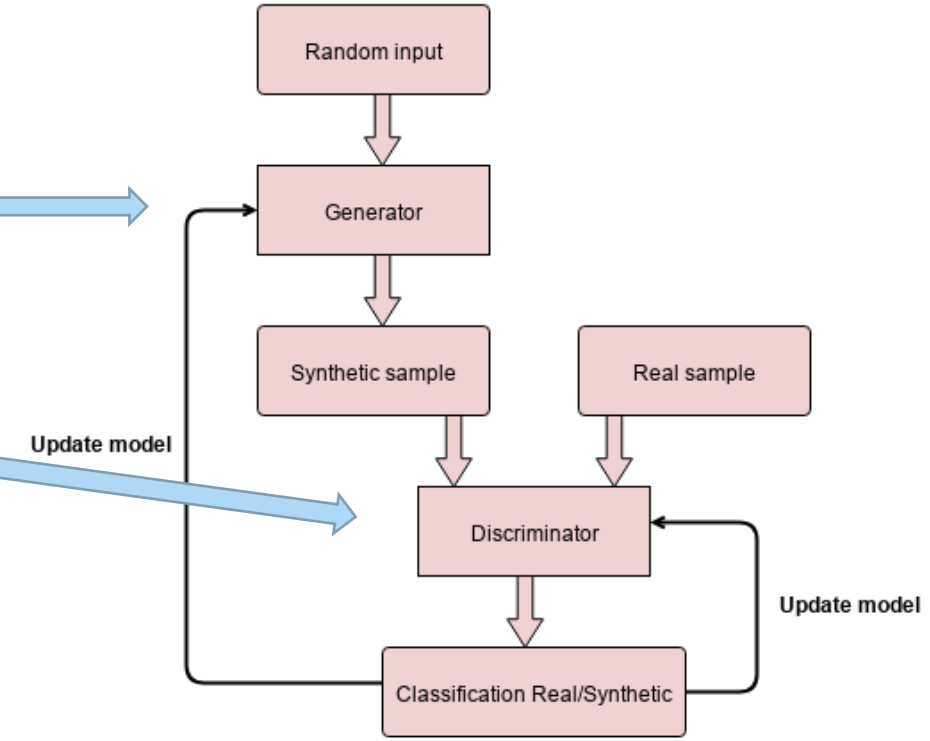


Health Campus Den Haag

B. Generative Adversial Networks (GAN)

- **GAN:** een spel tussen twee spelers: de Generator en de Discriminator.
  - trainings*voorbeelden* → waarschijnlijkheids*distributie*
- **Generator:** neuraal netwerk dat realistische synthetische data genereert uit willekeurige input.
  - Leert alleen van interactie met de discriminator
- **Discriminator:** neuraal netwerk dat de originele data en de synthetische data probeert te onderscheiden.
  - Ontvangt zowel originele data als synthetische data geproduceerd door de generator.
- De generator is succesvol als het adequaat de verdeling van de originele data leert via interactie met de discriminator.

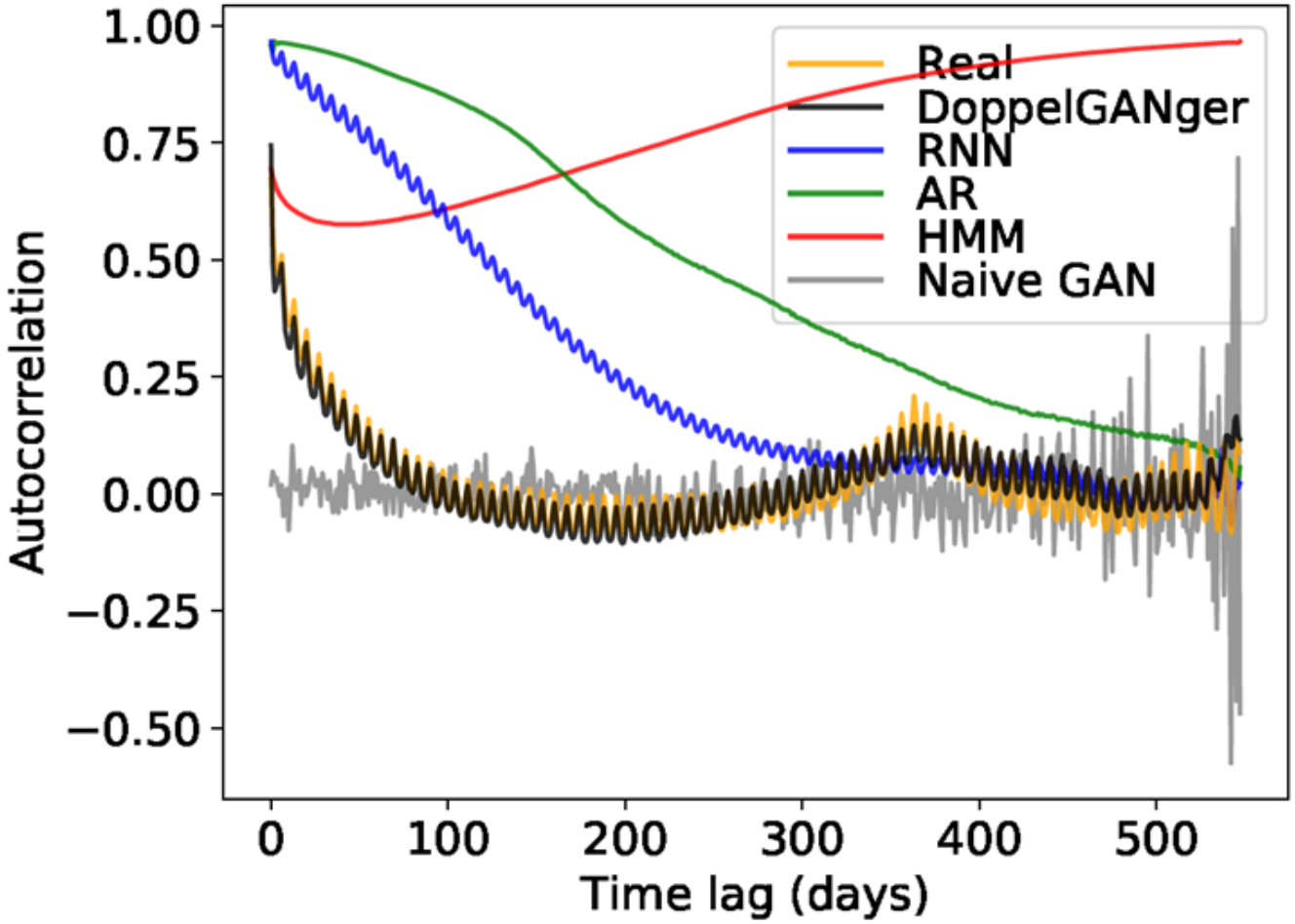
- GAN trainingsproces



Health Campus Den Haag

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139-144. [online] → 60K citaties!

Voorbeeld: DoppelGANger

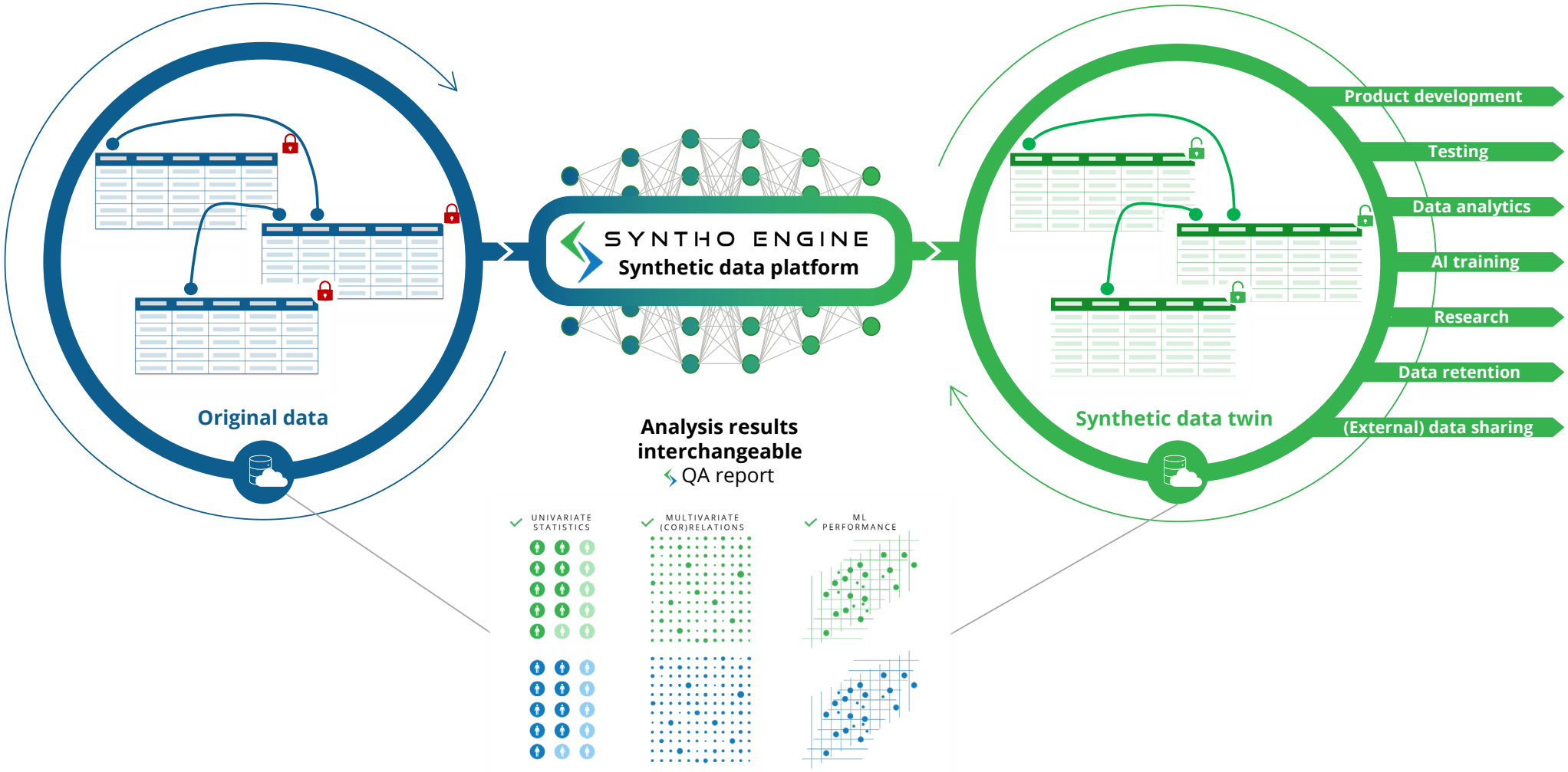


- 6 synthetische technieken voor het simuleren van het dagelijks bezoek van Wikipedia pagina's.
  - Alleen DoppelGANger pikt zowel wekelijkse als jaarlijkse patronne op ("Conditional GAN")

Health Campus Den Haag

Lin, Z., Jain, A., Wang, C., Fanti, G., & Sekar, V. (2020, October). Using gans for sharing networked time series data: Challenges, initial promise, and open questions. In *Proceedings of the ACM Internet Measurement Conference* (pp. 464-483). [online]

C. Off-the-shelf (*e.g.* Syntho *platform*)



Health Campus Den Haag

https://www.syntho.ai/syntho-engine/

Health Campus Den Haag

Status??

Uitdagingen??

Ambities??

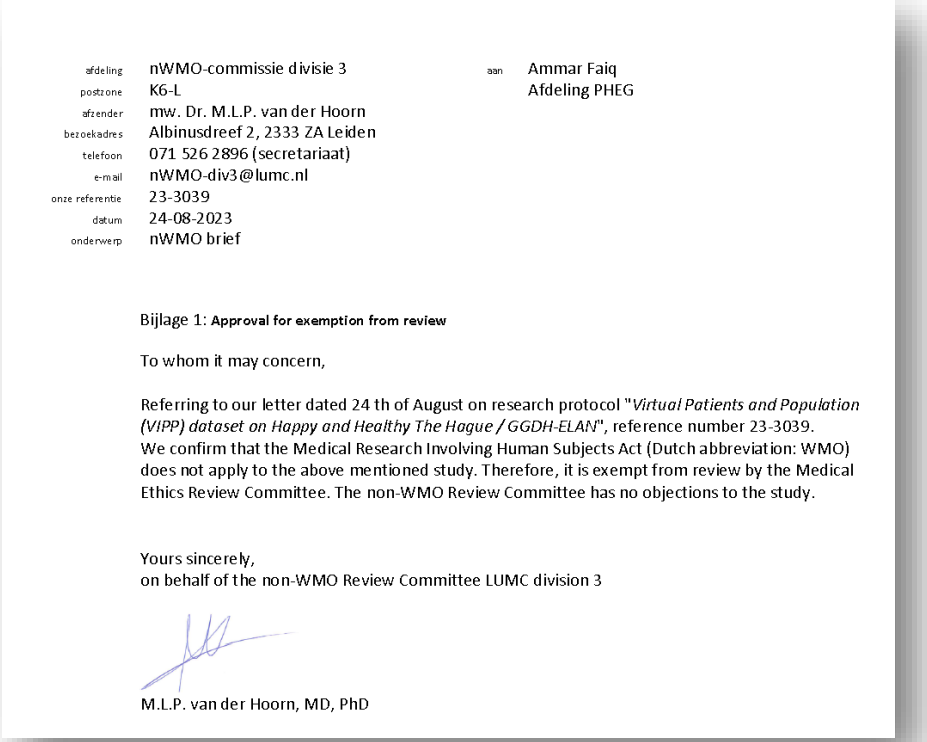
Synthetische data: Huidige status LUMC

1. VIPP pilot studie met CBS en Syntho
   - In speciale CBS IT-omgeving
   - Doel: reproductie Kist et al (2023)

     "SCORE2 cardiovascular risk prediction models in an ethnic and socioeconomic diverse population in the Netherlands: an external validation study"
2. Jaarlijkse PHM startvak opdracht
3. Begeleidingsethiek workshop
4. Project funding!
   - NWO OSF SENSYN: 50K, 8 months, 15/1/24

     "Making sensitive data reusable through synthetic data generation"
   - Horizon Europe INSAFEDARE: 570K, 3j, 1/11/23

     "Innovative applications of assessment and assurance of data and synthetic data for regulatory decision support"



afdeling	nWMO-commissie divisie 3	aan	Ammar Faiq
postzone	K6-L		Afdeling PHEG
afzender	mw. Dr. M.L.P. van der Hoorn		
bezoekadres	Albinusdreef 2, 2333 ZA Leiden		
telefoon	071 526 2896 (secretariaat)		
e-mail	nWMO-div3@lumc.nl		
onze referentie	23-3039		
datum	24-08-2023		
onderwerp	nWMO brief		

Bijlage 1: **Approval for exemption from review**

To whom it may concern,

Referring to our letter dated 24 th of August on research protocol "*Virtual Patients and Population (VIPP) dataset on Happy and Healthy The Hague / GGDH-ELAN*", reference number 23-3039.
We confirm that the Medical Research Involving Human Subjects Act (Dutch abbreviation: WMO) does not apply to the above mentioned study. Therefore, it is exempt from review by the Medical Ethics Review Committee. The non-WMO Review Committee has no objections to the study.

Yours sincerely,
on behalf of the non-WMO Review Committee LUMC division 3

M.L.P. van der Hoorn, MD, PhD

Health Campus Den Haag

Synthetische data: Enkele uitdagingen

1. Testen van *Goodness-of-Fit*: Zijn twee distributies gelijk of niet?
- Wat zijn bruikbare betrouwbare evaluatiemetrieken?
- Vergelijken via trainen voorspelmodel op synthetisch vs origineel
  - 'Fidelity'/*Gelijkenis*: Gelijkenis van distributies tov originele data
  - 'Utility'/*Nut*: Nut bij gebruik in praktijktaken tov originele data
  - 'Privacy risk'/*Onthullingsrisico*: Nut bij gebruik in praktijktaken tov originele data

2. *Multimodale* generatie → EPD data zijn zeer heterogeen van aard
   - *VRAAG: In hoeverre zijn LLM's bruikbaar voor genereren van klinische teksten?*
3. *Niet-technische* aspecten gladstrijken…

Health Campus Den Haag

Friedman, J. H. (2003). On multivariate goodness–of–fit and two–sample testing. *Statistical Problems in Particle Physics, Astrophysics, and Cosmology, 1*, 311. [online]

Synthetische data: Ambities

1. Nut aantonen bij gebruik in praktijktaken tov originele data
   Vertrouwen verkrijgen en behouden bij data leveranciers
   - Via gevalideerde metrieken?
   - Via afgeslagen 'Attribute Inference Attacks (AIA)'?
     - Voorspelmodel trainen op synthetische data, om gevoelige attributen uit originele test-data te achterhalen
2. *"Gouden standaard"* ontwikkelen via reeds gepubliceerde ELAN analyses
   - Niet enkel vergelijken met originele data, ook met origineel voorspelmodel
3. Meer data generatie technieken
   - *m.n.* Conditionele Probabilistische Autoregressieve Netwerken (PAN)
4. Meer datatypen generatie tbv meer realistische use cases
   - *m.n.* EPD tekstdata, ook EPD beelddata

Health Campus Den Haag

Health Campus Den Haag

Tot slot: Ethische aspecten

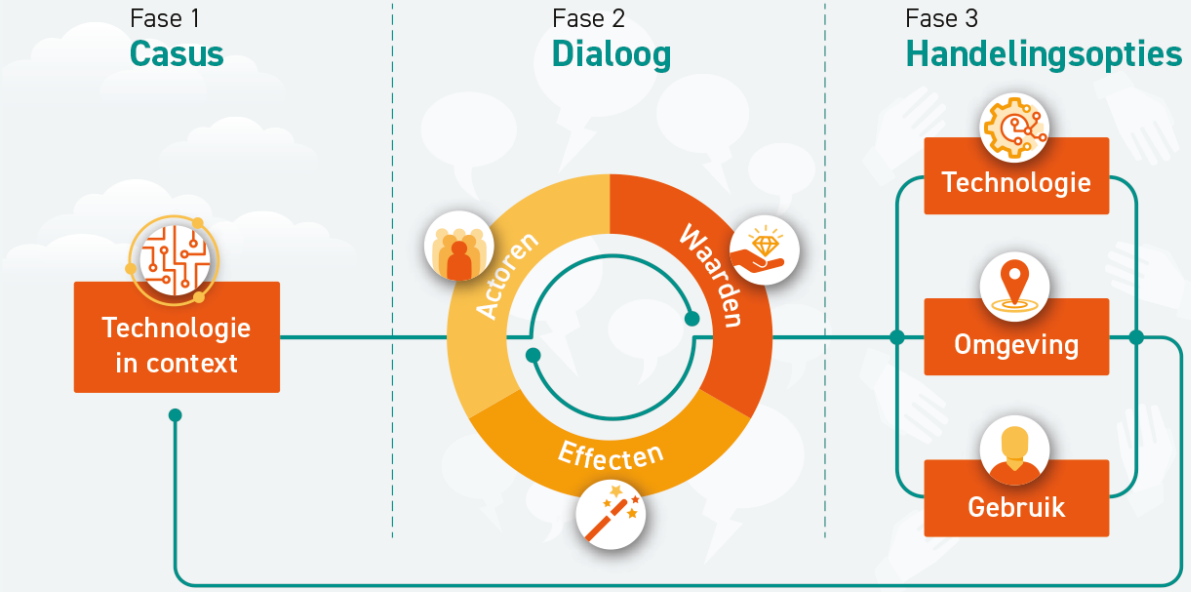
nav Workshop Aanpak Begeleidingsethiek

25 augustus 2023 olv Dr. André Krom (LUMC/R&E) met 14 stakeholders

- Welke effecten, waarden en actoren worden genoemd?
- Welke handelingsopties zijn er om deze toepassing te verbeteren?

Actoren: Wie is betrokken of wordt er door geraakt?

- 35 actoren geïdentificeerd: Banken (in relatie tot verzekeraars), Beleidsmakers (o.a. ministerie van VWS, Europese Commissie), Burgers, CBS, Data bezitters (bijv. patiënten, ziekenhuizen), Data leveranciers, Gemeenten, Toezichthouder, Verzekeraars, Zorgverleners, …



Health Campus Den Haag

Effecten: Welke positieve en negatieve effecten heeft invoering/gebruik?

Positieve effecten (33)

1. Nieuwe innovatieve toepassingen (K)
2. Versnelde innovatie van AI (OZ)
3. Betere kwaliteit van zorg (K)
4. Regionale samenwerking
5. Data compleet maken (O/OZ/K)
6. Gevaar tot data lek vermindert
7. Domein overstijgend werken
8. Onderzoek doen zonder beperkingen van privacy
9. Studenten kunnen met 'echte' data leren werken (O)

Negatieve effecten (21)

1. Toegenomen angst voor onthullingsrisico (K)
2. Gevaar tot datalek (risico)
3. Kan leiden tot bias (onwaarheden in dataset)
4. Ontstaan van verkeerde hypotheses
5. Klinische data kloppen niet met praktijk
6. Kan je onderscheid maken tussen echte en neppe data?
7. Gaat vaak om individuen en daar is synthetische data minder geschikt voor (K)
8. Meer technology push en niet meer beredeneerd vanuit behoefte (OZ)

Health Campus Den Haag

Waarden: Welke waarden vinden we belangrijk voor deze geïdentificeerde effecten?

Effect	Waarden
Gevaar tot data lek vermindert (+)	Vertrouwelijkheid (*confidentiality*) Privacy Weerlegbaarheid (*plausible deniability*) Institutioneel vertrouwen Reputatie Zelfbeschikking
Betere zorguitkomsten (K) (+)	Gezondheid Betaalbaarheid Werkplezier Kwaliteit van leven

Effect	Waarden
Toegenomen angst voor onthullingsrisico (K) (-)	Vertrouwen Betrouwbaarheid Uitlegbaarheid Transparantie Rechtvaardigheid (*justice*) Inclusiviteit Relaties
Hoe doe je onderzoek o.b.v. synthetische data → vertrouw je conclusies uit synthetische data? (-)	Betrouwbaarheid (van uitkomsten) Validiteit (van onderzoeksuitkomsten) Vertrouwen (in het onderzoeksproces)

Health Campus Den Haag

Handelingsopties: Hoe deze effecten te borgen mbt Technologie [1/3]

EFFECTEN	HANDELINGSOPTIES
1. Toegenomen angst voor onthullingsrisico (K)	• Fysieke presentatie van data (graphics) • De stappen kunnen laten zien/ documenteren • Gebruik een methode die geen echte data gebruikt (rule-based model) • Vereenvoudigde kopie van het proces van de technologie, inclusief media
2. Hoe doe je onderzoek o.b.v. synthetische data -> vertrouw je conclusies uit synthetische data	• Onderzoek doen naar evaluatie van metrieken om dit te kwantificeren / karakteriseren
3. Gevaar tot data lek vermindert	• Tijdens ontwikkelproces reeds juridische borging (AVG-compliant) en informatiebeveiliging (technisch)

Health Campus Den Haag

Handelingsopties: Hoe deze effecten te borgen mbt de Omgeving [2/3]

EFFECTEN	HANDELINGSOPTIES
1. Toegenomen angst voor onthullingsrisico (K)	• Veranderingen documenteren, documenteren hoe het model werkt en hierover afspraken maken • Proces documenteren → wat gebeurt er wanneer? • Protocollen opstellen die beschrijven hoe
2. Hoe doe je onderzoek o.b.v. synthetische data -> vertrouw je conclusies uit synthetische data	• Integreren in opleidingen t.b.v. toekomstige professionals
3. Gevaar tot data lek vermindert	• Ontwikkelen van gedragscodes, door het veld gedragen (en toetsen)

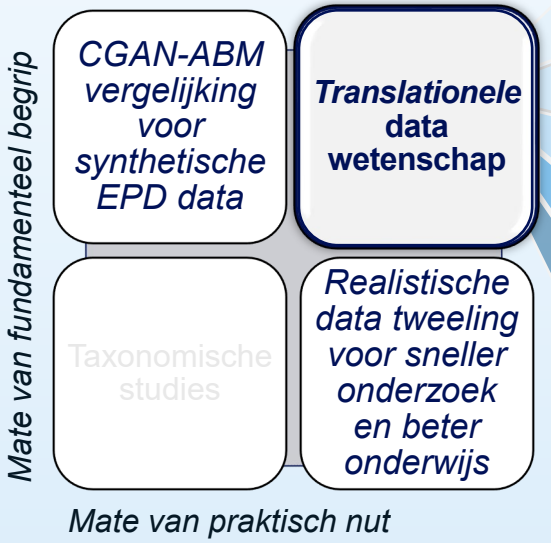
Health Campus Den Haag

Handelingsopties: Hoe deze effecten te borgen mbt de Mens [3/3]

EFFECTEN	HANDELINGSOPTIES
1. Toegenomen angst voor onthullingsrisico (K)	• Training • Workshop over veiligheid • Bewustzijn creëren (*leniency*) • “Hoe te publiceren” • Wetenschapsjournalisten die de technologie kunnen vertalen naar de gemiddelde burger
2. Hoe doe je onderzoek o.b.v. synthetische data -> vertrouw je conclusies uit synthetische data	• Communiceren nodig om bewustzijn/ competentie van publiek t.a.v. de technologie te stimuleren • Wetenschapscommunicatie
3. Gevaar tot data lek vermindert	• Responsible disclosure • Responsible stewardship • Aanspreken op en ‘opvoeden’ t.b.v. van morele verantwoordelijkheid

Health Campus Den Haag

ELAN-VIPP: Virtuele Patiënten en Populatie dataset



Contact: M.R.Spruit@lumc.nl

Health Campus Den Haag