Integrating population-wide molecular and clinical data with deep learning

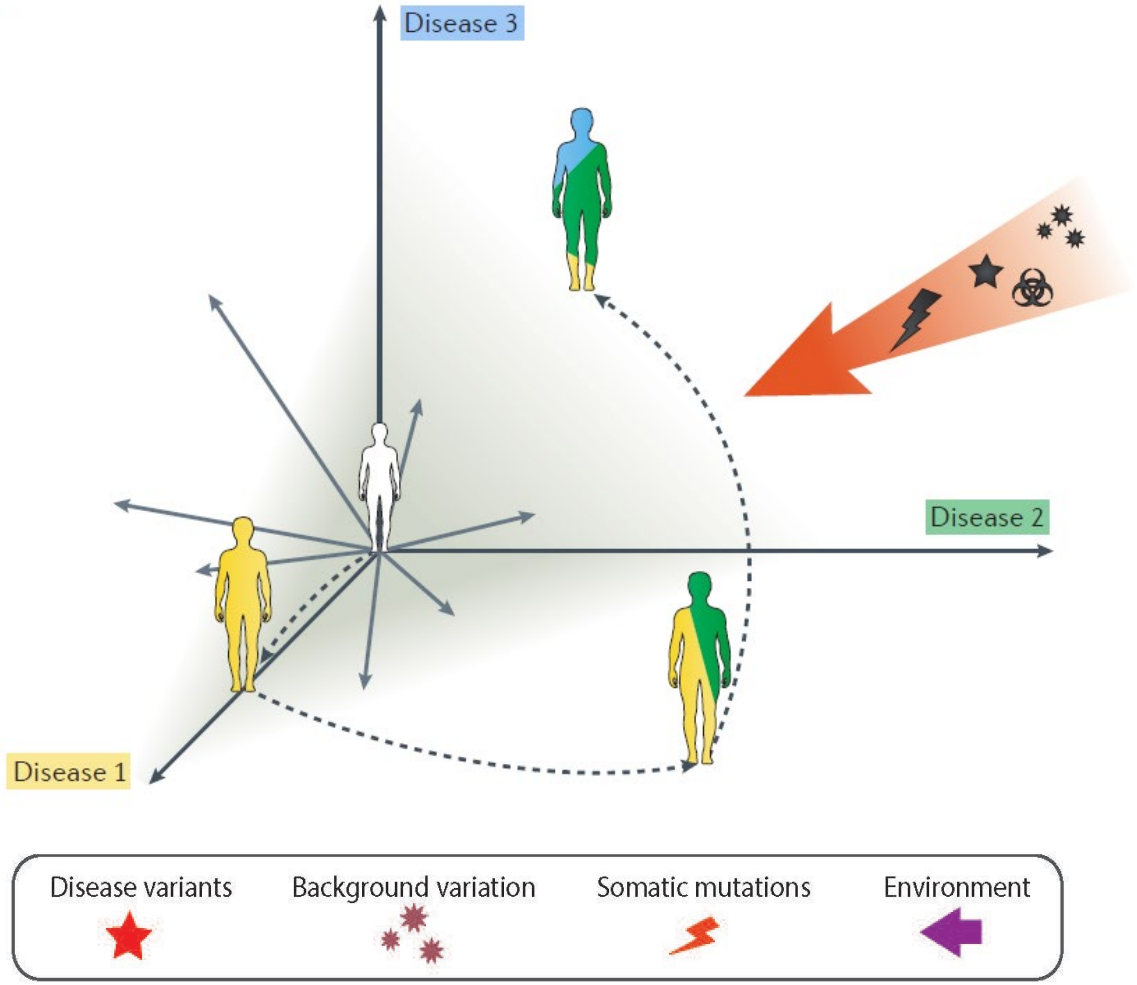
– Taking advantage of population-wide health data from the Nordic ecosystem

Søren Brunak

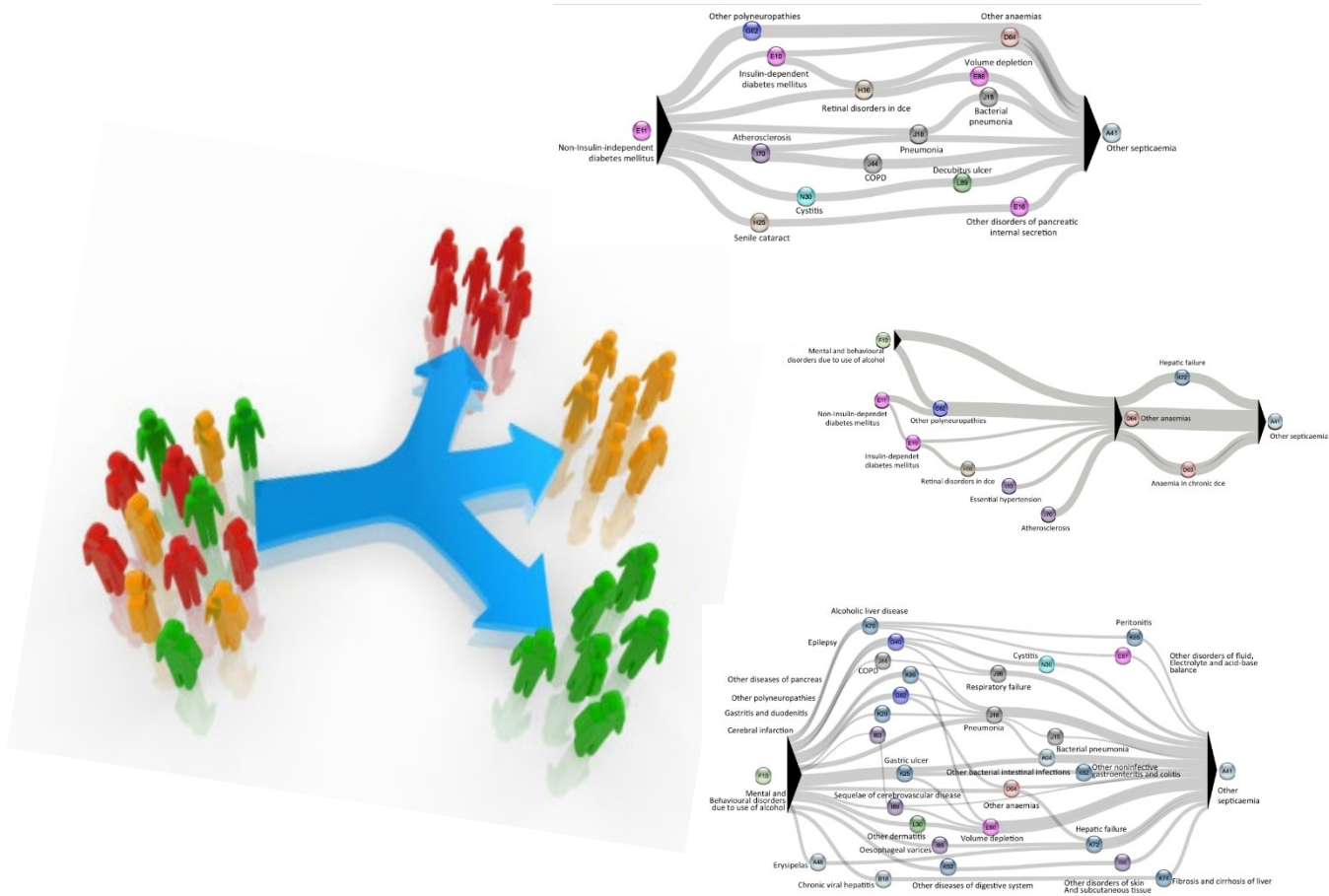
Novo Nordisk Foundation Center for Protein Research
University of Copenhagen
soren.brunak@cpr.ku.dk

Rigshospitalet
soeren.brunak@regionh.dk

Lifelong multimorbidity journeys in disease space



Hu, Thomas & Brunak
Nature Rev. Genetics 2016,


Other polyneuropathies
Other anaemias
Insulin-dependent diabetes mellitus
Volume depletion
Retinal disorders in dce
Bacterial pneumonia
Non-insulin-independent diabetes mellitus
Atherosclerosis
Pneumonia
Other septicaemia
COPD
Decubitus ulcer
Cystitis
Other disorders of pancreatic internal secretion
Senile cataract
Mental and behavioural disorders due to use of alcohol
Hepatic failure
Non-insulin-dependent diabetes mellitus
Other polyneuropathies
Other anaemias
Other septicaemia
Insulin-dependent diabetes mellitus
Retinal disorders in dce
Anaemia in chronic dce
Essential hypertension
Atherosclerosis
Alcoholic liver disease
Peritonitis
Epilepsy
Other disorders of fluid, Electrolyte and acid-base balance
Cystitis
COPD
Other diseases of pancreas
Respiratory failure
Other polyneuropathies
Gastritis and duodenitis
Pneumonia
Cerebral infarction
Bacterial pneumonia
Gastric ulcer
Other bacterial intestinal infections
Other noninfective gastroenteritis and colitis
Mental and Behavioural disorders due to use of alcohol
Other septicaemia
Sequelae of cerebrovascular disease
Other anaemias
Other dermatitis
Hepatic failure
Volume depletion
Oesophageal varices
Erysipelas
Other disorders of skin And subcutaneous tissue
Chronic viral hepatitis
Other diseases of digestive system
Fibrosis and cirrhosis of liver

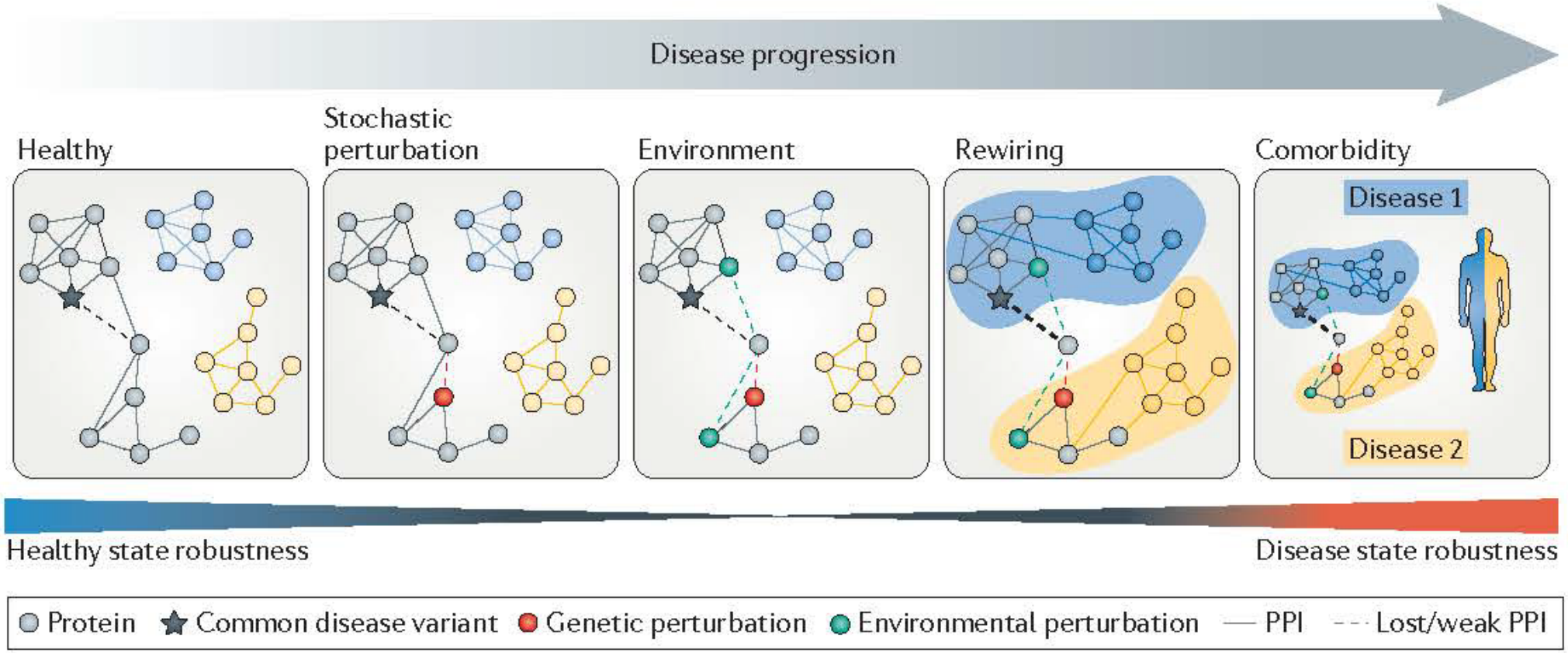
Longitudinal polypharmacy exposures are complex

The New York Times

This Teen Was Prescribed 10 Psychiatric Drugs. She's Not Alone.

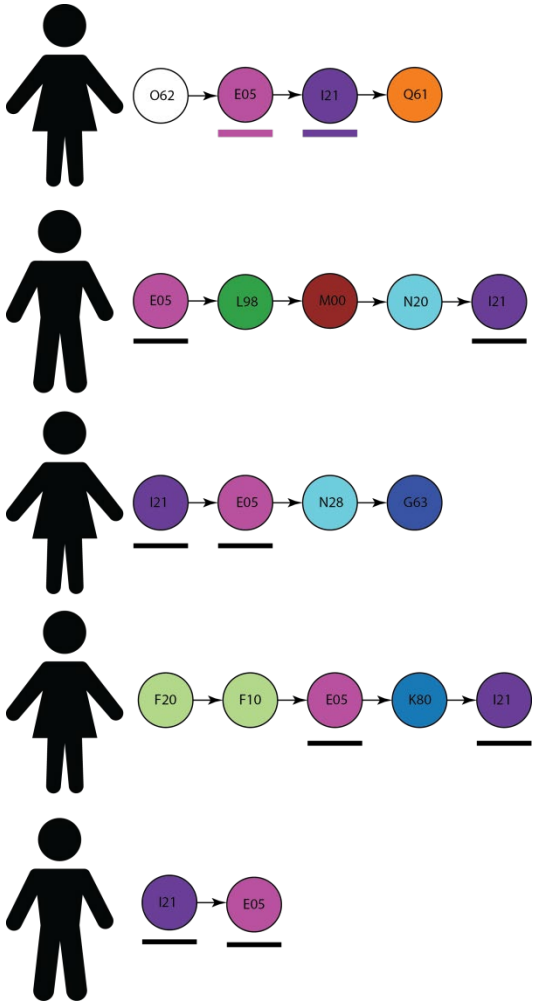
Increasingly, anxious and depressed teens are using multiple, powerful psychiatric drugs, many of them untested in adolescents or for use in tandem.

Clinical and socio-economic data are needed to interpret the molecular domain



Hu, Thomas & Brunak
Nature Rev. Genetics 2016

Diagnosis trajectories across millions of Danes
The ICD-10 system by WHO



ICD 10 chapter coloring

1: Certain infectious and parasitic diseases
2: Neoplasms
3: Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
4: Endocrine, nutritional and metabolic diseases
5: Mental and behavioural disorders
6: Diseases of the nervous system
7: Diseases of the eye and adnexa
8: Diseases of the ear and mastoid process
9: Diseases of the circulatory system
10: Diseases of the respiratory system
11: Diseases of the digestive system
12: Diseases of the skin and subcutaneous tissue
13: Diseases of the musculoskeletal system and connective tissue
14: Diseases of the genitourinary system
15: Pregnancy, childbirth and the puerperium
16: Certain conditions originating in the perinatal period
17: Congenital malformations, deformations and chromosomal abnormalities
18: Symptoms, signs and abnormal clinical and aboratory findings, not elsewhere classified
19: Injury, poisoning and certain other consequences of external causes
20: External causes of morbidity and mortality

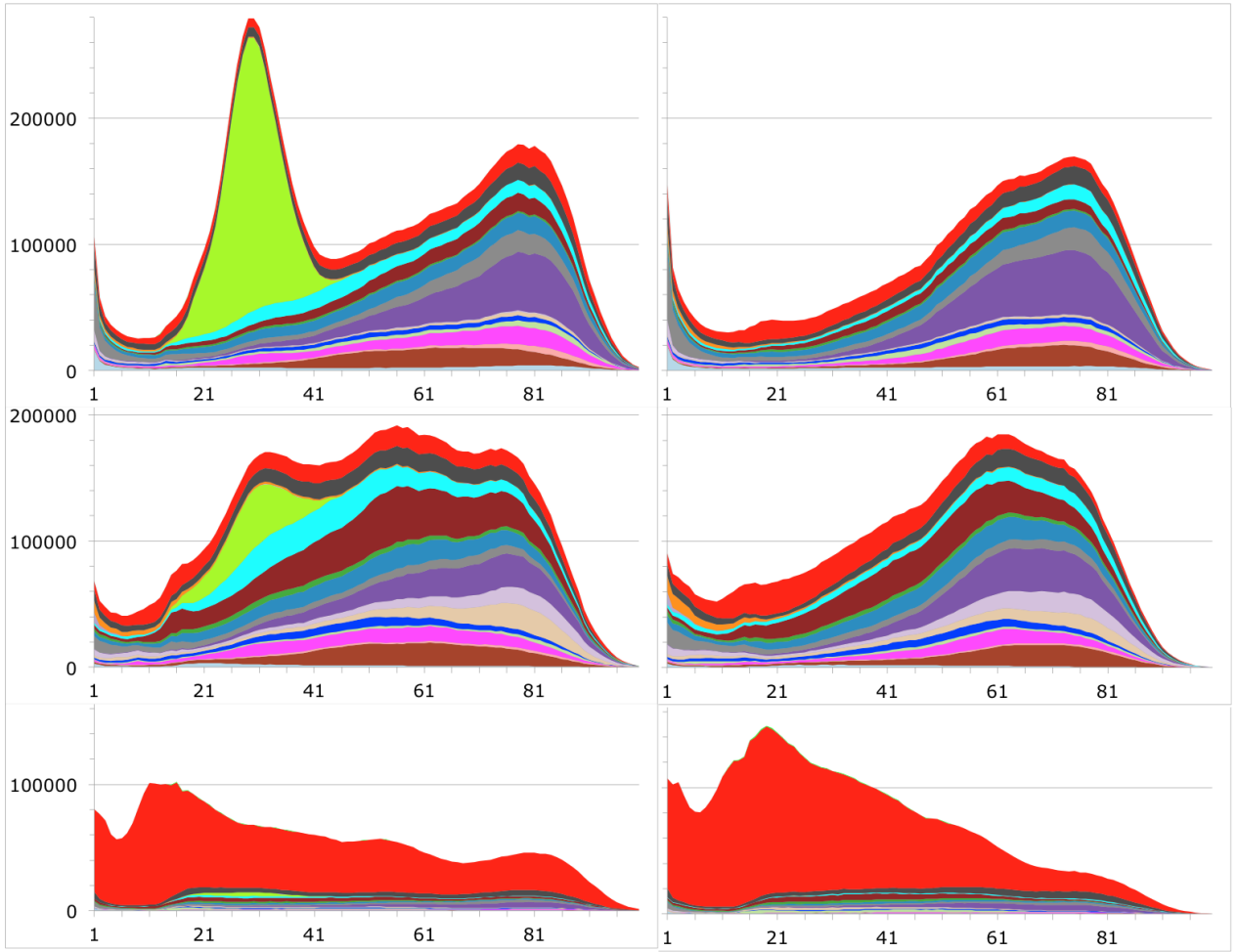
National Patient Registry (~7M Danes)
ICD-10 diagnoses as a function of age

(ICD-10 era, 1994-2019)

In patients

Out patients

IR patients



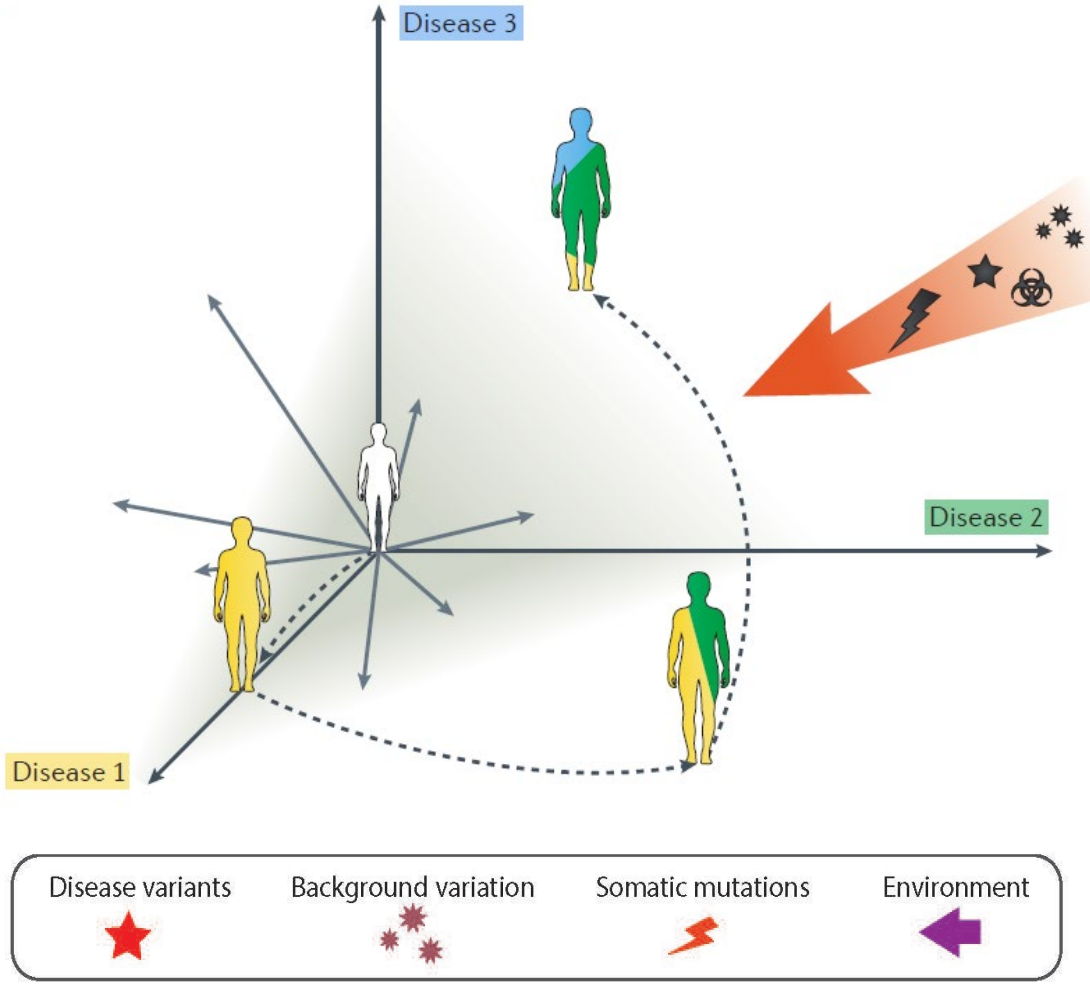
Females

Males

ICD 10 chapter coloring
1: Certain infectious and parasitic diseases
2: Neoplasms
3: Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
4: Endocrine, nutritional and metabolic diseases
5: Mental and behavioural disorders
6: Diseases of the nervous system
7: Diseases of the eye and adnexa
8: Diseases of the ear and mastoid process
9: Diseases of the circulatory system
10: Diseases of the respiratory system
11: Diseases of the digestive system
12: Diseases of the skin and subcutaneous tissue
13: Diseases of the musculoskeletal system and connective tissue
14: Diseases of the genitourinary system
15: Pregnancy, childbirth and the puerperium
16: Certain conditions originating in the perinatal period
17: Congenital malformations, deformations and chromosomal abnormalities
18: Symptoms, signs and abnormal clinical and aboratory findings, not elsewhere classified
19: Injury, poisoning and certain other consequences of external causes
20: External causes of morbidity and mortality

AB Jensen et al.,
Nature Comm., 2014

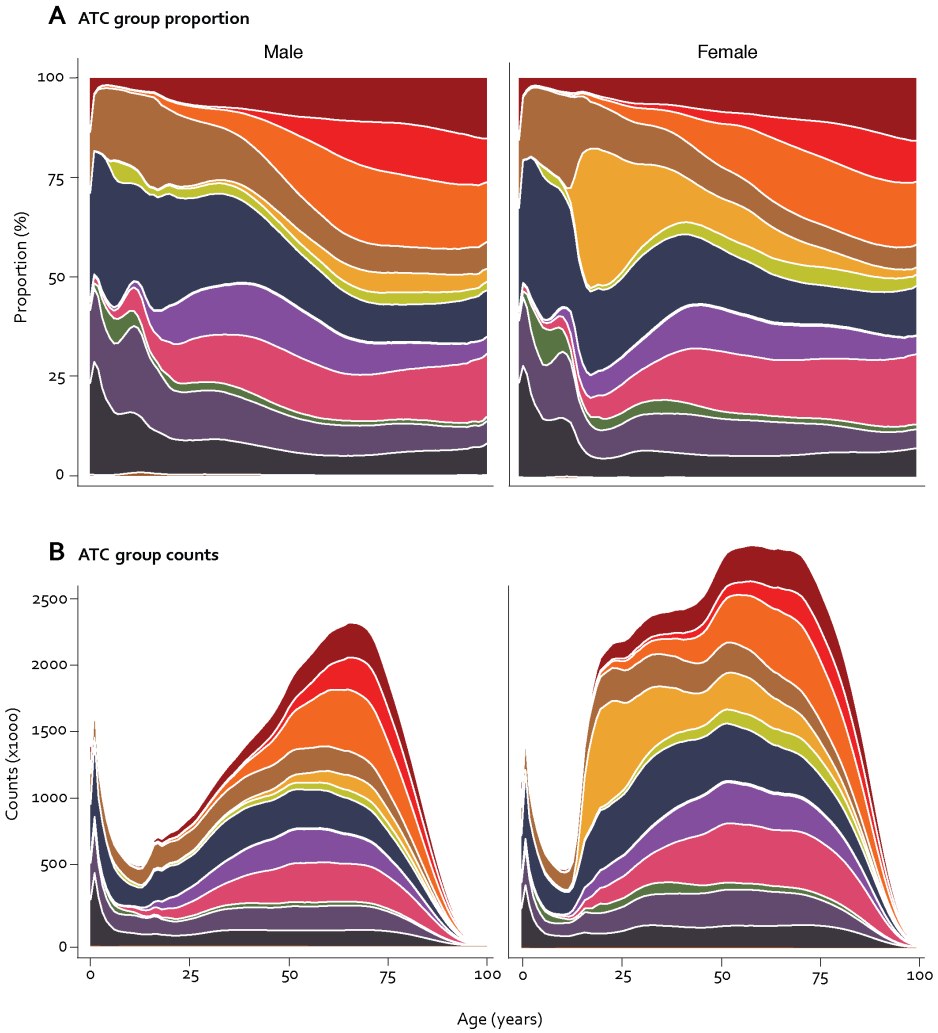
Linking multimorbidity journeys to genetics



Hu, Thomas & Brunak
Nature Rev. Genetics 2016,

TCCAAACCCAGGCTCTCTCCCAAACCAGTTTGCGGCAGATGGCCAGTGGAACCTCACTCTCCTC
ATCAGTAAAAAGGGGGCAGAGTGAGGGTCCTGAGAGCTAGTACAGGGACTGTGTGAAGTAGACA
ATGCCCAGTGTTTAGCGTAAGAATCAGGGTCCAGCTGGTGCTCCCTAAACAGCAGCTGCTGTTC
ACTGTTGAAAGGCGCTCTGGAAGGCCAGGCGCGGTGGCTCATGCTTGTAATCCCAGCACTGTGG
GAGGCCGAGGTGGGCGGATCACCTGAGGTAGGGAGTTCGAGACCAGCCTGACCAACGTGGAGAA
ACCCCATCTCTCCTAAAAATACAAAATTAGCCAGGCGTGGTAGCACATACCTGTAATCCCAGCG
ACTCGGGAGGCTGAGGCAAGAGAATTGCTTGAAACCAGCAGGGGAGGTTGTGGTGAGCCAAGAT
CGAGCCATTGCACTCCAGCCAGGGCAACAAGAGGCAAAATGGCGAAACTCCATCTCCGAGAAAA
AAAAAAAAAAAGAATACTTTCTGAAAGTATTTATTCATACAAATAAAGACTTGACCCATAAGGT
AGGAACGCAAATGGGCCACGGAATCACTCATTCCACAGTATACACCGAGTGCCCTTGAAGTGCT
GGGCACTGCTCCAGGATTGGGGGCATATTGGTGAAAAGAGAAGCAAGCCTGCCTGCTCAGATGG
CAGGGAATGGGGAAAAACAGGGAGACAGTTTCCTGTTTGAGATGTTGGGAGTCTGCTTCGAGTA
GTATATTTACTGGAAATAGACCACTAACTTGGATGTCCCTTTTTGGAAATGTGCCTGCGTCCAG
GGCTGGGTTGGGGCCCCAATGAACTTTGGCTCTGACATAGCTGTTGCCACACTCAGTGGAACTG
AATCCATGTTTGCCTTCACCCGGCATCCTTCACCCCAACTCTCCCCGCCACAACATACATCCCA
TGCCAGCCTGGGGACCCTCAAAGGTGCTTCATCATTAGGTTTGTGGCTGGGTCCTACTGAAGTA
ACAGACCCGGTGAGAGCCCCCATTCCAATGCACCCCCGATCTCAGCTGTCTGGCCAGAAGACCT
GAGCAAGTCCCTCCTTCTTCCTGGCCTTGGCCTTCCCATGGGTGGAACCGGGAGGGTTGGCTTT
AATCTCCACCAGAACTCTTGCCCCGGGACTGTGATGGGCGATTGGCCACTTCTCCTCGATAACA
TTACTGTTTTTCTTCCGCCTTCTGGTTGACTTTAGCCAGAACCAGTGCTTCTATAACTCCAGTT
ACCTGAATGTCCAGCGGGAGAATGGGACCGTCTCCAGATACGGTGAGGGCCAGCCCTCAGGCAG
GAGGGTTCACCGTGGGAACAGGGCAGGCCAGCATAAGGTGGGTTGAGCTGGTTCCACAGGCCAG
AGCTCATTCTGCCCTCTCCCCGGAAGACCTCCCACCCTGTCCCCATGCCTCTGCTTCTCCCTCA
CCCCAATTCCCCGCTGCCTTCTAGGATAAGTGTGAGCCACTGGAGAAGCAGCACGAGAAGGAGA
GGAAACAGGAGGAGGGGGAATCCTAGCAGGACACAGCCTTGGATCAGGACAGAGACTTGGGGGC
CATCCTGCCCCTCCAACCCGACATGTGTACCTCAGCTTTTTCCCTCACTTGCATCAATAAAGCT
TCGCATCGGCCTTTCGAAACGAGGAGTACAATAAGTCGGTTCAGGAGCCCTCAGGCAGGAGGGT
TCACCGTGGGAACAGGGCAGGCCAGCATAAGGTGGGGGCTGGATGTAGAGCCCTGGAGGCTTTG
GGCACAGAGGCCACCCTGGACCGGGTGAGTGCCTGGGCTAGCCCTGTCCTGAGCACATGGGCAG
CTGCCTCCCTTCTCTGGGCTTCCCTTTACCTGCTGGCTGTGGTCGCACCCCCACTCCCAGCCCC
CAACTCTCCCCGCCACAACATACATCCCATGCCCAGGAGGGTTCACCGTGGGAACAGGGCAGGC
CAGCATAAGGTGGGGGCTGGATGTAGAGCCCTGGAGGCTTTGGGCACAGAGGCCACCCTGGACC

ATC drug groups in 1.1. billion male and female prescriptions according to age



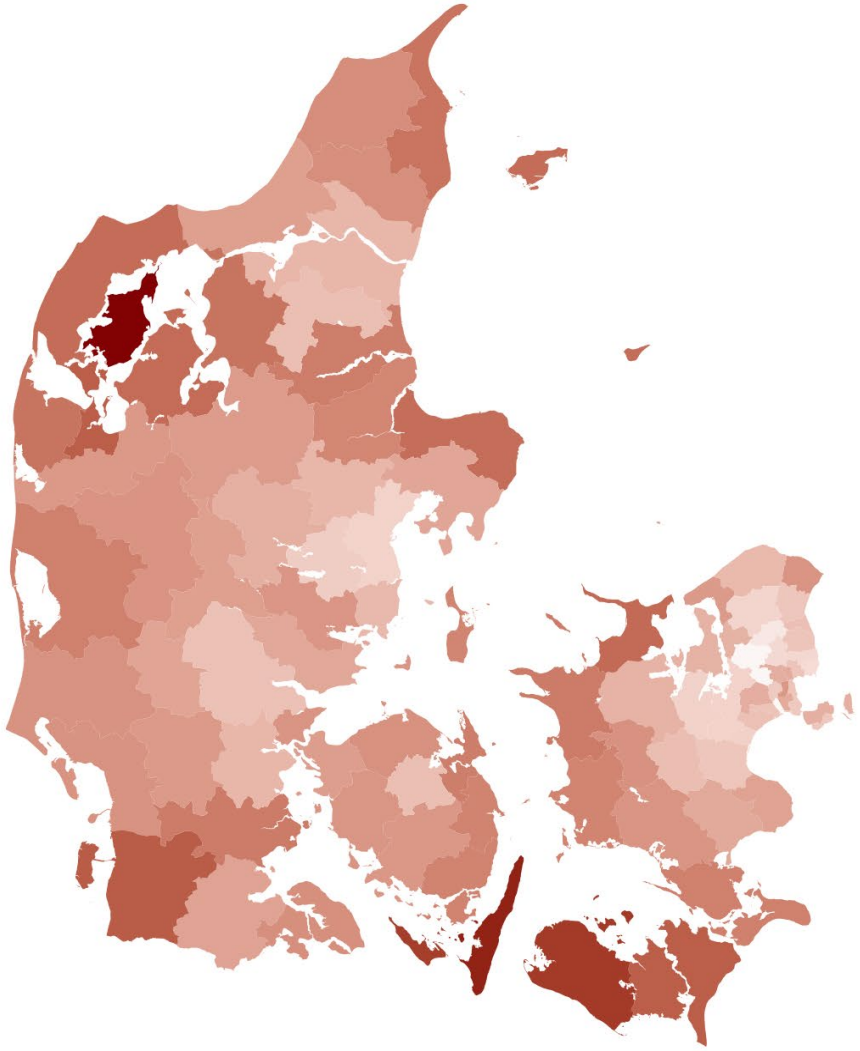
ATC group

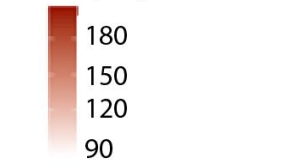
- A alimentary tract and metabolism
- B blood and blood forming organs
- C cardiovascular system
- D dermatologicals
- G genito-urinary system and sex hormones
- H systemic hormonal preparations
- J antiinfectives for systemic use
- L antineoplastic and immunomodulating agents
- M musculoskeletal system
- N nervous system
- P antiparasitic products, insecticides and repellents
- R respiratory system
- S sensory organs
- V various

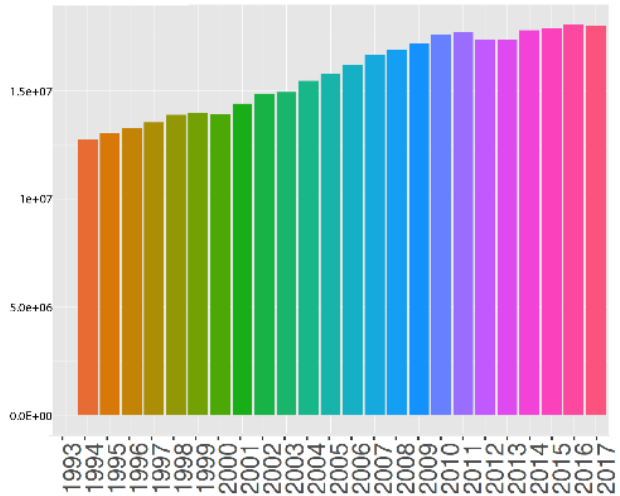
Aguayo Orozco et al., npj Digital Medicine 2021

1.1 billion prescriptions from Denmark

(map with density per individual 1993-2019)

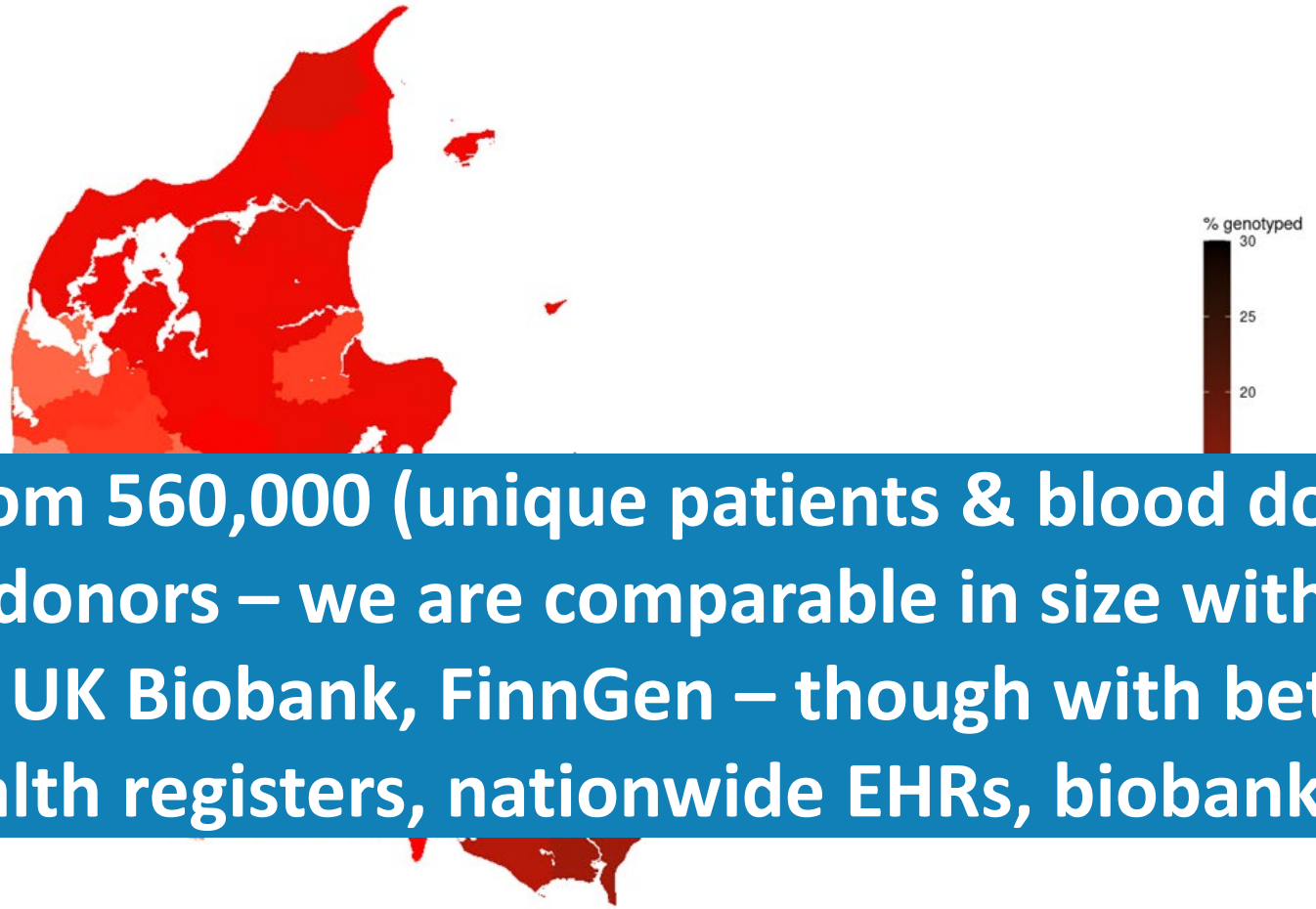






Aguayo Orozco et al., npj Digital Medicine 2021

Nationwide coverage of genotyped individuals in our data resource – in some municipalities up till 30% are genotyped!



With genotypes from 560,000 (unique patients & blood donors) & WGS from 60,000 blood donors – we are comparable in size with international initiatives e.g., The UK Biobank, FinnGen – though with better data from Danish health registers, nationwide EHRs, biobanks etc.

Copenhagen Hospital Biobank CHB (patients)

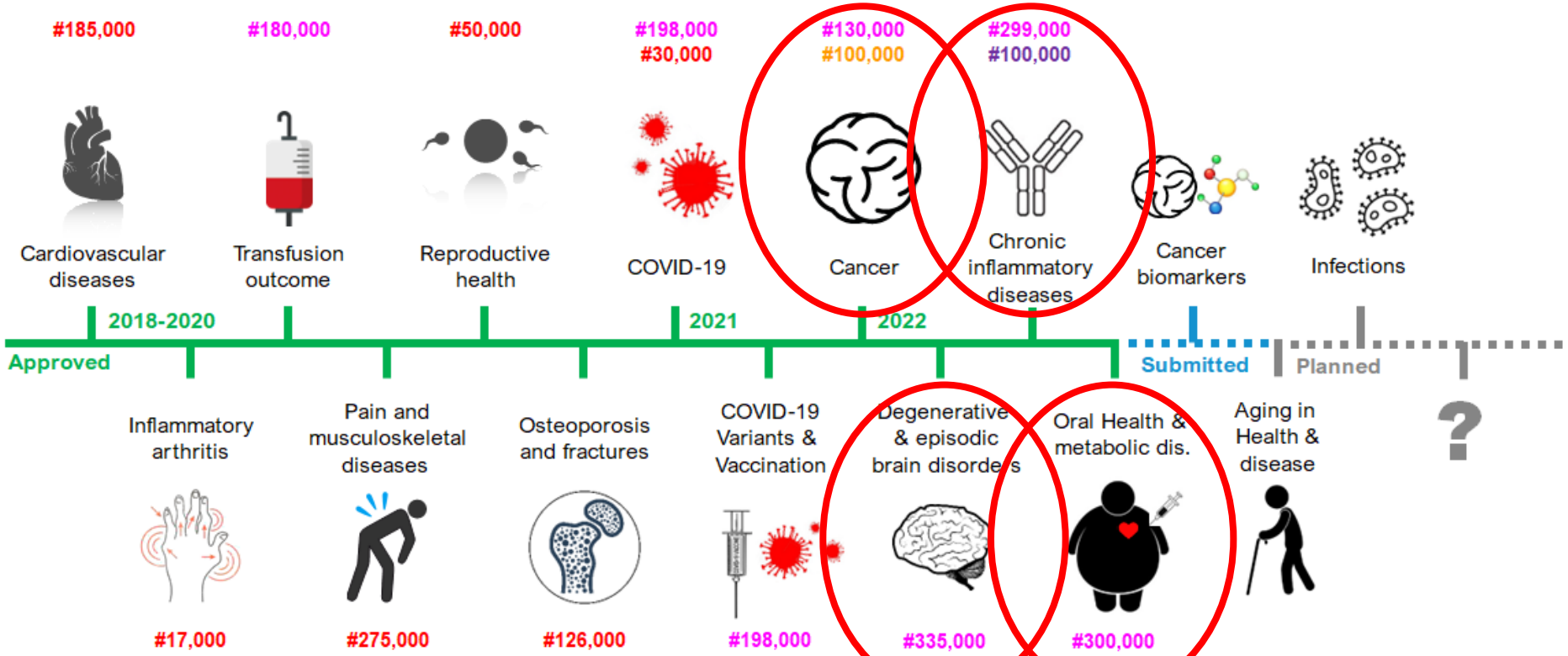
- Biobank samples from >**500,000 adult patients**
- **Genome-wide genomic data 360,000** (~200,000 alive)
- **N=90,000 cancer patients (DCB): WB genotypes & tissue biopsies WGS/targeted seq**
- **Health Register data**, electronic health records, laboratory data, imaging etc.
- **N=17,000 brain disease patients: CSF & plasma/serum sample proteomics analysis**
- **N=100,000 inflammatory disease patients: Plasma proteomics & SWAP microbiome**
- **N=20,000 cardiometabolic disease patients: Plasma and urine proteomics & SWAP microbiome analysis**

The Danish Blood Donor Study DBDS (healthy invividuals)

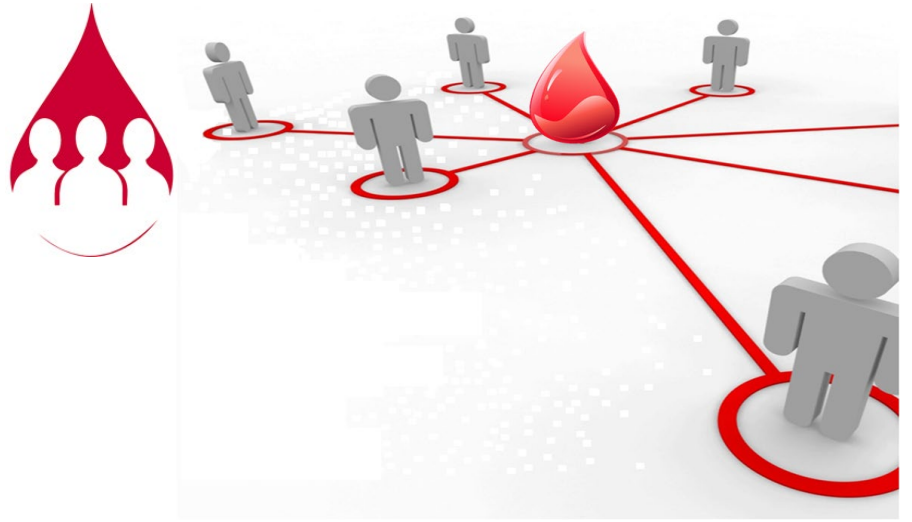
- Biobank samples from >**160,000 healthy adults**
- **Genome-wide genomic data 150,000**
- **Health Register data**, electronic health records, laboratory data, imaging etc.
- **~3 mill consecutive plasma samples**
- **Questionnaire data** – >400,000 responses
- **N=60,000 WGS**
- **N=30,000+ plasma olink Explore 3072 proteomics**
- **N=125,000 gen-5 questionnaires sent out**

***Ongoing or pending**

Copenhagen Hospital Biobank (CHB) – emerging multiomics data

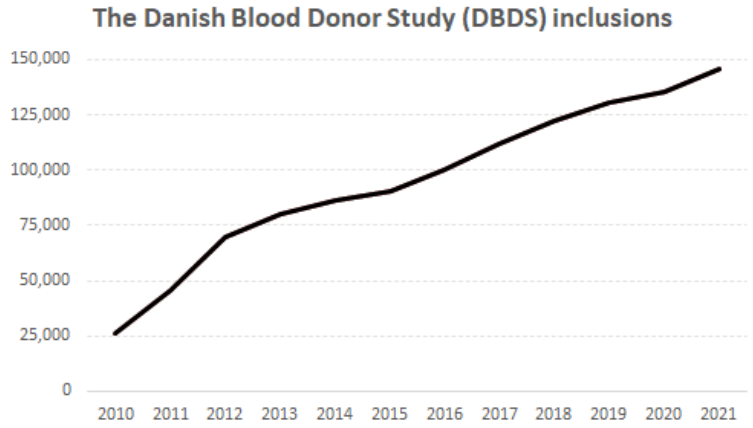


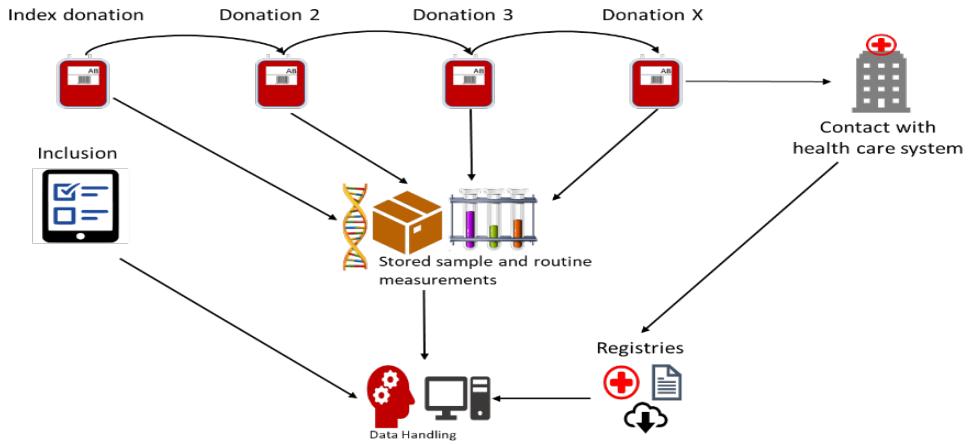
Research ethics protocol's approved (▬) submitted (▪▪) or in pipeline/planned (▪▪)
The protocols also include 116,000 genotyped DBDS participants as controls and cases
= newly genotyped **CHB** & **DCB** / reuse of genotyped **CHB** / **emerging plasma/CSF proteomics / swap microbiome**



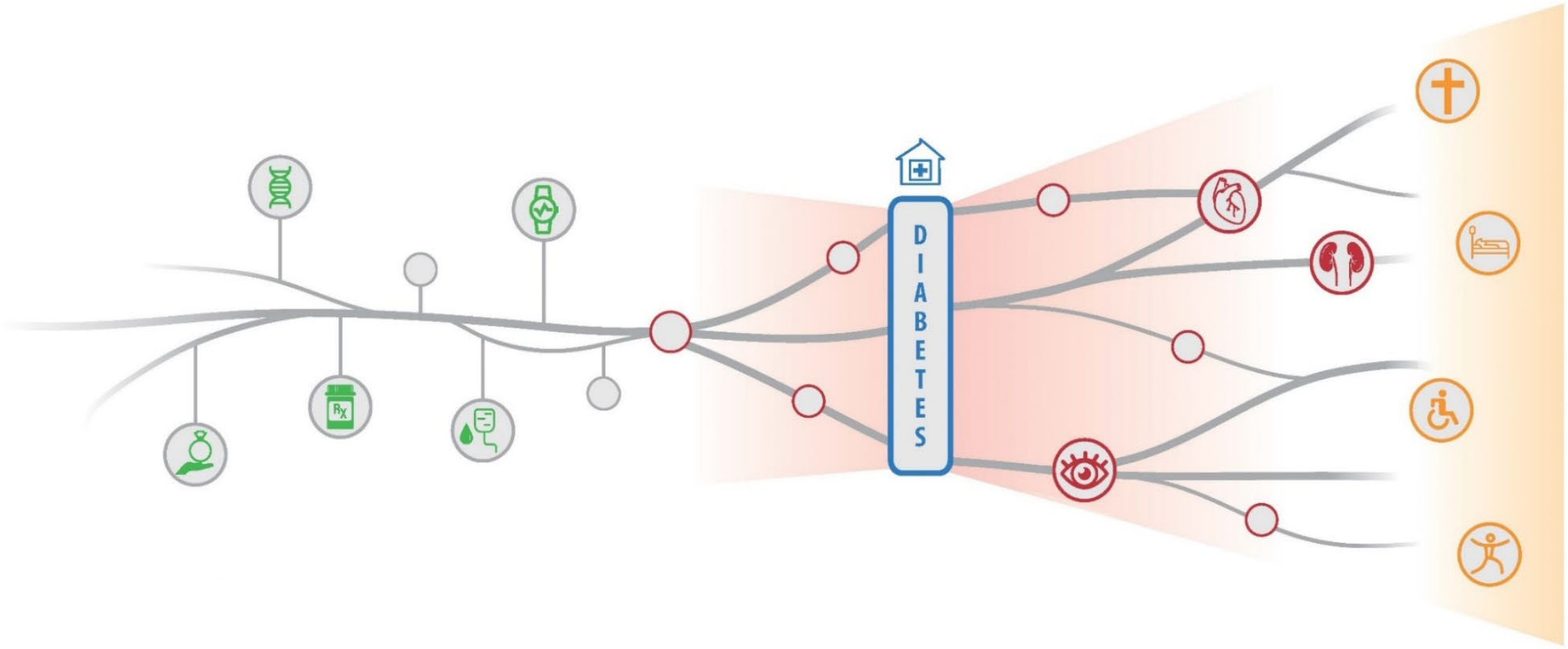
The Danish Blood Donor Study (DBDS)

- Established 2010
  - Nation-wide ongoing open cohort, biobank and research infrastructure, leveraging the Danish blood bank infrastructure: Donors are included and followed over years!

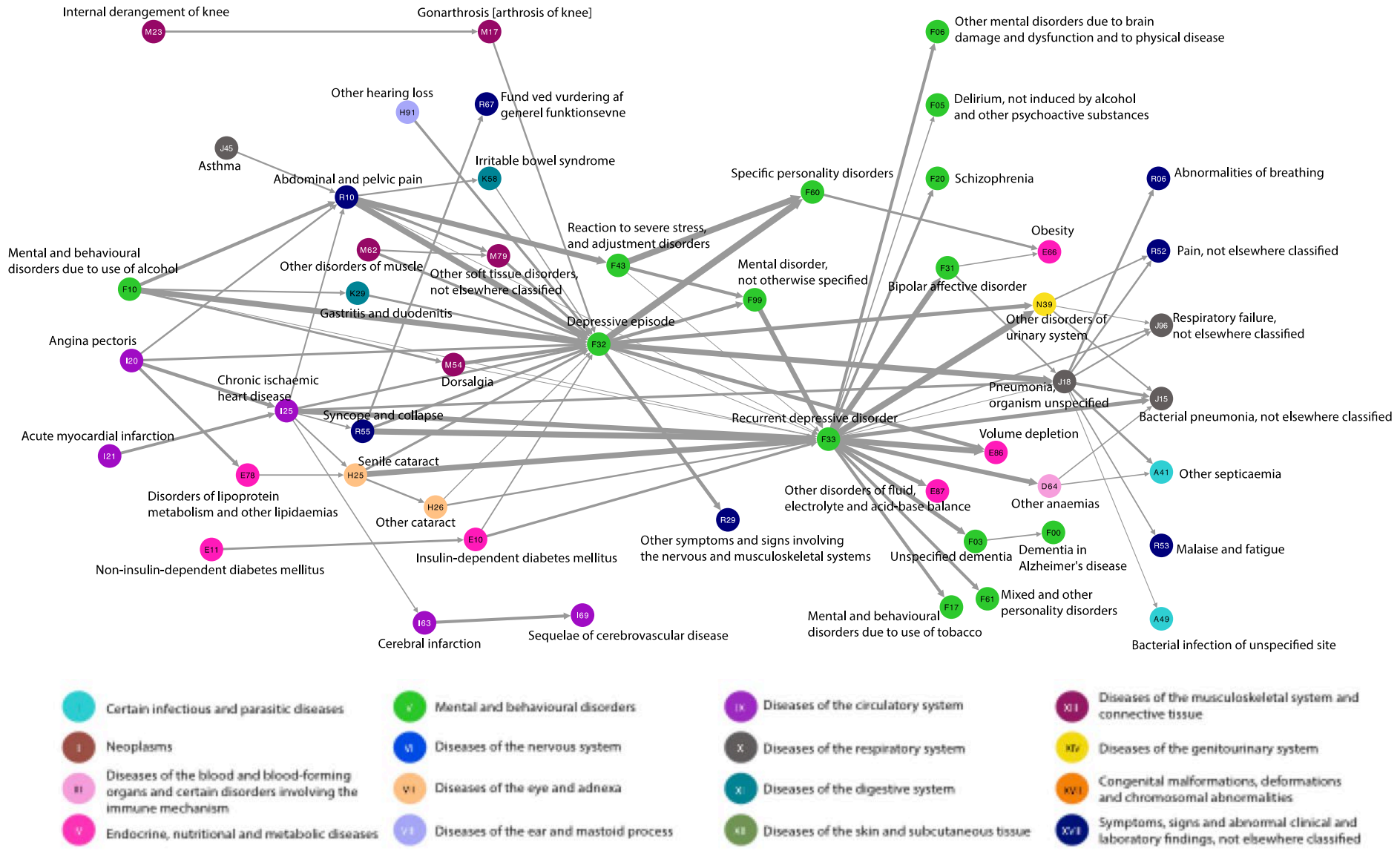




Life-course trajectories and health-to-disease transitions



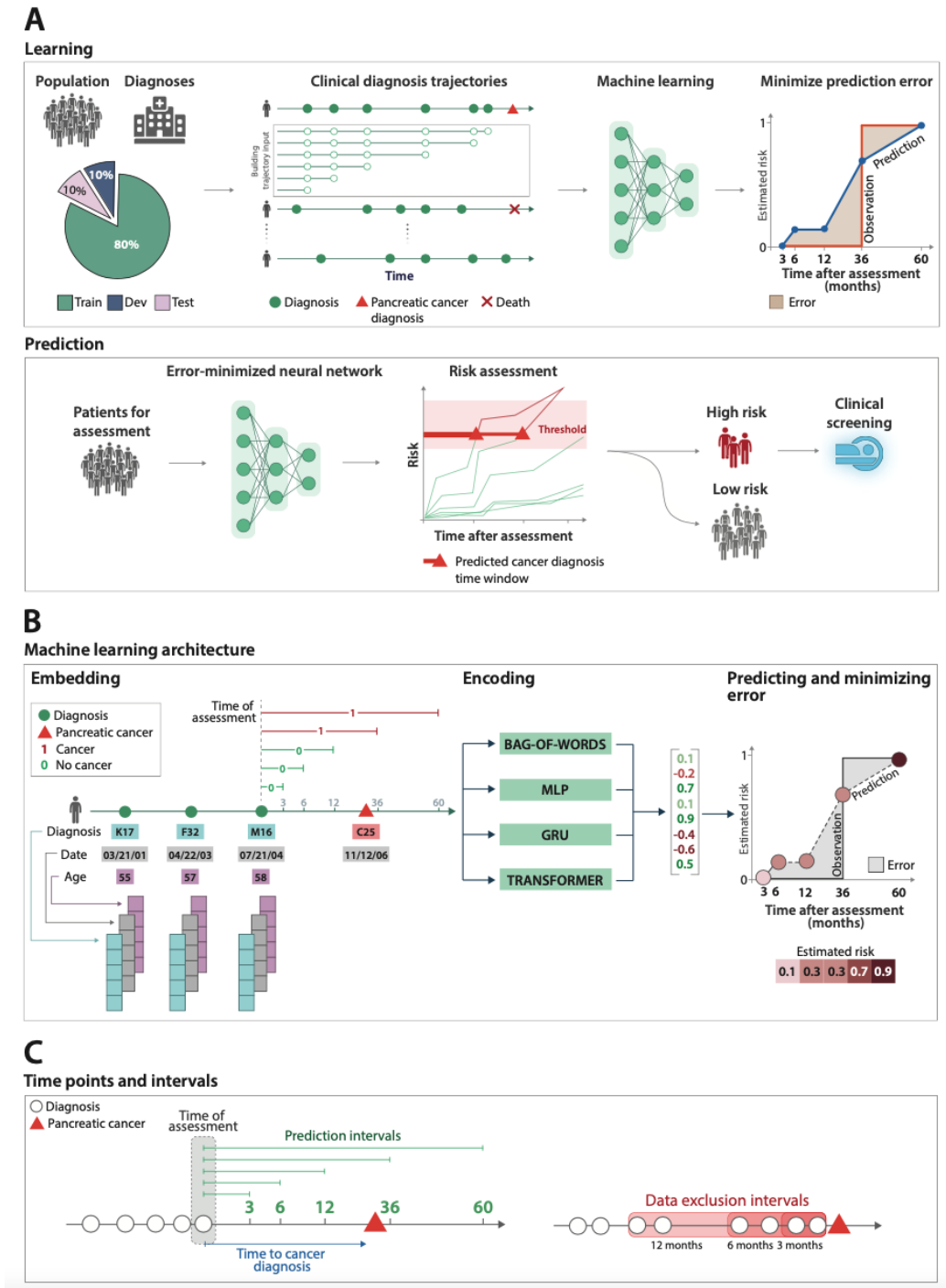
Disease trajectory network of depression patients



Prediction of pancreas cancer risk – training on Danish data, replication in US data

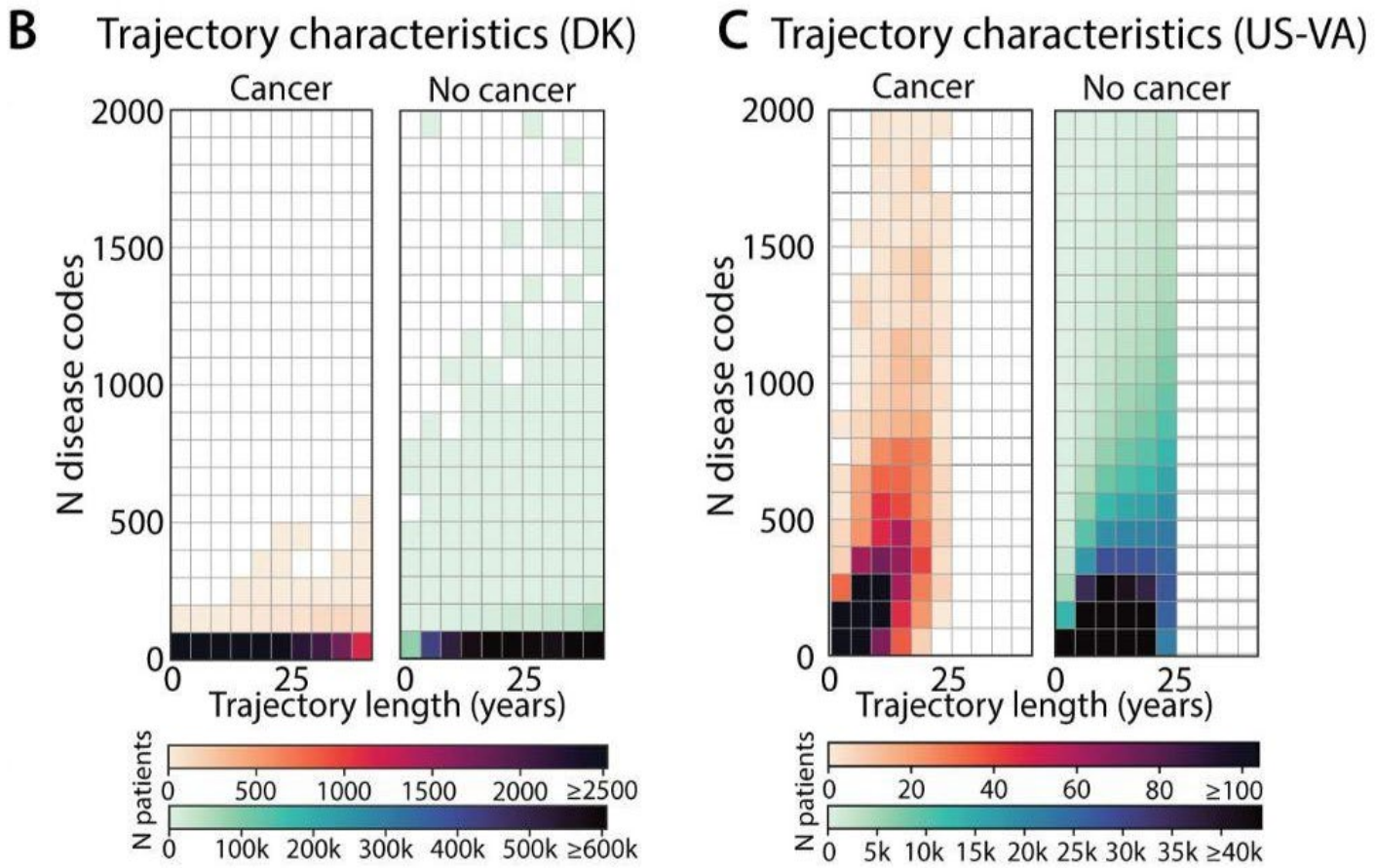
Disease histories from

- Danish National Patient Registry (DNPR), covering 8.6 M patients between 1977-2018 (6.1 M controls, 24,000 cases, av. 23 yrs of history)
- Veteran Affairs database, covering 2.9 M patients 1999-2020 (0.75 M controls, 3,800 cases, av. 12 yrs of history)



A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. Placido, Yuan, Hjaltelin, Zheng, …, Brunak & Sander, Nature Medicine 29: 1113; 2023

Different Denmark & US EHR structure



Many more codes in the US data per patient than in the DK data

Nature Medicine 2023

Feature importance ranking using explainability methods

Comparing features in the DK and US

C

Feature contributions - No exclusion (DK)

	Cancer in 0-6 months	Cancer in 6-12 months	Cancer in 12-24 months	Cancer in 24-36 months
1	Unspecified jaundice	Other diseases of biliary tract	Medical observation and evaluation for suspected diseases and conditions	Medical observation and evaluation for suspected diseases and conditions
2	Medical observation and evaluation for suspected diseases and conditions	Unspecified jaundice	Other diseases of biliary tract	Other diseases of pancreas
3	Other diseases of biliary tract	Medical observation and evaluation for suspected diseases and conditions	Other diseases of pancreas	Other diseases of biliary tract
4	Abdominal and pelvic pain	Other diseases of pancreas	Abdominal and pelvic pain	Non-insulin-dependent diabetes mell itus
5	Malignant neoplasm of other and uns pecified parts of biliary tract	Malignant neoplasm of other and uns pecified parts of biliary tract	Non-insulin-dependent diabetes mell itus	Unspecified jaundice
6	Other diseases of pancreas	Abdominal and pelvic pain	Malignant neoplasm of other and uns pecified parts of biliary tract	Abdominal and pelvic pain
7	Secondary malignant neoplasm of res piratory and digestive organs	Secondary malignant neoplasm of res piratory and digestive organs	Unspecified jaundice	Malignant neoplasm of other and uns pecified parts of biliary tract
8	Symptoms and signs concerning food and fluid intake	Non-insulin-dependent diabetes mell itus	Other functional intestinal disorde rs	Gastritis and duodenitis
9	Non-insulin-dependent diabetes mell itus	Malignant neoplasm without specific ation of site	Diseases of pancreas	Insulin-dependent diabetes mellitus
10	Other anaemias	Other anaemias	Secondary malignant neoplasm of res piratory and digestive organs	Other anaemias

D

Feature contributions - No exclusion (US-VA)

	Cancer in 0-6 months	Cancer in 6-12 months	Cancer in 12-24 months	Cancer in 24-36 months
1	Acute pancreatitis	Acute pancreatitis	Abdominal and pelvic pain	Diabetes mellitus
2	Abdominal and pelvic pain	Diabetes mellitus	Other diseases of biliary tract	Other diseases of liver
3	Other diseases of biliary tract	Other diseases of biliary tract	Diabetes mellitus	Persons encountering health services in other circumstances
4	Diabetes mellitus	Symptoms and signs concerning food and fluid intake	Persons encountering health services in other circumstances	Abdominal and pelvic pain
5	Other diseases of pancreas	Persons encountering health services in other circumstances	Acute pancreatitis	Other diseases of biliary tract
6	Symptoms and signs concerning food and fluid intake	Malignant neoplasm of trachea, bronchus or lung	Dependence of opioids, sedatives, cocaine, cannabinoids, hallucinogens, or other psychoactive substances	Nausea and vomiting
7	Disorders of social functioning with onset specific to childhood and adolescence	Abdominal and pelvic pain	Abuse of alcohol, tobacco, opioids, sedatives, cocaine, cannabinoids, hallucinogens, or other psychoactive substances	Abuse of alcohol, tobacco, opioids, sedatives, cocaine, cannabinoids, hallucinogens, or other psychoactive substances
8	Essential (primary) hypertension	Other diseases of pancreas	Cough, haemorrhage from respiratory passages	Unspecified jaundice, or skin eruption
9	Persons encountering health services in other circumstances	Dependence of opioids, sedatives, cocaine, cannabinoids, hallucinogens, or other psychoactive substances	Secondary malignant neoplasm of respiratory and digestive organs	Cataract
10	Examination and observation for other reasons	Other dermatitis	Cataract	Dependence of opioids, sedatives, cocaine, cannabinoids, hallucinogens, or other psychoactive substances

ICD-10 chapters

- II Neoplasms
- III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
- IV Endocrine, nutritional and metabolic diseases
- V Mental and behavioral disorders
- VII Diseases of the eye and adnexa
- IX Diseases of the circulatory system
- XI Diseases of the digestive system
- XII Diseases of the skin and subcutaneous tissue
- XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
- XXI Factors influencing health status and contact with health services

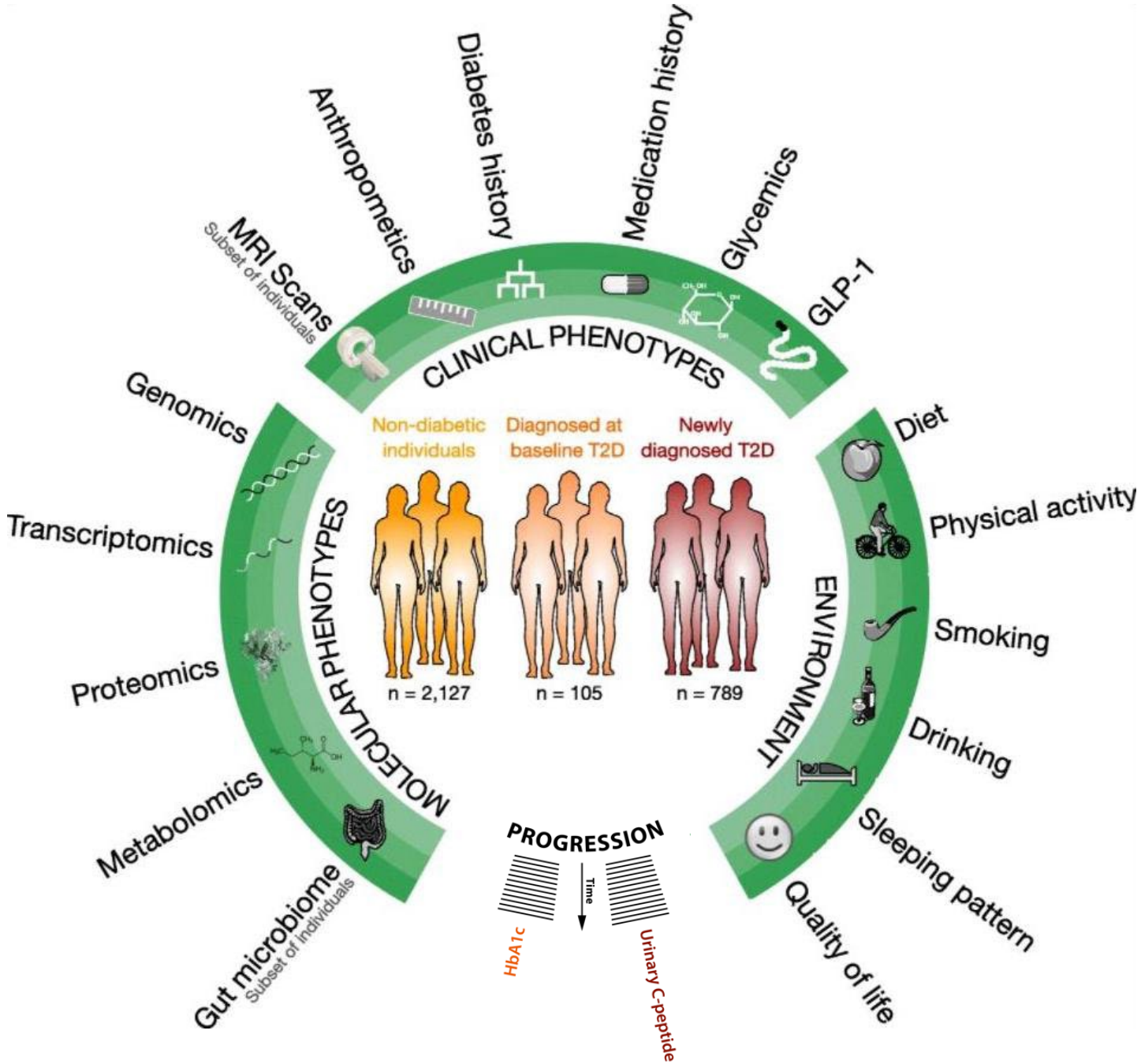
Nature Medicine 2023

DIRECT

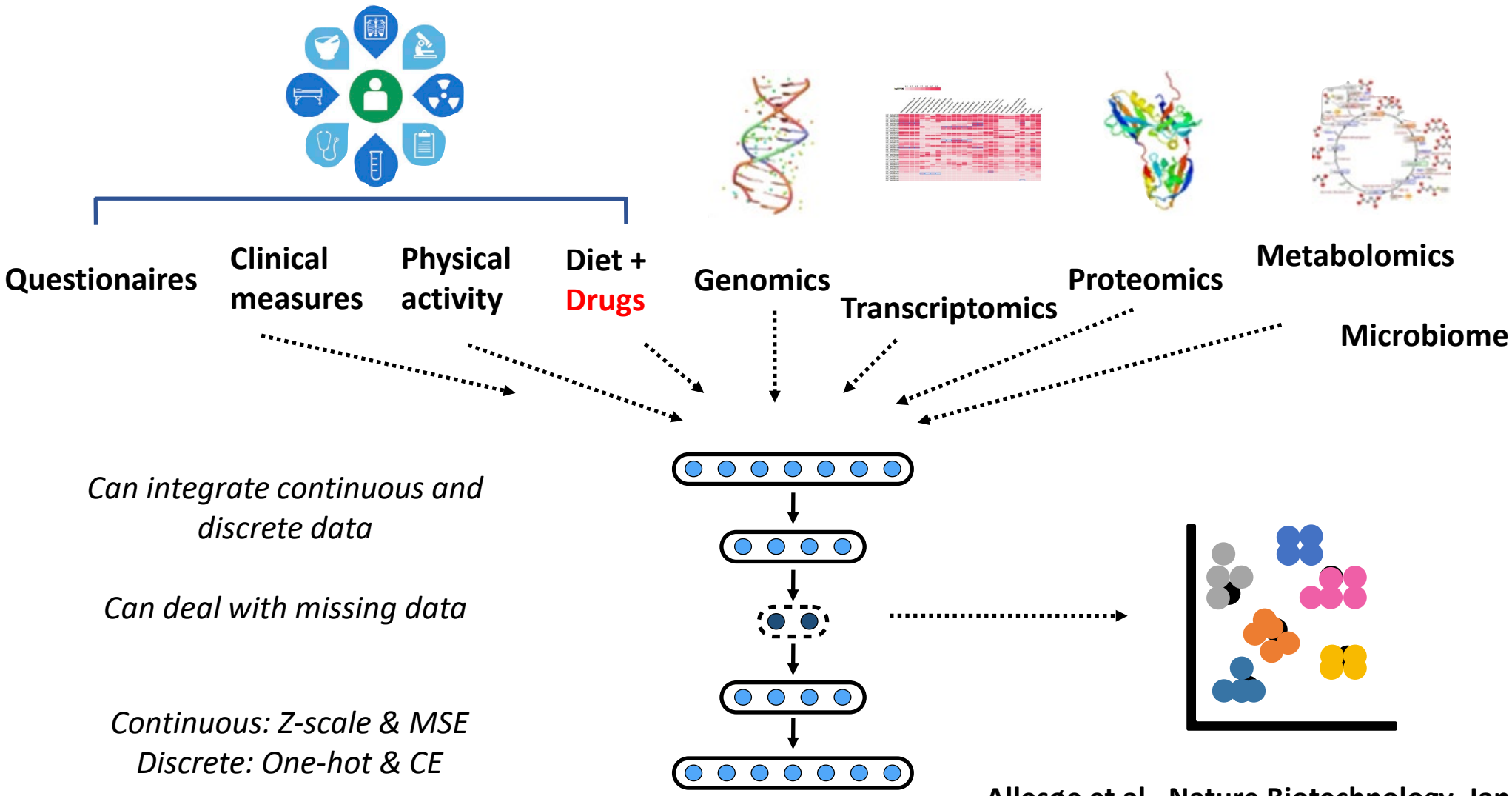
DIABETES RESEARCH ON PATIENT STRATIFICATION

DIRECT is a pan-European multi-omics consortium engaged in research on diabetes that was initially funded by the European Union's Innovative Medicines Initiative (IMI).

Twenty academic research institutions and five pharmaceutical research organizations launched the DIRECT project in 2012.



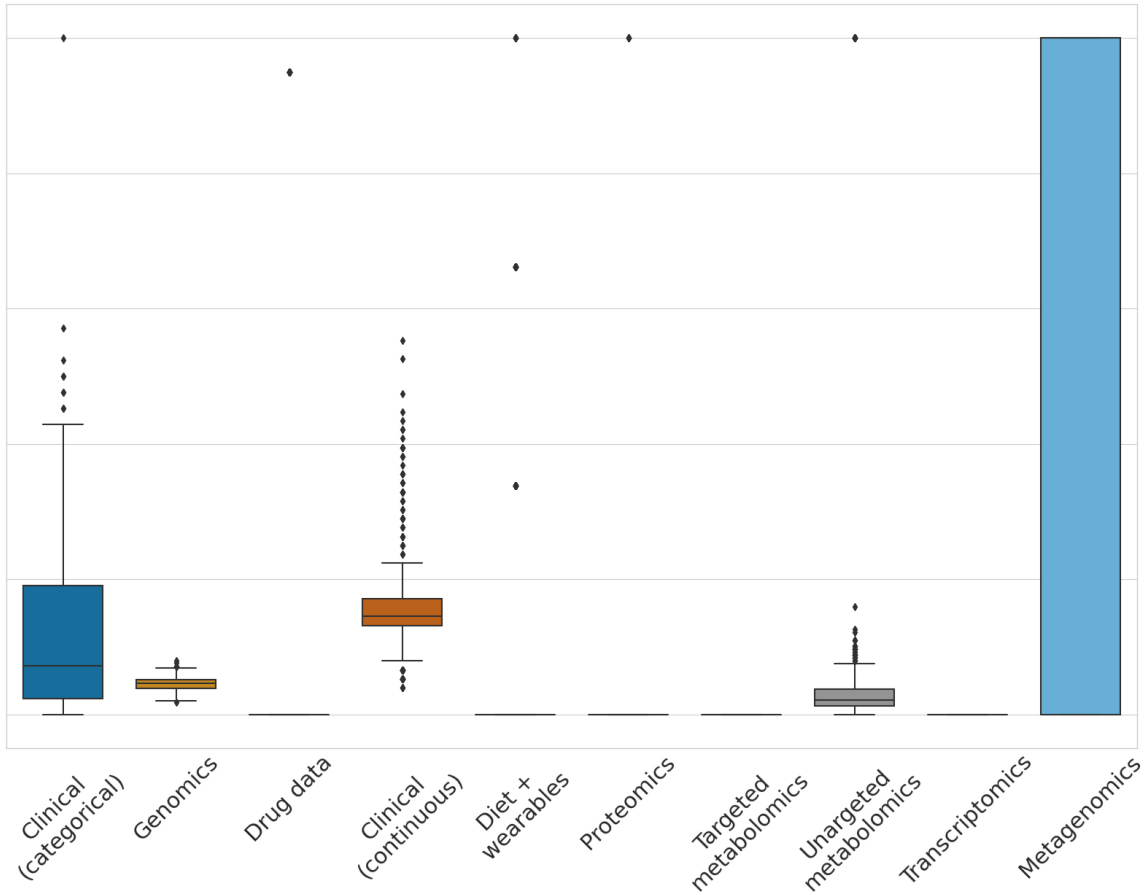
Full scale data integration in the DIRECT T2D cohort



Allesøe et al. Nature Biotechnology, Jan 2, 2023

Missing data can be handled

- Missing data is normally a big problem!
  - Take the samples out?
  - Remove the features?
  - Impute the data?

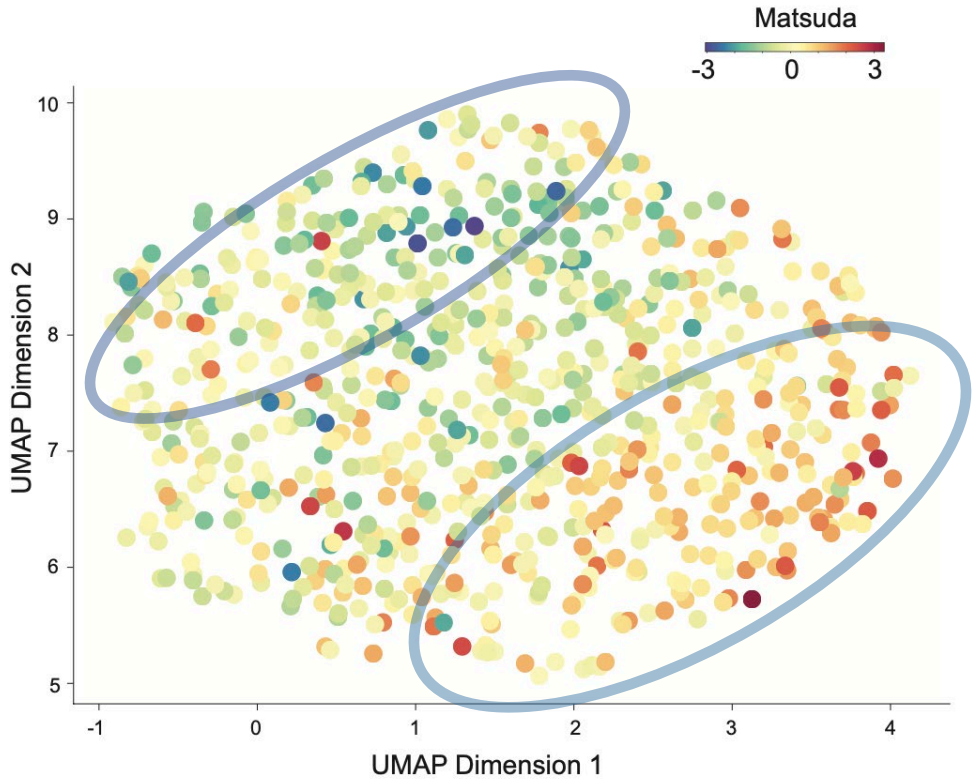


The latent space space of the autoencoder

UMAP: Newly diagnosed T2D

High insulin sensitivity

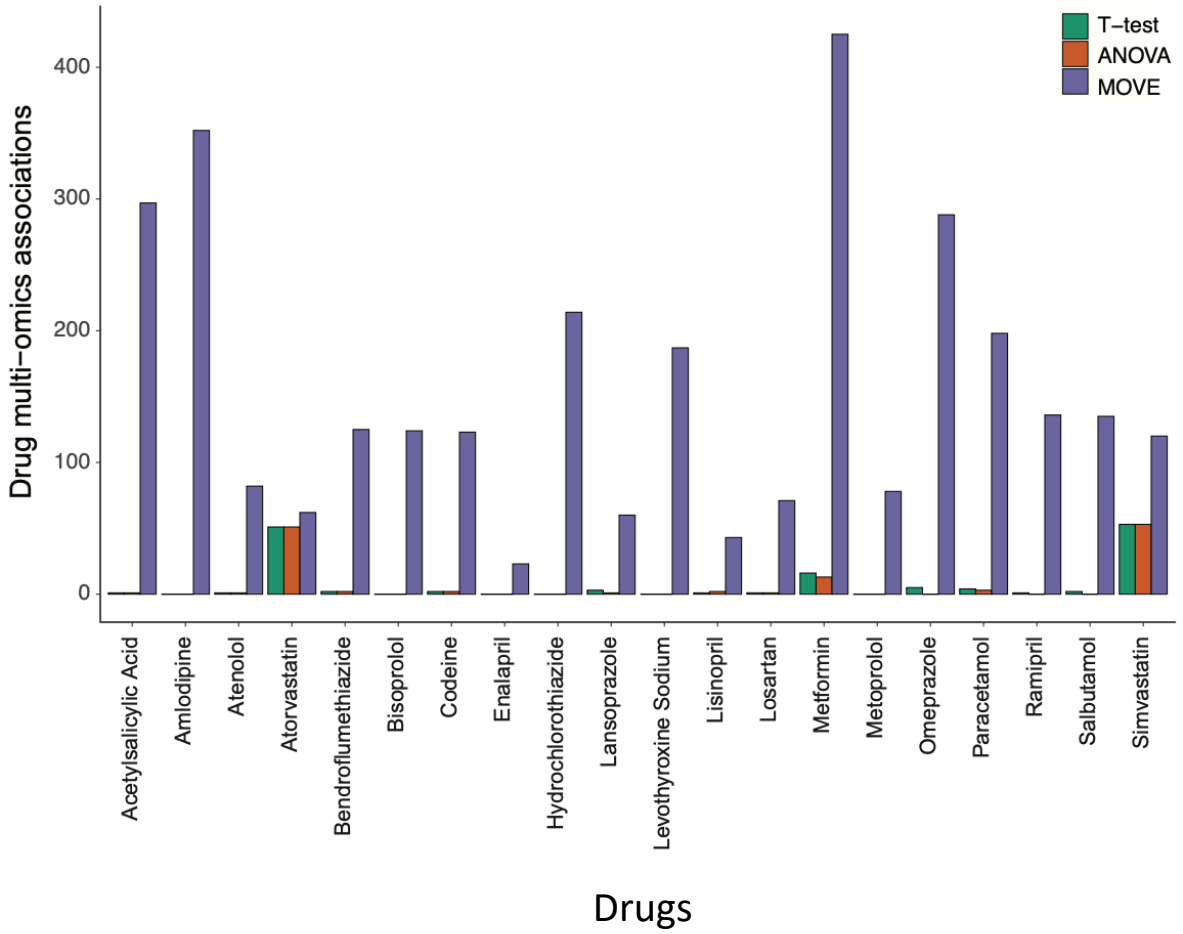
Low insulin sensitivity



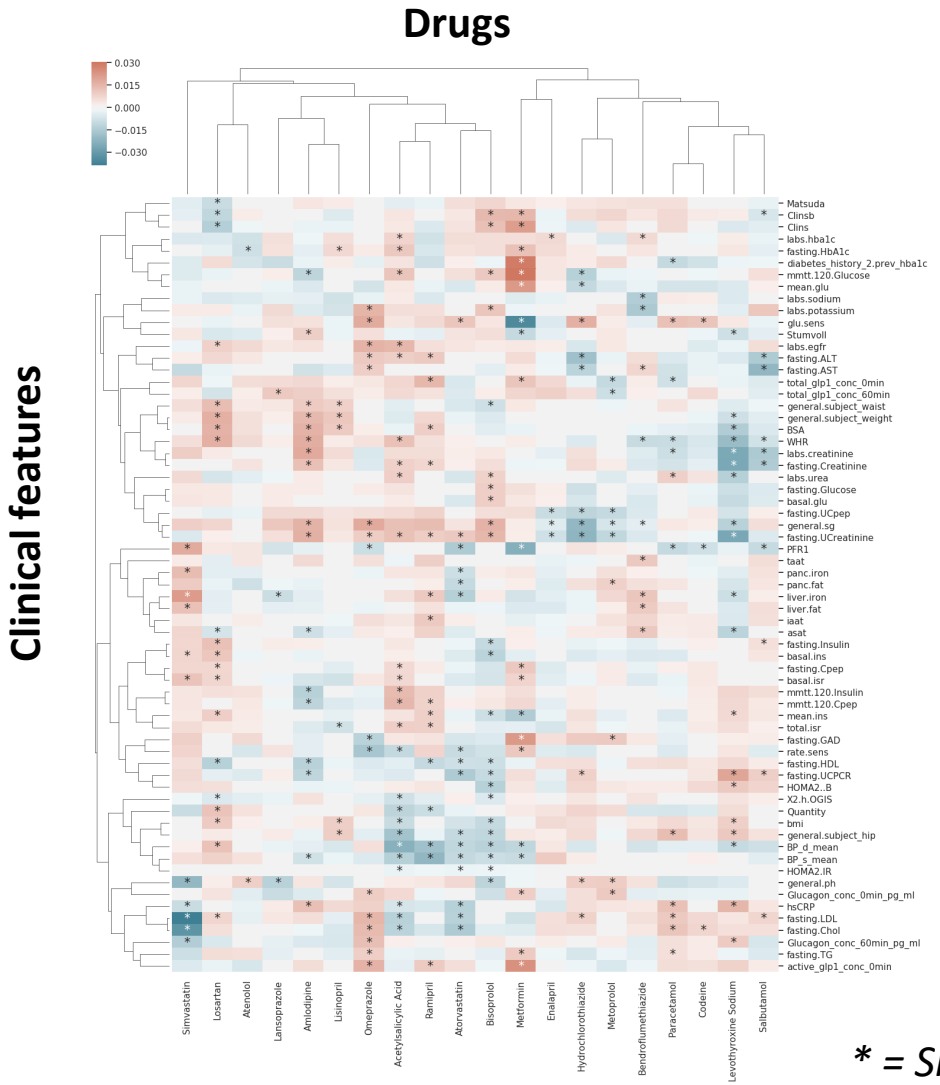
- Data is integrated in a meaningful way
- Newly diagnosed T2D is a continuum
- Note: Unsupervised!
- Raw data: Very little signal

The method is sensitive in finding associations

- **Drug ~ omics associations**
- Compared to standard statistical approaches
- MOVE: **3,000 more significant** associations



Drug ~ Clinical

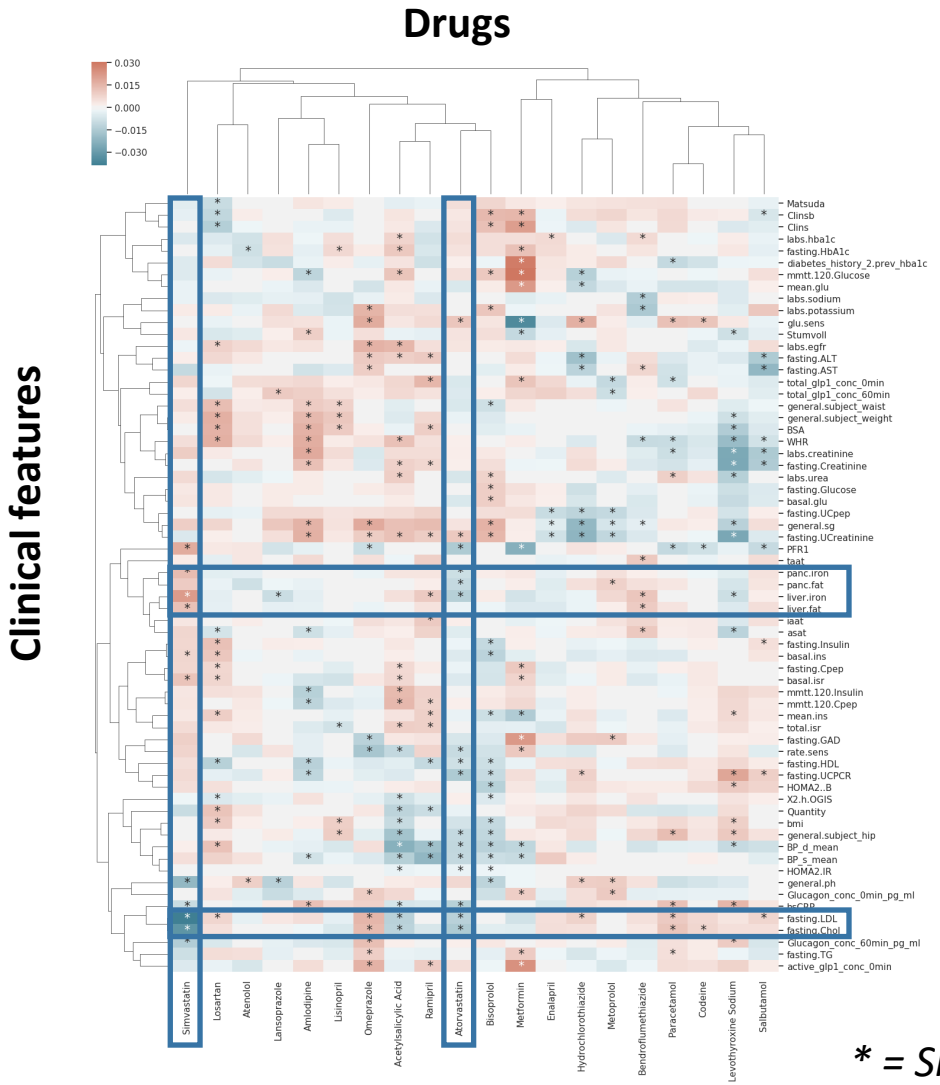


- *Model estimation of what will happen when given drug X*

- Compare patient not given a drug with the same patient given a drug – average over all patients
- Do we find what we expect?

** = Significant change*

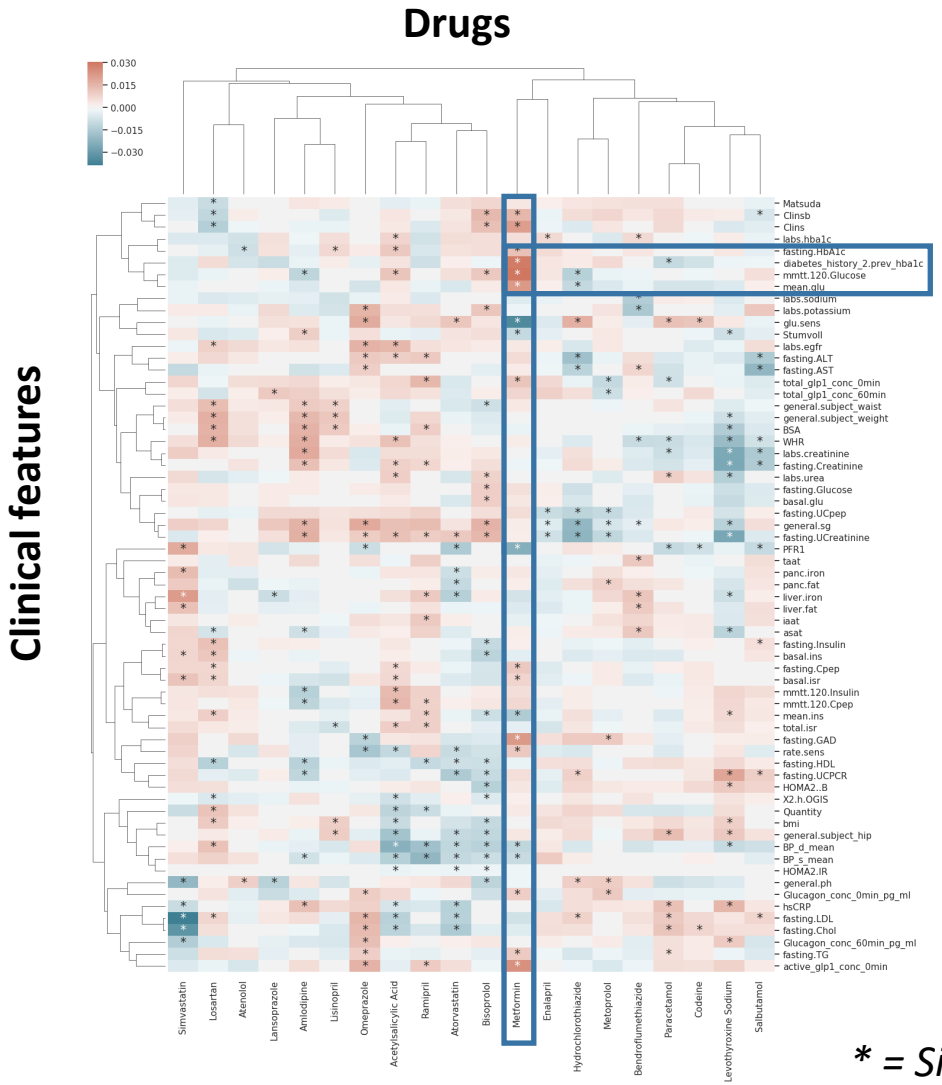
Statins ~ Clinical



- Simvastatin and Atorvastatin
- Used to treat high cholesterol
- Associated with fasting LDL and Chol
- Opposite associations on pancreatic fat

** = Significant change*

Metformin ~ Clinical

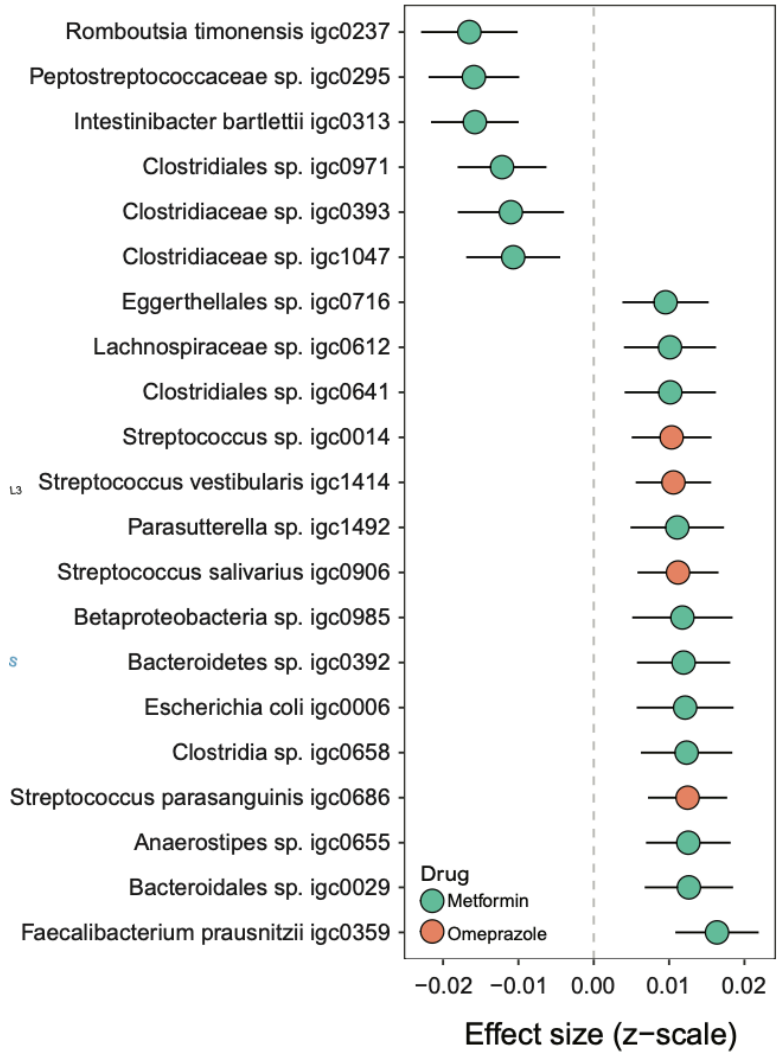


- Example Metformin associated with
  - HbA1c (blood glucose)
  - Other glucose measurements
- Confounding by indication

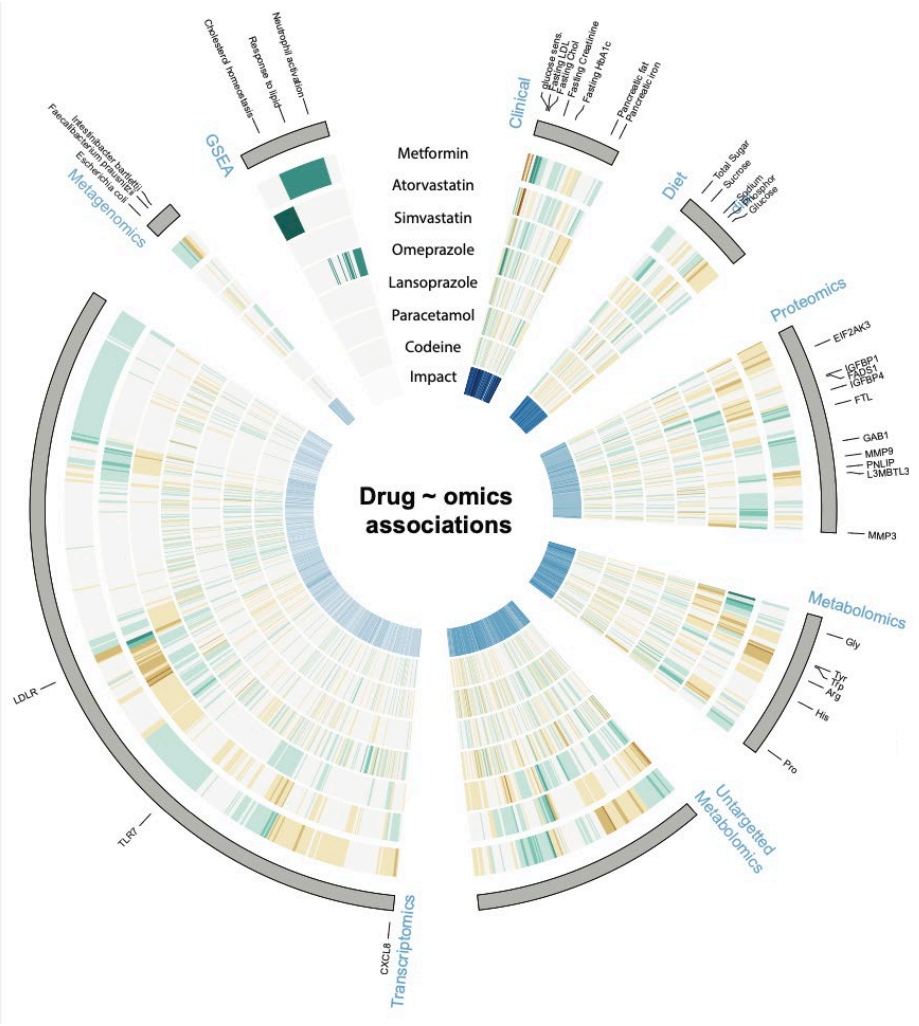
** = Significant change*

Drugs ~ microbiome

- Metformin:
  - 3 known associations (RCT)
  - 14 novel associations
- Omeprazole:
  - 4 Streptococcus sp. (previously shown towards genus)



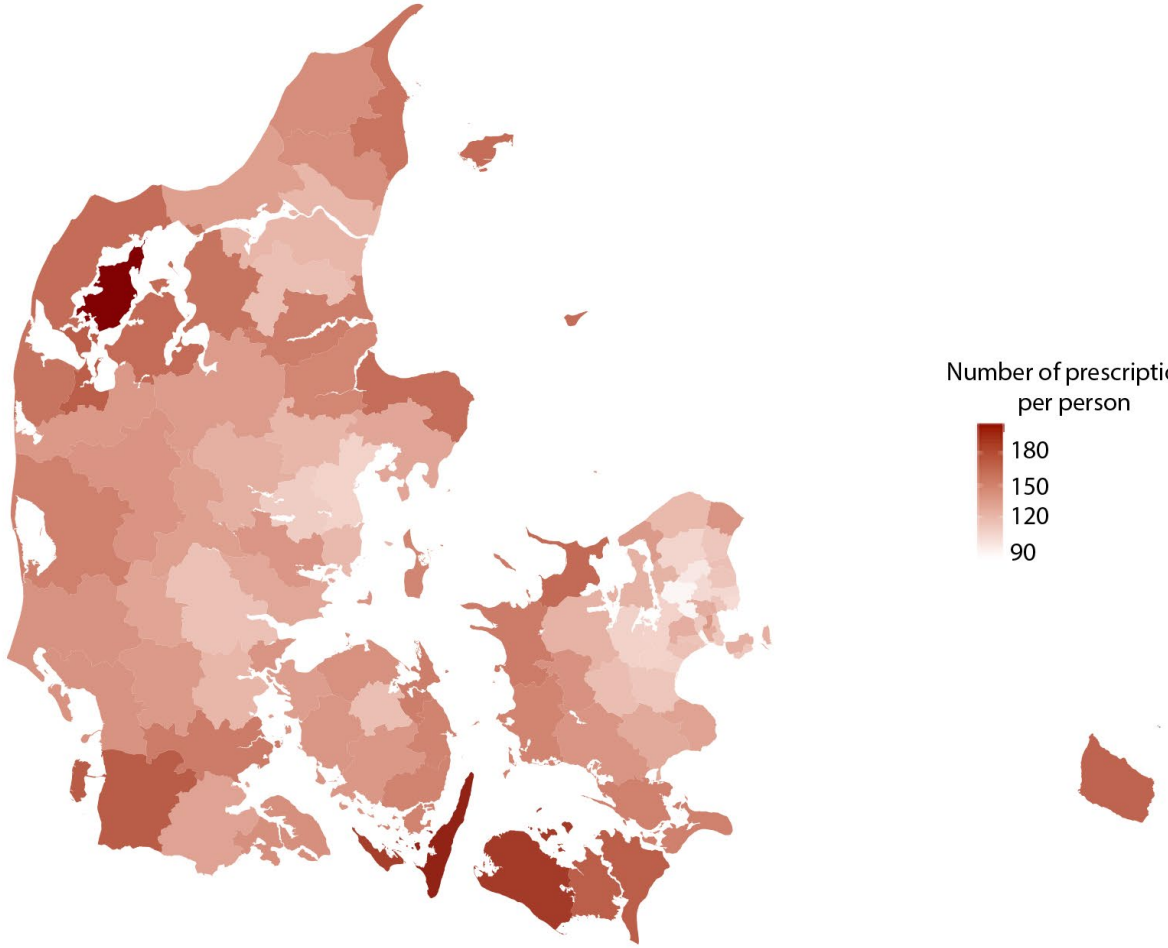
Multi-omics conclusions

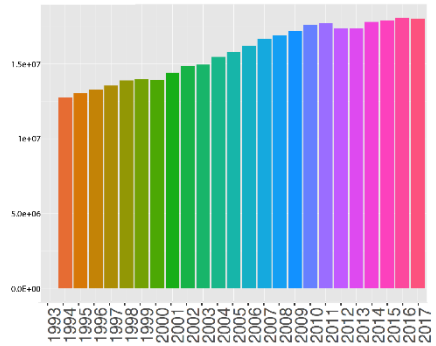


- DL based integration (MOVE):
  - Integrate continuous and categorial data
  - Extract orders of magnitudes more associations
- We only tested drugs!

Allesøe et al. Nature Biotechnology, Jan 2, 2023

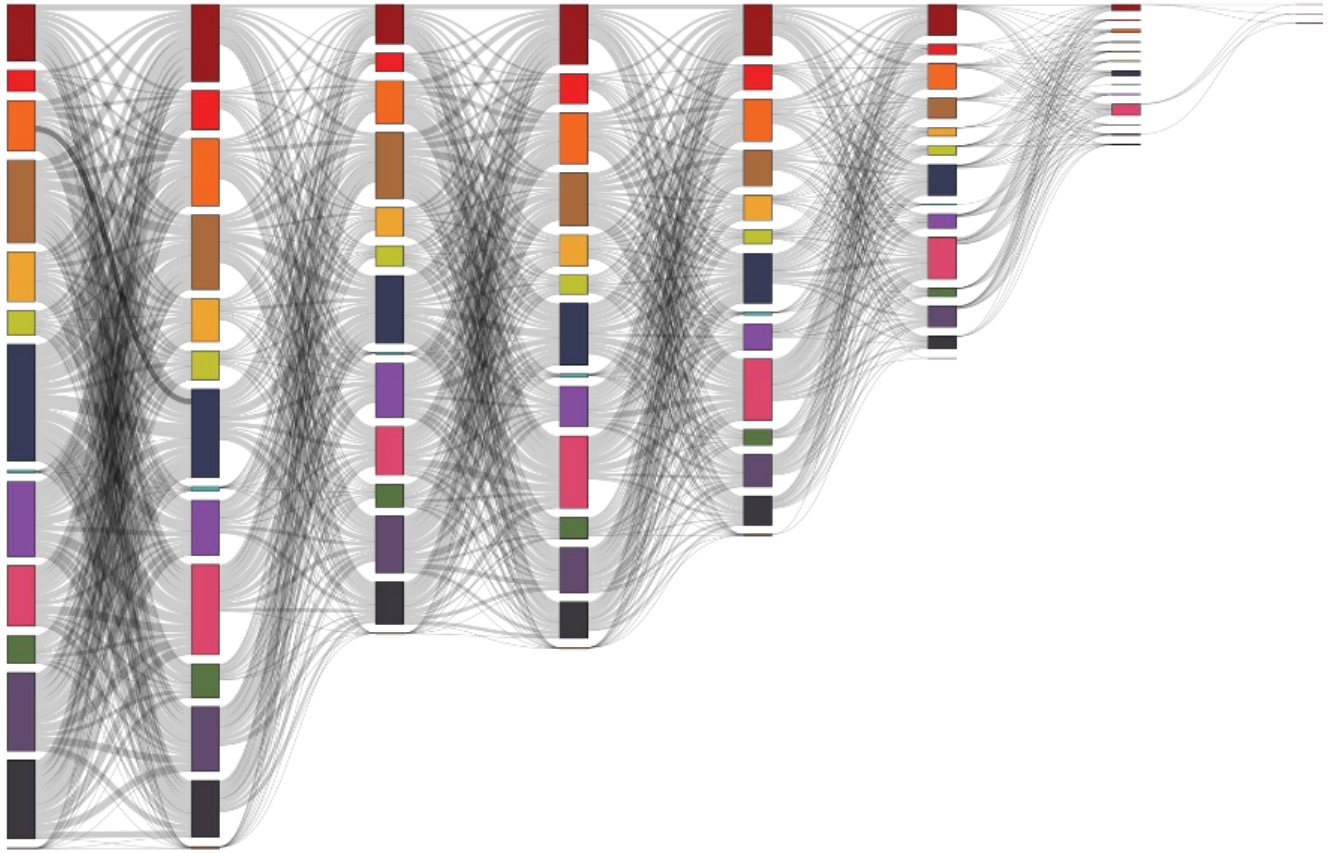
Longitudinal prescription data analysis
(per individual 1993-2018)





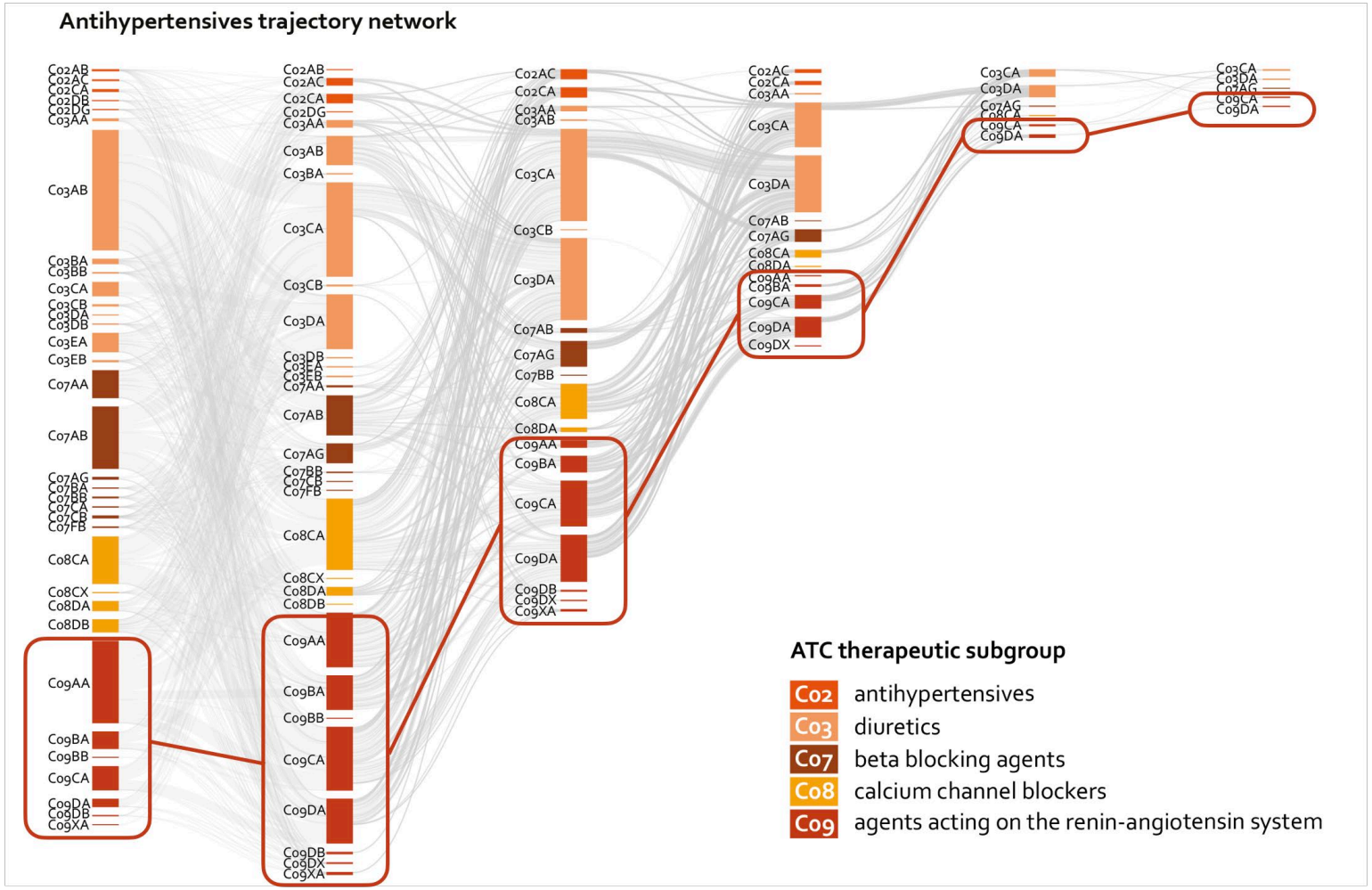
Prescription trajectories across ATC classes, 1.1 billion prescriptions, 24 years

Different ATC drug groups



Aguayo Orozco et al., npj Digital Medicine 2021

Multiple drug changes in patients treated with RAS

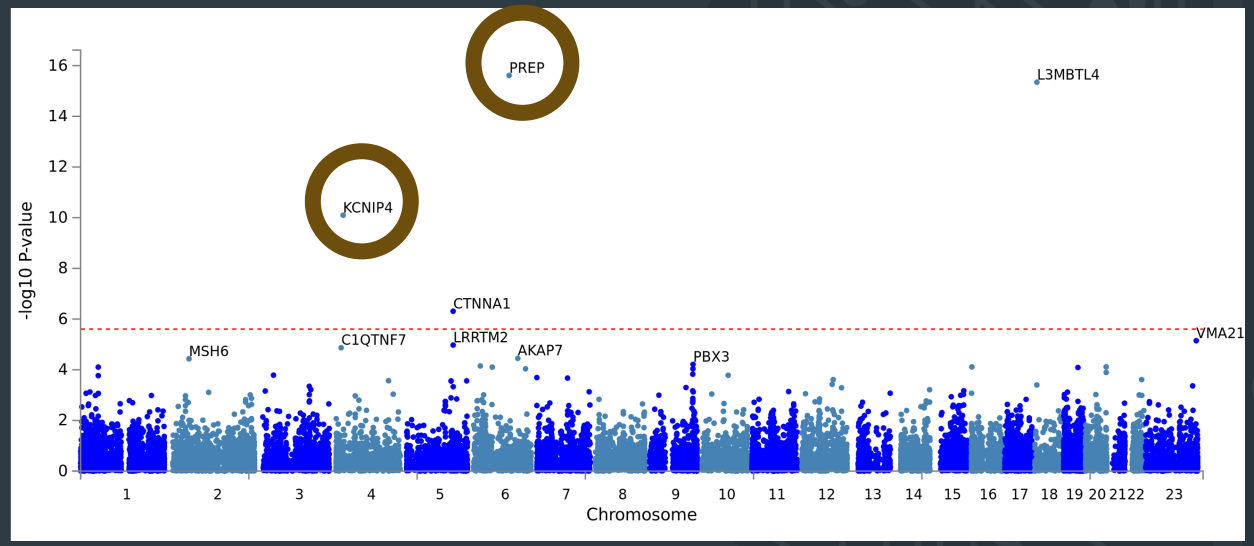


Renin-angiotensin system (RAS) drugs' first line treatment are ACE inhibitors (C09A and C09B)

These often require posterior changes to reach desired outcome

Other RAS used in posterior lines of treatment or by guideline defined patients are ARBs (C09C and C09D)

Genetic differences discovered via prescription trajectories



GWAS analysis of patients stratified according to wheather properly treated with ACE or wrongly treated with ACE (so they should change to ARB)

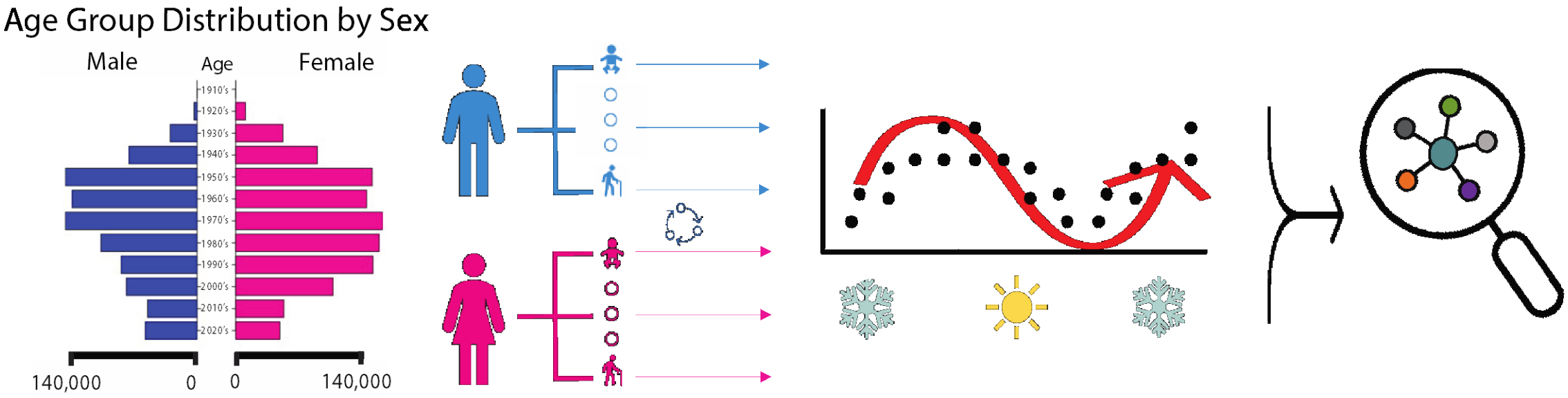
* PREP gene: responsible for maturation and breakdown of bradykinin

* KCNIP4 gene: previously associated with ACE associated side effects. These patients should be treated with ARB (1)

We identified genetic variants in the UKBB that potentially could explain the differences in the sub-stratified population.

1. JD Mosley et all, The Pharmacogenimics Journal, 2016

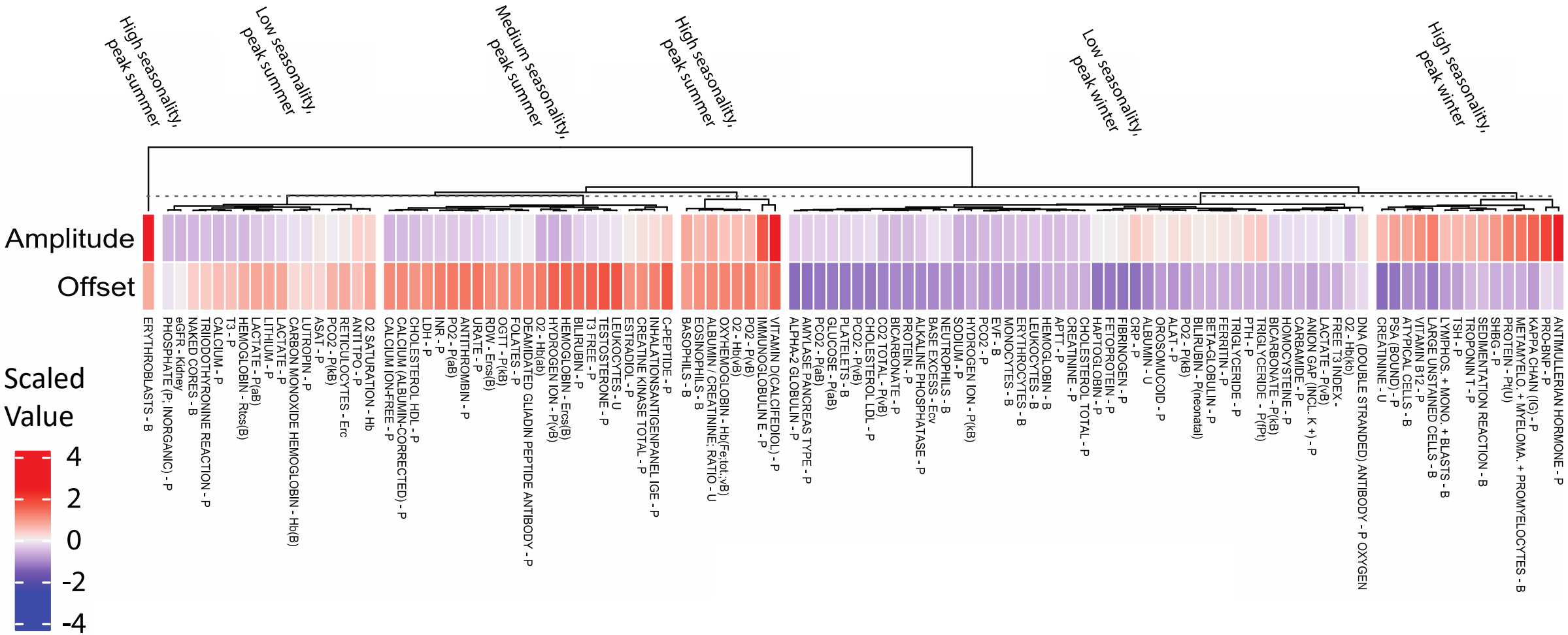
Lab value Seasonality Fitting Approach - 310M lab values



Muse et al., Cell Patterns 4, 100778, 2023

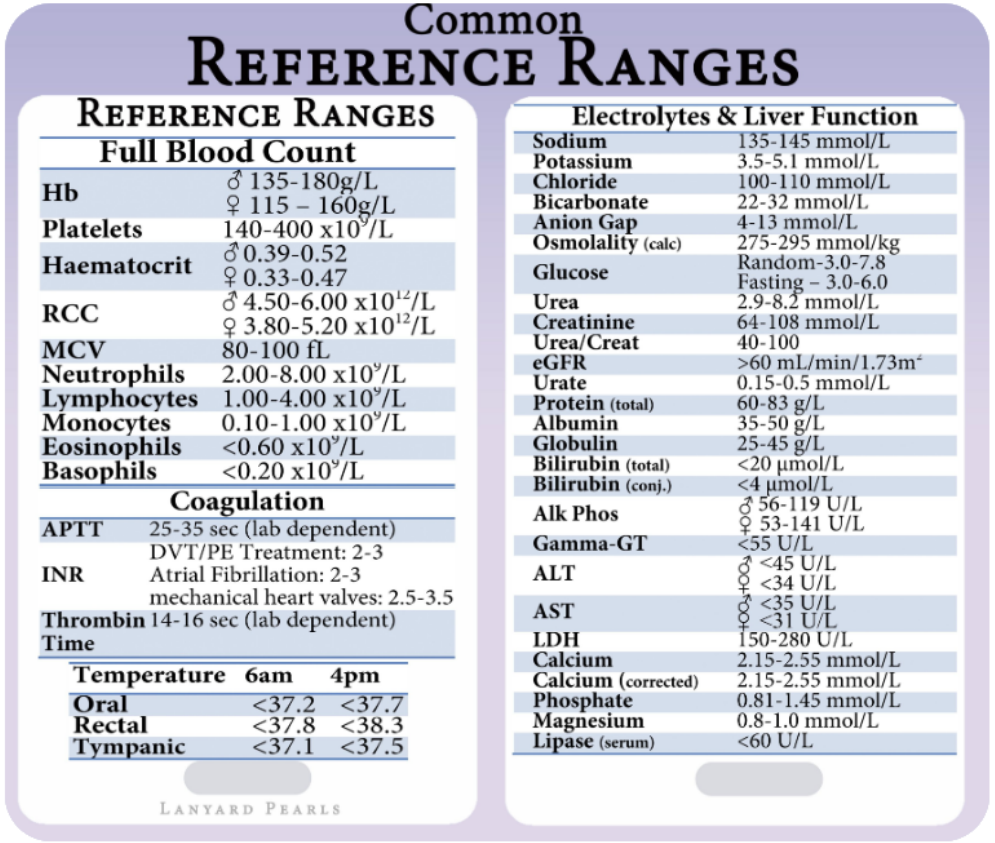
Muse et al., Cell Patterns, to appear 2023

Tests with significant parameter fits,
multiple testing corrected (FDR)



Note: Offset color scaled to what week of year peak occurs (red= summer, blue = winter)

Change reference values over the year



Common
REFERENCE RANGES

REFERENCE RANGES

Full Blood Count

Hb	♂ 135-180g/L ♀ 115 – 160g/L
Platelets	140-400 x10^9/L
Haematocrit	♂ 0.39-0.52 ♀ 0.33-0.47
RCC	♂ 4.50-6.00 x10^{12}/L ♀ 3.80-5.20 x10^{12}/L
MCV	80-100 fL
Neutrophils	2.00-8.00 x10^9/L
Lymphocytes	1.00-4.00 x10^9/L
Monocytes	0.10-1.00 x10^9/L
Eosinophils	<0.60 x10^9/L
Basophils	<0.20 x10^9/L

Coagulation

APTT	25-35 sec (lab dependent)
INR	DVT/PE Treatment: 2-3 Atrial Fibrillation: 2-3 mechanical heart valves: 2.5-3.5
Thrombin Time	14-16 sec (lab dependent)

Temperature	6am	4pm
Oral	<37.2	<37.7
Rectal	<37.8	<38.3
Tympanic	<37.1	<37.5

LANYARD PEARLS

Electrolytes & Liver Function

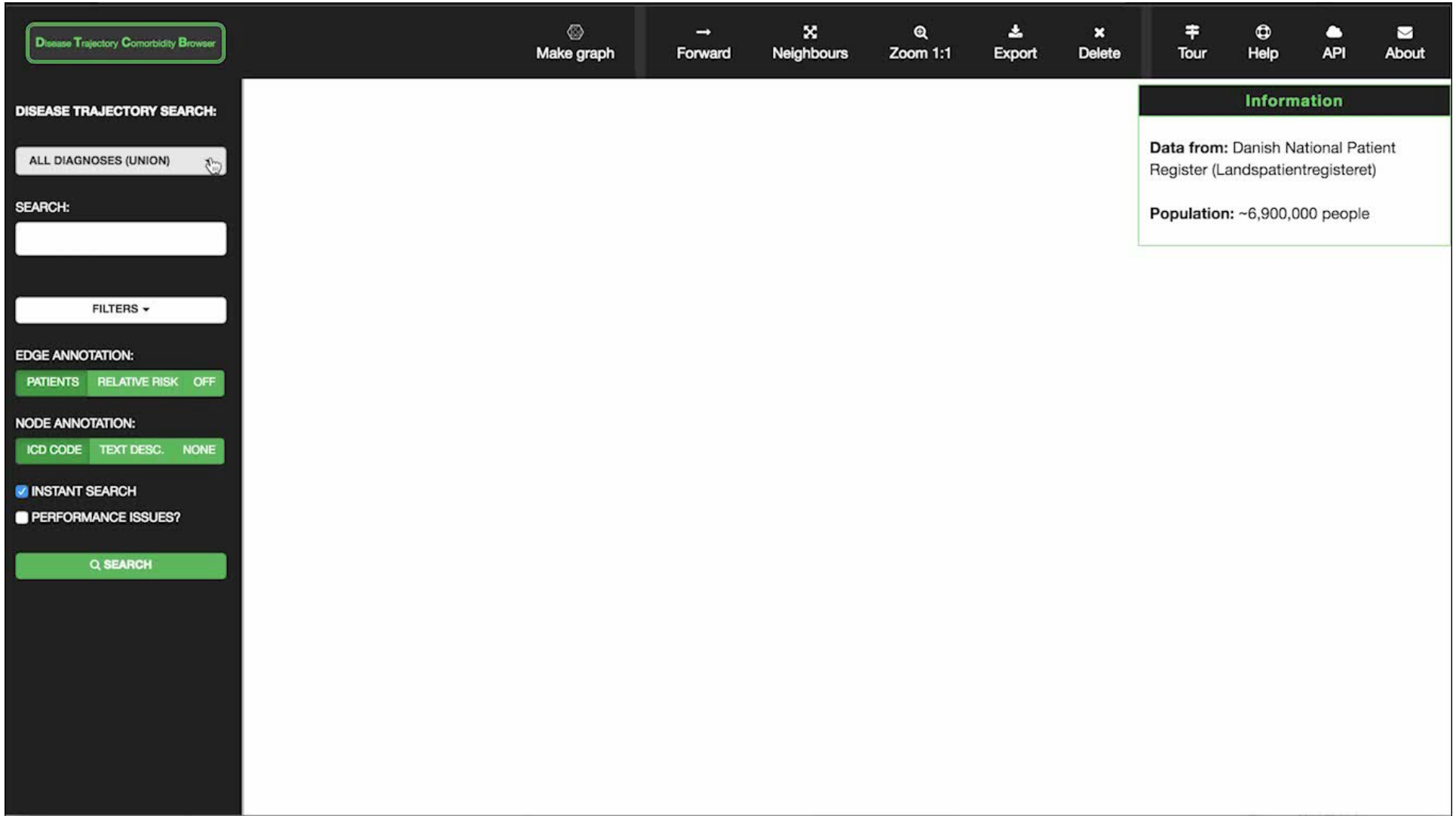
Sodium	135-145 mmol/L
Potassium	3.5-5.1 mmol/L
Chloride	100-110 mmol/L
Bicarbonate	22-32 mmol/L
Anion Gap	4-13 mmol/L
Osmolality (calc)	275-295 mmol/kg
Glucose	Random-3.0-7.8 Fasting – 3.0-6.0
Urea	2.9-8.2 mmol/L
Creatinine	64-108 mmol/L
Urea/Creat	40-100
eGFR	>60 mL/min/1.73m^2
Urate	0.15-0.5 mmol/L
Protein (total)	60-83 g/L
Albumin	35-50 g/L
Globulin	25-45 g/L
Bilirubin (total)	<20 µmol/L
Bilirubin (conj.)	<4 µmol/L
Alk Phos	♂ 56-119 U/L ♀ 53-141 U/L
Gamma-GT	<55 U/L
ALT	♂ <45 U/L ♀ <34 U/L
AST	♂ <35 U/L ♀ <31 U/L
LDH	150-280 U/L
Calcium	2.15-2.55 mmol/L
Calcium (corrected)	2.15-2.55 mmol/L
Phosphate	0.81-1.45 mmol/L
Magnesium	0.8-1.0 mmol/L
Lipase (serum)	<60 U/L

Improved estimation of mortality and diagnoses

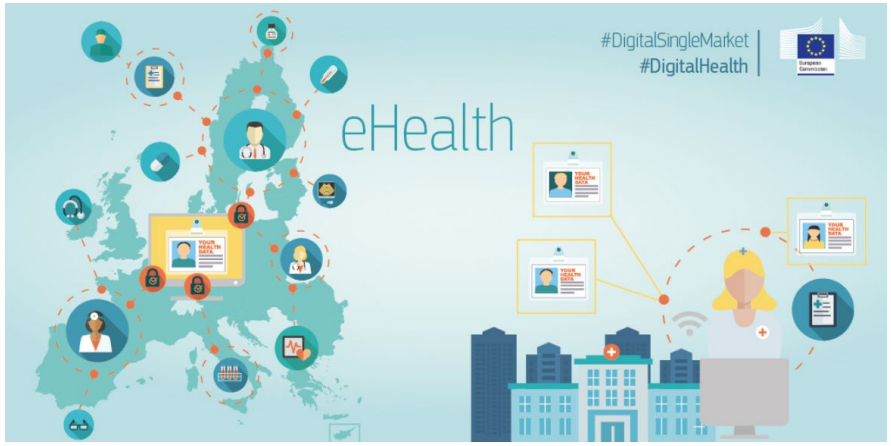
Muse et al., Cell Patterns 4, 100778, 2023

The Danish Disease Trajectory Browser:
http://dtb.cpr.ku.dk

Siggaard et al., Nature Comm, 2020




For a European
Health Data Space
#HealthUnion


#DigitalSingleMarket
#DigitalHealth
eHealth

European Commission

EHDS

欧州ヘルスデータスペースの規則案

～公表されたその概要について～

European Health Data Space

Δ population-wide health data

- Health data driven:
  - Redefine phenotypes as trajectories
  - Re-assign patients to the proper sub-category
  - Enable prediction using predictable trajectories?
  - Handle life long data capture
  - ”Live data” versus data dumps versus registers

- Include what is not in the hospital patient records in new ways:
  - Diet
  - Genetics
  - Income, …
  - Education, grades in exams, …
  - Wearable data (partly EHR included)
  - Patient generated data
  - Smart meter data

Acknowledgements