Big data indenfor sundhedsområdet og forbindelsen til ”personlig medicin”

Søren Brunak

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

Rigshospitalet
soeren.brunak@regionh.dk

The Novo Nordisk Foundation
Center for Protein Research

REGION H Rigshospitalet


DBU

IMPRECISION MEDICINE

For every person they do help (blue), the ten highest-grossing drugs in the United States fail to improve the conditions of between 3 and 24 people (red).

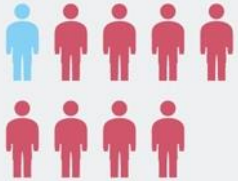
1. ABILIFY (aripiprazole)
Schizophrenia

2. NEXIUM (esomeprazole)
Heartburn

3. HUMIRA (adalimumab)
Arthritis

4. CRESTOR (rosuvastatin)
High cholesterol

5. CYMBALTA (duloxetine)
Depression



6. ADVAIR DISKUS (fluticasone propionate)
Asthma

7. ENBREL (etanercept)
Psoriasis

8. REMICADE (infliximab)
Crohn's disease

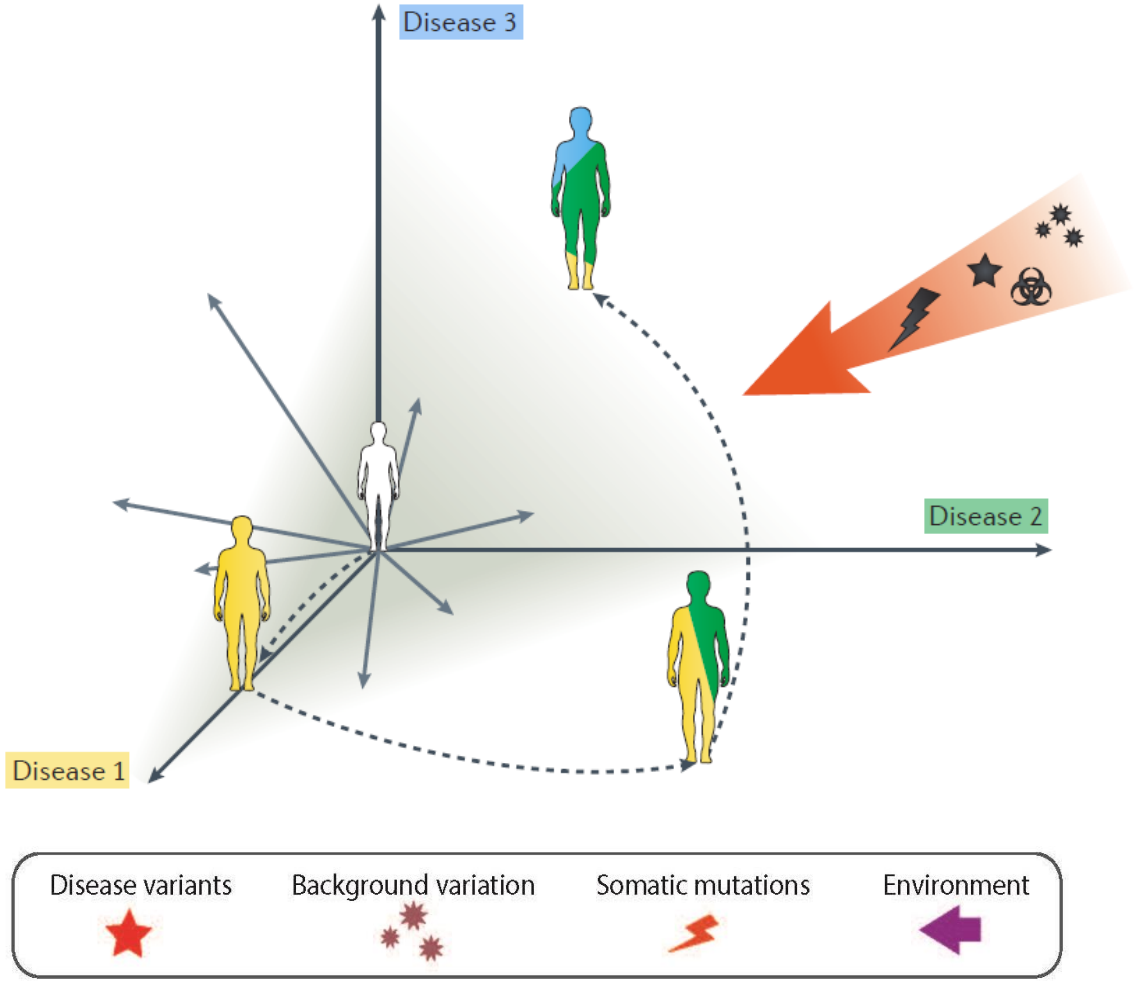
9. COPAXONE (glatiramer acetate)
Multiple sclerosis

10. NEULASTA (pegfilgrastim)
Neutropenia

Based on published number needed to treat (NNT) figures. For a full list of references, see Supplementary Information at go.nature.com/4dr78f.

Nature 520, 609–611 (30 April 2015)

Lifelong multimorbidity journeys in disease space

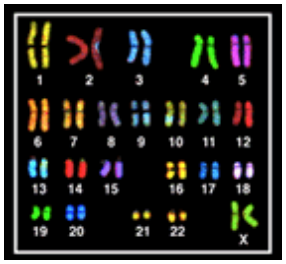


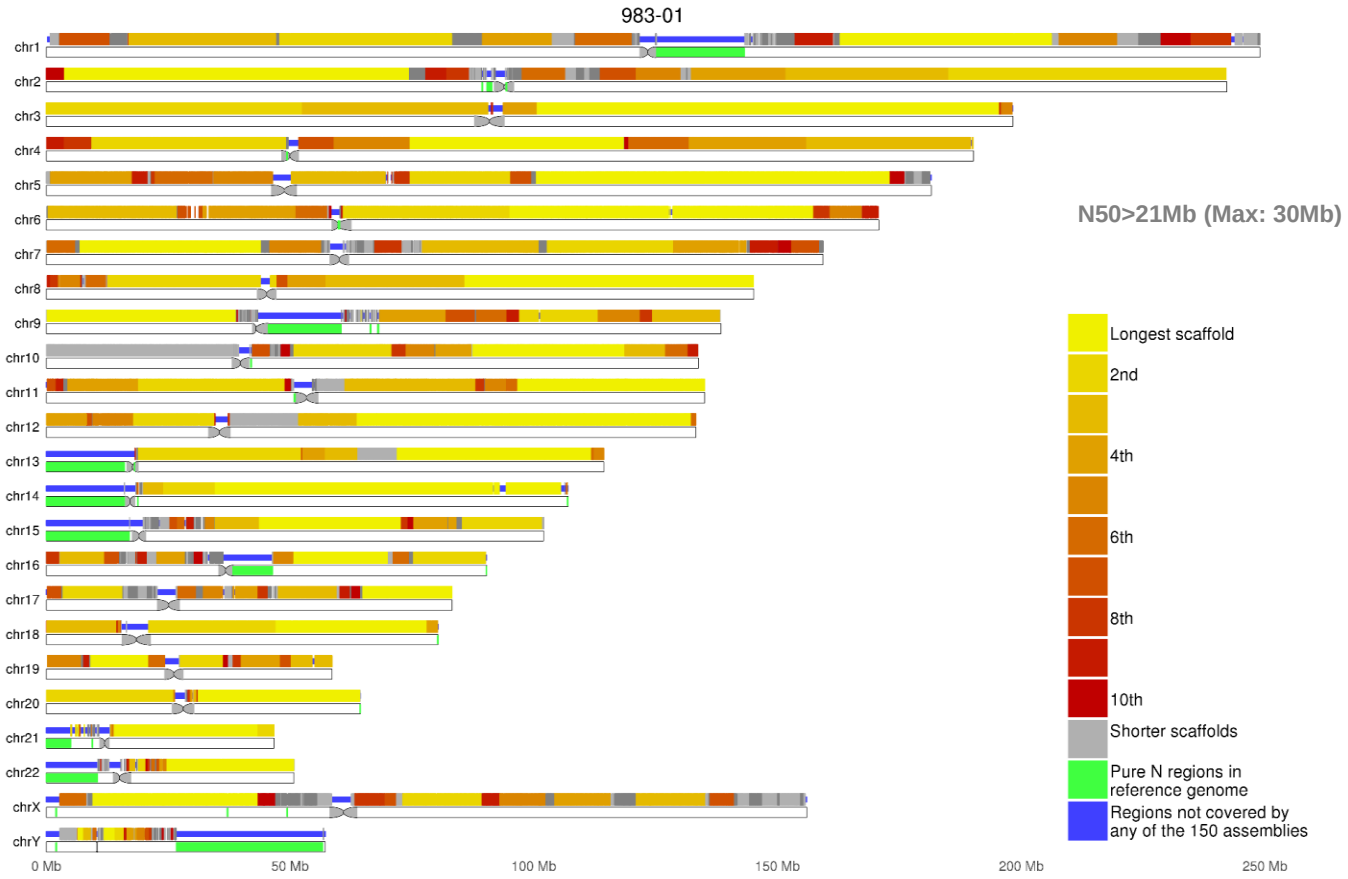
Hu, Thomas & Brunak
Nature Rev. Genetics 2016,

TCCAAACCCAGGCTCTCTCCCAAACCAGTTTGCGGCAGATGGCCAGTGGAACCTCACTCTCCTCATCAGTAAAAAGGGGGCAGAGTGAGGGTCCTGAGAGCTAGTACAGGGACTGTG
TGAAGTAGACAATGCCCAGTGTTTAGCGTAAGAATCAGGGTCCAGCTGGTGCTCCCTAAACAGCAGCTGCTGTTCACTGTTGAAAGGCGCTCTGGAAGGCCAGGCGCGGTGGCTCAT
GCTTGTAATCCCAGCACTGTGGGAGGCCGAGGTGGGCGGATCACCTGAGGTAGGGAGTTCGAGACCAGCCTGACCAACGTGGAGAAACCCCATCTCTCCTAAAAATACAAAATTAGC
CAGGCGTGGTAGCACATACCTGTAATCCCAGCGACTCGGGAGGCTGAGGCAAGAGAATTGCTTGAAACCAGCAGGGGAGGTTGTGGTGAGCCAAGATCGAGCCATTGCACTCCAGCC
AGGGCAACAAGAGGCAAAATGGCGAAACTCCATCTCCGAGAAAAAAAAAAAAAAAAGAATACTTTCTGAAAGTATTTATTCATACAAATAAAGACTTGACCCATAAGGTAGGAACGCA
AATGGGCCACGGAATCACTCATTCCACAGTATACACCGAGTGCCCTTGAAGTGCTGGGCACTGCTCCAGGATTGGGGGCATATTGGTGAAAAGAGAAGCAAGCCTGCCTGCTCAGAT
GGCAGGGAATGGGGAAAAACAGGGAGACAGTTTCCTGTTTGAGATGTTGGGAGTCTGCTTCGAGTAGTATATTTACTGGAAATAGACCACTAACTTGGATGTCCCTTTTTGGAAATG
TGCCTGCGTCCAGGGCTGGGTTGGGGCCCCAATGAACTTTGGCTCTGACATAGCTGTTGCCACACTCAGTGGAACTGAATCCATGTTTGCCTTCACCCGGCATCCTTCACCCCAACT
CTCCCCGCCACAACATACATCCCATGCCAGCCTGGGGACCCTCAAAGGTGCTTCATCATTAGGTTTGTGGCTGGGTCCTACTGAAGTAAGTCTTGGCACTCAGAGGGATAGGAATTG
AATGAAGACATGAGATTCCTCTGCGGGAGGCCTCTCTAGGAAATCTGTGGACTCACACGTTTACTAATGTTGCTGCAGCCCCGCACCCACCTTGGCCTTGGGCAGCCATACTCTAGG
GCTTTTGTAACCTCTCCATGTGAGGAACTCAAATTAGACCTGGGTTTGGAGGCGGTGCTCCGAGCTGGCCTTTGGGGGAGGTTTTGTGCGAGGCATTTCCCAAGTGCTGGCAGGATT
GTGTCACAGACACAGAGTAAACTTTTGCTGGGCTCCAAGTGACCGCCCATAGTTTATTATAAAGGTGACTGCACCCTGCAGCCACCAGCACTGCCTGGCTCCACGTGCCTCCTGGTC
TCAGTATGGCGCTGTCCTGGGTTCTTACAGTCCTGAGCCTCCTACCTCTGCTGGAAGCCCAGATCCCATTGTGTGCCAACCTAGTACCGGTGCCCATCACCAACGCCACCCTGGACC
GGGTGAGTGCCTGGGCTAGCCCTGTCCTGAGCACATGGGCAGCTGCCTCCCTTCTCTGGGCTTCCCTTTACCTGCTGGCTGTGGTCGCACCCCCACTCCCAGCTCTGCCTTTTTCTC
TTCTGGGTCCCCAGGGTGAAATTCTCACCAGCCCAGGGGACTCTGGAGGCACCCCCTGCCTCCAAACACAGAAGCCTCACTGCAGAGTCCTTCACGGAGGACGGTTCTGTGCTGGGC
CTGGAGGGGCTGCCTGGGGGGCAATGACTGATCCTCAGGGTGAGCTCCTGCATGCGCACTGCCCACCAGGGGCCTCATCTCCCCATCTGCAAAATCAGGGAGAGATCTGCCTGAGTC
TCCTCCCAGCTGACAGTCAAAGATTCAGCATCAAGCCCCCATCACCAGCTCCCCCCTTCTCCCCAGATCACTGGCAAGTGGTTTTATATCGCATCGGCCTTTCGAAACGAGGAGTAC
AATAAGTCGGTTCAGGAGATCCAAGCAACCTTCTTTTACTTTACCCCCAACAAGACAGAGGACACGATCTTTCTCAGAGAGTACCAGACCCGGTGAGAGCCCCCATTCCAATGCACC
CCCGATCTCAGCTGTCTGGCCAGAAGACCTGAGCAAGTCCCTCCTTCTTCCTGGCCTTGGCCTTCCCATGGGTGGAACCGGGAGGGTTGGCTTTAATCTCCACCAGAACTCTTGCCC
CGGGACTGTGATGGGCGATTGGCCACTTCTCCTCGATAACATTACTGTTTTTCTTCCGCCTTCTGGTTGACTTTAGCCAGAACCAGTGCTTCTATAACTCCAGTTACCTGAATGTCC
AGCGGGAGAATGGGACCGTCTCCAGATACGGTGAGGGCCAGCCCTCAGGCAGGAGGGTTCACCGTGGGAACAGGGCAGGCCAGCATAAGGTGGGGGCTGGATGTAGAGCCCTGGAGG
CTTTGGGCACAGAGAAATAACCACTAACATTTTTGAGCTCTTACCACGTGCTCAGAAAAAATCCCTAAGAAGACACTGAGAGAATTAGATGAGGAAACATAAGAACAGAGACCTCAA
ATAGTTTCCCCAAGGTCACACAGCTTATAATTAGAACTAGAATTGGAACTCCAGGCTGGCTTCAGATCTGCCTCTCTCTCACGCCCTCTTTAAGATCCTTTGCAAACCAATGGTAGA
AGCCTGTATGTTGGAGAGGTGGTACCTTCAACTATGTCCCCCATCACCGCAGAGGTGGCACATGGCAGGGATCTGATGGAGCTGAACTGACATCATTTAGCATCCCGAGCCTCCTCT
CTGGGCCTCATTTTCCTCCTCTGTAAAACGGGGAGAAAGGCCCTGACAGCCACAGTCTGTGTGAGGCTCCTGAGATCTCATGTACAGAAAGTGCTTGGCGTGGAGCTGGGCACGCAG
CAGGGGCTGGGCACACGGTGGCCCAAAGGAGACCCGGGCCTTCACTGATGGGCTTTGTGGCCCCGGACACATTTCTCTTCCAGAGGGAGGCCGAGAACATGTTGCTCACCTGCGTTC
CTTAGGGACACCCCTAGGACTCCTCACCTGTAAGACAGGCACCATTGTGCCATCCCATGTTCTCACCCAGAGGCTCTTAAGACCTTGATGTTTGGTTCCTACCTGGACGATGAGAAG
AACTGGGGGCTGTCTTTCTATGGTAGGCATGCTTAGCAGCCCCAAACTCATGCCCCTCTCAGGCCTCACCCCCCATTCACCCACCCCTGGGCTGGCCCCTAGAACCCCAGCCCTCCC
TGGCCTCCGCCGGGCCCCACCATGTCCCCAGTCAGTCTCCTTGCTCCCCCTGCAGCTGACAAGCCAGAGACGACCAAGGAGCAACTGGGAGAGTTCTACGAAGCTCTCGACTGCTTG
TGCATTCCCAGGTCAGATGTCATGTACACCGACTGGAAAAAGGTAAACGCAAGGGATTGGACATTGCCCACCTTGTCCATGGCCCAACTTGGGCAGCCCCAGAGGCCCAGAGCAGGA
AAGCTGCCAGGCAAGGCTGCACAGCTAGGCAGATCTTCTGCTTTTAGGCACCTGCCTCACTGTAGGGACAGCTGAGCTCTACAGAGGCCCAGGGGTGGTGGATGAGAGCCCAGGAGG
GAGAAGTCCCTGTGAAACCAGGGAGGACCTGAAAGCTAACAGGAGGGAACAGCGTGAGCCACGGGGTTGGGGGATTGGCAATTGGAGGGGACGTAATGCGGGGAGTTACCACCTACA
GACGCGTCCCAAACCCCAGGCTTTCACCCCAACCTCCACTCCCCGCTCATTTTTAATACCCGTGCAGTGGGGAATTGATACTGTGGTTTTCAATGTCACCCACACTGCAGCACGGCC
ACAGTCACCATCCCGATTTTTGCTACAAATGAAAATTACTGTATAATGAGCTCCTTAACACTTTTCTTTAAACCTGTGTTTGGAAGACTTGTGTTGGTGTGGCCCTGTGCCCTAATA
CCTGTGAAATCACAGCACCGATGAGCTGGTTCCAATTTTTAAAATATATACATGCAGTACTTCCATGACTATTCAAAGAAAAACAATTCCTTCCATTTGCCACCTGAGATGACCACC
AGGGATGTGAACTACCTCCTGCCCCATCCCCAGCCCCAGGATCCTGGGACAGGGCTTATGAACGCAACCACTGTAGTCAGCTCACTTGATCCACAGCCTGGCACCTCCACTGTCTGG
CTAGGGAGCCTCGAATGGGTCCCAAGGCCACCCTGCTCCTCAGTTACATCATCTGCATAGTAGTGGTGGTTGTGAGGAATTCAGGAGCTGCAGCATAAGGGCCCTGCAGGTACTATG
TGCTCAGTAAATGCCAGTGGTTCTTAAGGGTCTGAGCTCCCATTGTAGAGGCAAGTAAGCTGAGGTTCAGAGAAGAAAATGACTTGCCCAAGATCACCCAGCTGGGAAGTGACAGTG
CCAGGGTTGGAGCCCTGGTTGAGCTGGTTCCACAGGCCAGAGCTCATTCTGCCCTCTCCCCGGAAGACCTCCCACCCTGTCCCCATGCCTCTGCTTCTCCCTCACCCCAATTCCCCG
CTGCCTTCTAGGATAAGTGTGAGCCACTGGAGAAGCAGCACGAGAAGGAGAGGAAACAGGAGGAGGGGGAATCCTAGCAGGACACAGCCTTGGATCAGGACAGAGACTTGGGGGCCA
TCCTGCCCCTCCAACCCGACATGTGTACCTCAGCTTTTTCCCTCACTTGCATCAATAAAGCTTCGCATCGGCCTTTCGAAACGAGGAGTACAATAAGTCGGTTCAGGAGCCCTCAGG
CAGGAGGGTTCACCGTGGGAACAGGGCAGGCCAGCATAAGGTGGGGGCTGGATGTAGAGCCCTGGAGGCTTTGGGCACAGAGGCCACCCTGGACCGGGTGAGTGCCTGGGCTAGCCC
TGTCCTGAGCACATGGGCAGCTGCCTCCCTTCTCTGGGCTTCCCTTTACCTGCTGGCTGTGGTCGCACCCCCACTCCCAGCCCCAACTCTCCCCGCCACAACATACATCCCATGCC
CAGGAGGGTTCACCGTGGGAACAGGGCAGGCCAGCATAAGGTGGGGGCTGGATGTAGAGCCCTGGAGGCTTTGGGCACAGAGGCCACCCTGGACCGGGTGAGTGCCTGGGCTAGCCC

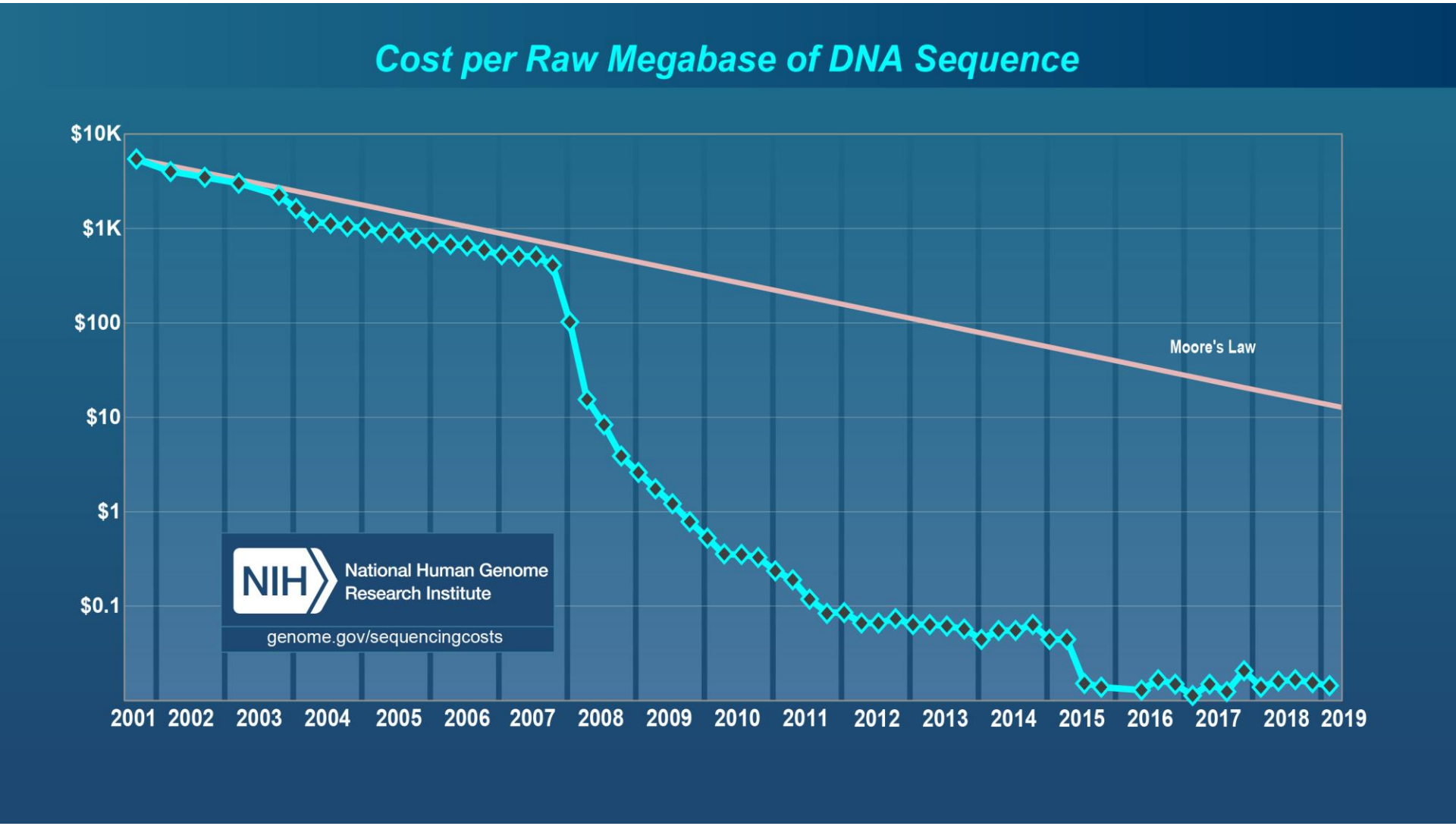
Genome based biomarkers for “all” diseases in one go

Individual high quality genome shown on the official reference for the human genome





Sequencing and de novo assembly of 150 genomes from Denmark as a population reference.
Maretty L, ..., Kristiansen K (*), Brunak S (*), Schierup MH (*). Nature, August 2017

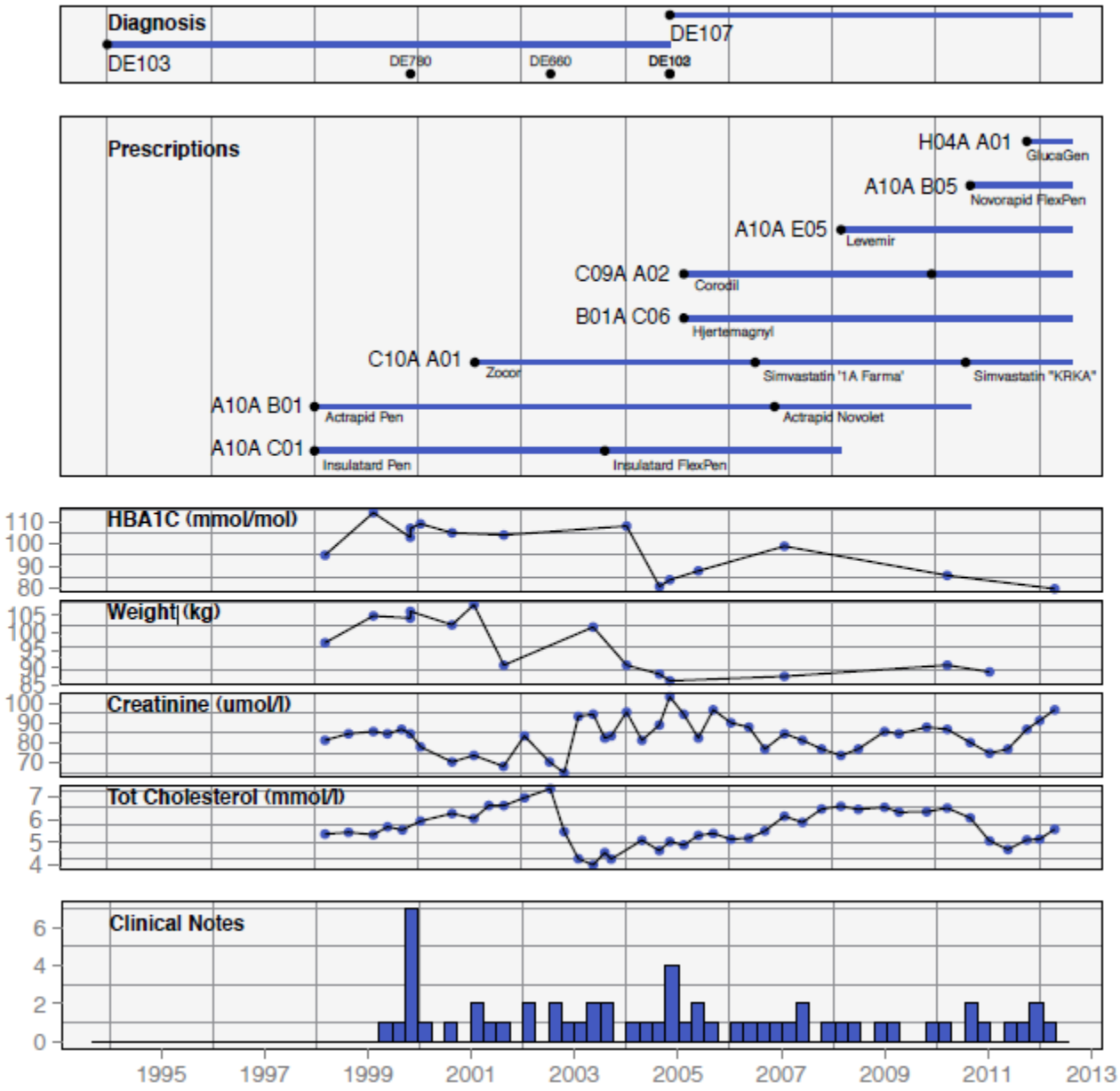

Cost per Raw Megabase of DNA Sequence
$10K
$1K
$100
$10
$1
$0.1
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019

Diagnoses

Drugs

Lab tests

Text



National Strategy for Personalised Medicine and the build-up of National Genome Center

GRAPH, June 8th 2018

Session on Big data and Personalised Medicine

Gert Sørensen

General Director, National Genome Center, Denmark

FOLKETINGSTIDENDE C

FOLKETINGET

Vedtaget af Folketinget ved 3. behandling den 29. maj 2018

Forslag

til

Lov om ændring af sundhedsloven

(Organiseringen i Sundheds- og Ældreministeriet, oprettelse af Nationalt Genomcenter m.v.)

§ 1

I sundhedsloven, jf. lovbekendtgørelse nr. 191 af 28. februar 2018, som ændret ved § 39 i lov nr. 620 af 8. juni 2016, § 1 i lov nr. 254 af 6. april 2018 og § 20 i lov nr. 503 af 23. maj 2018, foretages følgende ændringer:

1. I *§ 17, stk. 3,* indsættes før 1. pkt. som nyt punktum:
»En patient, der er fyldt 15 år, kan træffe beslutning om anvendelse af biologisk materiale og genetiske oplysninger, der er udledt af biologisk materiale, efter bestemmelserne i §§ 28-35.«

2. I *§ 17, stk. 3, 1. pkt.,* der bliver 2. pkt., indsættes efter », er«: »endvidere«.

3. *Overskriften* til kapitel 7 affattes således:

»Kapitel 7

Selvbestemmelse over biologisk materiale og genetiske oplysninger«.

4. I *§ 28, stk. 1,* 1. pkt., indsættes efter »§ 29, stk. 1,«: »1. pkt.,«.

5. I *§ 28, stk. 1,* 2. pkt., indsættes efter »§ 29, stk. 1,«: »1. pkt.,«.

6. I *§ 28, stk. 2,* 1. pkt., indsættes efter »§ 29, stk. 1,«: »1. pkt.,«, og »opbevaret biologisk materiale« ændres til: »oplysninger som nævnt i stk. 1«.

7. Efter § 29 indsættes før overskriften før § 30:

»Information om selvbestemmelse over genetiske oplysninger

§ 29 a. Forud for indhentning af patientens samtykke efter §§ 15 og 16 til en behandling, der omfatter genetisk analyse, skal den behandlende sundhedsperson informere patienten om retten til at træffe beslutning efter § 29, stk. 1, 2. pkt.

Stk. 2. Sundhedsministeren fastsætter nærmere regler om, hvordan patienten skal informeres om retten til at træffe beslutning efter § 29, stk. 1, 2. pkt.«

8. I *§ 32* indsættes efter »§ 29, stk. 1,«: »1. pkt.,«.

9. I *§ 32* indsættes som *stk. 2:*

»Stk. 2. Genetiske oplysninger, der er udledt af biologisk materiale i forbindelse med patientbehandling, og som opbe-

Creates a
National Genome Center
for precision medicine

~An Agency under the
Danish Health Ministry

- Written consent from the patient
- Right to self-determination of personal genetic data (opt-out)

22. Efter kapitel 67 indsættes i *afsnit XVII:*

»Kapitel 68

Nationalt Genomcenter

§ **223.** Nationalt Genomcenter er en institution under sundhedsministeren, som bistår ministeren med den centrale forvaltning af forhold vedrørende udviklingen af personlig medicin. Nationalt Genomcenter understøtter udviklingen af personlig medicin i samarbejde med det danske sundhedsvæsen, forskningsinstitutioner, patientforeninger m.v.

Stk. 2. Nationalt Genomcenter udvikler og driver fælles, landsdækkende informationsinfrastruktur for personlig medicin, herunder en landsdækkende infrastruktur til udførelse af genomsekventering og opbevaring af oplysningerne i en national genomdatabase. Nationalt Genomcenter stiller oplysninger til rådighed for personer inden for sundhedsvæsenet og patienter, herunder oplysninger fra den fælles, nationale genomdatabase til brug for patientbehandling m.v.

§ **223 a.** Sundhedsministeren kan fastsætte regler om, at der påhviler institutioner under Sundheds- og Ældreministeriet, regionsråd, kommunalbestyrelser, autoriserede sundhedspersoner og de private personer eller institutioner, der driver sygehuse m.v., en pligt til at give Nationalt Genomcenter genetiske oplysninger, som er udledt af biologisk materiale efter oprettelsen af Nationalt Genomcenter, og oplysninger om helbredsmæssige forhold, i det omfang oplysningerne er nødvendige for gennemførelsen af centerets opgaver.

Stk. 2. Sundhedsministeren kan fastsætte nærmere regler om, at borgere frivilligt kan overlade genetiske oplysninger, som er udledt af biologisk materiale, til Nationalt Genomcenter.

§ **223 b.** Oplysninger, der tilgår Nationalt Genomcenter, herunder genetiske oplysninger og oplysninger om helbredsmæssige forhold, må kun behandles, hvis det er nødvendigt med henblik på forebyggende sygdomsbekæmpelse, medicinsk diagnose, sygepleje, patientbehandling eller forvaltning af læge- og sundhedstjenester og behandlingen af oplysningerne foretages af en person inden for sundhedssektoren, der efter lovgivningen er undergivet tavshedspligt, eller hvis behandlingen alene sker med henblik på at udføre statistiske eller videnskabelige undersøgelser af væsentlig samfundsmæssig betydning og behandlingen er nødvendig af hensyn til udførelsen af undersøgelserne.

Stk. 2. Oplysninger som nævnt i stk. 1 er ikke genstand for edition efter retsplejelovens § 804, medmindre der er tale om efterforskning af en overtrædelse af straffelovens § 114 eller § 114 a.«

Establish a national infrastructure for whole genome sequencing and a single National database for storage

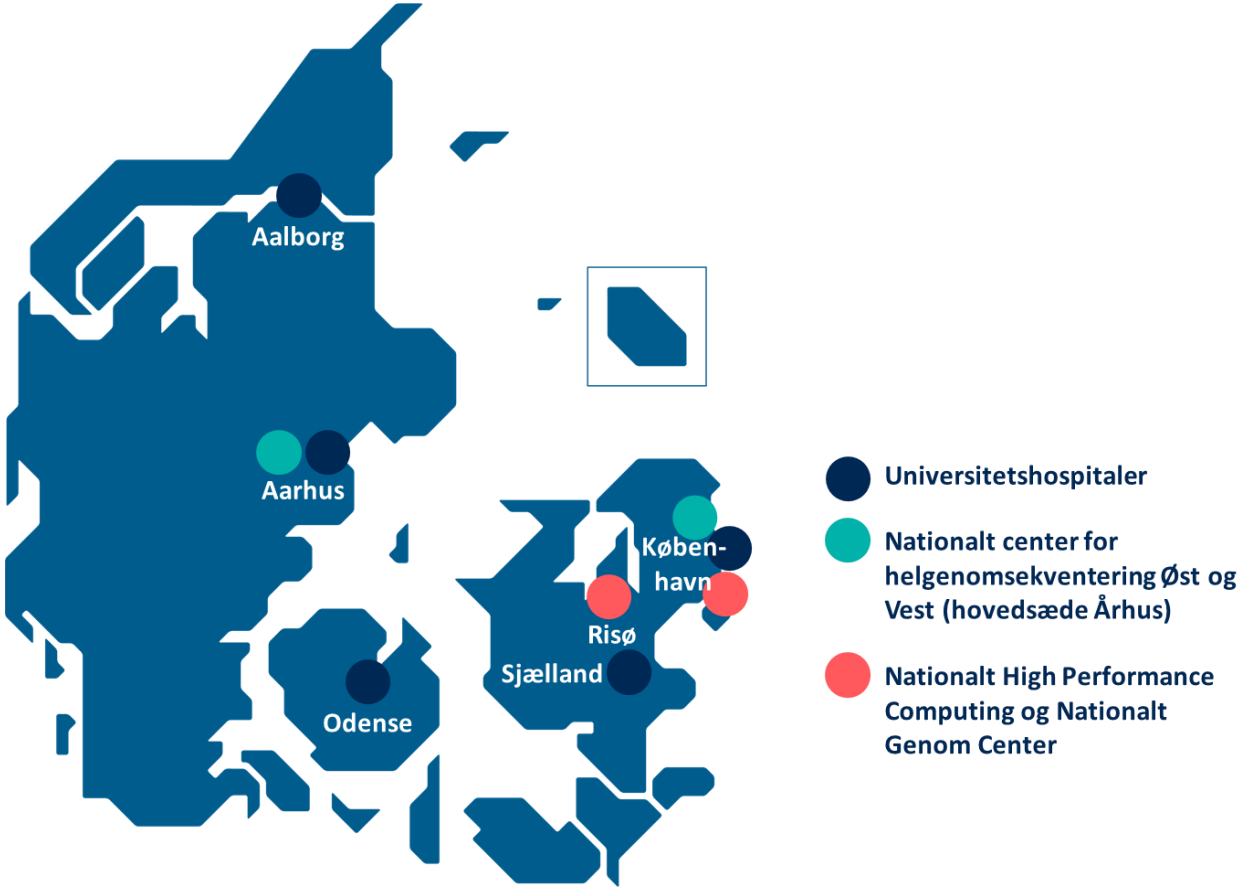
Compulsory storing of data generated in the health system

May also receive “citizen” generated data

Make data available to the healthcare system and for research

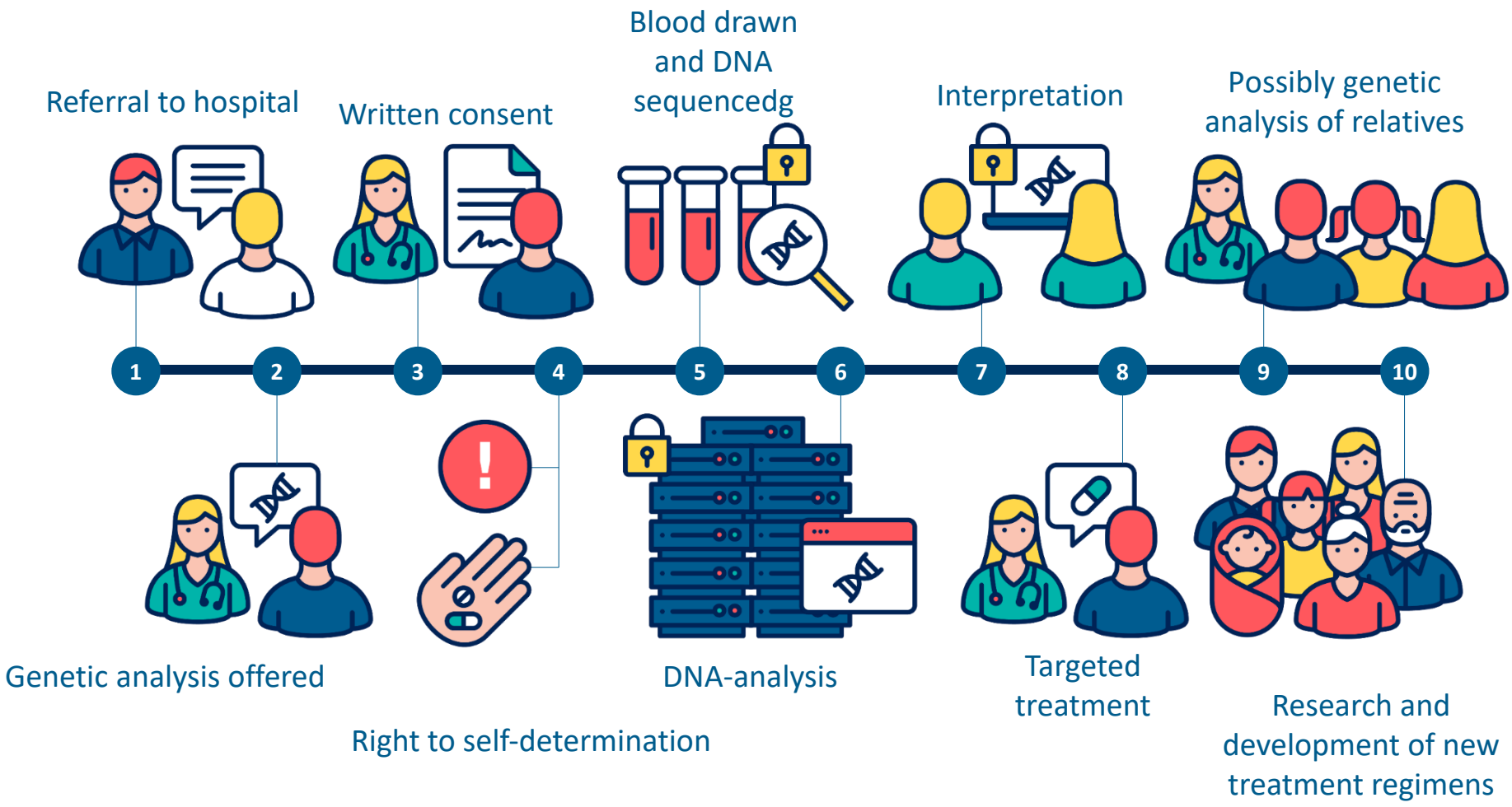
No police access except in cases of terrorism

NGC National HPC and WGS Centers – cooperation and governance



How do we ensure that the services of the National WGS Center are attractive for all relevant Danish hospitals, and with time the GP sector?

Steps related to genetic analysis via the National Genome Center in Denmark



COMPUTEROME 1.0

#1 Powerful supercomputer in Denmark 2014 - dedicated to life science

Computing Power: 19,000+ cores

- 525 x Thin nodes
- 28 core Intel CPUs with 128 GB DDR4
- GPU Node – NvidiaK40,P100
- 29 x FAT nodes
- 32 core Intel CPUs with 1TB DDR3 RAM
- 1 x High Memory node with 8TB memory

Storage: 8PB+

- 8 PB (8192 TB) extremely high performance storage
- +50GB/s aggregated bandwidth

COMPUTEROME 2019

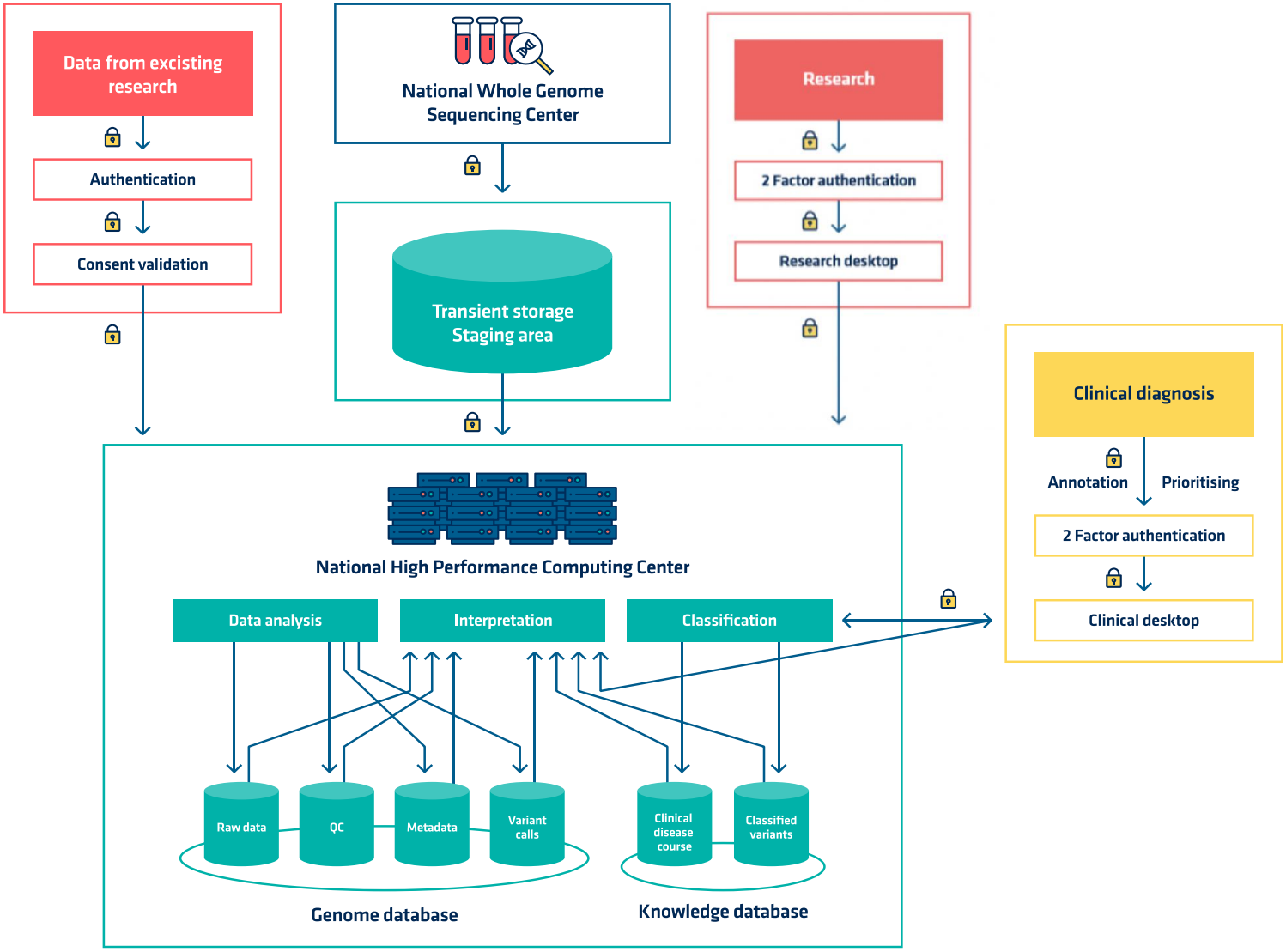

DANISH NATIONAL
GENOME CENTER
Brandvej
102, 134


DANISH NATIONAL
GENOME CENTER


MobilLoft A/S
44 66 38 00
DTU
DeiC


DTU
DTU

NGC – different routes for patient data and research project data



When One Diagnosis Is Not Enough

Kym M. Boycott, M.D., Ph.D., and A. Micheil Innes, M.D.

An accurate diagnosis is essential for effective medical management; in the case of rare genetic disease, it also guides genetic counseling. Nevertheless, clinical assessments and conventional genetic testing lead to a diagnosis in less than half of patients.[1] The introduction of whole-exome sequencing has substantially improved our ability to provide patients with a molecular diagnosis and is increasingly the approach of choice for patients with rare diseases. With the genes known for more than half of the predicted 7000 rare diseases,[2] such hypothesis-free approaches to diagnosis are attractive and have the potential to end the "diagnostic odyssey."

As clinicians, we are trained to identify a single explanation for a patient's clinical presentation; not only is the idea of more than one genetic disease in a patient difficult to fathom

N ENGL J MED 376;1 NEJM.ORG JANUARY 5, 2017

The NEW ENGLAND JOURNAL of MEDICINE

— it also seems unfair. Posey and colleagues, however, now report in the *Journal* a retrospective analysis of 2076 patients with a molecular diagnosis (28% of the patients referred to them for whole-exome sequencing), 101 (4.9%) of whom had two to four diagnoses.[3] This finding challenges the notion that a genetic investigation is complete after a single diagnosis is obtained.

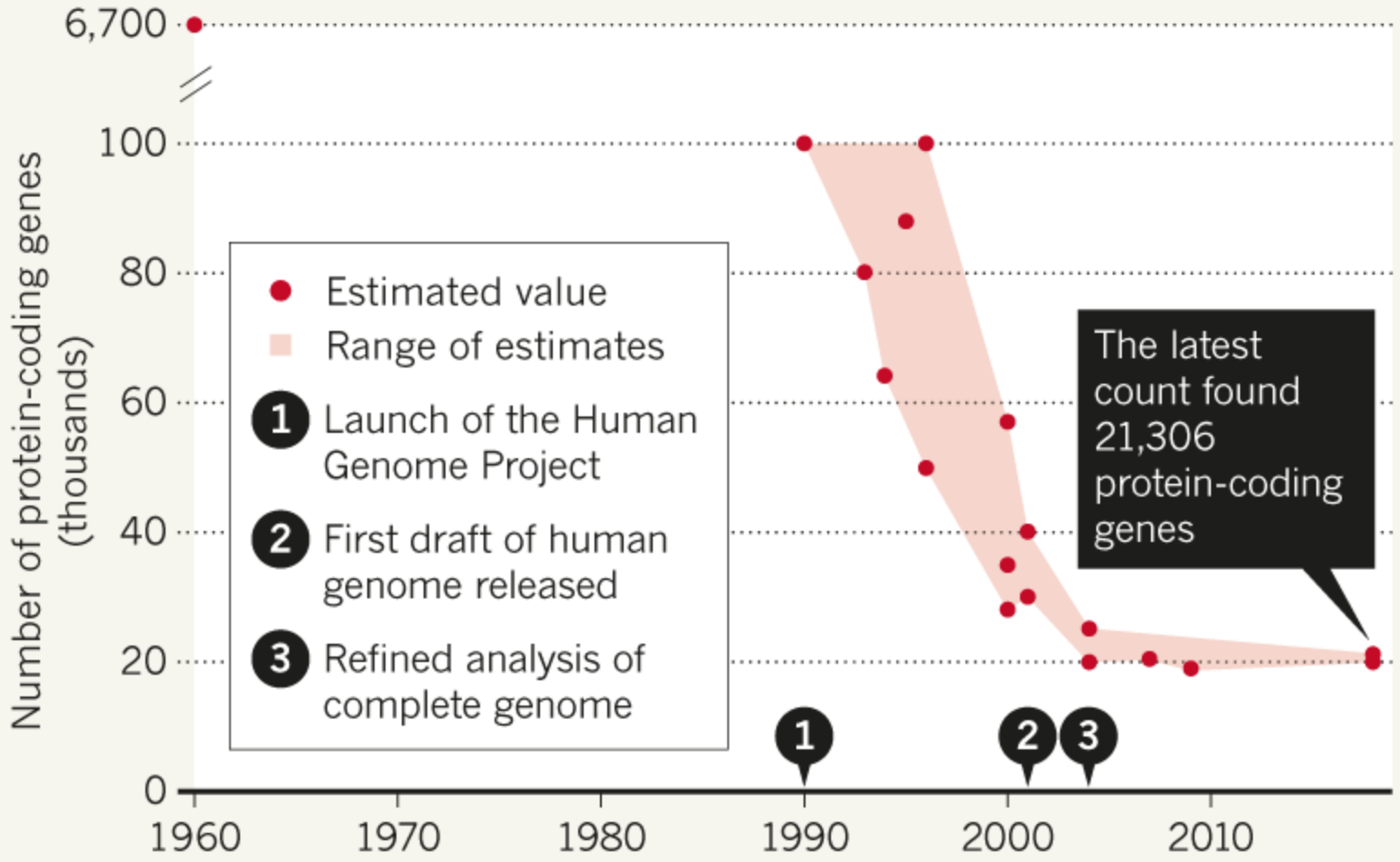
It is intuitive that in a patient with two or more genetic diagnoses, the features of the diseases will manifest in a "blended" fashion, in which

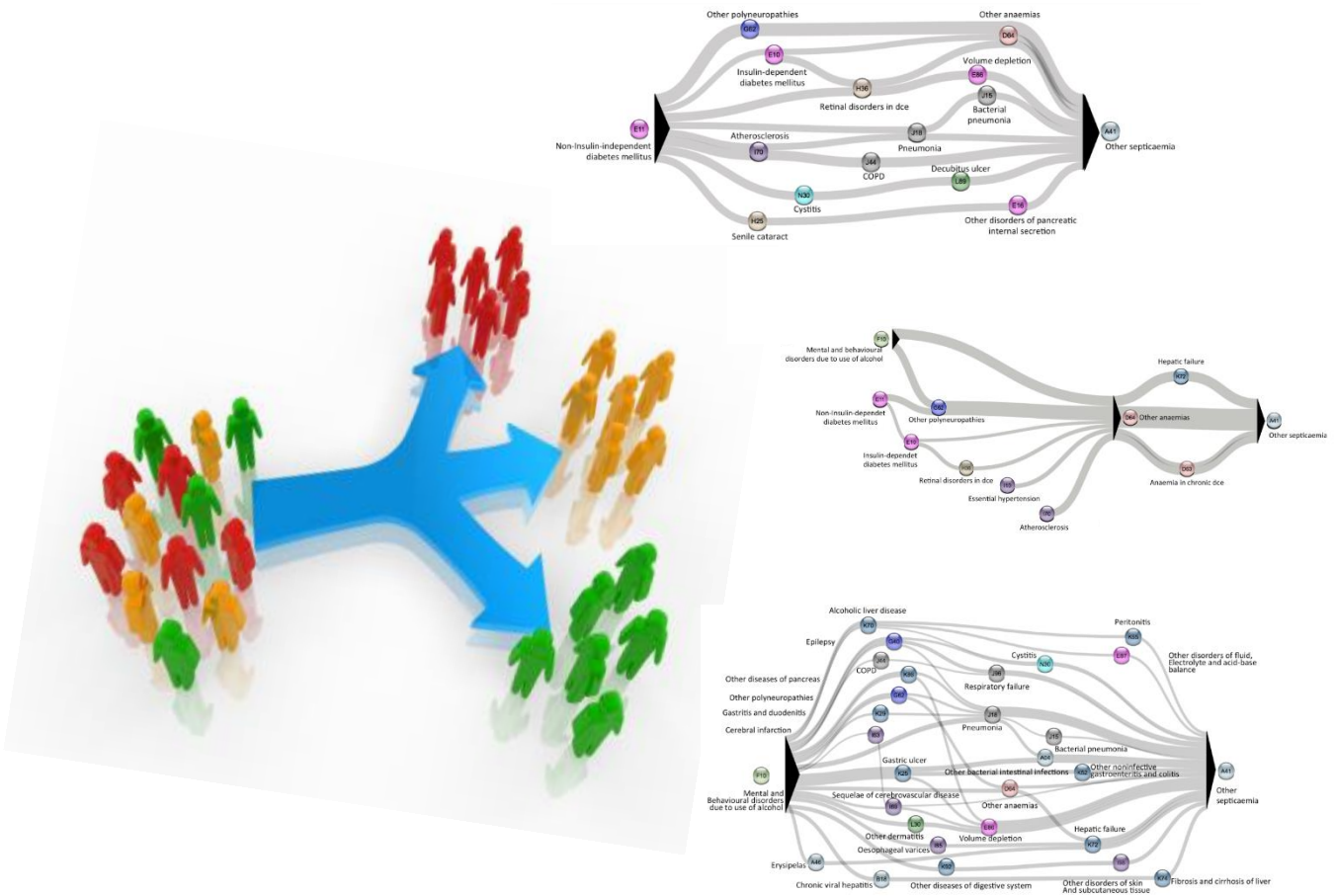
found in their patients with dual diagnoses; scores were lower if the two diseases affected different organ systems than they were when the diseases had overlapping features. The HPO also provides a structured vocabulary to fully document a patient's clinical presentation, termed "deep phenotyping." Taking this one step further, one can envision a computational mechanism by which a patient's HPO description is compared against a disease-phenotype annotation list and a prediction is made as to whether this patient's

NEJM 2017

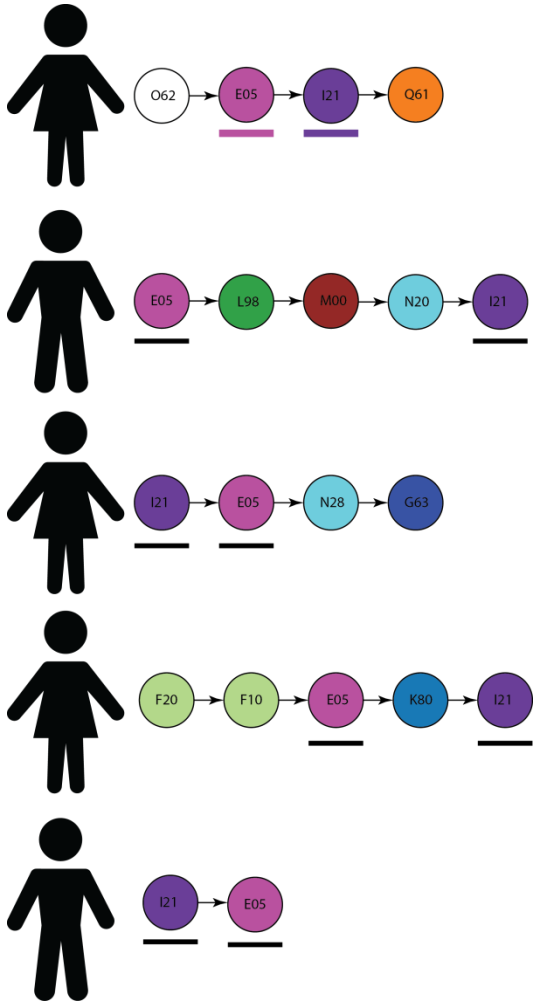
GENE TALLY

Scientists still don't agree on how many protein-making genes the human genome holds, but the range of their estimates has narrowed in recent years.




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
Respiratory failure
Other diseases of pancreas
Other polyneuropathies
Gastritis and duodenitis
Pneumonia
Cerebral infarction
Bacterial pneumonia
Gastric ulcer
Other noninfective gastroenteritis and colitis
Other bacterial intestinal infections
Mental and Behavioural disorders due to use of alcohol
Other septicaemia
Sequelae of cerebrovascular disease
Other anaemias
Hepatic failure
Other dermatitis
Volume depletion
Oesophageal varices
Erysipelas
Chronic viral hepatitis
Other diseases of digestive system
Other disorders of skin And subcutaneous tissue
Fibrosis and cirrhosis of liver

Diagnosis trajectories across millions of Danes using ICD

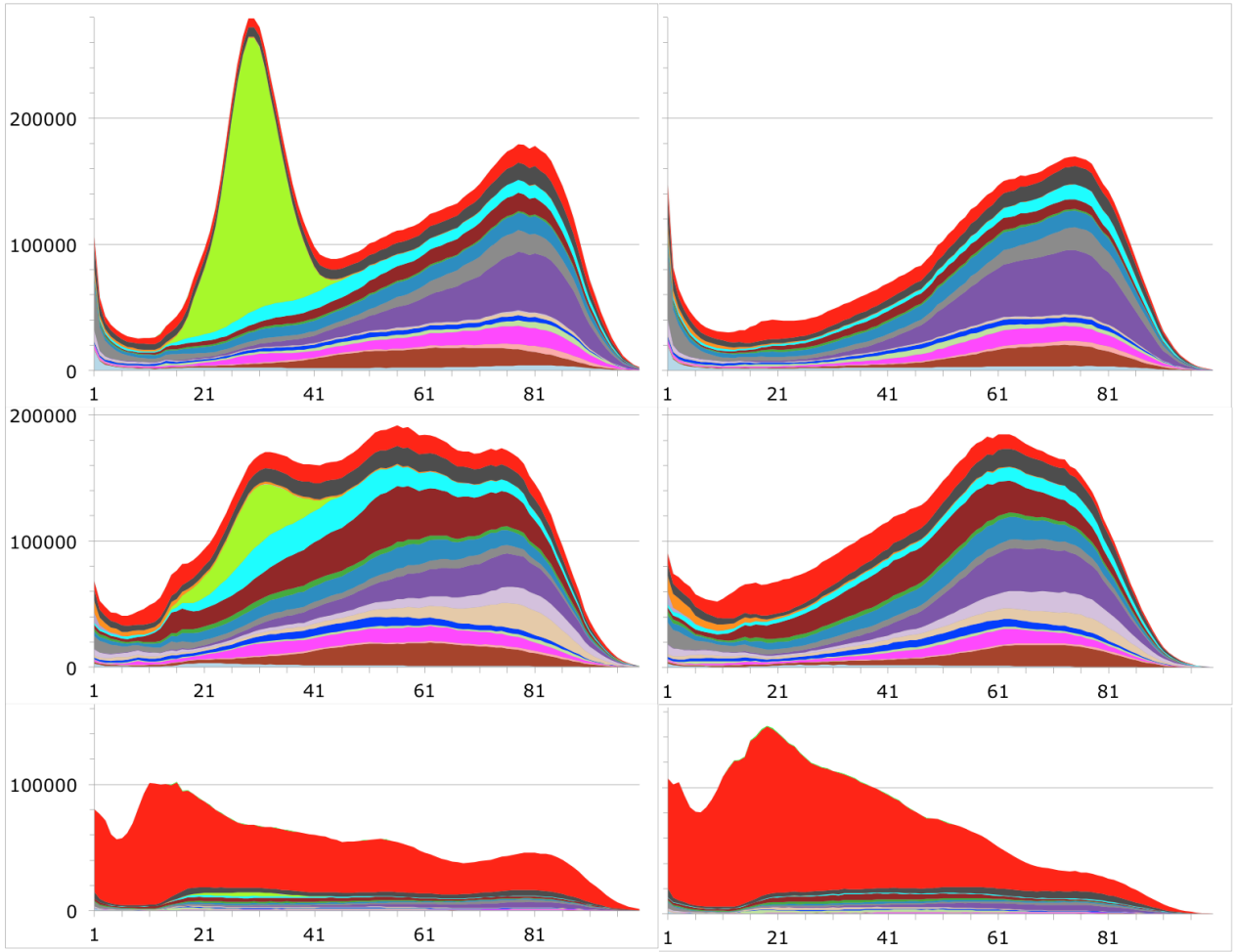


ICD 10 chapter coloring

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

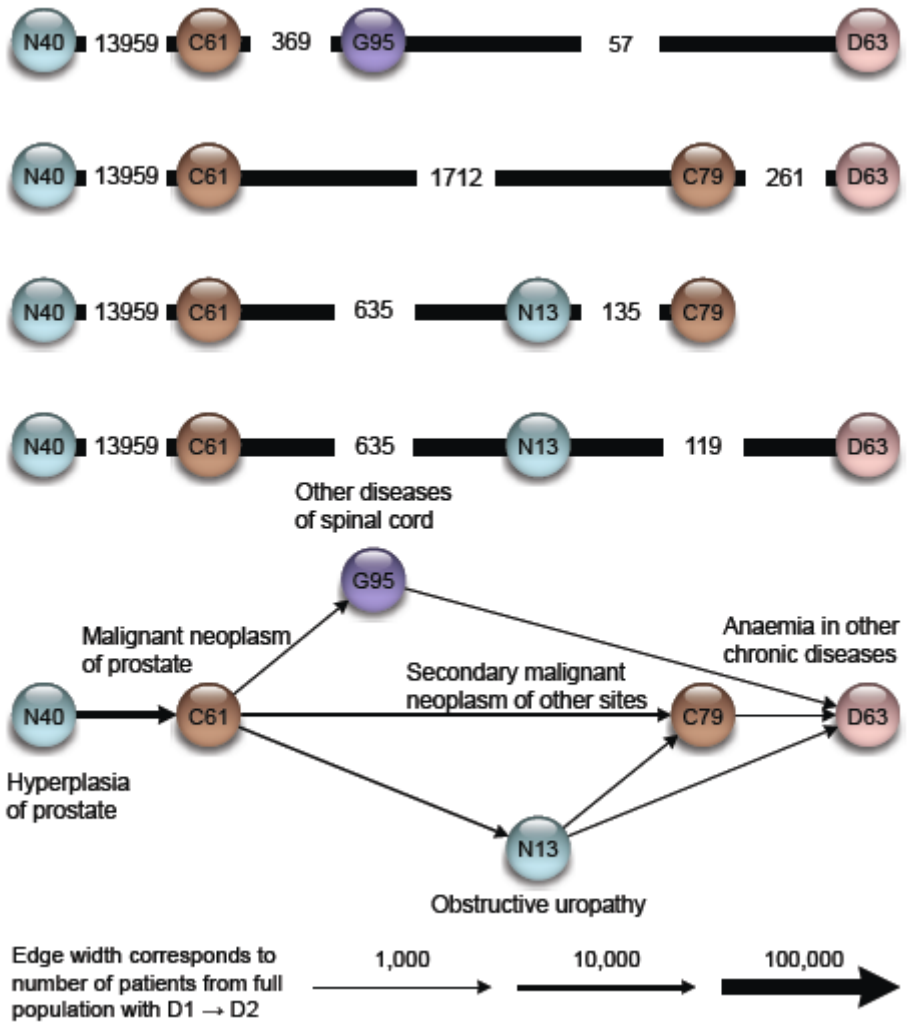
(ICD-10 era, 1994-2017)



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

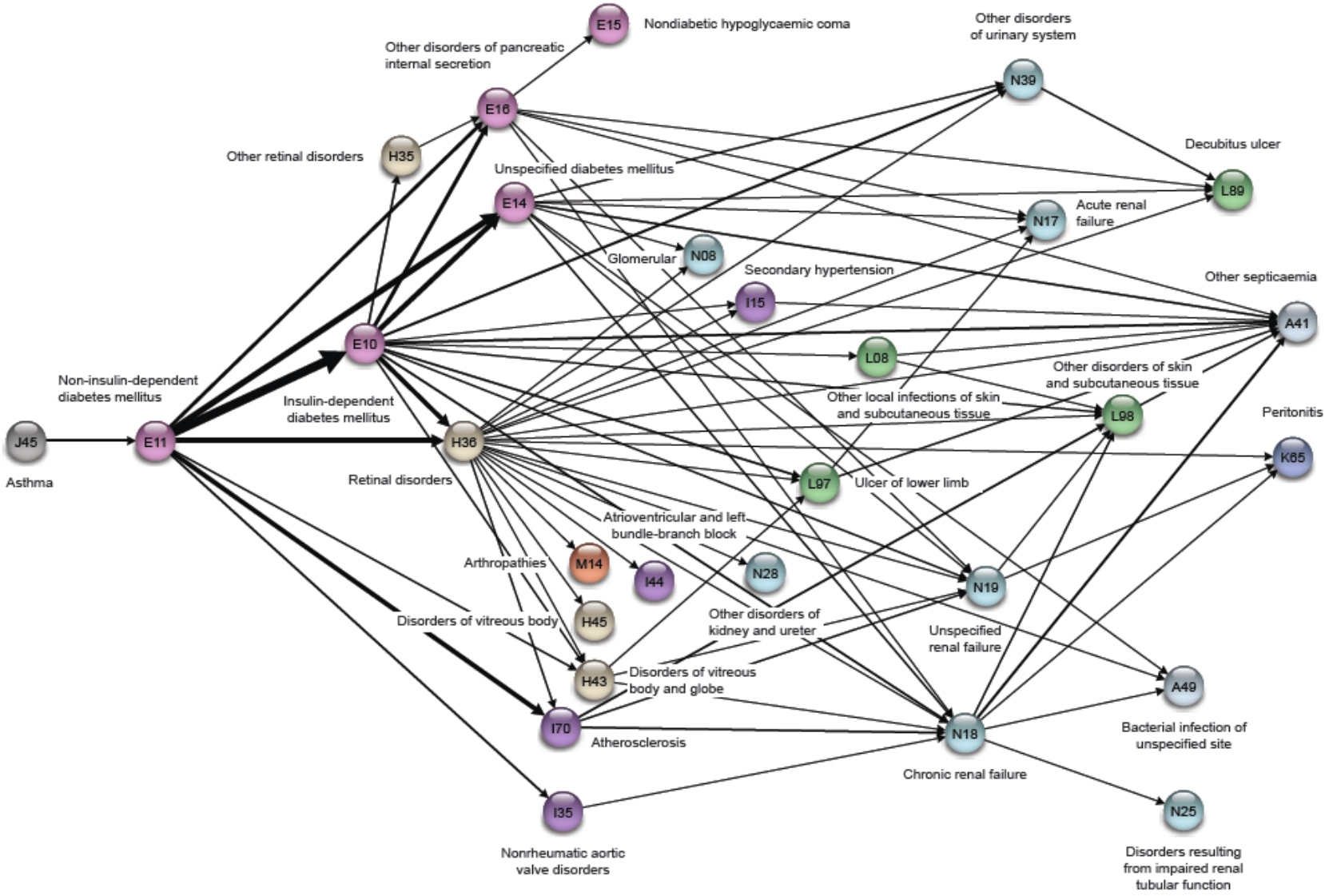
AB Jensen et al., Nature Comm., 2014

Disease trajectories and trajectory-networks



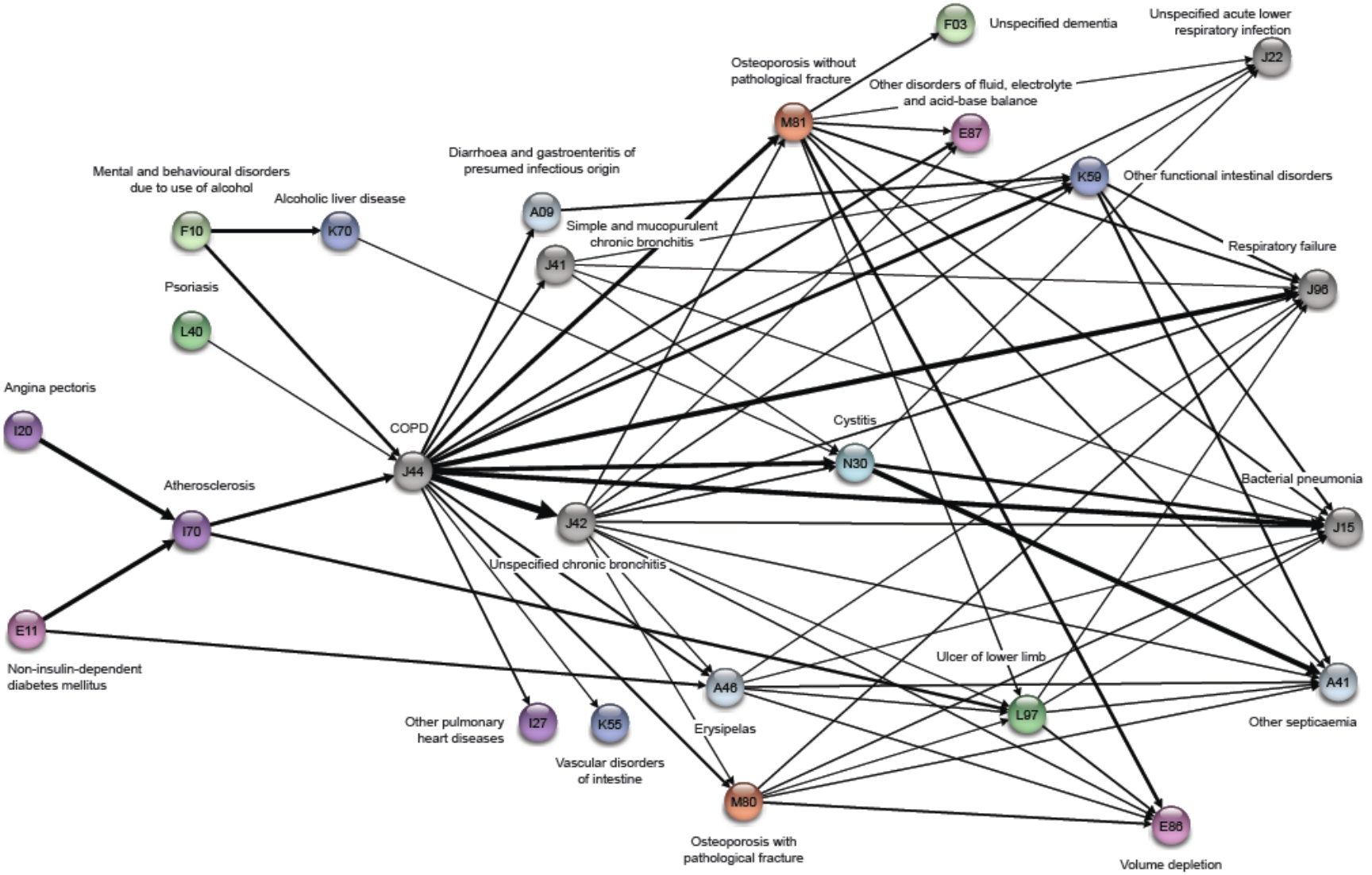
AB Jensen et al.,
Nature Comm., 2014

Diabetes trajectory network

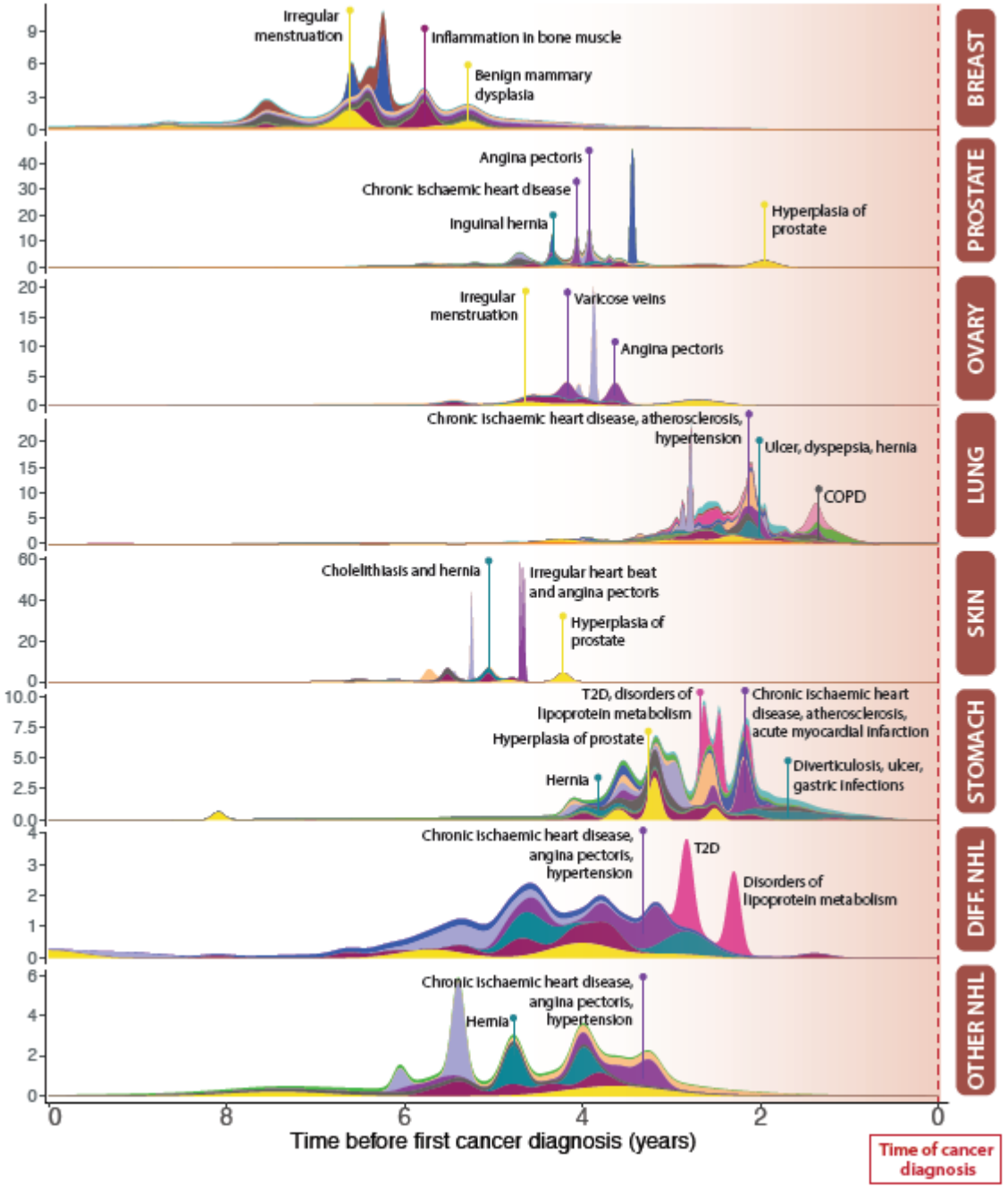


COPD trajectory cluster

with five preceding diagnoses leading to COPD and some of the possible outcomes

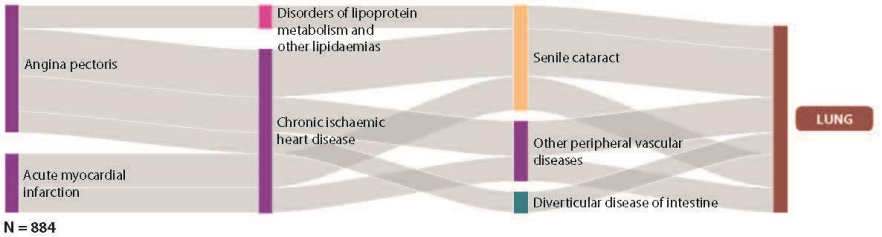


Temporal pre-cancer diagnoses patterns

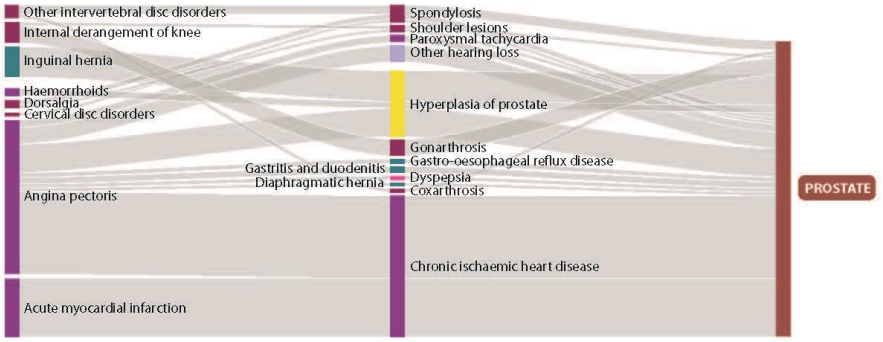


Hu et al., Cancer Res. 2019

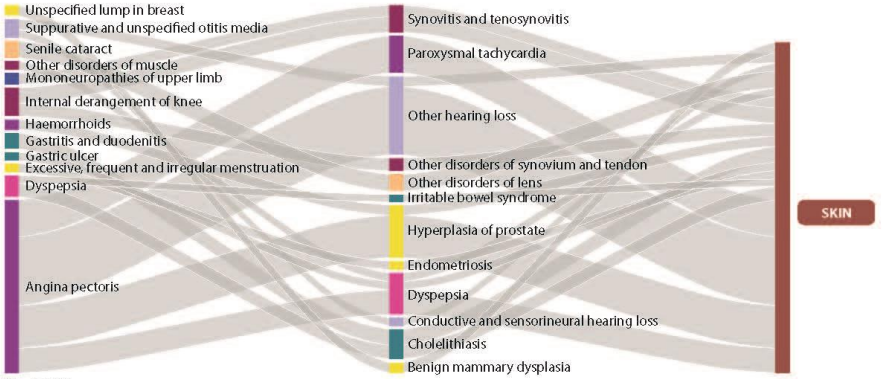
Temporal pre-cancer patient trajectories



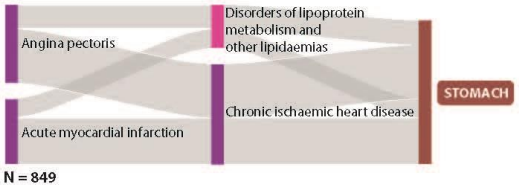
N = 884



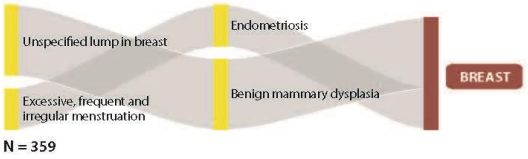
N = 8,373



N = 4,677



N = 849

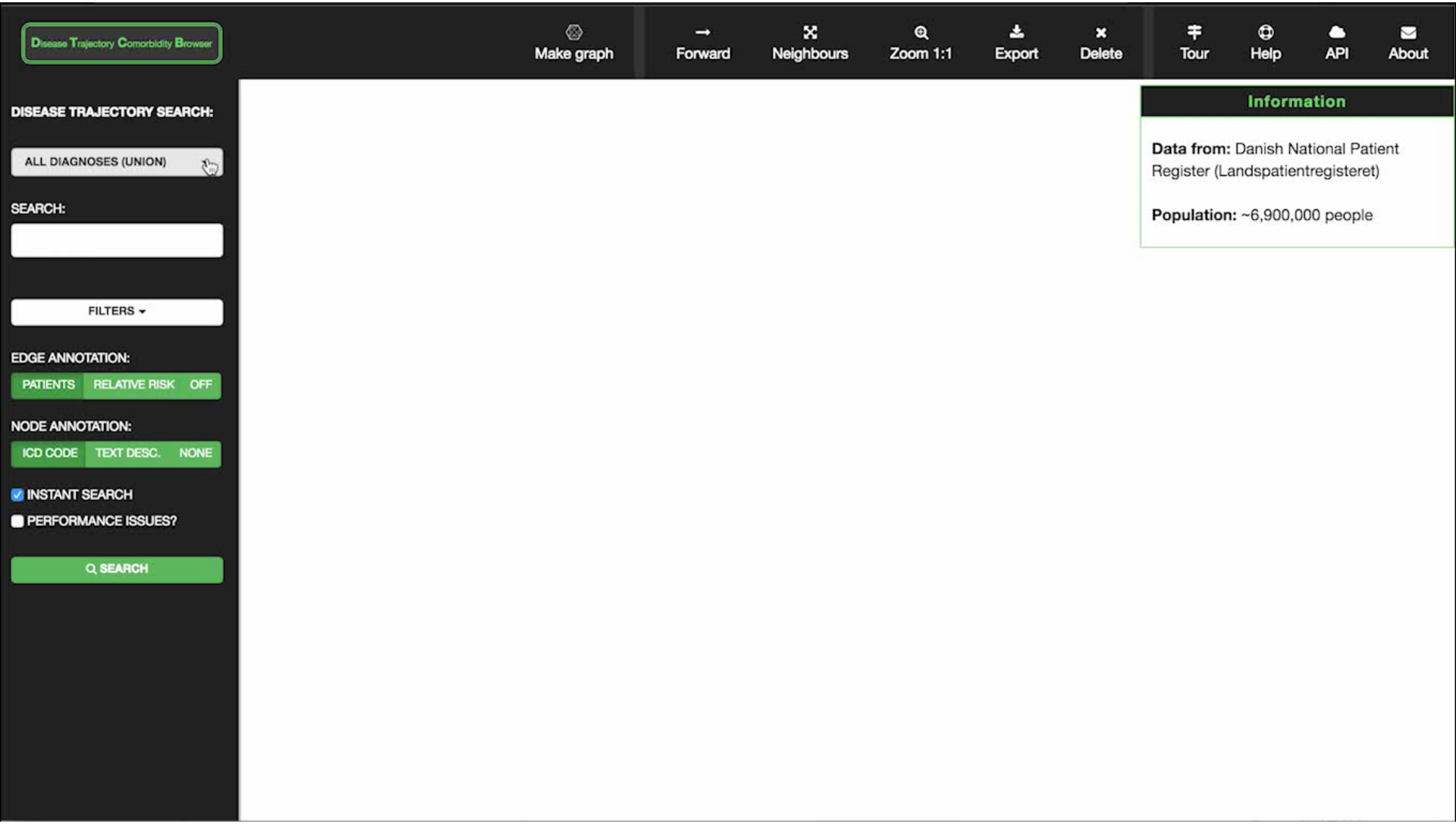


N = 359

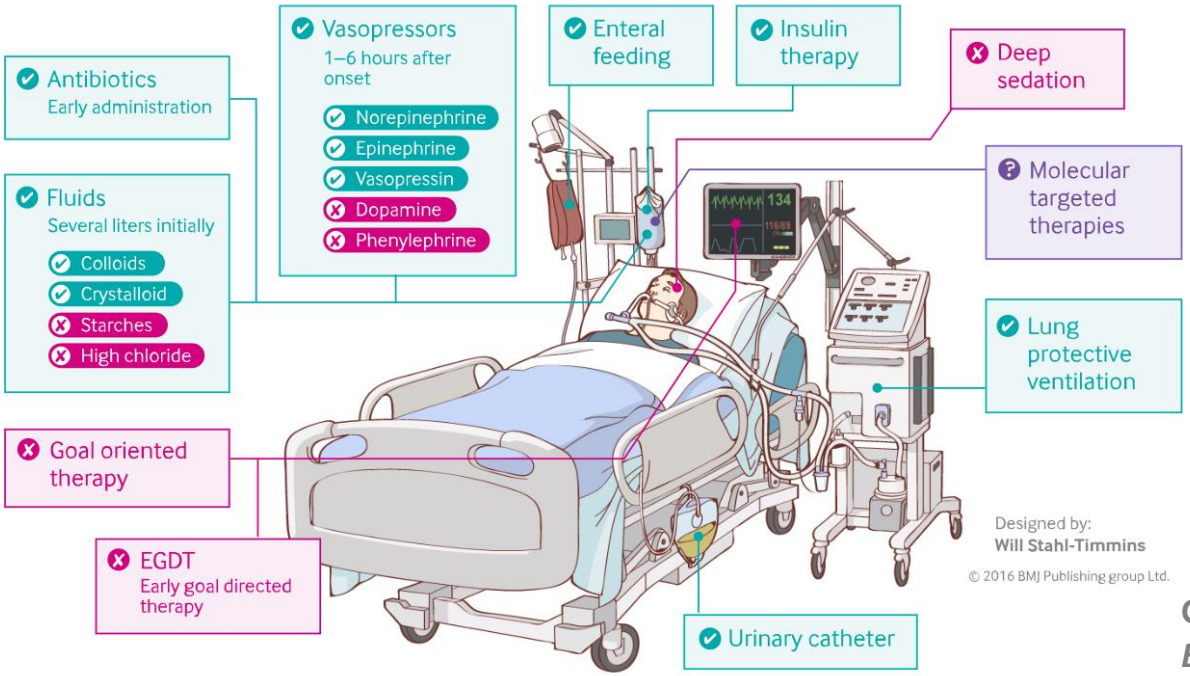
N = 126

N = 115

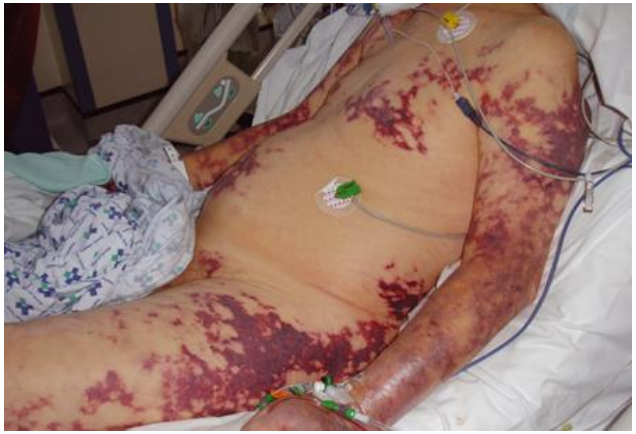
Hu et al., Cancer Res. 2019


Disease Trajectory Comorbidity Browser
Make graph
Forward
Neighbours
Zoom 1:1
Export
Delete
Tour
Help
API
About
DISEASE TRAJECTORY SEARCH:
ALL DIAGNOSES (UNION)
SEARCH:
FILTERS
EDGE ANNOTATION:
PATIENTS
RELATIVE RISK
OFF
NODE ANNOTATION:
ICD CODE
TEXT DESC.
NONE
INSTANT SEARCH
PERFORMANCE ISSUES?
SEARCH
Information
Data from: Danish National Patient Register (Landspatientregisteret)
Population: ~6,900,000 people

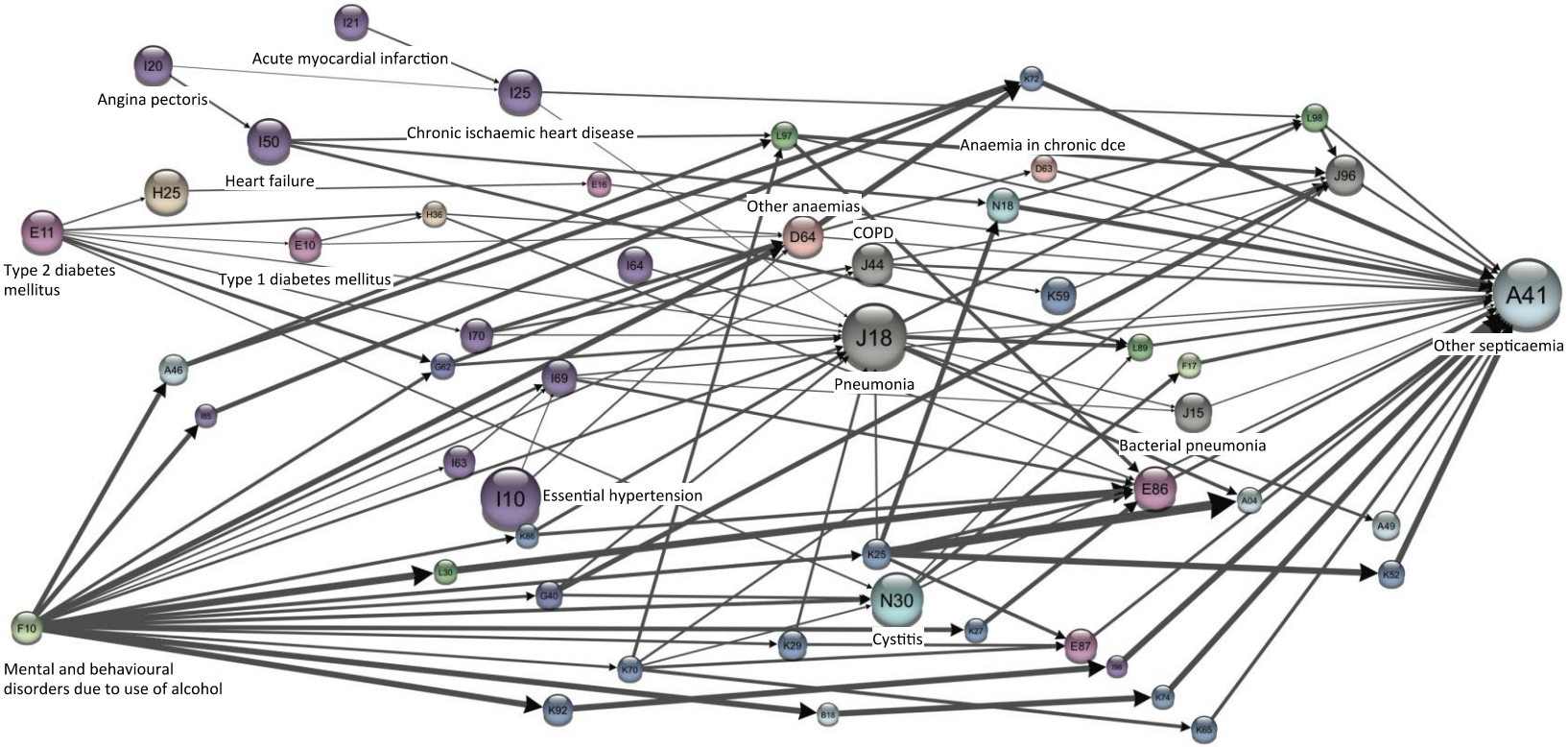
Intensive care patients



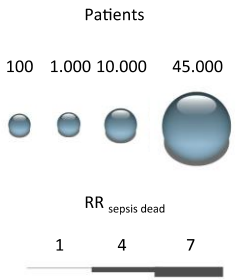
Gotts & Matthay
BMJ 2016; 353:i1585



Sepsis survival across pre-history
(120,000 patients, 56 significant trajectories)



A04	Other bacterial intestinal infections
A46	Erysipelas
A49	Bacterial infection of unspecified site
B18	Chronic viral hepatitis
E16	Other disorders of pancreatic internal secretion
E87	Other disorders of fluid, electrolyte and acid-base balance
G40	Epilepsy
G62	Other polyneuropathies
H25	Senile cataract
H36	Retinal disorders in diseases classified elsewhere
I63	Cerebral infarction
I69	Sequelae of cerebrovascular disease
I70	Atherosclerosis
I85	Oesophageal varices
I98	Other disorders of circulatory system in diseases classified elsewhere
J96	Respiratory failure, not elsewhere classified
K25	Gastric ulcer
K27	Peptic ulcer, site unspecified
K29	Gastritis and duodenitis
K52	Other noninfective gastroenteritis and colitis
K59	Other functional intestinal disorders
K65	Peritonitis
K70	Alcoholic liver disease
K72	Hepatic failure, not elsewhere classified
K74	Fibrosis and cirrhosis of liver
K86	Other diseases of pancreas
K92	Other diseases of digestive system
L30	Other dermatitis
L89	Decubitus ulcer
L97	Ulcer of lower limb, not elsewhere classified
N18	Chronic renal failure



Beck et al, 2016

Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records

Annelaura B Nielsen, Hans-Christian Thorsen-Meyer, Kirstine Belling, Anna P Nielsen, Cecilia E Thomas, Piotr J Chmura, Mette Lademann, Pope L Moseley, Marc Heimann, Lars Dybdahl, Lasse Spangsege, Patrick Hulsen, Anders Perner, Søren Brunak

Summary

Background Intensive-care units (ICUs) treat the most critically ill patients, which is complicated by the heterogeneity of the diseases that they encounter. Severity scores based mainly on acute physiology measures collected at ICU admission are used to predict mortality, but are non-specific, and predictions for individual patients can be inaccurate. We investigated whether inclusion of long-term disease history before ICU admission improves mortality predictions.

Methods Registry data for long-term disease histories for more than 230 000 Danish ICU patients were used in a neural network to develop an ICU mortality prediction model. Long-term disease histories and acute physiology measures were aggregated to predict mortality risk for patients for whom both registry and ICU electronic patient record data were available. We compared mortality predictions with admission scores on the Simplified Acute Physiology Score (SAPS) II, the Acute Physiologic Assessment and Chronic Health Evaluation (APACHE) II, and the best available multimorbidity score, the Multimorbidity Index. An external validation set from an additional hospital was acquired after model construction to confirm the validity of our model. During initial model development data were split into a training set (85%) and an independent test set (15%), and a five-fold cross-validation was done during training to avoid overfitting. Neural networks were trained for datasets with disease history of 1 month, 3 months, 6 months, 1 year, 2·5 years, 5 years, 7·5 years, 10 years, and 23 years before ICU admission.

Findings Mortality predictions with a model based solely on disease history outperformed the Multimorbidity Index (Matthews correlation coefficient 0·265 *vs* 0·065), and performed similarly to SAPS II and APACHE II (Matthews correlation coefficient with disease history, age, and sex 0·326 *vs* 0·347 and 0·300 for SAPS II and APACHE II, respectively). Diagnoses up to 10 years before ICU admission affected current mortality prediction. Aggregation of previous disease history and acute physiology measures in a neural network yielded the most precise predictions of in-hospital mortality (Matthews correlation coefficient 0·391 for in-hospital mortality compared with 0·347 with SAPS II and 0·300 with APACHE II). These results for the aggregated model were validated in an external independent dataset of 1528 patients (Matthews correlation coefficient for prediction of in-hospital mortality 0·341).

Interpretation Longitudinal disease-spectrum-wide data available before ICU admission are useful for mortality prediction. Disease history can be used to differentiate mortality risk between patients with similar vital signs with more precision than SAPS II and APACHE II scores. Machine learning models can be deconvoluted to generate novel understandings of how ICU patient features from long-term and short-term events interact with each other. Explainable machine learning models are key in clinical settings, and our results emphasise how to progress towards the transformation of advanced models into actionable, transparent, and trustworthy clinical tools.

Funding Novo Nordisk Foundation and Innovation Fund Denmark.

Lancet Digital Health 2019; 1: e78–89

See Comment page e68

Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark (A B Nielsen PhD, H-C Thorsen-Meyer MD, K Belling PhD, A P Nielsen MD, C E Thomas PhD, P J Chmura MSc, M Lademann PhD, Prof P L Moseley MD, Prof S Brunak PhD); Department of Intensive Care, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark (H-C Thorsen-Meyer, Prof A Perner PhD); Centre for IT, Medical Technology and Telephony Services, Capital Region of Denmark, Copenhagen, Denmark (M Heimann MSc); and Dalntel, Lyngby, Denmark (L Dybdahl MSc, L Spangsege MSc, P Hulsen BSc)

Correspondence to: Prof Søren Brunak, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark soren.brunak@cpr.ku.dk

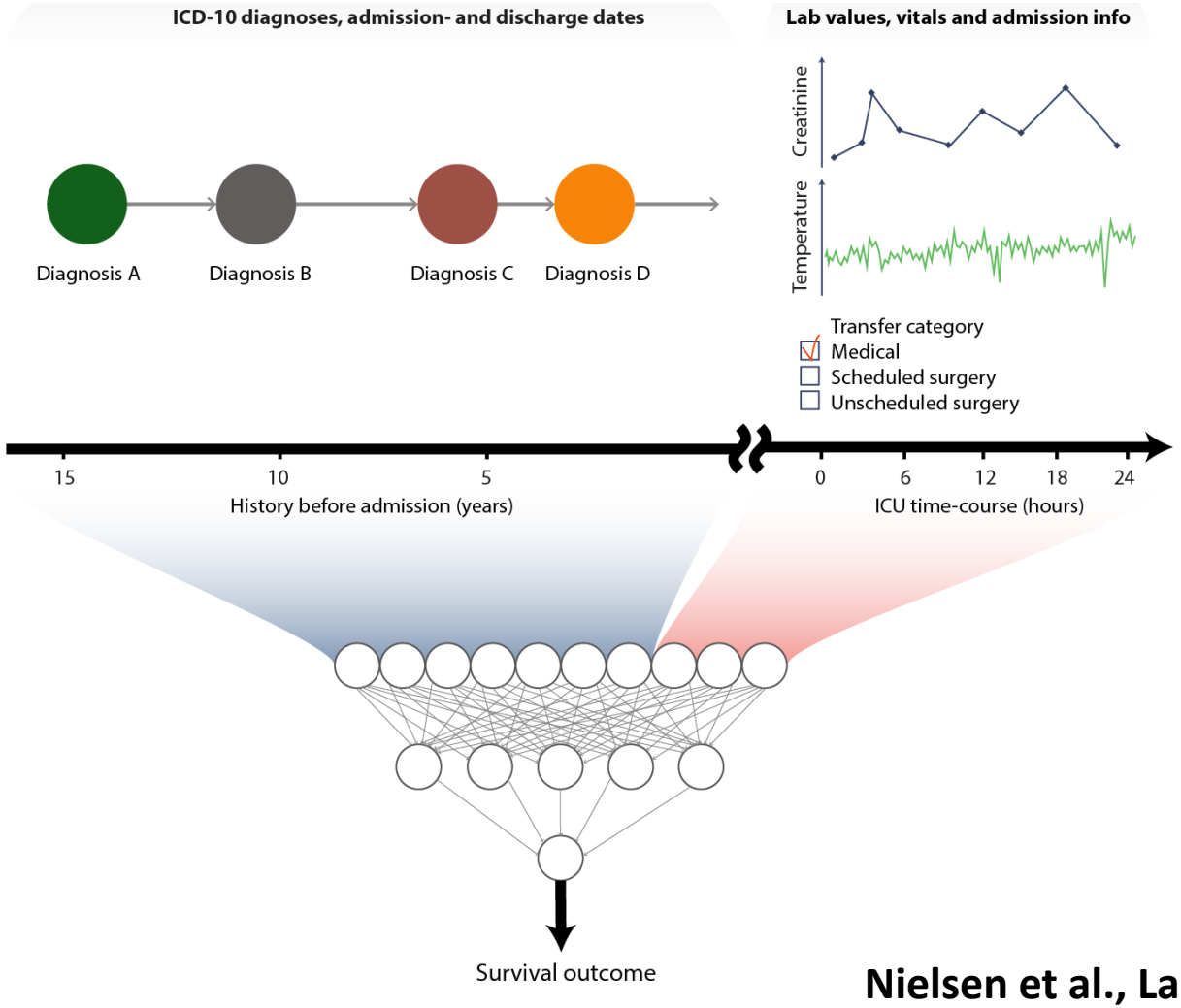
Introduction

Intensive-care units (ICUs) handle patients from all medical and surgical specialties. Therefore, their populations are highly heterogeneous, and consist of mainly elderly patients who often have a long history of disease. Prediction of prognosis to inform decision making in the ICU is difficult because of the severity of patients' current illness and their disease history.[1]

Mortality risk estimates based on acute physiology scores—such as the Simplified Acute Physiology Score (SAPS) and the Acute Physiologic Assessment and Chronic Health Evaluation (APACHE)—are sometimes used in clinical practice to assess disease severity.[2,3] They are based on logistic regression of specific markers of patient physiology that are recorded during the first hours after ICU admission. In the past 10 years, advanced

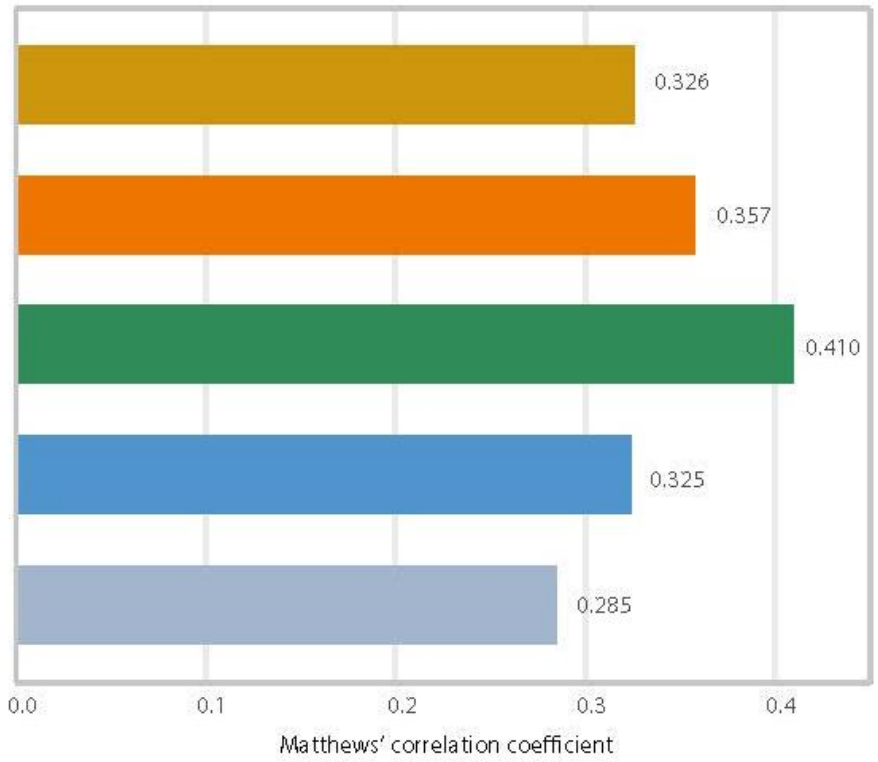
Nielsen et al., Lancet Díg Health, 2019

Mortality prediction from machine learning based aggregation of time scales



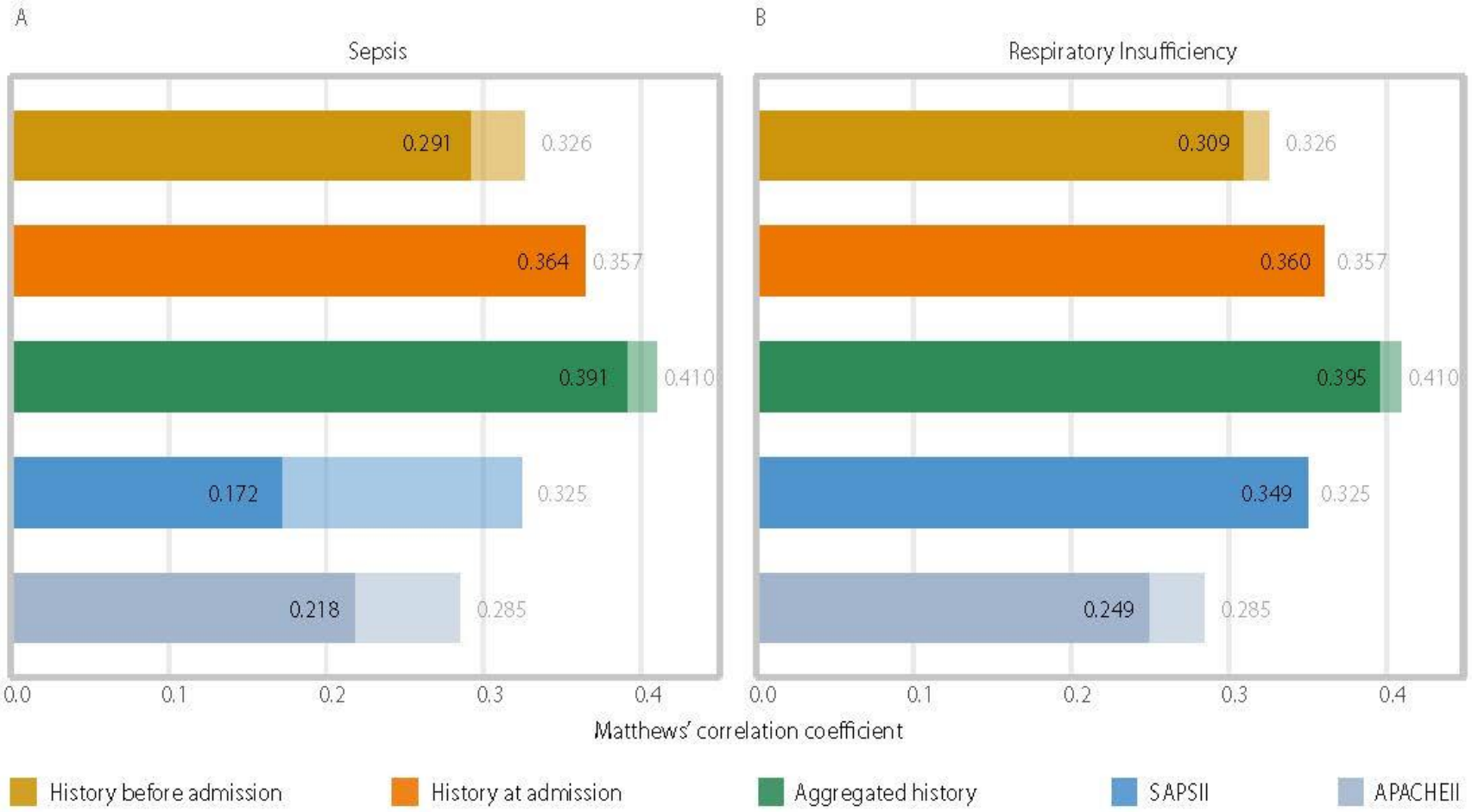
Nielsen et al., Lancet Díg Health, 2019

ICU mortality prediction performance



History before admission	History at admission	Aggregated history	SAPSII	APACHEII
Ten-year disease history Sex Age	Ten-year disease history Sex Age Length of stay Transfer category Hospital code	Ten-year disease history Sex Age Length of stay Transfer category Hospital code Acute physiology measures	Acute physiology measures	

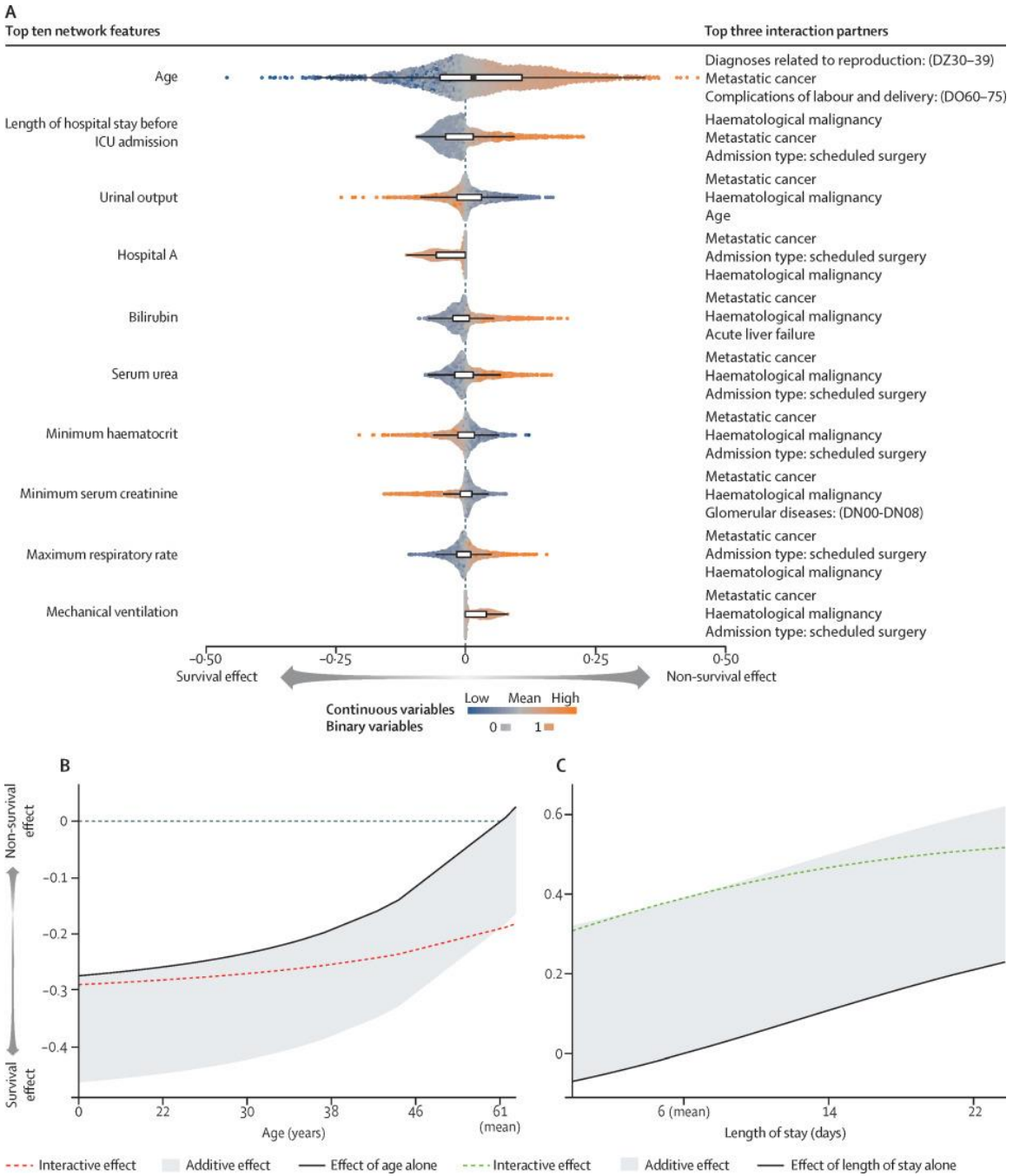
ICU prediction performance – patient subgroups



ICU mortality feature importance

(A) Each dot one patient

Interaction between age and history of diagnoses related to reproduction (B), and interaction between length of stay before ICU admission and history of haematological malignancy (C)

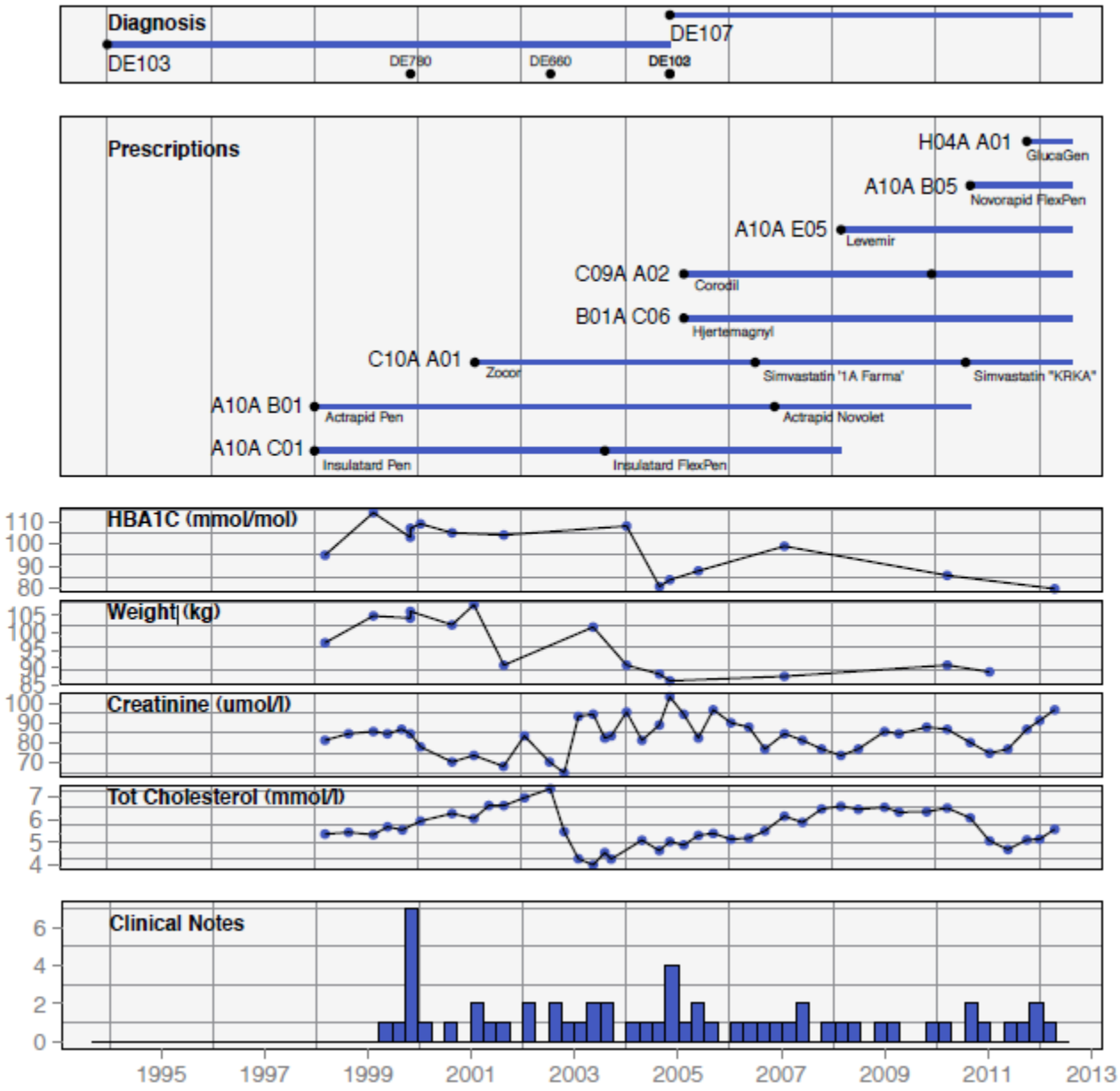


Diagnoses

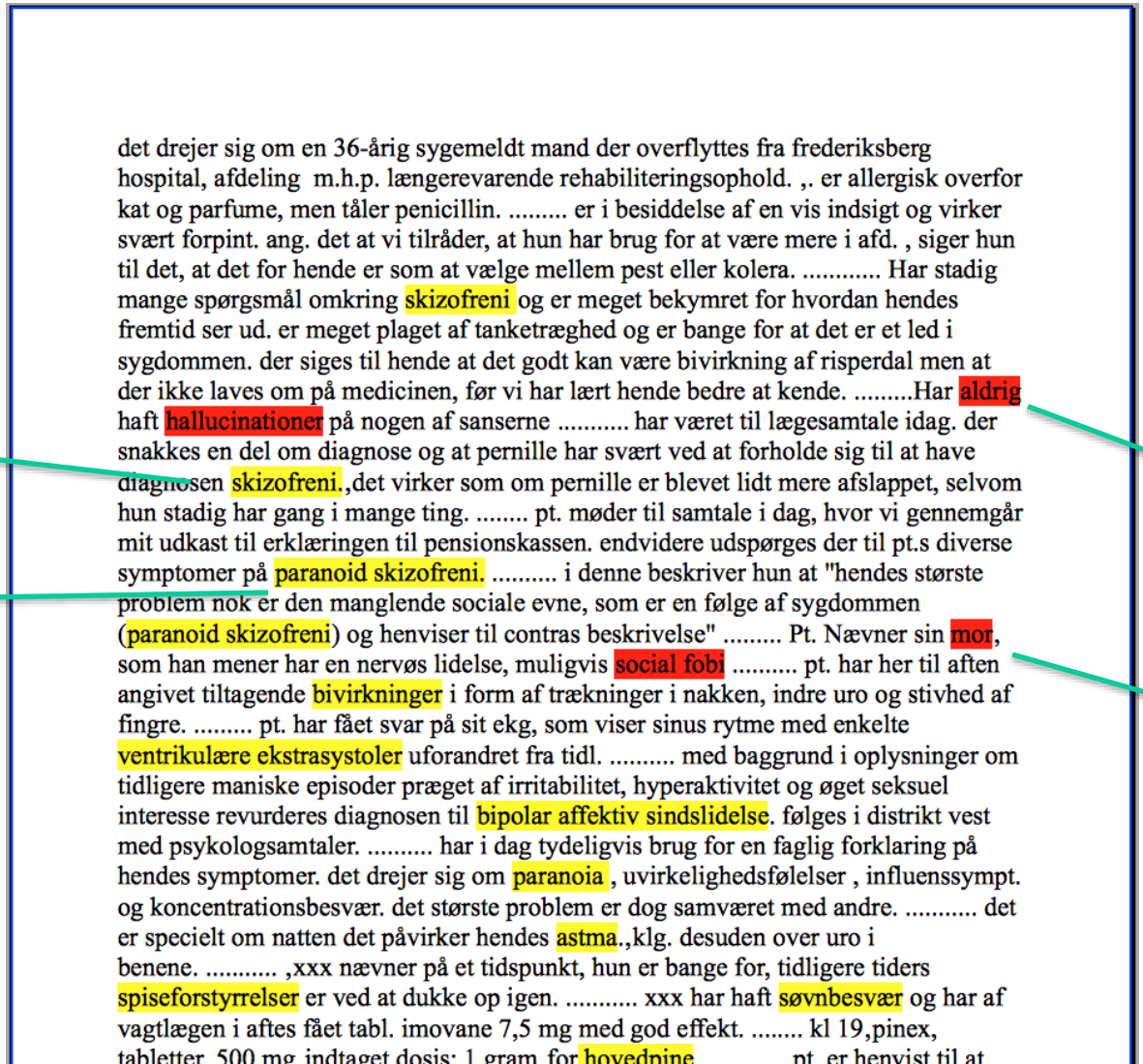
Drugs

Lab tests

Text



Deep-phenotyping by text mining of ICD-10 terms in patient records



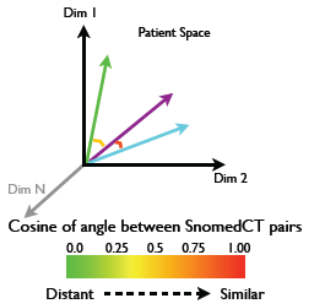
det drejer sig om en 36-årig sygemeldt mand der overflyttes fra frederiksberg hospital, afdeling m.h.p. længerevarende rehabiliteringsophold. ,. er allergisk overfor kat og parfume, men tåler penicillin. ..... er i besiddelse af en vis indsigt og virker svært forpint. ang. det at vi tilråder, at hun har brug for at være mere i afd. , siger hun til det, at det for hende er som at vælge mellem pest eller kolera. ..... Har stadig mange spørgsmål omkring skizofreni og er meget bekymret for hvordan hendes fremtid ser ud. er meget plaget af tanketrægghed og er bange for at det er et led i sygdommen. der siges til hende at det godt kan være bivirkning af risperdal men at der ikke laves om på medicinen, før vi har lært hende bedre at kende. .....Har aldrig haft hallucinationer på nogen af sanserne ..... har været til lægesamtale idag. der snakkes en del om diagnose og at pernille har svært ved at forholde sig til at have diagnosen skizofreni.,det virker som om pernille er blevet lidt mere afslappet, selvom hun stadig har gang i mange ting. ..... pt. møder til samtale i dag, hvor vi gennemgår mit udkast til erklæringen til pensionskassen. endvidere udspørges der til pt.s diverse symptomer på paranoid skizofreni. ..... i denne beskriver hun at "hendes største problem nok er den manglende sociale evne, som er en følge af sygdommen (paranoid skizofreni) og henviser til contras beskrivelse" ..... Pt. Nævner sin mor, som han mener har en nervøs lidelse, muligvis social fobi ..... pt. har her til aften angivet tiltagende bivirkninger i form af trækninger i nakken, indre uro og stivhed af fingre. ..... pt. har fået svar på sit ekg, som viser sinus rytme med enkelte ventrikulære ekstrasystoler uforandret fra tidl. ..... med baggrund i oplysninger om tidligere maniske episoder præget af irritabilitet, hyperaktivitet og øget seksuel interesse revurderes diagnosen til bipolar affektiv sindslidelse. følges i distrikt vest med psykologsamtaler. ..... har i dag tydeligvis brug for en faglig forklaring på hendes symptomer. det drejer sig om paranoia , uvirkelighedsfølelser , influenssympt. og koncentrationsbesvær. det største problem er dog samværet med andre. ..... det er specielt om natten det påvirker hendes astma.,klg. desuden over uro i benene. ..... ,xxx nævner på et tidspunkt, hun er bange for, tidligere tiders spiseforstyrrelser er ved at dukke op igen. ..... xxx har haft søvnbesvær og har af vagtlægen i aftes fået tabl. imovane 7,5 mg med god effekt. ..... kl 19,pinex, tabletter, 500 mg, indtaget dosis: 1 gram, for hovedpine. pt. er henvist til at

F20

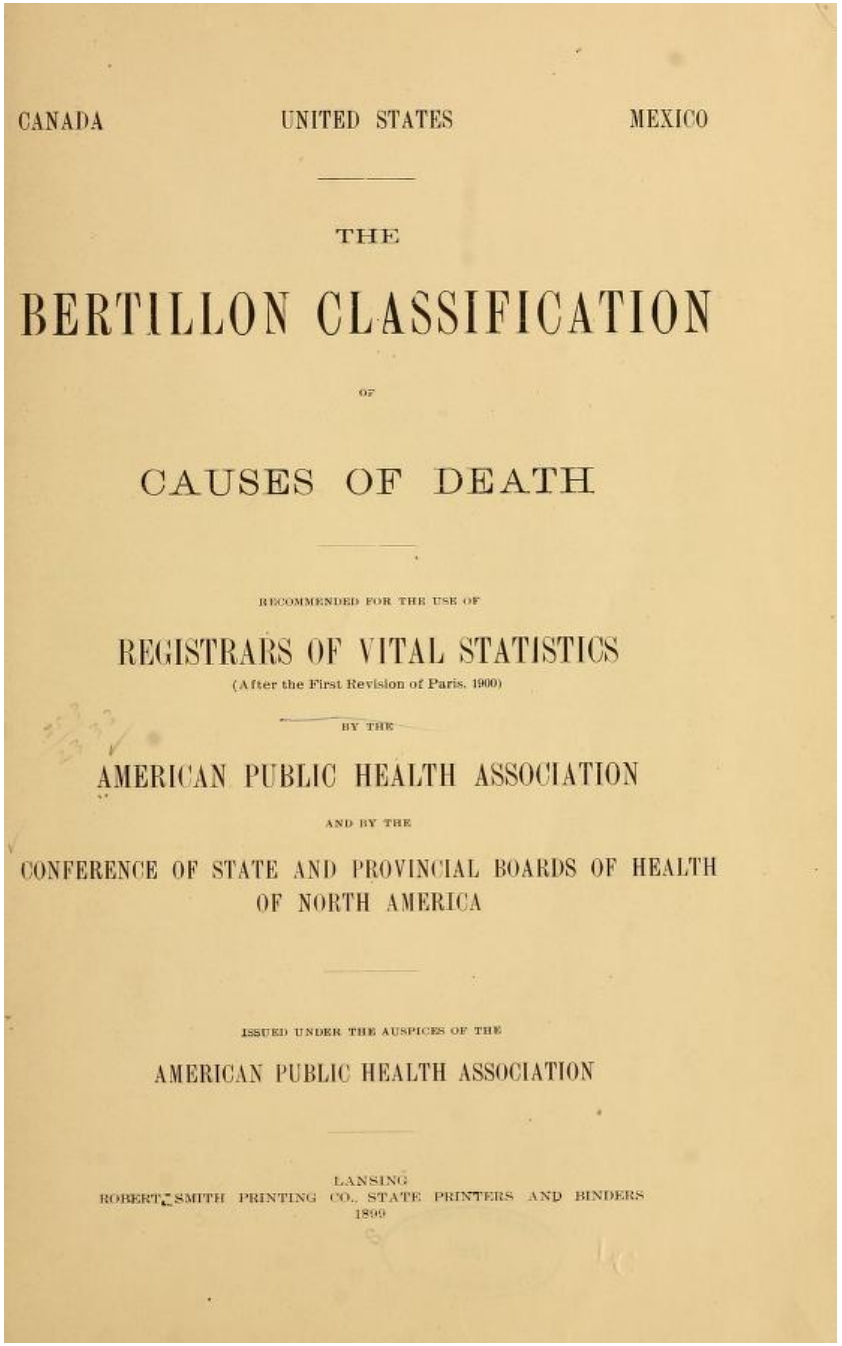
F200

Negation

Family



Roche et al.
PLoS Comp. Biol.
2011



CANADA UNITED STATES MEXICO

THE

BERTILLON CLASSIFICATION

OF

CAUSES OF DEATH

RECOMMENDED FOR THE USE OF

REGISTRARS OF VITAL STATISTICS

(After the First Revision of Paris, 1900)

BY THE

AMERICAN PUBLIC HEALTH ASSOCIATION

AND BY THE

CONFERENCE OF STATE AND PROVINCIAL BOARDS OF HEALTH OF NORTH AMERICA

ISSUED UNDER THE AUSPICES OF THE

AMERICAN PUBLIC HEALTH ASSOCIATION

LANSING
ROBERT SMITH PRINTING CO., STATE PRINTERS AND BINDERS
1899

The frequently used ICD diagnosis and disease classification systems dates back to 1893 where causes of death were "harmonized" based on English, German and Swiss classifications

"The Bertillon classification is not presented as by any means a perfect system of classification of causes of death.
No perfect system has ever been devised, and should there be, the progress of medical science would in time render it obsolete."

International Classification of Disease 1893-

Table 1 The development of ICD

Version	Year for revision	Number of codes
ICD-0	1893	161
ICD-1	1900	179
ICD-2	1909	189
ICD-3	1920	200
ICD-4	1929	200
ICD-5	1938	200
ICD-6	1948	952
ICD-7	1955	952
ICD-8	1965	1 040
ICD-9	1975	6 701
ICD-10	1989	12 420

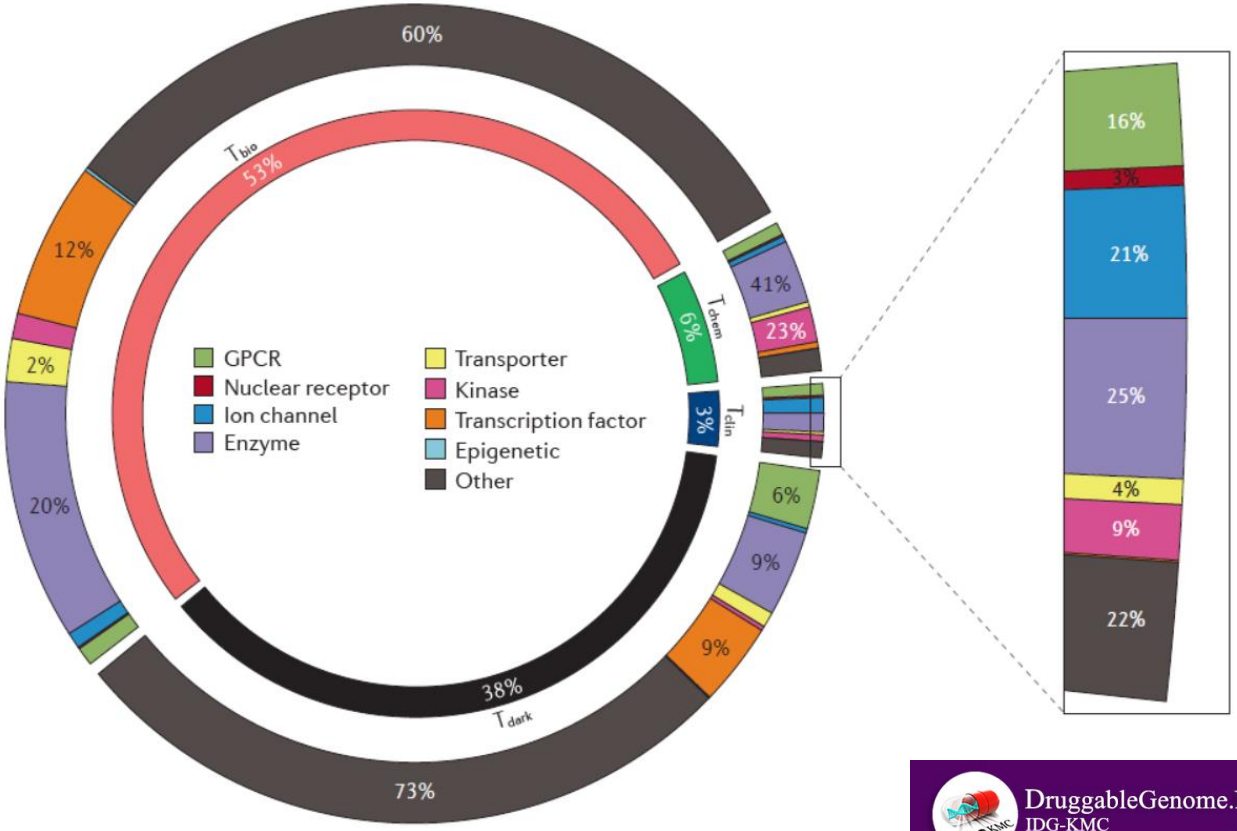
Adapted ICD versions may differ, some national editions have expanded the code set even further; with some going so far as to add procedure codes. ICD-10-CM, for example, has over 70,000 codes.

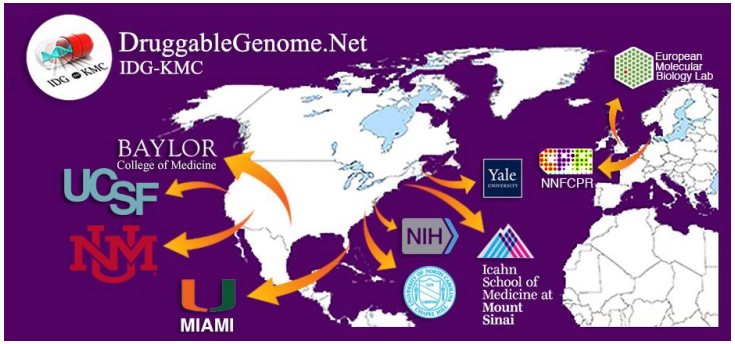
Δ population-wide health data

- 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 patient records in new ways:
  - Diet
  - Income, …
  - Education, grades in exams, …
  - Wearable data
  - Patient generated data

Illuminating the Druggable Genome
>33% of proteins “not” studied





Oprea et al.,
Nature Reviews Drug Discovery, 2018

Acknowledgements

EHR and registry data analysis

Anders Bøck Jensen, CPR/KU
Isa Kirk, CPR/KU
Christian Simon, CPR/KU
Karina Basanik, CPR/KU
Mette Beck, CPR/KU
Peter Bjødstrup Jensen, CPR/KU
Robert Eriksson, CPR/KU
Isabella Friis Jørgensen, CPR
David Westergaard, CPR
Annelaura Nielsen, CPR
Jessica Hu, CPR
Troels Siggard, CPR
Olivier Taboureau CPR/Paris Dederot
Alejandro Aguayo Orozco
Pope Mosely, CPR/U. New Mexico
Tudor Oprea, U. New Mexico
Lars Juhl Jensen, CPR/KU
Francisco Simões Roque, CBS/DTU

Danske Regioner
Karen Søeby, Hvidovre Hospital
Søren Bredkjær, **Region Zealand**
Mahad Huniche, Region Zealand
Stig Ejdrup, Region Zealand, Roskilde Hospital
Jens Lundgren, Rigshospitalet
Marie Helleberg, Rigshospitalet

Mental Centre Sct. Hans
Thomas Werge
Henriette Schmock
Thomas Hansen, Sct. Hans Hospital

Rigshospitalet
Anders Perner, Intensive care
Jan Bonde, Intensive care
Henrik Ullum, DBDS/CHB

Steno Diabetes Center, Hagedorn, Novo Nordisk A/S
Peter Rossing
Henrik Ullits Andersen
Regine Bergholdt
Thomas Almdal, now Gentofte Hospital
Flemming Pociot, now Herlev Hospital
Torben Hansen, now U. Copenhagen
Oluf Borbye Pedersen, now U. Copenhagen