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,
Longitudinal polypharmacy exposures are complex
Clinical and socio-economic data are needed to interpret the molecular domain.
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 laboratory 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

<table>
<thead>
<tr>
<th></th>
<th>In patients</th>
<th>Out patients</th>
<th>IR patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Females</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
<td><img src="image3" alt="Graph" /></td>
</tr>
<tr>
<td>Males</td>
<td><img src="image4" alt="Graph" /></td>
<td><img src="image5" alt="Graph" /></td>
<td><img src="image6" alt="Graph" /></td>
</tr>
</tbody>
</table>

(ICD-10 era, 1994-2019)

AB Jensen et al., Nature Comm., 2014

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 behavioral 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 laboratory findings, not elsewhere classified
19. Injury, poisoning and certain other consequences of external causes
20. External causes of morbidity and mortality
Linking multimorbidity journeys to genetics

Hu, Thomas & Brunak
Nature Rev. Genetics 2016,
ATC drug groups in 1.1 billion male and female prescriptions according to age

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
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 individuals)

- 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 protocols 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/swab microbiome

Cardiovascular diseases
Transfusion outcome
Reproductive health
COVID-19
Cancer
Chronic inflammatory diseases
Cancer biomarkers
Infections
Approved
2018-2020
2021
2022
Submitted
Planned
Inflammatory arthritis
Pain and musculoskeletal diseases
Osteoporosis and fractures
COVID-19 Variants & Vaccination
Degenerative & episodic brain disorders
Oral Health & metabolic dis.
Aging in Health & disease

#185,000
#180,000
#50,000
#198,000
#130,000
#299,000
#180,000
#50,000
#30,000
#100,000
#100,000
#17,000
#275,000
#126,000
#198,000
#335,000
#300,000
#17,000
#20,000

Sisse Rye Ostrowski
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
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)

Different Denmark & US EHR structure

Many more codes in the US data per patient than in the DK data
### Feature importance ranking using explainability methods

#### Comparing features in the DK and US

---

**C**

<table>
<thead>
<tr>
<th>Cancer in 0-6 months</th>
<th>Cancer in 6-12 months</th>
<th>Cancer in 12-24 months</th>
<th>Cancer in 24-36 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unspecified jaundice</td>
<td>Other disease of biliary tract</td>
<td>Medical observation and evaluation for suspected diseases and conditions</td>
<td>Medical observation and evaluation for suspected diseases and conditions</td>
</tr>
<tr>
<td>Medical observation and evaluation for suspected diseases and conditions</td>
<td>Unspecified jaundice</td>
<td>Other diseases of biliary tract</td>
<td>Other diseases of pancreas</td>
</tr>
<tr>
<td>Other diseases of biliary tract</td>
<td>Medical observation and evaluation for suspected diseases and conditions</td>
<td>Other diseases of pancreas</td>
<td>Other diseases of pancreas</td>
</tr>
<tr>
<td>Abdominal and pelvic pain</td>
<td>Other diseases of pancreas</td>
<td>Abdominal and pelvic pain</td>
<td>Non-insulin-dependent diabetes mellitus</td>
</tr>
<tr>
<td>Malignant neoplasm of other and unspecified parts of biliary tract</td>
<td>Malignant neoplasm of other and unspecified parts of biliary tract</td>
<td>Non-insulin-dependent diabetes mellitus</td>
<td>Unspecified jaundice</td>
</tr>
<tr>
<td>Other diseases of pancreas</td>
<td>Abdominal and pelvic pain</td>
<td>Malignant neoplasm of other and unspecified parts of biliary tract</td>
<td>Abdominal and pelvic pain</td>
</tr>
<tr>
<td>Secondary malignant neoplasm of respiratory and digestive organs</td>
<td>Secondary malignant neoplasm of respiratory and digestive organs</td>
<td>Unspecified jaundice</td>
<td>Malignant neoplasm of other and unspecified parts of biliary tract</td>
</tr>
<tr>
<td>Symptoms and signs concerning food and fluid intake</td>
<td>Non-insulin-dependent diabetes mellitus</td>
<td>Other functional intestinal disorder</td>
<td>Gastritis and gastroduodenitis</td>
</tr>
<tr>
<td>Non-insulin-dependent diabetes mellitus</td>
<td>Malignant neoplasm without specific location</td>
<td>Diseases of pancreas</td>
<td>Insulin-dependent diabetes mellitus</td>
</tr>
<tr>
<td>other anaemias</td>
<td>other anaemias</td>
<td>Secondary malignant neoplasm of respiratory and digestive organs</td>
<td>other anaemias</td>
</tr>
</tbody>
</table>

---

**D**

<table>
<thead>
<tr>
<th>Cancer in 0-6 months</th>
<th>Cancer in 6-12 months</th>
<th>Cancer in 12-24 months</th>
<th>Cancer in 24-36 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute pancreatitis</td>
<td>Acute pancreatitis</td>
<td>Abdominal and pelvic pain</td>
<td>Diabetes mellitus</td>
</tr>
<tr>
<td>Abdominal and pelvic pain</td>
<td>Diabetes mellitus</td>
<td>Other diseases of biliary tract</td>
<td>Other diseases of liver</td>
</tr>
<tr>
<td>Other diseases of biliary tract</td>
<td>Other diseases of biliary tract</td>
<td>Diabetes mellitus</td>
<td>Persons encountered health services in other circumstances</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>Symptoms and signs concerning food and fluid intake</td>
<td>Persons encountered health services in other circumstances</td>
<td>Abdominal and pelvic pain</td>
</tr>
<tr>
<td>Other diseases of pancreas</td>
<td>Symptoms and signs concerning food and fluid intake</td>
<td>Persons encountered health services in other circumstances</td>
<td>Abdominal and pelvic pain</td>
</tr>
<tr>
<td>Symptoms and signs concerning food and fluid intake</td>
<td>Malignant neoplasm of trachea, bronchi, or lung</td>
<td>Dependence of opioids, sedatives, or other psychoactive substances</td>
<td>Nausea and vomiting</td>
</tr>
<tr>
<td>Disorders of social functioning with other mental disorder</td>
<td>Abdominal and pelvic pain</td>
<td>Abuse of alcohol, heroin, sedatives, or other psychoactive substances</td>
<td>Abuse of alcohol, heroin, sedatives, or other psychoactive substances</td>
</tr>
<tr>
<td>Essential (primary) hypertension</td>
<td>Other diseases of pancreas</td>
<td>Cough, haemorrhage from respiratory passages</td>
<td>Unspecified jaundice, or skin erosion</td>
</tr>
<tr>
<td>Persons encountered health services in other circumstances</td>
<td>Secondary malignant neoplasm of respiratory and digestive organs</td>
<td>Cataract</td>
<td>Cataract</td>
</tr>
<tr>
<td>Examinations and observations for other reasons</td>
<td>Other dermatitis</td>
<td>Secondary malignant neoplasm of respiratory and digestive organs</td>
<td>Cataract</td>
</tr>
</tbody>
</table>

---

**ICD-10 chapters**

1. Neoplasms
2. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
3. Endocrine, nutritional and metabolic diseases
4. Mental and behavioural disorders
5. Diseases of the eye and adnexa
6. Diseases of the circulatory system
7. Diseases of the digestive system
8. Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
9. Factors influencing health status and contact with health services
10. Other personal health services
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

Can integrate continuous and discrete data
Can deal with missing data
Continuous: Z-scale & MSE
Discrete: One-hot & CE

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

- 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

**Clinical features**

**Drugs**
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

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

GWAS analysis of patients stratified according to whether 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)

Lab value Seasonality Fitting Approach - 310M lab values

Muse et al., Cell Patterns 4, 100778, 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

Improved estimation of mortality and diagnoses

Muse et al., Cell Patterns 4, 100778, 2023
The Danish Disease Trajectory Browser: 
[http://dtb.cpr.ku.dk](http://dtb.cpr.ku.dk)

Siggaard et al., Nature Comm, 2020
For a European Health Data Space

#HealthUnion

eHealth

#DigitalHealth
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