Disease Trajectory Analysis of Health Records and Registries
Disease

– journey versus destination

Cancer
Diabetes
Obesity
Mental disorders
Beyond single disease analysis  
(a la GWAS)

Disease-disease correlations

Disease-trajectories

What is potentially solely genetic and what is possibly treatment related?
From molecules to phenotypes ...
... from phenotypes to molecules

Molecular components
- Genomes
- Nucleotides
- Transcripts
- Proteins
- Domains
- Pathways
- Structures
- Small molecules

Integration
- Complexes
- Tissues and organs
- Biobanks
- Cells
- Human individuals

Translation
- Human populations
- Therapies
- Disease prevention
- Early Diagnosis

biomedbridges.eu
Precision medicine

Evolution of Medicine from Art to Scientifically Based

<table>
<thead>
<tr>
<th>High Efficacy</th>
<th>Treatment Effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heuristic</td>
<td>Emperical</td>
</tr>
<tr>
<td>Scientifically Based</td>
<td></td>
</tr>
</tbody>
</table>

Era of Precision Medicine
360 million diapers are changed every day

None of this health information has been used
Until *Smart Diapers*
Complexity of temporal health data – we all move towards becoming ICU patients

General practitioner  Outpatients  Inpatients  Intensive care
Personal identification number enables temporal analysis

From Wikipedia, the free encyclopedia

The Danish Personal Identification number is a national identification number, which is part of the personal information stored in the Civil Registration System.

Was established in 1968.

It is a ten-digit number with the format DDMMYY-SSSS, where DDMMYY is the date of birth and SSSS is a sequence number. The first digit of the sequence number encodes the century of birth (so that centenarians are distinguished from infants), and the last digit of the sequence number is odd for males and even for females.

Any person registered as of 2 April 1968 or later in a Danish civil register, receives a personal identification number.

The civil register list only persons who:
• Are born in Denmark of a mother already registered in the civil register, or
• Have their birth or baptism registered in a 'Dansk Elektronisk Kirkebog (DNK)' (Danish electronic church-book), or
• Reside legally in Denmark for 3 months or more (non-Nordic citizens must also have a residence permit)

Danish citizens, including newborn babies, who are entitled to Danish citizenship, but are living abroad, do not receive a personal ID number, unless they move to Denmark.
What is a precise phenotype?

Overweight and healthy: the concept of metabolically healthy obesity

POSTED SEPTEMBER 24, 2013, 4:31 PM
Patrick J. Skerrett, Executive Editor, Harvard Health

Carrying too many pounds is a solid signal of current or future health problems. But not for everyone. Some people who are overweight or obese may escape the usual hazards, at least temporarily. This weight subgroup has even earned its own moniker—metabolically healthy obesity.

Health professionals define overweight as a body mass index (BMI) between 25.0 and 29.9, and obesity as a BMI of 30 or higher. (BMI is a measure of weight that takes height into consideration. You can calculate your BMI here.)

Most people who are overweight or obese show potentially unhealthy changes in metabolism. These include high blood pressure or high cholesterol, which damage
Word frequencies in Danish patient records

Testis cancer

Mental disorders

Male infertility
Word frequencies in Danish patient records from 2 x ~6,000 T1D and T2D patients (Steno Diabetes Center)
Controlled vocabulary: ICD-10

ICD10 – International Classification of Disease

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Blocks</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>A00-B99</td>
<td>Certain infectious and parasitic diseases</td>
</tr>
<tr>
<td>II</td>
<td>C00-D48</td>
<td>Neoplasms</td>
</tr>
<tr>
<td>III</td>
<td>D50-D89</td>
<td>Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism</td>
</tr>
<tr>
<td>IV</td>
<td>E00-E90</td>
<td>Endocrine, nutritional and metabolic diseases</td>
</tr>
<tr>
<td>V</td>
<td>F00-F99</td>
<td>Mental and behavioural disorders</td>
</tr>
<tr>
<td>VI</td>
<td>G00-G99</td>
<td>Diseases of the nervous system</td>
</tr>
<tr>
<td>VII</td>
<td>H00-H59</td>
<td>Diseases of the eye and adnexa</td>
</tr>
<tr>
<td>VIII</td>
<td>H60-H95</td>
<td>Diseases of the ear and mastoid process</td>
</tr>
<tr>
<td>IX</td>
<td>I00-I99</td>
<td>Diseases of the circulatory system</td>
</tr>
<tr>
<td>X</td>
<td>J00-J99</td>
<td>Diseases of the respiratory system</td>
</tr>
<tr>
<td>XI</td>
<td>K00-K93</td>
<td>Diseases of the digestive system</td>
</tr>
<tr>
<td>XII</td>
<td>L00-L99</td>
<td>Diseases of the skin and subcutaneous tissue</td>
</tr>
<tr>
<td>XIII</td>
<td>M00-M99</td>
<td>Diseases of the musculoskeletal system and connective tissue</td>
</tr>
<tr>
<td>XIV</td>
<td>N00-N99</td>
<td>Diseases of the genitourinary system</td>
</tr>
<tr>
<td>XV</td>
<td>O00-O99</td>
<td>Pregnancy, childbirth and the puerperium</td>
</tr>
<tr>
<td>XVI</td>
<td>P00-P96</td>
<td>Certain conditions originating in the perinatal period</td>
</tr>
<tr>
<td>XVII</td>
<td>Q00-Q99</td>
<td>Congenital malformations, deformations and chromosomal abnormalities</td>
</tr>
<tr>
<td>XVIII</td>
<td>R00-R99</td>
<td>Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified</td>
</tr>
<tr>
<td>XIX</td>
<td>S00-T98</td>
<td>Injury, poisoning and certain other consequences of external causes</td>
</tr>
<tr>
<td>XX</td>
<td>V01-Y98</td>
<td>External causes of morbidity and mortality</td>
</tr>
<tr>
<td>XXI</td>
<td>Z00-Z99</td>
<td>Factors influencing health status and contact with health services</td>
</tr>
<tr>
<td>XXII</td>
<td>U00-U99</td>
<td>Codes for special purposes</td>
</tr>
</tbody>
</table>
Mine ICD10 dictionary terms from the medical record

det drejer sig om en 36-årig sygemeldt mand der overflyttes fra frederiksherg hospital, afdeling m.h.p. længereværende rehabiliteringsofhold. „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 tanketræthed 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 samsone ……… 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 vil 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 pensjonskassen. 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 fobie ………. pt. har her til aften angivet tillagende bivirkninger i form af trækninger i nakken, indre uro og stivhed af fingre. ………. pt. har fået svar på sit ekg, som viser sinus rytmee med enkelte ventrikulære ekstrasyteter uforandret fra tidlig. ……… med baggrund i oplysninger om tidligere maniske episoder præget af irritabilitet, hyperaktivitet og øget seksuel interesse revurderes diagnosen til bipolar affektiv sindsleidelse. følges i distrikt vest med psykologsamtaler. ………. har i dag tydeligt brug for en faglig forklaring på hendes symptomer. det drejer sig om paranoia, uviskerighedsfølelsler, influense sympt. og koncentrationsbesvær. det største problem er dog samarbejdet 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 afsls født tabl. imovane 7,5 mg med god effekt. ……… kl 19, pinex, tabletter, 500 mg indtaget dog: 1 gram for hovedpine ……… pt. er henviset til at
Compare patients by ICD10 terms mined and assigned in terminology space

Roque et al. PLoS Comp. Biol. 2011,
Mined and assigned ICD10 codes

Assigned Codes

Mined Codes

- 4947
- 3825
- 32626
Patient similarity metrics

Patient phenotype similarities:

- Patient similarity measures beyond “words” (using knowledge from ontologies to quantify the similarity of clinical features, biochemical data etc.)
- Semantic harmonization and phenotype harmonization, benchmarking and semantic interoperability e.g. across language barriers for meta-analysis
Clustering in medical terminology patient space

Roque et al.
PLoS Comp. Biol. 2011
Jensen et al.,
Alcohol and depressive disorders, anxiety disorders, and other personality disorders
Significant comorbidities among complex and Mendelian disorders (110M patients, registry data)

Blair et al., Cell 2013
Similarities of complex diseases computed from comorbidity profiles to Mendelian disorders
Which disease-disease and symptom correlations are treatment related?
Spontaneous reports

- Heavily trusted
- Underreporting and biases
- Data quality issues
Summaries of Product Characteristics (SPCs)

Every medicine pack includes a patient information leaflet (PIL). PILs are based on the Summaries of Product Characteristics (SPCs) which describe a medicinal product’s properties and the conditions attached to its use.

Controlled ADR/ADE vocabulary from SPCs
ADR dictionary characteristics

- Descriptions extracted from **7,446 drug SPCs.** Including drugs approved by both European Medicines Agency and Danish Medicines Agency.

- **21,342 uniquely spelled ADRs** was used to construct the dictionary.

- Current version can match about 8,000 different ADRs.

- Final dictionary can match > **4 \cdot 10^{12}** ways to describe ADRs before applying fuzzy matching.
Adverse effect concept linkage

- Elevated creatinine, creatinine elevation, elevation in plasma-Cr are recorded identically.
Text mining of drug names, ADE/ADRs, diagnoses, ...
Text mining Adverse Drug Reactions
(Using 7,500 drug names and 21,000 ADRs)

Eriksson et al. Drug Safety 2014
Frequency of Adverse Drug Reactions from text mining

Comparison between the 150 most statistically significant extracted ADR frequencies and the stated frequencies in the SPC by the manufacturer. Dots are showing the individual extracted frequencies.

Out of the 150 are 6 left out, according to the SPC these are having a frequency < 1/1000 and the corpus is not large enough to detect these with a satisfying frequency.

Recall of 75.1% and a precision of 95.0%.
(manual curation of 200 records)

Eriksson et al. Drug Safety 2014
ADR-dose dependencies
Dosages from structured medication data

Sedation is the most occurring ADR in the corpus.

Drugs selected are 10 antipsychotics (a class known to cause sedation).

ADRs and doses are normalized on multiples of the minimum dose prescribed of each drug.

Plot for 21 days steady dosage data is visualized, sample average slope 0.1105 (95% CI, 0.03085-0.1901), non-zero slope p-value was 0.0074, all individual drug slopes are positive except for haloperidol.

Eriksson et al. Drug Safety 2014
## Possible ADRs?

<table>
<thead>
<tr>
<th>Drug substance</th>
<th>ADE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dipyridamole</td>
<td>Visual impairment</td>
<td>4.375e-04</td>
</tr>
<tr>
<td>Simvastatin</td>
<td>Personality changes</td>
<td>8.408e-08</td>
</tr>
<tr>
<td>Citalopram</td>
<td>Psychosis</td>
<td>8.807e-04</td>
</tr>
<tr>
<td>Bendroflumethiazide</td>
<td>Apoplexy</td>
<td>8.46e-03</td>
</tr>
<tr>
<td>Chlordiazepoxide</td>
<td>Nystagmus</td>
<td>4.03e-08</td>
</tr>
</tbody>
</table>

p-values are multiple testing corrected

X1  Y1  p1
X2  Y2  p2
Drug-ADR similarities
(ADRs text mined by temporal analysis of EHRs)

13 years of drug use at Mental centre Sct. Hans ATC code coloring. Edges show Drug ADR profile similarity, darker edge indicates stronger similarity. Network contains 500 strongest edges (Jaccard index). “N03 Antiepileptics” are scattered, not showing any clustering. This is expected as antiepileptics is a very diverse drug class in terms of ADRs. Laxatives and antibacterials for systemic use group since both cause diarrhea and stomach ache and other gastrointestinal problems.
Comorbidity trajectories in individual patients

Typical development
E.g. type 2 diabetes > problems with foot > amputation

Yellow dot: Debut of a given ICD10 code, green debut of other diseases
Danish Discharge Registry, 6.2 million patients
Danish Patient Registry

Following up in registry data

15 years of **ICD10** hospitalization history for 6.2 million Danes
- 68 million records (in/out of ward)
- 45 million admissions (in/out of hospital)
- 119 million diagnosis-record-associations
- 18 million surgical procedures
- 131 million treatments & examinations

Non-hypothesis driven comorbidity analysis of registry data

---

Patient admission to hospital

Symptoms and clinical findings

Critical diagnoses written in patient journal

Diagnosis verified/declined

Procedures (operations and treatments)

Discharging physician makes a discharge summary including ICD-10 codes from patient journal

Entered in the hospital's registry by a medical secretary ICD-10 codes sent to National Dishcarge registry
National Patient Registry (6.2M Danes)
ICD10 diagnoses as a function of age

[Graphs showing ICD10 diagnoses by age for females and 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 laboratory findings, not elsewhere classified
19. Injury, poisoning and certain other consequences of external causes
20. External causes of morbidity and mortality
Disease trajectories and trajectory-cluster for prostate cancer

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
Cardiovascular trajectory network

AB Jensen et al., Nature Comm., 2014
Disease trajectory: towards an ontology

Reduced 6.2 million trajectories to

1,171 trajectories of 4 steps

Cover 140 diagnoses and >0.5 million unique patients

Clustering identified 5 major disease comorbidity clusters which incorporate ~600 of the trajectories
Musings

The $1,000 genome, the $100,000 analysis?

Elaine R Mardis

Correspondence: Elaine R Mardis emardis@wustl.edu

The Genome Center at Washington University School of Medicine, 4444 Forest Park Blvd, St Louis, MO 63108, USA


The electronic version of this article is the complete one and can be found online at: http://genomemedicine.com/content/2/11/84

Published: 26 November 2010

© 2010 BioMed Central Ltd
De-identification

In the Nordic countries and in some other countries we can minimize the effort on rediscovering what we already know.
EPR and registry analysis

Peter Bjødstrup Jensen, CPR/KU
Robert Eriksson, CBS/DTU/CPR/KU
Anders Bøck Jensen, CBS/DTU/CPR/KU
Teresa A Ajslev, U. Copenhagen
Pope Mosely, U. New Mexico
Tudor Oprea, U. New Mexico
Henriette Schmock, Sct. Hans Hospital
Lars Juhl Jensen, CPR/KU
Thomas Werge, Sct. Hans Hospital
Francisco Simões Roque, CBS/DTU
Eva Roitmann, CBS/DTU
Anders Juhl, Rigshospitalet, Copenhagen
Marlene Dalgaard, Rigshospitalet, Copenhagen
Massimo Andreatta, Copenhagen
Thomas Hansen, Sct. Hans Hospital
Karen Søeby, Hvidovre Hospital
Søren Bredkjær, Region Sealand
Thorkild IS Sørensen, U. Copenhagen

Steno Diabetes Center & Hagedorn
Peter Rossing
Henrik Ullits Andersen
Regine Bergholdt
Thomas Almdal
Flemming Pociot
Torben Hansen, now U. Copenhagen
Oluf Borbye Pedersen, now U. Copenhagen
Mining electronic health records: towards better research applications and clinical care

Peter B. Jensen¹, Lars J. Jensen¹ and Søren Brunak¹,²

Abstract | Clinical data describing the phenotypes and treatment of patients represents an underused data source that has much greater research potential than is currently realized. Mining of electronic health records (EHRs) has the potential for establishing new patient-stratification principles and for revealing unknown disease correlations. Integrating EHR data with genetic data will also give a finer understanding of genotype-phenotype relationships. However, a broad range of ethical, legal and technical reasons currently hinder the systematic deposition of these data in EHRs and their mining. Here, we consider the potential for furthering medical research and clinical care using EHR data and the challenges that must be overcome before this is a reality.

Information technology has transformed the way health care is carried out and documented. Presently, the practice of health care generates, exchanges and stores huge amounts of patient-specific information. In addition to the traditional clinical narrative, databases in modern health centres automatically capture structured data relating to all aspects of care, including diagnosis, medication, laboratory test results and radiological imaging data.

This transformation holds great promise for the individual patient as richer information, coupled with clinical decision support (CDS) systems, becomes readily available at the bedside to support informed decision making and to improve patient safety, especially interesting when traditional health-care-sector data is linked with biobanks and genetic data³.

Despite the great potential, researchers who wish to analyse large amounts of patient data are still faced with technical challenges of integrating scattered, heterogeneous data, in addition to ethical and legal obstacles that limit access to the data³. It is hoped that large-scale adoption of health information technology (HIT) infrastructure in the form of electronic health records (EHRs) and agreed standards for interoperability and schemes for privacy and consent, will improve this situation (Table 1). With incentives for improved public health and the expected health budget savings⁴, these matters