HEALTH BANK – A Workbench for Data Science Applications in Healthcare

Hercules Dalianis

Clinical Text Mining Group

Department of Computer and Systems Sciences (DSV)

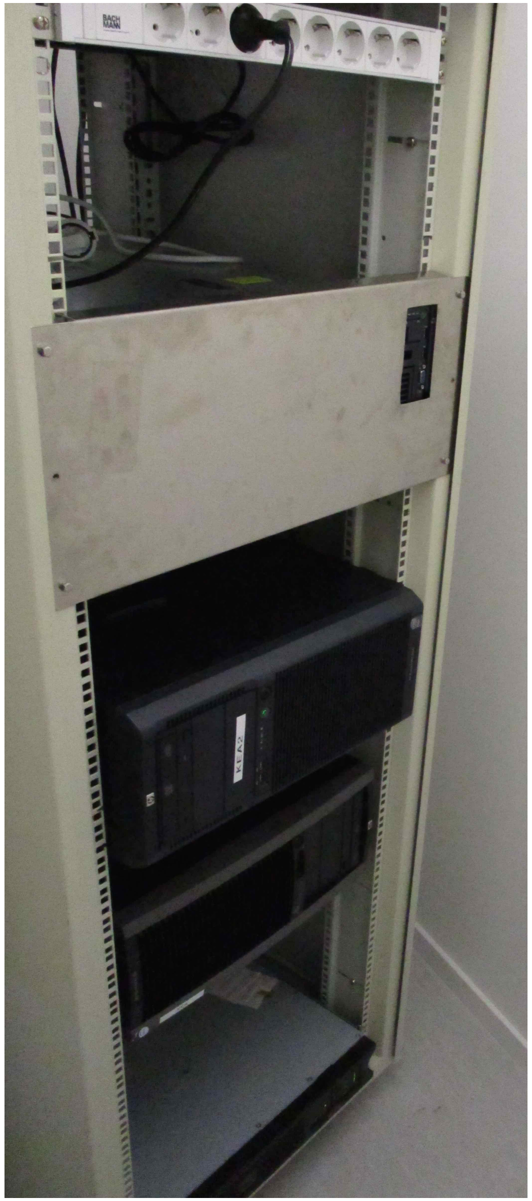
hercules@dsv.su.se

Stockholm University

Clinical text mining group 2007-2014

Aron Henriksson, Mia Kvist, Sara Brissman, Maria Skeppstedt, Gunnar Nilsson, Martin Hassel, Sumithra Velupillai, Hideyuki Tanushi and Hercules Dalianis (Not in photo Claudia Ehrentraut, and Rebecka Weegar)

Claudia Ehrentraut and Rebecka Weegar



HEALTH BANK
2 mil. patient records
2007-2014

Overview of talk

- Background
- Ethics
- The database
- The language in clinical text
- Applications for healthcare
- A workbench
- Collaborations

HEALTH BANK –The Swedish Health Record Research Bank

(Stockholm EPR Corpus)

- Karolinska University Hospital
  - TakeCare Intelligence
- First ethical permission 2008
- First database 2006-2008
- 4th database 2007-2014, ~ 2 million records
- And eight ethical permission!

Content in patient records

- Serial number, gender, age
- Admission, discharge date and time stamps
- Blood-, laboratory values, ICD-10 diagnosis codes
- Drugs - ATC-codes
- **AND LOTS OF!!!**
  Free text in Swedish
  – Physician’s notes, reasoning, nurses narratives, etc

Sensitive text

- In the free text under Social headings
  - Lots of personal names, phone numbers, addresses etc.
    - *The patient is assisted by her husband Anders, ph 070-567 32 55*

Medical language

- Noisy text
- Non-standard abbreviations
  - p5 - *pertrokantär femurfraktur (hip fracture)*
- Misspellings up to 10 percent of all words
  - *Parkisons, frctre*
- Missing subjects, incomplete sentences
- Physician expresses also herself very vague
  - *Possible infection, no signs of bacterias but..,*
  - *No feber, no pain, no Parkisons*

Medicinskt journalspråk

Septisk pat, oklart fokus,
rundodlas före Zinacef

=>

Patienten har sepsis med oklart ursprung,
bakterieodling tas från samtliga möjliga
infektionsfokus, inklusive blododling,
innan behandling med Zinacef inleds.

Medical language

Septicemic pat, unclear origin,
roundcultured before Zinacef.

=>

The patient has septicemia of unclear origin, bacterial culture samples taken from all possible foci for infection, including blood culture samples, before commencing treatment with Zinacef.

ICD-10 coding

- 22 classes and 35,000 codes
    - (J18) Pneumonia, organism unspecified
        - (J18.0) Bronchopneumonia, unspecified
        - (J18.1) Lobar pneumonia, unspecified
        - (J18.2) Hypostatic pneumonia, unspecified
        - (J18.8) Other pneumonia, organism unspecified
        - (J18.9) Pneumonia, unspecified

Challenges in Healthcare – Adverse events

Healthcare-associated infections (HAI)

- International studies have found that up to 10 per cent of patients at any given time has healthcare associated infections, (Humphreys and Smyths, 2006)
- 10 percent or more of the in-patients in Europe obtain an HAI
- Three million injured patients and 50 000 deaths yearly only in Europe.

Adverse drug events (ADE)

- ADEs causes 3.7% of hospital admissions worldwide.
- One of the most common causes of death
- Seventh most common cause of death in Sweden

Diagnosis code assignment and validation

- 20 percent of the assigned ICD-10 diagnosis codes are erroneous
- Many codes to choose among

Stockholm University

Early cancer symptoms

- The earlier cervical cancer, breast cancer or prostate cancer is detected the better
- Are there any early signs or early symptoms of cancer in the previous treatment of the patient?

Statistics and applications

- Costly to assign and validate ICD-10 diagnosis codes, over 20 percent of the assigned codes are erroneous

  => Automatic code assignment and validation using distributional semantics

- At least 10 percent of the admitted patients suffers from healthcare associated infections

  => Automatic detection and predictions of healthcare associated infections

- At least 4 percent of all admitted patient are admitted due to adverse drug events

  => Automatic detection of adverse drug events both known and unknown.

Statistics and applications (cont)

- Mining events preceding a cancer diagnosis.
  => Mining events in cervical cancer including negated events.
- Manual coding of pathology reports at Cancer Registry of Norway (Kreftregisteret) 25 full time coders and 180,000 reports yearly.
  => Automatic coding of pathology reports.
- Write discharge letter (epikris)
  => Automatic summarisation of health care episodes to a discharge letter.

Statistics and applications (cont)

- Patient/lay person reads his/her patient records

  => Automatic simplification of the patient record.
- Miss spellings and grammar errors in the patient record text.

  => Spelling and grammar checking of the patient record text.

Automatic code assignment and validation
(Unsupervised machine learning, random indexing - distributional semantics)

<u>Hosta (cough)</u>

J18.9 - Pneumoni, ospecificerad (Pneumonia, unspecified)
J15.9 - Bakteriell pneumoni, ospecificerad (Bacterial pneumonia, unspecified)
H66.9 - Mellanöreinflammation, ej specificerad som varig / icke varig (Otitis media, unspecified)
J20.9 - Akut bronkit, ospecificerad, (Acute bronchitis, unspecified)
B34.9 - Virusinfektion, ospecificerad, (Viral infection, unspecified)
G96.9 - Sjukdom i centrala nervsystemet, ospecificerad (Disorder of central nervous system, unspecified)
I50.9 - Hjärtinsufficiens, ospecificerad (Heart failure, unspecified)
F48.9 - Neurotiskt syndrom, ospecificerat (Neurotic disorder, unspecified)
C34.9 - Icke specificerad lokalisation av malign tumör i bronk & lunga (Bronchus or lung, unspecified)
L64.9 - Androgen alopeci, ospecificerad (Androgenic alopecia, unspecified)

Fig. 10.7 Example in ICD-10 code suggestion (© 2009 The authors - reprinted with permission from the authors. Published in Dalianis et al. (2009).)

Manual annotation and training with machine learning (Supervised learning)

- Manual annotation or classification
  - symptoms, diagnosis, drugs and body parts
- Training with different machine learning algorithms
  - Recognizing patterns – create rules

Detect-HAI

- Detect-HAI analyses clinical text and data for Healthcare-associated infections
  - Based on 213 manually classified health care episodes (128 HAI/85 NON-HAI)
  - Both text 1,300,000 tokens and structured information
  - 93.7% recall and 79.7% precision using the Gradient Tree Boosting

Manually annotation for detection of clinical named entities

- Program modules for detection of
  - Symptom and diagnosis
  - Negation
  - Uncertainty
  - Period of time

  76-årig kvinna med hypertoni och angina pectoris. Trolig hjärtinfarkt 2 år sedan. Inkommer med centrala bröstsmärtor utan utstrålning.

  - 76-year old woman with hypertension and angina pectoris. Possible heart attack 2 years ago. Admitted to hospital with central chest pain without radiation.

Manually annotation for detection of clinical named entities

- Program modules for detection of
  - Symptom and diagnosis
  - Negation
  - Uncertainty
  - Period of time

  76-årig kvinna med hypertoni och angina pectoris. Trolig hjärtinfarkt 2 år sedan. Inkommer med centrala bröstsmärtor utan utstrålning.

  - 76-year old woman with hypertension and angina pectoris. Possible heart attack 2 years ago. Admitted to hospital with central chest pain without radiation.

Manually annotation for detection of clinical named entities

- Program modules for detection o
  - Symptom and diagnosis
  - Negation
  - Uncertainty
  - Period of time

  76-årig kvinna med hypertoni och angina pectoris. Trolig hjärtinfarkt 2 år sedan. Inkommer med centrala bröstsmärtor utan utstrålning.

  - 76-year old woman with hypertension and angina pectoris. Possible heart attack 2 years ago. Admitted to hospital with central chest pain without radiation.

Manually annotation for detection of clinical named entities

- Program modules for detection of
  - Symptom and diagnosis
  - Negation
  - Uncertainty
  - Period of time

  76-årig kvinna med hypertoni och angina pectoris. Trolig hjärtinfarkt 2 år sedan. Inkommer med centrala bröstsmärtor utan utstrålning.

  - 76-year old woman with hypertension and angina pectoris. Possible heart attack 2 years ago. Admitted to hospital with central chest pain without radiation.

Manually annotation for detection of clinical named entities

Stockholm University

- Program modules for detection of
  - Symptom and diagnosis
  - Negation
  - Uncertainty
  - Period of time

76-årig kvinna med hypertoni och angina pectoris. Trolig hjärtinfarkt 2 år sedan. Inkommer med centrala bröstsmärtor utan utstrålning.

- 76-year old woman with hypertension and angina pectoris. Possible heart attack 2 years ago. Admitted to hospital with central chest pain without radiation.

Manually annotation of internal medicine emergency unit records

- Over 8,000 annotated entities
  - 1,998 Disorders
  - 3,822 Findings
  - 1,539 Drugs
  - 750 Body parts

Stockholm University

Pre-annotation trained on internal medicine emergency unit records, applied on obstetrics/gynecology and oncology Annotation and correction by manual annotators

Finding

1 VRU: Vulva ua.

Vulva, no remarks, abbreviated

Body_part

2 Vagina öppet 3 cm vid introitus, vaginas längd 6 cm.

Finding

3 Finner en rel fast striktur som en ring belägen cirka 2 cm från portioplanet.

Body_part Finding

4 Vid manipulation lättblödande slemhinnor men inget tumörsuspekt.

easily-bleeding mucous membranes but not tumor suspicious

5 Kan inspekteras eller palperas.

Finding Body_part

6 Debridierar digitalt sammanväxtningar på den högra och vänstra sidan av vaginalväggen och kan därefter

vaginal wall

prova ut en vaginaldilatator 30 mm som genomsläpplig ut med hela vaginas längd motsvarande 6 cm till

uretramynningen från portioplanet.

urethral orifice

7 Rekommenderas använda denna dilatator 3 gånger/ vecka tills att slemhinnorna slutar att blöda.

until mucous membranes stops bleeding

8 Får en återbesökstid till UT om cirka 4 månader.

9 Pat har fått en handskriven information med sig hem.

Annotation and corrections by manual annotator

Finding

1 VRU: Vulva ua.

Vulva, no remarks, abbreviated

Body_part Body_part Body_part New annotations

2 Vagina öppet 3 cm vid introitus, vaginas längd 6 cm.

Finding

3 Finner en rel fast striktur som en ring belägen cirka 2 cm från portioplanet.

stricture

New annotation Finding Body_part ~~Finding~~ Negation added

4 Vid manipulation lättblödande slemhinnor men inget tumörsuspekt.

easily-bleeding mucous membranes but not tumor suspicious

5 Kan inspekteras eller palperas.

Finding Body_part

6 Debridierar digitalt sammanväxtningar på den högra och vänstra sidan av vaginalväggen och kan därefter

vaginal wall

New annotation Body_part

prova ut en vaginaldilatator 30 mm som genomsläpplig ut med hela vaginas längd motsvarande 6 cm till

Body_part New annotation

uretramynningen från portioplanet.

urethral orifice

New annotations

Body_part Finding

7 Rekommenderas använda denna dilatator 3 gånger/ vecka tills att slemhinnorna slutar att blöda.

until mucous membranes stops bleeding

8 Får en återbesökstid till UT om cirka 4 månader.

9 Pat har fått en handskriven information med sig hem.

Table 7. The most frequent findings, disorders and negations found in the physicians' notes.

Most frequent findings and disorders	Nbr of instances	Most frequently negated findings and disorders	Nbr of negated instances	Findings and disorders with highest portion of negation	Portion negated
cervixcancer (cervical cancer)	873	besvär (trouble/problem)	338	gynekologiska besvär (gynecological problems)	1.0
besvär (trouble/problem)	790	feber (fever)	243	palpabla resistenser (palpable resistance)	1.0
illamående (nausea)	677	illamående (nausea)	198	särskilda besvär (particular problems)	1.0
mår bra (feels well)	662	blödningar (bleedings)	171	nytillkomna symtom (new symptoms)	0.96
smärta (pain)	656	smärta (pain)	150	nytillkomna besvär (new problems)	0.92
tumör (tumor)	642	smärtor (pains)	126	infektionstecken (signs of infection)	0.89
smärtor (pains)	629	blödning (bleeding)	99	biljud (murmur)	0.86
feber (fever)	562	infektionstecken (signs of infection)	91	tumörstrukturer (tumour structures)	0.83
cancer (cancer)	508	tumör (tumor)	83	tumörsuspekta förändringar (tumor suspicious changes)	0.82
blödningar (bleedings)	491	nytillkomna besvär (new troubles)	79	subjektiva besvär (subjective problems)	0.8
blödning (bleeding)	482	buksmärtor (pain of the abdomen)	72	tumörsuspekt (tumor suspicion)	0.80
skivepitelcancer (squamous cell carcinoma)	428	hydronefros (hydronephrosis)	65	spridning (spreading)	0.78

Kreftregisteret i Oslo
The Cancer Registry in Norway


KURSBEVIS
brystkreft

Pathology report for breast cancer in Norwegian

Mammaresektat (ve. side) med infiltrerende duktalt karsinom, histologisk grad 3
Tumordiameter 15 mm
Lavgradig DCIS med utstrekning 4 mm I kranial retning fra tumor
Frie reseksjonsrender for infiltrerende tumor (3 mm kranialt)
Lavgradig DCIS under 2 mm fra kraniale reseksjonsrand
ER: ca 65 % av cellene positive
PGR: negativ
Ki-67: Hot-spot 23% positive celler. Cold spot 8%.
Gjennomsnitt 15%
HER-2: negativ
Tidl. BU 13:
3 sentinelle lymfeknuter uten påviste patologiske forandringer

⇒ **Found diagnostic tests by a mockup system**

Progesteronreseptorer: 1 (1 is a table value that corresponds to "negative")
Østrogenresepttorer: 4 (4 is a table value that corresponds to "65 %")
KI67 Hotspot: 23
KI67 Gjennomsnitt hot/cold: 1
Tumors histologiske grad : 3

Samtidig Sentinell Node: 0
HER-2 Immunihistokjemi: 2 (table value)
Tumordiameter: 15

Courtesy by Rebecka Weegar

Translated to English:

```
Mamma specimen (le. side) with infiltrating ductal
carcinoma, histological grade 3
Tumor diameter 15 mm
Low-grade DCIS extending 4 mm in cranial direction
from the tumor
Free resection margins for infiltrating tumor (3 mm
cranially)
Low-grade DCIS less than 2 mm from the cranial
resection margin

ER: ca 65 % of the cells are positive
PGR: negative
Ki-67: Hot-spot 23% positive cells. Cold spot 8%.
Average 15%
HER-2: negative
Prev. BU 13:

3 sentinel lymph nodes without proven pathological
changes
```

Found concepts by a mockup system

```
Progesteronreseptorer (PGR): 1 (1 is a table value
that corresponds to "negative" in the text)
Samtidig Sentinell Node: 0
Østrogenreseptorer (ER): 4 (4 is a table value
that corresponds to "65 %" in the text)
KI67 Hotspot: 23
Tumors histologiske grad (Histological grade): 3
KI67 Gjennomsnitt hot/cold (Average): 15
Tumordiameter (Tumor diameter): 15
(In some cases the data is not found in the text
but in sketch attached to the pathology report).
```

Stockholm University

Automatic summarisation of health care episodes to a discharge letter.

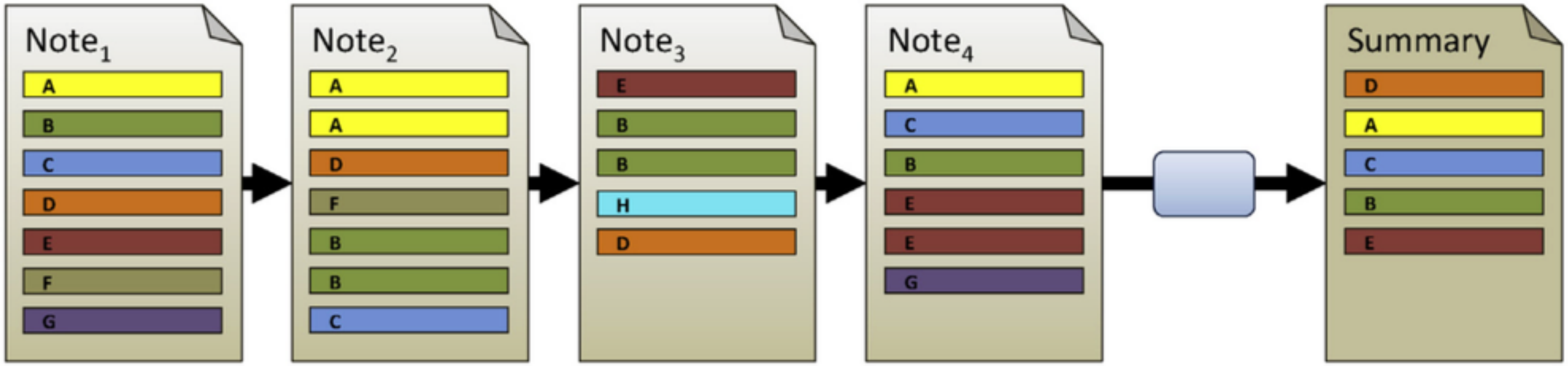


Fig. 10.5 Example on automatic discharge summary creation. Redundant information is removed and highscoring information is added in the beginning of the summary from highest to lowest, low scoring information G, F and H, are excluded. Taken from Figure 3 in (Moen et al., 2016), (Licensed under Creative Commons.)

Stockholm University

Data science applications in healthcare

- Automatic surveillance of healthcare-associated infections
- Detection and exploration of adverse drug events
- Diagnosis code assignment
- Text mining in the cancer domain
  - Cervical cancer- detect early symptoms
  - Pathology reports – extract diagnostic tests
- Text simplification of clinical narratives/discharge summary
- Text input, spell checking
- Comorbidity analysis

Comorbidity view

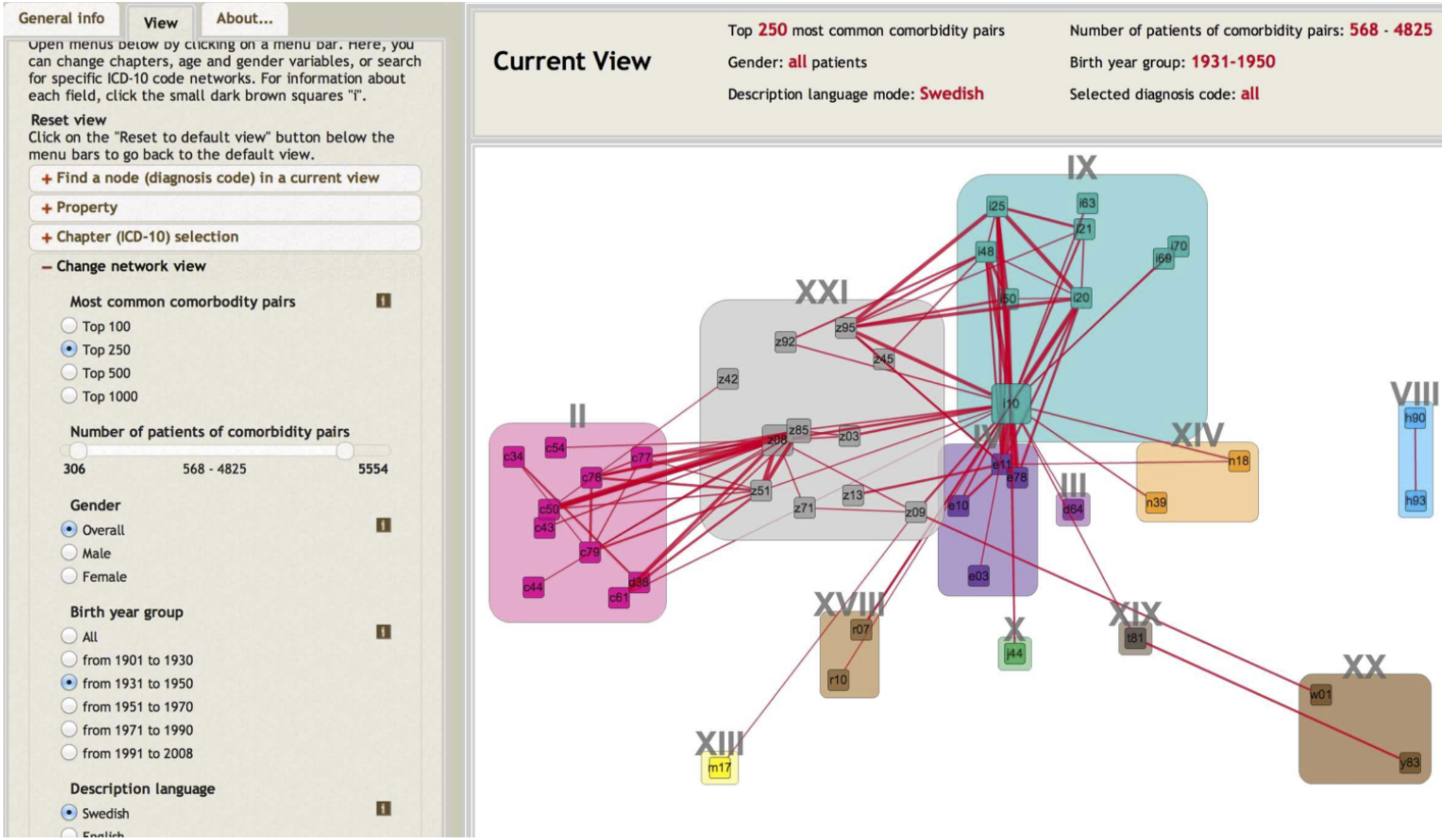
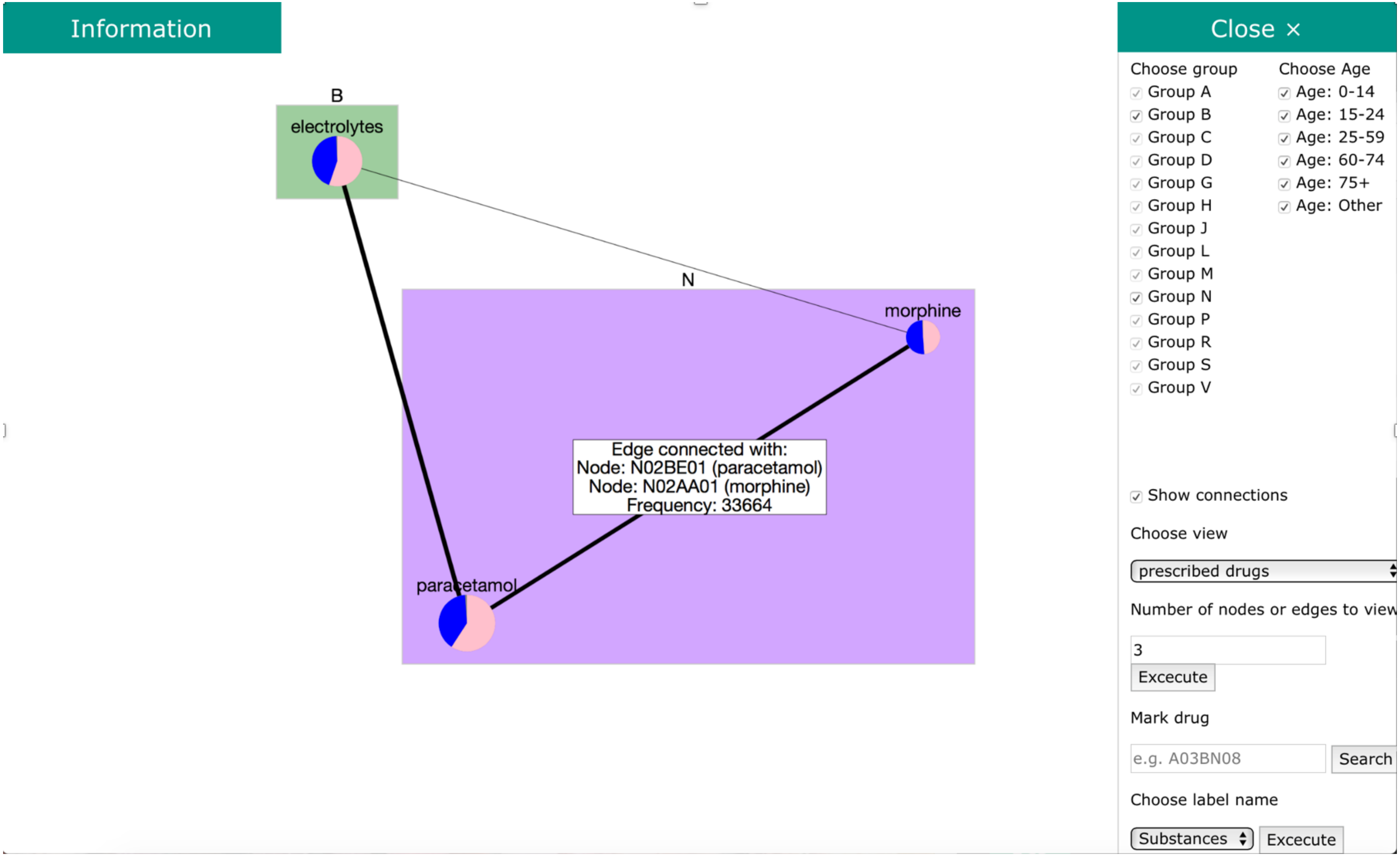


Fig. 10.10 Screen shoot of Comorbidity-view which is a visualising tool for comorbidity network. It contains all the disorders in ICD-10 code form that patient records have been assigned. The data contains 605,587 patients from 2006-2008 from the Karolinska University hospital. The thicker the line in Comorbidity-view the more patients have both ICD-10 codes (Tanushi et al., 2011).

Drug view



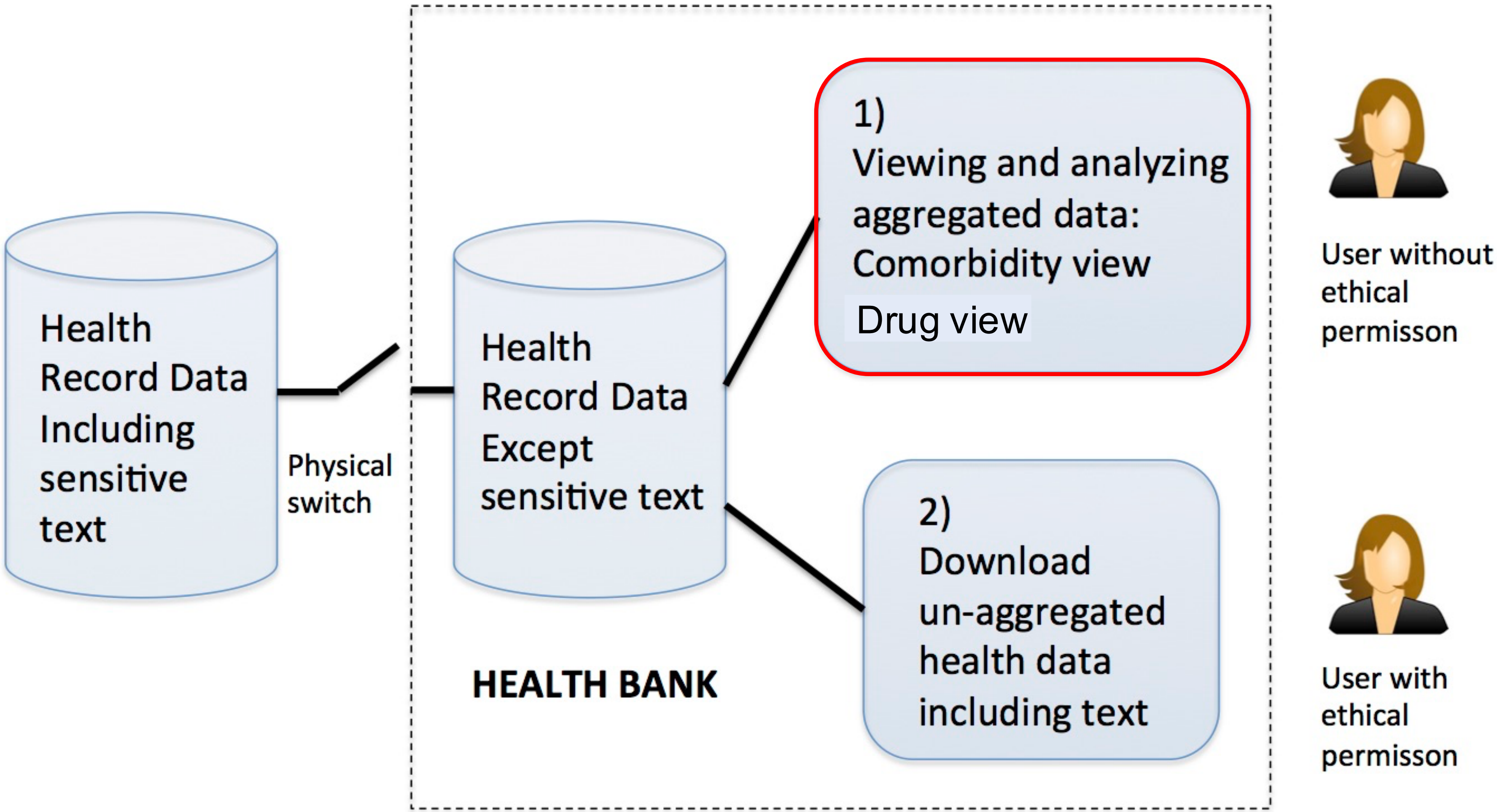
Information
B
electrolytes
N
morphine
Edge connected with:
Node: N02BE01 (paracetamol)
Node: N02AA01 (morphine)
Frequency: 33664
paracetamol
Close ×
Choose group
Group A
Group B
Group C
Group D
Group G
Group H
Group J
Group L
Group M
Group N
Group P
Group R
Group S
Group V
Choose Age
Age: 0-14
Age: 15-24
Age: 25-59
Age: 60-74
Age: 75+
Age: Other
Show connections
Choose view
prescribed drugs
Number of nodes or edges to view
3
Excecute
Mark drug
e.g. A03BN08
Search
Choose label name
Substances
Excecute

HEALTH BANK – the Swedish Health Record Research Bank

- An infrastructure
- A workbench
- A data exploration tool
- Give access to data to both researchers and industry

Researchers

- Epidemiologists
- Pharmacologists
- Medical researchers
  - To generate and evaluate hypotheses
- Data scientists
  - To build systems


Health
Record Data
Including
sensitive
text
Physical
switch
Health
Record Data
Except
sensitive text
HEALTH BANK
1)
Viewing and analyzing
aggregated data:
Comorbidity view
Drug view
2)
Download
un-aggregated
health data
including text
User without
ethical
permisson
User with
ethical
permisson

HEALTH BANK

- Contain all possible tools for visualisation of electronic patient records
- Contain all possible tools for processing clinical text
- For secondary use of data

Current academic users

- Sweden
  - Stockholm University
  - Karolinska Institutet
  - Karolinska University Hospital
  - Uppsala University
  - Gothenburg University
  - University of Borås

Abroad

- University of Turku, Finland
- University of Copenhagen and DTU- Danmarks Tekniske Universitet
- NTNU-Trondheim, Norway
- Vytautas Magnus University, Lithuania,
- UC San Diego and University of Utah, USA
- SAS Institute and Treat Systems, Denmark

New text book April 2018 open access

Hercules Dalianis

Clinical Text Mining

Secondary Use of Electronic Patient Records

Springer Open

Comments & Questions?

Stockholm University

Hercules Dalianis

Clinical Text Mining

Secondary Use of Electronic Patient Records

Open access here:

https://www.springer.com/gp/book/9783319785028

Springer Open

Stockholm University

Stockholm University

Stockholm University

Stockholm University

Stockholm University

Stockholm University