Data Mining of Medical Data: Opportunities and Challenges

Dan A. Simovici
IALS
Cecilienhof Potsdam
Brandemburg, Germany

UMB

Data Mining Processes

Mining Tabular Data

AR and Nosocomial Infections

Association Rules and Adverse Drug Reactions

Transitivity of Association Rules

An opening quotation by L. L. Weed

> *The beginning clinical clerk, the house officer and the practicing physician are all confronted with conditions that are frustrating in every phase of medical action. ... To deal effectively with these frustrations it will be necessary to develop a more organized approach to the medical record, a more rational acceptance and use of the paramedical personnel and a more positive attitude about the computer in medicine.*

L. L. Weed: Medical records that guide and teach, New England Journal of Medicine, 1968

Knowledge Discovery Through Data Mining

Data Mining (DM) is the process that discovers new patterns embedded in large data sets. DM makes use of this information to build predictive models.

DM is grounded in

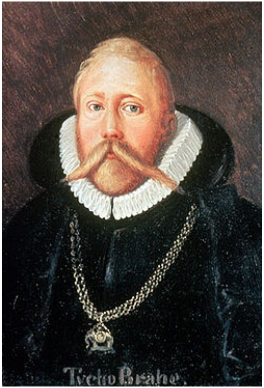
- artificial intelligence
- databases
- statistics

The Role of DM in Healthcare

- huge and complex volumes of data are generated by healthcare activities; un-automated analysis has become impractical;
- DM can generate information that can be useful to all stakeholders in health care, including patients by identifying effective treatments and best practices;
- the existence of insurance fraud and abuse impels insurers to use DM.

Early data collection and data mining

DM came into prominence in mid 90s, but the history is much older ....



Dec. 14, 1546, Knutstorp Castle
- Oct. 24, 1601, Prague
Univ. of Copenhagen
Univ. of Rostock
Accurate Collector of
Enormous Observation
Data Set

Dec. 27, 1571, Weil der Stadt
-Nov. 15, 1630, Regensburg
Tübinger Stift
Discover of Laws of
Planetary Motions

Main DM Activities

- description and visualization;
- association;
- clustering;
- classification and estimation.

Typical Healthcare DM Applications

- treatment effectiveness;
- healthcare management;
- improving customer relationship management;
- fraud and abuse detection;

Limitations of DM

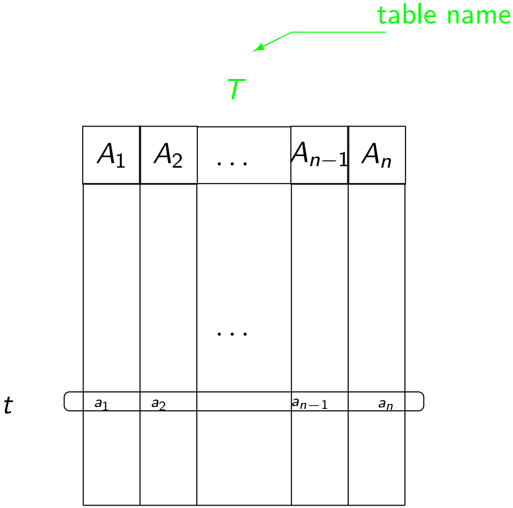
- DM can be limited by the accessibility to data that often is distributed in different settings (clinical, administration, insurers, labs, etc.);
- data may be incomplete, corrupted, noisy, or inconsistent;
- ethical, legal and social issues (data ownership, privacy concerns);
- many patterns find in DM may be the result of random fluctuations, so many such patterns may be useless;
- DM of medical data requires specific medical knowledge as well as knowledge of DM technology.
- DM requires institutional commitment and funding.

Tables

A table is an aggregate that consists of

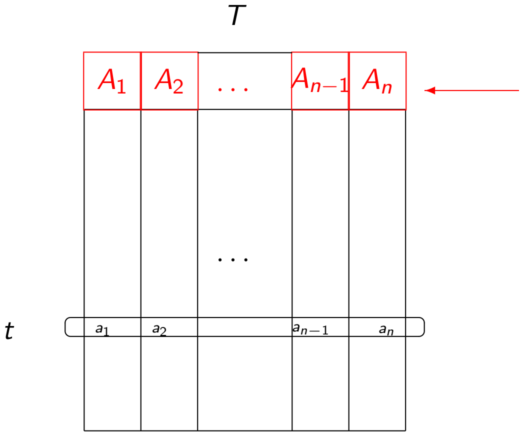
- a *table name*;
- a *heading* that contains a set $A_1, \ldots, A_n$ of symbols called *attributes*;
- a *content* that is a multiset of rows: we could have multiple copies of the same row;
- each attribute A has a *domain* $\mathrm{dom}(A_i)$, a set that contains at least two elements;
- a row is a sequence of values $(a_1, \ldots, a_n)$, such that a_i is a member of $\mathrm{dom}(A_i)$ for $1 \leq i \leq n$.

Components of Tables



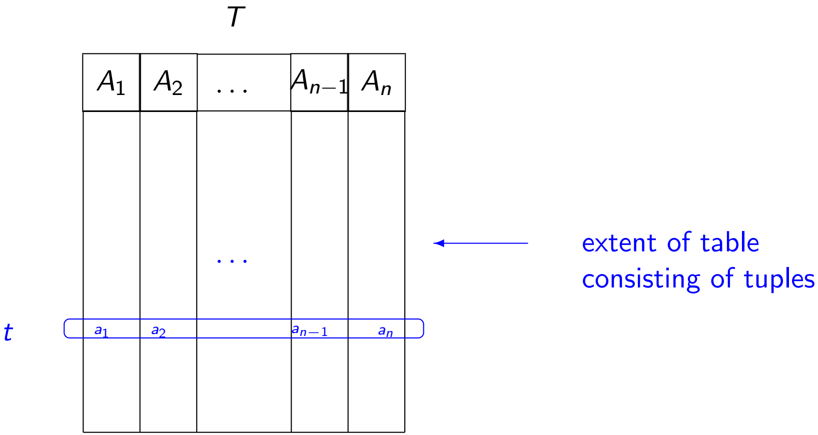
	A_1	A_2	$\dots$	A_{n-1}	A_n
			$\dots$		
t	a_1	a_2		a_{n-1}	a_n

Components of Tables



heading consisting of attributes $A_1, \ldots, A_n$

Components of Tables



T

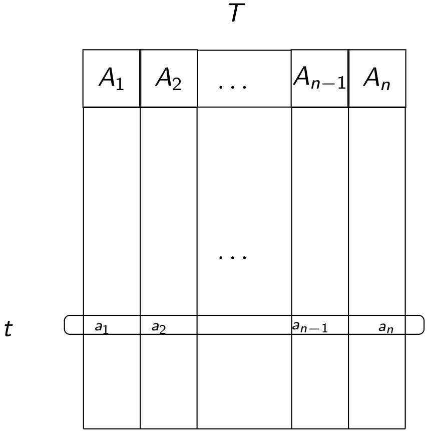
A_1 A_2 $\dots$ A_{n-1} A_n

$\dots$

t a_1 a_2 a_{n-1} a_n

extent of table
consisting of tuples

Conceptualization of Tables



T

	A_1	A_2	$\ldots$	A_{n-1}	A_n
			$\ldots$		
t	a_1	a_2		a_{n-1}	a_n

Example

OBJECTS

	shape	length	width	height	color
1	cube	5	5	5	red
2	sphere	3	3	3	blue
3	pyramid	5	6	4	blue
4	cube	2	2	2	red
5	sphere	3	3	3	blue

Attributes:

- categorial: shape, color
- numerical: length, width, height

Binary Tables

- all attributes have the domain $\{0, 1\}$;
- tuples are sequences of 0s and 1s;
- binary tables have the capability of recording collections of sets: items purchased at a supermarket, medicines prescribed for a treatment, preferred treats for dogs, etc.

Example

Customer basket	Content
1	{milk, bread, butter, diapers}
2	{bread, beer, diapers}
3	{milk, bread, butter, beer}
4	{bread, butter, diapers}
5	{milk, butter, beer, diapers}
6	{milk, butter}
7	{butter, beer}

An equivalent tabular form:

	milk	bread	butter	beer	diapers
t_1	1	1	1	0	1
t_2	0	1	0	1	1
t_3	1	1	1	1	0
t_4	0	1	1	0	1
t_5	1	0	1	1	1
t_6	1	0	1	0	0
t_7	0	0	1	1	0

Support of an attribute set

	A	B	C	D	R
t_1	1	1	0	0	0
t_2	1	0	1	0	0
t_3	0	0	0	1	0
t_4	1	1	1	0	0
t_5	1	0	0	0	1
t_6	1	0	0	0	1
t_7	1	1	0	0	0

The *support of X* is the number of tuples that have 1 components corresponding to all attributes of X.

$$\text{supp}(A) = 6, \text{supp}(B) = 3, \text{supp}(AB) = 3, \text{supp}(ABC) = 1$$

Support of an attribute set (cont'd)

- the larger the attribute set, the smaller the support: $X \subseteq Y$ implies $\mathrm{supp}(Y) \leq \mathrm{supp}(X)$;
- $\mathrm{supp}(X)$ estimates the probability that a randomly chosen transaction t contains all elements of X;

Association Rules as Embedded Knowledge

An *association rule* (AR) is a pair (X, Y) of sets of attributes, denoted by $X \rightarrow Y$. X is the *antecedent* and Y is the *consequent* of the rule $X \rightarrow Y$. Parameters of an AR:

- *the support of a rule* $X \rightarrow Y$ is the number of records that contain all items of X and Y:

$$\text{supp}(X \rightarrow Y) = \text{supp}(X);$$

- *the confidence of a rule* $X \longrightarrow Y$ is

$$\text{conf}(X \rightarrow Y) = \frac{\text{supp}(XY)}{\text{supp}(X)};$$

- the confidence of $X \longrightarrow Y$ is an estimation of the probability that a record that contains the items of X, chosen at random, will contain the items of Y.

Example

	A	B	C	D	R
t_1	1	1	0	0	0
t_2	1	0	1	0	0
t_3	0	0	0	1	0
t_4	1	1	1	0	0
t_5	1	0	0	0	1
t_6	1	0	0	0	1
t_7	1	1	0	0	0

For the AR $AB \rightarrow C$, support equals 3 and confidence is

$$\text{conf}(AB \rightarrow C) = \frac{\text{supp}(ABC)}{\text{supp}(AB)} = \frac{1}{3} = 0.33$$

An AR $X \rightarrow Y$ holds with support μ and confidence c if $\text{supp}(XY) \geq \mu$ and $\text{conf}(X \rightarrow Y) \geq c$.

Complexity of identifying AR

- AR of the form $X \longrightarrow Y$ with $Y \subseteq X$ are *vacuous* because

$$\text{conf}(X \longrightarrow Y) = \frac{\text{supp}(XY)}{\text{supp}(X)} = 1;$$

  such rules are not informative and they are also referred to as *trivial rules*;
- for n attributes there are $3^n - 2^n$ non-trivial possible rules;
- the number of possible rules is very large and the number of collection of possible rules is immense even for modest values of n; for $n = 20$ there exist more that one billion non-trivial AR and more that $10^{300000000}$ sets of AR (for comparison, there are 10^{80} atoms in the known universe!).

Frequent Item Sets

To find an AR $X \to Y$ with support μ and confidence c we need to:

- find an item set U that is at least μ-frequent, that is, $\text{supp}(U) \geq \mu$;
- find a subset V of U such that $\text{supp}(V) \leq \frac{\text{supp}(U)}{c}$.

U and V define the AR $X \to Y$, where $X = U - V$ and $Y = V$, having support at least equal to μ and confidence ar least equal to c.
Thus, computing association rule amounts to computing frequent item sets.

The Apriori Algorithm

The most common algorithm is the *Apriori algorithm* by Agrawal, Imielinski and Swami:

- detect all items that that have a support at least equal to μ;
- for successive numbers $i \geq 2$ join item sets that contain i items with individual item sets (candidate generation phase);
- evaluate the resulting item sets and retain only those who have sufficient support (evaluation phase).

Where It All Started...

- R. Agrawal, T. Imielinski, A. N. Swami: Mining Association Rules between Sets of Items in Large Databases, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, 207-216.
- R. Agrawal, H. Mannila, R. Srikant, H. Toivonan, A. I. Verkamo; Fast discovery of association rules. In: Fayyad UM, Piatetsky- Shapiro G, Smyth P, Uthurusamy R (eds). Advances in Knowledge Discovery and Data Mining. Cambridge, Mass.: MIT Press, 1996:30728.
- R. Agrawal, J. Schaffer: Parallel mining of association rules. IEEE Trans Knowl Data Eng. 1996;8:9629.
- M. J. Zaki, S. Parthasarathy, M. Ogihara, W. Li: New algorithms for fast discovery of association rules. In: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining. Menlo Park, Calif.: AAAI Press, 1997:2836.

AR in Hospital Infection Control

S. E. Brosette, A. P. Sprague, J. M. Hardin, K. B. Waites, W. T. Jones and S. A. Moser: "Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance", "J. of the American Medical Information Association", 1998, 5, 373–381.

- *Pseudomonas aeruginosa* is a common Gram-negative bacterium that can cause nosocomial infections in humans;
- symptoms of infections are generalized inflammation and sepsis;
- medical equipment, including catheters, contaminated by *P. aeruginosa* causes cross-infections in hospitals and clinics.

Data Format

Each record describes a single *Pseudomonas aeruginosa* isolate
Attributes:

- date reported;
- source of isolate (sputum, blood);
- location of patient in the hospital;
- patient's home zip code;
- resistant (R), intermediate resistance (I), susceptible (S) for
  piperacillin,
  ticarcillin/clavulanate,
  ceftazidime
  imipenem
  amikacin
  gentamicine
  tobramycine
  ciprofloxacin

Data Pre-processing

- duplicate records are removed, so each patient has one isolate per month;
- this results in about 80 non-duplicate records per month;
- the system is designed to detect patterns of increasing resistance to antimicrobials; therefore, items of the form *S*-antimicrobial were removed.

Algorithmic Approach

- data is partitioned horizontally in time-slices; in each slice AR with high support are sought and their confidence is computed;
- the variation between the confidence of a rule $X \rightarrow Y$ in the current time-slice and the confidence of the same in previous time slices is calculated;
- if a substantial increase in the confidence occurs (as verified using a statistical test) relative to the previous partition(s), this finding constitutes an *event*.

Experimental Design

Data was partitioned horizontally in
A. 12 one-month fragments: 2,000 rules;
B. 4 three-month fragments: 12,000 rules;
C. 2 six-month fragments: 20,000 rules.
Minimum support for an item was 2 and minimum support for an AR was 10.
Patterns sought:

1. short-lived interesting patterns in slices of type A;
2. long-lived interesting patterns in slices of type C.

Experimental Design (cont'd)

A relatively small number of rules were presented to the user as shown below:

Experiment	A	B	C
	34	57	28

AR of the form $\emptyset \rightarrow Y$ have

$$\text{supp}(\emptyset \rightarrow Y) = \text{supp}(\emptyset) = n;$$

and

$$\text{conf}(\emptyset \rightarrow Y) = \frac{\text{supp}(Y)}{n},$$

which shows that only confidence is significant and equals the probability of Y.
Thus, conf(R-antimicrobial) gives the probability that *Pseudomonas aeruginosa* develops resistance to the antimicrobial; variation in the level of confidence are evaluated on an monthly, quarterly, and semestrial basis.

Evaluation of Changes in Confidence

Rules are selected based on the variance in their confidence.

- For each AR $X \rightarrow Y$ the confidence in P_c, $\text{conf}(X \rightarrow Y, P_c)$ is compared with the confidence of $X \rightarrow Y$ in the last data set P_d in which $X \rightarrow Y$ was found which precedes P_c, $\text{conf}(X \rightarrow Y, P_d)$.
- The comparison of confidences is done using a χ^2-square comparison of two proportions, or when the number of expected value is small, by the Fisher exact test.
- If $\text{conf}(X \rightarrow Y, P_c) \geq \text{conf}(X \rightarrow Y, P_d)$ and the probability that the difference between the proportions occurred by chance is less than 5%, then this finding is presented to the user.

AR Found

$\emptyset \rightarrow$ R-ticarcillin/clavulanate R-ceftazidime R-piperacillin

INTERPRETATION: a jump from 4%(Oct) to 8%(Nov) to 11%(Dec)suggests that the isolate is resistant to ticarcillin/clavulanate, ceftazidime and piperacillin

R-ceftazidime R-piperacillin $\rightarrow$ sputumR-ticarcillin/clavulanate

INTERPRETATION: 8%(Feb)-32%(Aug) it is likely that the isolate is from sputum and is ticarcillin resistent given that is resistant to ceftazidime and piperacillin

R-piperacillin $\rightarrow$ sputumR-ticarcillin/clavulanateR - ceftazidime

INTERPRETATION: an increase from 6% (Q3) to 26% (Q4)in the probability that the isolate is from sputum, is ticarcillin/clavulanate and ceftazidime resistant given that is piperacillin resistant

R-ticarcillin/clavulanate $\longrightarrow$ sputumR-ceftazidimeR-piperacillin

INTERPRETATION: an increase from 7% (Q3) to 24% (Q4) in the probability that isolate is from sputum, is ceftazidime and piperacillin resistant given that is ticarcilline/clavulanate resistant

R-ticarcillin/clavulanateR-ceftazidimeR-piperacillin $\longrightarrow$ sputum

INTERPRETATION: an increase from 12% (Q3) to 42% (Q4)in the probability that the isolate is from sputum given that it is resistent to ticarcillin/clavulanate, ceftazidime, and piperacillin

Adverse Effects of Drug Reactions

Adverse drug reactions are monitored at

- Uppsala Monitoring Center, Sweden, since 1978), a unit of WHO that mines data originating from individual case safety reports (ICSRs) and maintains Vigibase, a WHO case safety report (access costs money);
- at Food and Drug Administration (FDA), a US federal unit that maintains the AERS database (Adverse Event Reporting System) where access is free.
- at various pharma units who, by US law, must record adverse reactions to drugs.

Adverse Effects of Drug Reactions (cont'd)

Sources:

- R. Harpaz, H. S. Chase, C. Friedman: *Mining multi-item drug adverse effect associations in spontaneous reporting systems*, 2010 at the AMIA Summit on Translational Bioinformatics San Francisco, and
- A. M. Wilson, L. Thabane, A. Holbrook: *Application of data mining techniques in pharmacovigilance*, British Journal of Clinical Pharmacology, 2003, 57, 127–134.

The Seriousness of ADE

Wilson et al.:
ADE account for

- 5% of hospital admissions (Pirmohamed);
- 28% of af emergency department visits (Patel);
- 5% of hospital deaths (Juntti-Patinen).

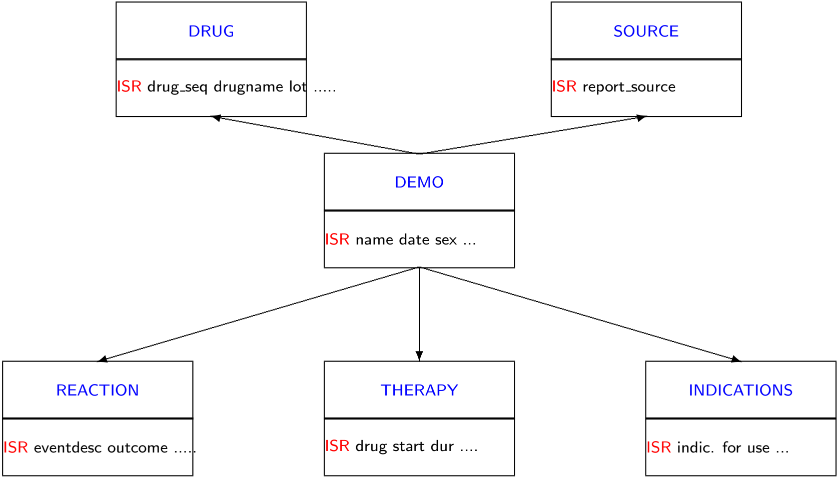
Adverse Effects of Drug Reactions (cont'd)

- multi-item adverse drug event (ADE) associations are associations relating multiple drugs to possibly multiple adverse events.
- ADEs result in losses of several billion dollars annually, and may cause harm to patients.
- pushed the current standard in pharmacovigilance from binary association rules , where each single drug-adverse effect combination is studied separately to general association rules (ADEs) of the form $X \longrightarrow A$, where X describes a set of drugs rather than a single drug;
- based on a set of 162,744 reports of suspected ADEs reported to AERS and published in the year 2008, this approach identified 1167 multi-item ADE associations.

Computational Aspects of Seeking ADE Association Rules

- Items are partitioned into two broad classes: drugs and symptoms;
- AR have the form $X \longrightarrow Y$, where X is a set of drugs and Y is a set of symptoms;
- system goes beyond seeking simple rules of the form Vioxx $\rightarrow$ Heart Attack; rules that have larger sets of drugs and adverse symptoms are especially interesting;
- given a set of drugs X it is important to find the largest set of symptoms Y such that $X \longrightarrow Y$ has a certain level of support and confidence;
- indexing (hashing) based on drugs and on symptoms was used to speed up searches in the data.

Architecture of the ADE Data



DRUG

ISR drug_seq drugname lot .....

SOURCE

ISR report_source

DEMO

ISR name date sex ...

REACTION

ISR eventdesc outcome .....

THERAPY

ISR drug start dur ....

INDICATIONS

ISR indic. for use ...

ISR: unique number for identifying an EARS report

- A taxonomy that characterizes the associations was developed based on a representative sample.
- 67 percentages of potential multi-item ADE associations identified were characterized and clinically validated by a domain expert as previously recognized ADE associations.

Filtering of AR

Filtering can be done based on interestingness measures (confidence is just one of them).

In this case confidence is inappropriate since rules of the form $X \rightarrow$ NAUSEA will have high confidence due to the high frequency of NAUSEA.

Alternatives for a rule $X \longrightarrow Y$:

	Y	$\bar{Y}$
X	a	b
$\bar{X}$	c	d

Int. Measure	Formula
support	a
confidence	$\frac{a}{a+c}$
χ^2	$\frac{(ad-bc)^2(a+b+c+d)}{(a+b)(a+c)(b+c)(b+d)}$
interest (lift)	$\frac{a(a+b+c+d)}{a+b)^2}$
conviction	$\frac{(a+c)(b+d)}{(a+b+c+d)c}$

Relative Reporting Ratio (RR) as an interestingness measure

n: the total number of records.

$$\text{RR} = \text{conf}(X \to Y) \cdot \frac{n}{\text{supp}(Y)}$$

is the confidence of the AR $A \longrightarrow Y$ normalized by the relative support of the consequent Y.

- RR is symmetric relative to X and Y.
- Large values of RR suggest that the occurrence of drugs-adverse reactions is larger than in the general collection of drugs.

Taxonomy of multi-item ADE associations

1a	Drug-drug interactions found that are known
1b	Drug-drug combinations known to be given together or treat same indi
1c	Drug-drug combinations that seem to be due to confounding
1d	Drug-drug interactions that are unknown

Associations

2a	Associations (drug[s]-event) that are known	67%
2b	Associations (drug[s]-event) that are unknown	33%

Sample of multi-item ADE associations

Association Support RR

1a-2a	metformin metoprolol → NAUSEA	50	7.4
1b-2a	cyclophosphamide, prednisone, vincristine → FEBRILE NEUTROPENIA	78	45
1c-2a	cyclophosphamide, doxorubicin, prednisone, rituximab → FEBRILE NEUTROPENIA	63	59
1b-2b	atorvastatin, lisinopril → DYSPNOEA	55	3.5
1a-2b	omeprazole simvastatin → DYSPNOEA	58	12
1d-2b	varenicline darvocet → ABNORMAL DREAMS, FATIGUE, INSOMNIA,MEMORY IMPAIRMENT, NAUSEA	52	2668

Specific Association Rules

- metformin metoprolol → NAUSEA: each drug triggers nausea, so this is forseable;
- cyclophosphamide, prednisone, vincristine → FEBRILE NEUTROPENIA: combination used in cancer; known complication;
- atorvastatin, lisinopril → DYSPNOEA: nice personal surprize!;
- varenicline darvocet → ABNORMAL DREAMS, FATIGUE, INSOMNIA,MEMORY IMPAIRMENT, NAUSEA had an RR of 2668: 62 reports contain both drugs, and 53 contained all the AE affecting the central nervous system; this suggests a high degree of duplication in reporting!

An Example

Wright-Chen-Maloney (at BWH in Boston):

Sample data:

p_1	(lisinopril, multivitamin, hypertension)
p_2	(insulin, metformin, lisinopril,diabetes, hypertension)
p_3	(insulin, diabetes)
p_4	(metformin, diabetes)
p_5	(metformin, polycystic ovarian syndrome)
$\vdots$	$\vdots$

Study included 100,000 patients and involved medications, lab results, and problems
Encoding was done using proprietary terminologies that eventually mapped to:

- problems were coded using SNOMED CT;
- laboratory results were encoded using LOINC;
- medications were encoded using RxNorm.

Co-morbidities

- surprisingly strong associations between apparently unrelated items atributable to co-morbidities, e.g, insulin $\longrightarrow$ hypertension due to the coexistence of hypertension and diabetes;
- previous remark highlights the need of mining for co-morbidities;
- AR do not enjoy transitivity: this means that if $X \longrightarrow Y$ and $Y \longrightarrow Z$ are AR with sufficient support and confidence, no conclusion can be drawn about $X \longrightarrow Z$.

Example

For the data set

	A	B	C
t_1	1	1	0
t_1	0	1	1

and $A \rightarrow B$ and $B \rightarrow C$ we have

$$\text{supp}(A \rightarrow B) = 50\%, \text{conf}(A \rightarrow B) = 100\%,$$
$$\text{supp}(B \rightarrow C) = 100\%, \text{conf}(B \rightarrow C) = 50\%.$$

but

$$\text{supp}(A \rightarrow C) = 50\% \text{ and } \text{conf}(A \rightarrow C) = 0\%.$$

Yet another example

For the data set

	A	B	C
t_1	1	0	1
t_1	0	1	0

and $A \rightarrow B$ and $B \rightarrow C$ we have

$$\begin{aligned} \text{supp}(A \rightarrow B) = 50\%, \text{conf}(A \rightarrow B) = 0\%, \\ \text{supp}(B \rightarrow C) = 50\%, \text{conf}(B \rightarrow C) = 0\%. \end{aligned}$$

but

$$\text{supp}(A \rightarrow C) = 50\% \text{ and } \text{conf}(A \rightarrow C) = 100\%.$$

So, the confidence of $A \rightarrow C$ is unrelated to either $\text{conf}(A \rightarrow B)$ or to $\text{conf}(B \rightarrow C)$.

Dealing with the lack of transitivity

- starting from existent AR $X \rightarrow Y$ and $Y \rightarrow Z$ which have a satisfactory medical interpretation investigate if $X \rightarrow Z$ is supported by data (the TransMiner system of Narayanasamy et al.);
- the reverse approach: starting from an AR $X \rightarrow Z$ (e.g. insulin $\rightarrow$ hypertension identify candidate item sets Y such that $X \rightarrow Y$ and $Y \rightarrow Z$ are plausible association rules (Wright et al.);

Y could be diabetes or other co-morbidities of hypertension; once these cases are excluded the confidence of insulin $\rightarrow$ hypertension decreases sharply.

Conclusions and Open Problems

- DM cannot replace the human factor in medical research; however it can be a precious instrument in epidemiology, pharmacovigilance.
- Interaction between DM and medical research is beneficial for both domains; biology and medicine suggest novel problems for DM and ML.
- Extending association mining to unstructured data (progress reports, radiology reports, operative notes, outpatient notes)
- Developing information-theoretical techniques for AR evaluation.

Yet another quotation from L. L. Weed:

35 years later:

> *Knowledge should be held in tools that are kept up to date and used routinely–not in heads, which are expensive to load and faulty in the retention and processing of knowledge.*

(L.L. Weed, M.D.: *New connections between medical knowledge and patient care*, British Medical Journal, 1997)

🇺🇸 🇬🇧 Thank you for your attention!

🇩🇪 Vielen Dank für Ihre Aufmerksamkeit!

🇧🇷 Obrigado pela vossa atençõ!

Presentation and text can be found at

www.cs.umb.edu/˜dsim