Master thesis presentation: Mining adverse events from healthcare data

Dries Van Daele

KU Leuven, DTAI

6 september, 2013

Problem description

Problem: voluntary reporting records a fraction of the adverse events

manual detection

- is costly
- is limited in scope
- is driven by intuition

data mining

- treats all patient data uniformly
- can reason over all relevant data
- enables automation
- makes biases explicit

Problem description

Task: Apply data mining on a given database to detect adverse events or discover novel adverse event triggers

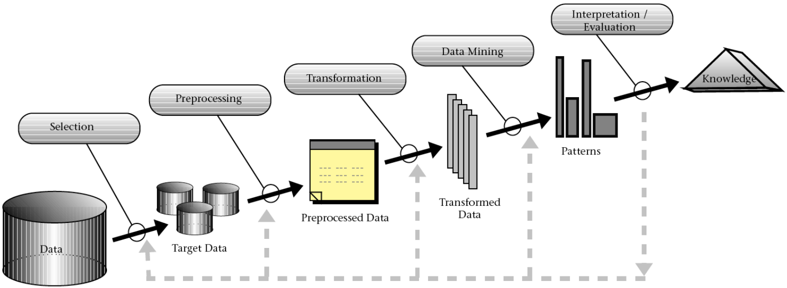
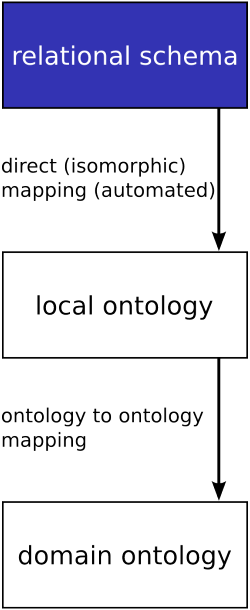


Figure: http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

Data Preparation

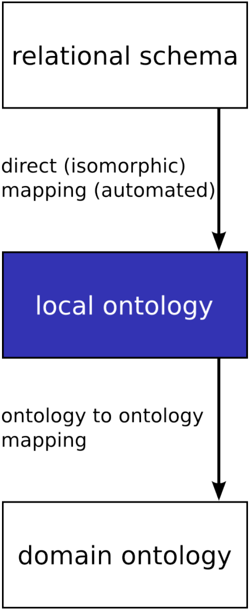


patient_id	birth_date	gender
patient_1	07-MAR-1965	M
...	...	...

medical_case_id	patient_id	admission_date	discharge_date
medical_case_100	patient_1	12-JUL-2007	28-JUL-2007
medical_case_101	patient_1	03-FEB-2008	06-FEB-2008
...	...	...	...

diagnose_id	medical_case_id	ICD_code
diagnose_1	medical_case_101	J01
diagnose_2	medical_case_101	N18
...	...	...

Data Preparation




```
@prefix diagnose: <http://www.example.org/ddo/diagnose#>.
@prefix medical-case: <http://www.example.org/ddo/medical-case#>.
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.

diagnose:Diagnose a rdfs:Class.
diagnose:diagnose-id a rdf:Property ;
                       rdfs:domain diagnose:Diagnose ;
                       rdfs:range xsd:Literal.
diagnose:medical-case-id a rdf:Property ;
                           rdfs:domain diagnose:Diagnose ;
                           rdfs:range medical-case:Medical-Case.
diagnose:icd-code a rdf:Property ;
                    rdfs:domain diagnose:Diagnose ;
                    rdfs:range xsd:Literal.

...
```

relational schema

direct (isomorphic) mapping (automated)

local ontology

ontology to ontology mapping

domain ontology

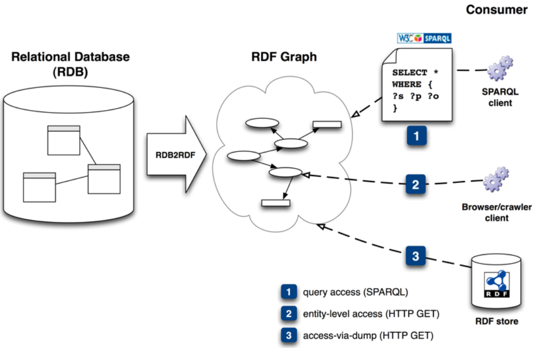
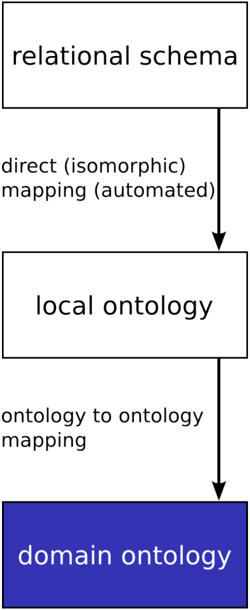


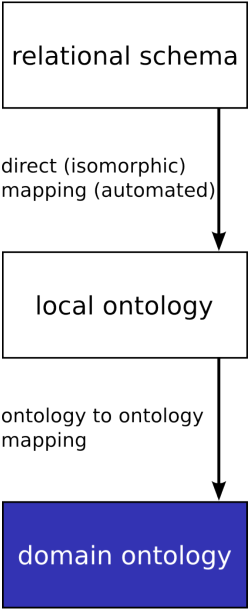
Figure:
http://www.w3.org/2001/sw/rdb2rdf/use-cases/

Data Preparation



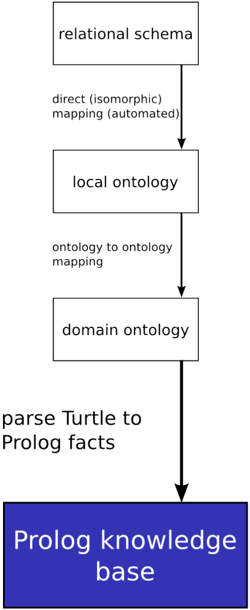

```
@prefix patient: <http://www.example.org/ddo/patient#>.
@prefix heca: <http://www.agfa.com/w3c/2009/healthCare#>.
{
    ?patient a patient:Patient.
} => {
    ?patient a heca:Patient.
}.

...
```

Data Preparation



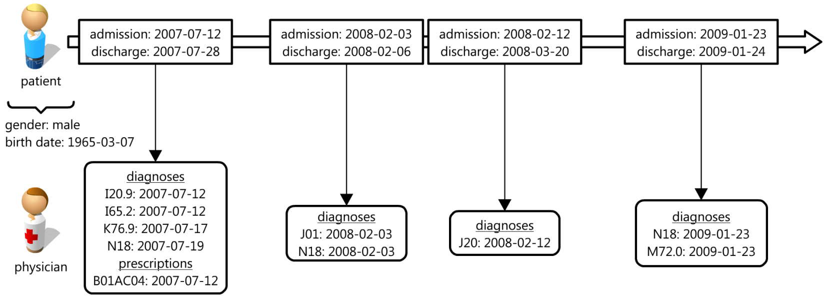

```
@prefix orgStructure: <http://www.example.org/ddo/orgStructure#>.
@prefix space: <http://eulersharp.sourceforge.net/2003/03swap/space#>
{
    _:structure
        orgStructure:structID ?s ;
        orgStructure:innerStructID ?is .
} => {
    ?is space:containedBy ?s .
}.

{
    ?startNode space:containedBy ?middleNode .
    ?middleNode space:containedBy ?endNode .
} => {
    ?startNode space:containedBy ?endNode .
}.

...
```

Data Preparation



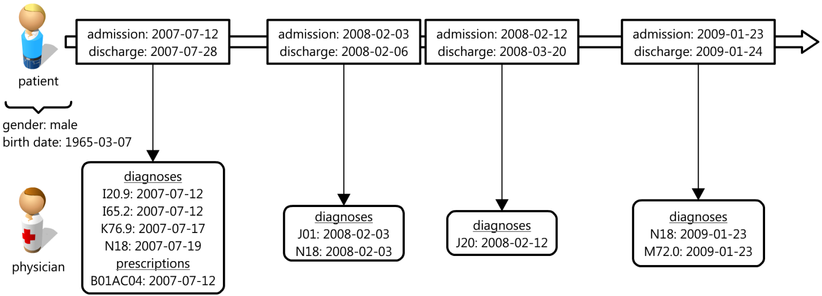
- include ATC and ICD hierarchy
- add ad hoc predicates ($\approx$ 150 e.g. discretise age, distinguish acute from chronic disorders)
- final theory $\approx 43,000,000$ clauses (103,256 patients, 15649 drugs prescribed)

Data mining approach




```
patient_gender(patient_1, male).
patient_birth_date(patient_1, '1965-03-07').
patient_medical_case_dates(patient_1, medical_case_100, '2007-07-12','2007-07-28').
patient_disorder(patient_1, 'I20.9', '2007-07-12').
patient_disorder(patient_1, 'I65.2', '2007-07-12').
patient_disorder(patient_1, 'K76.9', '2007-07-17').
patient_disorder(patient_1, 'N18', '2007-07-19').
patient_drug(patient_1, 'B01AC04', '2007-07-12').
...
```

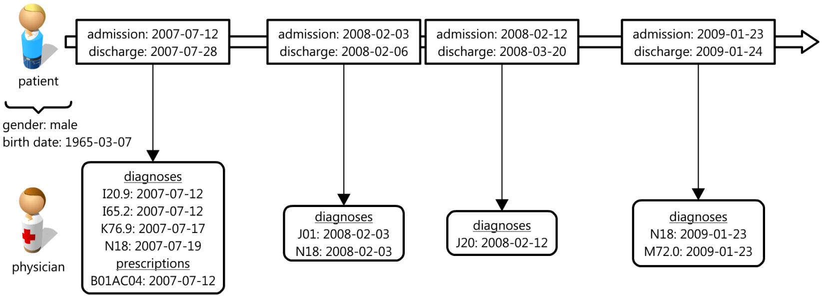
Data mining approach



Problem:

- whether a disorder is an adverse event depends entirely on context, difficult to capture in a monolithic model

Data mining approach



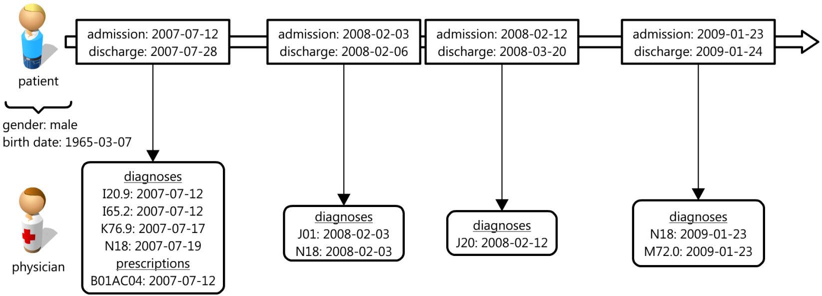
Problem:

- whether a disorder is an adverse event depends entirely on context, difficult to capture in a monolithic model

Solution:

- split detection task up by treatment or trigger

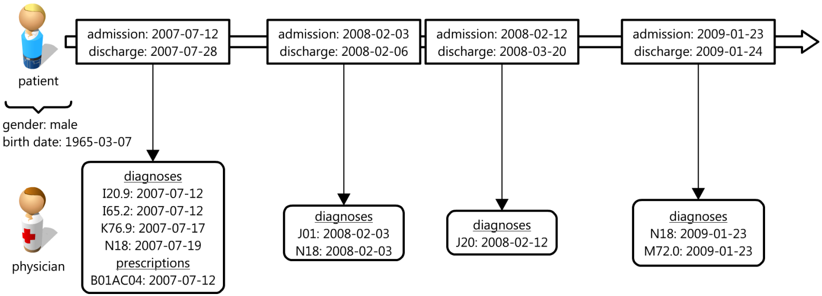
Data mining approach



Problem:

- no labelled data

Data mining approach



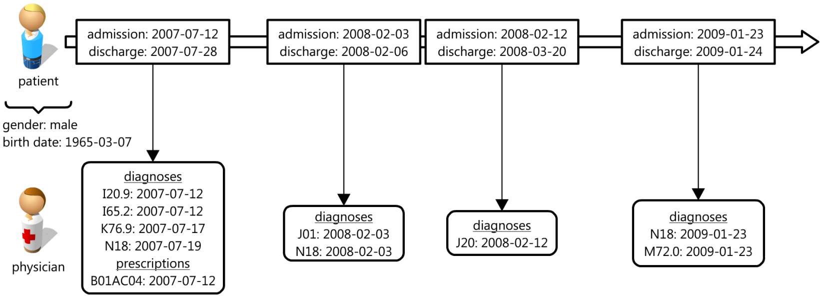
Problem:

- no labelled data

Solution:

- reverse machine learning to target relevant patterns

Data mining approach



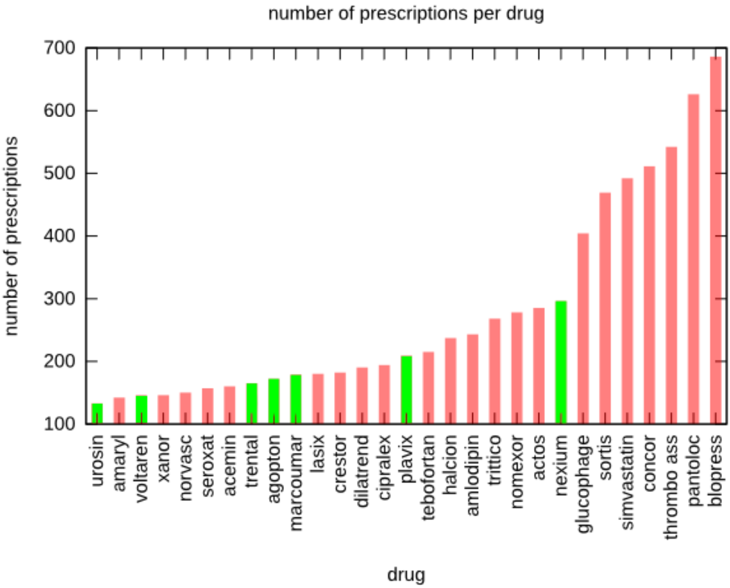
Problem:

- no labelled data

Solution:

- reverse machine learning to target relevant patterns
- performed MONADIC experiment

Collecting evidence



Collecting evidence

	positive examples	negative examples
BUN>60 mg/dl	531	1477
readm. within 30 days	37058	222388
Agopton	141	141
Marcoumar	166	166
Nexium	224	224
Plavix	174	174
Trental	80	80
Urosin	121	121

Table: Target attribute characteristics

Exploratory manual detection of adverse events

Problem:

- No notion of amount of adverse events in data

Exploratory manual detection of adverse events

Problem:

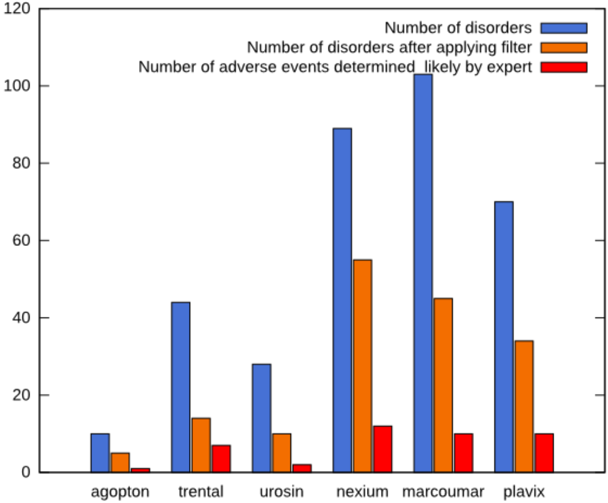
- No notion of amount of adverse events in data

Solution:

- Manually evaluate portion of the drug data (one drug retention period after prescription)

Exploratory manual detection of adverse events

Used FAERS data to filter disorders:



Data mining technique characteristics (Aleph)

- support relational input data

Data mining technique characteristics (Aleph)

- support relational input data
  - examples are Prolog facts
  - hypothesis is a set of clauses (learning a predicate)
  - background knowledge can rely on all Prolog features

Data mining technique characteristics (Aleph)

- support relational input data
  - examples are Prolog facts
  - hypothesis is a set of clauses (learning a predicate)
  - background knowledge can rely on all Prolog features
- perform induction

Data mining technique characteristics (Aleph)

- support relational input data
  - examples are Prolog facts
  - hypothesis is a set of clauses (learning a predicate)
  - background knowledge can rely on all Prolog features
- perform induction
  - Progol algorithm

Data mining technique characteristics (Aleph)

- support relational input data
  - examples are Prolog facts
  - hypothesis is a set of clauses (learning a predicate)
  - background knowledge can rely on all Prolog features
- perform induction
  - Progol algorithm
- incorporate reverse machine learning

Data mining technique characteristics (Aleph)

- support relational input data
  - examples are Prolog facts
  - hypothesis is a set of clauses (learning a predicate)
  - background knowledge can rely on all Prolog features
- perform induction
  - Progol algorithm
- incorporate reverse machine learning
  - using positive and negative examples of the chosen target attribute

Data mining technique characteristics (Aleph)

- support relational input data
  - examples are Prolog facts
  - hypothesis is a set of clauses (learning a predicate)
  - background knowledge can rely on all Prolog features
- perform induction
  - Progol algorithm
- incorporate reverse machine learning
  - using positive and negative examples of the chosen target attribute
- handle probability

Data mining technique characteristics (Aleph)

- support relational input data
  - examples are Prolog facts
  - hypothesis is a set of clauses (learning a predicate)
  - background knowledge can rely on all Prolog features
- perform induction
  - Progol algorithm
- incorporate reverse machine learning
  - using positive and negative examples of the chosen target attribute
- handle probability
  - clause evaluation functions rely on frequency counts of the true and false positives

Aleph usage

space of legal clauses is extremely large, search is not sufficiently targeted:

- vocabulary consisting of derivative predicates
- user-defined prune statements (e.g. limit number of disorders referenced in a clause)
- integrity constraints

acceptable clauses need to

- not be trivial: cover at least 2 positive examples
- have a precision of at least 60%

Conclusions: BUN

		Predicted example class		
		Positive	Negative	Total
Actual example class	Positive	65	466	531
	Negative	1	1476	1477
	Total	66	1942	2008

Table: confusion matrix for the theory on the BUN$>$60 mg/dl trigger

- predict as positive only those examples that contain likely adverse events
- return rules that strongly correlate with the targeted treatment/trigger: only 1 FP

Conclusions: Medication & Readmission

- Medication:
  - Problem: limited amount of data + heterogeneous set of adverse events

Conclusions: Medication & Readmission

- Medication:
  - Problem: limited amount of data + heterogeneous set of adverse events
  - Conclusion: strongest correlation with other treatments (combination therapy), majority of likely AEs indistinguishable from random events

Conclusions: Medication & Readmission

- Medication:
  - Problem: limited amount of data + heterogeneous set of adverse events
  - Conclusion: strongest correlation with other treatments (combination therapy), majority of likely AEs indistinguishable from random events
- Readmission within 30 days:
  - Problem: rules focus on patterns involving chronic disorders

Conclusions: Medication & Readmission

- Medication:
  - Problem: limited amount of data + heterogeneous set of adverse events
  - Conclusion: strongest correlation with other treatments (combination therapy), majority of likely AEs indistinguishable from random events
- Readmission within 30 days:
  - Problem: rules focus on patterns involving chronic disorders
  - Solution: introduce predicate that excludes chronic disorders & neoplasms

Conclusions: Medication & Readmission

- Medication:
  - Problem: limited amount of data + heterogeneous set of adverse events
  - Conclusion: strongest correlation with other treatments (combination therapy), majority of likely AEs indistinguishable from random events
- Readmission within 30 days:
  - Problem: rules focus on patterns involving chronic disorders
  - Solution: introduce predicate that excludes chronic disorders & neoplasms

  - Problem: limiting available diagnoses to specific period ignores useful information

Conclusions: Medication & Readmission

- Medication:
  - Problem: limited amount of data + heterogeneous set of adverse events
  - Conclusion: strongest correlation with other treatments (combination therapy), majority of likely AEs indistinguishable from random events
- Readmission within 30 days:
  - Problem: rules focus on patterns involving chronic disorders
  - Solution: introduce predicate that excludes chronic disorders & neoplasms

  - Problem: limiting available diagnoses to specific period ignores useful information
  - Solution: allow diagnoses that capture relevant health status (e.g. diagnosed with allergies)

Conclusions: Medication & Readmission

- Medication:
  - Problem: limited amount of data + heterogeneous set of adverse events
  - Conclusion: strongest correlation with other treatments (combination therapy), majority of likely AEs indistinguishable from random events
- Readmission within 30 days:
  - Problem: rules focus on patterns involving chronic disorders
  - Solution: introduce predicate that excludes chronic disorders & neoplasms

  - Problem: limiting available diagnoses to specific period ignores useful information
  - Solution: allow diagnoses that capture relevant health status (e.g. diagnosed with allergies)

  - Conclusion: reasonable set of interesting rules can be retrieved

Questions

Thanks for your attention