Data Quality in Biomedical Text Mining

Ulf Leser, Humboldt-Universität zu Berlin

Data Quality [en.wikipedia.org/wiki/Data_quality]

- Degree of excellence exhibited by the data in relation to the portrayal of the actual scenario.
- The state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use.
- The totality of features and characteristics of data that bears on its ability to satisfy a given purpose.

- Fitness for use

	General information (How to humans function?)	**Specific information (What happens in this human?)**
Structured data	Public Biomedical Databases	Clinical Information Systems
Unstructured data	Mining Scientific Articles	Mining Medical Documentation

Table of Content

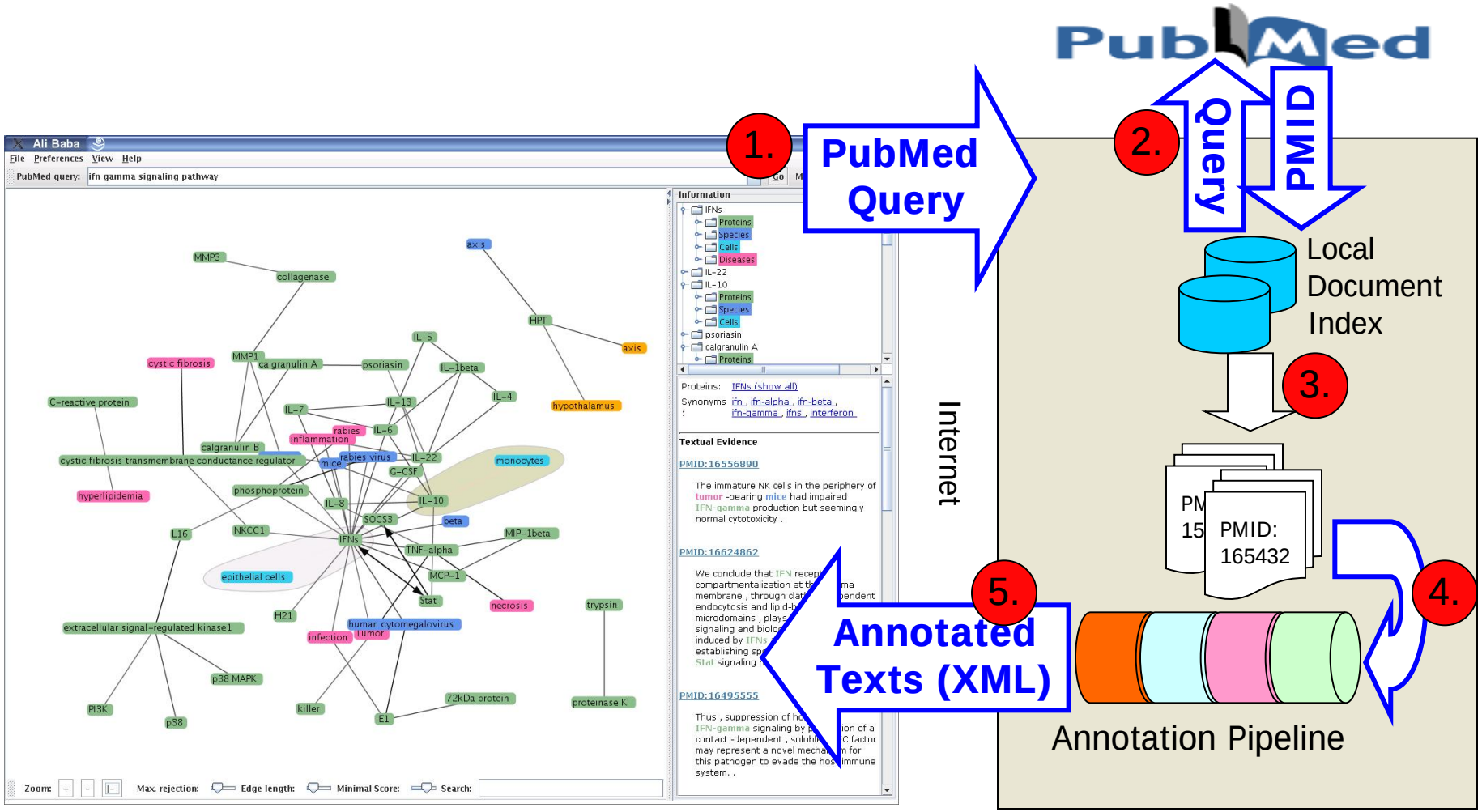
- Biomedical Text Mining
- Quality in BTM of Scientific Articles
- Quality in BTM Medical documentation

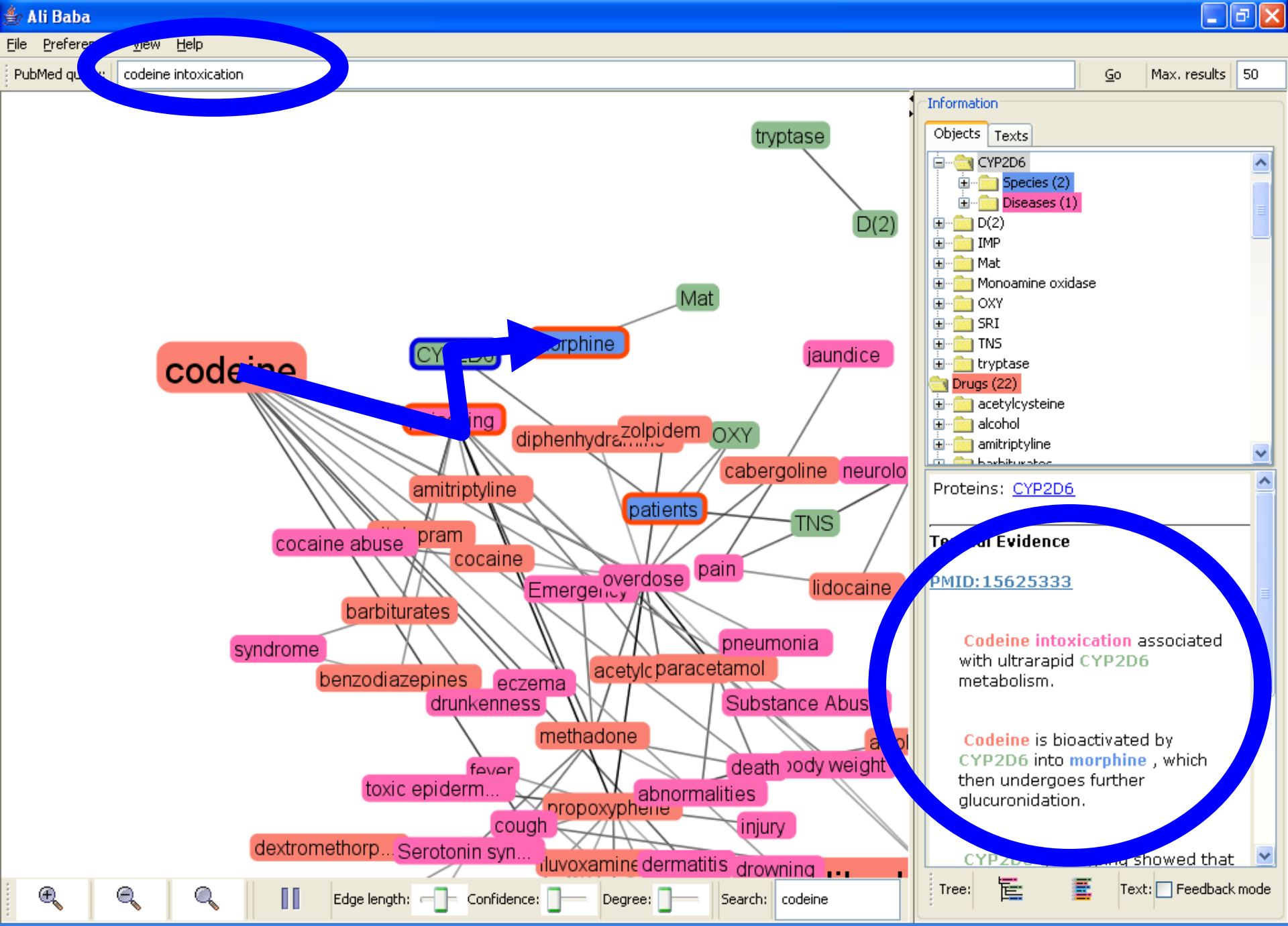
Case Report

- Patient with pneumonia and cough
- Normal dosage of codeine
- Patient not responding any more at day 4
- What‘s going on?
  - PubMed „Codeine intoxication“ -> 70 abstracts
  - Aren’t there better ways?

Case report from Univ. Hospital Geneva, thanks to Christian Meisel, Roche

AliBaba (Plake et al. 2006, Hakenberg et al. 2010)




Ali Baba
File
Help
PubMed
codeine intoxication
Go
Max. results
50
Information
Objects
Texts
CYP2D6
Species (2)
Diseases (1)
D(2)
IMP
Mat
Monoamine oxidase
OXY
SRI
TNS
tryptase
Drugs (22)
acetylcysteine
alcohol
amitriptyline
Proteins: CYP2D6
PMID:15625333
Codeine intoxication associated with ultrarapid CYP2D6 metabolism.
Codeine is bioactivated by CYP2D6 into morphine , which then undergoes further glucuronidation.
tryptase
D(2)
Mat
codeine
jaundice
zolpidem
OXY
cabergoline
amitriptyline
patients
TNS
cocaine abuse
cocaine
pain
lidocaine
overdose
barbiturates
syndrome
pneumonia
benzodiazepines
eczema
paracetamol
drunkenness
Substance Abus
methadone
death
abnormalities
toxic epiderm...
cough
injury
dextromethorp...
Serotonin syn...
dermatitis
drowning
Edge length:
Confidence:
Degree:
Search:
codeine
Tree:
Text:
Feedback mode

Input

Z-100 is an arabinomannan extracted from Mycobacterium tuberculosis that has various immunomodulatory activities, such as the induction of interleukin 12, interferon gamma (IFN-gamma) and beta-chemokines. The effects of Z-100 on human immunodeficiency virus type 1 (HIV-1) replication in human monocyte-derived macrophages (MDMs) are investigated in this paper. In MDMs, Z-100 markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed amphotropic Moloney murine leukemia virus or vesicular stomatitis virus G envelopes. Z-100 was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the env gene is defective and the nef gene is replaced with the firefly luciferase gene) when this vector was transfected directly into MDMs. These findings suggest that Z-100 inhibits virus replication, mainly at HIV-1 transcription. However, Z-100 also downregulated expression of the cell surface receptors CD4 and CCR5 in MDMs, suggesting some inhibitory effect on HIV-1 entry. Further experiments revealed that Z-100 induced IFN-beta production in these cells, resulting in induction of the 16-kDa CCAAT/enhancer binding protein (C/EBP) beta transcription factor that represses HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of p38 mitogen-activated protein kinases (MAPK), indicating that the p38 MAPK signalling pathway was involved in Z-100-induced repression of HIV-1 replication in MDMs. These findings suggest that Z-100 might be a useful immunomodulator for control of HIV-1 infection.

Find Entities

Z-100 is an ***arabinomannan*** extracted from Mycobacterium tuberculosis that has various immunomodulatory activities, such as the induction of **interleukin 12**, **interferon gamma** (**IFN-gamma**) and beta-chemokines. The effects of ***Z-100*** on human immunodeficiency virus type 1 (HIV-1) replication in human **monocyte-derived macrophages** (**MDMs**) are investigated in this paper. In **MDMs**, ***Z-100*** markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed amphotropic Moloney murine leukemia virus or vesicular stomatitis virus G envelopes. ***Z-100*** was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the ***env*** gene is defective and the ***nef*** gene is replaced with the ***firefly luciferase*** gene) when this vector was transfected directly into **MDMs**. These findings suggest that ***Z-100*** inhibits virus replication, mainly at HIV-1 transcription. However, ***Z-100*** also downregulated expression of the cell surface receptors **CD4** and **CCR5** in **MDMs**, suggesting some inhibitory effect on HIV-1 entry. Further experiments revealed that ***Z-100*** induced **IFN-beta** production in these cells, resulting in induction of the 16-kDa **CCAAT/enhancer binding protein** (**C/EBP**) **beta transcription factor** that represses HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of **p38 mitogen-activated protein kinases** (**MAPK**), indicating that the **p38 MAPK** signalling pathway was involved in ***Z-100***-induced repression of HIV-1 replication in **MDMs**. These findings suggest that ***Z-100*** might be a useful immunomodulator for control of HIV-1 infection.

Normalize Entities

Tax: 1773

Entrez: 3458

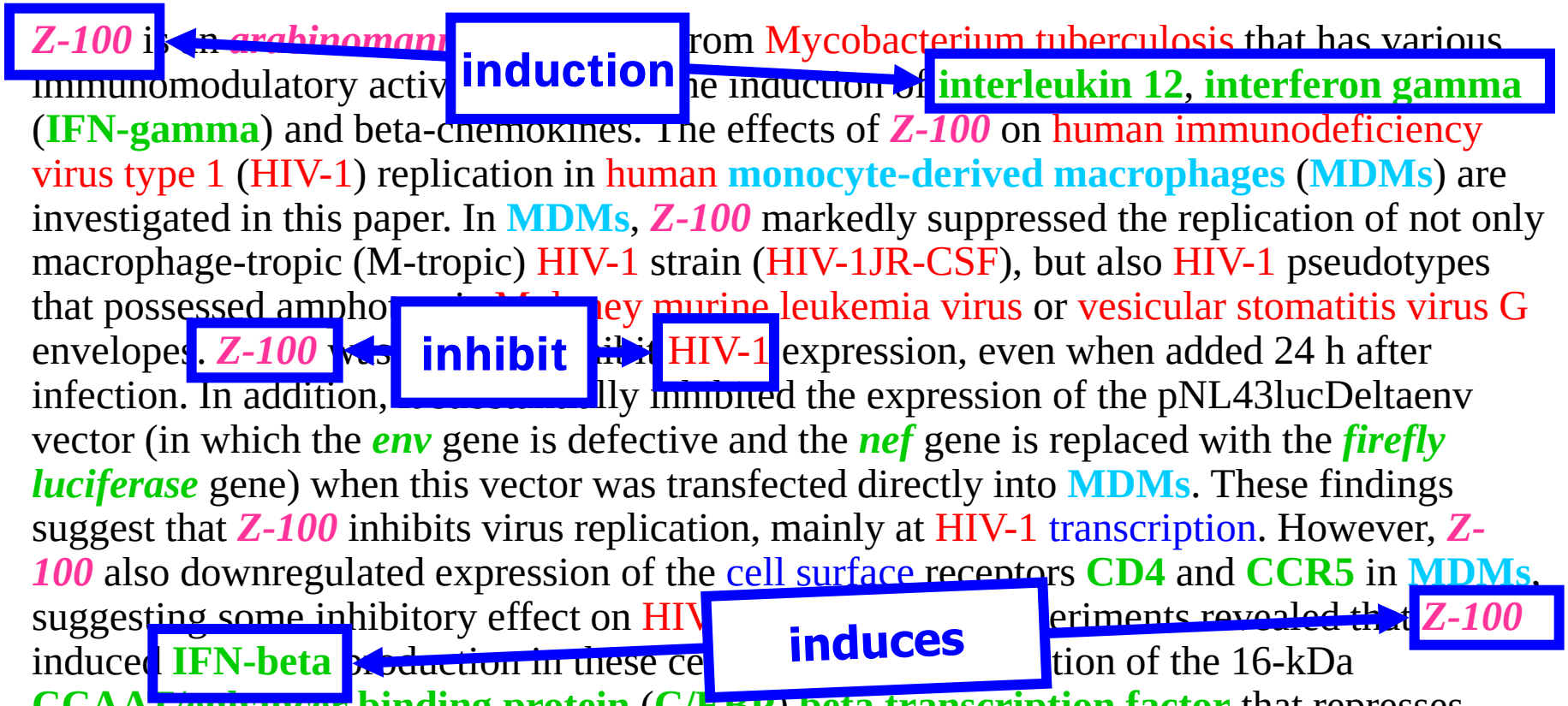
Tax: 9606

UMLS: C0001175

GO:0009986

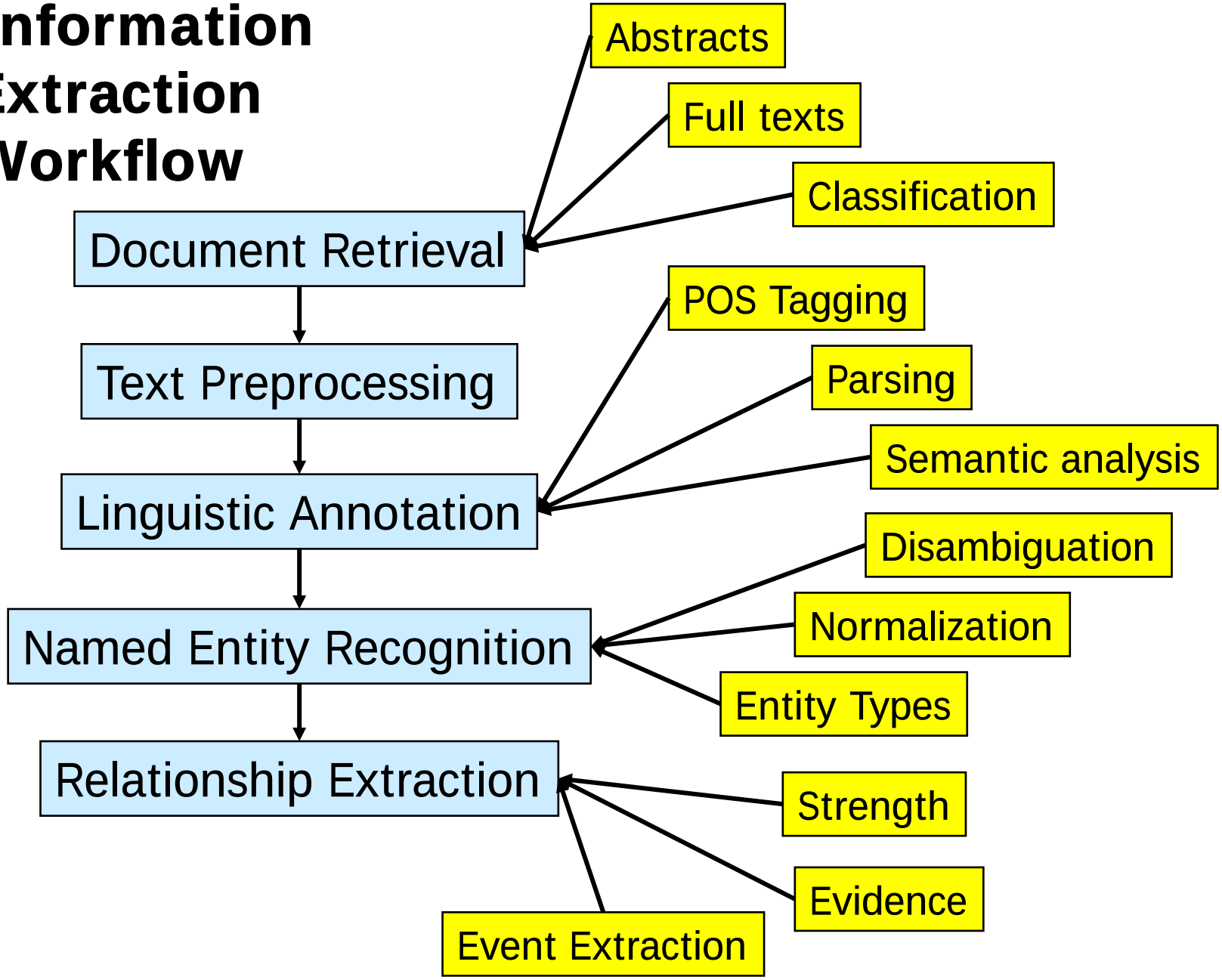
Z-100 is an ***arabinomannan*** ex[tracted from Mycobacterium tuberculosis th]at has various immunomodulatory activities, [ction of **interleukin 12**, **interferon gamma** (**IFN-gamma**) and beta-chemokin[e. The effects of ***Z-100*** on [y virus type 1 (HIV-1) replication in human **monocyte-derived** [are investigated in this paper. In **MDMs**, ***Z-100*** markedly suppressed the [p]lication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed amphotropic Moloney murine leukemia virus or vesicular stomatitis virus G envelopes. ***Z-100*** was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the ***env*** gene is defectiv[e] [a]ced with the ***firefly luciferase*** gene) when this vector was tr[ansduced into] [M]DMs. These findings suggest that ***Z-100*** inhibits virus replication, ma[i]nly at HIV-1 transcription. However, ***Z-100*** also downregulated expression of the cell surface receptors **CD4** and **CCR5** in **MDMs**, suggesting some inhibitory effect on HIV-1 entry. Further experiments revealed that ***Z-100*** induced **IFN-beta** production in these cells, resulting in induction of the 16-kDa **CCAAT/enhancer binding protein** (**C/EBP**) **beta transcription factor** that represses HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of **p38 mitogen-activated protein kinases** (**MAPK**), indicating that the **p38 MAPK** signalling pathway was involved in ***Z-100***-induced repression of HIV-1 replication in **MDMs**. These findings suggest that ***Z-100*** might be a useful immunomodulator for control of HIV-1 infection.

Find Relationships



Z-100 is [illegible] from Mycobacterium tuberculosis that has various immunomodulatory activ[illegible]e induction of **interleukin 12**, **interferon gamma** (**IFN-gamma**) and beta-chemokines. The effects of ***Z-100*** on human immunodeficiency virus type 1 (HIV-1) replication in human **monocyte-derived macrophages** (**MDMs**) are investigated in this paper. In **MDMs**, ***Z-100*** markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed ampho[illegible]ney murine leukemia virus or vesicular stomatitis virus G envelopes. ***Z-100*** was [illegible]ibit HIV-1 expression, even when added 24 h after infection. In addition, [illegible]ally inhibited the expression of the pNL43lucDeltaenv vector (in which the ***env*** gene is defective and the ***nef*** gene is replaced with the ***firefly luciferase*** gene) when this vector was transfected directly into **MDMs**. These findings suggest that ***Z-100*** inhibits virus replication, mainly at HIV-1 transcription. However, ***Z-100*** also downregulated expression of the cell surface receptors **CD4** and **CCR5** in **MDMs**, suggesting some inhibitory effect on HIV [illegible]eriments revealed that ***Z-100*** induced **IFN-beta** [illegible]bduction in these ce[illegible]tion of the 16-kDa **CCAAT/enhancer binding protein** (**C/EBP**) **beta transcription factor** that represses HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of **p38 mitogen-activated protein kinases** (**MAPK**), indicating that the **p38 MAPK** signalling pathway was involved in ***Z-100***-induced repression of HIV-1 replication in **MDMs**. These findings suggest that ***Z-100*** might be a useful immunomodulator for control of HIV-1 infection.

Information Extraction Workflow



Why Text-Mine Scientific Articles?

- Search
  Let users find specific information faster
- Curation
  Support construction of high quality knowledge bases
- Biomedical Research
  Provide background information for specific types of biomedical analysis (network / systems biology)

Curation: A DB of Human TF-TF relationships
[Thomas et al., Bioinformatics, 2016]

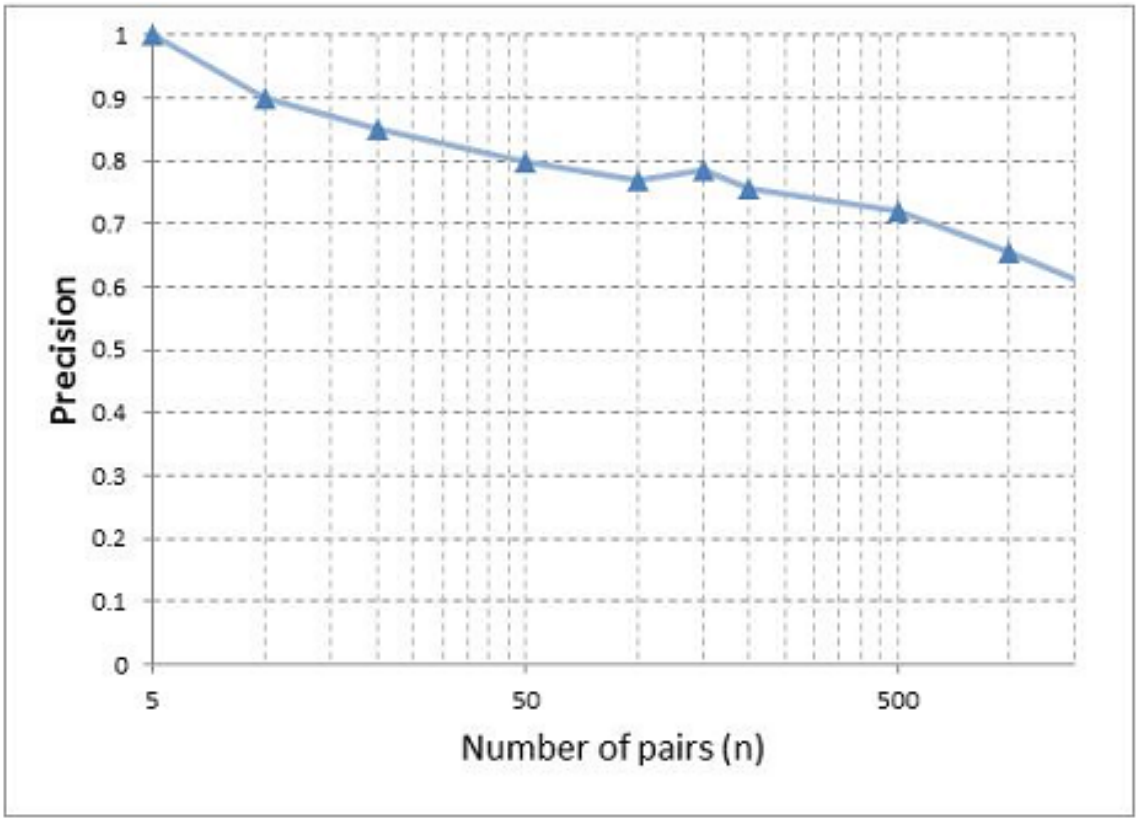
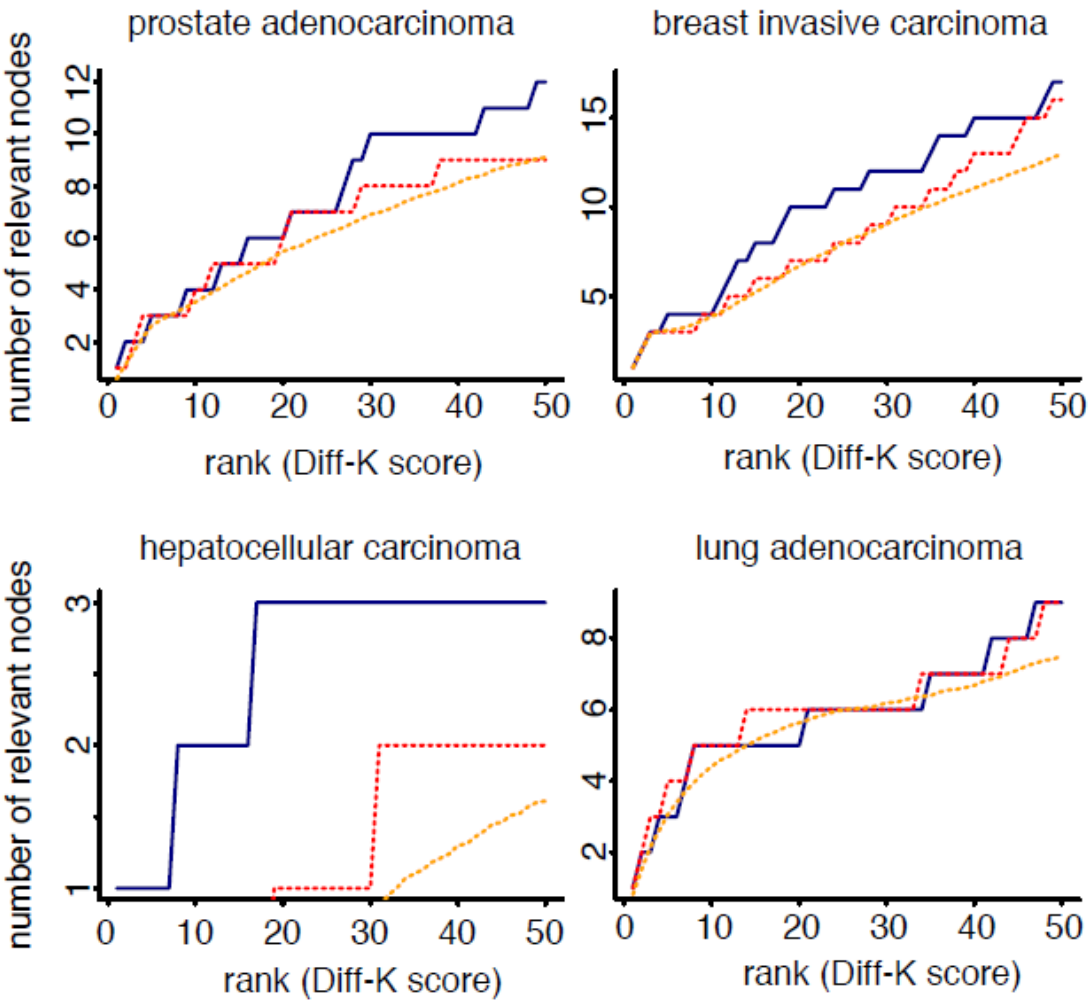


Fig. 4. Precision of our workflow for the n most confidently classified and manually curated sentences. Pairs already contained in a regulatory database are ignored (see Table 3).

Biomedical Research: Biomarker Ranking

[Thomas et al., Bioinformatics, 2016]



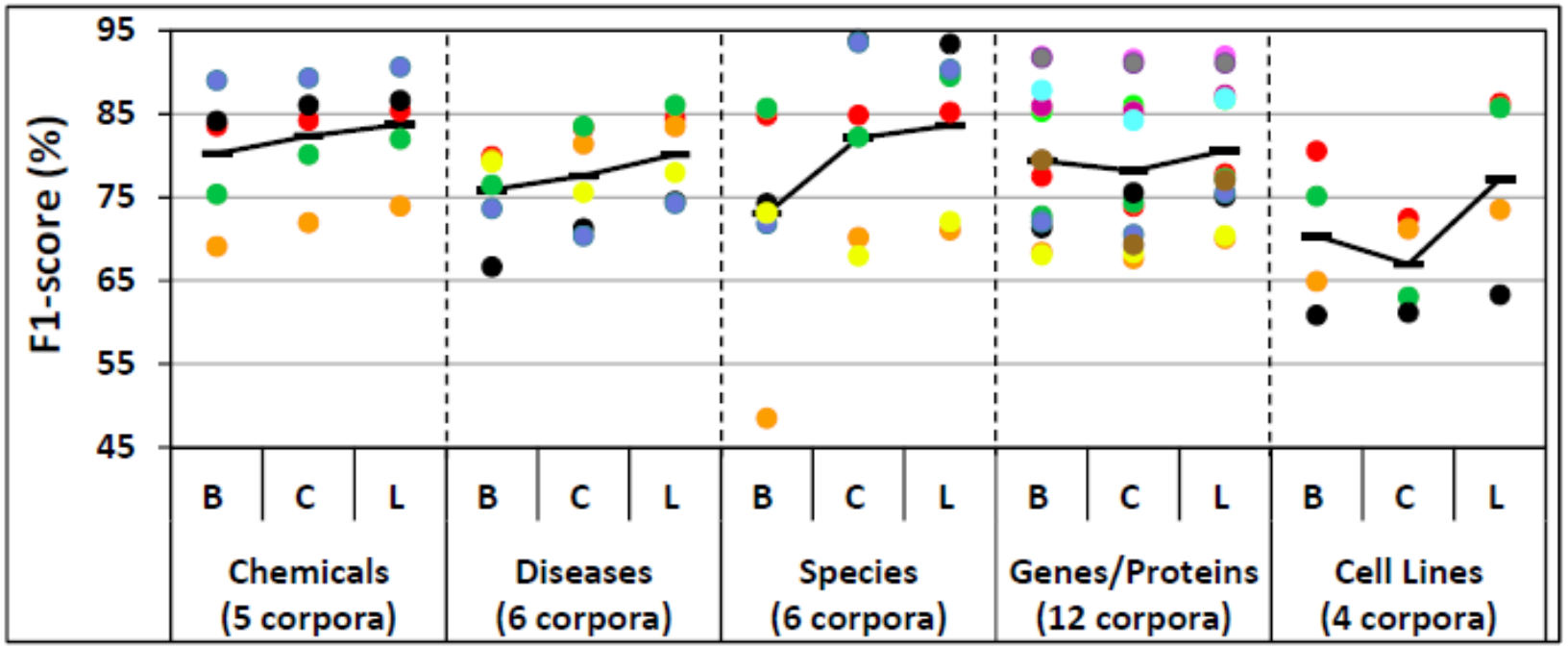
- With TM_REG
- Current network
- Randomized

Table of Content

- Biomedical Text Mining
- Quality in BTM of Scientific Articles
- Quality in BTM Medical documentation

Quality?
[Habibi et al., ISMB, 2017]

- Typically measured by comparison of results to a gold standard corpus



But …

- Methods are trained on one corpus – cross-corpus results?
- NER does not include normalization – impact?
- Methods evaluated on abstracts – full texts or patents?
- Relationships require >1 entity
- In complex tasks (chemicals), IAA is often low

- Entities: ~80%-90% for common types on abstracts
- Relationships: ~40-60%
- Plethora of corpora and open source tools available

Data Quality

- Definition: „Fitness for use“, „Suitability for a purpose“
- “Fitness” in text mining: Correctness / completeness
  - Precision / recall / F1
  - Sensitivity / specificity

- Classical Information Extraction: Extract certain correct and complete information from single sentences
- Different: Extract certain correct and complete knowledge from the literature

High Accuracy BTM?

We have shown that in the knock-out mice strain RX32 expression of FOX3 is positively correlated to downregulation of STX, which hints towards a regulatory relationship between STX and FOX3.

This data shows that in rats after application of normal dosage of XYZ expression of STX decreased, while expression of STX increased, providing evidence that STX is not an activator of FOX3

As shown by others, STX is a regulator of FOX3 ...

Based on simulations on our MAP-ERK pathway model, we conclude that a direct relationship between STX and FOX3 does not exist, but ...

Limits of current BTM

- Statement extraction
- Fact or guess?
- Scientific evidence?
- Context-dependent?
- What did others find?

Focus of current BTM

Under-researched
Rely on human wisdom

Under-researched
Aggregation / integration

Summary

- Quality of Text Mining in scientific articles
  - Recognition of basic entities 
  - Normalization of basic entities 
  - Extraction of general relationships 
  - Extraction of specific relationships 
  - Fact assessment and aggregation 
  - Down-stream usage in critical applications 

Table of Content

- Biomedical Text Mining
- Quality in BTM of Scientific Articles
- Quality in BTM Medical documentation

Reasons for Mining Medical Documentation

- Consistency of structured data (DRG codes) and textual documentation
- Curation: Improving capture of structured data (DRG)
- Coherence of real treatment paths to planned ones
- Extraction of best practices
- Consistency of treatments within clinic, across time
- Medical findings: Adverse drug effects, DDI, ...
- Case-based treatment (similar patients)
- ...

Comparison to Mining Scientific Articles

- NER tasks: Medication, diseases, therapies, anatomy, ...
  - Less relationship extraction: Centered around patient
- Specific problems
  - Temporal alignment of findings: Filtering and aggregation
    - Especially in discharge summary
  - Non-grammatical sentences: Special NLP methods necessary
  - Many negations and coordinated expression: Special treatment
  - Idiosyncratic terms / abbreviations: Adaptation and configuration
    - Differences between clinics, regions, countries

State-of-the-Art

	English	German
Corpora available?	Some large (MIMIC III)	
Freely available tools?	Some (cTakes, Metamap, ...)	
Basic NLP tools freely available?	Many (openNLP, NLTK, ...)	
Comprehensive ontologies?	Many (SNOMED, UMLS, ...)	
Clear rules for data access and privacy?	Guess so (HIPPA)	

State-of-the-Art

	English	German
Corpora available?	Some large (MIMIC III)	None
Freely available tools?	Some (cTakes, Metamap, ...)	Almost none
Basic NLP tools freely available?	Many (openNLP, NLTK, ...)	Few (Treetagger, JCore)
Comprehensive ontologies?	Many (SNOMED, UMLS, ...)	Few (incomplete translations)
Clear rules for data access and privacy?	Guess so (HIPPA)	Great uncertainty

Summary

- Quality of Text Mining in German medical documentation
  - Recognition of basic entities 
  - Normalization of basic entities 
  - Temporal alignment 
  - Fact assessment and filtering 
  - Down-stream usage in critical applications 

Die WorkshopteilnehmerInnen führen die Schwächen der deutschen Forschung zu medizinischem Text Mining vor allem auf die strengen und zwischen den Bundesländern uneinheitlichen Datenschutzbeschränkungen zurück. Dies führt zu fehlendem Austausch, einem Ausschluss von Forschern außerhalb der Kliniken und Nichtvergleichbarkeit und damit fehlenden Wettbewerb bei Einzellösungen. Dadurch fehlen Success Stories, wie sie in anderen Ländern existieren, was wiederum zu fehlender Finanzierung in den Kliniken und bei Mittelgebern führt. Als weitere Folge sind dazu allgemein Text Mining Technologien in der Medizin unterentwickelt, was sich auch in unzureichenden deutschen medizinischen Terminologien niederschlägt.

Open Issues

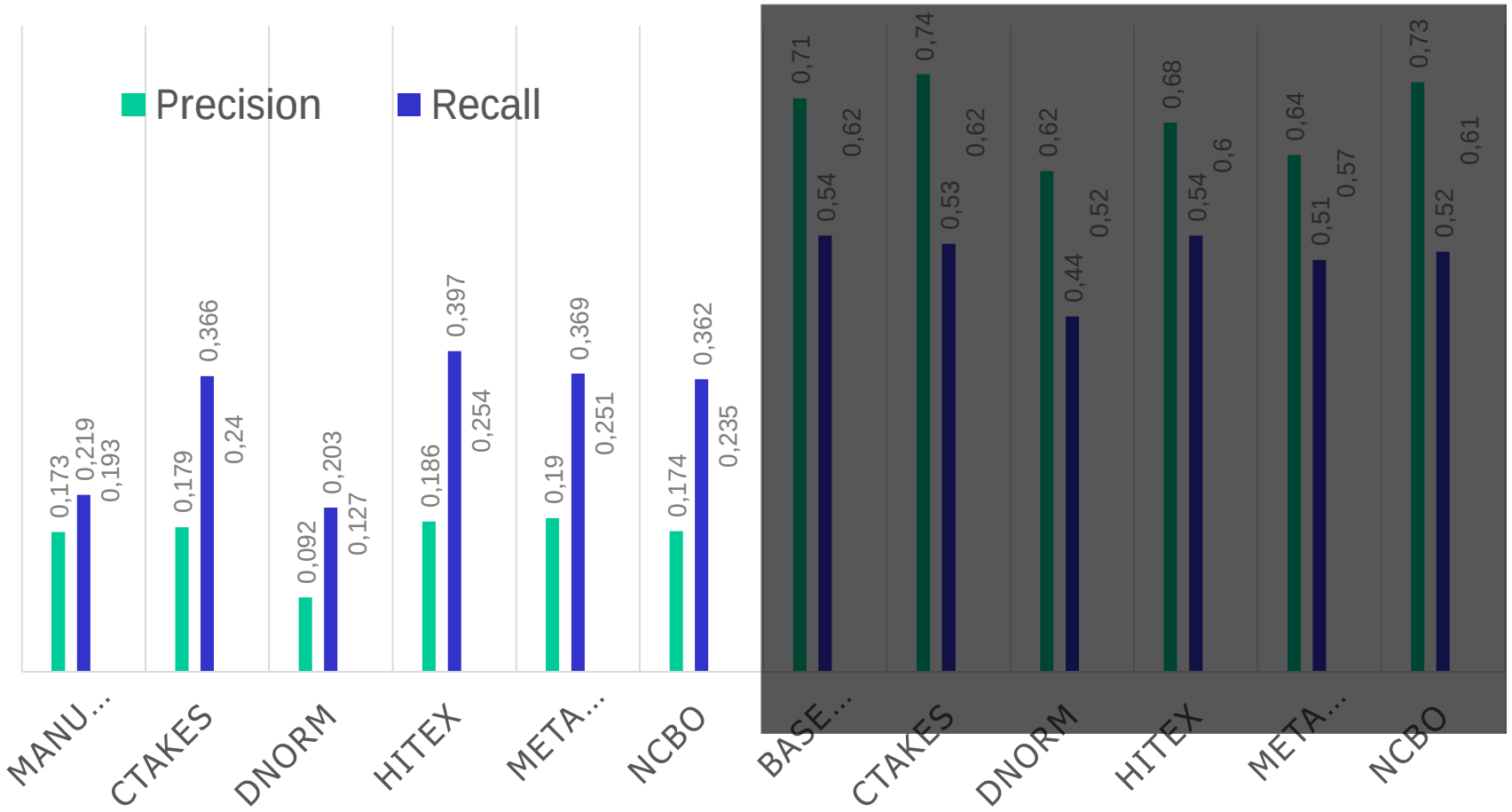
- What quality will we reach in automatically extracting medication / therapy / diagnosis?

- Will we be able to share our corpus?
- Will others be able to confirm our results?
- Will others be able to improve our results?
- Will we be able to confirm our results using other docs?
- Will we be able to co-operate with other groups?
- Will Charité be able to use our tools?

Fitness for Use (MIMIC-III)

[J. Bräuer, 2017]

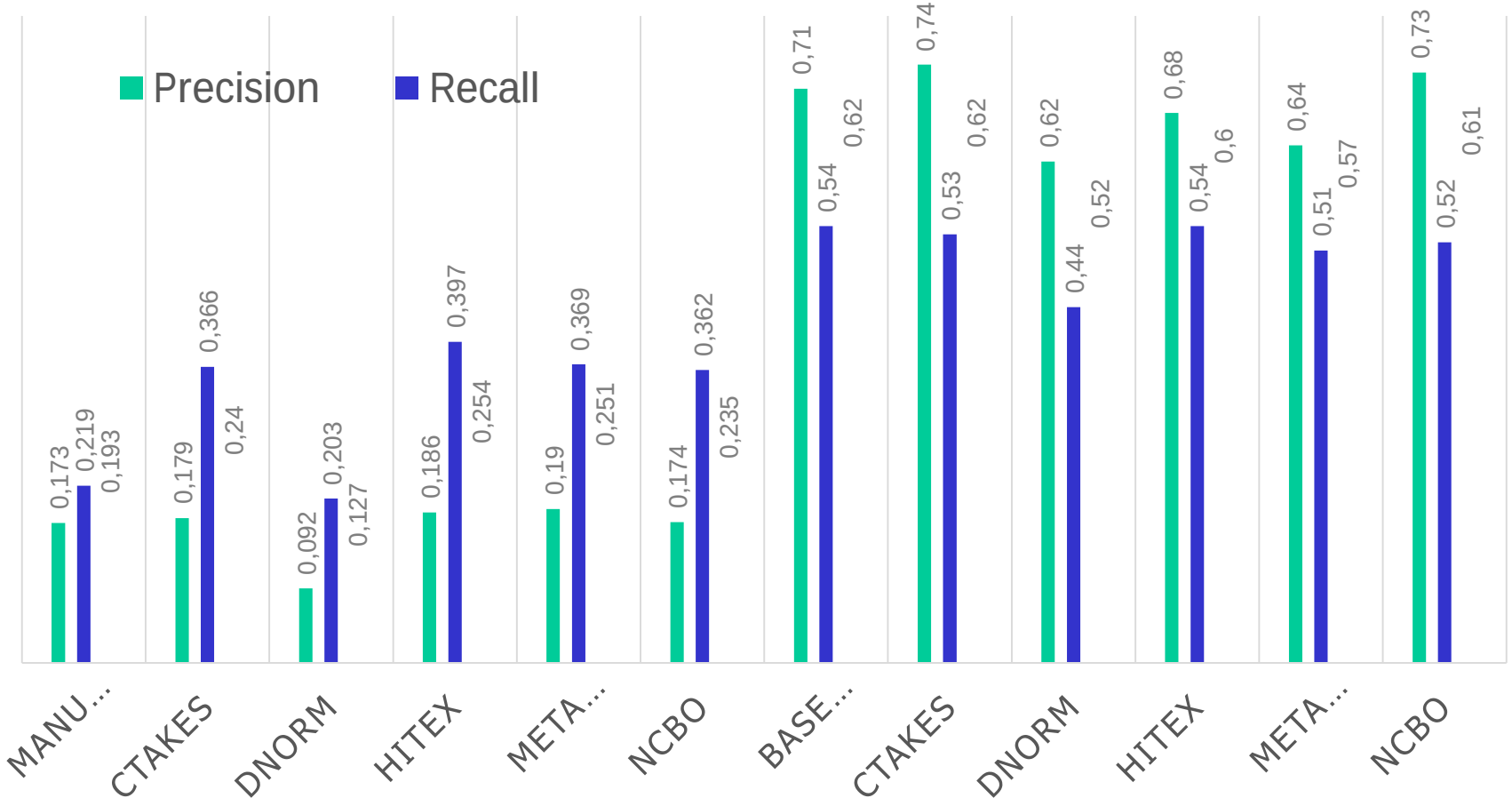
- 50 k discharge summaries
- 7 k diagnosis codes



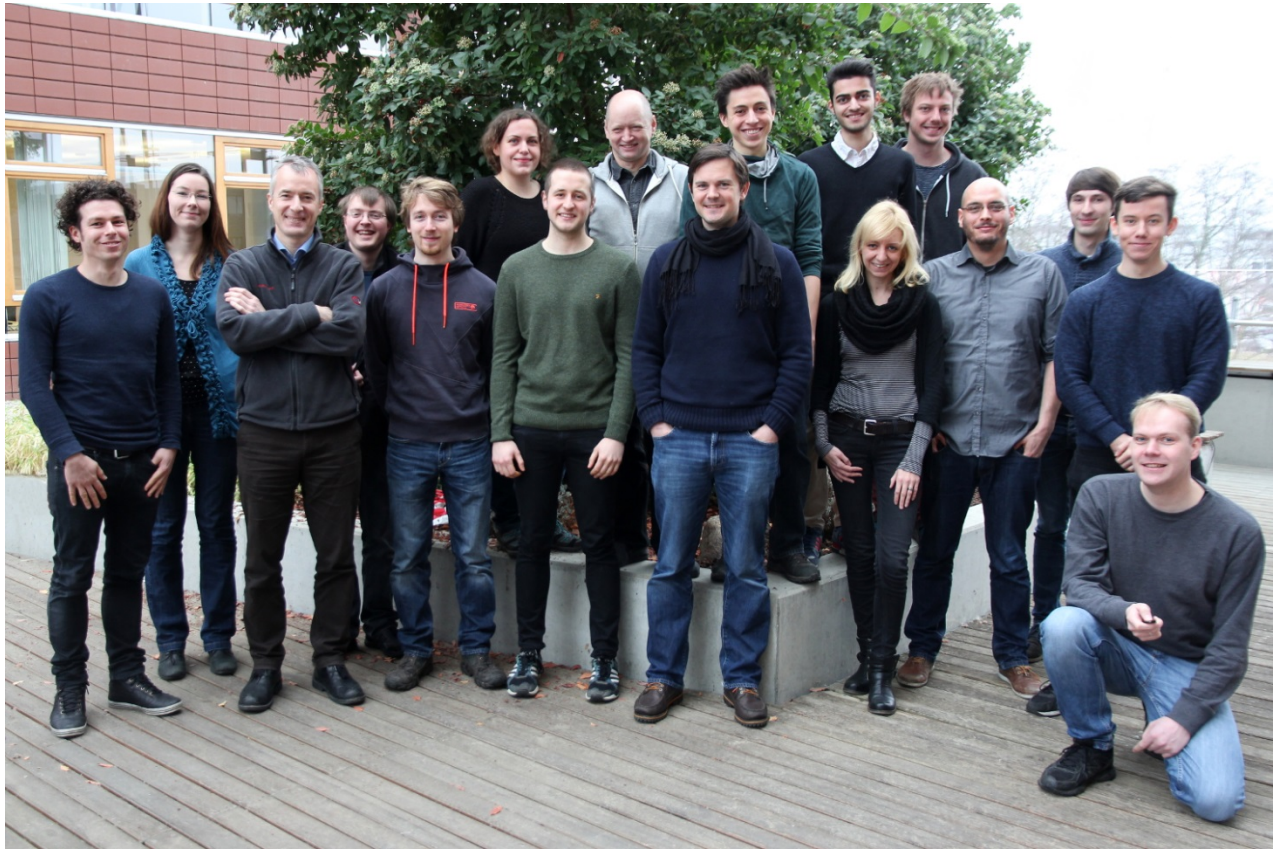
Fitness for Use (MIMIC-III)

[J. Bräuer, 2017]

- 50 k discharge summaries
- 7 k diagnosis codes



Acknowledgements



Deutsche Forschungsgemeinschaft DFG

Alexander von Humboldt Stiftung/Foundation

HUMBOLDT-UNIVERSITÄT ZU BERLIN

DAAD
Deutscher Akademischer Austausch Dienst
German Academic Exchange Service

Bundesministerium für Bildung und Forschung

Bundesministerium für Wirtschaft und Technologie

SEVENTH FRAMEWORK PROGRAMME