Digital Biomarkers for Precision Medicine DBM4PM

- **Project 1.20**
- **Stefan Schulz, Markus Kreuzthaler**

Basel: Aug 29, 2019

Partners from Graz, Austria

CBmed

- Competence center for systematic medical biomarker research, brings together scientific experts with leading pharmaceutical, diagnostic, medical-technology and IT industry partners. Established 2014
- CBmed research projects will identify new biomarkers, validate potential biomarkers and conduct translational biomarker research
- Focus: Cancer, cardio-metabolic health

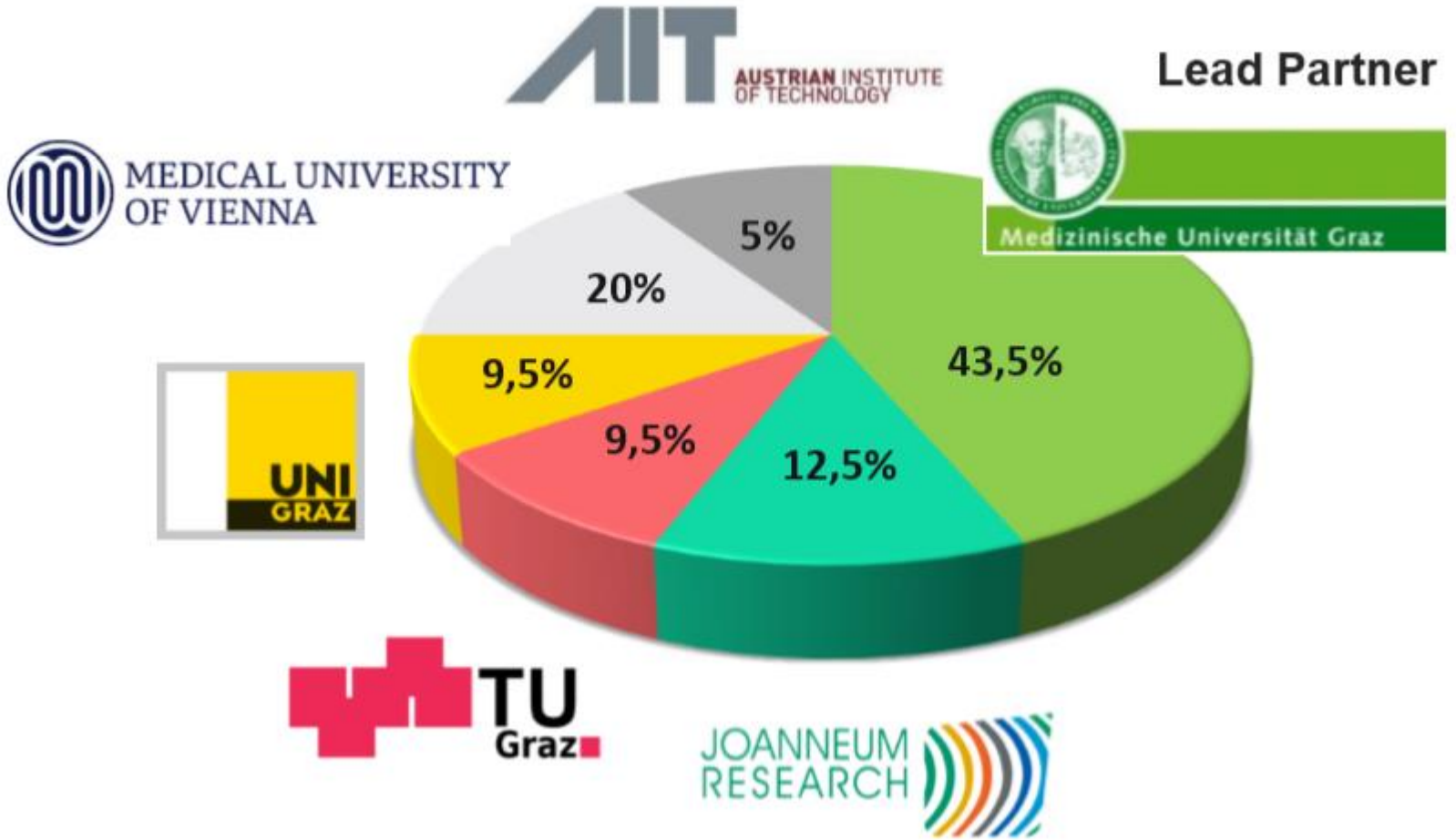
Medical University of Graz

- Public Medical School and research centre for Life Sciences
  - IMI – Institute for Medical Informatics, Statistics and Documentation

KAGes

- Public Hospital Network, covering most of in-patient care in the Federal state of Styria (1.2M inhabitants)

Shareholders / Main Scientific Partners



Scientific Consortium Members

Columbia University Medical Center
DLR
Aarhus University
名古屋大学 Nagoya University
Georg-August-Universität Göttingen
School of Medicine Temple University
Humboldt-Universität zu Berlin
Hôpitaux Universitaires Genève
VUmc
Otto von Guericke Universität Magdeburg
vetmeduni vienna
research center pharmaceutical engineering

Main industry partners

Pharma Industry

Diagnostic Industry

Biotech Industry

Information Technology

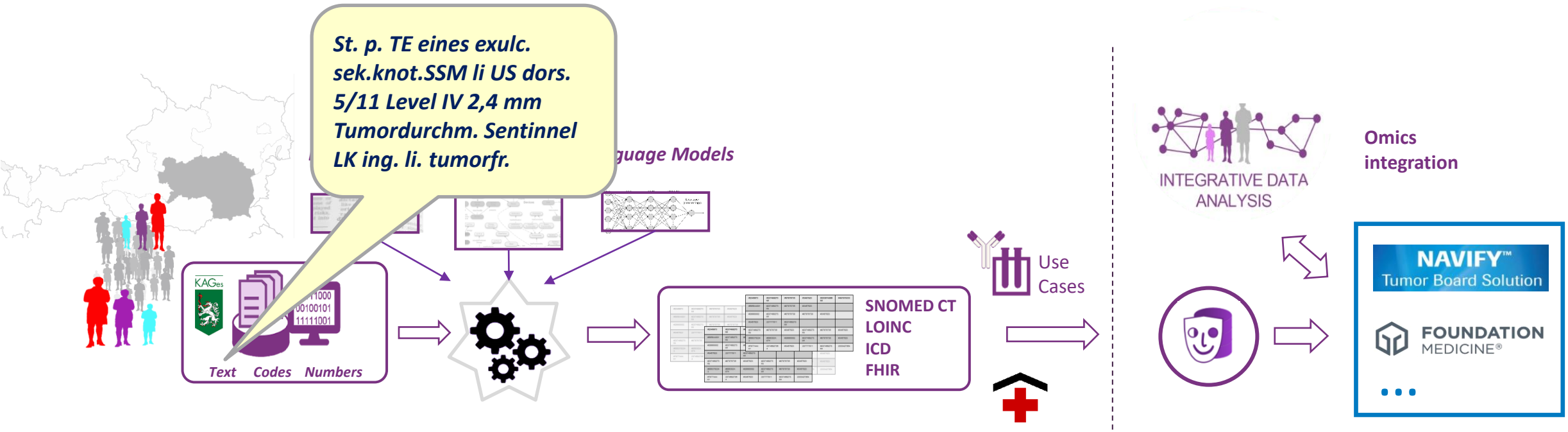
Nutrition & Probiotics

> 50 international partners on 4 continents

Digital Biomarkers for Precision Medicine (DBM4PM)

- The primary goal of DBM4PM is to make clinical real-world data in hospitals available for a clinical and research use cases related to CBmed
- This requires extracting semantically explicit information from highly heterogeneous RWD sources, mostly low-structured
- “Digital biomarkers” – in our understanding: RWD (extracts) of various levels of structure and validity characterising patient phenotypes and health care processes and supporting the above use cases
- This requires the use and the construction of human language and semantic technologies and resources
- In particular, in the co-operation with Roche, the NAVIFY use case will aim at processing data extracts from patient records to support Tumour Board processes

DBM4PM - Overview



Real-Word-Data from Electronic Patient records

Human Language Technologies (NLP)

De-identified, semantically standardised interoperable ("FAIR") data extracts

Anonymization & Import (AIM)

Tumor Board

Core Semantic Technologies

- **RESOURCES:**
  Terminologies / Ontologies / Information models
  - Management (enhancement, mapping, localisation -> German)
  - Representations, reasoning, semantic interoperability

- **No official translations** / interface vocabularies for important terminology standards (SNOMED CT, LOINC)

- **METHODS: *Clinical* NLP**
  e.g., term recognition / concept mapping:
  assigning ontology IDs to one or more words within documents
  - Procedure
  - Condition
  - Substance

  Multiple hierarchies in ontology allows for aggregating data along different axes and at different levels

  Context recognition, relation extraction

Core Semantic Technologies

- **METHODS (continued): *Clinical* NLP**
  Information extraction: filling of a template with
  - NER e.g. de-identification of clinical narratives
  - a small set of pre-defined values, e.g.
    - TNM codes
    - Smoking status
    - Radiation yes / no
  - numeric values
    - Tumour volume
    - Date of 1st diagnosis
    - Date of surgery

- **Rule-based**
  Knowledge Engineering Approach
  Top Down
- **Model-based**
  Data Driven Approach
  Bottom Up
- **Terminology resources**
  Term matching

- **No „open“ annotated German clinical corpora** like for English (MIMIC III, n2c2, TREC challenges)

Named entity recognition

Example: De-identification

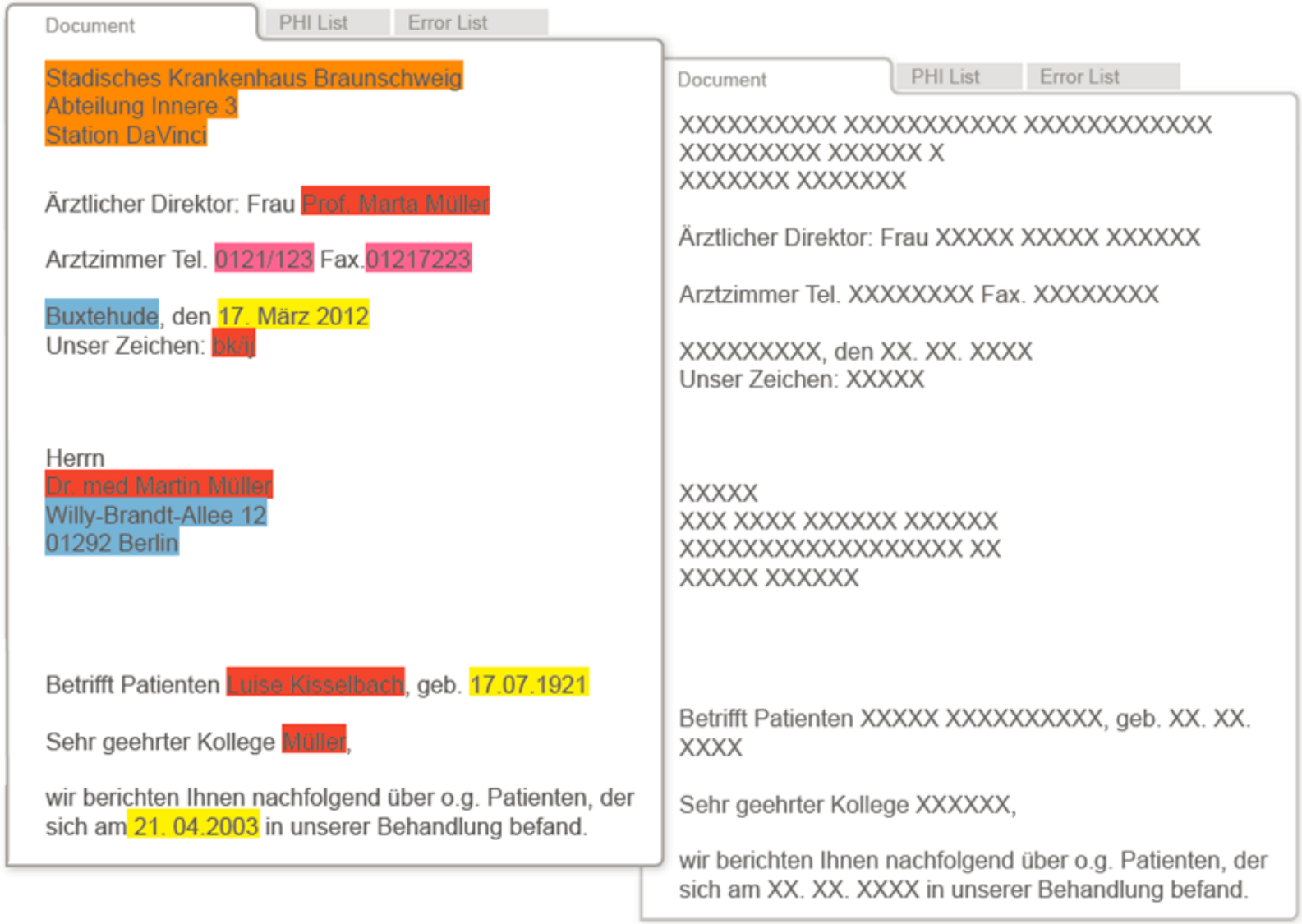
deID
averbis deidentification

Metadata

Name Lists

Patterns

Machine Learning



Smoking Status

Examples

- strikte Alkohol - und Nikotinabstinenz
- strikte Nikotinkarenz!
- strikte Nikotinkarenz.

(Nieraucher) und diskreter Erhöhung der Eosinophilen zu Beginn, empfehle ich - ergänzend zur bereits
(seitdem Ulcusantherapie). VHFA, I48 arterieller Hypertonus, I10 St.p. chron. Nikotinabusus, F17.1 ICMP,
10 % bei bekannter Amaurose, bekannte Lebermetastasen bei Colon-CA. Nikotin negativ. Alkohol negativ.
10 bis 15 Zig. pro Tag. Miktion: Kontinenz, Dranginkontinenz, sonst unauffällig. Letzte Gyn-Untersuchung
1x wöchentlich, Nikotin - Rauchstopp vor einem Jahr, davor 10 py. Caput/Collum: unauffällig,
2. Optimierung der kardiovaskulären Risikofaktoren. Eine strikte Nikotinkarenz ist dringend empfohlen.
2kg während des letzten Monats. Nikotinanamnese leer, auch keine Allergie bekannt, Alkoholkonsum von 1
3) Nikotinkarenz!
4. Strikte Nikotinkarenz!
4. Strikte Nikotinkarenz.
65jährige Pat. in reduziertem AEZ, Inappetenz und Obstipation. Harn unauff. Nikotin-/Alkoholanamnese negativ,
7. Chron. Nikotinabusus
87-jähriger Pat. in gutem AZ u. normalem EZ, Nikotin- u. Alkoholanamnese negativ.
abgenommen), Ex-Nikotinabusus (Ex seit 21 Jahren) mit insgesamt etwa 35 py, keine Dyspnoe oder AP, keine
absolute Nikotinkarenz
Adipositas, chron. Leberparenchymschaden, Hyperlipidämie, Nikotinabusus, Z.n.
Alkohol gelegentl. Nikotin wird neg. Allergien keine bekannt.
Alkohol negativ. Nikotin negativ. Allergie negativ.
Alkohol regelmäßig, kein Nikotin, keine Drogen. Caput/Collum: unauffällig. Pulmo: Vesikuläratmung bds. Cor:
Alkohol und Nikotin: negiert.
Alkohol- und Nikotinabusus werden verneint.
Alkohol, Nikotin: negiert.
Alkohol/Nikotin negativ.
Alkohol/Nikotin: neg.
Alkohol/Nikotin: werden negiert.

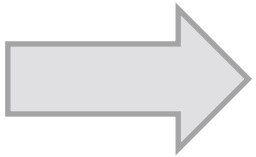
Smoking Status

Annotation example

Nikotin und Alkohol: verneint.

- BA a.c.e.t.Smoking [1]
  - BA Nikotin
    - begin: 437
    - end: 444
    - sctid: 365981007
    - preferred: Finding of tobacco smoking behavior (finding)
    - BA qualifier: verneint
      - begin: 458
      - end: 466
      - value: negative
      - sctid: 260385009
      - preferred: Negative (qualifier value)

NLP pipeline output



Clinical information model, e.g. based on FHIR resources*

* Philipp Daumke, Kai U. Heitmann, Simone Heckmann, Catalina Martínez-Costa, Stefan Schulz. Clinical Text Mining on FHIR. MEDINFO 2019

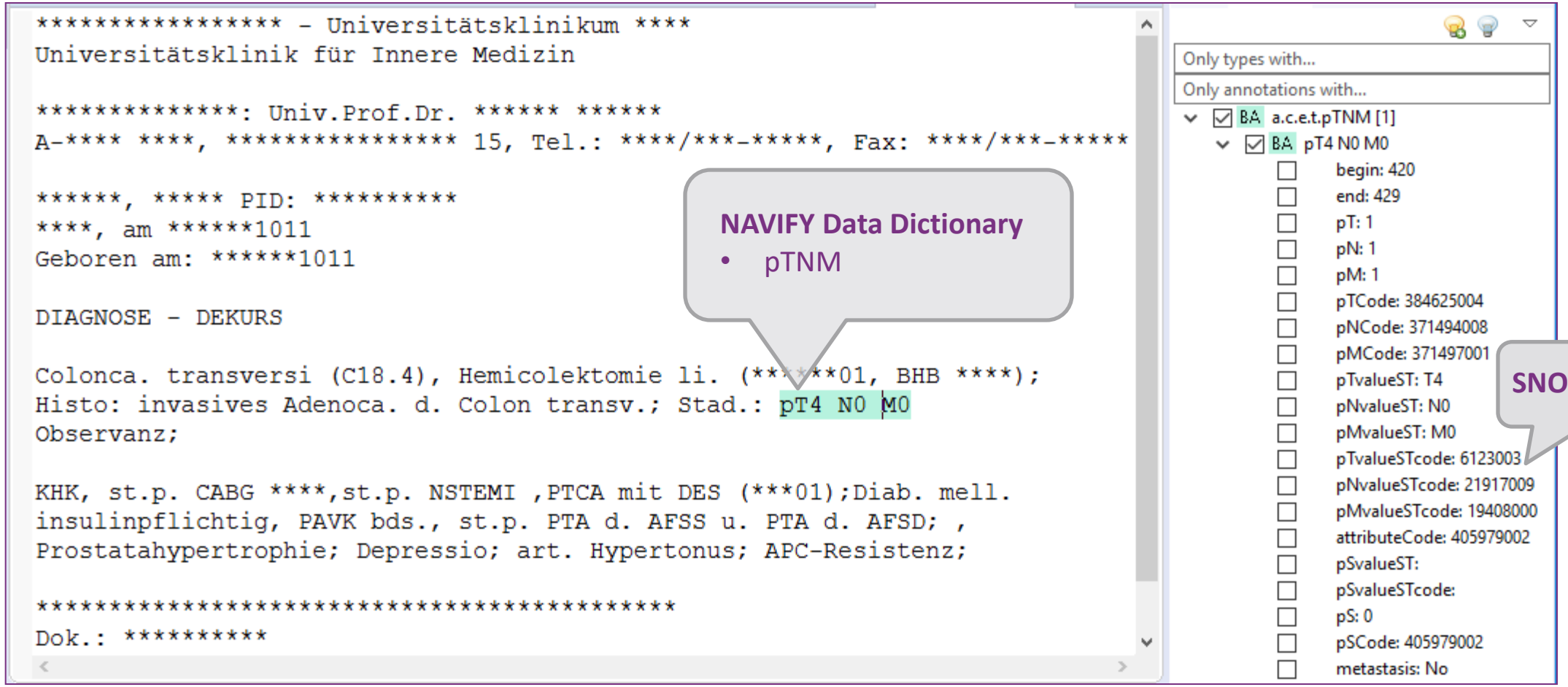
pTNM

Entity examples

Stad.: pT3bN1a(1/15)MXG1
Stad.: pT3N1(1/17)M1G2R0L1 sowie Leberanteil mit AdenoCA-formationen,pM1 (HEP),R0
UICC 2009: G2 pT3 N1(1/17) R0 L1
Stad.: pT3 N1 (1/17)M1 G2 R0 L1 sowie Leberanteil mit AdenoCA-formationen,pM1 (HEP),R0
Stad.: pT2N2a(5/17)MXG2R0
Stad.: pT2 N2a (5/17)MX G2 R0
Histo.: Adenokarzinom, G2 Stad.: pT2 pN2a (5/17), R0 M0.
Stad.: pT3N1(1/46)G3R0
AdenokarzinompT3, N1 (1/46),G3, R0.
Stad.: pT2N2a(6/15)MXG2R0
Stad.: pT2 N2a (6/15)MX G2 R0
Stad.: pT3pN0(0/21)G3R0
Klassifikation:G-2;pT-3;pN-0; pM-X.
Stad.: pT4aN1bG2R0
Stad.: pT4a N1b G2 R0
Stad.: pT3N0(O/4)G2R0
Stad.: pT3bN2a(4/33)G2 (Metastase im Grenzlymphknoten A. colica med.) HNPCC-Testung: Lynch-Syndrom
Stad.: pT3b N2a (4/33) G2 (Metastase im Grenzlymphknoten A. colica med.) HNPCC-Testung:
Stad.: pT3bN0(0/14)MXG2R0L1
Stad.: pM1R0 M1=fil.hep.
Stad.: pT3cN1b(2/12)MXG3R0
Stad.: pT3N0(0/14)G2R0
Stad.: pM1 M1=fil.hep.
UICC 2009: G2 pT3 N0
Stad.: pT3 N0 (0/14) G2 R0
Stad.: pT4aN1a(1/19)MXG1R0 +hyperplastischer Colonpolyp
Stad.: pT4aN0(0/15)G3L1V0

pTNM

Automatic annotation in the narrative



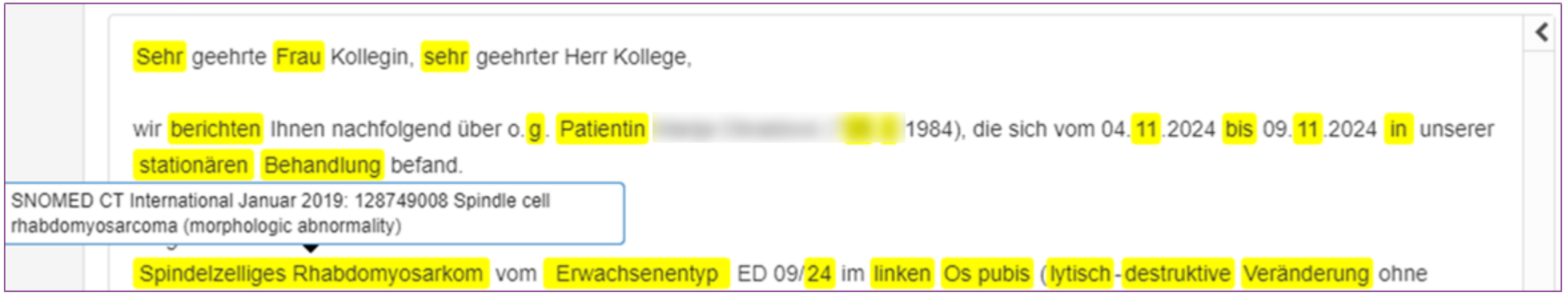
Resources

Terminologies

- **Bottleneck: German language**
  - SNOMED CT introduced in Austria, but no official German translation
- **SNOMED CT Interface Terminology: work in progress at MUG-IMI**
  - Combination of automated and manual term translation
  - Currently 5 Million term candidates, scored

363410008	0.071	482620011_000001	Malignant tumor of sigmoid colon	bösartiger Tumor des Colon sigmoideum
363410008	0.069	482620011_000002	Malignant tumor of sigmoid colon	Malignom des Colon sigmoideum
363410008	0.062	3288335015_000006	Sigmoid colon cancer	Sigmakarzinom
363410008	0.056	3288335015_000002	Sigmoid colon cancer	Colon sigmoideum Karzinom
363410008	0.056	3288335015_000001	Sigmoid colon cancer	Colon sigmoideum Krebs
363410008	0.031	3288335015_000004	Sigmoid colon cancer	Sigmoidkarzinom
363410008	0.031	3288335015_000005	Sigmoid colon cancer	Sigmakrebs
363410008	0.000	3288335015_000003	Sigmoid colon cancer	Sigmoidkrebs

Concept mapping
SNOMED CT



Resources

Existing corpora / gold standards, currently available

- **De-identified annotated corpora**
  - Clinical problem lists annotated with ICD codes (~ 1,7 M unique entries)
  - ~ 3500 reports annotated with inflammation / neoplasm (pathology)
  - ~ 400 discharge summaries annotated with medication information (dermatology)
- **Non-annotated corpora**
  - KAGes non-de-identified texts (millions)
  - De-identified KAGes dumps for experiments
    - ~ 32000 discharge summaries (cardiology)
    - ~ 1300 discharge summaries (dermatology)
    - ~ 13000 discharge summaries (colon cancer)
  - Wikipedia extracts

Example: machine learning resource acquisition

- Resource: millions of short problem list entries from HIS, partly annotated with administrative codes (ICD-10)

SNOMED CT
The global language of healthcare

map

N04.0	Glomerulopathie mit Minimalveränderung
E11.9	Diab. mell. Typ II - OAD (aktueller HbA1c 58 mmol/
G93.0	Arachnoidalzyste
I25.0	KHK III, Z. n. CTR bei cardiopulmonaler Reanimatio
R31	Denovo Proteinurie und Hämaturie zur Abklärung -
	Soor genital
R99	Sonstige ungenau oder nicht näher bezeichnete Tode
K21.9	Refluxösophagitis III°

map

- German ICD terms and synonyms
- English ICD terms and synonyms
- Wikipedia pages (via ICD codes)

Machine learning:
- Lexicon acquisition
- Word sense disambiguation

NLP Pipeline

Clinical Texts

Reader	Preprocessing	Information Extraction	Writer
File Readers	Document Zoning	Diagnosis	**File Consumer**
DB Readers	Sentence Detection	Medication	Search Engine
…	Tokenizing	SNOMED CT	I2B2/tranSMART
	POS-Tagging	Breslow Level	Tableau, Pentaho, Spotfire, Qlikview
	Chunking	Clark Level	**NAVIFY**
	Stemming	pTNM	
	Decompounding	Smoking Status	
	…	…	

NAVIFY Data Dictionary

Unstructured Information Management Architecture
An Apache Project

averbis text analytics

HEALTH discovery

SNOMED CT
LOINC
ICD
FHIR

HL7® FHIR®

OMOP

- De-identified, semantically standardised interoperable (“FAIR”) data extracts
- Open Clinical Information MODELS (FHIR)
- Open data warehouses (OMOP)

Evaluation

- **Integration aspects**
  - ***NAVIFY***
  - On-premise
  - Cloud-based (AWS)
  - Clinical NLP as a service
- **Real time constraints**
  - Documents / minute
- **Scale out**
  - Apache Spark (Amazon EMR)
- **NLP quality**
  - Gold standards
  - F-measure
- **Terminologies / Ontologies**

- **Internal reports / Steering committee**
- **Publications**
- **Quality assured prototypes**
- **Science vs. Engineering**
- **General show cases**
- **NAVIFY show cases**

Deliverables

Overview

WP-0 Project Management

Deliverable	Description	Month
D-0.1	Internal management structure described, steering board established	M09
D-0.2	Kick-Off Workshop (Roche, MUG, CBmed, KAGes)	M09
D-0.3	Licence management plan competed (Roche, Averbis, third-party tools, terminology resources)	M10
D-0.4	Project environment (ownCloud, freedcamp, Trello, SharePoint, GitLab, Slack, Google services) and communication strategy set up and functional	M12
D-0.5	Ethics committee proposal approved	M15
D-0.6	Mid-term assessment workshop	M30
D-0.7	Sustainability plan	M42

WP-1 Data Platforms / Management

Deliverable	Description	Month
D-1.1	Secure Data Lake installed and functional	M15
D-1.2	ETL tooling installed and functional	M15
D-1.3	Manual data decoration toolset installed	M15
D-1.4	Concept for scalable clinical data warehouse elaborated	M15
D-1.5	Definition of data dictionary completed	M18
D-1.6	ETL workflows for structured and unstructured data running	M18
D-1.7	NAVIFY staging area loaded	M18
D-1.8	First NAVIFY showcase	M21
D-1.9	Final NAVIFY showcase	M42

Deliverables

Overview

WP-2 Documentation and Communication Ontologies and Standards

Deliverable	Description	Month
D-2.1	Inventory of semantic resource standards	M10
D-2.2	Semantic standards management plan including quality metrics	M12
D-2.3	Identification of FHIR resources needed for the project	M15
D-2.4	Report of standards use in the project and assessment of their fitness	M24
D-2.5	Assessment of the potential of semantic post-processing of coded data	M42
D-2.6	Final Report on use of semantic standards	M48

WP-3 Natural Language Processing

Deliverable	Description	Month
D-3.1	NLP tech stack specification (Spacy, AllenNLP, UIMA, Averbis, Roche)	M07
D-3.2	Installation NLP engine(s) on premise	M08
D-3.3	Inventory of needed semantic extractors	M12
D-3.4	HIPAA based de-identification of clinical narratives implemented	M18
D-3.5	Rule engines adapted and validated	M21
D-3.6	Machine learning models adapted and validated	M24
D-3.7	Complete NLP pipeline adapted and validated	M36
D-3.8	Final gold standard (manual data curation)	M36
D-3.9	Semantic production integration in system environment	M42

Deliverables

Overview

WP-4 Terminology Management

Deliverable	Title	Month
D-4.1	Definition of interface terminology needs and value sets.	M12
D-4.2	Specification of terminology services needed for the project (cloud-based versus on-premise)	M12
D-4.3	Survey of terminology management platforms	M15
D-4.4	Local terminologies and value sets – first release	M21
D-4.5	Terminology services implemented	M27
D-4.6	Terminology management platform implemented and in use	M30
D-4.7	Semi-automatic terminology expansion using large data sets: completed and assessed	M36
D-4.8	Final release of complete set of terminologies and value sets assessed	M45

WP-5 Manual Data Curation

Deliverable	Title	Month
D-5.1	Ethics committee agreement (CBmed) and verify eligibility for project (Roche) is provided	M12
D-5.2	Source data and metadata on patient staging is provided (CBmed)	M18
D-5.3	Target data fields linked to terminology / value sets are provided	M21
D-5.4	Access to EHR system extracts (source data) to manual data curators provided (CBmed)	M27
D-5.5	Curation completed, manual data curation results reported (e.g. availability of fields, clustered by stage, time to curate, time effort)	M42

Financial

a. Half-year breakdown of financial shares of Cooperation Partner Roche Diagnostics

Total costs: € 1.623.067,28		Cost contribution Company Partner Roche		
Half-year	Half-year costs	Cash	Personnel and in-kind services (non-cash)	Total
1st HY 2019	€ 202.883,41	€0,00	€57.952,67	€57.952,67
2ndHY 2019	€ 202.883,41	€300.000,00	€57.952,67	€357.952,67
1st HY 2020	€ 202.883,41	€33.333,33	€79.303,65	€112.636,98
2nd HY 2020	€ 202.883,41	€33.333,33	€79.303,65	€112.636,98
1st HY 2021	€ 202.883,41	€33.333,33	€73.203,37	€106.536,70
2ndHY 2021	€ 202.883,41	€33.333,33	€73.203,37	€106.536,70
1st HY 2022	€ 202.883,41	€33.333,33	€61.002,81	€94.336,14
2ndHY 2022	€ 202.883,41	€33.333,33	€61.002,81	€94.336,14
Total	€ 1.623.067,28	€499.999,98	€542.925,00	€1.042.924,98

b. Cost overview for research project 1.20

Personnel costs CBmed 1)	€	467.513,00
Material costs CBmed including Biobank samples 2)	€	-
Other costs CBmed (investment, travel etc.) 3)	€	78.100,00
Scientific Partner costs (personnel and material costs) 4)	€	241.845,00
In-kind other Company Partner	€	500.000,00
Total project costs	€	1.330.385,02
Plus overhead contribution (22% of total project costs)	€	292.684,26
Total costs	€	1.623.067,28