Mining the electronic health record
Linguistic and ontological challenges

Stefan Schulz, Medical University of Graz, Austria

June 11, 2021
Conflict of Interest Disclosure

- Professor for Medical Informatics, Medical University of Graz, Austria
- Project co-leader CBmed Biomarker Research GmbH, Graz Austria
- Head of Medical Research Averbis GmbH, Freiburg, Germany
“Precision medicine’ has emerged as a computational approach to functionally interpret omics and big data, and facilitate their application to health care provision. In this new era, patients are not segregated by disease, or disease subtype. Instead, the aim is to treat every patient as an individual case, incorporating a range of personalized data including genomic, epigenetic, environmental, lifestyle and medical history”
Data as “Fuel” for precision medicine

Source: CBmed – Center for Biomarker Research in Medicine, Graz, Austria
Clinical data

Phenotype
Environment
Lifestyle
Clinical History

Next Generation Sequencing (NGS)

Digital Pathology
Flow Cytometry
in vivo Imaging

Integrative Data Analysis

FUSION Technology

Proteomics
Metabolomics
MALDI-MS

Where is this data?

Source: CBmed – Center for Biomarker Research in Medicine, Graz, Austria
EHRs
Electronic Health Records

Phenotype
Environment
Lifestyle
Clinical History
What is in EHRs?

- Clinical Information Systems
- Clinical Prediction
- Clinical Decision Support
- Clinical Research Support
- Clinical Quality Assessment
- Enhanced Clinical Data Use

How can it be used for PM?
PM requires precision clinical data
PM requires precision clinical data

• **FAIR data:**  
  Findable, Accessible, Interoperable, Reusable

• **Barriers:**
  - **Technical:** clinical information systems not designed for data export and secondary use
  - **Legal / ethical:** patient consent, de-identification
  - **Structure:** Lack of structured data, unstructured data produced for humans, not for machines
  - **Contexts and provenance:** data generation workflows, data creators, intent, motivation and purposes for data collection
  - **Standardisation:** standards for meaning (ontologies), standards for information collection and exchange templates

**The EHR heat map**

<table>
<thead>
<tr>
<th></th>
<th>Completeness</th>
<th>Correctness</th>
<th>Granularity</th>
<th>Structure</th>
<th>Standardisation</th>
<th>Data Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics, ADT (Admission/Discharge/Transfer)</td>
<td>💲💲</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Administrative Codes (ICD...)</td>
<td>💲</td>
<td>💲</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical Lab</td>
<td>💲</td>
<td>💲</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prescriptions</td>
<td>💲</td>
<td>💲</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problem List</td>
<td>💲</td>
<td>💲</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical Registries</td>
<td>💲</td>
<td>💲</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Findings Reports</td>
<td>💲</td>
<td>💲</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discharge Summaries</td>
<td>💲</td>
<td>💲</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Schulz S.* Clinical Informatics Challenges in Precision Medicine. Pathways to Precision Medicine. To Appear 2021
## The EHR heat map

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Completeness</th>
<th>Correctness</th>
<th>Granularity</th>
<th>Structure</th>
<th>Standardisation</th>
<th>Data Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics, ADT (Admission/Discharge/Transfer)</td>
<td>💲💲</td>
<td>💲</td>
<td>💲</td>
<td>💲</td>
<td>💲</td>
<td>☢</td>
</tr>
<tr>
<td>Administrative Codes (ICD...)</td>
<td>💲</td>
<td>💲</td>
<td>💲</td>
<td>🍒</td>
<td>🍒</td>
<td>☢</td>
</tr>
<tr>
<td>Clinical Lab</td>
<td>💲</td>
<td>💲</td>
<td>🍒</td>
<td>☢</td>
<td>☢</td>
<td>☢</td>
</tr>
<tr>
<td>Prescriptions</td>
<td>💲</td>
<td>💲</td>
<td>🍒</td>
<td>☢</td>
<td>☢</td>
<td>☢</td>
</tr>
<tr>
<td>Problem List</td>
<td>💲</td>
<td>💲</td>
<td>🍒</td>
<td>☢</td>
<td>☢</td>
<td>☢</td>
</tr>
<tr>
<td>Clinical Registries</td>
<td>💲</td>
<td>💲</td>
<td>🍒</td>
<td>☢</td>
<td>☢</td>
<td>☢</td>
</tr>
<tr>
<td>Findings Reports</td>
<td>💲</td>
<td>💲</td>
<td>🍒</td>
<td>☢</td>
<td>☢</td>
<td>☢</td>
</tr>
<tr>
<td>Discharge Summaries</td>
<td>💲</td>
<td>💲</td>
<td>🍒</td>
<td>☢</td>
<td>☢</td>
<td>☢</td>
</tr>
</tbody>
</table>

Schulz S. Clinical Informatics Challenges in Precision Medicine. Pathways to Precision Medicine. To Appear 2021
PM requires precision extraction tools

Structured data in context
PM requires precision extraction tools

• Automated analysis of unstructured data:
  ▪ Images
  ▪ Biosignals
  ▪ Natural language: information extraction by natural language processing:
    large parts of EHR content is free text:
    - Findings reports (radiology, pathology,...)
    - Progress notes
    - Nursing notes
    - Problem lists
    - Discharge summaries and letters
Large parts of information only in free text

St. p. TE eines exulc. sek.knot.SSM li US dors. 5/11 Level IV 2,4 mm Tumordurchm. Sentinnel LK ing. li. tumorfr.

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°
K21.9 ; Refluxösophagitis III°
N17.9 ; protrahiertes akutes Nierenversagen - delayed Graft
N39.0 ; Komplizierter Katheter-assoziierter Harnwegsinfekt
E05.9 ;

Primary Care Physician: Dr Dianna Miller
Referring Physician:
Consulting Physician(s): Dr Gary Marshall - hospitalist
Condition on Discharge: stable
Final Diagnosis: RLL pneumonia, COPD exacerbation, mild CHF, osteoarthritis
Procedures: none

History of Present Illness 72 year old thin white male presented to emergency on 8/1/14 with shortness of breath, weakness and dehydration. Chest X-ray showed right lower lobe infiltrate, ABGs unremarkable.
Pulse ox on RA was 79%.

1) Pneumonia: treated with ceftriaxone and azithromycin iv. Switched to PO after 72 hours.
2) Exacerbation of COPD: patient treated with inhaled and oral steroids, O2 at 2L/nc. On RA at time of discharge
3) Weakness and dehydration: secondary to pneumonia and COPD. Responded well to strengthening with PT and regular meals.

Discharge Medications Zithromycin daily until gone, inhalers # of puffs.

Discharge Instructions: no activity restriction, regular diet, follow up in two to three weeks
Natural language processing (NLP)

Source data (text)

St. p. TE eines exulc. sek.knot.SSM li US dors. 5/11 Level IV 2,4 mm Tumordurchm. Sentin nel LK ing. li. tumorfr.

Semantic Resources

Ontologies
Terminologies

ML Models
Rules
Reference Corpora
St. p. TE eines exulc. sek.knot.SSM li US dors. 5/11 Level IV 2,4 mm Tumordurchm. Sentinzel LK ing. li. tumorfr.

<table>
<thead>
<tr>
<th>Code (SNOMED CT)</th>
<th>Value</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>2547300000</td>
<td>Superficial spreading malignant melanoma of skin</td>
<td>History of</td>
</tr>
<tr>
<td>301889008</td>
<td>Excision of malignant skin tumour</td>
<td>History of</td>
</tr>
<tr>
<td>47224004</td>
<td>Skin of posterior surface of lower leg</td>
<td>Current</td>
</tr>
<tr>
<td>7771000</td>
<td>Left</td>
<td>Current</td>
</tr>
<tr>
<td>81827009</td>
<td>Diameter</td>
<td>2.4</td>
</tr>
<tr>
<td>258673006</td>
<td>Millimetre</td>
<td>Current</td>
</tr>
<tr>
<td>94339008</td>
<td>Secondary malignant neoplasm of inguinal lymph nodes</td>
<td>Current Absent</td>
</tr>
</tbody>
</table>
Natural language processing (NLP)

Source data (text)
- Hastily written or dictated
- Typos
- Transcription errors
- Telegram style
- Acronyms, abbreviations
- Dialects
- Sublanguages
- It’s not going to change substantially!

Semantic Resources
- Standardised Target Representation
- Ontologies
- Terminologies

ML Models
Rules
Reference Corpora

Text Mining
De-Identification
Semantic Enrichment

<table>
<thead>
<tr>
<th>Code (SNOMED CT)</th>
<th>Value</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>254730000</td>
<td>Superficial spreading malignant melanoma of skin</td>
<td>History of</td>
</tr>
<tr>
<td>301889008</td>
<td>Excision of malignant skin tumour</td>
<td>History of</td>
</tr>
<tr>
<td>47224004</td>
<td>Skin of posterior surface of lower leg</td>
<td>Current</td>
</tr>
<tr>
<td>7771000</td>
<td>Left</td>
<td></td>
</tr>
<tr>
<td>81827009</td>
<td>Diameter</td>
<td>2.4 Current</td>
</tr>
<tr>
<td>258673006</td>
<td>Millimetre</td>
<td></td>
</tr>
<tr>
<td>94339008</td>
<td>Secondary malignant neoplasm of inguinal lymph nodes</td>
<td>Current Absent</td>
</tr>
</tbody>
</table>
Natural language processing (NLP)

Source data (text)
- Hastily written or dictated
- Typos
- Transcription errors
- Telegram style
- Acronyms, abbreviations
- Dialects
- Sublanguages
- It’s not going to change substantially!

Semantic Resources
- Clinical NLP lagging behind
- Privacy vs. sharing of annotated corpora
- Reliability of de-identification
- Data ownership vs. sharing of models
- Low adherence to standards (e.g. SNOMED CT)
- Quality issues of standards
- Coverage of clinical jargon by terminologies: Translation vs. interface terminology creation → (PMID 29295238)

Standardised Target Representation

<table>
<thead>
<tr>
<th>Code (SNOMED CT)</th>
<th>Value</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>254730000</td>
<td>Superficial spreading malignant melanoma of skin</td>
<td>History of</td>
</tr>
<tr>
<td>301889008</td>
<td>Excision of malignant skin tumour</td>
<td>History of</td>
</tr>
<tr>
<td>47224004</td>
<td>Skin of posterior surface of lower leg</td>
<td>Current</td>
</tr>
<tr>
<td>7771000</td>
<td>Left</td>
<td></td>
</tr>
<tr>
<td>81827009</td>
<td>Diameter</td>
<td>Current</td>
</tr>
<tr>
<td>258673006</td>
<td>Millimetre</td>
<td>2.4</td>
</tr>
<tr>
<td>94339008</td>
<td>Secondary malignant neoplasm of inguinal lymph nodes</td>
<td>Current</td>
</tr>
<tr>
<td></td>
<td>Absent</td>
<td></td>
</tr>
</tbody>
</table>
Natural language processing (NLP)

Source data (text)
- Hastily written or dictated
- Typos
- Transcription errors
- Telegram style
- Acronyms, abbreviations
- Dialects
- Sublanguages
- It’s not going to change substantially!

Standardised Target Representation
- Competing representations of same content
  - Low inter-coder agreement
    \(\Rightarrow\) ASSESS CT (PMID: 30654902)
- Meaning vs. context:
  - Negation
  - Plan
  - Uncertainty
  - Other subjects (family history)
- Ontologies (e.g. SNOMED CT) vs. information models (e.g. FHIR)

Semantic Resources
- Clinical NLP lagging behind
- Privacy vs. sharing of annotated corpora
- Reliability of de-identification
- Data ownership vs. sharing of models
- Low adherence to standards (e.g. SNOMED CT)
- Quality issues of standards
- Coverage of clinical jargon by terminologies: Translation vs. interface terminology creation
  \(\Rightarrow\) (PMID 29295238)
Precision medicine requires precision representations of clinical language.
Precision medicine requires precision representations of clinical language

• **Subtle difference in spelling – large difference in meaning**
  - "Sodium chloride", "Sodium chlorite", "Sodium chlorate"
  - "AIDS", "ARDS", "STEMI", "NSTEMI"
  - "Hepatitis A", "Hepatitis B", "Hepatitis C"

• **Synonyms**

• **Homonyms**
  - "RTA": "road traffic accident" vs. "renal tubular acidosis"

• **Neologisms**
  - Single-word compounds, e.g. in German: "Mediainfarktverdacht", "Botulismustoxinvergiftung"
Example: SNOMED CT Interface Terminology for German

<table>
<thead>
<tr>
<th>SNOMED ID</th>
<th>Score</th>
<th>Fully Specified Name (English)</th>
<th>German Interface Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>99451000119105</td>
<td>0.833</td>
<td>Cerebral infarction due to stenosis of carotid artery (disorder)</td>
<td>Hirninfarkt verursacht durch Stenose der A. carotis</td>
</tr>
<tr>
<td>99451000119105</td>
<td>0.833</td>
<td>Cerebral infarction due to stenosis of carotid artery (disorder)</td>
<td>Hirninfarkt verursacht durch Stenose der A. karotis</td>
</tr>
<tr>
<td>99451000119105</td>
<td>0.833</td>
<td>Cerebral infarction due to stenosis of carotid artery (disorder)</td>
<td>Schlaganfall wegen Stenose der Halsschlagader</td>
</tr>
<tr>
<td>99451000119105</td>
<td>0.833</td>
<td>Cerebral infarction due to stenosis of carotid artery (disorder)</td>
<td>Insult wegen Stenose der Halsschlagader</td>
</tr>
<tr>
<td>99451000119105</td>
<td>0.833</td>
<td>Cerebral infarction due to stenosis of carotid artery (disorder)</td>
<td>Schlaganfall wegen Karotisstenose</td>
</tr>
<tr>
<td>99451000119105</td>
<td>0.833</td>
<td>Cerebral infarction due to stenosis of carotid artery (disorder)</td>
<td>Insult wegen Karotisstenose</td>
</tr>
<tr>
<td>99451000119105</td>
<td>0.800</td>
<td>Cerebral infarction due to stenosis of carotid artery (disorder)</td>
<td>Gehirninfarkt verursacht durch Verengung der Halsschlagader</td>
</tr>
</tbody>
</table>

Example: annotation for smoking status

- Text snippets from discharge summaries.
  Annotations: {current smoker, past smoker, never smoked}
Smoking Status Custom annotators

7242 manually annotated context lines

Parameter optimized **shallow neural network**
Annotator integrated into Averbis health discovery platform

<table>
<thead>
<tr>
<th>Smoking Status</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>CURRENT-NON-SMOKER</td>
<td>0.80</td>
<td>0.94</td>
<td>0.86</td>
<td>64</td>
</tr>
<tr>
<td>CURRENT-SMOKER</td>
<td>0.93</td>
<td>0.97</td>
<td>0.95</td>
<td>958</td>
</tr>
<tr>
<td>NEVER-SMOKER</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>2</td>
</tr>
<tr>
<td>PAST-SMOKER</td>
<td>0.95</td>
<td>0.84</td>
<td>0.89</td>
<td>423</td>
</tr>
<tr>
<td>UNKNOWN</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>accuracy</td>
<td>0.93</td>
<td></td>
<td></td>
<td>1448</td>
</tr>
<tr>
<td>macro avg</td>
<td>0.64</td>
<td>0.65</td>
<td>0.64</td>
<td>1448</td>
</tr>
<tr>
<td>weighted avg</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>1448</td>
</tr>
</tbody>
</table>

https://fasttext.cc/
https://averbis.com/de/health-discovery/
Precision medicine requires precision standards
Precision medicine requires precision standards

- **Two kinds of semantic standards for interoperable representation of EHR content**
  - Information models (models of use):
    Standardised templates for recurring clinical documentation needs, e.g.
    - condition, observation, procedure, medication administration
  - Ontologies (models of meaning)
    Standardised formal and informal descriptions for types of entities that are referred to by the EHR
    - diseases, procedures, substances,
      - body parts, organisms, lab observables
    - linked to technical terms in several languages

- **Ontology IDs provide standardised meaning for the patient-specific instantiations of FHIR resources**
Standards require precise definitions

- **Problem: ill-defined primitives**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Element Id</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Condition</td>
<td>A clinical condition, problem, diagnosis, or other event, situation, issue, or clinical concept that has risen to a level of concern.</td>
</tr>
</tbody>
</table>

[Clinical finding] represents the result of a clinical observation, assessment or judgment and includes normal and abnormal clinical states e.g. [asthma], [headache], [normal breath sounds]). The [clinical finding] hierarchy includes concepts used to represent diagnoses.

Appendicitis ≡ Disease ⊑

∃ Role_Group.(∃ Finding_site.Appendix_structure ⊑

∃ Associated_morphology.Inflammatory_morphology)

Adolescent ⊑ Minor

Infant ⊑ Minor

(no text definition, no formal definition)
Standards should support the detection of "isosemantic" expressions

Text 1: "in the nail of the right great toe, candida species were found as cause of infection"
Text 2: "candida onychomycosis, right great toe"

<table>
<thead>
<tr>
<th>Text 1</th>
<th>Asserted SNOMED concepts</th>
<th>Implied SNOMED CT concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Im rechten</td>
<td>Right (qualifier value)</td>
<td>is-a</td>
</tr>
<tr>
<td>Großzehennagel</td>
<td>Structure of nail unit of great toe (body structure)</td>
<td>is-a</td>
</tr>
<tr>
<td>fanden sich</td>
<td>Structure of nail unit of toe (body structure)</td>
<td>is-a</td>
</tr>
<tr>
<td>Candida-Spezies</td>
<td>Genus Candida (organism)</td>
<td>is-a</td>
</tr>
<tr>
<td>als Ursache der</td>
<td>Infectious process (qualifier value)</td>
<td>is-a</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text 2</th>
<th>Asserted SNOMED concepts</th>
<th>Implied SNOMED CT concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candida-Onychomykose</td>
<td>Candidiasis of nails (disorder)</td>
<td>causative agent</td>
</tr>
<tr>
<td>Großzeh rechts</td>
<td>Structure of right great toe (body structure)</td>
<td>is-a</td>
</tr>
<tr>
<td></td>
<td></td>
<td>is-a</td>
</tr>
</tbody>
</table>
Standards + data should allow detecting semantically close expressions

Problems of large ontologies and terminologies: semantically close, undefined classes
Take-home messages

- Clinical data are overly heterogeneous
- Much information needs to be extracted from free text
- NLP-based information extraction requires costly resources
- Lack of openly-accessible clinical text
- Precision medicine needs
  - Precision information extraction tools
  - Precision language resources
  - Precision semantic standards
Thank you!

Stefan Schulz
stefan.schulz@medunigraz.at

References:

- Schulz S. Clinical Informatics Challenges in Precision Medicine. Pathways to Precision Medicine. To Appear 2021

Acknowledgements:
This work has been carried out with the K1 COMET Competence Center CBmed, which is funded by the Federal Ministry of Transport, Innovation and Technology (BMVIT); the Federal Ministry of Science, Research and Economy (BMWFW); Land Steiermark (Department 12, Business and Innovation); the Styrian Business Promotion Agency (SFG); and the Vienna Business Agency. The COMET program is executed by the FFG. KAGes and SAP provided significant resources, manpower and data as basis for research and innovation.