Clinical Informatics Challenges in Precision Medicine

Stefan Schulz, Medical University of Graz
Conflict of Interest Disclosure

• Professor for Medical Informatics, Medical University of Graz, Austria
• Project leader at CBmed Biomarker Research GmbH, Graz Austria
• Head of Medical Research at Averbis GmbH, Freiburg, Germany
“Precision medicine’ has emerged as a computational approach to functionally interpret omics and big data, and facilitate their application to healthcare 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.”
“Fuel” for precision medicine

Source: CBmed – Center for Biomarker Research in Medicine, Graz, Austria
“Fuel” for precision medicine

Phenotype
Environment
Lifestyle
Clinical History

Next Generation Sequencing (NGS)

Integrative Data Analysis

Proteomics
Metabolomics
MALDI-MS

Digital Pathology
Flow Cytometry

in vivo Imaging

Where is this data?

Source: CBmed – Center for Biomarker Research in Medicine, Graz, Austria
Digital footprints

Phenotype
Environment
Lifestyle
Clinical History
Can Facebook’s Machine-Learning Algorithms Accurately Predict Suicide?

March 10, 2017

**News Type:** Weekly Spark, Weekly Spark News

**Scientific American**

Facebook has just expanded the array of tools it provides to reach users at risk for suicide and connect them with mental health resources. The menu of options that allows Facebook users to report posts with content indicating potential thoughts of suicide or self-harm will now be available for Facebook live streams as well. The social media company is also piloting a pattern recognition algorithm that it hopes will automatically identify posts of concern even if they have not yet been reported by users. According to Facebook spokesperson William Nevis, the algorithm will use words or phrases related to suicide or self-harm in a user’s post, and in comments added by friends, to determine if the person may be at risk. The system will automatically alert Facebook's Community Operations team about posts of concern so that the team can quickly review them. If the team determines that support is warranted, they will ensure that information about helping resources will appear in the user's news feed.

**Spark Extra!** Check out a [community guide for Facebook users](#).

**Planning and Implementing:** New and Social Media
Digital footprints

Health Records

Phenotype
Environment
Lifestyle
Clinical History
EHRs
Electronic Health Records

CLINICAL INFORMATICS

Phenotype
Environment
Lifestyle
Clinical History
What is in EHRs?          How can it be used for PM?

EHRs
Electronic Health Records
# The EHR heat map

<table>
<thead>
<tr>
<th></th>
<th>Completeness</th>
<th>Correctness</th>
<th>Granularity</th>
<th>Structure</th>
<th>Interoperability</th>
<th>Data Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics / ADT</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>
<tr>
<td>Demographics / ADT</td>
<td>Completeness</td>
<td>Correctness</td>
<td>Granularity</td>
<td>Structure</td>
<td>Interoperability</td>
<td>Data Volume</td>
</tr>
<tr>
<td>-------------------</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>TEXT</td>
<td>TEXT</td>
<td>TEXT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prescriptions</td>
<td>TEXT</td>
<td>TEXT</td>
<td>TEXT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problem List</td>
<td>TEXT</td>
<td>TEXT</td>
<td>TEXT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical Registries</td>
<td>TEXT</td>
<td>TEXT</td>
<td>TEXT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Findings Reports</td>
<td>TEXT</td>
<td>TEXT</td>
<td>TEXT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discharge Summaries</td>
<td>TEXT</td>
<td>TEXT</td>
<td>TEXT</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
St. p. TE eines exulc. sek.knot.SSM li US dors. 5/11 Level IV 2,4 mm Tumordurchm. Sentinell LK ing. li. tumorfr.

Acute kidney failure, unspecified

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/inh. 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) pipeline

Source data (text)

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

Semantic Enrichment

Text Mining De-Identification

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></td>
</tr>
<tr>
<td>94339008</td>
<td>Secondary malignant neoplasm of inguinal lymph nodes</td>
<td>Current Absent</td>
</tr>
</tbody>
</table>

Semantic Resources

Ontologies
Terminologies

ML Models
Rules
Reference Corpora
Natural language processing (NLP) pipeline

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
- Ontologies
- Terminologies

ML Models
- Rules
- Reference Corpora

Text Mining
- De-Identification
- Semantic Enrichment

Standardised Target Representation

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

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

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 in France, Germany)
- 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>2.41</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>
NLP issues

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 → ASSESS CT (PMID: 30654902)
- Meaning vs. context:
  - Negation
  - Plan
  - Uncertainty
  - Other subjects (family history)
- Ontologies vs. information models
- Technical issues: data warehousing, querying, (poly)hierarchical expansions

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 in France, Germany)
- Quality issues of standards
- Coverage issues of clinical jargon by terminologies: Translation vs. interface terminology creation → (PMID 29295238)
Example: DBM4PM
(Digital Biomarkers for Precision Medicine)
Example: DBM4PM
(Digital Biomarkers for Precision Medicine)

Source data (text)
- Discharge Summaries
- Problem lists from KAGes hospital network, Austria

Text Mining
De-Identification

Semantic Enrichment

Semantic Resources
- SNOMED CT
- Drug DB
- German Interface Lexicon
- Lexicon maintenance

Rules

Short form resolution

Deep Learning

Clinical texts
WWW content

Semi-automated lexicon acquisition

Standardised Target Representation
- Use cases
  - Annotations of biobank samples with standardised clinical features
  - Support cohort building for clinical research
  - Use extracts from clinical documents for automatically generated “EHR Quick View”
  - Feed prediction model for risk of acute delirium in hospitalised patients
  - Improve quality of coding

R
The concept of “Digital Biomarkers”

• “data or data extracts that can be obtained from all kinds of artefacts related to an individual, and on which health-related predictions can be grounded”

• Heterogeneous in format, quality, correctness, completeness, structure
  ▪ Not primarily acquired for prediction of conditions / events
  ▪ Extracted from EHRs, social networks, mobile devices
  ▪ Often implicit contexts

• Different levels of complexity
  ▪ Simple: single data points
  ▪ Complex: series of data points
  ▪ Algorithmic: data + multivariate prediction models

• Predictive value:
  ▪ Allows prediction of conditions or events to a relevant degree
  ▪ Good Predictions possible from noisy data

1. M. S. Lim et al. Advancing biomarker development through convergent engagement. Summary Report of the 2nd International Danube Symposium on Biomarker development, Molecular Imaging and Applied Diagnostics; March 14-16, 2018; Vienna, Austria, UNDER REVIEW
# Examples of digital biomarkers

<table>
<thead>
<tr>
<th>Digital Biomarker</th>
<th>Condition / Event</th>
<th>Specificity</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>*GAITRite® signals</td>
<td>Bradykinesia</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>*Wii Balance Board signals</td>
<td>Postural instability</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>**HITEx algorithm</td>
<td>Current Smoker</td>
<td>+++</td>
<td>++</td>
</tr>
<tr>
<td>Mention of &quot;Metformin&quot; in the EHR</td>
<td>Diabetes mellitus type 2</td>
<td>++</td>
<td>-</td>
</tr>
<tr>
<td>Administrative ICD codes I10 or I11 or I12 or I13 or I15</td>
<td>Hypertensive disease</td>
<td>++</td>
<td>+/-</td>
</tr>
<tr>
<td>substring &quot;malign&quot; in pathology report</td>
<td>malignancy</td>
<td>--</td>
<td>+</td>
</tr>
<tr>
<td>*** Regular expression pattern matching</td>
<td>Gleason score, Clark level, Breslow depth</td>
<td>++</td>
<td>++</td>
</tr>
</tbody>
</table>


** Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Medical Informatics and Decision Making. 2006;6:30

Conclusions

• Information about phenotype, clinical history, lifestyle: “buried” in clinical narratives

• Clinical texts primarily written for inter-professional communication

• High diversity and idiosyncrasy of medical (sub)languages and look & feel of clinical documents
  ▪ Not likely to change significantly
  ▪ Dependent on tools, workflows, institutional cultures

• Clinical Informatics, particularly NLP approaches promising but their usability for precision medicine highly dependent on
  ▪ Community-maintained resource (corpora, dictionaries), bottlenecks: accessibility, shareability of clinical corpora, dictionary creation / maintenance
  ▪ Semantic standards (coding systems, ontologies), quality issues, adherence

• Notion of “digital biomarker”: even simple language extracts or low-quality codes may be useful for predictions
Thank you!

Stefan Schulz
stefan.schulz@medunigraz.at

References:

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.