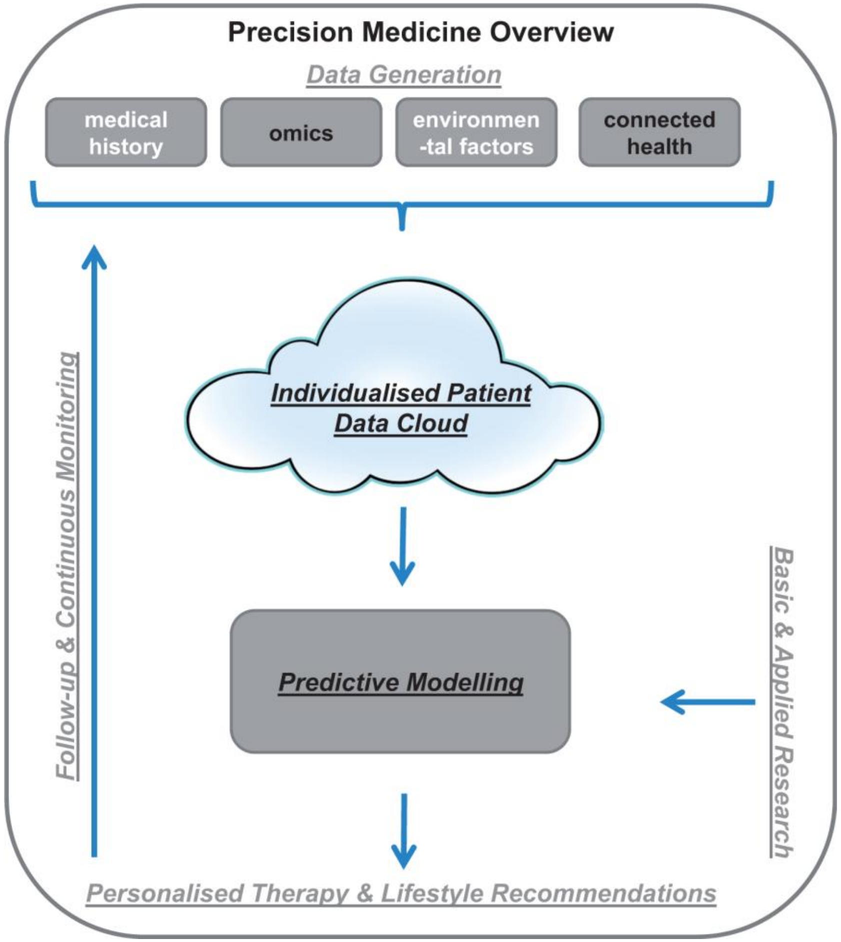
Clinical Informatics Challenges in Precision Medicine

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

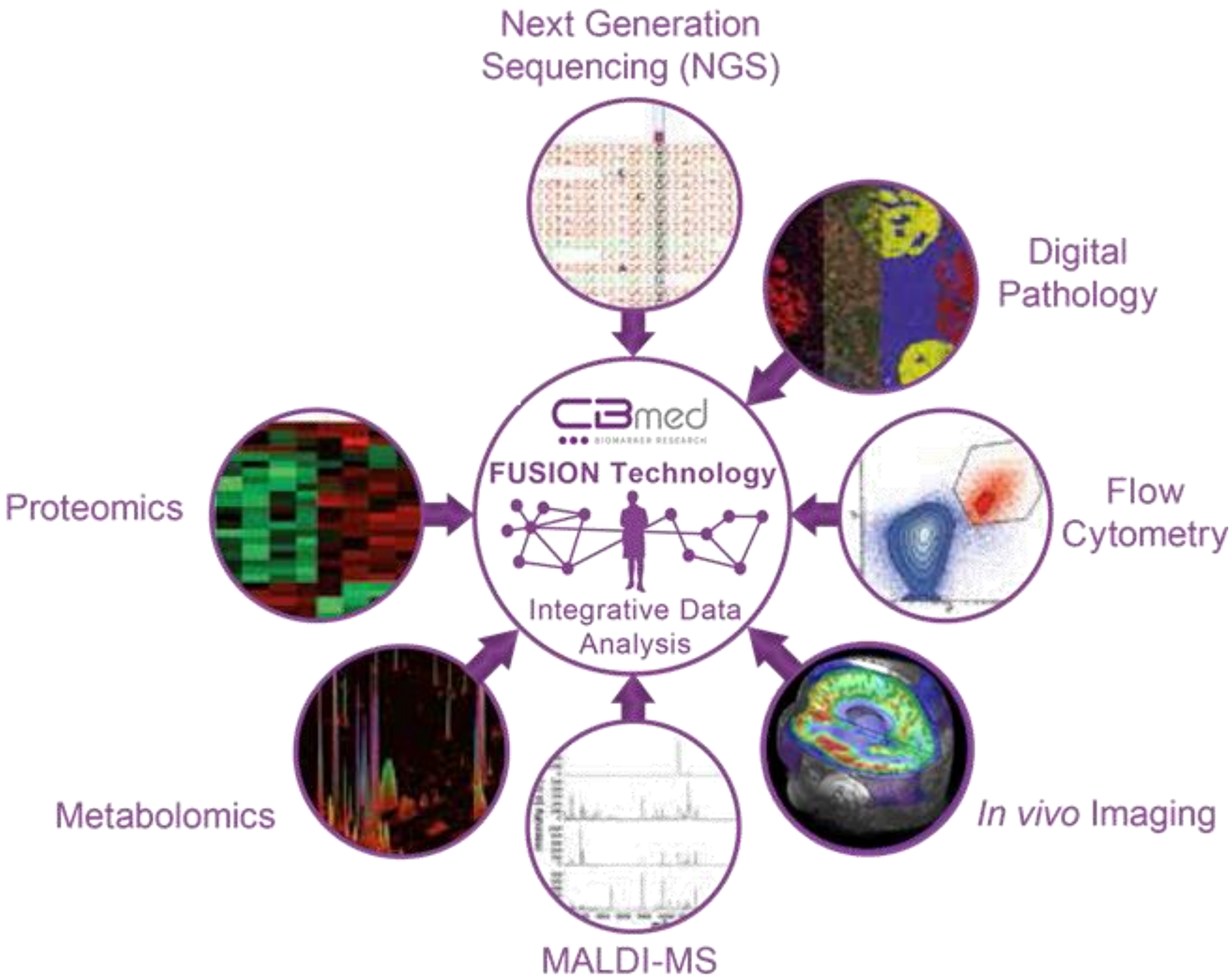
Precision Medicine is data centred



"'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**"

David J. Duffy. Problems, challenges and promises: perspectives on precision medicine. Briefings in Bioinformatics, Volume 17, Issue 3, May 2016, Pages 494–504, https://doi.org/10.1093/bib/bbv060

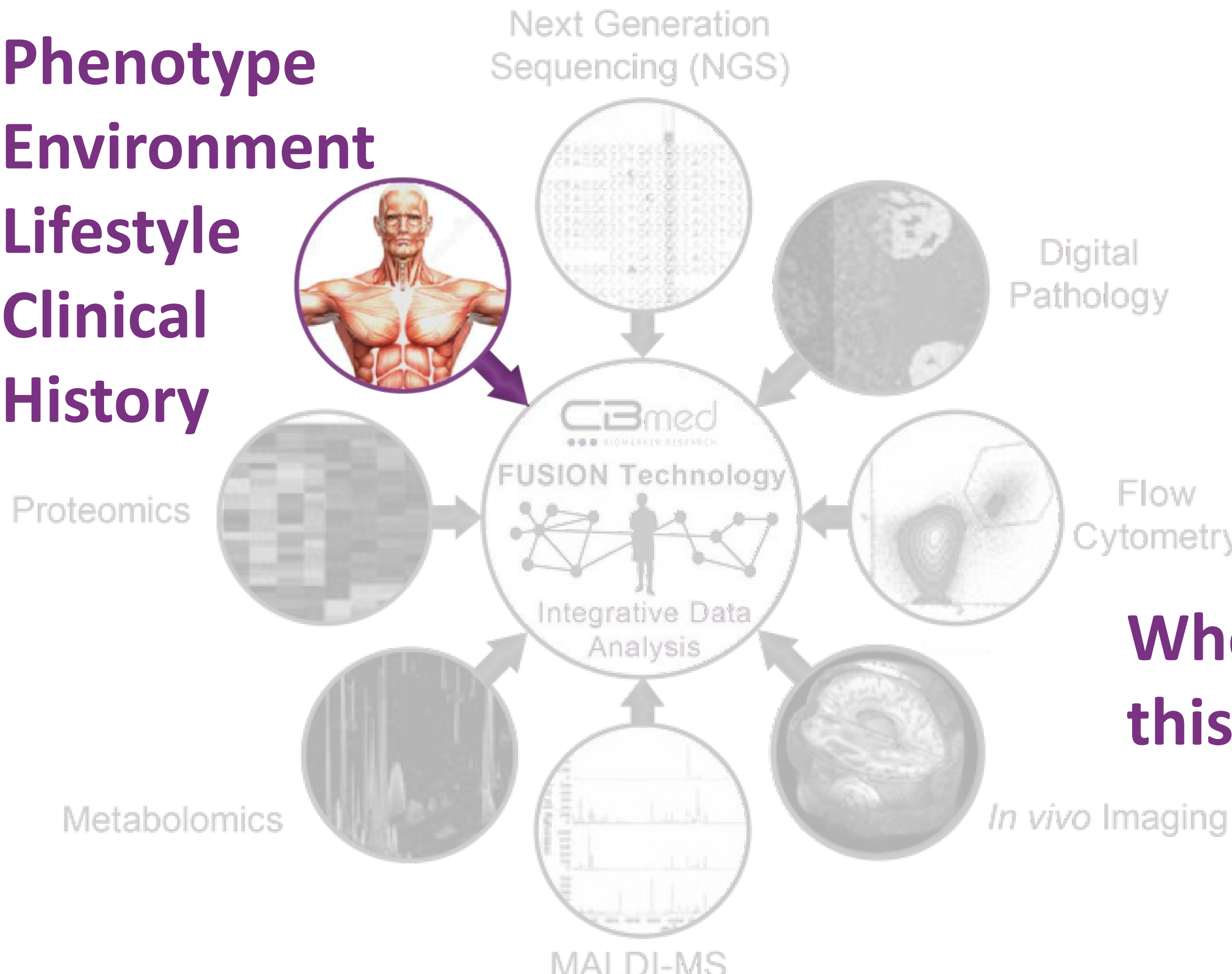
“Fuel” for precision medicine



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

“Fuel” for precision medicine

Phenotype
Environment
Lifestyle
Clinical
History

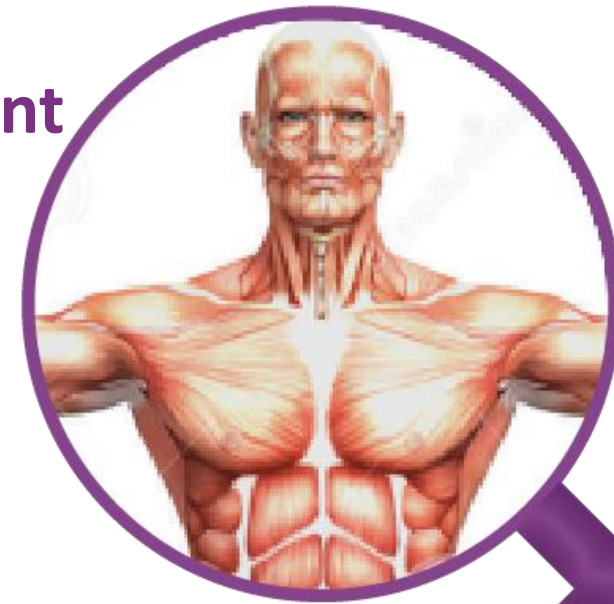


Where is this data?

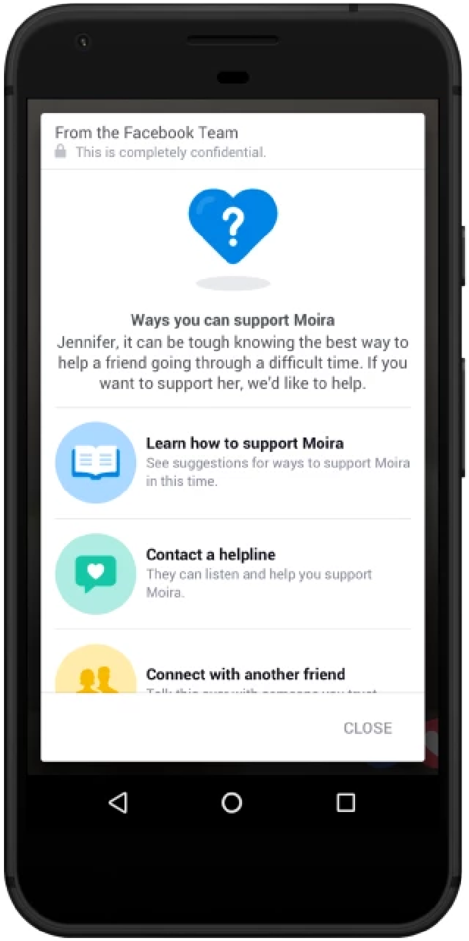
Source: CBmed – Center for Biomarker Research in Medicine, Graz, Austria

Digital footprints

Phenotype
Environment
Lifestyle
Clinical
History



New from the *Weekly Spark*



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 Nevius, 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.

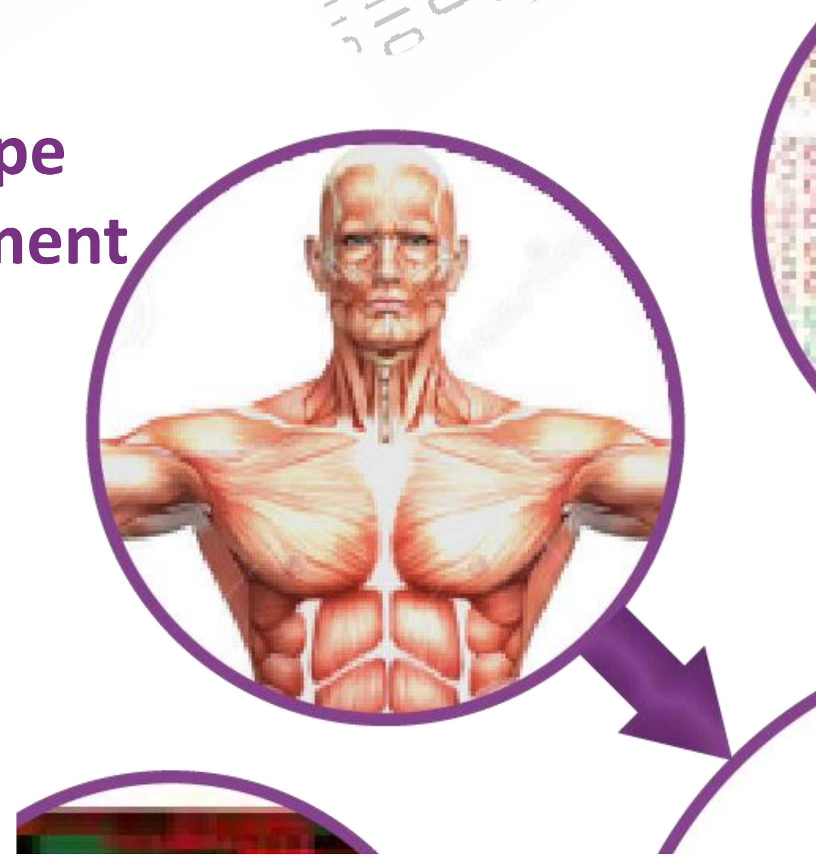
Spark Extra! Check out a community guide for Facebook users.

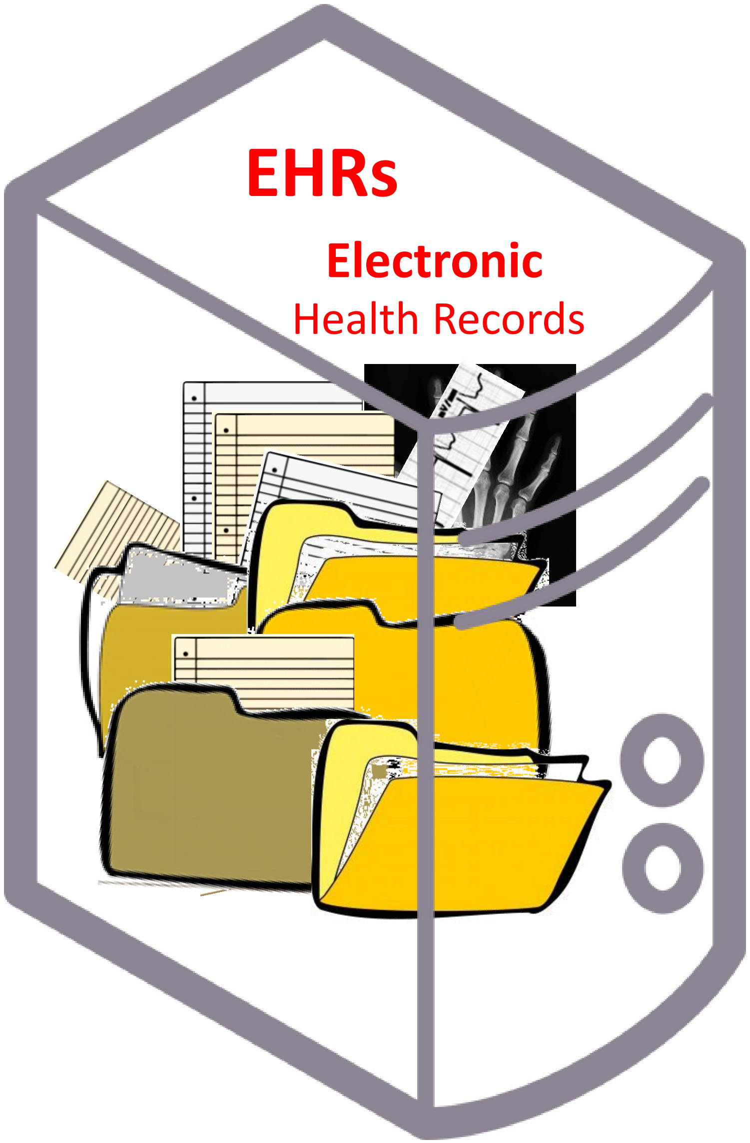
Planning and Implementing: New and Social Media

Digital footprints

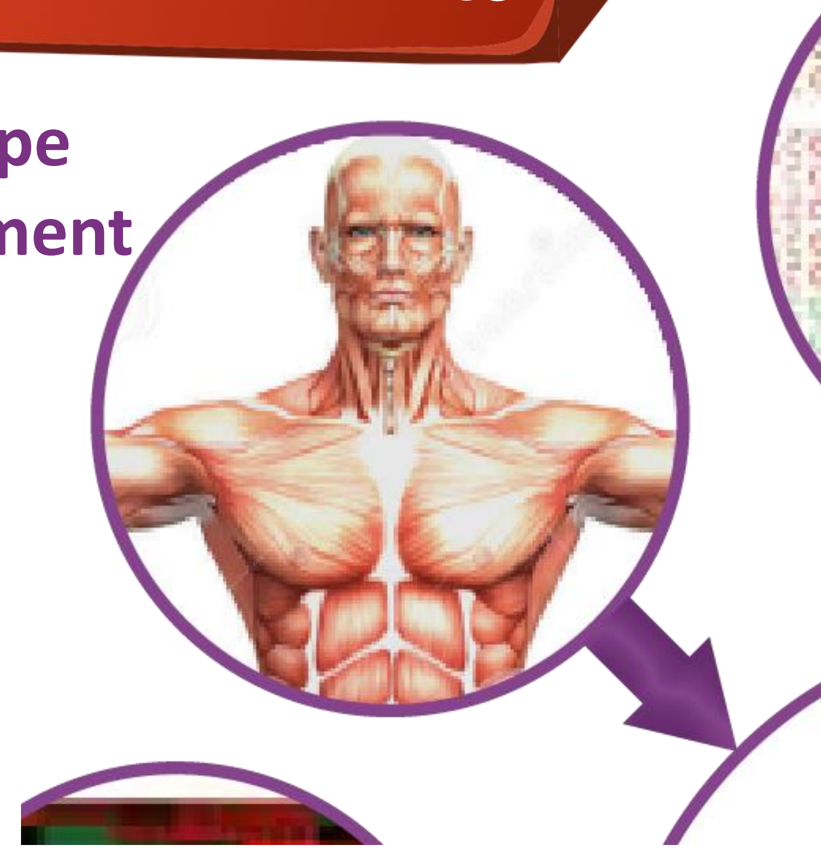
Health Records

Phenotype
Environment
Lifestyle
Clinical
History

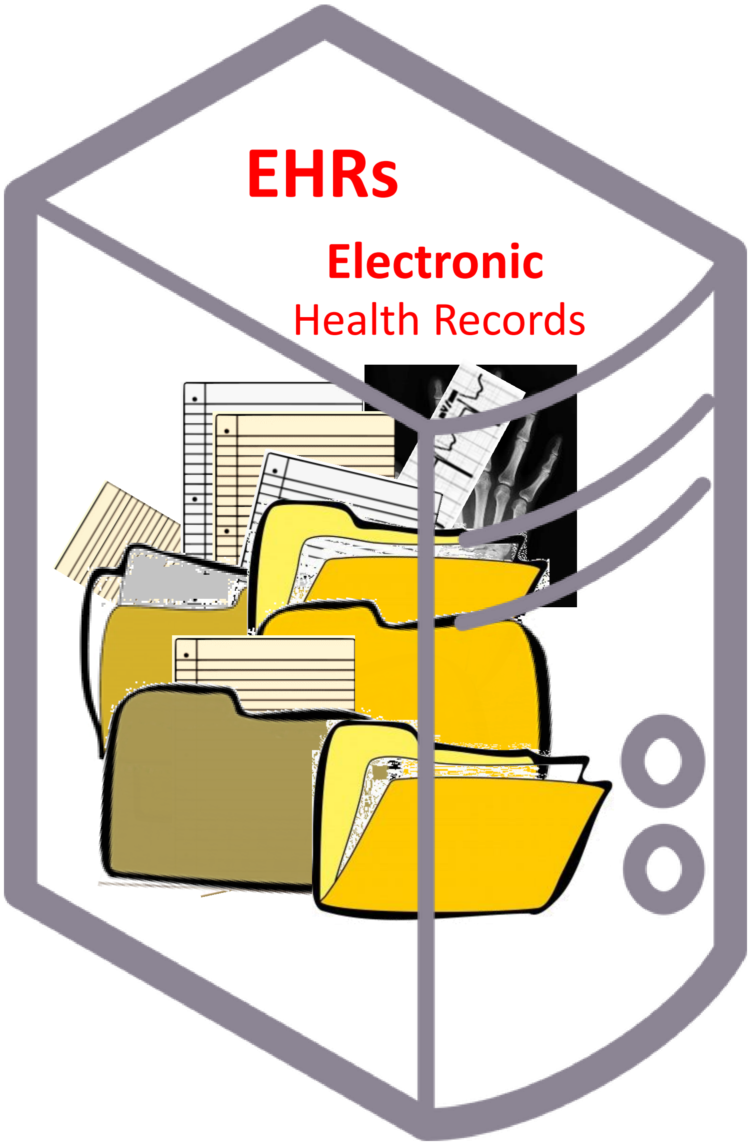




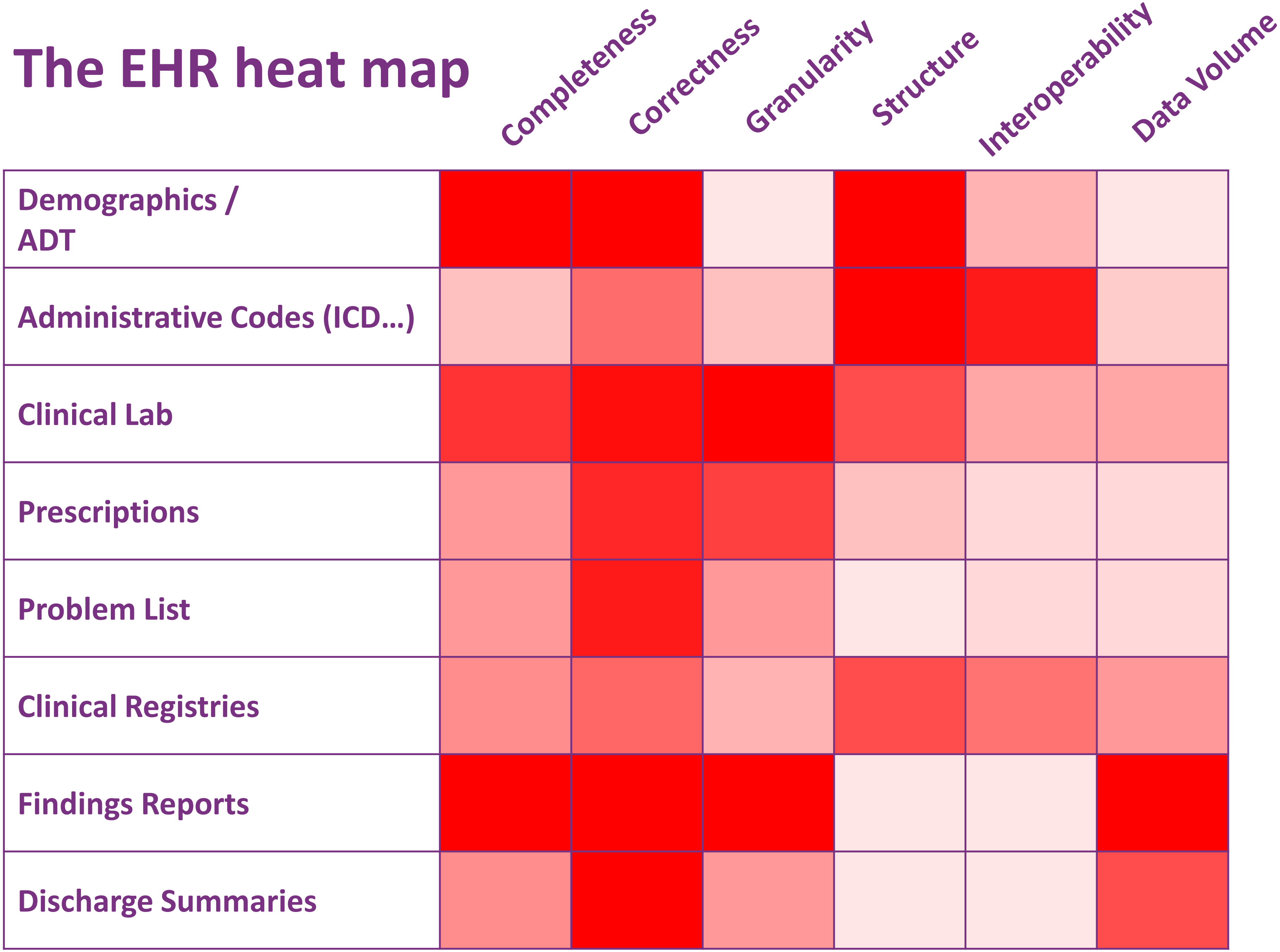
Phenotype
Environment
Lifestyle
Clinical
History



What is in EHRs? How can it be used for PM?



The EHR heat map



The EHR heat map

	Completeness	Correctness	Granularity	Structure	Interoperability	Data Volume
Demographics / ADT						
Administrative Codes (ICD…)						
Clinical Lab						
Prescriptions				TEXT		
Problem List				TEXT		
Clinical Registries						
Findings Reports				TEXT		
Discharge Summaries				TEXT		

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 ;
```

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*

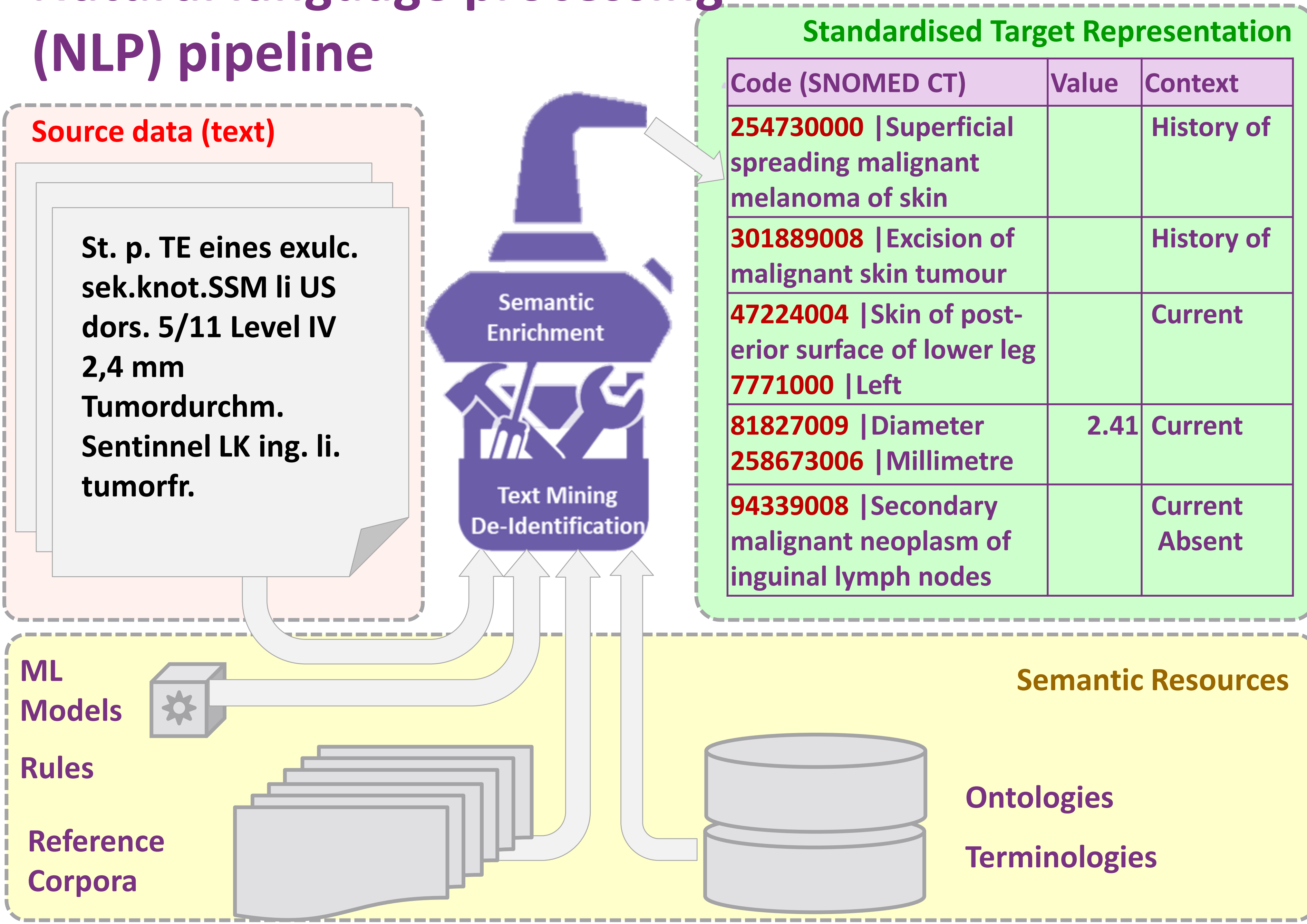
History of Present Illness *72 year old thin white male presented to emergency on 8/1/14 with shortness of breath, weakness and dehydration. Chest X-ray showed right lower lobe infiltrate, ABGs unremarkable. Pulse ox on RA was 79%.*

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/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) pipeline



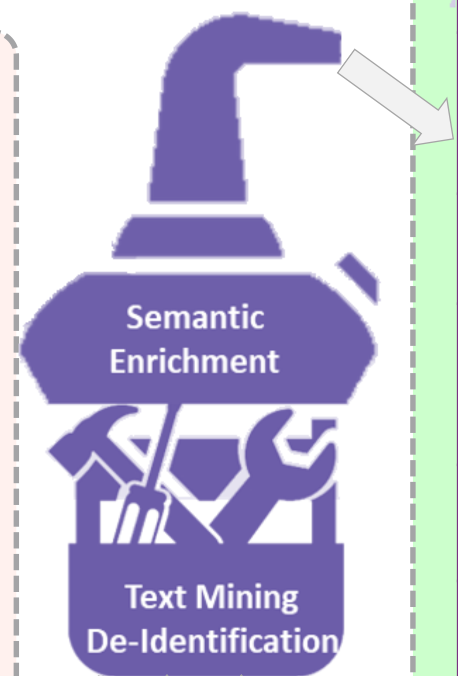
Code (SNOMED CT)	Value	Context
254730000 |Superficial spreading malignant melanoma of skin		History of
301889008 |Excision of malignant skin tumour		History of
47224004 |Skin of post-erior surface of lower leg 7771000 |Left		Current
81827009 |Diameter 258673006 |Millimetre	2.41	Current
94339008 |Secondary malignant neoplasm of inguinal lymph nodes		Current Absent

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!**

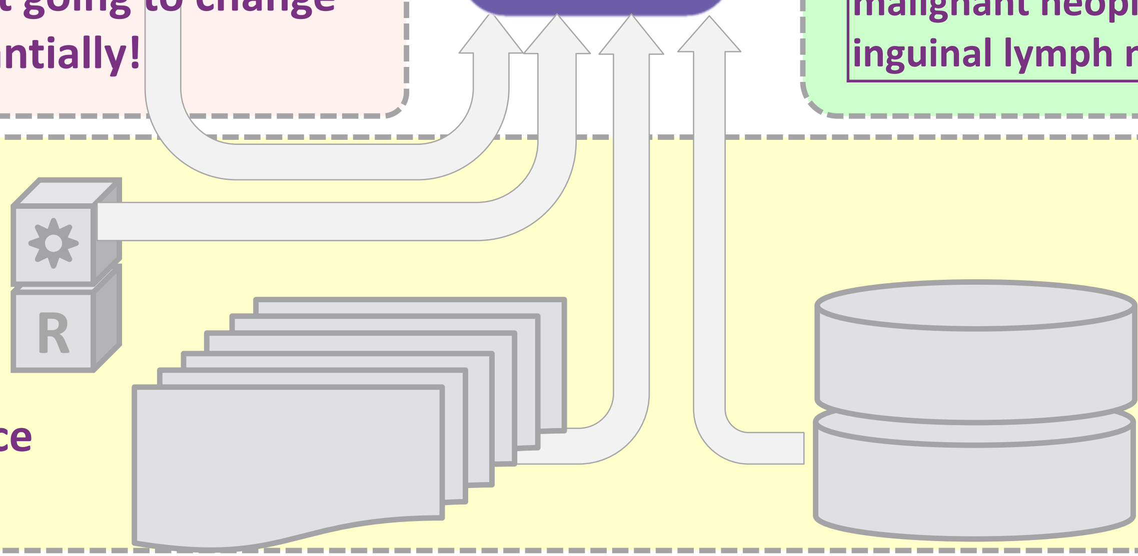


Standardised Target Representation

Code (SNOMED CT)	Value	Context
254730000 |Superficial spreading malignant melanoma of skin		History of
301889008 |Excision of malignant skin tumour		History of
47224004 |Skin of post-erior surface of lower leg 7771000 |Left		Current
81827009 |Diameter 258673006 |Millimetre	2.41	Current
94339008 |Secondary malignant neoplasm of inguinal lymph nodes		Current Absent

ML Models

Rules



Reference Corpora

Semantic Resources

Ontologies

Terminologies

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

Code (SNOMED CT)	Value	Context
254730000 |Superficial spreading malignant melanoma of skin		History of
301889008 |Excision of malignant skin tumour		History of
47224004 |Skin of post-erior surface of lower leg 7771000 |Left		Current
81827009 |Diameter 258673006 |Millimetre	2.41	Current
94339008 |Secondary malignant neoplasm of inguinal lymph nodes		Current Absent

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)

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 of clinical jargon by terminologies: Translation vs. interface terminology creation → (PMID 29295238)

Example: DBM4PM

(Digital Biomarkers for Precision Medicine)

Standardised Target Representation

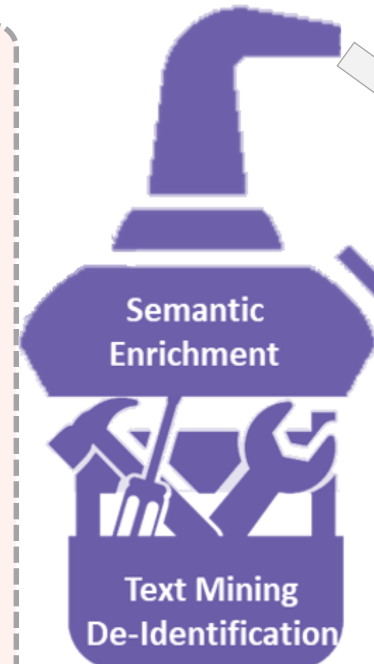
Example: DBM4PM

(Digital Biomarkers for Precision Medicine)

Source data (text)

- Discharge Summaries
- Problem lists

from KAGes hospital network, Austria



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

Semantic Resources

Deep Learning

Rules

Short form resolution

Clinical texts
WWW content

Semi-automated lexicon acquisition

- **SNOMED CT**
- **Drug DB**

German Interface Lexicon

Lexicon maintenance

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[1]”
- 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[2]

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
2. Jauk, S; Kramer, D; Schulz, S; Leodolter, W. Evaluating the Impact of Incorrect Diabetes Coding on the Performance of Multivariable Prediction Models. Stud Health Technol Inform. 2018; 251: 249-252.

Examples of digital biomarkers

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

*Godinho C, Domingos J, Cunha G, et al. A systematic review of the characteristics and validity of monitoring technologies to assess Parkinson's disease. Journal of NeuroEngineering and Rehabilitation. 2016;13:24.
** 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 Cancer
*** Napolitano G, Fox C, Middleton R, Connolly D. Pattern-based information extraction from pathology reports for cancer registration. Causes Control. 2010 Nov;21(11):1887-94.

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:

- Kreuzthaler M, Pfeifer B, Vera Ramos JA, Kramer D, Grogger V, Bredenfeldt S, Pedevilla M, Krisper P, Schulz S. EHR Text Categorization for Enhanced Patient-Based Document Navigation.
- Stud Health Technol Inform. 2018;248:100-107.Miñarro-Giménez, JA; Cornet, R; Jaulent, MC; Dewenter, H; Thun, S; Gøeg, KR; Karlsson, D; **Schulz, S**. Quantitative analysis of manual annotation of clinical text samples. Int J Med Inform. 2019; 123:37-48
- Jauk, S; Kramer, D; **Schulz, S**; Leodolter, W. Evaluating the Impact of Incorrect Diabetes Coding on the Performance of Multivariable Prediction Models. Stud Health Technol Inform. 2018; 251: 249-252.
- **Schulz, S**; Kreuzthaler, M; Huppertz, B; Sargsyan K; Kaiser, P; Fasching, R; Pieber, T. Secondary Use of Clinical Routine Data for Enhanced Phenotyping of Biobank Sample Data. Proceedings of the 1st Global Biobank Week. 2017; 1(1):53-53.-Global Biobank Week; SEP 13-15, 2017; Stockholm, SWEDEN.
- Kreuzthaler M, Martínez-Costa C, Kaiser P, Schulz S. Semantic Technologies for Re-Use of Clinical Routine Data.
- Stud Health Technol Inform. 2017;236:24-31.
- Kramer D, Veeranki S, Hayn D, Quehenberger F, Leodolter W, Jagsch C, Schreier G. Development and Validation of a Multivariable Prediction Model for the Occurrence of Delirium in Hospitalized Gerontopsychiatry and Internal Medicine Patients. Stud Health Technol Inform. 2017;236:32-39.
- Schulz S, Rodrigues JM, Rector A, Chute CG. Interface Terminologies, Reference Terminologies and Aggregation Terminologies: A Strategy for Better Integration.
- Oleynik M, Kreuzthaler M, Schulz S. Unsupervised Abbreviation Expansion in Clinical Narratives. Stud Health Technol Inform. 2017;245:539-543.

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.