

Retrieval in clinical data extracts: the role of semantic standards, text mining platforms and language models

Graz, Austria, 2025-06-04



Stefan Schulz

<https://purl.org/steschu>

Retrieval – performance measures



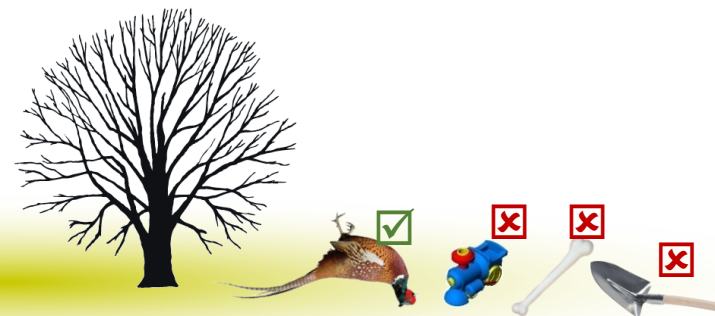
Retrieval – performance measures



Precision = 60% (3 of 5)
(two false positives)

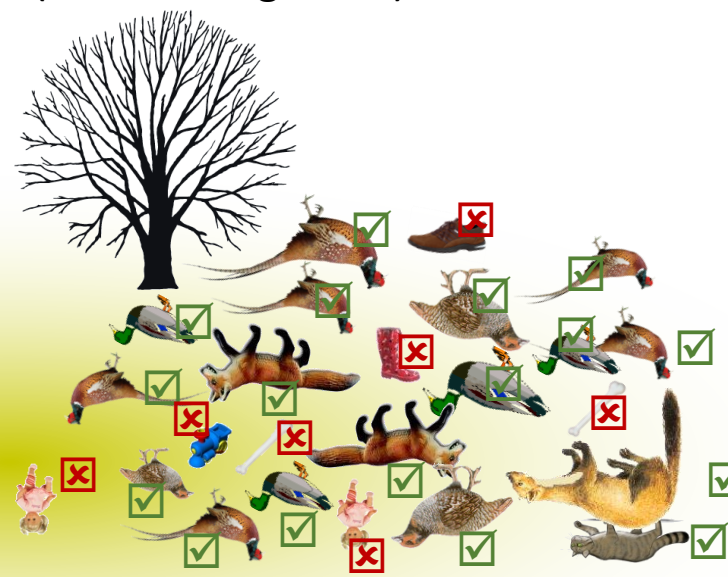
Precision = found relevant / found
Recall = found relevant / relevant

$$\text{F1-Measure} = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}$$



Recall= 75% (3 of 4)
(one false negative)

Recall= 15% (3 of 20)
(17 false negatives)



Text Retrieval - Googling as a cultural technology

googeln



Der deutsche Wortschatz von 1600 bis heute.

Lesezeichen

zitieren/teilen

ausklappen

Grammatik	Verb · <u>googelt</u> , <u>googelte</u> , <u>hat gegoogelt</u>
Nebenform	seltener googlen · Verb · <u>googlet</u> , <u>googlete</u> , <u>hat gegooglet</u>
Aussprache	['gu:ɡlɪn]
Worttrennung	goo-geln · goog-len
Grundform	Google _{Eigenname}
Wortbildung	mit ›googeln‹/›googlen‹ als Letztglied: ↗ nachgoogeln / nachgooglen

Bedeutungsgeschichte

Das Verb basiert auf dem Namen der zur Zeit der Prägung mit Abstand am weitesten verbreiteten Internet-Suchmaschine **Google**.

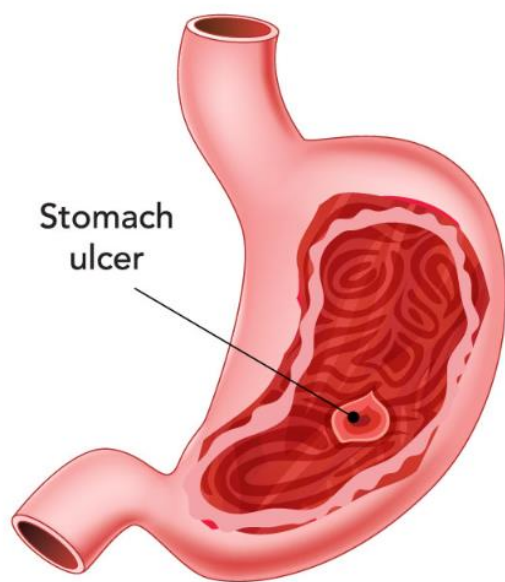
Retrieval in clinical data – example

Example:

Count alle KAgas patients in 2024
with a stomach ulcer
(confirmed diagnosis)

Problem:

- Not always coded (K25) in case of multimorbid patients
- May be coded (K25) in case of unconfirmed (treated for ulcer)



Simple free text search for "Magenulkus" in clinical narratives

- "florides Magenulcus"
- "Stressulkus im Corpusbereich":
- "Magenulkus kann nicht ausgeschlossen werden"
- "Familienanamnese: Vater Magenulkus"
- "Gabe von Omeprazol zur Prophylaxe eines Magenulkus"

→ false negative
→ false negative
→ false positive
→ false positive
→ false positive

Clinical data is heavily based on non-standard text

DM2

CABG n HWI

Bantobrazol

Polenallergie

Fischgerät im Hals

ZW. EBA PAE re. ,TVT. li

Acutes Coronar Syndrom

Nagelkranzfraktur Damen li

V. mors fel dig IV man dex

chron perfor. Gabla DD: Gabla CA

Bursitis Olecrani sin mit begl.Erisypel

Zst.n.Lux. Daumensattelgel.protese li.

Bulli-Trauma: Rippenserienfraktur links

Außenmeniscusscheibenmeniscusdeformität

Zahnex in AN bei Entwicklungsret. unkl.Genese

St. post Pneumonektomie links propter Neo Bronchi

Intervalle, in denen die Beschwerden nahezu beschwerdefrei sind

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

- spelling errors (systematic)
- typing errors (accidental)
- transcription errors
- spelling variations
- acronyms, often ambiguous
- ad-hoc-Compounds
- ad-hoc-Abbreviations
- local contexts
- idiosyncrasies, mannerisms
- Latinisms, "denglish"
- 2nd language professionals

Terminology standards provide preferred labels and few synonyms but clinicians speak and write differently



Frequency of SNOMED CT Preferred Terms and translations	Hits Google*
– English: "Secondary malignant neoplasm of liver"	100
– Swedish: "sekundär malign levertumör"	1
– German: "Sekundäre maligne Neoplasie der Leber"	1
Frequency of common synonyms	
– English: "liver metastases"	1,230,000
– Swedish: "levermetastaser"	217,000
– Swedish: "Lebermetastasen"	204,000

Similar findings in clinical text collection or literature abstracts:
e.g. no single hit for “Elektrokardiogramm” in 30.000 discharge letters from cardiology

*<https://www.google.com/search?q=%22Secondary+malignant+neoplasm+of+liver%22>

Semantic retrieval in unstructured clinical data

- Semantic retrieval: retrieving relevant content by **understanding the intent and meaning of queries**, rather than just matching words

Semantic retrieval in unstructured clinical data

- Semantic retrieval: retrieving relevant content by **understanding the intent and meaning of queries**, rather than just matching words
- Desiderata
 - Represent synonymous expressions by the same terminology codes
"U. ventriculi", "Magenulcus", "Magenulkus", "Magengeschwür" →

 Gastric ulcer (disorder)
SCTID: 397825006

Semantic retrieval in unstructured clinical data

- Semantic retrieval: retrieving relevant content by **understanding the intent and meaning of queries**, rather than just matching words
- Desiderata
 - Represent synonymous expressions by the same terminology codes
"U. ventriculi", "Magenulcus", "Magenulkus", "Magengeschwür" →
 - Identify contexts
"Omeprazol zur Prophylaxe eines Magenulkus"

☰ Gastric ulcer (disorder)
SCTID: 397825006

☰ Gastric ulcer (disorder)
SCTID: 3978250

unlikely

Semantic retrieval in unstructured clinical data

- Semantic retrieval: retrieving relevant content by **understanding the intent and meaning of queries**, rather than just matching words
- Desiderata
 - Represent synonymous expressions by the same terminology codes
"U. ventriculi", "Magenulcus", "Magenulkus", "Magengeschwür" →
 - Identify contexts
"Omeprazol zur Prophylaxe eines Magenulkus"
 - Disambiguate
"das Corpus-Ca wurde histologisch gesichert. Eine Hysterektomie wird empfohlen"

☰ Gastric ulcer (disorder)
SCTID: 397825006

☰ Gastric ulcer (disorder)
SCTID: 397825006

unlikely

☰ Malignant epithelial neoplasm of uterus (disorder)
SCTID: 446022000

Semantic retrieval in unstructured clinical data

- Semantic retrieval: retrieving relevant content by **understanding the intent and meaning of queries**, rather than just matching words
- Desiderata
 - Represent synonymous expressions by the same terminology codes
"U. ventriculi", "Magenulcus", "Magenulkus", "Magengeschwür" →
 - Identify contexts
"Omeprazol zur Prophylaxe eines Magenulkus"
 - Disambiguate
"das Corpus-Ca wurde histologisch gesichert. Eine Hysterektomie wird empfohlen"
 - Exploit taxonomic and ontological information
"Entlassmedikation: Urbason 16mg Tbl."

 Gastric ulcer (disorder)
SCTID: 397825006

 Gastric ulcer (disorder)
SCTID: 3978250

unlikely

 Malignant epithelial neoplasm of uterus (disorder)
SCTID: 446022000

 Product containing precisely 
methylprednisolone 16
milligram/1 each conventional release
oral tablet (clinical drug)
SCTID: 325413007

owl:subClassOf

 Product containing steroid
(product)

sct:hasActiveIngredient some

 Steroid (substance)
SCTID: 116566001

A query on "steroid" should retrieve documents with "Urbason 16mg Tbl."

Technologies and resources for semantic retrieval

- Terminologies / Ontologies / Information models
 - Terminologies: vocabularies that provide structured lists of domain terms in a certain (domain) language, assigning unique codes to synonyms and related expressions
 - Ontologies categorize and formally describe the entities of a domain
 - Information models are structured frameworks that integrate terminology/ontology codes within clinical contexts, defining how data should be represented, exchanged, and used (e.g., HL7 FHIR, openEHR)
- Text Mining
 - extracts meaningful information from text using natural language processing (NLP) techniques
- Machine Learning & AI
 - support in concept recognition, context analysis, disambiguation, syntactic and discourse analysis, relation detection, to construct precise semantic representations of narrative content

Typical use cases (IMI data management group)

Samples \ Queries	"Diabetes Typ 1"	"Diabetes Typ 2"	"Pankreaserkrankung"	"Nierenerkrankung"
"KHK II, St. p. 2 fach stenting 29.09.2018 Persist. VHF, Hypercholesterinämie, Statinmyopathie, Pankreaslipomatose HbA1c nd % kein Diabetes"				
"Kombinierte Hyperlipidämie DM Typ 2, sek insulinisiert inaktive Psychose seit 2000, med. Prophylaxe"				
"Diabetes mellitus Typ 2, sek insulinisiert Hypercholesterinämie, CKD IV" St.p.Anti Faktor VIII, Hemmkörperhämophilie ACE Hemmerunverträglichkeit Unverträglichkeit von Januvia Cardiale Stauungsleber				

IMI Semantics Group – Terminology creation and maintenance

- "Graz Interface Terminology":
 - Links everyday German clinical language to terminology standard SNOMED CT, by bridging the gap between free-text documentation and structured coding
 - Based on core vocabulary of about 440,000 short English-German term translations
 - 4,1700,000 German interface terms generated from core vocabulary covering 330,000 SNOMED CT concepts, prioritised by corpus token frequency and filtered by 1,539,000 negative patterns

- Schulz S, Rodrigues JM, Rector A, Chute CG. Interface Terminologies, Reference Terminologies and Aggregation Terminologies: A Strategy for Better Integration. Stud Health Technol Inform. 2017;245:940-944.
- Hashemian Nik D, Kasáč Z, Goda Z, Semlitsch A, Schulz S. Building an Experimental German User Interface Terminology Linked to SNOMED CT. Stud Health Technol Inform. 2019 Aug 21;264:153-157.

```
of pancreas <CNT> 913
of pancreas <SECOND> TRUE
of pancreas <TRANS> _Bauchspeicheldrüsen_
of pancreas <TRANS> _Pancreas_
of pancreas <TRANS> _Pankreas_
of pancreas <TRANS> der Bauchspeicheldrüse
of pancreas <TRANS> des Pancreas
of pancreas <TRANS> des Pankreas
of pancreas <TRANS> %LEFT% pankreatisches|JJ %LEFT%
of pancreas <TRANS> %LEFT% pankreatisches|JJ %LEFT%
```

[Term]
id: 123669009

 Pseudolipomatosis hypertrophy of pancreas (disorder)
SCTID: 123669009

```
name: Pseudolipomatosis hypertrophy of pancreas (disorder)
synonym: "Pseudolipomatosis hypertrophy of pancreas (disorder)" EXACT PREF []
synonym: "Pankreaslipomatose" EXACT []
synonym: "Pancreaslipomatose" EXACT []
synonym: "pseudolipomatöse Hypertrophie des Pankreas" EXACT []
synonym: "pseudolipomatöse Hypertrophie des Pancreas" EXACT []
synonym: "pseudolipomatöse Bauchspeicheldrüsenhypertrophie" EXACT []
synonym: "pseudolipomatöse Pankreashypertrophie" EXACT []
synonym: "pseudolipomatöse Pancreashypertrophie" EXACT []
synonym: "pseudolipomatöse Hypertrophie der Bauchspeicheldrüse" EXACT []
synonym: "pseudolipomatöse pankreatische Hypertrophie" EXACT []
synonym: "pseudolipomatöse pankreatische Hypertrophie" EXACT []
synonym: "pseudolipomatöse Pancreashypertrophie" EXACT []
```

Cooperation with ELGA GmbH – Austria's clinical data and standards hub



Med Uni Graz / News

KOOPERATION ZWISCHEN DER MED UNI GRAZ UND DER ELGA GMBH

22.05.2025

*Ab 2026 sollen in der ambulanten Versorgung in Österreich standardisierte Diagnosecodes eingeführt werden, was eine Herausforderung für Ärzt*innen darstellt. Im hektischen Praxisalltag bleibt kaum Zeit für aufwendige Recherchen nach dem passenden Code. Sekundenschnelle Codierung ist daher ein Muss.*

Der Austausch klinischer Daten soll in Österreich künftig durch Codes aus SNOMED CT erleichtert werden. SNOMED CT ist der weltweit umfangreichste Terminologiestandard im Gesundheitswesen. SNOMED CT enthält über 370.000 Codes und Beschreibungen für Erkrankungen, Befunde, Prozeduren, Wirkstoffe, Organismen und vieles mehr. Die deutschsprachige Übersetzung wird von der ELGA GmbH – als österreichisches SNOMED CT National Release Center (NRC) – in Zusammenarbeit mit den NRCs Deutschlands und der Schweiz betreut. Dank einer Lizenz des Bundesministeriums für Arbeit, Soziales, Gesundheit, Pflege und Konsumentenschutz (BMASGPK) ist die Nutzung von SNOMED CT kostenfrei.

Österreichweite Kooperation

Im Rahmen einer im Jahr 2024 gestarteten Kooperation mit dem Institut für Medizinische Informatik, Statistik und Dokumentation der Med Uni Graz baut die ELGA GmbH auf den Vorarbeiten und der Expertise des international anerkannten Terminologie-Experten Stefan Schulz auf, Universitätsprofessor an der Med Uni Graz für das Fach Medizinische Informatik. Ziel der Zusammenarbeit ist die SNOMED-Codierung von Diagnoseangaben – zunächst im ambulanten Bereich, perspektivisch auch in weiteren Sektoren des Gesundheitssystems.

Im Laufe von zwölf Jahren haben Stefan Schulz und sein Team Hunderttausende sprachliche Ausdrücke aus dem vor allem in Österreich, aber auch in Deutschland und der Schweiz gebräuchlichen medizinischen Fachjargon gesammelt und mithilfe teilautomatisierter Verfahren SNOMED-Codes zugeordnet. Darunter befinden sich zahlreiche in Arztbriefen und Befunden vorkommende Abkürzungen. Als „Graz Interface Terminology for SNOMED CT“ wird diese Ressource nun gemeinsam weiterentwickelt. Sie ist eine wichtige Grundlage für das vom BMASGPK entwickelte und betriebene e-Health-Codierservice. Dieses steht kostenfrei als Unterstützung für die strukturierte Erfassung von Diagnosen zur Verfügung. Für Softwarehersteller gibt es bereits einen technischen Demobetrieb.

Meilenstein des Gesundheitswesens

Engmaschige Anpassungen an die sich stetig weiterentwickelnde Medizinsprache erfordern eine langfristige Kooperation und den Einsatz moderner Technologien – darunter auch künstliche Intelligenz – zur Inhaltspflege und Qualitätssicherung. Für die ELGA GmbH ist dabei die kontinuierliche Einbindung der Ärzteschaft von zentraler Bedeutung. So konnten bereits in den ersten Projektphasen zentrale Anliegen der Österreichischen Ärztekammer, wie Benutzerfreundlichkeit, die Abbildung des medizinischen Sprachgebrauchs und das Prinzip „Dokumentieren statt Codieren“, berücksichtigt werden.

Stefan Schulz hebt hervor: „Die gemeinsame Arbeit an der Einführung von SNOMED CT ist ein Meilenstein für die Digitalisierung des österreichischen Gesundheitswesens und richtungsweisend für den gesamten deutschsprachigen Raum. SNOMED CT ermöglicht eine feingranulare und bedeutungserhaltende klinische Dokumentation über System- und Sprachgrenzen hinweg. Die Interface-Terminologie der Med Uni Graz bildet die Brücke zwischen Medizinjargon und einem internationalen Standard, der die Qualität der ärztlichen Dokumentation auf ein neues Level hebt. Das Gesundheitssystem ist durch die Optimierung der Zusammenarbeit der verschiedenen Akteur*innen besser für weitere Herausforderungen gerüstet. Die kontinuierliche Pflege und Weiterentwicklung dieser Terminologiesysteme werden die digitale Gesundheitslandschaft nachhaltig prägen.“

Weitere Informationen zu SNOMED CT, den aktuellen Entwicklungen in Österreich sowie den Forschungsarbeiten am Institut für Medizinische Informatik, Statistik und Dokumentation finden Sie unter: snomed.org; codierservice.ehealth.gv.at; elga.gv.at; und imi.medunigraz.at/forschung

Steckbrief: Stefan Schulz

Stefan Schulz ist ausgebildeter Humanmediziner mit Promotion in theoretischer Medizin, seit über 30 Jahren wissenschaftlich tätig und in zahlreichen internationalen und interdisziplinären Projekten im Bereich der medizinischen Informatik aktiv. Seit 2010 vertritt er dieses Fachgebiet als Universitätsprofessor am Institut für Medizinische Informatik, Statistik und Dokumentation der Medizinischen Universität Graz. Seine Forschung konzentriert sich auf die semantische Modellierung medizinischer Daten, den Einsatz künstlicher Intelligenz zur Analyse von Textinhalten in elektronischen Patient*innenakten sowie auf die Entwicklung und Anwendung biomedizinischer Standards wie SNOMED CT.

Kontakt

Univ.-Prof. Dr. med.
Stefan Schulz
Institut für Medizinische Informatik, Statistik und Dokumentation
Medizinische Universität Graz
T: +43 316 385 16939
E: [stefan.schulz\(at\)medunigraz.at](mailto:stefan.schulz(at)medunigraz.at)



Forschungsprofil >

Zur Presseinformation >

Stefan Schulz >

Institut für Medizinische
Informatik, Statistik und
Dokumentation >

SNOMED CT >

ELGA >

IMI Semantics Group – Use of text mining tool



■ Customization of Averbis Health Discovery

SNOMED
CT - Graz
Interface
Terminology

SCT-GIT

20240802

Übergabe an
Textanalyse
abgeschlossen (329097
/ 329097)

Texteingabe

Text Analyse Ergebnisse

DeidentifiedDocument

DocumentAnnotation

☒ LaboratoryParameter

☒ LaboratoryValue

search

Filter

LaboratoryParameter

LVEF

begin: 0

end: 4

conceptID: 10230-1

dictCanon: Linksventrikuläre Ejektionsfraktion

matched term: LVEF

source: Averbis-Lab-Terminology_2.3

uniqueID: Averbis-Lab-Terminology_2.3:10230-1

Texteingabe

Text Analyse Ergebnisse

Anatomy

☒ Concept

LaboratoryParameter

linksventrikuläre Ejektionsfraktion

Concept

linksventrikuläre Ejektionsfraktion

begin: 0

end: 35

conceptID: 250908004

dictCanon: Left ventricular ejection fraction (obse...

matchedTerm: linksventrikuläre Ejektionsfraktion

source: SNOMED CT - Graz Interface Terminology

uniqueID: SNOMED CT - Graz Interface Terminology:2...

LaboratoryValue

LVEF 59%

begin: 0

end: 8

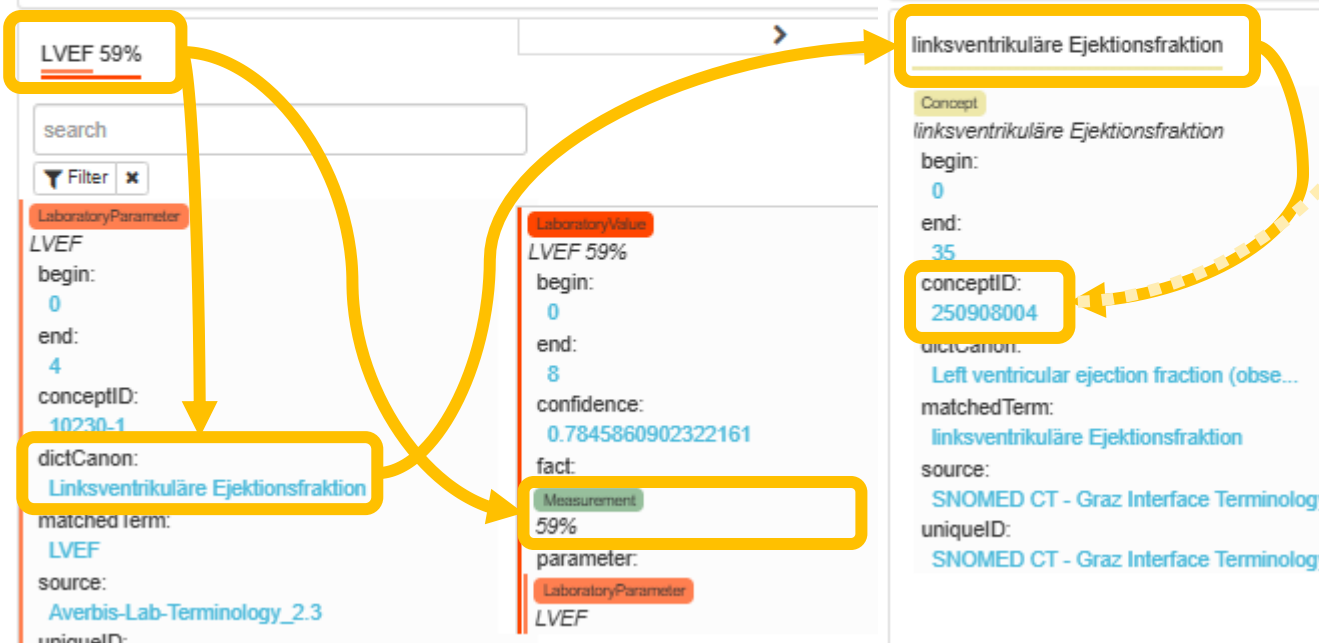
confidence: 0.7845860902322161

fact: Measurement

59%

parameter: LaboratoryParameter

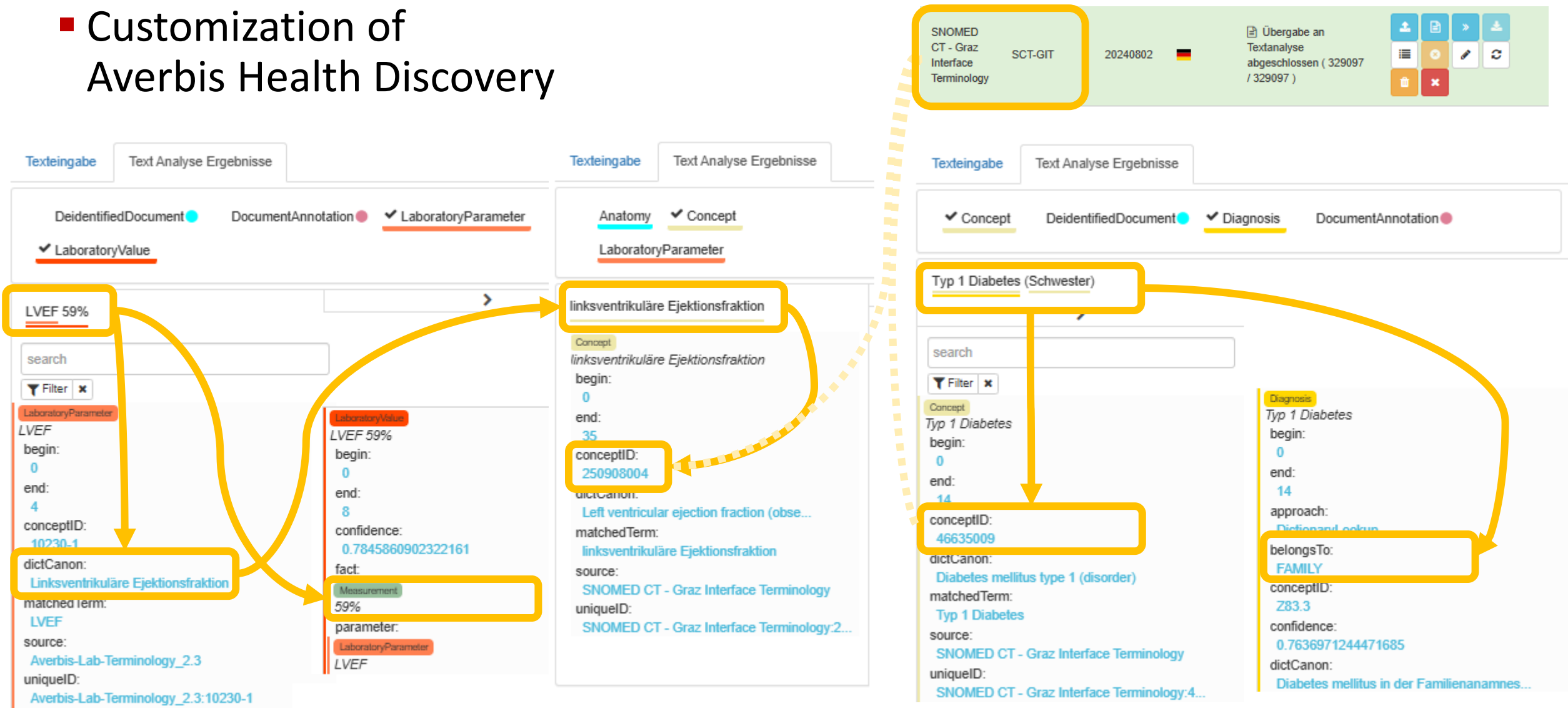
LVEF



IMI Semantics Group – Use of text mining tool

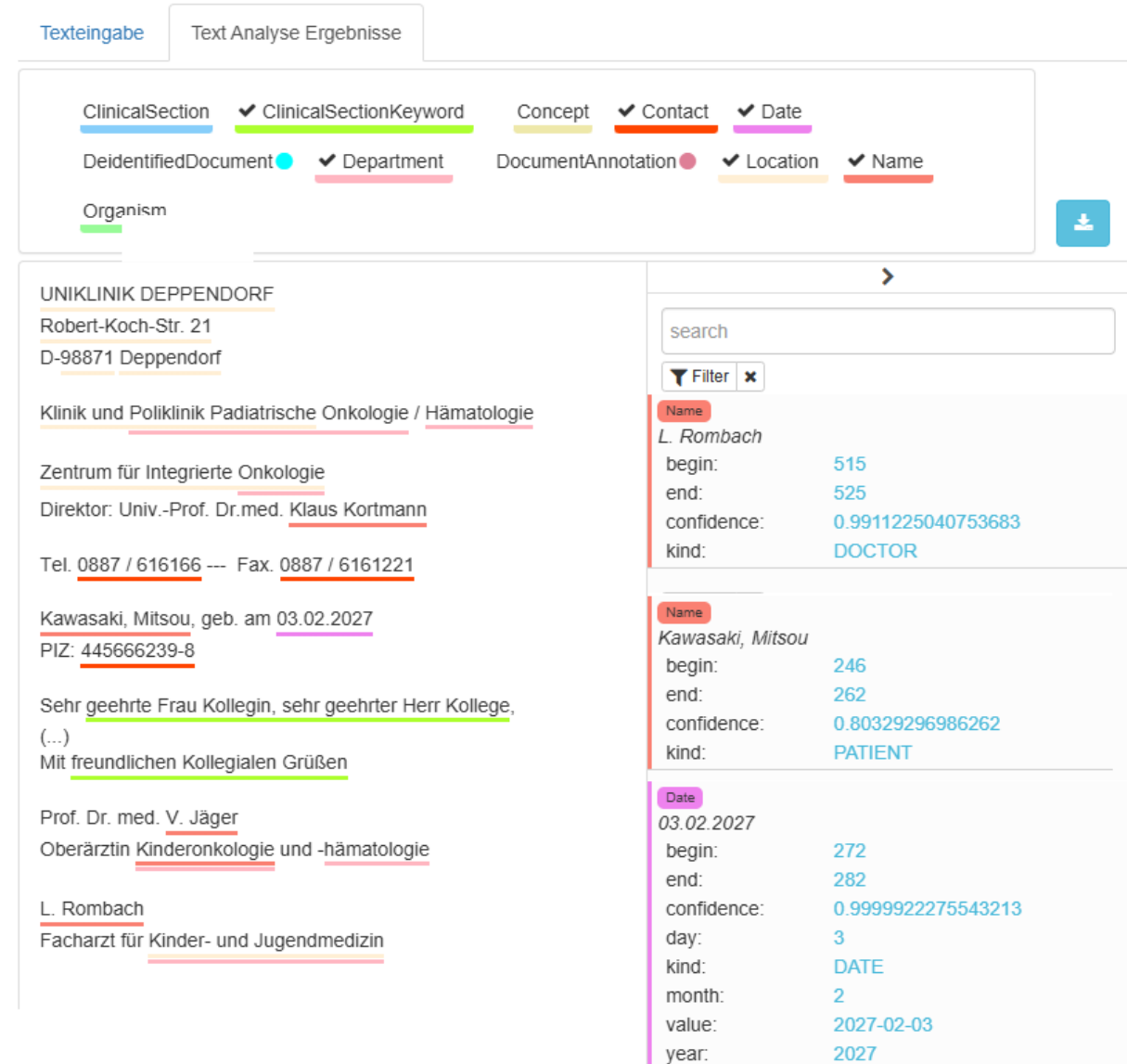


■ Customization of Averbis Health Discovery



IMI Semantics Group – Use of deidentification tool

- Customization of Averbis Health Discovery
- Deidentification:
 - Person (patient, physician)
 - Institutions
 - Addresses
 - ID



The screenshot displays the Averbis Health Discovery deidentification tool interface. It features a 'Texteingabe' (Text Input) tab and a 'Text Analyse Ergebnisse' (Text Analysis Results) tab. The interface includes a filter bar with various categories and their corresponding counts, a search bar, and a list of identified entities with their respective confidence scores and types.

Filter Bar:

- ClinicalSection (blue bar)
- ✓ ClinicalSectionKeyword (green bar)
- Concept (yellow bar)
- ✓ Contact (orange bar)
- ✓ Date (purple bar)
- DeidentifiedDocument (cyan bar)
- ✓ Department (pink bar)
- DocumentAnnotation (magenta bar)
- ✓ Location (orange bar)
- ✓ Name (red bar)
- Organism (green bar)

Search Bar: search

Filter: Filter x

Entities:

- Name:** L. Rombach
 - begin: 515
 - end: 525
 - confidence: 0.9911225040753683
 - kind: DOCTOR
- Name:** Kawasaki, Mitsou
 - begin: 246
 - end: 262
 - confidence: 0.80329296986262
 - kind: PATIENT
- Date:** 03.02.2027
 - begin: 272
 - end: 282
 - confidence: 0.9999922275543213
 - day: 3
 - kind: DATE
 - month: 2
 - value: 2027-02-03
 - year: 2027

Text Input:

UNIKLINIK DEPPENDORF
Robert-Koch-Str. 21
D-98871 Deppendorf

Klinik und Poliklinik Padiatrische Onkologie / Hämatologie

Zentrum für Integrierte Onkologie

Direktor: Univ.-Prof. Dr.med. Klaus Kortmann

Tel. 0887 / 616166 --- Fax. 0887 / 6161221

Kawasaki, Mitsou, geb. am 03.02.2027
PIZ: 445666239-8

Sehr geehrte Frau Kollegin, sehr geehrter Herr Kollege,
(...)
Mit freundlichen Kollegialen Grüßen

Prof. Dr. med. V. Jäger
Oberärztin Kinderonkologie und -hämatologie

L. Rombach
Facharzt für Kinder- und Jugendmedizin

Large Language Models for SNOMED coding



I am a clinical author who wants to improve the readability of clinical documents such as discharge summaries written by colleagues". Please do the following with the German language text snippet below:

- Assign SNOMED CT codes to each clinical concept in the following text sample:

"KHK II, St. p. 2 fach stenting 29.09.2018
Persist. VHF, Hypercholesterinämie, Statinmyopathie,
Pankreaslipomatose HbA1c nd % kein Diabetes
Medikation

Amiogamma 200 1-1-1"

Here is the improved text with SNOMED CT codes assigned to each clinical concept:

KHK II (Koronare Herzkrankheit Stadium II) – SNOMED CT: 275100004	
Zustand nach zweifachem Stenting am 29.09.2018 – SNOMED CT: 429937008 (Zustand nach Einsetzen eines Stents in die Koronararterie)	
Persistierendes Vorhofflimmern – SNOMED CT: 49436004	
Hypercholesterinämie – SNOMED CT: 13644009	
Statininduzierte Myopathie – SNOMED CT: 447682000	
Pankreaslipomatose – SNOMED CT: 302869004	
HbA1c nicht bestimmt – SNOMED CT: 165679005 (HbA1c-Wert nicht gemessen)	
Kein Diabetes mellitus – SNOMED CT: 302866003 (Kein Nachweis von Diabetes mellitus)	
Medikation:	
Amiodaron 200 mg 1-1-1 – SNOMED CT: 387174006 (Amiodaron)	

correct code

wrong code

code non-existing

Large Language Models for text cleansing

Rewriting and improving



I am a clinical author who wants to improve the readability of clinical documents such as discharge summaries written by colleagues". Please do the following with the German language text snippet below:

- use standard medical terminology such as used in medical literature
- substitute short forms by long forms
- fix writing and typing errors
- substitute brand names by ingredients they contain
- use German language

Here is the text sample:

"KHK II, St. p. 2 fach stenting 29.09.2018
Persist. VHF, Hypercholesterinämie, Statinmyopathie,
Pankreaslipomatose HbA1c nd % kein Diabetes
Medikation

Amiogamma 200 1-1-1"

Hier ist der überarbeitete Text mit standardisierter medizinischer Terminologie und verbesserter Lesbarkeit:

"Koronare Herzkrankheit (KHK) Stadium II, Zustand nach zweifachem Stenting am 29.09.2018.



Persistierendes Vorhofflimmern, Hypercholesterinämie, statininduzierte Myopathie,



Pankreaslipomatose.

Glykosylierter Hämoglobinwert (HbA1c) nicht bestimmt; kein Diabetes mellitus.



Medikation:

Amiodaron 200 mg 1-1-1"



Diese Version nutzt medizinisch präzise Begriffe, korrigiert Fehler und ersetzt den Markennamen durch den Wirkstoff. Lass mich wissen, falls du weitere Anpassungen benötigst!

LLMs for SNOMED matching using text cleansing + RAG

Received 12 September 2024, accepted 25 September 2024, date of publication 2 October 2024, date of current version 18 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3472500



Large Language Models for Clinical Text Cleansing Enhance Medical Concept Normalization

AKHILA ABDULNAZAR^{1,2}, ROLAND ROLLER³, STEFAN SCHULZ¹, AND MARKUS KREUZTHALER¹

¹Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, 8036 Graz, Austria

²CBmed GmbH—Center for Biomarker Research in Medicine, 8010 Graz, Austria

³German Research Center for Artificial Intelligence (DFKI), 10559 Berlin, Germany

Corresponding author: Markus Kreuzthaler (markus.kreuzthaler@medunigraz.at)

ABSTRACT Most clinical information is only available as free text. Large language models (LLMs) are increasingly applied to clinical data to streamline communication, enhance the accuracy of clinical documentation, and ultimately improve healthcare delivery. This study focuses on a corpus of anonymized clinical narratives in German. On the one hand it evaluates the use of ChatGPT for text cleansing, i.e., the automatic rephrasing of raw text into a more readable and standardized form, and on the other hand for retrieval-augmented generation (RAG). In both tasks, the final goal was medical concept normalization (MCN), i.e., the annotation of text segments with codes from a controlled vocabulary using natural language processing. We found that ChatGPT (GPT-4) significantly improves precision and recall compared to simple dictionary matching. For all scenarios, the importance of the underlying terminological basis was also demonstrated. Maximum F1 scores of 0.607, 0.735 and 0.754 (i.e., for top 1, 5 and 10 matches) were achieved through a pipeline including document cleansing, bi-encoder-based term matching based on a large domain dictionary linked to SNOMED CT, and finally re-ranking using RAG.

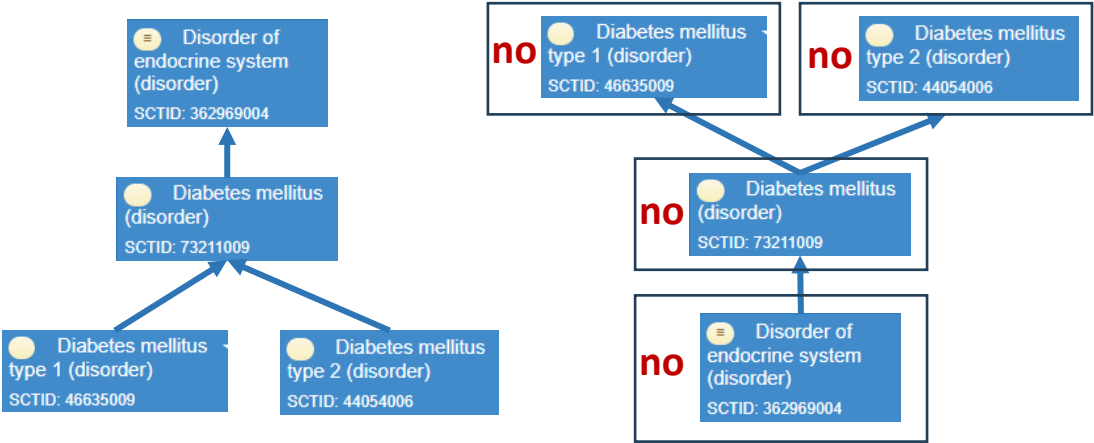
Method	Source	Data	P@1	R@1	F1@1
Dictionary matching	UMLS_DE	raw text	0.137	0.115	0.122
		cleansed text (GPT-4)	0.195	0.173	0.178
	IT_DE	raw text	0.330	0.284	0.296
		cleansed text (GPT-4)	0.339	0.282	0.297
Bi-encoder	UMLS_DE	raw text	0.252	0.232	0.232
		cleansed text (GPT-4)	0.306	0.285	0.286
	IT_DE	raw text	0.593	0.523	0.542
		cleansed text (GPT-4)	0.618	0.552	0.568
RAG	UMLS_DE	raw text	0.253	0.244	0.241
		cleansed text (GPT-4)	0.321	0.300	0.298
	IT_DE	raw text	0.615	0.558	0.572
		cleansed text (GPT-4)	0.646	0.595	0.607

- UMLS_DE: German terms and SNOMED CT mapping via the UMLS Metathesaurus (dictionary baseline)
- IT_DE: Graz Interface Terminology for SNOMED CT
- Bi-Encoder: SapBERT model (transformer)
- RAG: Retrieval-augmented generation using GPT

Outlook

- Bringing Semantics into routine
 - Work with IMI Data Management team
 - Practical use of text mining tool and interface terminology
 - Important feedback re clinical language and use cases
 - Combination of several indexing strategies
 - Interesting challenges of reasoning via taxonomies for optimized retrieval result

Q1	Q2	Q3	Q4	Q5	Q6	DOKUMENT_NR	DOKUMENT	Diagnose	Relevanter Befund
	Diabetes Typ 2	Body Mass Index				000000001000000005858412	00	Diabetes mellitus Typ 2 Adipositas III Kombinierte Hyperlipidaemie Leberfunktionsstörung Idiopathisches Lymphödem	Labor: BMI 41.5 kg/m² 170 cm, 120 kg : Cholesterin 193 mg/dl
		Body Mass Index				000000001000000005855186	00	Wohl familiäre Hypercholesterinaemie KHK, Z. n. STEMI, PCI & STENT LAD/ALA 2016 Ischaemische CMP, NYHA II St.p. Ventrikeltrombus, NAK Therapie Infarctuelles Aortenaneurysma ZAVK I	Labor: BMI 27.0 kg/m² 169 cm, 77 kg : Cholesterin 93 mg/dl
		Body Mass Index				000000001000000005855118	00	Adipositas I Diabetes mellitus 2b arterieller Hypertonus Hypercholesterinaemie Leberfunktionsstörung milde Hyperuricaemie	Labor: BMI 31.4 kg/m² 160 cm, 80.5 kg (-2kg) : Cholesterin 164 mg/dl
Diabetes Typ 1		Body Mass Index				000000001000000005854942	00	Diabetes mellitus Typ 1 Hypercholesterinaemie, Lp(a)-Erhöhung arterieller Hypertonus, Mikroalbuminurie LPS St.p. TE, St.p. Fract calcanei sin 10/2001 RPD proliferans mit Amaurose rechts, stabil	Labor: BMI 26.5 kg/m² 176 cm, 82 kg : Cholesterin 158 mg/dl
		Body Mass Index				000000001000000005854684	00	Kombinierte Hyperlipidaemie Lp(a) Erhöhung Diabetes mellitus 2b seit 2010 Adipositas I arterieller Hypertonus Stabile AP bei KHK II und St. p. 4x ACBG, St. p.	Labor: BMI 30.4 kg/m² 171 cm, 89 kg : Cholesterin 162 mg/dl
		Body Mass Index	kein Typ 2 Diabetes			000000001000000005853064	00	<Stoffwechseldiagnosen: </> Hypercholesterinaemie KHK, St.p. LAD Stent 2015 Coronasklerose bds. Statintunverträglichkeit Nikotin negativ	Labor: BMI kg/m² : Cholesterin mg/dl
		Body Mass Index	kein Typ 2 Diabetes			000000001000000005852832	00	<Stoffwechseldiagnosen: </> Familiäre Hypercholesterinaemie mit hochpositiver Familienanamnese KHK II mit deutlicher Progredienz von 2014 bis 2018 (Coronar CT)	Labor: BMI kg/m² Cholesterin mg/dl



- Research
 - Use and validation of language models for several purposes
 - Combination of ML with ontologies
 - Representation of clinical document content as knowledge graphs

LLM – Medical prompting examples

Terminology management

I am a medical domain expert, lexicographer and translator. Please find good and clinically common translations of

"drug delivery"

to German

Don't add any comment.

I am a medical domain expert, lexicographer and translator. A generative approach to terminology translation creates for the English term

"drug delivery"

the German translations

"Wirkstoffentbindung"

"Drogenentbindung"

"Medikamentenentbindung"

"Wirkstoffgabe"

"Drogengabe"

"Medikamentengabe"

"Medikamentenverabreichung"

"Wirkstoffverabreichung"

"Drogenverabreichung"

Please identify good, bad, and borderline translations and list them!

Don't add any comment.

Clinical text preprocessing

I am a medical domain expert and I would like you to make clinical texts more readable. A big problem with clinical texts is their compact form with co-ordinations, anaphors and listed items of which the head word needs to be identified.

Please rewrite the following text, aiming at separated small sentences with subject, predicate, object. Rewrite each concept as precise as possible, and avoid anaphora and coordinations.

Please rewrite in German:
Familienanamnese:
Mutter:
- Diabetes mellitus Typ 2
- Demenz
Metabolisches Syndrom
Vater:
- Bronchial-Ca
- COPD

I am a medical domain expert and I would like you to make clinical texts more readable. A big problem with clinical texts is their compact form with co-ordinations, anaphors and listed items of which the head word needs to be identified.

Please rewrite the following text, aiming at separated small sentences with subject, predicate, object. Rewrite each concept as precise as possible, and avoid anaphora and coordinations.

Please rewrite in German:
Hämatome:
Oberschenkel links
- Präpatellar links
- Unter dem linken Rippenbogen
- linker Ellenbogen
- Linke Schulter

Abschürfungen:
- Handrücken links
- Handgelenk links
- Unterarm links

Keine Auffälligkeiten:
- Abdomen
- Thorax
- Kopf/Hals