# Case Study: Natural-Language-Processing (NLP) with Open Data for Drug Repositioning on Glioblastoma Therapy

**Curdin Marxer** (DAViS Center, University of Applied Sciences of the Grisons)
Heiko Rölke (DAViS Center, University of Applied Sciences of the Grisons)
Alex Alfieri (Department of Neurosurgery, Cantonal Hospital of Winterthur)
Marc-Eric Halatsch (Department of Neurosurgery, Cantonal Hospital of Winterthur)

## Agenda

- Introduction
- Methodology
- Results & Discussion
- Limitations & Further research

# Speaker Introduction

**Curdin Marxer**

- BSc. Information Science
- MSc. in Business Administration: Information and Data Management
- Research assistant at the Centre for Data Analytics, Visualization and Simulation (DAViS) of the University of Applied Sciences of the Grisons in Chur, Switzerland

Introduction

# Background of our research

- Part of a collaborate research project between the DAViS Centre and two neurosurgeons of the Cantonal Hospital of Winterthur
- CUSP9v3 - an innovative combined regimen of nine repurposed non-oncological drugs with metronomic temozolomide for the treatment of glioblastoma (malignant brain tumor)
- Our overarching research goal is to predict new possible drug candidates for repositioning by harvesting multiple different data sources using ML/DL techniques
- This research started as a master's thesis with the main research question:

***How can unstructured text data be used to identify new drug repositioning candidates and how can they complement databases?***

# Drug development and the role of drug repositioning

- The development and discovery of new drugs is costly, risky and takes a lot of time
- The success rate of a newly developed drug to be clinically approved and to reach world-wide markets is below 10%
- Many newly developed medical compounds end up abandoned

- **Drug repositioning/repurposing** describes the process of identifying and developing new uses for already existing drugs or known active compounds
- Discover new use cases of known drugs by harvesting knowledge of previous research
- Highly beneficial (commercially and for patients), especially for rare diseases

# Some strategies for drug repositioning

- **side-effect-based approaches**
  - based on the idea that therapeutically observed side effects of already developed drugs can also provide information about possible alternative uses
- **similarity-based approaches**
  - similarity between two drugs can be determined based on the culmination of multiple similarities in chemical structure, overlap of molecular target structures and side effects.
  - similarities between diseases can be concluded through ontologies or shared treatment profiles

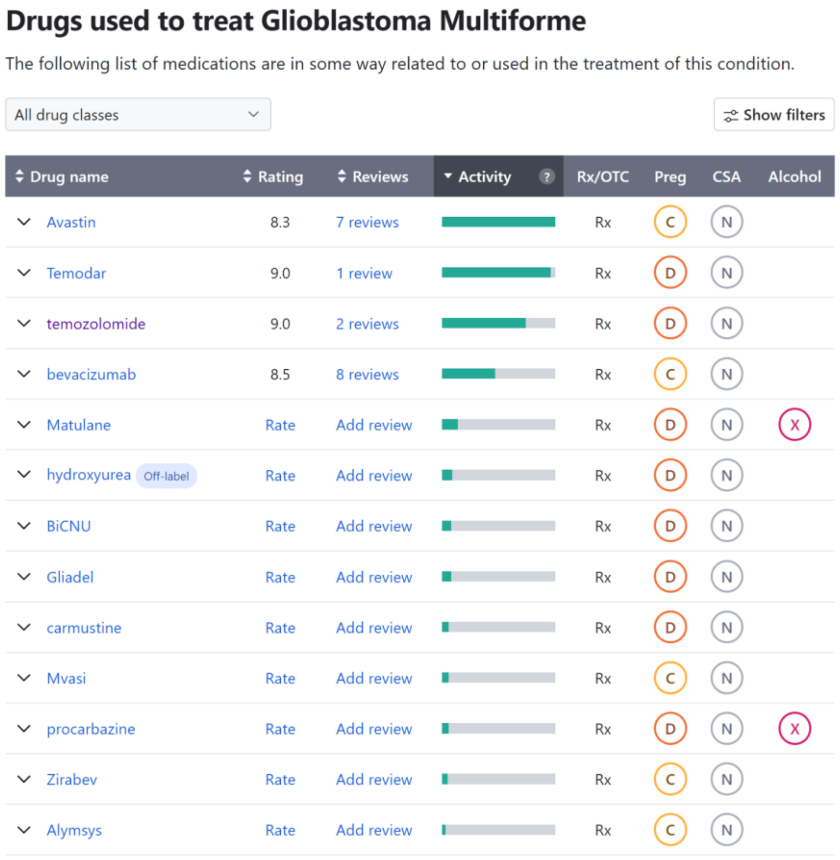
Introduction

# Challenges of working with drug data

- Highly complex and strongly interconnected data
- Amounts of biomedical data on drugs and the number of available repositories and databases are constantly increasing
- Data of these repositories or databases differ significantly in terms of scope, quality and reliability – causing inconsistencies
- Raise the challenge in selecting the adequate database(s) containing the required information

Introduction

# Examples of these inconsistencies based on our Case Study

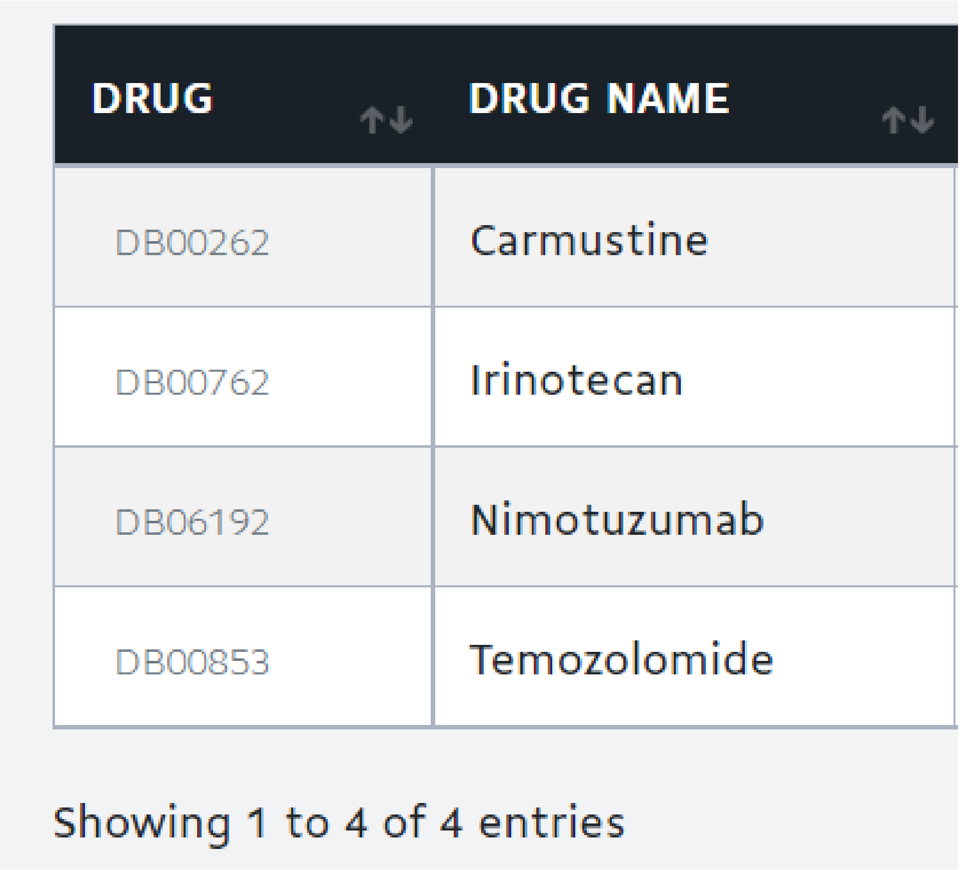


## Drugs used to treat Glioblastoma Multiforme

The following list of medications are in some way related to or used in the treatment of this condition.

| Drug name | Rating | Reviews | Activity | Rx/OTC | Preg | CSA | Alcohol |
|---|---|---|---|---|---|---|---|
| Avastin | 8.3 | 7 reviews | | Rx | C | N | |
| Temodar | 9.0 | 1 review | | Rx | D | N | |
| temozolomide | 9.0 | 2 reviews | | Rx | D | N | |
| bevacizumab | 8.5 | 8 reviews | | Rx | C | N | |
| Matulane | Rate | Add review | | Rx | D | N | X |
| hydroxyurea Off-label | Rate | Add review | | Rx | D | N | |
| BiCNU | Rate | Add review | | Rx | D | N | |
| Gliadel | Rate | Add review | | Rx | D | N | |
| carmustine | Rate | Add review | | Rx | D | N | |
| Mvasi | Rate | Add review | | Rx | C | N | |
| procarbazine | Rate | Add review | | Rx | D | N | X |
| Zirabev | Rate | Add review | | Rx | C | N | |
| Alymsys | Rate | Add review | | Rx | C | N | |

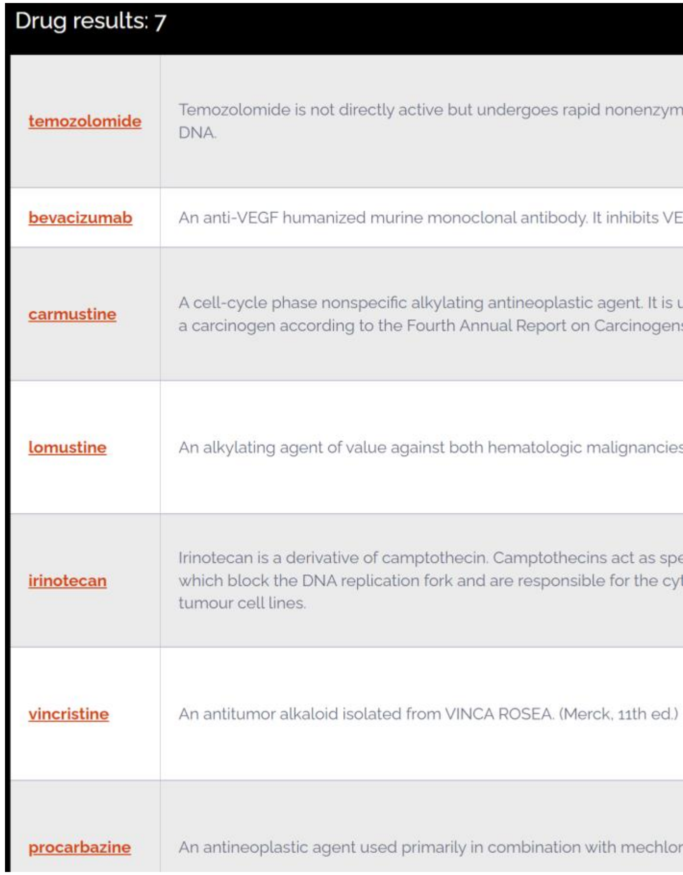
drugs.com



| DRUG | DRUG NAME |
|---|---|
| DB00262 | Carmustine |
| DB00762 | Irinotecan |
| DB06192 | Nimotuzumab |
| DB00853 | Temozolomide |

Showing 1 to 4 of 4 entries

go.drugbank.com
(4 “drug” indications and 316 “drug trial” indications)



Drug results: 7

| | |
|---|---|
| temozolomide | Temozolomide is not directly active but undergoes rapid nonenzym<br>DNA. |
| bevacizumab | An anti-VEGF humanized murine monoclonal antibody. It inhibits VE |
| carmustine | A cell-cycle phase nonspecific alkylating antineoplastic agent. It is u<br>a carcinogen according to the Fourth Annual Report on Carcinogens |
| lomustine | An alkylating agent of value against both hematologic malignancies |
| irinotecan | Irinotecan is a derivative of camptothecin. Camptothecins act as spe<br>which block the DNA replication fork and are responsible for the cyt<br>tumour cell lines. |
| vincristine | An antitumor alkaloid isolated from VINCA ROSEA. (Merck, 11th ed.) |
| procarbazine | An antineoplastic agent used primarily in combination with mechlor |

drugcentral.org

FH
GR

Introduction

# The potential of text data

- A huge amount of exclusive medical knowledge is hidden in various types of unstructured text data such as clinical reports, scientific research documents or journals
- Especially clinical reports provide new knowledge that is not yet recorded in standardized databases or summarized medical literature, e.g., on new side effects of individual drugs or drug-drug interactions
- Potential buried information on abandoned medical compounds

## Goal of this research

- Test and evaluate different approaches and methods to predict new drug repositioning candidates using NLP on open and publicly available unstructured text data
- Determine the potential of unstructured text data to combat database inconsistencies by filling existing data gaps
- Development of two different workflows to identify drug repositioning candidates for possible therapeutic use for glioblastoma
  - ➢ **Combined use of pre-trained biomedical Named Entity Recognition (NER) models to identify possible relationships between biomedical entities**

**Methodology**

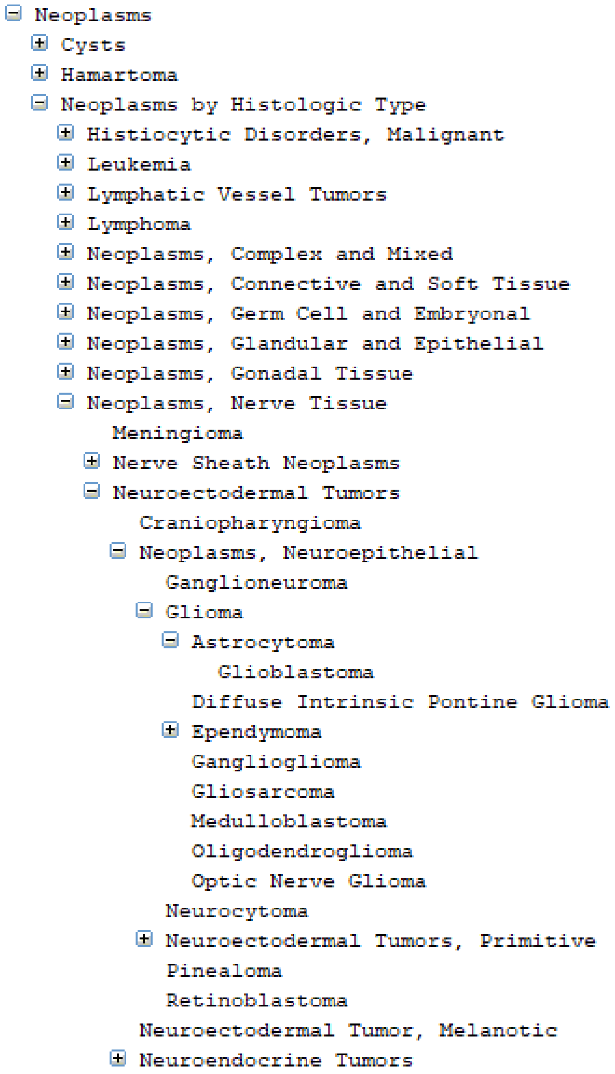
# Tools

- ScispaCy v0.5.1
- Provides fast, easy-to-use and robust biomedical NER-models
- Offers four different specialized biomedical NER models which enable a wide subject-specific scope of applications

Methodology

# Selection of open text data

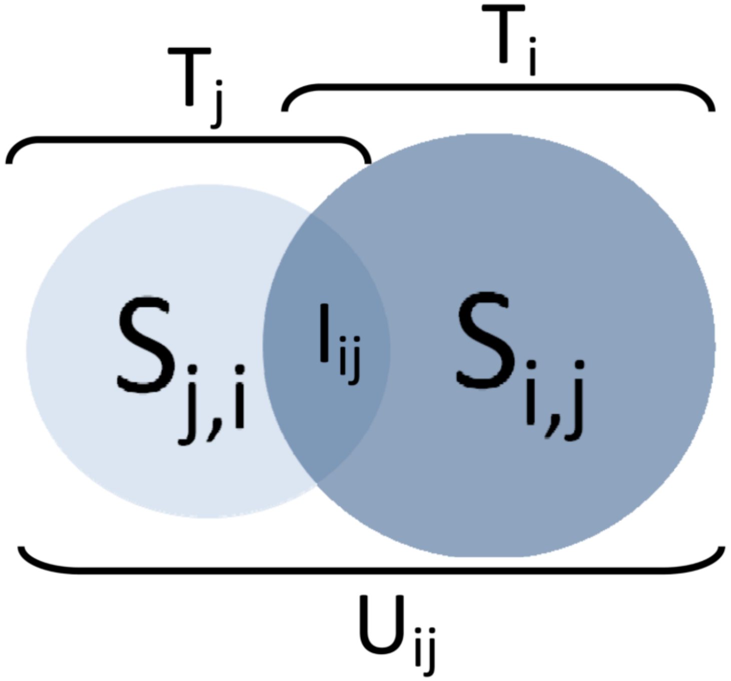
- Extracted by using the term “Neuroectordermal Tumors” of the controlled vocabulary MeSH
- 6’741 clinical studies from ClinicalTrials.gov
- Supplemented with 3’259 abstracts from PubMed
- resulting test data set of 10’000 documents



```
⊟ Neoplasms
  ⊞ Cysts
  ⊞ Hamartoma
  ⊟ Neoplasms by Histologic Type
    ⊞ Histiocytic Disorders, Malignant
    ⊞ Leukemia
    ⊞ Lymphatic Vessel Tumors
    ⊞ Lymphoma
    ⊞ Neoplasms, Complex and Mixed
    ⊞ Neoplasms, Connective and Soft Tissue
    ⊞ Neoplasms, Germ Cell and Embryonal
    ⊞ Neoplasms, Glandular and Epithelial
    ⊞ Neoplasms, Gonadal Tissue
    ⊟ Neoplasms, Nerve Tissue
        Meningioma
      ⊞ Nerve Sheath Neoplasms
      ⊟ Neuroectodermal Tumors
          Craniopharyngioma
        ⊟ Neoplasms, Neuroepithelial
            Ganglioneuroma
          ⊟ Glioma
            ⊟ Astrocytoma
                Glioblastoma
              Diffuse Intrinsic Pontine Glioma
            ⊞ Ependymoma
              Ganglioglioma
              Gliosarcoma
              Medulloblastoma
              Oligodendroglioma
              Optic Nerve Glioma
            Neurocytoma
          ⊞ Neuroectodermal Tumors, Primitive
            Pinealoma
            Retinoblastoma
          Neuroectodermal Tumor, Melanotic
        ⊞ Neuroendocrine Tumors
```

# Method 1: Co-occurrence analysis

- Method 1: Co-occurrence analysis based on the “Guilt by Association (GBA)” principle
- Using the co-occurrence of NER-extracted biomedical entities to determine most similar document pairs
- Use these most similar document pairs to extract and associate new drug-disease pairs



"Guilt by Association" scheme for the discovery of new uses for known drugs (Chiang & Butte, 2009)

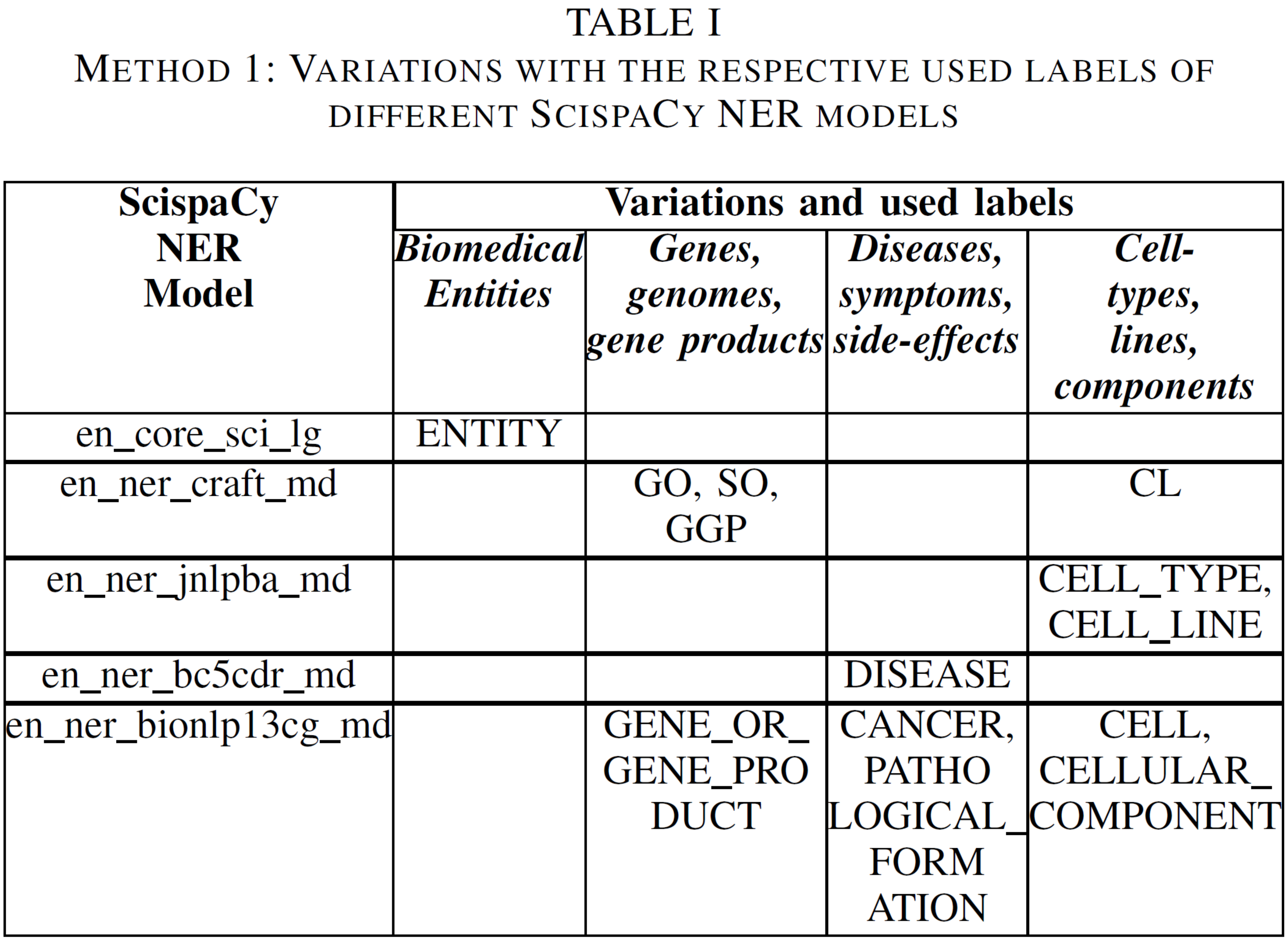


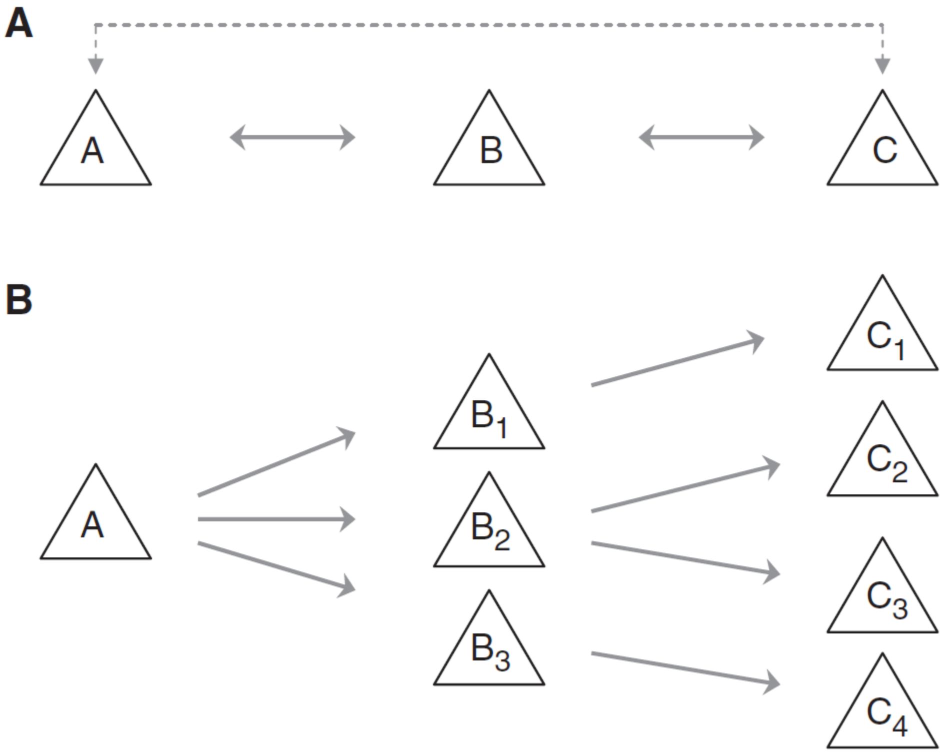
TABLE I

METHOD 1: VARIATIONS WITH THE RESPECTIVE USED LABELS OF DIFFERENT SCISPACY NER MODELS

| **ScispaCy NER Model** | **Variations and used labels** | | | |
|---|---|---|---|---|
| | ***Biomedical Entities*** | ***Genes, genomes, gene products*** | ***Diseases, symptoms, side-effects*** | ***Cell-types, lines, components*** |
| en_core_sci_lg | ENTITY | | | |
| en_ner_craft_md | | GO, SO, GGP | | CL |
| en_ner_jnlpba_md | | | | CELL_TYPE, CELL_LINE |
| en_ner_bc5cdr_md | | | DISEASE | |
| en_ner_bionlp13cg_md | | GENE_OR_GENE_PRODUCT | CANCER, PATHOLOGICAL_FORMATION | CELL, CELLULAR_COMPONENT |

# Method 2: Chains of association

- Method 2: Chains of association using “Swanson’s ABC-model”
- By building association chains to identify new possible transitive relations to determine new repurposing candidates
- Extraction of A-B relations from state-of-the-art biomedical databases
- Using extracted B-terms as search terms for the text data, identify hit documents, determine entity type C as possible repurposing candidate for A



"Swanson's ABC model" (Andronis et al., 2011)

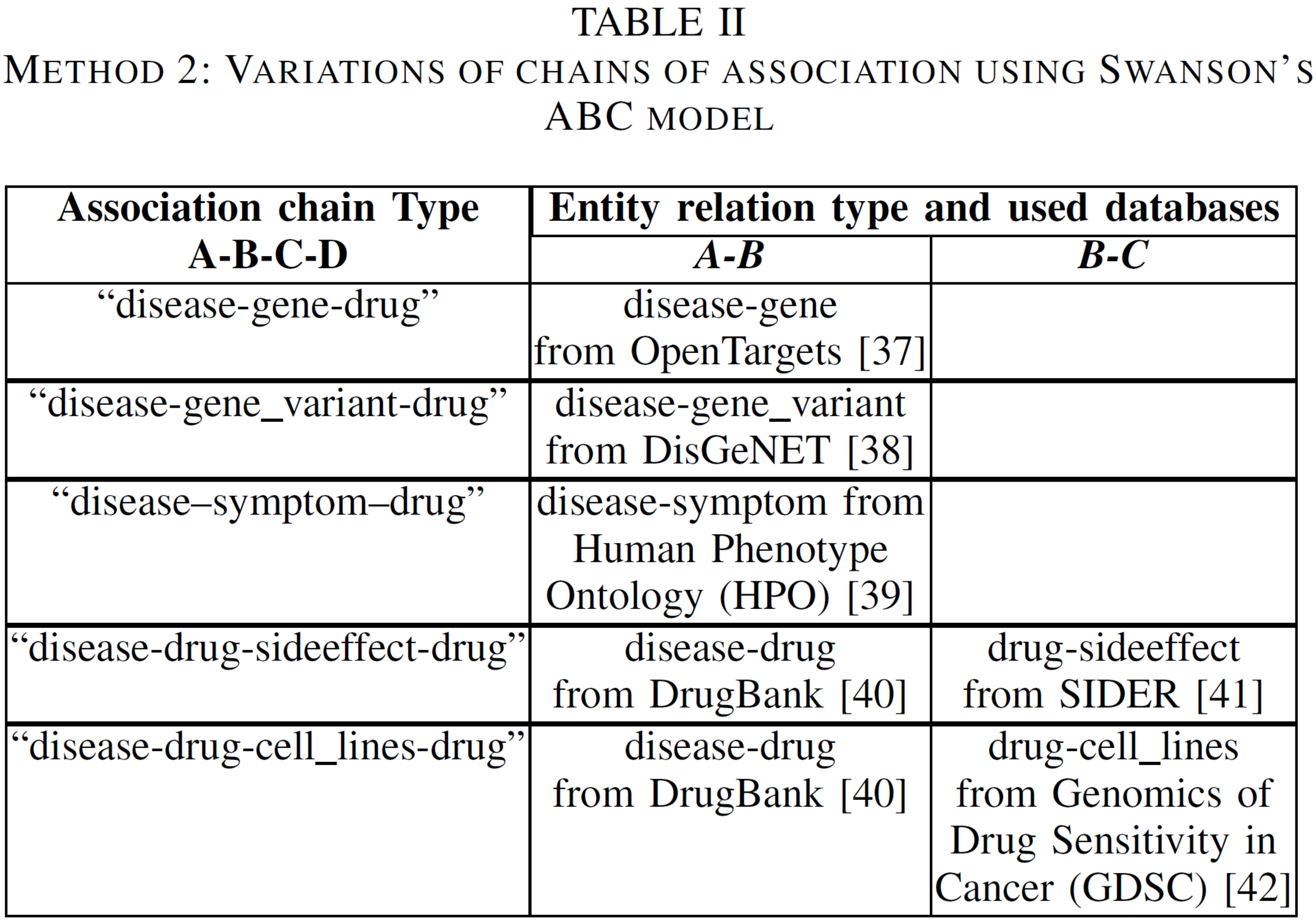


TABLE II
METHOD 2: VARIATIONS OF CHAINS OF ASSOCIATION USING SWANSON'S ABC MODEL

| **Association chain Type A-B-C-D** | **Entity relation type and used databases** | |
|---|---|---|
| | ***A-B*** | ***B-C*** |
| "disease-gene-drug" | disease-gene from OpenTargets [37] | |
| "disease-gene_variant-drug" | disease-gene_variant from DisGeNET [38] | |
| "disease–symptom–drug" | disease-symptom from Human Phenotype Ontology (HPO) [39] | |
| "disease-drug-sideeffect-drug" | disease-drug from DrugBank [40] | drug-sideeffect from SIDER [41] |
| "disease-drug-cell_lines-drug" | disease-drug from DrugBank [40] | drug-cell_lines from Genomics of Drug Sensitivity in Cancer (GDSC) [42] |

Results & Discussion

# Our predicted candidates for repurposing

- Types and examples of results as predicted candidates for repurposing on glioblastoma therapy:
  - **Chemical elements:** calcium, indium
  - **Chemical compounds:** most observed type, e.g., O6-benzylguanine
  - **Experimental vaccines:** “DNX-2401” (tasadenoturev)
  - **Hormones:** estrogen, steroids
  - **Various therapeutics:** TT-Fields

Results & Discussion

# Our evaluation of candidates

- Predicted repurposing candidates labeled in three categories for evaluation:
  - **“Known in DrugBank”**
  - **“Potential unknown candidates”**
  - **“Invalid”**

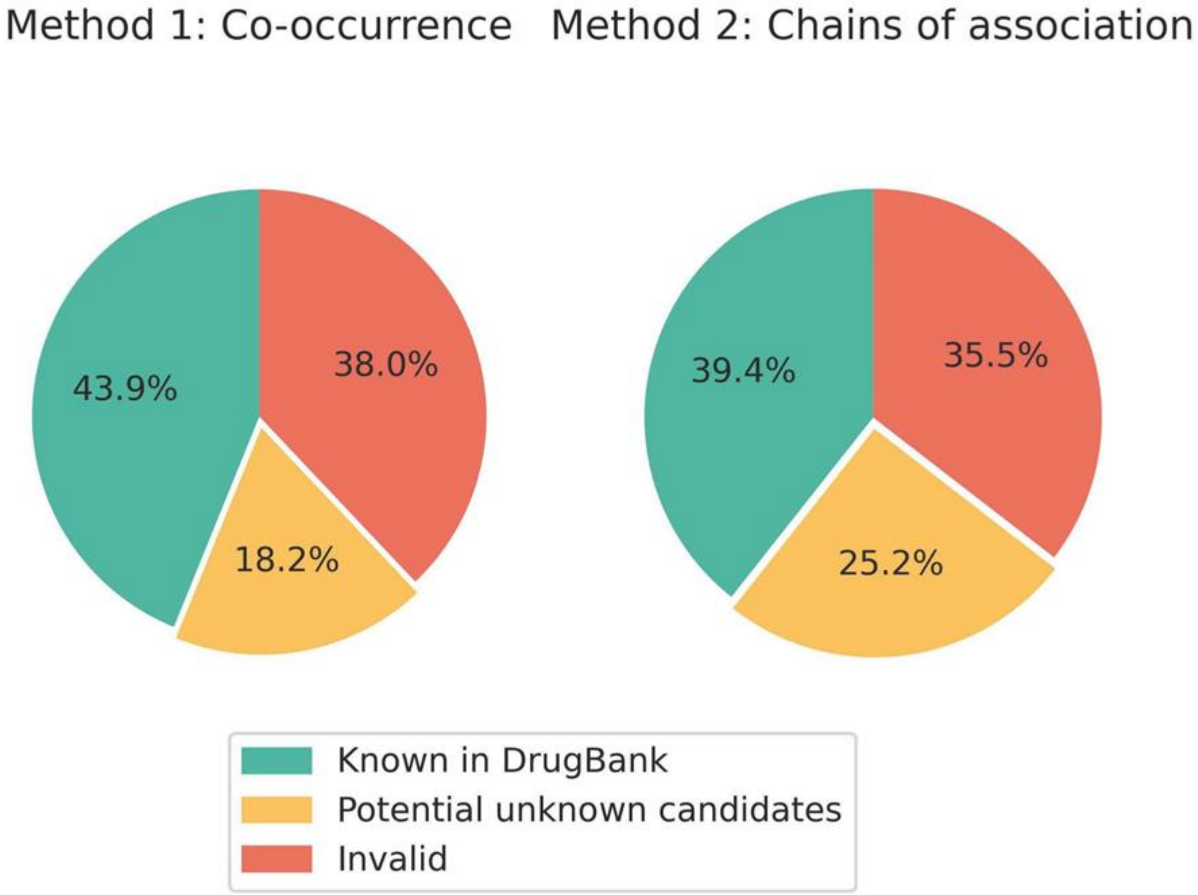
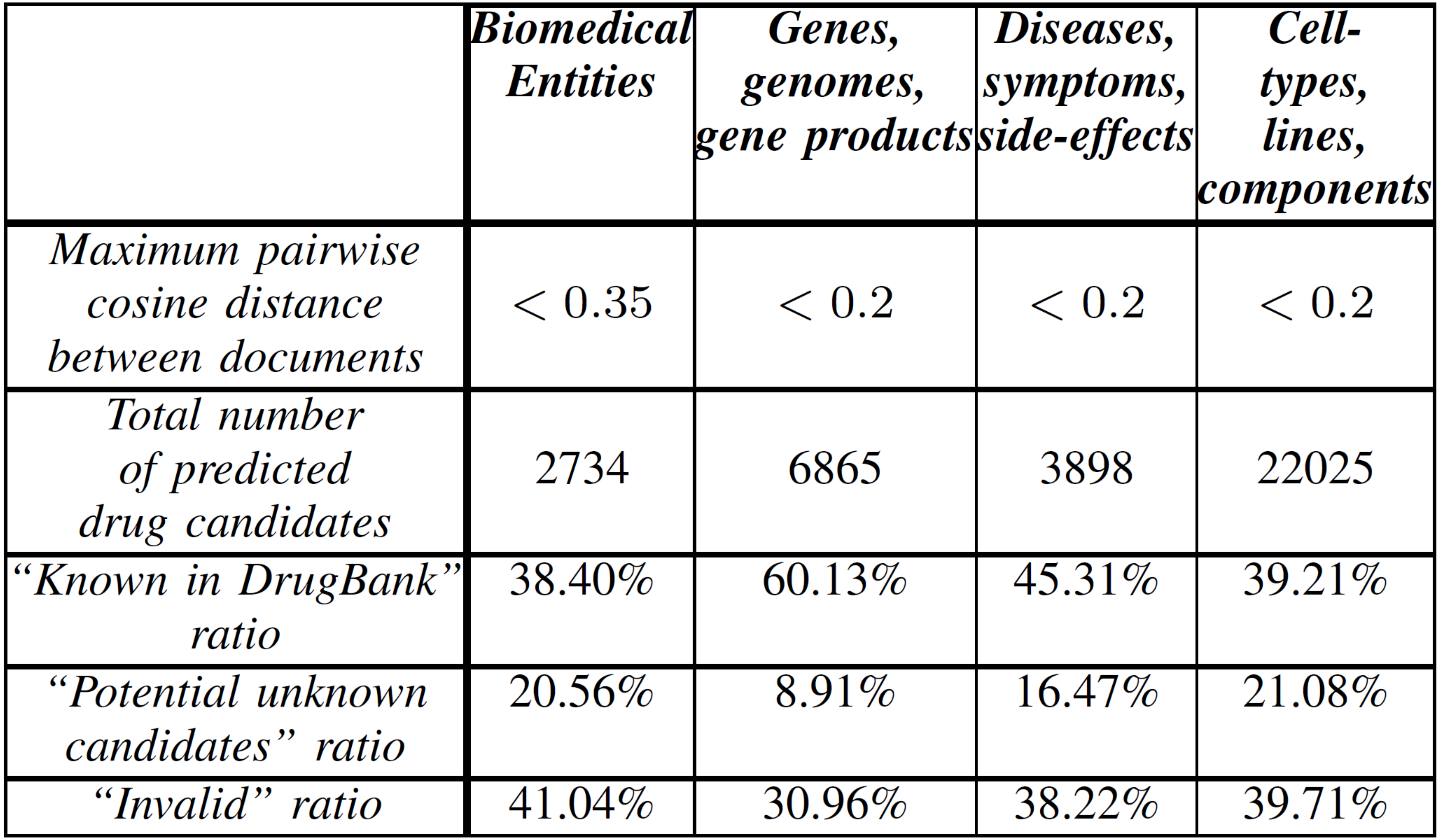


TABLE III

METHOD 1: SUMMARY OF RESULTS OF ALL EMPLOYED VARIATIONS



| | ***Biomedical Entities*** | ***Genes, genomes, gene products*** | ***Diseases, symptoms, side-effects*** | ***Cell-types, lines, components*** |
|---|---|---|---|---|
| *Maximum pairwise cosine distance between documents* | $< 0.35$ | $< 0.2$ | $< 0.2$ | $< 0.2$ |
| *Total number of predicted drug candidates* | 2734 | 6865 | 3898 | 22025 |
| *"Known in DrugBank" ratio* | 38.40% | 60.13% | 45.31% | 39.21% |
| *"Potential unknown candidates" ratio* | 20.56% | 8.91% | 16.47% | 21.08% |
| *"Invalid" ratio* | 41.04% | 30.96% | 38.22% | 39.71% |

## TABLE IV
METHOD 2: SUMMARY OF RESULTS OF ALL EMPLOYED VARIATIONS

| | ***“disease-gene-drug”*** | ***“disease-gene_variant-drug”*** | ***“disease-symptom-drug”*** | ***“disease-drug-sideeffect-drug”*** | ***“disease-drug-cell_line-drug”*** |
|---|---|---|---|---|---|
| *Number of extracted search terms from the selected database* | 118 | 291 | 24 | 200 | 68 |
| *Total number of predicted drug candidates* | 2226 | 8 | 975 | 7021 | 47 |
| *“Known in DrugBank” ratio* | 38.63% | 0% | 39.90% | 39.52% | 44.68% |
| *“Potential unknown candidates” ratio* | 20.44% | 37.50% | 24.51% | 26.81% | 14.89% |
| *“Invalid” ratio* | 40.93% | 62.50% | 35.59% | 33.67% | 40.43% |

# Some conclusions

- The analysis of unstructured text data can enable a more comprehensive overview of potential repurposing candidates
- Case examples, like "Cixutumumab" and "Paracetamol" were identified as possible repurposing candidates. However, these compounds are currently not listed in DrugBank, although they have been part of clinical trials for the treatment of glioblastoma
- Using overlapping genes as markers showed most reliable potential candidates for repositioning through our methods and use case
- Rarely any gene variants or cell-lines were documented in our selected text data, so only symptoms, side-effects or genes seem suitable as possible embeddings for future analyses of text data

# Limitations

- Publicly available clinical text data are mostly only provided as short summaries or abstracts
- Lack of an objective qualitative assessment of our identified repositioning candidates
- Vague association approach, prone to false positives
  - most associations were not analyzed based on their exact semantic connections, such as their possible causalities and their positive or negative relationships
- Poor performance in precision and recall of ScispaCy NER-models in comparison to models from other biomedical NER-Taggers, e.g., “Stanza” (StanfordNLP) or “SparkNLP”

# Further research

- Use of better specialized NER models, continued training of available models on new data via Transfer Learning through SparkNLP
- Focus on the use of more full text data from PubMed-Central, scope not solely limited to “Neuroectodermal tumors” for “soft-repurposing”
- Extract drug combinations and regimen and profiling them with scores based on multiple data sources on side-effects, drug-drug interactions and drug-characteristics

**University of Applied Sciences of the Grisons**
Pulvermühlestrasse 57
7000 Chur
T +41 81 286 24 24
info@fhgr.ch

# Thank you very for your attention.
# Feel free to ask any questions!

Fachhochschule Graubünden
Scola auta spezialisada dal Grischun
Scuola universitaria professionale dei Grigioni
University of Applied Sciences of the Grisons

swissuniversities

SCHWEIZERISCHER AKKREDITIERUNGSRAT
CONSEIL SUISSE D'ACCRÉDITATION
CONSIGLIO SVIZZERO DI ACCREDITAMENTO
SWISS ACCREDITATION COUNCIL

Institutionell akkreditiert nach HFKG 2018-2025