Normalizing Therapy Concepts with TheraPy

James Stevenson, Kori Kuzma, Matthew Cannon, Susanna Kiwala, Jason Walker, Jeremy L. Warner, Obi L. Griffith, Malachi Griffith, Alex H. Wagner

Introduction

Biomedical literature and knowledgebases may reference drug and therapy concepts in a number of ambiguous or duplicative forms (fig. 1), including:

- Drug labels
- Brand or generic names
- Experimental codes
- Database IDs
- Colloquial terms and nicknames

We introduce **TheraPy**, a Python software package and web API that constructs searchable normalized concepts for drugs and other therapeutics. TheraPy provides mappings from a variety of referent types to a stable concept identifier and aggregated set of traits, enabling more refined data processing in applications ranging from clinical decision-making augmentation to computational drug discovery.

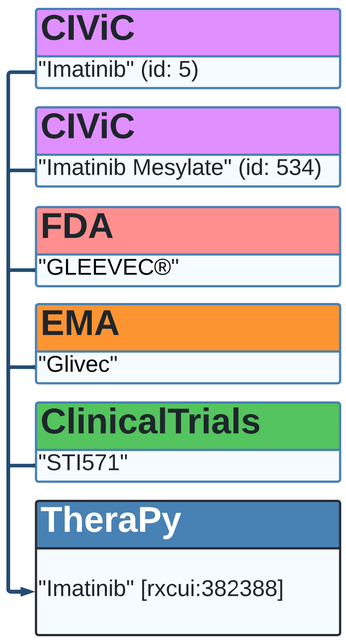

Figure 1: Ambiguous reference terms, and normalized concept ID from TheraPy

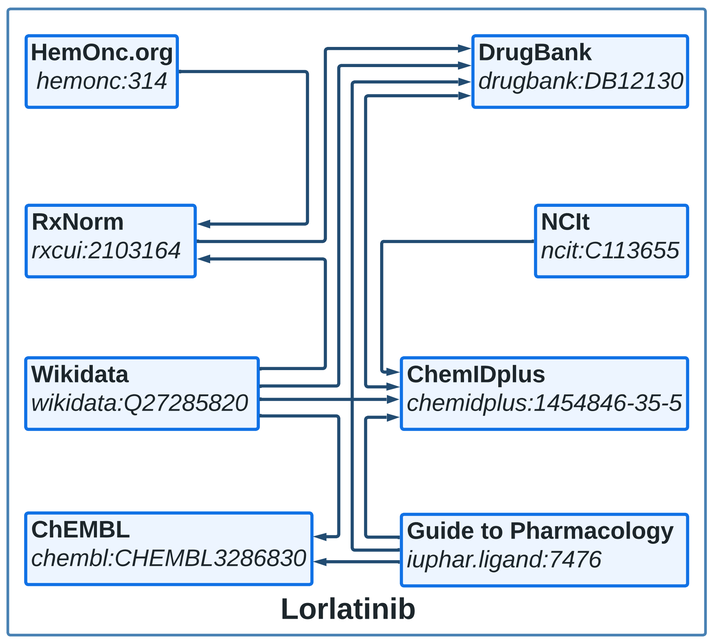

Figure 2: Induced subgraph of connected source records constituting normalized record for *Lorlatinib*

Methods

Terms are extracted from community-generated resources such as Wikidata and the HemOnc.org vocabulary, as well as ChEMBL, the National Cancer Institute Thesaurus, RxNorm, ChemIDplus, Drugs@FDA, DrugBank, and the IUPHAR Guide to Pharmacology. Normalized concepts are generated in a two-step process (fig. 2):

1. A **directed graph** is constructed from source data, where records from each source are nodes, and cross-references from those records to records in other sources are edges
2. The transitive closure of each connectivity relation, i.e. **each set of connected components**, is related as **a distinct therapy concept** and assigned a common identifier. Referents such as aliases and trade names, as well as annotations like regulatory approval and indication data, are merged under this header.

TheraPy is a Python software package and web API that normalizes drug references to support genomic knowledge harmonization.

Architecture and Design

TheraPy is installable locally as a PyPI-distributed library, or via a web API served under the Variant Interpretation for Cancer Consortium (VICC) domain at https://normalize.cancervariants.org/therapy/. Requests are routed via the Python FastAPI package, and employs AWS Elastic Beanstalk for load-balancing and AWS DynamoDB for scalable data storage (fig. 3).

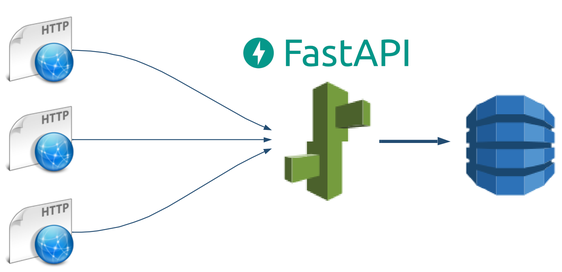

Figure 3: TheraPy web service architecture

Responses are structured according to the VRSATILE specification (fig. 4) for interoperability and consistency with other VICC services such as the ongoing MetaKB project.


```
"therapy_descriptor": {
  "therapy_id": "rxcui:1875534"
  "id": "normalize.therapy:Avelumab",
  "type": "TherapyDescriptor",
  "label": "avelumab",
  "xrefs": [
    "ncit:C116870",
    "hemonc:53",
    "drugbank:DB11945",
    "chemidplus:1537032-82-8",
    "wikidata:Q21083261"
  ],
  "alternate_labels": [
    "anti-PD-L1 monoclonal antibody",
    "Bavencio",
    "MSB0010718C",
    "anti-PD-L1 mAb",
    ...
  ],
  "extensions": [
    {
      "type": "Extension",
      "name": "regulatory_approval",
      "value": {
        "has_indication": [
          {
            "id": "hemonc:569",
            "type": "DiseaseDescriptor",
            "label": "Bladder cancer",
            "disease_id": "ncit:C9334",
          },
        ...
    },
    {
      "type": "Extension",
      "name": "trade_names",
      "value": [
        "Bavencio"
      ]
    ...
}
```


***Figure 4*:** Abbreviated description of normalized concept for cancer drug Avelumab

Analysis

An analysis of the breadth of normalization capabilities was performed on a selected group of external knowledgebases. Results for successful normalization of therapeutic terms from CIViC evidence items, Molecular Oncology Almanac assertions, Genomics of Drug Sensitivity in Cancer measurements, and PharmGKB annotations are shown in figure 5.

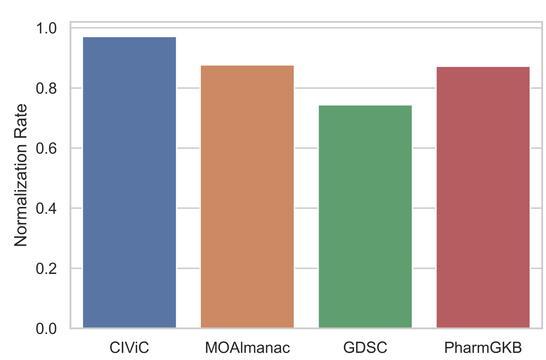

***Figure 5*:** Normalization rates for evidence items from selected external knowledgebases.

Conclusion

TheraPy utilizes cross-source mappings to rapidly and automatically normalize therapy terms. It is currently used in upcoming versions of two major software projects:

- Genomic medical data harmonization in the 2.0 release of the VICC Meta-Knowledgebase
- Drug grouping in the 5.0 release of the Drug-Gene Interaction Database (DGIdb)

Potential limitations include:

- Mapping errors in source databases
- Inconsistent definitions, scopes, or meanings behind cross-references
- Insufficient or non-existant cross-references

Future directions for development include refinement of the normalization algorithm and use of supplementary manual annotations. More expressive source metadata regarding the scope of cross-source mappings could also enable greater rigorousness and accuracy of normalized groupings.

james.stevenson@nationwidechildrens.org
@jamessharpsteen

NATIONWIDE CHILDREN'S
When your child needs a hospital, everything matters.

THE OHIO STATE UNIVERSITY
COLLEGE OF MEDICINE

VICC Variant Interpretation for Cancer Consortium