

12-1-2023

Normalization of drug and therapeutic concepts with Thera-Py

Matthew Cannon

Nationwide Children's Hospital

James Stevenson

Nationwide Children's Hospital

Kori Kuzma

Nationwide Children's Hospital

Susanna Kiwala

Washington University School of Medicine in St. Louis

Jeremy L Warner

Brown University

See next page for additional authors

Follow this and additional works at: https://digitalcommons.wustl.edu/oa_4



Part of the [Medicine and Health Sciences Commons](#)

Please let us know how this document benefits you.

Recommended Citation

Cannon, Matthew; Stevenson, James; Kuzma, Kori; Kiwala, Susanna; Warner, Jeremy L; Griffith, Obi L; Griffith, Malachi; and Wagner, Alex H, "Normalization of drug and therapeutic concepts with Thera-Py." JAMIA Open. 6, 4. ooad093 (2023).

https://digitalcommons.wustl.edu/oa_4/2553






This Open Access Publication is brought to you for free and open access by the Open Access Publications at Digital Commons@Becker. It has been accepted for inclusion in 2020-Current year OA Pubs by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.

Authors

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

Application Notes

Normalization of drug and therapeutic concepts with Thera-Py

Matthew Cannon , PhD¹, James Stevenson, BA¹, Kori Kuzma, BS¹, Susanna Kiwala , MS²,
Jeremy L. Warner, MD, MS³, Obi L. Griffith , PhD², Malachi Griffith , PhD²,
Alex H. Wagner , PhD^{1,4,*}

¹The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, United States, ²Division of Oncology, Department of Medicine, Washington University School of Medicine, St. Louis, MO, United States, ³Department of Medicine, Brown University, Providence, RI, United States, ⁴Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH, United States

*Corresponding author: Alex H. Wagner, PhD, Nationwide Children's Hospital, Room WB3155, Research Building 3, 575 Children's Crossroad, Columbus, OH 43210 (alex.wagner@nationwidechildrens.org)

Abstract

Objective: The diversity of nomenclature and naming strategies makes therapeutic terminology difficult to manage and harmonize. As the number and complexity of available therapeutic ontologies continues to increase, the need for harmonized cross-resource mappings is becoming increasingly apparent. This study creates harmonized concept mappings that enable the linking together of like-concepts despite source-dependent differences in data structure or semantic representation.

Materials and Methods: For this study, we created Thera-Py, a Python package and web API that constructs searchable concepts for drugs and therapeutic terminologies using 9 public resources and thesauri. By using a directed graph approach, Thera-Py captures commonly used aliases, trade names, annotations, and associations for any given therapeutic and combines them under a single concept record.

Results: We highlight the creation of 16 069 unique merged therapeutic concepts from 9 distinct sources using Thera-Py and observe an increase in overlap of therapeutic concepts in 2 or more knowledge bases after harmonization using Thera-Py (9.8%-41.8%).

Conclusion: We observe that Thera-Py tends to normalize therapeutic concepts to their underlying active ingredients (excluding nondrug therapeutics, eg, radiation therapy, biologics), and unifies all available descriptors regardless of ontological origin.

Lay Summary

Working with therapeutic terminology in medicine is challenging due to the ambiguity associated with different naming strategies. A therapeutic can have many different types of identifiers across many vocabularies: natural product names, chemical structures, development codes, generic names, brand names, product formulations, or treatment regimens. This diversity of nomenclature makes therapeutic terminology uniquely difficult to manage and the need for harmonized cross-resource mappings is becoming increasingly apparent. To support these mappings, we introduce Thera-Py, a Python package and web API that constructs stable, searchable therapeutic concepts for drugs and therapeutic terminology. By using a directed graph approach, Thera-Py captures commonly used aliases, trade names, annotations, and associations for any given therapeutic and harmonizes them under a single merged concept record. Using this approach, we found that Thera-Py tends to normalize therapeutic concepts to their underlying active ingredients (excluding nondrug therapeutics, eg, radiation therapy, biologics) and unifies all available descriptors regardless of ontological origin. In this report, we highlight the creation of 16 069 unique merged therapeutic concepts from 9 distinct sources and observe an increased overlap of therapeutic concepts in commonly used knowledge bases after harmonization using Thera-Py.

Key words: therapeutics; medical informatics; biological ontologies; knowledge bases; health information interoperability.

Background and significance

Harmonizing all existing names for any one given therapeutic concept has been a challenging problem for medical informatics in recent decades.¹⁻³ In modern medical practice, medical professionals are frequently expected to synthesize therapeutic knowledge of drug mechanisms, effectiveness, and other metrics to design treatment regimens that achieve the best possible outcomes for their patients. Databases and other resources exist that allow medical professionals to collect information about a therapeutic, but this process can be hampered due to the ambiguity associated with therapeutic

naming strategies. Harmonizing even a single therapeutic requires curated knowledge of all possible identifiers of active ingredients, chemical structures, developmental aliases, and generic or brand names.⁴ This problem is exacerbated in clinical genomics, where ambiguity (or a lack of standardization) can confound treatment decision-making.

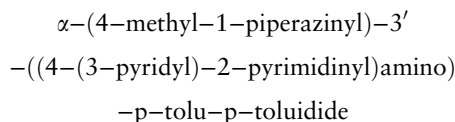
Consider imatinib, a tyrosine kinase inhibitor that was first used to treat Philadelphia chromosome-associated chronic myelogenous leukemia.⁵ This same drug was initially marketed as Gleevec in the United States and Glivec in the EU, by the Swiss-American pharmaceutical company Novartis;

Received: August 31, 2023; Revised: October 11, 2023; Editorial Decision: October 15, 2023; Accepted: October 16, 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of the American Medical Informatics Association.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

additional generic brand names now include Celonib, Enliven, Gleevac, Imalek, Imatib, Mesylonib, Mitinab, Plivatib, Shantinib, Temsan, and Veenat. Before any of these brand names were assigned to the therapeutic, it was published under the identifier “STI-571” in the medical literature.⁶ It can additionally be referenced by the different salt formulations present on the market (imatinib mesylate or imatinib methanesulfonate), or by its chemical structure:



Despite their different ontological origins, the preceding examples are all contextually equivalent when referenced with respect to drug–gene or drug–variant interaction annotations, even if there may be subtle distinctions in other, non-therapeutic contexts. Standards and naming conventions exist at every level of development, from internal pharmaceutical development compound identifiers (eg, AZD-####) and chemical structure names employed in early development pipelines, to fully-realized brand and marketing names with myriad formulations defined by subgroups of additives and delivery mechanisms.⁷ This notion has driven regulatory bodies and programs (such as the United States Adopted Names program) to assign generic names reflecting the underlying active ingredients prior to their marketing. Changes such as these were made in an effort to unify ambiguously named products and protect consumers.⁸ Thus, no matter the stage of development, all assigned names have some tangible link to one another through their relation as a descriptor to the underlying active ingredient(s).

To bridge therapeutic ambiguity, we introduce Thera-Py, a Python package and web API that constructs searchable merged concepts for drugs and therapeutic terminologies using publicly available therapeutic resources and thesauri. Merged concepts are constructed from an aggregate set of traits, trade names, and aliases that act as a cross-resource mapping to enable more refined data processing for downstream clinical and research applications. In this report, we outline the methodology behind Thera-Py and provide an analysis on normalization rates across different data sources. Further, we examine the challenges of normalization of therapeutic terminologies and provide suggestions on improving data standards to support improving data harmonization.

Results

Normalization/grouping routine

Thera-Py utilizes community-driven vocabularies to generate stable concept mappings between identifiers (Figure 1). We aggregated concept codes from 9 therapeutic ontologies and vocabularies. Terms were extracted from: Wikidata,⁹ HemOnc,¹⁰ ChEMBL,¹¹ the National Cancer Institute Thesaurus¹² (NCIt), RxNorm,¹³ ChemIDplus,¹⁴ Drugs@FDA,¹⁵ DrugBank,¹⁶ and the IUPHAR Guide to Pharmacology.¹⁷ These sources were chosen due to their high public use as well as the diversity of scope and knowledge contained within each source. We then developed an algorithm to cross-map extracted concept codes and link together records. Normalized identity records are generated in a 2-step process:

- 1) Directed graphs are constructed from source data, where records from each source act as nodes and “has reference to” relationships act as edges between nodes. These relationships are explicit, curated references (xrefs) from one record to another (eg, the record rxncui: 282388 explicitly references drugbank: DB00619) (shown in Figure 1).
- 2) Each set of connected nodes is related as a distinct, unified therapeutic concept and assigned a common identifier. All aliases, trade names, annotations, associations, regulatory approvals, and indications are merged under this identifier.

Starting nodes were chosen according to an internal source trust ranking, where records with higher priority were used to initialize groups whenever possible. Sources were ordered according to their perceived therapeutic scope where those designed and annotated primarily for clinical decision-making (usually through expert curation) ranked higher than generalized sources. Thusly, the source priority order used for anchor node decision-making, from most preferred to least preferred, was: RxNorm, NCIt, HemOnc, Drugbank, Drugs@FDA, IUPHAR Guide to Pharmacology, ChEMBL, ChemIDplus, followed by Wikidata.

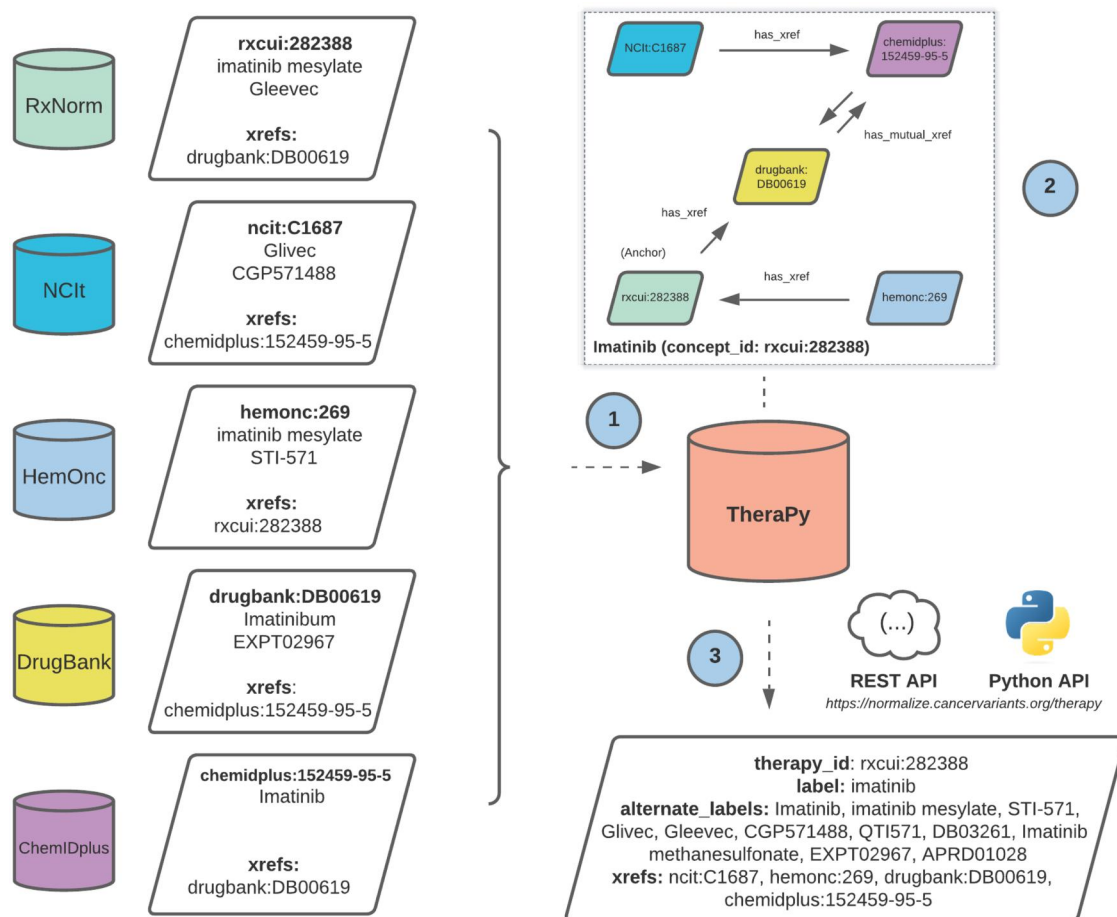
Creation and access of normalized concepts

We ran our normalization routine via Thera-Py as described previously in the Methods section (also available from <https://go.osu.edu/TPY>). Distinct sets of nodes were assigned a stable merged concept identifier with all associated aliases, trade names, and other therapeutic descriptors associated (Figure 1). All merged therapeutic descriptors and cross-references remained accessible via their assigned stable concept identifier.

A total of 16 069 merged groups were created for different therapeutic concepts. These merged groups were assigned identifiers reflective of the anchor node (Figure 1) used to create each group: NCIt (6574 groups, 40.9%), RxNorm (4647 groups, 28.9%), GuideToPharmacology (3490 groups, 21.7%), DrugBank (1214 groups, 7.6%), ChEMBL (93 groups, 0.5%), ChemIDplus (51 groups, 0.3%) (Figure S1a). Of all merged concepts created, 84.7% of groups contained between 2 and 5 records (Figure S1b). The remaining 15.3% of groups contained anywhere from 6 to 86 records. The merged groups with the largest number of combined records are highlighted in Table S1.

Analysis of concept normalization rates

To evaluate the ability of Thera-Py to successfully harmonize therapeutic terminology across resources, we obtained searchable drug vocabularies from 7 different publically available knowledgebases to act as our test set. These knowledgebases are distinct from those used to build Thera-Py and comprised Memorial Sloan Kettering (MSK) Precision Oncology Knowledge Base (OncoKB),¹⁸ Pharmacogenomics Knowledgebase (PharmGKB),¹⁹ Clinical Interpretation of Variants in Cancer (CIViC),²⁰ Cancer Genome Interpreter Cancer Biomarkers Database (CGI),²¹ Molecular Oncology Almanac (MOAlmanac),²² Tumor Alterations Relevant for Genomics-Driven Therapy (TARGET),²³ and the Drug-Gene Interaction Database (DGIdb).²⁴ Prior to normalization, therapeutic terminology was compared via string matching to obtain the intersection of common terminology across resources (Figure 2). DGIdb was not included in this preliminary analysis due to its nature as an aggregate resource. Our



analysis showed a total of 1198 terms unique to a single resource, with 115 and 58 terms being shared across 4 and 5 resources, respectively.

The unique set of terms from each source was then normalized using a local installation of Thera-Py (v.0.3.6) and merged concept identifiers were obtained for each term (Figure S2). The lack of a merged concept identifier for a unique term was deemed as a failure to normalize. This analysis showed high normalization rates for 4 of 7 sources: PharmGKB (95.6%), CGI (91.2%), OncoKB (86.7%), and CIVIC (85.1%) (Figure 3A). The remaining 3 sources saw lower rates of normalization: MOAlmanac (69.2%), DGIdb (65.4%), and TARGET (36.5%). Examples of terms that failed to normalize are highlighted in Tables 1 and 2. The anchor nodes for each successfully retrieved merged concept were also recorded. Our analysis of anchor distributions showed RxNorm to be the most frequently-occurring anchor node for drug terms within 6 out of 7 drug sets: OncoKB, PharmGKB, CIVIC, CGI, MOAlmanac, and TARGET (Figure 3B). In contrast, ChEMBL was the most frequently occurring anchor node for drug terms obtained from DGIdb.

Discussion

Therapeutic vocabularies from public sources were subjected to directed graph construction to construct stable merged concepts for all descriptors for any given therapeutic concept. Our results showed the construction of 16 069 unique therapeutic concept groups from our import set. We found that 84.7% of all merged concepts created with this methodology contained between 2 and 5 records per group (Figure S1c). The remaining 15.3% merged concepts contained >5 records per group with the largest 25 groupings shown in Table S1. The size of these larger groupings can likely be attributed to the contributions of Drugs@FDA. This resource was added to Thera-Py to capture more accurate notions of regulatory approval for therapeutic concept groups through association with all active Abbreviated New Drug Application (ANDA) and New Drug Application (NDA). In doing so, however, this has inflated some therapeutic groups to larger sizes as evidenced by the group “rxcuri: 21245” containing 84 records (79 of which are ANDA/NDA application records).

Our analysis of publicly available drug vocabularies showed high rates of normalization for terms obtained from 5 of 7 sources, with TARGET and DGIdb seeing lower rates

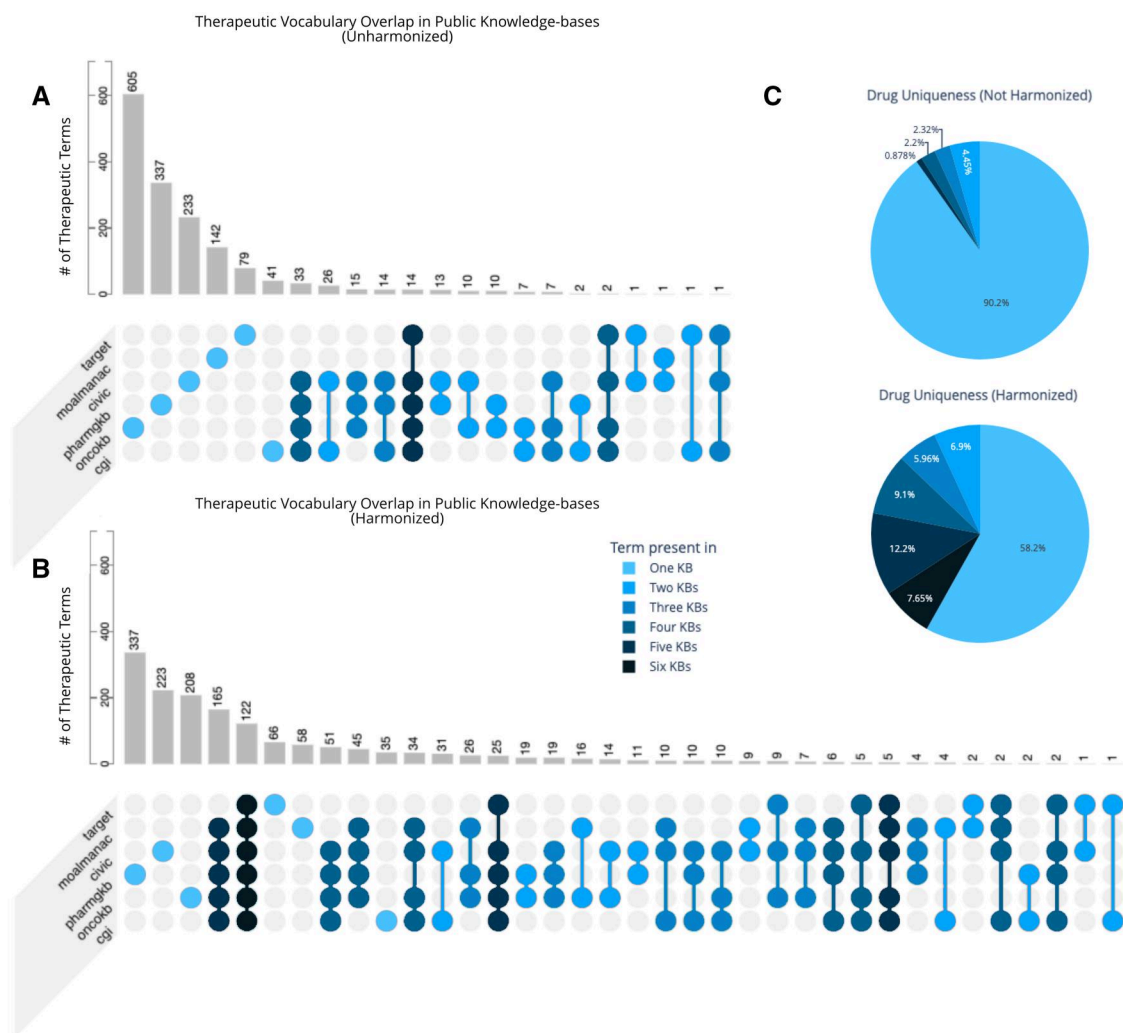
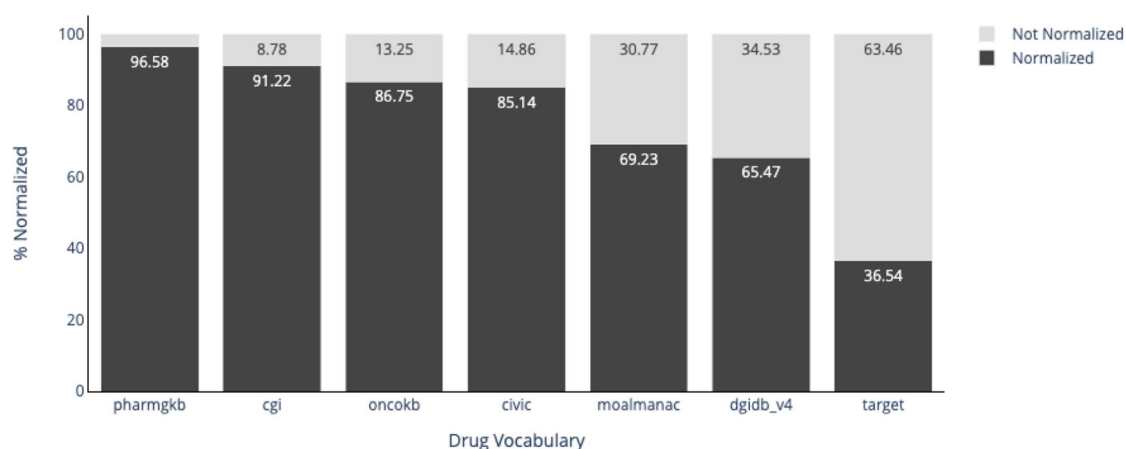


Figure 2. Intersection of therapeutic vocabulary from public knowledge-bases, pre-, and postharmonization using Thera-Py. Therapeutic terminology was obtained from 6 different publicly available drug vocabularies as a test set to evaluate cross-resource therapeutic overlap. (A) Test sets of therapeutic terminology were compared via string matching to quantify the number of exact matches present across resources. The intersections of resources with exact matches are highlighted and colored by the number of contributing resources. (B) Test sets of therapeutic terminology were harmonized using Thera-Py and then compared via concept ID to evaluate the number of matches across resources. Terminologies with exact matches for their concept IDs (irrespective of their original vocabulary term) were quantified. The intersections of resources with matches are highlighted and colored by the number of contributing resources. (C) Drug uniqueness of therapeutic vocabulary across resources pre- and postharmonization using Thera-Py. Uniqueness is quantified as the number of terms present in various knowledge-bases intersection sizes.

of normalization for vocabularies (30%, 65.3%, respectively). We expect the lower rates of normalization in these 2 sources to be likely due to more frequent occurrences of general categories (eg, HDAC Inhibitors, MEK Inhibitors), non-specific identifiers (eg, Pyrrolidine derivative 3, Carbamate derivative 3), misspelled or uncaptured multilanguage labels (eg, “Vandetinib,” Cysplatyna), or unlisted experimental compound identifiers (eg, EVT-103, ADR-851). Additionally, some terms present within these datasets proved to be therapeutically descriptive but difficult to normalize (eg, CD19 CAR Gene Transduced T Lymphocytes, Anti-PD-L1 CSR T Cells, Long-acting erythropoietin conjugate). While Thera-Py does not support fuzzy checks or approximate string matching in its current form, these techniques could be implemented later to handle some of these difficult terminologies. Additionally, the recent development of large language model (LLM) based methodologies could potentially enhance our ability to handle difficult therapeutic terminologies.

We found that our methodology tended to favor normalizing therapeutic concepts to their active ingredient (as defined by USAN generic naming standards). Thera-Py was able to reliably capture relationships between the most used therapeutics at the level of generic names, brand names, and even developmental codes or chemical structures in some cases. Conversely, however, it was unable to capture broader therapeutic groupings such as “tyrosine kinase inhibitor” or “antibody therapy.” Using our approach, attempts to capture broader descriptors would lead to unintended downstream effects whereby all therapeutics would normalize to their broader therapeutic definition regardless of underlying ingredients (ie, erlotinib, dasatinib, or gefitinib all normalizing to “tyrosine kinase inhibitor”). The capture of these broader therapeutic classes likely has practical benefits for downstream applications, though their implementation would need to be defined to a different conceptual space within therapeutic concepts. For example, an additional field called

A TheraPy Performance across Drug Vocabularies



B

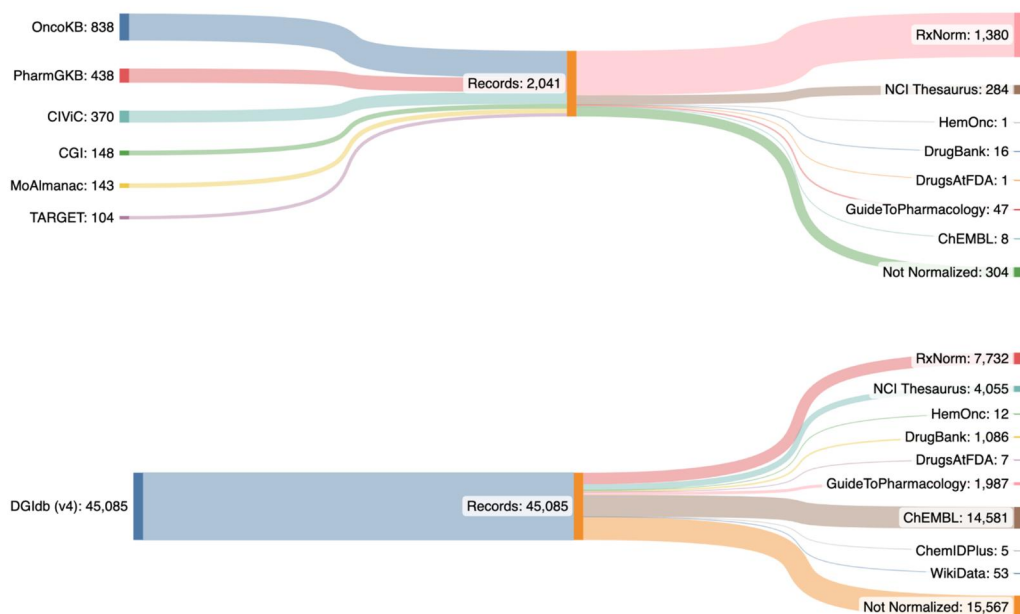


Figure 3. Thera-Py normalization performance using publicly available drug vocabularies. (A) Normalization performance for therapeutic terms obtained from 7 different publicly available resources. (B) Parent node representation for normalized therapeutic terms taken from different resources. Source priority is represented via verticality with records that failed to normalize at the bottom.

“drug class” could be implemented that attaches groupings such as “tyrosine kinase inhibitor” to their relevant therapeutic concepts. This information could then enable harmonization of therapeutics from the point-of-view of drug classes as opposed to explicit underlying ingredients.

Interestingly, among the vocabularies used to create groups and subsequently test Thera-Py, we observed many different types of therapeutic categories all co-occurring within the same fields. These types included: natural products, chemical structures, development codes, generic names, brand names, product formulations, and treatment regimens. With all terms carrying a similar weight despite connotations of maturity, it is important to consider the nuances of what defines a “therapeutic” when applying a normalization strategy such as the one we introduced in Thera-Py.

Our results highlight a critical step for harmonizing therapeutic vocabularies in a computationally digestible format. By merging available records for any therapeutic concept, we are able to create a corresponding identifier that contains all aliases, trade names, and descriptors for commonly used therapeutics. These identifiers can be incorporated within bioinformatic and clinical workflows to unify therapeutic terminology regardless of origin, brand, or maturity stage. Merged records also have potential applications within machine learning workflows, where grouped descriptors can be used to aid in the generation of embeddings for downstream tasks.

More work remains to disambiguate the nuances between therapeutic concept domains and provide additional avenues for quality control of therapeutic concept groups. Future

Table 1. Normalization failure terminology from publicly available drug vocabularies.

Terminology	Occurrences
chemotherapy	6
carboplatin-taxol regimen	3
r3mab	3
car-t cells targeting mesothelin	3
anti-vegf	2
parp inhibitors	2
hdac inhibitors	2
src inhibitors	2
egfr inhibitors	2
opium derivatives and expectorants	2
oxymetazoline and tetracaine	2
antithrombotic agents	2
paracetamol, combinations excl. psycholeptics	2
vitamin b-complex with vitamin c	2
parathyroid hormones and analogues	2
high dose chemotherapy	2
flourouracil	2
pc regimen	2
car-t cells targeting muc1	2
cat regimen	2
sb3	2
carbo-tax regimen	2
liposomal doxorubicin	2
sgk1-inh	2
radiation ionizing radiotherapy	2

Most frequently occurring terms (top 25) taken from drug vocabularies that were not able to return matches from Thera-Py and thus failed to normalize.

Table 2. Additional examples of normalization failure terminology from publicly available drug vocabularies.

Terminology	Description
HDAC inhibitors	General Category
MEK Inhibitors	General Category
beta blocking agents, nonselective	General Category
Tyrosine kinase inhibitors	General Category
Pyrrolidine derivative 3	Compound Identifier
Carbamate derivative 3	Compound Identifier
Tetra-hydro-isoquinoline derivative 4	Compound Identifier
Benzene sulfonamide derivative 3	Compound Identifier
EVT-103	Exp. Compound Identifier
ADR-851	Exp. Compound Identifier
APN-201	Exp. Compound Identifier
Cysplatyna	Multi-language Label
Flourouracil	Misspelled Label
Vandetinib	Misspelled Label
Radiation Therapy	Therapeutic Description
CD19 CAR Gene Transduced T Lymphocytes	Therapeutic Description
Anti-PD-L1 CSR T Cells	Therapeutic Description
Long-acting erythropoietin conjugate	Therapeutic Description
Interferon-alpha lozenge	Therapeutic Description
coxsackievirus type a21	Therapeutic Description
249565746	Unknown Identifier

Select terms taken from drug vocabularies that were not able to return matches from Thera-Py and thus failed to normalize. Descriptions of type of drug terminology have been provided next to each term.

effort will require more precise encodings of semantic relations between classes, leveraging recent specifications such as SSSOM for unambiguous, standardized sharing of cross-domain concept mappings. We look forward to these developments, as success in this area will pave the way for applications such as Thera-Py to assist inference engines and

the development of AI-driven clinical decision support capable of relating disparate therapeutic knowledge resources.

Materials and methods

Extraction of therapeutic concepts from resources

Records for drugs, therapeutics, and chemicals were obtained from individual publicly available resources: Terms were extracted from: Wikidata,⁹ HemOnc,¹⁰ ChEMBL,¹¹ the National Cancer Institute Thesaurus,¹² RxNorm,¹³ ChemID-plus,¹⁴ Drugs@FDA,¹⁵ DrugBank,¹⁶ and the IUPHAR Guide to Pharmacology.¹⁷ Further detail on extraction from each individual source can be found within [Supplementary Methods](#). Records were imported directly as identity records and stored in a locally deployed DynamoDB instance. For each record, aliases, trace names, and database cross-references were extracted and stored as pointers to their original identity. Records within the DynamoDB instance are updated from parent knowledge bases on a quarterly basis.

Analysis of normalization success rates

Drug terminology sets were obtained from 7 different publicly available resources: the Memorial Sloan Kettering (MSK) Precision Oncology Knowledge Base (OncoKB),¹⁸ Pharmacogenomics Knowledgebase (PharmGKB),¹⁹ Clinical Interpretation of Variants in Cancer (CIVIC),²⁰ Cancer Genome Interpreter Cancer Biomarkers Database (CGI),²¹ Molecular Oncology Almanac (MOAlmanac),²² Tumor Alterations Relevant for Genomics-Driven Therapy (TARGET),²³ and the Drug-Gene Interaction Database (DGIdb).²⁴ All drug terms from each source were normalized using a local installation of Thera-Py (v.0.3.6). Successful normalization was determined by the retrieval of a merged concept for each term. If a merged concept was not identified, that term was recorded as a failure of normalization.

Author contributions

M.C. designed experiments, ran data collection, performed data analysis, and drafted the article. A.H.W., M.G., and O.L.G. conceived and supervised the study, and contributed to study design. J.S., K.K., and S.K. wrote the software behind Thera-Py. J.S. additionally assisted with data collection, data analysis, and drafting the article. J.L.W. contributed datasets for Thera-Py and helped conceive the study. All authors contributed to manuscript revision and gave approval for the final version to be published.

Supplementary data

[Supplementary material](#) is available at *JAMIA Open* online.

Funding

This work was supported by the National Human Genome Research Institute (NHGRI) (grant number R00HG010157).

Conflicts of interest

None declared.

Data availability

Thera-Py is an open-source python package and is available for download and use at: <https://github.com/cancervariants/therapy-normalization>. An interactive demo is available from: <https://normalize.cancervariants.org/therapy>. Individual terms can be searched on this page to allow for manual inspection of therapeutic concepts without the need for locally hosted software. All therapeutic concepts were constructed from publicly accessible knowledge bases and are available for access via Thera-Py endpoints. The DynamoDB instance supporting Thera-Py is updated on a quarterly basis.

References

- Peters L, Kapusnik-Uner JE, Bodenreider O. Methods for managing variation in clinical drug names. *AMIA Annu Symp Proc*. 2010;2010:637-641.
- McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. *Proc Annu Symp Comput Appl Med Care*. 1994:235-239.
- Eccher C, Ferro A, Pisanelli DM. An ontology of therapies. In: Kostkova P, ed. *Electronic Healthcare*. Springer Berlin Heidelberg; 2010:139-146.
- Quist AJL, Hickman TTT, Amato MG, et al. Analysis of variations in the display of drug names in computerized prescriber-order-entry systems. *Am J Health Syst Pharm*. 2017;74(7):499-509.
- Iqbal N, Iqbal N. Imatinib: a breakthrough of targeted therapy in cancer. *Chemother Res Pract*. 2014;2014:357027-357029.
- Verweij J, van Oosterom A, Blay JY, et al. Imatinib mesylate (STI-571 Glivec®, Gleevec) is an active agent for gastrointestinal stromal tumours, but does not yield responses in other soft-tissue sarcomas that are unselected for a molecular target: results from an EORTC Soft Tissue and Bone Sarcoma Group Phase II Study. *Eur J Cancer*. 2003;39(14):2006-2011.
- Lester CA, Flynn AJ, Marshall VD, Rochowiak S, Rowell B, Bagian JP. Comparing the variability of ingredient, strength, and dose form information from electronic prescriptions with RxNorm drug product descriptions. *J Am Med Inform Assoc*. 2022;29(9):1471-1479.
- Karet GB. How do drugs get named? *AMA J Ethics*. 2019;21(8):E686-E696.
- Vrandečić D, Krötzsch M. Wikidata: a free collaborative knowledgebase. *Commun ACM*. 2014;57(10):78-85.
- Warner JL, Dymshyts D, Reich CG, et al. HemOnc: a new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. *J Biomed Inform*. 2019;96:103239.
- Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. *Nucleic Acids Res*. 2017;45(D1):D945-D954.
- Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform*. 2007;40(1):30-43.
- Liu S, Ma W, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. *IT Prof*. 2005;7(5):17-23.
- Tomasulo P. ChemIDplus-super source for chemical and drug information. *Med Ref Serv Q*. 2002;21(1):53-59.
- Center for Drug Evaluation and Research (U.S.). Drugs@FDA. 2004. Accessed December 10, 2022. <http://www.accessdata.fda.gov/scripts/cder/drugsatfda/>.
- Wishart DS, Knox C, Guo AC, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*. 2006;34(90001):D668-D672.
- Harding SD, Armstrong JF, Faccenda E, et al.; NC-IUPHAR. The IUPHAR/BPS guide to PHARMACOLOGY in 2022: curating pharmacology for COVID-19, malaria and antibacterials. *Nucleic Acids Res*. 2022;50(D1):D1282-D1294.
- Chakravarty D, Gao J, Phillips SM, et al. OncoKB: a Precision Oncology Knowledge Base. *JCO Precis Oncol*. 2017;2017:PO.17.00011. <https://doi.org/10.1200/PO.17.00011>
- Whirl-Carrillo M, Huddart R, Gong L, et al. An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther*. 2021;110(3):563-572.
- Griffith M, Spies NC, Krysiak K, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet*. 2017;49(2):170-174.
- Tamborero D, Rubio-Perez C, Deu-Pons J, et al. Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med*. 2018;10(1):25.
- Reardon B, Moore ND, Moore NS, et al. Integrating molecular profiles into clinical frameworks through the molecular oncology almanac to prospectively guide precision oncology. *Nat Cancer*. 2021;2(10):1102-1112.
- Van Allen EM, Wagle N, Stojanov P, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med*. 2014;20(6):682-688.
- Freshour SL, Kiwala S, Cotto KC, et al. Integration of the drug-gene interaction database (DGIdb 4.0) with open crowdsourcing efforts. *Nucleic Acids Res*. 2021;49(D1):D1144-D1151.