Drug Repurposing for Cancer: An NLP Approach to Identify Low-Cost Therapies

Shivashankar Subramanian[1,2] **Ioana Baldini**[1] **Sushma Ravichandran**[1]

Dmitriy A. Katz-Rogozhnikov[1] **Karthikeyan Natesan Ramamurthy**[1]

Prasanna Sattigeri[1] **Kush R. Varshney**[1] **Annmarie Wang**[3,4] **Pradeep Mangalath**[3,5]

Laura B. Kleiman[3]
[1]IBM Research, Yorktown Heights, NY, USA
[2]University of Melbourne, VIC, Australia
[3]Cures Within Reach for Cancer, Cambridge, MA, USA
[4]Massachusetts Institute of Technology, Cambridge, MA, USA
[5]Harvard Medical School, Boston, MA, USA

Abstract

More than 200 generic drugs approved by the U.S. Food and Drug Administration for non-cancer indications have shown promise for treating cancer. Due to their long history of safe patient use, low cost, and widespread availability, repurposing of generic drugs represents a major opportunity to rapidly improve outcomes for cancer patients and reduce healthcare costs worldwide. Evidence on the efficacy of non-cancer generic drugs being tested for cancer exists in scientific publications, but trying to manually identify and extract such evidence is intractable. In this paper, we introduce a system to automate this evidence extraction from PubMed abstracts. Our primary contribution is to define the natural language processing pipeline required to obtain such evidence, comprising the following modules: querying, filtering, cancer type entity extraction, therapeutic association classification, and study type classification. Using the subject matter expertise on our team, we create our own datasets for these specialized domain-specific tasks. We obtain promising performance in each of the modules by utilizing modern language modeling techniques and plan to treat them as baseline approaches for future improvement of individual components.

1 Introduction

Each year around the world, nearly 10 million people die from cancer [2] and the cost of cancer exceeds USD $1 trillion [1]. Finding new therapeutic uses for inexpensive generic drugs ("drug repurposing") can rapidly create affordable new treatments. Hundreds of non-cancer generic drugs have shown promise for treating cancer, but it is unclear which are the most worthwhile repurposing opportunities to pursue. Scientific publications such as preclinical laboratory studies and small clinical trials contain evidence on generic drugs being tested for cancer use. The Repurposing Drugs in Oncology (ReDO) project, through manually inspecting research articles indexed by PubMed, found anti-cancer evidence for more than 200 non-cancer generic drugs [15, 4, 19]. However, manual review to identify and analyze potential evidence is time-consuming and intractable to scale. As PubMed

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Workshop on Machine Learning for Health Vancouver, Canada.

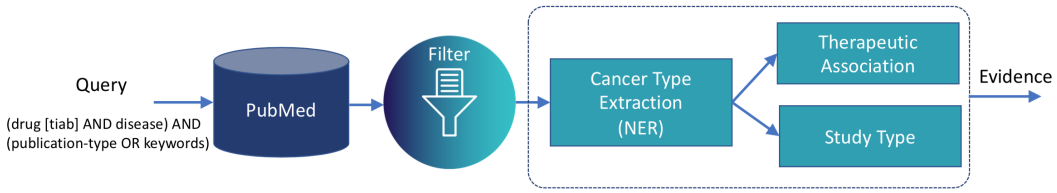

Figure 1: Solution overview: Input to the evidence discovery pipeline is a list of non-cancer generic drugs, and the output is published scientific evidence for each drug for treating various cancer types.

indexes millions of articles and the collection is continuously updated, it is imperative to devise (semi)automated techniques to synthesize the existing evidence. Machine learning (ML)-powered evidence synthesis would provide a comprehensive and real-time view of drug repurposing data and enable actionable insights. To this end, we must achieve task automation, algorithmic accuracy, and technical scalability in three key areas: evidence identification, extraction, and synthesis.

The work presented in this paper is part of an ambitious initiative to synthesize the plethora of scientific and real-world data on non-cancer generic drugs to identify the most promising therapies to repurpose for cancer. This type of endeavor requires close collaboration between experts in different disciplines, such as cancer research (to provide guidance, annotate datasets, and verify results), machine learning (to devise machine learning tasks, select datasets to be annotated, devise models, and evaluate performance), and software engineering (to incorporate models in end-to-end online applications). Furthermore, implementing repurposed therapies as the standard of care in medical practice requires definitive clinical trials, new incentives and business models to fund them, and engagement by various stakeholders such as patients, doctors, payers, and policymakers. In this paper, we focus on the key aspect of identifying and extracting relevant evidence from PubMed articles.

Methods for synthesizing drug repurposing evidence can be divided into three major categories: network-based methods, natural language processing (NLP) approaches, and semantic techniques [20]. Network-based approaches aim to infer relationships between biological entities (drug–disease or drug–target relationships), inspired by the fact that biologic entities (disease, drug, protein, etc.) in the same module of biological networks share similar characteristics [12]. NLP approaches aim to both identify biological entities and mine new knowledge from scientific literature [10]. Semantic approaches require a semantic network to be built first, which can be used with various approaches to mine relationships between entities [14]. We focus on NLP approaches. Our primary contributions, described in detail in the remainder of this paper, are as follows: formulating the pipeline of NLP tasks required to identify relevant evidence of generic drug repurposing for cancer from PubMed articles, precisely specifying the NLP tasks in terms of input and output (not an easy endeavor), Creating domain-specific datasets that support the task definition, designing and evaluating initial models for each of the domain-specific tasks.

2 NLP Pipeline and Dataset for Drug-Cancer Evidence Extraction

PubMed, provided by the National Center for Biotechnology Information (NCBI), is a comprehensive source of biomedical studies, comprising more than 30 million biomedical abstracts and citations from various sources such as MEDLINE, life science journals, and online books. Given a list of generic drugs, the goal of our work is to automatically select abstracts from the large PubMed collection, that measure cancer-relevant phenotypic outcomes of interventions with generic drugs.[1]

We propose an evidence discovery pipeline, shown in Figure 1. First we query PubMed using a strategy inspired by the Cochrane highly sensitive search (CHSS) strategy [6] to narrow the collection of articles we analyze. Note that querying PubMed, even with a sophisticated search string, may not yield only *relevant* articles. Hence we have a (shallow) filtering stage to reject the easy irrelevant

[1]Phenotype is the observable physical properties of an organism; these include the organism's appearance, development, and behavior. We focus on phenotypic outcomes (such as proliferation/death of cells grown in culture or tumor progression/overall survival rates for clinical trials) since they are a more direct measure of outcomes that matter to cancer patients and represent stronger therapeutic evidence (as opposed to, for example, the effects of drugs on protein levels).

Retinoids can block cell proliferation and induce apoptosis in tumor cells. The antitumoral effect of synthetic retinoids like **Adapalene (ADA)** on **hepatoma** cells (HepG2, Hep1B) was investigated. **ADA** at 10(-4)M efficiently induced apoptosis, reaching 61.7% in HepG2 and 79.1% in Hep1B after 72 h incubation. This was accompanied by up-regulation of pro-apoptotic bax and caspase 3, while bcl-2 was down-regulated, shifting the bax/bcl-2 ratio to >2.3 in hepatoma cells. ***ADA** inhibits **hepatoma** cell growth in vitro and is a powerful inducer of **hepatoma** cell apoptosis.*

Figure 2: Sample *relevant* abstract annotation. PubMed # 15105045, *Adapalene* is the non-cancer generic drug, used to treat *hepatoma cancer*. It is a preclinical evidence (in vitro study), and has an *effective* association. Evidence for association with phenotypic outcome measured is italicized.

Coarse-level	Association	Count
Irrelevant	No relation to cancer	553
	No phenotypic outcome	155
Relevant	Effective	555
	Detrimental	50
	Inconclusive	216
	No effect	68

Study type	Count
Preclinical	318
Clinical observational study	107
Clinical trial	69
Other	28

Table 1: Dataset statistics for therapeutic association distribution (left) and study type distribution (right) for the relevant abstracts.

cases. Using the resulting abstracts, cancer types are identified using a named entity recognition (NER) model. With the abstract and pairs of drug-cancer types, therapeutic association is classified and also the type of study is categorized. We refer to this collection of information (i.e., drug, cancer, therapeutic association, study type) as the *evidence* discussed in the PubMed abstract.

The therapeutic association schema contains the following classes: *1. Irrelevant:* A. Drug has no relation to cancer (cases where either the drug or the cancer is not the focus of the study); B. Abstract does not discuss a *phenotypic outcome* and *2. Relevant:* A. Effective: the drug was shown to be effective for treating the cancer; B. Detrimental: the drug has a detrimental effect on the cancer; C. No effect: the drug has no effect on the cancer; D. Inconclusive: the results of the study are inconclusive.

The study types we consider are defined as follows: preclinical studies (in vitro, in vivo), observational studies (including case reports), and clinical trials.

Identifying such evidence from scientific abstracts is not trivial. The articles that discuss cancer interventions use domain-specific jargon which makes the text hard to comprehend by both humans with non-expert background and machines that are not trained with domain-specific data. Hence a strong collaboration between domain-experts and data scientists is required to define machine learning tasks, collect and annotate the appropriate information and design and evaluate machine learning models that address the designed tasks. Due to space limitations, we cannot elaborate in this paper on all difficulties in manually annotating datasets for all these tasks. We will refer to them during the presentation of our work. An example of an annotated abstract is shown in Figure 2.

Our team of machine learning and biomedical scientists worked closely together to fine-tune the querying and filtering strategy and to annotate cancer types, along with the therapeutic associations and study types. In the interest of space, we present the results of the dataset creation in Table 1 without the details of the iterative process of producing it.

3 Models for Cancer Type, Therapeutic Association, and Study Type

We briefly discuss the models and their performance for cancer entity extraction, therapeutic association classification, and study type classification, which form the key components of the proposed evidence discovery pipeline.

3.1 Cancer Type Extraction

For cancer entity identification, we use two main named entity recognition (NER) methods. We train sequential token-level IOB (inside, outside, beginning) tag prediction model using the BioNLP13CG dataset [17]. Tokens that are not of interest are treated as 'O'. We use the well-known conditional

Class	Log. Reg	DAN	SciBERT
Irrelevant	**0.83**	0.81	0.81
Relevant	0.80	0.79	**0.85**

Table 2: Binary therapeutic association classification

Class	Log. Reg	DAN	SciBERT
No relation	0.74	0.71	**0.80**
No phenotypic outcome	0.30	0	**0.34**
Effective	0.72	0.67	**0.73**
Detrimental	0.12	0	**0.33**
No effect	**0.18**	0	**0.18**
Inconclusive	**0.30**	0.03	0.25

Class	PT	BoW	MeSH	All
Preclinical	0.86	**0.96**	0.95	**0.96**
Clinical observational study	0.25	0.81	0.66	**0.84**
Clinical trial	0.78	0.78	0.60	**0.80**
Other	0.38	0.34	0.37	**0.40**

Table 3: Fine-grained therapeutic association classification (left) and Study type classification using logistic regression (right).

random field (CRF) [18] and convolutional neural network (CNN) based SpaCy models [7] for entity extraction. We evaluate the performance on the 1085 abstracts using recall[2] with exact match and a token-level overlap score [13], where the predicted entity with highest overlap is used to compute the score. The CRF-based model obtains a recall of 54.2% and an overlap score of 66.4%, while the SpaCy-based model presents higher performance of 67.7% recall and 77.6% overlap score.

3.2 Therapeutic Association Classification

Given a drug-cancer pair and the corresponding abstract text, we build three different models for therapeutic association classification: *1. Logistic Regression* with feature vectors that are a concatenation of term frequency bag-of-words representations of abstracts, drug and cancer type [9]; *2. Deep Averaging Networks (DAN)*: similar to logistic regression, except that the tokens of abstract, drug and cancer type are initialized with word vectors trained using skip-gram objective over a large set of PubMed abstracts [16]; the text of a given abstract is passed through deep averaging networks [8] where the word vectors are re-trained (with the training data and classification objective), and the representations of abstract, drug and cancer type are concatenated, and passed through a final logistic layer; and *3. SciBERT [5, 3]*: The drug and cancer type entities are encapsulated with special characters and concatenated with the input abstract text. The task is framed as a multi-class classification problem. The sequence representation is obtained using SciBERT's encoding of the [CLS] token[3] (from the last hidden layer). This encoding captures the entire sequence representation and is used for the multi-class classification with a logistic layer.

We perform a 5-fold cross-validation split at the document level, and evaluate performance for drug-cancer type pairs, given the abstract text. Note that we use gold-standard cancer type annotations for this analysis. We evaluate two different settings to understand the complexity of the task: (1) irrelevant vs. relevant binary classification (Table 2) and (2) all six classes (Table 3, left side). Performance is measured using F-score. SciBERT performs the best in most cases.

3.3 Study Type Classification

We train logistic regression models with different choices of features [11]: bag-of-words (BoW), publication type (PT), MeSH terms and combining all. Results using logistic regression with different choices of features are given in Table 3 (right side). Performance is measured using F-score. Using all the features together provides the best performance.

4 Conclusion and Future Work

We proposed an end-to-end evidence discovery pipeline that fetches potential candidate abstracts from PubMed for further evaluation with the goal of identifying non-cancer generic drug activity against different cancer types. We discuss the components in the pipeline, and use NLP approaches along with a number of well-thought-out heuristics to provide solutions for each component.

[2]Ratio between *count of unique cancer entities predicted correctly by the model* and *number of unique cancer entities*.

[3][CLS] is inserted as a special beginning token for every input sequence.

References

[1] The economics of cancer prevention and control. https://issuu.com/uicc.org/docs/wcls2014_economics_of_cancer_final?e=10430107/10454633. [Online; accessed 09-09-2019].

[2] Worldwide cancer statistics. https://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer. [Online; accessed 09-09-2019].

[3] I. Beltagy, A. Cohan, and K. Lo. SciBERT: Pretrained contextualized embeddings for scientific text. arXiv:1903.10676, 2019.

[4] G. Bouche, P. Pantziarka, and L. Meheus. Beyond aspirin and metformin: The untapped potential of drug repurposing in oncology. *Eur. J. Cancer*, 172:S121—-S122, 2017.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[6] K. Dickersin, R. Scherer, and C. Lefebvre. Identifying relevant studies for systematic reviews. *BMJ*, 309:1286, 1994.

[7] M. Honnibal and I. Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017.

[8] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *ACL-IJCNLP*, 2015.

[9] E. Lehman, J. DeYoung, R. Barzilay, and B. C. Wallace. Inferring which medical treatments work from reports of clinical trials. In *NAACL-HLT*, 2019.

[10] J. Li, X. Zhu, and J. Y. Chen. Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. *PLoS Comput. Biol.*, 5(7):e1000450, 2009.

[11] I. J. Marshall, A. Noel-Storr, J. Kuiper, J. Thomas, and B. C. Wallace. Machine learning for identifying randomized controlled trials: An evaluation and practitioner's guide. *Res. Synth. Methods*, 9:602–614, 2018.

[12] V. Martínez, C. Navarro, C. Cano, W. Fajardo, and A. Blanco. DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data. *Artif. Intell. Med.*, 63(1):41–49, 2015.

[13] E. Moreau, F. Yvon, and O. Cappé. Robust similarity measures for named entities matching. In *COLING*, 2008.

[14] G. Palma, M.-E. Vidal, and L. Raschid. Drug-target interaction prediction using semantic similarity and edge partitioning. In *ISWC*, 2014.

[15] P. Pantziarka, V. Sukhatme, L. Meheus, V. Sukhatme, and G. Bouche. Repurposing non-cancer drugs in oncology — how many drugs are out there? bioRxiv:197434, 2017.

[16] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou. Distributional semantics resources for biomedical text processing. In *LBM*, 2013.

[17] S. Pyysalo, T. Ohta, R. Rak, A. Rowley, H.-W. Chun, S.-J. Jung, S.-P. Choi, J. Tsujii, and S. Ananiadou. Overview of the cancer genetics and pathway curation tasks of BioNLP shared task 2013. *BMC Bioinformatics*, 16(Suppl. 10):S2, 2015.

[18] H.-J. Song, B.-C. Jo, C.-Y. Park, J.-D. Kim, and Y.-S. Kim. Comparison of named entity recognition methodologies in biomedical documents. *BioMed. Eng. OnLine*, 17(Suppl. 2):158, 2018.

[19] C. Verbaanderd, L. Meheus, I. Huys, and P. Pantziarka. Repurposing drugs in oncology: Next steps. *Trends Cancer*, 3(8):543–546, 2017.

[20] H. Xue, J. Li, H. Xie, and Y. Wang. Review of drug repositioning approaches and resources. *Int. J. Biol. Sci.*, 14(10):1232–1244, 2018.