Improving a text-mining algorithm to update the Pediatric Molecular Target List

Ainiwan, Nuerye, FDA/CDER/OTS/OCP/DARS and Oak Ridge Institute for Science and Education; Samuels, Sherbet, FDA/CDER/OTS/OCP; Hyland, Paula, FDA/CDER/OTS/OCP/DARS; Akalu, Alemayehu, FDA/CDER/OND/OHOP/OCE; Reaman, Gregory, FDA/CDER/OND/OHOP/OCE; Racz, Rebecca, FDA/CDER/OTS/OCP/DARS

Abstract

The Pediatric Molecular Target List (PMTL) was created in 2018 by the Food and Drug Administration (FDA) to facilitate pediatric oncology drug development. This study uses the rule-based text mining software, Linguamatics, to develop and analyze a query to identify abstracts and titles containing molecular targets related to pediatric oncology. A previous version of this query was created in 2019, and this study integrates new phrase patterns and cancers to increase performance. The precision of the current version is near 100%, an increase of approximately 10% compared to the previous query, and specificity is near 100%. The sensitivity is 69%, meaning some relevant abstracts were missed. When all of MEDLINE was searched, a total of 3967 new molecular targets that are not currently on the PMTL were retrieved. After manual curation completed which is used to determine the performance of the algorithm and identify new molecular targets for the PMTL, there are 2142 newly discovered pediatric cancer markers: 129 diagnostic biomarkers, 366 prognostic biomarker, 250 other type of biomarker, and 1397 potential therapeutic targets that should be further evaluated for inclusion on the PMTL. With further filtering and prioritization, the results of this query will decrease the manual labor required to update the PMTL and aid as a reviewer tool during receipt and review of pediatric oncology studies

Introduction

- Cancer is the leading cause of death in the pediatric population, and pediatric cancer types, molecular features, and pathogenesis can differ from adults
- The PMTL was created to guide the pharmaceutical industry in pediatric oncology drug development
    - Currently contains over 200 targets that show evidence of impacting the growth or progression of at least one pediatric cancer.
    - This list is required to be updated regularly and is currently updated manually
- To reduce the manual curation required for PMTL updates, a text-mining query was developed in 2019 to identify molecular targets associated with pediatric cancer in MEDLINE abstracts
    - This study aims to update the 2019 query with the goal of improving query performance

Materials and Methods

Query Development and Validation (Figure 1)

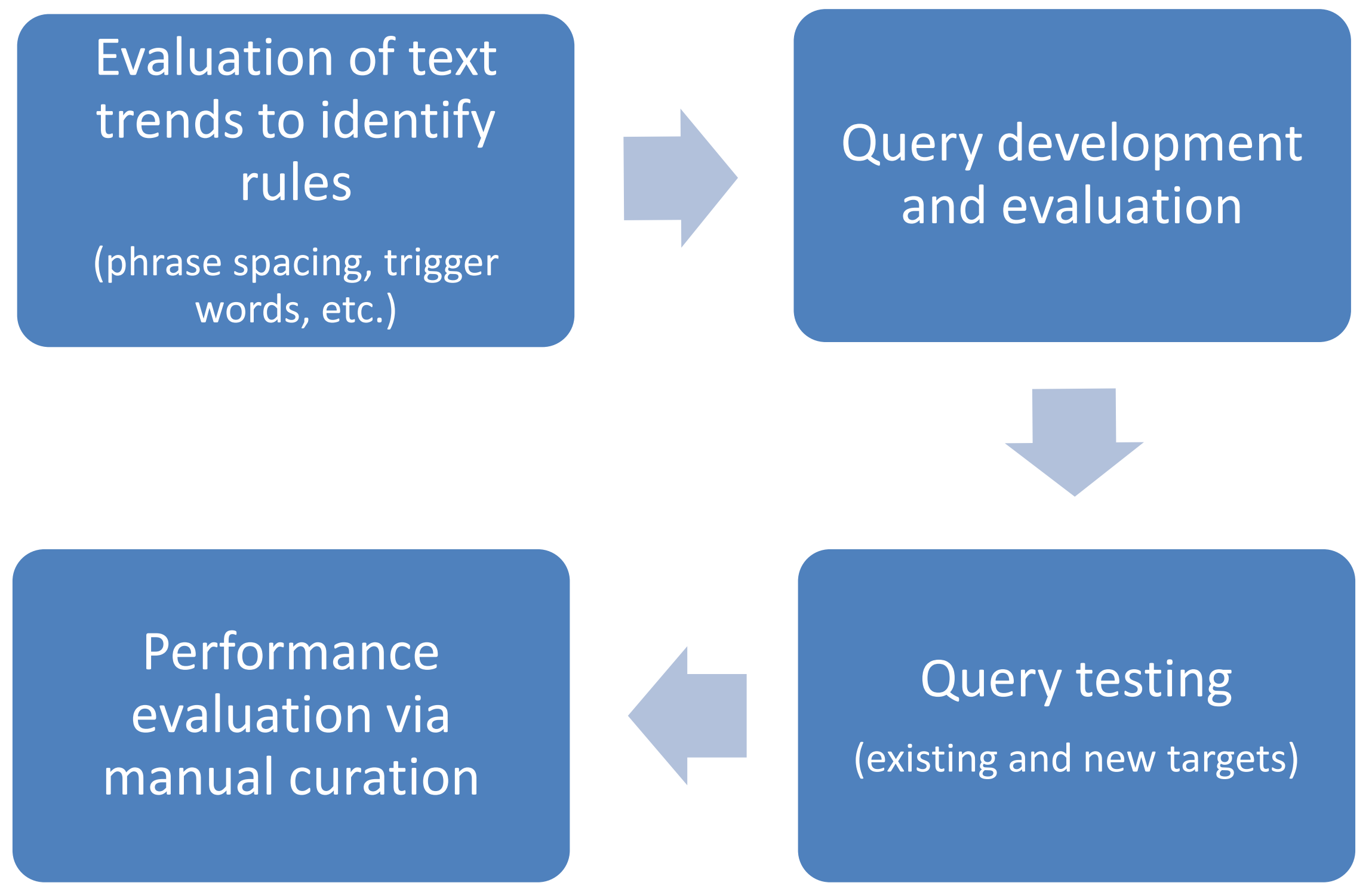
- Linguamatics OnDemand rule-based NLP software was used to develop the query and extract molecular targets related to pediatric cancer
    - The MEDLINE abstracts index in Linguamatics was used for query development
- The original query was developed in 2019 by manually identifying "rules" from relevant abstracts for 50 targets on the PMTL
    - These 50 targets were used to validate the query once it was developed
    - Rules include:
        - A gene and pediatric cancer within 20 words and within the same sentence
        - A "Gene-Disease" linker between the gene and cancer
        - Negation of medical terms that share the same abbreviation of a gene
- Updates were made to enhance performance, including:
    - Negation of additional medical terms
    - Addition of multiple pediatric cancers

Query Testing

- The query was tested using targets and abstracts that are currently on the PMTL as well as new targets found in MEDLINE abstracts
    - Test 1: Retrieval of current PMTL targets (Table 1)
    - Test 2: Evaluation of 100 current PMTL abstracts (Table 2)
    - Test 3: Evaluation of new targets retrieved from MEDLINE (Figure 2)

Statistical Analysis

- Sensitivity, specificity, and precision were calculated and compared to prior query versions



Evaluation of text trends to identify rules

(phrase spacing, trigger words, etc.)

Query development and evaluation

Query testing

(existing and new targets)

Performance evaluation via manual curation

Figure 1: Workflow for query development and evaluation process.

Materials and Methods (Continued)

Manual curation

- Three team members manually reviewed the new molecular targets identified by the query in Test 3.
    - An annotation protocol was created to ensure agreement on true and false positives, and interrater reliability was calculated (Table 3)
- Positive targets were classified as a "therapeutic target", "prognostic biomarker", "diagnostic biomarker", or "other relevant targets"

Results and Discussion

Table 1: Comparison of the number of current PMTL genes returned by the original and updated queries. The number of current genes retrieved by the updated query in MEDLINE was less than the original query. This could be due to more stringent negation criteria.

	Original Query	Updated Query
Returned (true positive)	112	90
Omitted (false negative)	9	30

Table 2: Comparison of the number of current PMTL abstracts returned by the original and updated queries. One hundred abstracts on the current PMTL were curated to evaluate query performance. An increase in precision and specificity was balanced by a decrease in sensitivity, suggesting that relevant abstracts may be missed.

	Original Query	Updated Query
True positive	36	35
True negative	55	59
False positive	4	0
False negative	15	16
Precision	90%	100%
Specificity	93%	100%
Sensitivity	71%	69%

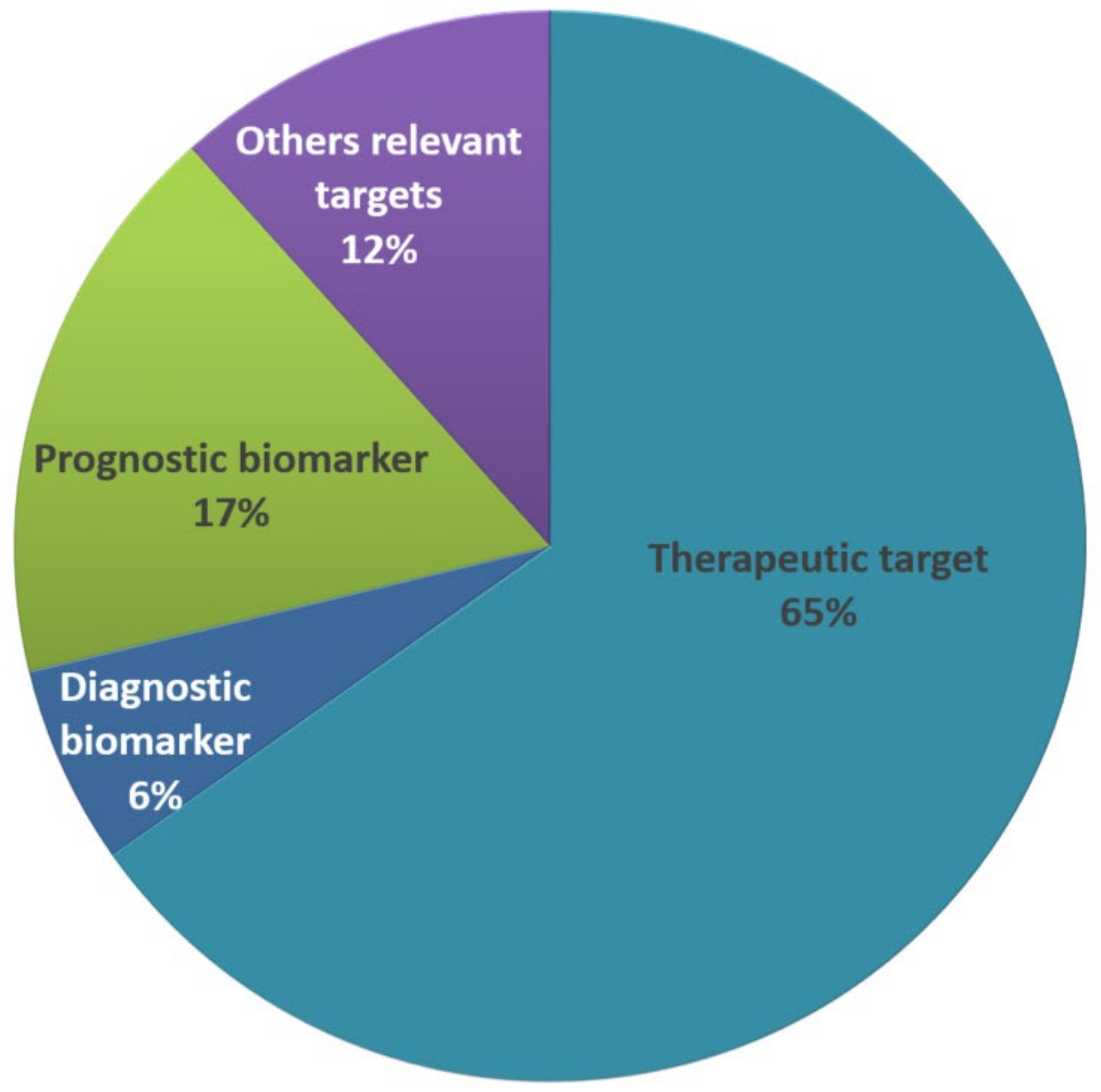


Figure 2: New pediatric cancer associated targets retrieved by the updated query. Targets were grouped into 4 categories: diagnostic biomarker, prognostic, therapeutic target and others relevant targets.

Reviewer Pair	Cohen's Kappa	Interpretation
Reviewer 1-Reviewer 2	0.76	Substantial agreement
Reviewer 1-Reviewer 3	0.72	Substantial agreement
Reviewer 2-Reviewer 3	0.79	Substantial agreement

Table 3: Interrater reliability for manual curation. Interrater reliability was calculated using Cohen's kappa and resulted in substantial agreement between reviewers.

Conclusions

- This project updated a standardized text-mining algorithm for molecular targets in pediatric cancer. Compared to the previous query version, more new pediatric cancer-related molecular targets were discovered.
- The updated query demonstrates improved precision and specificity but may miss relevant abstracts.
    - Further updates to the query will aim to improve the sensitivity
- Minimizing false positives will reduce the need for manual curation when updating the PMTL, provide evidence-based information to FDA oncology reviewers, and guide future pediatric oncology research

Acknowledgements

The authors would like to thank Raajan Naik (OTS/OCP) for developing the original query modified in this project and Terry Tsui (OND/ORO) for her valuable insight and suggestions for improvements.

This project was supported in part by an appointment to the Research Participation Program at the Office of New Drugs/Center for Drug Evaluation and Research, U.S. Food and Drug Administration, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and FDA.

Disclaimer: *The findings and conclusions in this presentation reflect the views of the authors and should not be construed to represent FDA's views or policies.*