Developing a Portfolio Analysis Tool for Prevention Science: A Work in Progress

Patricia Mabry[1], Jessica Wu[1], Jennifer Villani[1]*, Fai Chan[2], B. Ian Hutchins[2], Ifeyinwa Udo[3], George Santangelo[2], Calvin Johnson[4], David Murray[1]

[1]Office of Disease Prevention and [2]Office of Portfolio Analysis, Division of Program Coordination, Planning, and Strategic Initiatives, and [4]Center for Information Technology, National Institutes of Health, Bethesda, MD 20892; [3]IQ Solutions, Inc., Rockville, MD 20852; *Corresponding author: villani@nih.gov

Goal

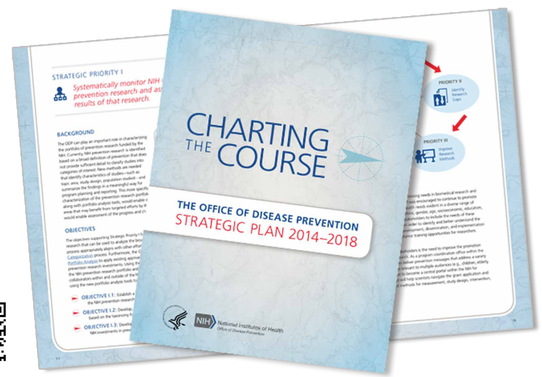
To create a computer-based automated tool capable of objectively classifying the NIH prevention research grant portfolio on a number of dimensions of interest.

Background

The mission of the NIH Office of Disease Prevention (ODP) is to improve the public's health by increasing the scope, quality, dissemination, and impact of prevention research supported by the NIH. Recently, the ODP released its first Strategic Plan consisting of six strategic priorities, covering fiscal years 2014–2018 (https://prevention.nih.gov/about/strategic-plan).

ODP STRATEGIC PRIORITY I: *Systematically monitor NIH investments in prevention research and assess the progress and results of that research. Objectives:*

- *Establish a taxonomy for prevention research that the ODP can apply to analyze the NIH prevention research portfolio.*
- *Develop, test, and implement portfolio analysis tools to classify NIH awards based on the taxonomy for prevention research.*
- *Develop and implement a process to regularly assess the progress and results of NIH investments in prevention research.*



Project Rationale and Summary

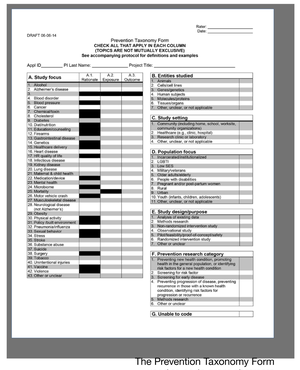
Existing NIH tools for portfolio analysis include the RCDC system, a computerized process that categorizes and reports NIH spending in 233 categories. While prevention is one of those categories, further characterization of grants within prevention is limited. RCDC does not provide sufficient detail to classify studies into the approximately 200 diseases, conditions, and research topics that the ODP has identified as being of interest to the NIH prevention research community. Manual categorization is impractical due to the large volume of grants to be reviewed and human factors that may introduce variability in the coding. Therefore, the ODP sought a computer-based solution to enable an automated, standardized, rapid, and objective characterization of NIH prevention research funding. This tool will enable identification of patterns and trends in NIH prevention research funding, as well as research areas that may benefit from targeted investments by the NIH Institutes and Centers. Such categorization will enable assessment of the progress and changes in NIH-funded prevention research over time to inform program planning and reporting.

Methods

Developing this portfolio analysis tool is a multi-step process involving four partner organizations—the ODP at the NIH, the ODP's contractor IQ Solutions, Inc. (IQS), the NIH Office of Portfolio Analysis (OPA), and the NIH Center for Information Technology (CIT). The tool development process is expected to take about two years to complete and involves:

Developing a taxonomy for coding abstracts: *The Prevention Taxonomy: a framework for classifying research grants, developed by the ODP*

- First draft of the taxonomy was developed through consultation with the trans-NIH Prevention Research Coordinating Committee (PRCC).
- Refined based on feedback from 3–4 trained coders who applied the taxonomy to >200 abstracts previously categorized as "Prevention" by RCDC.
- Six categories: Study focus, Entities studied, Study setting, Population focus, Study design, and Prevention research category.
- Within each category are topics; one or more topics may apply to any given abstract.
- Taxonomy protocol: provides instructions, definitions, and examples to facilitate correct, standardized use of the taxonomy by coders.



The Prevention Taxonomy Form used to code grant abstracts

Coding a training set of exemplars:

- Staff at IQS will apply the taxonomy to thousands of NIH R01 grant abstracts.
- ODP staff will provide instruction and supervision.
- The ODP will verify accuracy of coding on a subsample of abstracts.
- The resulting manual coding will serve as a training data set with both positive and negative examples of Prevention studies.

Using manually coded data to inform the Portfolio Learning Tool (PLT):

- The OPA will use the training data exemplars as "gold standards."
- Using machine learning, the PLT will train a model representing each category in the taxonomy and retrieve grants meeting the model criteria from a larger set of candidates from IMPAC II.

Validating the PLT output:

- Identification of a modest selection of Type 1 (false positive) and Type 2 (false negative) errors in the retrieval set of grants will be used to enrich the training set.
- The PLT will be re-run against the refined training set until validation criteria are satisfied.

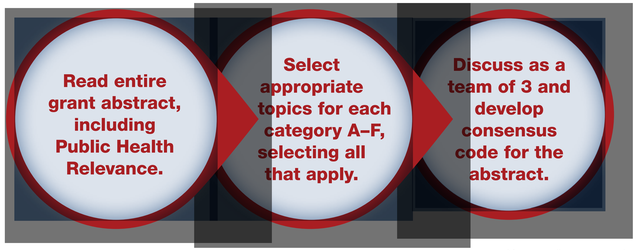
What Is the Portfolio Learning Tool?

The PLT is a prototype application for classifying NIH research grants, contracts, intramural projects, and interagency agreements. It has been under development for a few years and is approaching a beta rollout. The PLT allows users to train a model of a research category of interest and then retrieve a list of all grants in IMPAC II that are deemed to be relevant to that category. Users supply the PLT with a training set containing a list of positive examples of the category of interest as well as a list of negative examples. Grants are represented in the PLT as indexed vectors of the scientific concepts extracted from the title, abstract, and specific aims sections of the application. The PLT incorporates an enhancement to the popular Support Vector Machine learning algorithm that improves classification performance. When building a retrieval set, the PLT compares the "feature vector" representation of a candidate grant with the model representation and provides the user with a score representing the distance of that grant to the decision boundary. Although the retrieval result also includes a suggested threshold value that optimizes the performance in the training set, users may choose their own threshold.

How Are Prevention Research Exemplars Identified?

Developing the training data set for the PLT is the most time-consuming and staff-intensive aspect of the project. Deriving a set of taxonomy topics applicable to a given abstract requires having three taxonomy experts simultaneously read a funded NIH grant abstract and independently select applicable topics for each of the six categories on the taxonomy form. After each individual has completed his/her coding of the abstract, the group discusses their selections and develops a consensus for how to best classify the abstract according to the prevention research taxonomy. Throughout the individual and consensus coding process, the protocol is consulted for guidance. It provides descriptions of each topic, keywords that may be used in abstracts, and examples. Coders are not permitted to use any other resources to inform their coding.

The ODP maintains vigilant oversight of the IQS manual coding process because it is fundamental to the PLT's ultimate ability to characterize the NIH's prevention research portfolio. In fact, the ODP implements an identical independent coding process for a subset of abstracts in order to compare the consensus codes from IQS coders to the consensus codes from the ODP coders. Kappa statistics are calculated as an indicator of interrater reliability between each individual coder and their group's consensus as well as between the two consensus sets of codes. Kappa statistics for each coder are monitored to ensure an appropriate level of skill has been achieved before the coder is permitted to code abstracts for the training data set. Abstracts for which the kappa statistics indicate little agreement among coders will be flagged for coding and review by the ODP experts. Reconciliation of the IQS consensus with the ODP consensus will be done by ODP experts, and a "final consensus" resulting from reconciliation efforts will serve as the "gold standard" for the PLT.



Manual coding process using the Prevention Taxonomy

Progress to Date and Next Steps

Progress:

- The ODP has developed and refined a taxonomy and protocol for prevention research.
- The ODP has developed an instructional program for ODP and IQS staff to learn to apply the taxonomy.
- IQS has developed a computerized system to capture taxonomy coding and consensus.
- The ODP has developed quality control procedures and coder performance metrics.
- The OPA and CIT have developed a working version of the PLT.

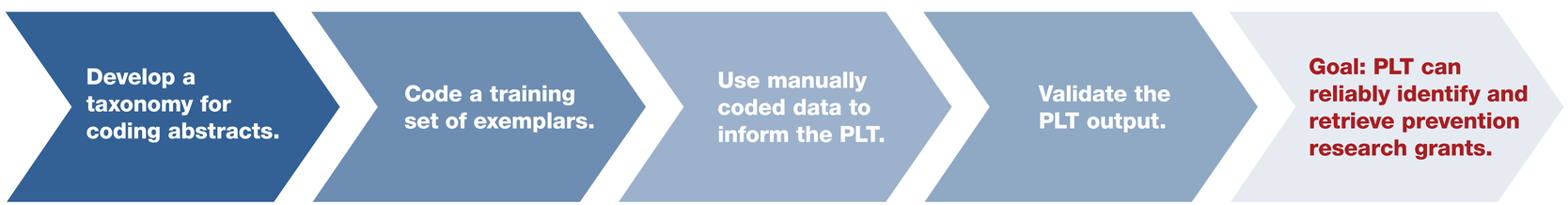
Next steps:

- The ODP and IQS are coding a large number of exemplars as training data.
- The OPA will use the training data to inform and train the PLT.
- The ODP and IQS will review and validate the PLT output to refine the PLT algorithm.
- Upon completion (estimated 2 years), the tool will be used to explore the NIH's investments in prevention research.

Follow ODP @NIHprevents

prevention.nih.gov

Development Process for the Prevention Portfolio Analysis Tool



NIH National Institutes of Health, Office of Disease Prevention
NIH National Institutes of Health, Office of Portfolio Analysis
CIT Center for Information Technology
iQ solutions