Richer, deeper, stronger: New analytical approaches for enhanced analyses of the NIH behavioral and social sciences research portfolio

W.N. Elwood,[1] C.A. Johnson,[2] S.H. Jonas ,[3] R.M. Kaplan,[1] K.M. Kulinowski,[3] W.W. Lau,[2] E.L. Stover,[1] & E.M. Talley,[4] [1]NIH/OD/DPCPSI/OBSSR, [2]NIH/CIT/DCB/HPCIO, [3]IDA/STPI, [4]NIH/NINDS/DER/ERP

Abstracts

To enhance the understanding of NIH's portfolio in behavioral and social sciences research (BSSR), we engaged program directors (PDs) to categorize NIH BSSR at a level of granularity that is finer than is available through RCDC. The charge of RCDC is to report research funding transparently and comprehensively, not to report on every emerging scientific direction or trans-NIH initiative. There are currently two relevant broad and trans-NIH categories, namely behavioral and social sciences research (BSSR) and basic behavioral and social sciences research (b-BSSR). The focus of the current pilot investigation is on five BSSR-centric categories, namely decision sciences, social epidemiology, measurement development, and mobile health (mHealth). We have developed classifiers that were trained on the entire NIH extramural portfolio including grants selected by eleven Institutes as representative to each proposed category. The results of these classifiers were then sampled and validated by the participating PDs. We intend to access available personal PI information to determine whether and how NIH's overall BSSR portfolio contributes to ameliorating recent, publicly-reported extramural research funding disparities. Both qualitative and quantitative results will be reported that reflect the most significant trends—as well as how results will be used for strategic planning across the agency.

Introduction

Our goal is to document the state and trends in NIH-supported behavioral and social sciences research (BSSR) from FY 2007 through 2011. BSSR consists of multiple fields and disciplines that do not appear in RCDC. A subgroup of the BSSR Coordinating Committee identified four categories not currently represented in RCDC to serve as subjects of a pilot for this method. The most frequently occurring topics were *decision sciences, measurement development, mobile (m-) health*, and *social epidemiology*. We added *prevention* as a fifth term given its use in RCDC as well as BSSR.

With expert help from the members of the Behavioral and Social Sciences Coordinating Committee, we created definitions for each of these scientific areas.

Representatives from 11 ICs volunteered to examine their respective portfolios using these definitions. They collectively provided over 1200 grants (over two rounds) that they found to be most representative of each research topic.

BSSR pilot categories:

Decision sciences *(DS)*: Judgment and (medical) decision-making; decision neuroscience, behavioral economics, reward processing.

Measurement development *(MD)*: Activities akin to PROMIS and Toolbox efforts, measures of sociobehavioral phenomena for assays, labs, trials, assays.

M-health *(mH)*: Use of mobile technology to capture real-time data longitudinally, date- and time-stamped; and personalized information to study behavioral, clinical, and social research.

Prevention *(P)*: Design, implementation, and evaluation, and social interventions to prevent occurrence, recurrence, or progression of health problems.

Social epidemiology *(SE)*: Social-behavioral influences on population health.

Methods and Materials

ICs coordinated to identify five categories of BSSR research to use in a pilot: Decision Science, Measurement Development, Mobile (m)Health, Prevention, and Social Epidemiology. These categories represent an emerging area of research (mHealth), a broader, RCDC (NIH-wide) defined area (Prevention), and other boutique categories that are prevalent in BSSR.

Eleven ICs that represent ~80% (RCDC-FY2011) of all BSSR-funded research volunteered to participate in the pilot.

Training Data Collection

- Nearly 600 annotated funded grant applications (FGAs) were collected across the 11 ICs
- Each FGA was annotated on the degree of representativeness for each of the 5 pilot categories

Algorithm 1	Algorithm 2
• An adaption of libSVM • Trained on positive and negative samples in a feature space that included direct text from: o Weighted titles o Abstracts o Specific aims o N-grams (single word strings to many word strings) o Topics derived from NIHmaps.org	• An adaptation of libSVM. • Trained on positive and negative samples as well as a large number of "unknowns", which were sampled at random from funded Type 1 and Type 2 projects across all ICs over the same time period as the training data. • Retrieval performed on an "or" ensemble of three classifiers as described above with three different feature spaces. The score assigned was the maximum score of the three feature spaces: i) RCDC Thesaurus; ii) topics derived from NIHmaps.org; and iii) a merger of I and ii.

Sampling for Expert Validation

For each category:

- Classified FGAs from each algorithm were independently ranked by score
- The retrieval lists were ranked such that the same number of high-scoring FGAs was evaluated by each algorithm
- Lists were combined into a two dimensional space such that the coordinate of each classified FGA is (x,y) = (Score-A1, Score-A2)
  - Partitions were demarcated by the median score from each list
  - The orthogonal median bisectors form four quadrants with the following scores:
    - Q1 (High, High)
    - Q2 (Low, High)
    - Q3 (Low, Low)
    - Q4 (High, Low)
- Samples for expert validation were taken from each of the four quadrants

Each of the 11 ICs were:

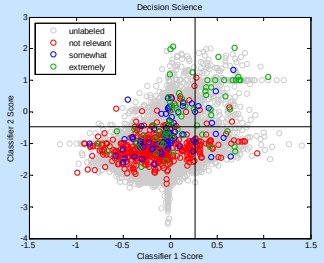
- Given 60 FGAs from their IC to validate (660 total)
  - 12 FGAs from each category, where possible
  - 3 FGAs from each quadrant, where possible
- Asked to annotate each FGAs for every category (i.e., 5 responses per FGA)

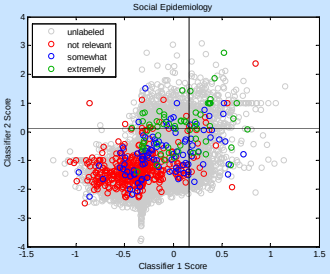
This diagram is a schematic illustration of the process used to develop the classifiers for the five BSSR and basic BSSR subcategories. This same process was followed for the two classification algorithms, which were developed independently on the same data with the same expert annotations.

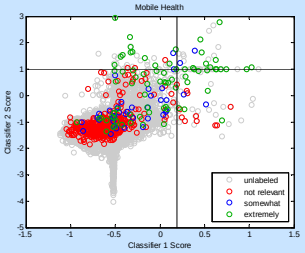
In the training round of expert annotation, participants from 11 ICs rated nearly 600 FGAs as either extremely relevant, very relevant, somewhat relevant, or not relevant to each of the subcategories. (Note: each FGA was rated against all five subcategories). In the validation round, participants annotated 660 FGAs, and the 'very relevant' rating was also dropped. The table below provides the number of FGAs in each rating for both training and validation.

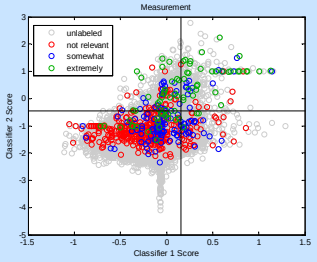
Cat.	Training				Validation		
	Extr.	Very	Some.	Not	Extr.	Some.	Not
DS	107	55	42	283	67	68	450
SE	114	20	57	283	63	99	422
mH	83	23	19	344	78	39	396
MD	107	43	58	278	78	116	396
P	119	49	66	258	125	120	339

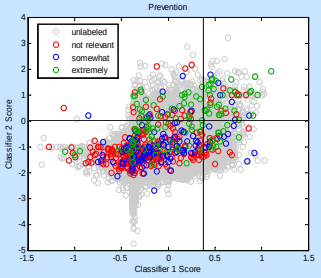
The following scatter plots compare classifier scores against expert ratings on the retrieved data including the validated FGAs.











The top table in this panel provides classification performance metrics against the expert validations when only 'Extremely Representative' ratings are taken as positive. The lower table corresponds to either 'Extremely Representative' or 'Somewhat Representative' projects being considered positive. Recall is the proportion of positives that were retrieved. Precision is the proportion of retrievals that were positive. The F-score is the harmonic mean of recall and precision (F=2*R*P/(R+P)). AUC is the area under the recall-precision curve. The scores shown for recall, precision, and F were taken at the best performing threshold.

Cat.	Classifier 1				Classifier 2			
	Recall	Prec.	F	AUC	Recall	Prec.	F	AUC
DS	0.30	0.41	0.34	0.31	0.52	0.51	0.52	0.54
SE	0.79	0.28	0.41	0.34	0.60	0.54	0.56	0.49
mH	0.60	0.53	0.56	0.54	0.61	0.70	0.65	0.65
MD	0.53	0.30	0.38	0.32	0.63	0.48	0.55	0.54
P	0.56	0.40	0.47	0.42	0.67	0.55	0.61	0.52

Cat.	Classifier 1				Classifier 2			
	Recall	Prec.	F	AUC	Recall	Prec.	F	AUC
DS	0.69	0.38	0.49	0.40	0.62	0.56	0.59	0.62
SE	0.65	0.58	0.61	0.61	0.71	0.61	0.66	0.64
mH	0.55	0.71	0.62	0.63	0.73	0.63	0.68	0.71
MD	0.70	0.50	0.59	0.56	0.59	0.56	0.58	0.63
P	0.87	0.49	0.63	0.60	0.78	0.55	0.65	0.63

Discussion and Conclusions

This pilot work has identified five key areas of BSSR and validated the utility of the candidate algorithms. Next steps involve incorporating grants from the remaining 13 ICs for a comprehensive analysis and consulting with program directors on additional emerging trends and terms for future analyses. Algorithmic enhancements to be evaluated include combining Classifier 1 and Classifier 2 into a super-ensemble, and augmenting the feature-space with IMPAC II data.

When evaluating classifier performance, it can be useful to consider inter-rater agreement as an upper bound of sorts on expected performance. The table below provides inter-rater agreement on 10 projects from one participating IC. As another point of comparison, relevant RCDC category performance on these data was similar or modestly lower than novel subcategory classifiers.

DS	SE	mH	MD	P
9/10	7/10	8/10	7/10	8/10

This project demonstrates a collaborative approach that empowers tracking of scientific trends within existing NIH data. The methods described here can be used to obtain a more complex and detailed understanding of two scientific research areas (BSSR and basic BSSR) than RCDC or other central resources were designed to provide.

Acknowledgments

The authors are indebted to the contributions of Guoli Wang and Arun Ravindran (CIT); Gina Walejko, Rashida Nek, Gilbert Watson, Mario Nunez, and Alyson Wilson (IDA); and participants from 11 ICs: AA, AG, AT, CA, DA, DC, DK, EY, HD, HL, and MH.

This project has been funded in part with federal funds from the Office of Behavioral and Social Sciences Research (OBSSR) and the Office of Program Evaluation and Performance (OPEP) Division of Program Coordination, Planning, and Strategic Initiatives, National Institutes of Health, under contract HHSN276201100441U. This project was supported in part by the Center for Information Technology's Intramural Research Program. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

http://dcb.cit.nih.gov/hpcio/OBSSR_Researchfest_2012.pdf