



HHS Public Access

Author manuscript

Am J Prev Med. Author manuscript; available in PMC 2019 December 01.

Published in final edited form as:

Am J Prev Med. 2018 December ; 55(6): 926–931. doi:10.1016/j.amepre.2018.07.024.

A Machine Learning Approach to Identify NIH-Funded Applied Prevention Research

Jennifer Villani, PhD, MPH¹, Sheri D. Schully, PhD¹, Payam Meyer, MS², Ranell L. Myles, PhD, MPH, CHES¹, Jocelyn A. Lee, PhD, MPH¹, David M. Murray, PhD¹, and Ashley J. Vargas, PhD, MPH, RDN¹

¹Office of Disease Prevention, NIH, Rockville, Maryland ²Office of Portfolio Analysis, NIH, Bethesda, Maryland

Abstract

Introduction: To fulfill its mission, the NIH Office of Disease Prevention systematically monitors NIH investments in applied prevention research. Specifically, the Office focuses on research in humans involving primary and secondary prevention, and prevention-related methods. Currently, the NIH uses the Research, Condition, and Disease Categorization system to report agency funding in prevention research. However, this system defines prevention research broadly to include primary and secondary prevention, studies on prevention methods, and basic and preclinical studies for prevention. A new methodology was needed to quantify NIH funding in applied prevention research.

Methods: A novel machine learning approach was developed and evaluated for its ability to characterize NIH-funded applied prevention research during fiscal years 2012–2015. The sensitivity, specificity, positive predictive value, accuracy, and F1 score of the machine learning method, the Research, Condition, and Disease Categorization system, and a combined approach were estimated. Analyses were completed during June–August 2017.

Results: Because the machine learning method was trained to recognize applied prevention research, it more accurately identified applied prevention grants (F1=72.7%) than the Research, Condition, and Disease Categorization system (F1=54.4%), and a combined approach (F1=63.5%) with $p < 0.001$.

Conclusions: This analysis demonstrated the use of machine learning as an efficient method to classify NIH-funded research grants in disease prevention.

Address correspondence to: Jennifer Villani, PhD, MPH, Office of Disease Prevention, NIH, 6100 Executive Blvd., Rockville MD 20892. villani@nih.gov.

JV, SDS, RLM, JAL, DMM, and AJV conceptualized and designed the study. PM developed the machine learning approach. JV, SDS, RLM, JAL, and AJV collected the data. AJV led the statistical analyses. All authors contributed to this manuscript and approved the final version.

The views expressed in this paper are solely the opinions of the authors and do not necessarily reflect the official policy of NIH. No financial disclosures were reported by the authors of this paper.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

INTRODUCTION

The mission of NIH is to foster biomedical discoveries and apply that knowledge to “enhance health, lengthen life, and reduce illness and disability.”¹ One of the ways that the NIH advances population health is through its investment in prevention research. The NIH Office of Disease Prevention (ODP) was established in 1986 and is responsible for assessing, facilitating, and stimulating research in disease prevention and health promotion.² Achieving these goals requires a detailed and precise understanding of the agency’s investment in prevention research to inform decisions about scientific program planning.

Since 2008, the NIH has used the Research, Condition, and Disease Categorization (RCDC) system to report funding across 282 categories of biomedical research.³ The RCDC system was designed to provide consistent and transparent information to the public about the level of support in each category. It uses text mining to distill the terms and concepts in each funded grant’s title, abstract, specific aims, and public health relevance statement. It ranks the terms that have a match in the RCDC thesaurus, and each grant is classified into RCDC categories via an expert-defined process when a threshold score is met.⁴ NIH scientific experts validate a sample of the grants in each category using a trans-NIH framework and have the option of nominating grants that were missed by the RCDC system for a given category.

The RCDC system defines prevention research to include studies of primary and secondary prevention, prevention-related methods, as well as basic and preclinical studies that have a prevention focus; by contrast, the ODP definition is more nuanced and includes only applied research in humans limited to primary and secondary prevention and prevention-related methods.⁵ Primary prevention includes the study of risk and protective factors for the onset of a new health condition, and interventions delivered to prevent a new health condition. Secondary prevention includes the identification of risk and protective factors for the recurrence of a health condition, early detection of an asymptomatic health condition, and interventions to prevent recurrence or slow progression of a health condition. In addition, RCDC Categorical Spending reports exclude grants funded using non-NIH appropriations, such as grants awarded under the Tobacco Regulatory Science Program, which uses funds from the Food and Drug Administration. Grants awarded with funds from gifts to NIH, Breast Cancer Stamps, the President’s Emergency Plan for AIDS Relief, the Assistant Secretary for Preparedness and Response, and any funds transferred to NIH from other federal agencies are also excluded from RCDC reports. The ODP wanted to characterize prevention research supported by NIH regardless of the funding source.

To address these issues, the ODP collaborated with the NIH Office of Portfolio Analysis to develop a new method for portfolio analysis that uses machine learning (ML). This paper describes the process used to evaluate three methods for identifying prevention research grants funded at the NIH based on the ODP definition. Results from (1) the RCDC system, (2) the ODP ML method, and (3) a hybrid approach utilizing both methods are presented.

METHODS

Data

The most common type of research grant awarded by NIH is the R01, which represents about half of all new NIH-funded research project grants.⁶ Therefore, this study focused on new R01 grants (i.e., 1R01s) awarded by the NIH in fiscal years 2012–2015 (FY2012–2015). Three discrete methods for classifying 1R01s as prevention research according to the ODP definition were tested.

To create the gold standard dataset against which to measure the performance of the three classification methods, a team of three research analysts read the title, abstract, and public health relevance statement of each 1R01 and individually coded each grant as prevention or non-prevention according to the ODP definition. Then, in a triad, the research analysts were required to reach a consensus on whether each 1R01 was prevention or non-prevention research. A separate team of NIH staff scientists blindly reviewed 10%–20% of all coded grants using the same coding process. Both the research analysts and the NIH staff scientists underwent 8 weeks of training and were required to demonstrate 70% agreement on identifying prevention grants before beginning to code grants. The joint probability of agreement was calculated comparing the research analysts' consensus decision to the NIH staff scientists' consensus decision. Disagreements between the research analysts' team and the NIH staff team were resolved through discussion to reach a final consensus for each 1R01 grant that was reviewed. The data from this manual coding process were used as the gold standard.

Research, Condition, and Disease Categorization Grant Classification Method

The RCDC system was assessed as a method (herein called the RCDC method) for classifying grants as prevention or non-prevention according to the ODP definition, though the grants were identified using the RCDC definition of prevention. Although the RCDC method provided predictions for all NIH 1R01s from FY2012 to 2015, it was not feasible to manually code all 14,566 grants. Therefore, the predictions were verified by manually coding 100% of 1R01s that were classified as prevention research by the current RCDC method and a 5% random sample of the 1R01s that were classified as non-prevention for FY2012–2015 ($n=3,814$; gold standard).

To reduce the classification error that was expected based on the different definitions of prevention between RCDC and the ODP, the results of the RCDC method were subsequently restricted by removing the 1R01 grants that were flagged during NIH peer review as involving only animal subjects. As above, classifications of the resulting dataset were validated by manually coding 100% of grants still classified as prevention research and a random 5% sample of the grants classified as non-prevention research ($n=3,005$; gold standard).

Machine Learning Grant Classification Method

An ensemble of ML algorithms (herein called the ML method) was used to classify grants as prevention research or non-prevention according to the ODP definition. The ML algorithms

were trained using the previous FY's manually coded grants (training set, $n=4,529$ across FY2012–2015; average text length of 2,729 characters), then each grant in the FY of interest (test set, $n=14,566$ across FY2012–2015) was run through each algorithm (Figure 1). Because of randomization differences in initial values for the algorithms, it is possible for them to make different predictions for the same grant after different training runs on the same data set. To accommodate this variation, each algorithm was trained and tested five times. The results of each run through an algorithm were either prevention or nonprevention for each grant. Initial experiments with these algorithms showed differences in their predictive ability and further indicated that aggregating the predictions of the individual algorithms would improve overall classification of prevention grants. Therefore, an ensemble classifier was used to predict prevention research and non-prevention research grants using the majority of predictions.^{7,8} Predictions were made on all 14,566 1R01 grants from FY2012 to 2015, but it was not possible to manually code all of them. Instead, the results were validated by manually coding 100% of 1R01s that were classified as prevention research by the ML method and a 5% random sample of the remaining 1R01s ($n=3,572$; gold standard).

Multiple algorithms (e.g., support vector machines, LIBLINEAR, random forest, k-nearest neighbors, OpenNLP maximum entropy, neural networks, scaled relative frequency ratios)^{9,10} were tested early in the process on a set of NIH grants. The four algorithms that consistently performed the best were selected— LIBLINEAR (F1 score, 81.8%),¹¹ OpenNLP MaxEnt (F1 score, 83.0%),¹² neural networks (F1 score, 84.3%),¹³ and scaled relative frequency ratios (F1 score, 80.0%; Appendix, available online, provides parameters). During pilot testing, the predictive ability from the neural networks algorithm was superior, so it was used again as a tie-breaker, bringing the total number of algorithms in the ensemble to five. Running the test set through each algorithm five times resulted in a total of 25 predictions per grant. If 13 of 25 total runs predicted that the grant was prevention, then the grant was classified as prevention; otherwise it was classified as non-prevention.

Combined Research, Condition, and Disease Categorization and Machine Learning Grant Classification Method

The RCDC and ML approaches were combined as a method (herein called the RCDC–ML method) for classifying grants as prevention or non-prevention. Any grant classified as prevention by either method was classified as prevention for the combined method, whereas those that were not classified as prevention by either method were classified as non-prevention for the combined method. Again, predictions were made for all 14,566 1R01 grants from FY2012 to 2015. The results were verified by manually coding 100% of 1R01s that were classified as prevention research using the RCDC–ML method, and a 5% random sample of the remaining 1R01s ($n=4,986$; gold standard).

Statistical Analysis

The grants in each method's random 5% sample of coded non-prevention grants were weighted and their prevention coding was extrapolated to reflect the entire non-prevention portfolio of 1R01s. The observed true positives and false positives with the extrapolated true negatives and false negatives were used to calculate the sensitivity, specificity, positive

predictive value (PPV), accuracy, and F1 score for each method. In the ML field, sensitivity is known as “recall” and PPV is known as “precision.” The harmonic mean of these two metrics is the F1 score and is often used in ML as a single measure of an algorithm’s overall performance.^{14,15} In this analysis, sensitivity represents the proportion of prevention grants that were identified as such by the method. Specificity is the proportion of non-prevention grants that were identified as such by the method. PPV is the probability that grants identified by a method as prevention were truly prevention awards. Accuracy is the proportion of grants correctly identified as either prevention or non-prevention by the method.

F1 results for each approach to classifying the 1R01 grants (RCDC versus ML versus RCDC–ML) were bootstrapped¹⁶ 5,000 times to generate a normal distribution of F1 scores using RStudio, version 1.0.136, and the results were compared using paired *t*-tests to test for significance in a pair-wise fashion using Stata, version 15.

RESULTS

A total of 14,566 NIH 1R01 grants were awarded during FY2012–2015 at the time data were pulled for this analysis. Manual coding of these grants was completed over a period of 15 months. The initial joint probability of agreement between the research analysts and the NIH staff scientists was >0.7, indicating good concurrence between these two teams. Disagreements were discussed in-person by both teams for all validated grants until a final agreement was reached by consensus. These discussions resulted in prevention coding changes for \cong 3% of the validated grants.

RCDC classified 3,227 as prevention research using the RCDC definition and the remaining 11,339 grants as non-prevention research. All of the prevention grants identified by the RCDC method and a 5% random sample ($n=587$) of the RCDC method’s non-prevention grants were manually coded using the ODP definition of prevention research. For FY2012–2015, the average sensitivity was 57.2% and the average specificity was 86.6% when grants identified using the RCDC definition of prevention were evaluated against the ODP definition (Table 1). In addition, the average accuracy was 80%, the average PPV was 51.8%, and the average F1 score was 54.4%. These results reflect both the commonalities and differences between the ODP and RCDC definitions of prevention research.

To better align the RCDC method with the ODP definition of prevention, all 1R01 grants involving only animal subjects were automatically classified as non-prevention. This modification improved the PPV and the F1 score to 70.0% and 62.6%, respectively. It was also more specific (93.9%) and more accurate (86.3%) than the RCDC method, but not as sensitive (56.6%).

Table 1 shows the classification results for the ensemble ML algorithm. The ML method classified a total of 2,994 grants as prevention research using the ODP definition and the remaining 11,572 grants as non-prevention research. All of the prevention grants and a 5% random sample ($n=578$) of the non-prevention grants identified by the ML method were manually coded. For the ML method, average sensitivity was 71.5% and average specificity

was 93.2%. The ML method achieved the highest average accuracy (88%), PPV (73.9%), F1 score (72.7%) of all the methods tested.

Table 1 also shows the classification results for the combined, RCDC–ML approach. All 1R01 awards that were identified as prevention by either the RCDC or ML methods were considered prevention awards ($n=4,481$) and awards that were not identified by either method were considered non-prevention awards ($n=10,085$). As before, all of the prevention grants and a 5% random sample ($n=505$) of the non-prevention grants identified by the RCDC–ML method were manually coded. This combined approach yielded the highest average sensitivity, at 78.4% (a result guaranteed by the method of combining), but the lowest average specificity, at 81.8%. The average PPV was low as well, at 53.3%, and the average F1 score was 63.5% with an average accuracy of 81%.

The predictions from all three methods were sampled with replacement 5,000 times and then the resulting F1 values were compared using post-hoc paired t -tests. Statistically significant differences in the F1 values of the ML method, the RCDC method, and the combined RCDC–ML method were observed (all $p<0.001$), indicating that the F1 values were higher for the ML method than the other two methods.

DISCUSSION

Since 2008, NIH has utilized the RCDC system to provide consistent and transparent information to the public about the level of support across 282 research categories. This system provides a wealth of information across topics, but does not cover all of the prevention topics of interest to the ODP and may not define individual topics in the same nuanced way that the ODP defines them. The observed sensitivity and specificity of using RCDC alone to classify prevention research in 1R01s across NIH suggests that the current RCDC categorization for prevention is not adequate for addressing the needs of the ODP. Given the differences in the definition of prevention research between the ODP and RCDC, it was not surprising that the sensitivity, specificity, PPV, accuracy, and F1 score of the ML method were higher than the RCDC method. This held true even when grants involving only animal studies were removed from the RCDC-identified pool of grants.

Prevention is a broad and complex topic that spans research in etiology, epidemiology, and intervention effectiveness to reduce morbidity and mortality. Many have struggled with how to define the concept of prevention.^{17,18} This ambiguity makes categorizing it across the NIH grant portfolio challenging. Therefore, the ML method and manual coding process used in this study were designed specifically for identifying primary and secondary prevention research grants involving human subjects, along with prevention-related methods grants, across the NIH extramural 1R01 grant portfolio. By leveraging this method, the ODP was able to curate a large database of research funding despite the vast amount of data, and the heterogeneity of the research supported. This method can serve as a blueprint for quantifying research support in other complex scientific areas, like health disparities and social and behavioral research.

Because the ML method described here relies on recognizing patterns in large text data sets, it can also be adapted for other prevention applications. For example, ML algorithms can predict adverse birth outcomes,¹⁹ and identify soldiers at risk for sexual assault perpetration.²⁰ Rose²¹ used a similar ensemble ML approach to predict mortality risk in an elderly California population. ML approaches have also been used by researchers to study the etiology of depression,²² predict recurrence of complex diseases,²³ and detect interactions between risk factors for complex diseases.²⁴

Limitations

A limitation of this study for the ML method is that it relied solely on the text provided by the principal investigator in each grant's title and abstract. Although it was usually possible to discriminate prevention research from non-prevention research based on the grants' abstracts, there were some grants that did not describe the proposed research clearly enough in those sections of the application to enable this type of discrimination. The ML algorithms cannot be expected to perform better given the same text. Another potential limitation is human error in the consistency of the manually coded data used as gold standards for evaluating the classification methods. To ensure a high degree of agreement, grants were coded by teams of three, with each coder (i.e., research analysts and NIH staff scientists) completing an extensive training and passing a test before joining a team of coders. Research analysts, who manually coded the grants, and NIH staff scientists, who validated the manually coded grants, met on a weekly basis to discuss any discrepancies and reach a consensus on the final coding decision of prevention or non-prevention for all validated grants.

CONCLUSIONS

The application of ML algorithms is an efficient method to characterize prevention research funding at NIH. The ODP plans to extend this ML approach to additional grant mechanisms in order to characterize the broader NIH applied prevention research portfolio according to the ODP definition. Additionally, work is underway to train the ML algorithms to identify finer details about individual prevention research grants based on study characteristics, such as the health conditions measured, the populations studied, the research designs used, and the types of prevention research examined. The goals of these efforts are to reveal trends in NIH-funded prevention research, and to identify prevention research gaps that could benefit from additional investments.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors would like to thank Dr. George Santangelo, Dr. Kirk Baker, and Matthew Davis of the NIH Office of Portfolio Analysis for their conceptual contributions, machine learning expertise, and technical support in this project. We would also like to acknowledge Dr. Pamela Carter-Nolan and Dr. Lamyaa Yousif, and the research analysts who skillfully coded the NIH grants in our dataset—Dr. Shahina Akter, Arielle Dolegui, Dr. Luis Ganoza, Allison Gottwalt, Alexis Hall, Jimmy Ngo, Adeola Olufunmilade, Priti Patel, Agnieszka Roman, Rachel Tolbert, Dr. Sundeep Vikraman, and Taylor Walter. Finally, we would like to recognize David Tilley and Kat Schwartz from

the NIH Office of Disease Prevention, and Dr. Richard Ikeda, Judy Riggie, and the rest of the NIH Research, Condition, and Disease Categorization team for their comments which greatly improved the manuscript.

REFERENCES

1. NIH. Mission and goals. www.nih.gov/about-nih/what-we-do/mission-goals. Updated July 27, 2017. Accessed December 19, 2017.
2. NIH Office of Disease Prevention. About Us. <https://prevention.nih.gov/about>. Updated March 16, 2017. Accessed December 19, 2017.
3. NIH Office of Extramural Research. Estimates of funding for various Research, Condition, and Disease Categories (RCDC). https://report.nih.gov/categorical_spending.aspx. Published July 3, 2017. Accessed December 19, 2017.
4. NIH Office of Extramural Research. Categorization process. <https://report.nih.gov/redc/process.aspx>. Published May 16, 2012. Accessed December 19, 2017.
5. Murray DM, Cross WP, Simons-Morton D, et al. Enhancing the quality of prevention research supported by the National Institutes of Health. *Am J Public Health*. 2015;105(1):9–12. <https://doi.org/10.2105/AJPH.2014.302057>. [PubMed: 25393184]
6. NIH Office of Extramural Research. Funding Facts. <https://report.nih.gov/fundingfacts/fundingfacts.aspx>. Published August 29, 2014. Accessed December 19, 2017.
7. Dietterich TG. Ensemble methods in machine learning. In: Kittler J, Roli F, eds. *Multiple Classifier Systems Vol 1857* 2000:1–15. https://doi.org/10.1007/3-540-45014-9_1.
8. Dong Y-S, Han K-S. A comparison of several ensemble methods for text categorization. *Proceedings of the 2004 IEEE International Conference on Services Computing 2004* <https://doi.org/10.1109/SCC.2004.1358033>.
9. Harrington P *Machine Learning in Action*. Shelter Island, NY: Manning; 2012.
10. Patterson J, Gibson A. *Deep Learning: A Practitioner's Approach*. Sebastopol, CA: O'Reilly Media; 2017.
11. LIBLINEAR: A library for large linear classification. www.csie.ntu.edu.tw/~cjlin/liblinear/. Accessed August 11, 2017.
12. Apache OpenNLP. <http://opennlp.apache.org/>. Accessed August 11, 2017.
13. Deeplearning4j: Open-source, distributed deep learning for the JVM. <https://deeplearning4j.org/>. Accessed August 11, 2017.
14. Carrell DS, Halgrim S, Tran DT, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. *Am J Epidemiol*. 2014;179(6):749–758. <https://doi.org/10.1093/aje/kwt441>. [PubMed: 24488511]
15. Zhang Q, Yu H. Computational approaches for predicting biomedical research collaborations. *PLoS One*. 2014;9(11):e111795 <https://doi.org/10.1371/journal.pone.0111795>. [PubMed: 25375164]
16. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci*. 1986;1(1):54–75. <https://doi.org/10.1214/ss/1177013815>.
17. Starfield B, Hyde J, Gervas J, Heath I. The concept of prevention: a good idea gone astray? *J Epidemiol Community Health*. 2008;62(7):580–583. <https://doi.org/10.1136/jech.2007.071027>. [PubMed: 18559439]
18. Ataguba JE, Mooney G. Building on “The concept of prevention: A good idea gone astray?”. *J Epidemiol Community Health*. 2011;65(2):116–118. <https://doi.org/10.1136/jech.2008.082818>. [PubMed: 19934170]
19. Pan I, Nolan LB, Brown RR, et al. Machine learning for social services: a study of prenatal case management in Illinois. *Am J Public Health*. 2017;107(6):938–944. <https://doi.org/10.2105/AJPH.2017.303711>. [PubMed: 28426306]
20. Rosellini AJ, Monahan J, Street AE, et al. Predicting sexual assault perpetration in the U.S. Army using administrative data. *Am J Prev Med*. 2017;53(5):661–669. <https://doi.org/10.1016/j.amepre.2017.06.022>. [PubMed: 28818420]
21. Rose S. Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol*. 2013;177(5):443–452. <https://doi.org/10.1093/aje/kws241>. [PubMed: 23364879]

22. Roetker NS, Page CD, Yonker JA, et al. Assessment of genetic and nongenetic interactions for the prediction of depressive symptomatology: an analysis of the Wisconsin Longitudinal Study using machine learning algorithms. *Am J Public Health*. 2013;103(suppl 1):S136–S144. <https://doi.org/10.2105/AJPH.2012.301141>. [PubMed: 23927508]
23. Abreu PH, Santos MS, Abreu MH, et al. Predicting breast cancer recurrence using machine learning techniques: a systematic review. *ACM Comput Surv*. 2016;49(3):52 <https://doi.org/10.1145/2988544>.
24. Koo CL, Liew MJ, Mohamad MS, et al. A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. *Biomed Res Int*. 2013;2013:432375 <https://doi.org/10.1155/2013/432375>. [PubMed: 24228248]
25. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. *CoRR: Computing Research Repository*. 2013:1–12.
26. Smith L, Rindfleisch T, Wilbur WJ. MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics*. 2004;20(14):2320–2321. <https://doi.org/10.1093/bioinformatics/bth227>. [PubMed: 15073016]
27. NIH National Library of Medicine. Medical subject headings. www.nlm.nih.gov/mesh/. Published September 1, 1999. Updated November 14, 2017 Accessed December 19, 2017.
28. Manning CD, Schütze H. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press; 1999.

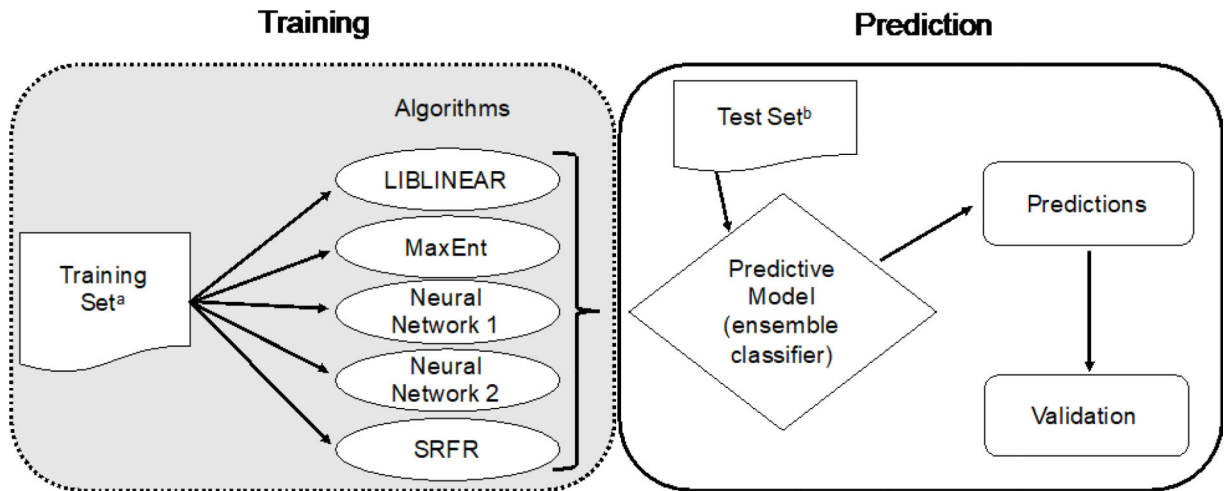


Figure 1.

Overview of the machine learning grant classification method.

^aEach training set was derived from manually coded grants from the previous fiscal year.

^bEach test set was the entire NIH 1R01 portfolio for that specific fiscal year.

SRFR, Scaled Relative Frequency Ratio

Table 1.

Performance of Three Approaches to Identify NIH-Funded Prevention Research in Grants from FY2012–2015, Mean

Method	Sensitivity, %	Specificity, %	PPV, %	Accuracy	F1, %
RCDC	57.2	86.6	51.8	80	54.4
ML	71.5	93.2	73.9	88	72.7
RCDC–ML	78.4	81.8	53.3	81	63.5

ML, machine learning; PPV, positive predictive value; RCDC, Research, Condition, and Disease Categorization system.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript