Metric selection and promotional language in health artificial intelligence

S. Scott Graham[1*] & Trisha Ghotra[2]

[1]Department of Rhetoric & Writing, Center for Health Communication, The University of Texas at Austin, Austin, Texas, United States of America

[2]College of Natural Sciences, The University of Texas at Austin, Austin, Texas, United States of America

*Corresponding author:
Email: ssg@utexas.edu

Abstract

Background

Recent advances in Artificial intelligence (AI) have the potential to substantially improve healthcare across clinical areas. However, there are concerns health AI research may overstate the utility of newly developed systems and that certain metrics for measuring AI system performance may lead to an overly optimistic interpretation of research results. The current study aims to evaluate the relationship between researcher choice of AI performance metric and promotional language use in published abstracts.

Methods and findings

This cross-sectional study evaluated the relationship between promotional language and use of composite performance metrics (AUC or F1). A total of 1200 randomly sampled health AI abstracts drawn from PubMed were evaluated for metric selection and promotional language rates. Promotional language evaluation was accomplished through the development of a customized machine learning system that identifies promotional claims in abstracts describing the results of health AI system development. The language classification system was trained with an annotated dataset of 922 sentences. Collected sentences were annotated by two raters for evidence of promotional language. The annotators achieved 94.5% agreement ($\kappa = 0.825$). Several candidate models were evaluated and, the bagged classification and regression tree (CART) achieved the highest performance at Precision = 0.92 and Recall = 0.89. The final model was used to classify individual sentences in a sample of 1200 abstracts, and a quasi-Poisson framework was used to assess the relationship between metric selection and promotional language rates. The results indicate that use of AUC predicts a 12% increase (95% CI: 5-19%, p = 0.00104) in abstract promotional language rates and that use of F1 predicts a 16% increase (95% CI: 4% to 30%, p = 0. 00996).

Conclusions

Clinical trials evaluating spin, hype, or overstatement have found that the observed magnitude of increase is sufficient to induce misinterpretation of findings in researchers and clinicians. These results suggest that efforts to address hype in health AI need to attend to both underlying research methods and language choice.

Introduction

Popular and scientific accounts describing the potential of Artificial Intelligence (AI) for health and medicine promise fundamental transformations in the nature and quality of care.[1,2] Accounts of the near future of health AI promise full life-span benefits from reproductive planning through end-of-life care.[1-2] as well as transformation in related areas like health policy.[3] Recently developed and currently available health AI technologies have shown great potential for cancer diagnosis,[3] intensive care unit admission prediction,[4] health policy,[5] and even mitigating systemic biases in medicine.[6] These kinds of

promising results are fueling unprecedented investments in the sector, with over $14 billion in US venture capital in 2020.[7] The enthusiasm for health AI is often warranted. However, many in medicine and bioethics are concerned that if practitioners or hospital systems put too much trust in health AI's promotional claims, it may lead to significant patient harm.[7-13] Overly enthusiastic adoption of AI may lead to misdiagnoses, medical errors, and inequitable delivery of care.

To help address these issues, researchers in medicine and bioethics are advancing new standards for health AI research and reporting.[14-16] Nevertheless, the ability to effectively identify extravagant and promotional claims will remain an important part of vetting new technological innovations. Appropriate validity methods and related performance metrics are often seen as the cornerstone of these initiatives. The ability to precisely determine the accuracy of a given AI system offers an important tool for separating the hype from reality, in terms of both AI performance and potential utility. Specificity and sensitivity, for example, have been primary metrics in diagnostic medicine for nearly 75 years.[17] These metrics are relatively intuitive ratios confusion matrix values (true positives, true negatives, false positives, false negatives). Specificity is defined by the number of true negatives over the sum of true negatives and false positives; whereas, sensitivity represents the number of true positives over the sum of true positives and false negatives. In short, these metrics offer clinicians critically important information about how likely a given test is to correctly diagnose a patient and how likely that same test is to correctly clear a patient. Other common metrics such as accuracy, precision, and recall are derived similarly based on a simple confusion matrix. However, health AI researchers increasingly use composite metrics that mathematically aggregate these ratios. Area Under the Receiver Operating Characteristic Curve (AUROC or AUC) and F1 are among the most popular. AUC is determined by plotting a ROC curve defined by sensitivity and 1-specificity and then calculating the two-dimensional area under that curve. F1 is the harmonic mean of precision and recall. While composite metrics can be helpful when comparing the performance of different candidate models built under the same framework, many in clinical medicine have expressed concerns that these metrics are often misunderstood and do not offer healthcare providers critically important information about diagnostic performance.[18-20] The underlying question of this study,

therefore, is to evaluate if use of these composite metrics might associate with the kinds of increases in promotional language that lead to overconfidence in new AI systems.

Methods

To evaluate this question, we assessed the relationship between metric selection and promotional language usage rates in a random sample of 1200 abstracts collected from PubMed. Promotional language was identified using a custom machine learning classifier trained on human-annotated sentences from previously collected abstracts reporting the results of newly developed health AI systems. In what follows, we describe our search strategy and sampling technique. Subsequently, we describe the development and validation of our promotional language classifier.

Search strategy and data collection

Our goal in this study was to curate a dataset that would allow us to assess any potential relationship between use of composite metrics and rates of promotional language related to health AI. Specifically, our aim was to collect research on health-related machine learning systems that included relevant benchmarking data in the PubMed-indexed abstract. In order to do so, we began by implementing a search strategy that identified PubMed-indexed articles containing (1) “machine learning” within the Medical Subject Heading (MeSH) ontology, and (2) one of the following terms in either the published article title or abstract: Accuracy, AUC, AUROC, F1, F-1, Negative Predictive Value, NPV, Positive Predictive Value, PPV, Precision, Recall, Sensitivity, Specificity, TNR, TPR, True Negative Rate, or True Positive Rate. This search yielded a total of 15,481 papers. Collected papers were subsequently screened for English language, presence of an abstract, and discussion of specific metrics within the abstract. Since the available PubMed search protocol bundles the title and abstract fields, this secondary screening focused on abstracts alone was necessary to locate papers that fit the study inclusion criteria. 7,421 records remained after screening, and a random sample of 1200 was extracted for

subsequent analysis. This sample size was identified prospectively in order to assure 95% power.[21] See Fig. 1 for additional details on the search strategy and selection of articles.

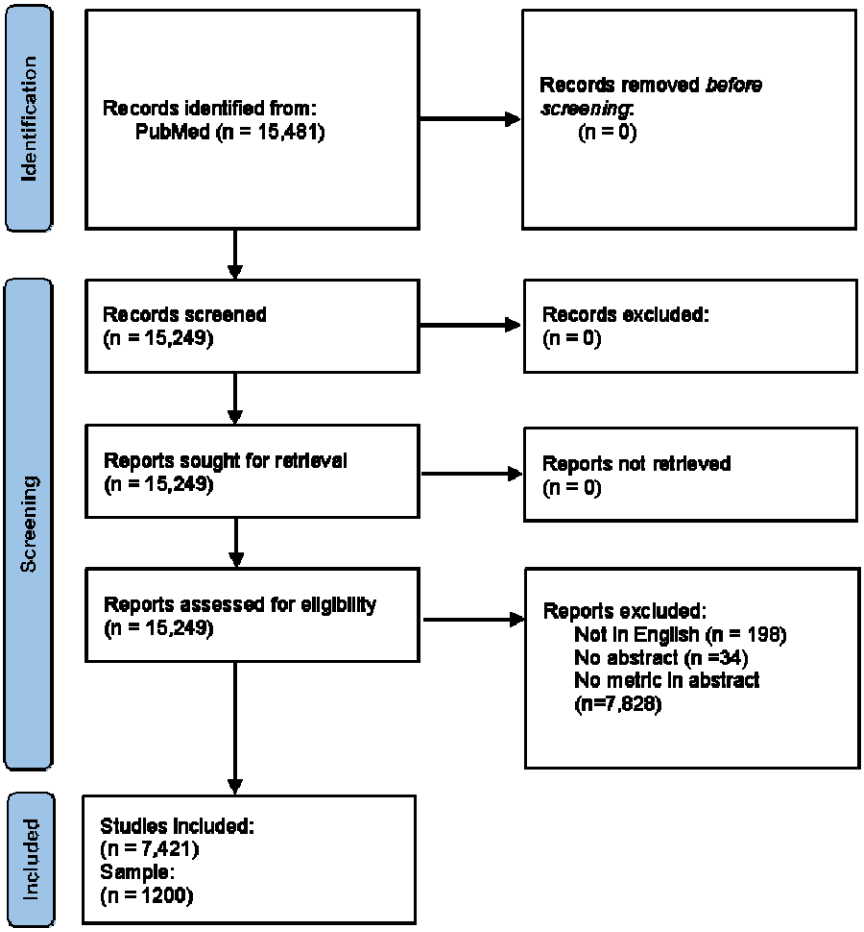

Fig. 1. Identification, Screening, and Inclusion of Studies. This flowchart details the identification, screening, and inclusion of studies for this paper. The chart is an adaptation Adapted from the PRISMA 2020 Guidelines.[22]

For each abstract in the sample, we collected data on which and how many performance metrics were being used to evaluate tested systems. Since AUC and F1 are composite metrics, they are often used alongside their constituent metrics. Each new metric presented in an abstract, therefore, affords writers the opportunity to use more promotional language. Subsequently, it was important to know if the number of metrics alone was sufficient to predict changes in the relative frequency of promotional language use. Metric identification was based on the calculation or framework rather than simply the name of the

metric. So, for example, instances of true positive rate (TPR), sensitivity, and recall were all identified as TPR since these metrics use identical calculations. Table 1 details the regular expressions used to identify and classify metrics in each abstract.

Table 1. Metric identification and classification regular expressions

Metric	Regex Query
True Positive Rate	(?i)true postive rate|TPR|(?i)sensitivity|(?i)recall
True Negative Rate	(?i)true negative rate|TNR|(?i)specificity
Positive Predictive Value	(?i)positive predictive value|PPV|(?i)precision
Negative Predictive Value	(?i)negative predictive value|NPV
Accuracy	(?i)balanced accura|(?i)percent accura|(?i)precentage accura|(((?i)acc)(\\s|:|=))
AUC	(?i)area under the curive|(?i)area under the receiver operating characteristic curve|AUC|AUROC
F1	F1|F-1

Annotation and Model Training

For the promotional language analysis, we built a natural language processing machine learning model capable of identifying promotional claims in the collected abstracts. The model was trained on human identification of promotional and non-promotional statements in research abstracts for 82 articles collected as part of a pre-existing meta-study on diagnostic AI performance.[23] The meta-study inclusion criteria selected for articles that (1) evaluated a diagnostic classification task for a specific disease, (2) used deep learning models, and (3) compared system performance to healthcare professionals. The study authors qualitatively evaluated 82 publications and included 25 in the statistical meta-analysis. Collected articles include evaluations of AI systems designed to support diagnostic imaging in oncology, dermatology, ophthalmology, cardiology, and other assorted subspecialties. To curate the training set, we searched PubMed for each article in the original meta-study and collected available abstracts. Seventy-six

of the abstracts are publicly available via PubMed. The remaining six abstracts were collected directly from publisher websites.

Collected abstracts were tokenized by sentence creating a dataset of 922 sentences for annotation. Each of these sentences was annotated by two annotators for evidence of promotional claims about developed AI system(s). The promotional language annotation was assigned when one of the following were present: (1) favorable comparisons to human annotators or previously developed health AI systems, (2) positive superlative qualifying adjectives describing the performance or efficiency of the system, (3) claims to generalizability or clinical applicability, or (4) assertions that system performance meets the standards for Food and Drug Administration medical device clearance or approval. Many sentences classified as promotional met multiple requirements for annotation. For example, it was common for favorable comparisons to include superlative qualifiers. A given sentence was only annotated as "promotional language" if it described a system under evaluation in the article. Promotional language about AI, in general, was not assigned to the "promotional" category. Additionally, claims about system performance that might be considered objectively good (e.g., AUC=.9997) were not classified as "promotional" unless the sentence also included favorable comparisons, positive superlative qualifiers, claims to generalizability/applicability, or claims to meeting regulatory standards. Further details about these features and illustrative examples are available in Table 2.

Table 2. Common features of promotional language and illustrative examples.

Promotional Feature	Description and Examples
Favorable Comparison	Sentence asserts that an ML system performs as well as or better than a qualified medical expert or previously developed ML system. • For the first time, dermatologist-level image classification was achieved on a clinical image classification task without training on clinical images. • For the whole-slide image classification task, the best algorithm (AUC, 0.994 [95% CI, 0.983-0.999]) performed significantly better than the pathologists WTC in a diagnostic simulation (mean AUC, 0.810 [range, 0.738-0.884]; P < .001).
Superlative Qualifier	Sentence uses positive-valence superlative adjectives or adverbs to

	qualify claim. • The significant improvements in diagnostic accuracy that we observed in this study show that deep learning methods are a mechanism by which senior medical specialists can deliver their expertise to generalists on the front lines of medicine, thereby providing substantial improvements to patient care. • Further, RADnet achieves higher recall than two of the three radiologists, which is remarkable.
Generalizability or Applicability	Sentence claims that findings are generalizable or warrant use in clinical contexts. • These methods could be of benefit to centres at which thoracic imaging expertise is scarce, as well as for stratification of patients in clinical trials. • Collectively, the current system may have capabilities for screening purposes in general medical practice, particularly because it requires only a single clinical image for classification.
Regulatory Standards	Sentence asserts that findings meet standards for regulatory approval (e.g., superior or noninferior). • The three-dimensional convolutional neural network described in this article demonstrated both high sensitivity and high specificity in classifying pulmonary nodules regardless of diameters as well as superiority compared with manual assessment. • The sensitivity of the DL algorithm for diagnosing ONFH using digital radiography was noninferior to that of both less experienced and experienced radiologist assessments.

Initial annotation was completed on a subsample of 288 sentences. Annotations were applied independently by each annotator and inter-rater reliability was assessed using Cohen's kappa. Initial inter-rater agreement was 94.1% ($\kappa = 0.812$). The two annotators used these results to conduct an additional norming exercise where they discussed points of disagreement. They then re-annotated the original 288 sentences independently, along with the remaining 634 sentences to create the final dataset. Final interrater reliability was almost perfect ($\kappa = 0.825$ with 94.5% agreement) according to previously published guidelines for qualitative interpretation of Cohen's κ.[24] The few remaining annotation disagreements were resolved by a third annotator prior to model training.

We extracted relevant features from the annotated sentences and assessed several competing approaches to training in order to develop the final model. All feature extraction and training was

implemented in R version 4.0.2, although some feature extraction techniques made use of python virtual environments via the reticulate library.[25] All feature extraction and modeling was performed on a desktop workstation (Dell G5, core i3-9300, 16 GB RAM). Part-of-speech (POS) average location (aveloc) was used for primary feature extraction.[26] POS aveloc uses spaCyr to classify each word according to POS type and identifies the average location within each sentence.[27] The method was designed to provide a purely syntactic approach to feature engineering in a computational social science context.[26] For POS aveloc to work well, it generally requires cases where (1) the unit of analysis is a sentence, (2) the collected sentences are fairly homogeneous in terms of content, and (3) the label of interest is a discursive or linguistic strategy/technique. POS aveloc feature engineering occurs in two steps: (1) The text of interest is parsed to identify parts-of-speech, and (2) the average location (within each sentence) of each part of speech is identified. POS aveloc feature engineering was augmented with a sentence order variable that resulted in moderate increases in accuracy. Candidate models included k-nearest neighbor (KNN), Bagged classification and regression trees (CART), naïve bayes, and neural network (NNet). All training was implemented in caret using a random 80/20 train/test split and 10-fold cross-validation.[28] The bagged CART models had the highest accuracy across metrics. Performance, recall, and AUC measures are available in Table 3, and the ROC curves for the final models are available in Figure 2.

Table 3. Precision, recall, and AUC for candidate models.

Model	Precision	Recall	AUC
Bagged CART	0.92	0.89	0.8947
NNet	0.86	0.85	0.7685
KNN	0.84	0.96	0.7501
Naive Bayes	0.8	0.944	0.7614

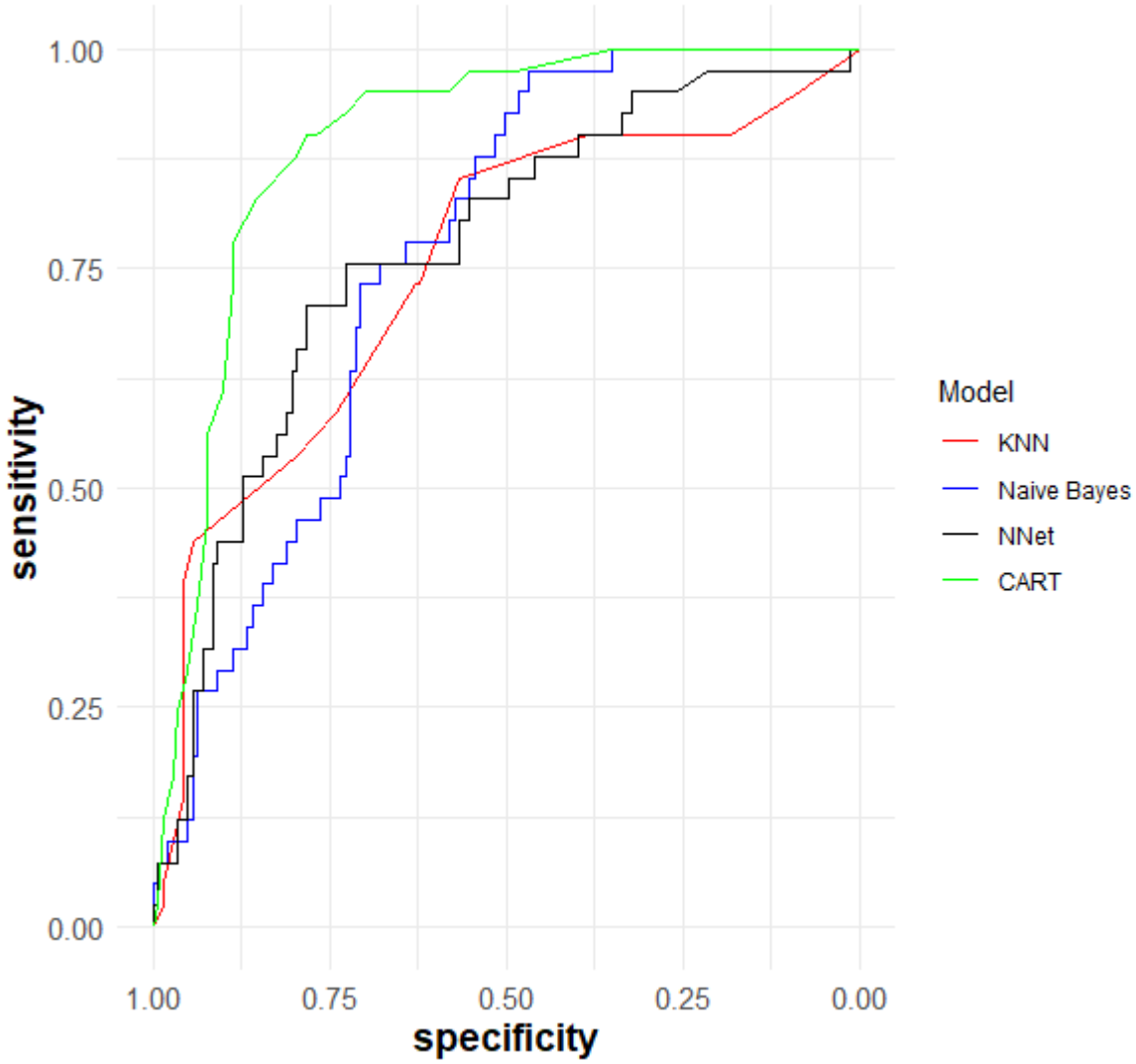

Fig. 2. ROC curves for candidate models. Figure shows the differential performance of KNN (red), Naive Bayes (blue), Neural Net (black), and bagged CART (Green) classification models. ROC curves were generated with the pROC R library.[29] The bagged CART model was the most performant with a total area under the curve of 0.8947.

Results

The 1200 studies included in this study focused on the development or validation of health AI systems in a wide variety of clinical areas. The studies were all published between 2010 and 2021 in 421

distinct journals. The most commonly represented journals included *Scientific Reports* (58 articles), *PLOS One* (56 articles), *Computers in Biology and Medicine* (36 articles), *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (31 articles), *Sensors* (27 articles), *Computer Methods and Programs in Biomedicine* (23 articles), *European Radiology* (23 articles), *BMC Bioinformatics* (18 articles), *BMC Medical Informatics and Decision Making* (15 articles), and the *International Journal of Medical Informatics* (13 articles).

Each evaluated abstract reported between 1 and 5 metrics, with an average of 1.66. The number of sentences identified as containing promotional language ranged from 0 to 12, with an average of 2.67. Journal conventions for abstract length vary widely with unstructured abstracts often being as few as five sentences and structured clinical abstracts sometimes being as many as 20. Therefore, we used the promotional language flagged sentences to determine the percentage of promotional sentences in each abstract. The proportion of promotional claims in each abstract ranged from 0 to 60% with an average of 22.85%, slightly higher than the 20% norm identified in linguistic studies of biomedical abstracts.[30] Table 3 provides additional details regarding distribution of promotional statements and number of metrics per abstract.

Table 3: Distribution of promotional statements and number of metrics.

	Low	Mean (SD)	High
Promotional Statements (N)	0	2.67 (1.89)	12
Promotional Statements (%)	0	22.85 (13.1)	60.0
Number of Metrics	1	1.66 (0.79)	5

To evaluate the relationship between use of AUC, use of F1, total number of metrics, and the proportion of promotional sentences, we used a quasi-Poisson framework. While both use of AUC and F1 proved to be significant predictors of promotional language use ($p = 0.00149$ and $p = 0.01564$), the number of metrics used was not a significant predictor ($p = 0.53216$), So we removed the number of metrics variable from the model. The final model predicts a 12% increase (95% CI: 5-19%, $p = 0.00104$)

in the promotional language rates for abstracts that report AUC and a 16% increase (95% CI: 4% to 30%, p = 0. 00996) for abstracts that use F1. See fig. 3 for additional details.

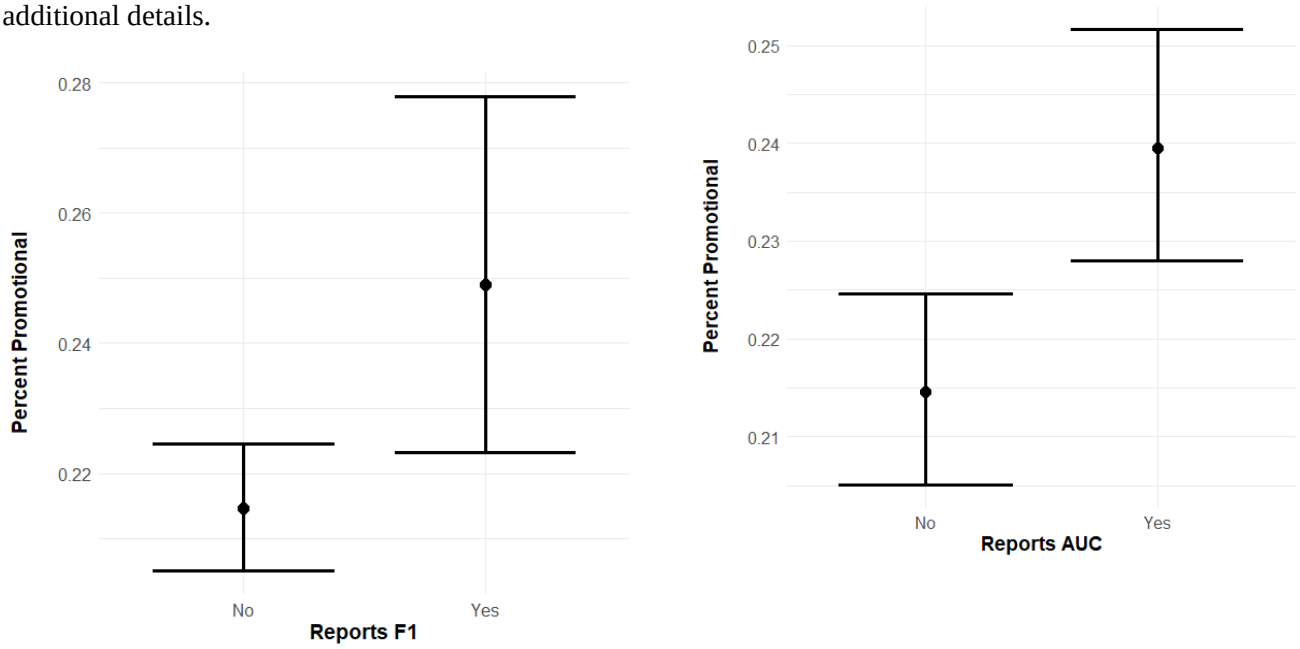

Fig. 3. Promotional language rates by F1 (left) and AUC (right) usage. Figure details the incident rate ratios and 95% confidence intervals for the proportion of promotional language in each abstract by composite metric usage.

Discussion

While the overall observed magnitude of the increase is not large, it is consistent with promotional language differentials seen in randomized clinical trials of researcher and clinician responses to spin or overstatement. In one study, a single sentence of overstatement was enough to prompt improper assessment of research results in 44.4% of cases.[31] Importantly, the effect of an overstated sentence was not the same across reader demographics. While clinicians who had led research projects were less likely to accept a face-value reading of an overstated claim, time since graduation for practicing clinicians predicted a greater likelihood of accepting overstated claims. Modest increases in promotional language have also been shown to lead researchers and clinicians to be more likely to evaluate new treatments as potentially beneficial to patients, even though the presence of spin led researchers to rate studies as less methodologically rigorous.[32]

Advances in machine learning and AI have the potential to substantially improve biomedical research and clinical practice. However, the adoption of new AI innovations that do not live up to the promises made in research reports can lead to both adverse events for patients and overall distrust in the potential benefits of clinical AI. Excessive promotional language, hype, or spin helps to create the conditions for these adverse outcomes. Efforts to study and address promotional language use in scientific and biomedical research tend to focus on hype in popular media.[33-36] Within this framework, much of the research on hype or overstatement evaluates mismatches between the underlying research and the presentation of findings in press releases and news articles. Similar research also evaluates inconsistencies between research results and the presentation of findings in abstracts or published articles.[20-21] These common methodologies for addressing hype assume a readily identifiable division between the underlying research and the linguistic presentation of results. The results presented in this article indicate that some methods themselves may lead to measurable increases in promotional language. Subsequently, these findings suggest that efforts to address hype in health AI need to attend to both underlying research methods and language choice in the presentation of findings. Given the established threats to clinical utility and the results of this study that indicate use of composite performance metrics can increase promotional language rates, health AI researchers and editorial boards may wish to reconsider ideal reporting practices in these areas. While composite metrics are quite useful when it comes to comparing the performance of candidate models within a study, authors and editors should be on guard against the attendant risks of increased promotional language that comes with the use of these metrics.

Acknowledgments

S.S.G. and T.G. thank Vivian Tran for her assistance with promotional language annotations.

References

1. Topol E. Deep medicine: how artificial intelligence can make healthcare human again. Hachette UK; 2019.
2. Gennatas ED, Chen JH. Artificial intelligence in medicine: past, present, and future. InArtificial Intelligence in Medicine 2021 Jan 1 (pp. 3-18). Academic Press.

3. Li X, Zhang S, Zhang Q, Wei X, Pan Y, Zhao J, Xin X, Qin C, Wang X, Li J, Yang F. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. The Lancet Oncology. 2019 Feb 1;20(2):193-201.
4. Romero-Brufau S, Gaines K, Nicolas CT, Johnson MG, Hickman J, Huddleston JM. The fifth vital sign? Nurse worry predicts inpatient deterioration within 24 hours. JAMIA open. 2019 Dec;2(4):465-70.
5. Ashrafian H, Darzi A. Transforming health policy through machine learning. PLoS Medicine. 2018 Nov 13;15(11):e1002692.
6. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. Nature Medicine. 2021 Jan;27(1):136-40.
7. Micca P, Chang C, Shukla M, Gisby S. Trends in health tech investments: funding the future of health. Deloitte Insights. 2021.
8. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. PLoS medicine. 2018 Nov 6;15(11):e1002689.
9. Saria S, Butte A, Sheikh A. Better medicine through machine learning: What's real, and what's artificial. PLoS Med. 2018;15(12):e1002721.
10. Car J, Sheikh A, Wicks P, Williams MS. Beyond the hype of big data and artificial intelligence: building foundations for knowledge and wisdom. BMC Medicine. 2019 Jul 17;17(1):143.
11. Toh TS, Dondelinger F, Wang D. Looking beyond the hype: Applied AI and machine learning in translational medicine. EBioMedicine. 2019 Sep 1;47:607-15.
12. Bhattacharya S, Pradhan KB, Bashar MA, Tripathi S, Semwal J, Marzo RR, Bhattacharya S, Singh A. Artificial intelligence enabled healthcare: A hype, hope or harm. Journal of family medicine and primary care. 2019 Nov;8(11):3461.
13. Matheny M, Israni ST, Ahmed M, Whicher D. Artificial intelligence in health care: the hope, the hype, the promise, the peril. NAM Special Publication. Washington, DC: National Academy of Medicine. 2019.
14. McCradden MD, Stephenson EA, Anderson JA. Clinical research underlies ethical integration of healthcare artificial intelligence. Nature Medicine. 2020 Sep;26(9):1325-6.
15. Taylor M, Liu X, Denniston A, Esteva A, Ko J, Daneshjou R, Chan AW. Raising the Bar for Randomized Trials Involving Artificial Intelligence: The SPIRIT-Artificial Intelligence and CONSORT-Artificial Intelligence Guidelines. The Journal of investigative dermatology. 2021 Sep;141(9):2109-11.
16. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, Arnaout R, Kohane IS, Saria S, Topol E, Obermeyer Z. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nature medicine. 2020 Sep;26(9):1320-4.
17. Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. Public Health Reports (1896-1970). 1947 Oct 3:1432-49.
18. Carter JV, Pan J, Rai SN, Galandiuk S. ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. Surgery. 2016 Jun 1;159(6):1638-45.
19. Mallett S, Halligan S, Thompson M, Collins GS, Altman DG. Interpreting diagnostic accuracy studies for patient care. Bmj. 2012 Jul 2;345.
20. Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. European radiology. 2015 Apr;25(4):932-9.
21. Signorini DF. Sample size for Poisson regression. Biometrika. 1991 Jun 1;78(2):446-50.

22. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Bmj. 2021 Mar 29;372.
23. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, Ledsam JR. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The lancet digital health. 2019 Oct 1;1(6):e271-97.
24. McHugh ML. interrater reliability: the kappa statistic. Biochemica Medica, 22 (3), 276–282.
25. Ushey K, Allaire JJ, Tang Y. reticulate: interface to 'Python'. R package version 1.18-9006. 2021.
26. Graham SS, Hopkins HR. AI for Social Justice: New Methodological Horizons in Technical Communication. Technical Communication Quarterly. 2021 Aug 6:1-4.
27. Benoit K, Matsuo A. spacyr: Wrapper to the 'spaCy' 'NLP' library. R package version 1.2.1. 2020.
28. Kuhn M. caret: classification and regression training. R package version 6.0-86. 2020.
29. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics. 2011 Dec;12(1):1-8.
30. Swales JM, Feak CB. Abstracts and the writing of abstracts. University of Michigan Press ELT; 2009.
31. Tsujimoto Y, Aoki T, Shinohara K, So R, Suganuma AM, Kimachi M, Yamamoto Y, Furukawa TA. Physician characteristics associated with proper assessment of overstated conclusions in research abstracts: A secondary analysis of a randomized controlled trial. PloS one. 2019 Jan 25;14(1):e0211206.
32. Boutron I, Altman DG, Hopewell S, Vera-Badillo F, Tannock I, Ravaud P. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. Journal of Clinical Oncology. 2014 Dec 20;32(36):4120-6.
33. Lynch J, Bennett D, Luntz A, Toy C, VanBenschoten E. Bridging science and journalism: Identifying the role of public relations in the construction and circulation of stem cell research among laypeople. Science Communication. 2014 Aug;36(4):479-501.
34. Li Y, Zhang J, Yu B. An nlp analysis of exaggerated claims in science news. InProceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism 2017 Sep (pp. 106-111).
35. Yu B, Wang J, Guo L, Li Y. Measuring Correlation-to-Causation Exaggeration in Press Releases. In Proceedings of the 28th International Conference on Computational Linguistics 2020 Dec (pp. 4860-4872).
36. Patro J, Baruah S. A Simple Three-Step Approach for the Automatic Detection of Exaggerated Statements in Health Science News. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume 2021 Apr (pp. 3293-3305).

Competing Interests

The authors declare that there are no competing interests.